Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Graduate School
Humanities

Genre Analysis and Corpus Design: Nineteenth Century Spanish-American Novels (1830–1910)

Inaugural-Dissertation
zur Erlangung der Doktorwürde der
Graduiertenschule für die Geisteswissenschaften /
Graduate School of the Humanities (GSH)
der
Julius-Maximilians-Universität Würzburg

vorgelegt von
Ulrike Henny-Krahmer
aus Rostock

Würzburg
2023

This document is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0):
http://creativecommons.org/licenses/by/4.0 This CC license does not apply to third party material (attributed to another source) in this publication.

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Graduate School
Humanities

Gutachter / Mitglieder des Promotionskomitees:

Vorsitz des Promotionsprüfungsverfahrens:
Professor Dr. Jörn Müller
Julius-Maximilians-Universität Würzburg, Fakultät für Humanwissenschaften

Gutachter und Erstbetreuer im Promotionskomitee:
Professor Dr. Christof Schöch
Universität Trier, Fachbereich Sprach-, Literatur- und Medienwissenschaften

Gutachter und Zweitbetreuer im Promotionskomitee:
Professor Dr. Fotis Jannidis
Julius-Maximilians-Universität Würzburg, Philosophische Fakultät

Zweitbetreuer im Promotionskomitee:
Professor Dr. Hanno Ehrlicher
Eberhard Karls Universität Tübingen, Philosophische Fakultät

Tag des Promotionskolloquiums: 7. Juni 2021

Lizenz: Creative Commons Attribution 4.0 (BY)
Satz: Bernhard Assmann, Ulrike Henny-Krahmer, LuaTeX

Acknowledgements

My thanks go to my three supervisors: Christof Schöch, Fotis Jannidis, and Hanno Ehrlicher, who supported me with their professional ideas, hints, and feedback during the creation of this thesis, from the first beginnings of the definition of the topic to the guidance during the working process and the finalization of the text. This work has involved much more than reading, thinking, excerpting, structuring, and writing. The project also consisted of collecting, modeling, managing, analyzing, visualizing, and publishing data, working with digital tools and programming. My supervisors assisted and encouraged me in all these activities and are role models for me to follow established as well as new paths in academic work and in the Digital Humanities.

This work on the computational analysis of subgenres of nineteenth-century Spanish-American novels was undertaken within the framework of the BMBF-funded project "Computational Literary Genre Stylistics" (CLiGS) at the Chair of Computational Philology and Newer German Literary History of the University of Würzburg. I am grateful for the fact that I was able to concentrate on my dissertation by working in the project and for the exciting and instructive time and cooperation with the other staff members in CLiGS and at the chair. I very much hope that our paths will continue to cross. I would like to thank, in particular, my co-doctoral students José Calvo Tello, Daniel Schlör, and Stefanie Popp, as well as the project members Robert Hesselbach, Katrin Betz, and Steffen Pielström. For their support in the preparation of my corpus of novels, I would like to thank the student assistants Constanze Ludewig and Jakob Stahl. The CLiGS project has advanced computational genre analysis with corpora in Romance languages, in particular, and I am glad to have been a part of this initiative.

Further supporters have made the elaboration, completion, and publication of this dissertation possible: Thank you to the Ibero-American Institute (IAI) in Berlin for their support in the digitization of several of the Spanish-American novels, which are now part of the IAI's digital library and of the Conha19 corpus that I created as part of this dissertation project. I would also like to thank Thomas Schmid and the Graduate School of the Humanities (GSH) of the University of Würzburg, which not only guided me through the administrative process of doctoral studies but also accompanied the steady progress of the project and opened up additional qualification opportunities. I thank DARIAH-EU for creating the Open Access Monograph Bursary for Early Career Researchers in Digital Humanities. I am very honored to be the first one to receive it. It encouraged me to publish the book immediately in Open Access and, above all, to pursue the approach of linking research data, program code, and text even more consistently. It is only because of the DARIAH bursary that there are now also TEI and HTML versions of this dissertation.

For proofreading and valuable advice on content and form, I thank my supervisors, the members of the Institute for Documentology and Scholarly Editing (IDE), Sean Winslow from the University of Graz and Rebecca Collin from the Academic Writing Consultancy at the University of Rostock. Thank you to Bernhard Assmann for his support in preparing the PDF version of this monograph and to Christopher Pollin for helping me with HTML and CSS complexities. Thank you to the Ehrenfelder Musikschule and the Urania Theater in Cologne, where I was able to sit, work and write during the pandemic when all libraries were closed.

I would also like to thank those who marked my professional path before my doctoral studies. These are, above all, the staff of the Cologne Center for eHumanities (CCeH) at the University of Cologne, where my work as a digital humanities researcher began and where I later, as a guest, always had a place in an office. I also thank the members of the IDE, which I joined in 2012. The IDE members are colleagues and friends, a group that, regardless of age, locality, or institutional affiliation, advances the topic of digital documentology and editing through joint activities. I would like to thank especially Frederike Neuber, Martina Scholger, Patrick Sahle, and Franz Fischer, with whom I have collaborated the most in the past and present.

Looking back even further, I thank my host families, classmates, and friends in Mexico who welcomed me during my year abroad there in 1999 and 2000, through whom I learned Spanish and developed a desire to further engage with their language, culture, and literature. Without them, the topic of this thesis would certainly be very different – or this thesis would not exist at all.

Finally, my thanks go to my friends and family. Thanks to my parents for letting me go out into the world and do what I was interested in, even to the far away places and on the uncertain paths. Thanks to my parents-in-law for taking care of me and having my back when it matters. Thank you, Constantin and Ivo, for bearing with me when I work, for distracting me, as well, and for being there. I dedicate this work to you.

Genre Analysis and Corpus Design: Nineteenth-Century Spanish-American Novels (1830–1910)

Summary

This work in the field of digital literary stylistics and computational literary studies is concerned with theoretical concerns of literary genre, with the design of a corpus of nineteenth-century Spanish-American novels, and with its empirical analysis in terms of subgenres of the novel. The digital text corpus consists of 256 Argentine, Cuban, and Mexican novels from the period between 1830 and 1910. It has been created with the goal to analyze thematic subgenres and literary currents that were represented in numerous novels in the nineteenth century by means of computational text categorization methods. The texts have been gathered from different sources, encoded in the standard of the Text Encoding Initiative (TEI), and enriched with detailed bibliographic and subgenre-related metadata, as well as with structural information.

To categorize the texts, statistical classification and a family resemblance analysis relying on network analysis are used with the aim to examine how the subgenres, which are understood as communicative, conventional phenomena, can be captured on the stylistic, textual level of the novels that participate in them. The result is that both thematic subgenres and literary currents are textually coherent to degrees of 70–90 %, depending on the individual subgenre constellation, meaning that the communicatively established subgenre classifications can be accurately captured to this extent in terms of textually defined classes.

Besides the empirical focus, the dissertation also aims to relate literary theoretical genre concepts to the ones used in digital genre stylistics and computational literary studies as subfields of digital humanities. It is argued that literary text types, conventional literary genres, and textual literary genres should be distinguished on a theoretical level to improve the conceptualization of genre for digital text analysis.

Análisis de género y diseño de corpus: Novelas hispanoamericanas del siglo XIX (1830–1910)

Resumen

Este trabajo en el campo de la estilística literaria digital y los estudios literarios computacionales se ocupa de las preocupaciones teóricas del género literario, del diseño de un corpus de novelas hispanoamericanas del siglo XIX y de su análisis empírico en términos de subgéneros de la novela. El corpus de textos digitales consta de 256 novelas argentinas, cubanas y mexicanas del período comprendido entre 1830 y 1910. Ha sido creado con el objetivo de analizar los subgéneros temáticos y las corrientes literarias que estaban representadas en numerosas novelas del siglo XIX mediante métodos de categorización computacional de textos. Los textos han sido recogidos de diferentes fuentes, codificados en el estándar de la Iniciativa de Codificación de Textos (TEI), y enriquecidos con detallados metadatos bibliográficos y de subgéneros, así como con información estructural.

Para la categorización de los textos se utiliza una clasificación estadística y un análisis de semejanza familiar basado en el análisis de redes, con el fin de examinar cómo los subgéneros, entendidos como fenómenos comunicativos y convencionales, pueden ser captados en el plano estilístico y textual de las novelas que participan en ellos. El resultado es que tanto los subgéneros temáticos como las corrientes literarias son textualmente coherentes en grados del 70–90 %, dependiendo de la constelación individual de subgéneros, lo que significa que las clasificaciones de subgéneros establecidas comunicativamente pueden ser capturadas con precisición hasta este punto en términos de clases textualmente definidas.

Además del enfoque empírico, la disertación también pretende relacionar los conceptos teóricos de género literario con los utilizados en la estilística de género digital y los estudios literarios computacionales como subcampos de las humanidades digitales. Se argumenta que los tipos de texto literario, los géneros literarios convencionales y los géneros literarios textuales deberían distinguirse a nivel teórico para mejorar la conceptualización del género para el análisis de textos digitales.

Gattungsanalyse und Korpusaufbau: Hispanoamerikanische Romane im 19. Jahrhundert (1830–1910)

Zusammenfassung

Diese Arbeit ist in den Forschungsfeldern der digitalen literaturwissenschaftlichen Stilistik und der Computational Literary Studies angesiedelt und setzt sich mit theoretischen Gattungsproblemen, mit der Erstellung eines Korpus von hispanoamerikanischen Romanen des 19. Jahrhunderts und mit ihrer empirischen Analyse nach Untergattungen auseinander. Das digitale Textkorpus umfasst 256 argentinische, kubanische und mexikanische Romane aus der Zeit von 1830 bis 1910 und ist mit dem Ziel erstellt worden, thematische Untergattungen und literarische Strömungen, die im 19. Jahrhundert durch zahlreiche Romane repräsentiert waren, mit Hilfe computergestützter Methoden der Textkategorisierung zu analysieren. Die Texte wurden aus verschiedenen Quellen zusammengetragen und gemäß dem Standard der Text Encoding Initiative (TEI) codiert, wobei die Dokumente mit detaillierten bibliographischen und untergattungsbezogenen Metadaten sowie mit textstrukturellen Informationen angereichert wurden.

Um die Texte zu kategorisieren werden Verfahren der statistischen Klassifikation und eine Familienähnlichkeitsanalyse verwendet, die auf einer Netzwerkanalyse basiert. Das Ziel der Analysen ist es zu untersuchen inwieweit die Untergattungen, die primär als Phänomene der Kommunikation und Konvention verstanden werden, auf der stilistischen, textlichen Ebene der Romane, die an ihnen teilhaben, erfasst werden können. Das Ergebnis ist, dass sowohl die thematischen Untergattungen als auch die literarischen Strömungen zu 70–90 % textlich kohärent sind, in Abhängigkeit der gewählten Untergattungskonstellation, womit gemeint ist, dass die kommunikativ etablierten Untergattungsklassifikationen in diesem Maß an Genauigkeit auch als textlich definierte Klassen erfasst werden können.

Über die empirische Ausrichtung hinaus ist ein weiteres Ziel der Dissertation, literaturtheoretische Gattungskonzepte zu denjenigen in Beziehung zu setzen, die in der digitalen Gattungsstilistik als einer Teildisziplin der Digital Humanities verwendet werden. Es wird argumentiert, dass literarische Texttypen, konventionelle literarische Gattungen und textliche literarische Gattungen auf einer theoretischen Ebene unterschieden werden sollten, um die Konzeption von Gattung für die digitale Textanalyse zu verbessern.

Contents

1 Introduction

If people are asked about the objects, beings, or events around them, they will most probably name the categories that the things belong to, for example, a book on a shelf, a bird in a tree, a dancer on stage, or a thunderstorm in the sky. Research in cognitive development has shown that even small babies begin to recognize what is around them in terms of categories when they are about a year old (Gopnik, Meltzoff, and Kuhl 2009, 79–83). Paradoxically, however, all objects and beings are unique: "All you ever see are individual objects: this particular sweet pea, this individual dollar bill. There is no 'sweet-peaness' or 'dollarhood' in the world. So how could it ever be informative to say that this individual thing belongs to this nonexistent, mythical category, when the individual thing itself is all we ever actually experience?" (Gopnik, Meltzoff, and Kuhl 2009, 79). Categorizing serves a basic need of humans to confer meaning to what they perceive and to leave aside the individuality of things. It helps them to grasp the world around them. However, the perception of the individual is also dependent on the understanding of the general, so that what is special about something emerges from the background of the familiar.

Literary texts are no exception. One type of category that they are commonly associated with is genre. A poem is something different than a drama, and a science fiction novel is not to be confused with a sentimental one. One comes across literary genres in everyday life, for example, as an organizing principle in a bookstore, in a library, or on the covers of the books themselves. Experiences in daily life usually suggest that the assignment of individual texts to genres does not cause particular problems, only that one might be disappointed, surprised, or impressed when the book bought is different than expected by its genre label. In literary theory and its antecedents, however, the "genre problem" has been discussed intensely for thousands of years, starting with the attempt of Aristotle and Plato to formulate a theory of poetry (Zymner 2003, 10). Some of the main questions in the debates about genre are as what type of category they can be conceived: as logical classes with clear boundaries into which all the literary works can neatly be sorted? As prototypical categories with exemplary masterpieces at the center and mediocre imitations at the edge? As networks of related texts that form generic families? Or as some other kind of category that can be described as a combination of necessary and optional features?[1] Moreover, it has been debated if genres can be assumed to exist at all beyond pure naming conventions, given that the literary works associated with them can be so different. This is connected to the problem of genre change and also dependence on the cultural context because literary historians must deal with the phenomenon that the same names of genres are applied to phenomena with quite distinct textual characteristics across time and place. At times it has been tried to avoid the challenges that genres as categories of literary texts pose by denying their relevance altogether (for example, by Croce 1905). However, the practical relevance that genres have not only in daily life but also for students of literature and literary scholars cannot be denied. Topics of courses and exams are often defined in terms of genres, for example, a seminar on the Spanish picaresque novel or classic drama. Literary histories are also often structured in terms of genres that were important for certain periods. Finally, the interpretation of individual literary works does not happen in a vacuum. In order to assess the value of single texts, they are often examined with

[1] For an overview of the categorization aspect of genres, see Zymner (2003, 99–104).

regard to a specific literary tradition or genre (Keckeis and Michler 2020, 7–8). As a way to approach the genre problem theoretically, there is a tendency in recent literary genre theory to see the phenomenon as one that can be described in different dimensions that are linked to each other in cognitive, communicative, social, and textual dimensions (Gymnich and Neumann 2007).

This dissertation aims to enter the theoretical discussion about genres from an interdisciplinary perspective. It is located in the field of digital literary stylistics, which is part of the wider discipline of digital humanities, in which humanities research is combined with methods from information science and computer science, and which includes interdisciplinary disciplines such as computational linguistics, computational philology, or computational literary studies. The subfield of digital stylistics is concerned with the analysis of linguistic and literary style with computational methods. An important subject is the investigation of the style of individual authors, but genre style has also been the focus of digital stylists.[2] To examine genre on the level of style means that the approach is primarily text-centered, and it also entails empirical work.

Digital literary stylistics is not exclusively but predominantly applied research. The basis for it are digital corpora of literary texts, which are designed for a specific language, period, set of authors, or genre, or combinations of several ones of them if the aim is a contrastive analysis. The topic of this dissertation is, therefore, genre analysis and corpus design, both as a theoretical discussion of genre as a concept in literary theory and digital stylistics and as an empirical corpus study. A specific corpus was built for this purpose as a basis for an analysis of metadata and texts in terms of genre. The genres that the empirical part of this study is concerned with are the novel and its subgenres in the context of nineteenth-century Spanish-American literature, more precisely Argentine, Cuban, and Mexican novels that were published between 1830 and 1910. There is a growing number of digital stylistic studies concerned with texts in Romance languages, as the contributions to the conference "Digital Stylistics in Romance Studies and Beyond", which took place at the University of Würzburg in 2019, show.[3] Nevertheless, most of the digital stylistic studies on literary texts are still based on corpora of texts in English.[4] Comprehensive central repositories of digital literary texts, which are curated following scholarly standards, such as the Digital Library in the TextGrid Repository (TextGrid n.d.) or the German Text Archive (Berlin-Brandenburgische Akademie der Wissenschaften 2022) for German texts, are not yet available for Spanish literary works. There are, however, initiatives that also promote the building of corpora of literary texts in Spanish, many of which go back to individual work, community initiatives, or research projects, for example, the "Corpus of Spanish Golden-Age Sonnets" (Navarro-Colorado, Ribes Lafoz, and Sánchez 2016; Navarro-Colorado 2020) or the multi-language corpora DraCor (Fischer et al. 2019, n.d.) and ELTEC (Odebrecht, Burnard, and Schöch 2021), which include Spanish drama and novels, respectively. In addition, the project "Computational Literary Genre Stylistics" (CLiGS), the context in which this dissertation was written, was concerned with building and analyzing digital corpora of literary texts in French,

---

[2] For an introduction to the background and goals of digital literary stylistics, see the website (SIG-DLS n.d.) of the corresponding special interest group of the Alliance of Digital Humanities Organizations (ADHO).

[3] See the call for papers (CLiGS n.d.) and the conference proceedings to be published in 2023 (Hesselbach et al., forthcoming).

[4] See, for instance, the influential studies of Jockers (2013) and Underwood (2019).

Spanish, Italian, and also Portuguese.[5] Building such corpora is important for several reasons: it strengthens digital quantitative research in the respective disciplines and language areas, and it helps to make the empirical results of digital stylistics in general more reliable if they are based on findings derived from a broad range of different corpora and if they are not specific for certain languages or genres.

The Spanish-American nineteenth-century novel is well studied, and knowledge about it is consolidated in literary histories and monographs.[6] Various subgenres of this novel have also been analyzed in depth by literary scholars, for instance, the historical novel, the anti-slavery novel, or novels of the romantic, naturalistic, and modernist currents (Löfquist 1995; Peñaranda Medina 1994; Read 1939; Rivas 1990; Schlickers 2003; Suárez-Murias 1963). However, nineteenth-century Spanish-American novels and their subgenres have not yet been analyzed on the basis of a comprehensive digital text corpus and by means of stylistic computational methods. There are several reasons why a quantitative digital analysis of the subgenres of that novel is of interest. First, many studies on nineteenth-century Spanish-American novels focus on selected works that have a canonical status. The effect is that only a specific section of the whole literary production of the time forms the basis of the literary-historical knowledge about the novel and its subgenres in that period.[7] There is also qualitative research based on larger corpora, but in these cases, scholars mostly concentrate either on the novel as a whole or on one specific subgenre.[8]

A digital approach in which several subgenres of the novel are contrasted can contribute new insights into the characteristics of the texts that are distinctive for the different subgenres. Moreover, more novels can be taken into account if a comparatively large corpus is used – not only the well-known novels but also works that have thus far not received much critical attention. This can shed new light on the concepts of the subgenres, on the one side because the quantitative relevance of the subgenres becomes clearer, and on the other side because lesser-known works possibly represent the subgenres that they are associated with in a different way, by use of other textual traits and stylistic means. Third, a quantitative digital study is different from a qualitative approach even if the same number of novels and subgenres would be analyzed because the way knowledge about the texts is extracted and summarized is not directly dependent on the human reader but results from a mechanical treatment of the texts

---

[5] One outcome of the project is the Textbox, a collection of small to medium-sized corpora of literary texts in Romance languages of different genres, which are published on GitHub and free to reuse (Schöch, Calvo Tello et al. 2018, 2019). Beyond the Textbox, the following more extensive individual corpora resulting from the CLiGS project are worth mentioning: the "Corpus of Novels of the Spanish Silver Age" (CoNSSA, Calvo Tello 2021a) and a text collection of over 800 French dramatic texts (Schöch 2017b) derived from the corpus Théâtre Classique (Fièvre 2007–2022). The latter is also available as part of the multilingual DraCor corpus, where it is called FreDraCor (Milling, Fischer, and Göbel 2021).

[6] For general literary histories on Spanish-American literature that also cover the nineteenth-century novel and for specialized monographs, see, among others, Alegría (1959), Anderson Imbert (1954), Dill (1999), Gálvez (1990), Goić (2009), Íñigo Madrigal, Alvar, and Aínsa (1982), Lindstrom (2004), Rössner (2007), and Sánchez (1953).

[7] Rivas (1990), for instance, establishes the concept of the anti-slavery novel based on seven different novels. Gnutzmann (1998) as well studies the Argentine naturalistic novel with a corpus of seven texts.

[8] For example, Löfquist (1995) on the Chilean historical novel, Read (1939) on the Mexican historical novel, or Schlickers (Schlickers 2003) on the Spanish-American naturalistic novel. Another approach is to consider the novel as a whole for an individual country and for a certain period. Lichtblau (1959), for example, studies the nineteenth-century novel in Argentina, and Molina (2011) the Argentine novel between 1838 and 1872.

and computational processing. This can produce new findings about the subgenres that remain unrecognized by close reading methods. Even if the nineteenth century is past, the literature of that time is still of importance because that century marked the rise of the novel as a genre and the beginning of the national literatures of the different Spanish-American countries. Many of the subgenres that were practiced or emerged in the nineteenth century are still relevant in twenty-first-century literature, such as historical or crime novels. For digital genre stylistics in general, the subgenres of the Spanish-American novels are an interesting empirical case because they combined generic concepts of a European origin with specific local inventions. Especially regarding literary currents, a neat chronological succession is not given so that several currents were en vogue at once (on these aspects, see, for instance, Varela Jácome [1982] 2000). It is an interesting question to what extent different theoretical concepts of genre categories are suitable to capture the various nineteenth-century Spanish-American subgenres.

Although this dissertation is concerned with texts in Spanish, it is written in English because the field of digital genre stylistics and digital humanities in general is highly interdisciplinary. The aim is to provide results that can be appreciated by scholars of Spanish-American literature but also by digital humanists from around the world. A second linguistic and cultural background of this thesis is German, and much research literature from German-speaking countries has been taken into account, especially literature on genre theory but also digital stylistics papers and research on the Spanish-American novel. In general, quotes are not translated, assuming that the context provides enough information to grasp their meaning.

Before the specific goals and questions of this thesis and its structure are outlined, it must be clarified what is not covered here. The period that is covered is 1830 to 1910, the whole long nineteenth century, but the corpus of novels that is analyzed is treated as a synchronic one. In the discussion of results, the publication date of the novels has been taken into account to see if that had an influence on the results, but no inherently diachronic analysis of the subgenres is pursued here. A second aspect that is not addressed fully at the moment of the publication of this thesis is the one of sustainable research data management in connection with the publication of the corpus. Its basic publication strategy is presented, and it is published in Open Access and in standard formats in a public code repository on GitHub and Zenodo, including versioning, but the publication method is not discussed explicitly in relationship to the FAIR principles (Wilkinson et al. 2016) or other best practices for research data publication. In the longer term, it is planned to prepare the corpus for long-term preservation and accessibility in suitable institutional or subject-specific repositories, but in the context of this dissertation, the initial focus was on its creation, analysis, and basic availability for transparency and re-use.

As mentioned above, the main goals of this dissertation are firstly in the area of genre theory, secondly in the construction of a digital corpus of novels, and thirdly in its computer-assisted analysis. The theoretical foundations of the thesis are clarified in chapter 2, “Concepts”. On the level of genre theory (chapter 2.1, “Literary Genres”), the aim is to work out which of the existing concepts of the ontological status of genres (chapter 2.1.2, “Ontological Status and Relevance of Genres”) and their historical and theoretical nature (chapter 2.1.3, “System and History”) are relevant and applicable and how these concepts need to be adapted for digital genre stylistics. In this context, three aspects are specifically addressed. The first aspect is the question of how generic terms can be modeled and defined. This is an important issue because genre labels are the

main feature through which genre conventions enter a digital stylistic text analysis. This aspect is deepened in chapters 2.1.2.1 (“Semiotic Models of Genres”) and 2.1.2.2 (“Genres and Digital Genre Stylistics: The Roles of Corpora, Genre Labels, Features, and Text Style”). Second, definitions of genre and text types stemming from literary theory and linguistics are compared to see to what degree they are suitable for digital stylistics. In particular, the question of how and whether a conventional or historical level of genre and a textual one should be separated is discussed. An own proposal is made for conceptual differentiation in chapter 2.1.3.1 (“A Conceptual Proposal for Digital Genre Stylistics: Literary Text Types, Conventional Literary Genres, and Textual Literary Genres”), building on existing approaches. Third, three main concepts that have been proposed to conceptualize genres as categories, namely logical classes, prototypical structures, and family resemblance networks, are related to the distinction between conventional and textual levels of genre (chapter 2.1.4, “Categorization”). It is then outlined how these three concepts can be implemented in text-based digital genre analyses by referring to computational methods of text classification, clustering, and network analysis. The theoretical part also explains the concept of literary style (chapter 2.2, “Style”) that underlies the analyses in the empirical part. Furthermore, the part on concepts is closed by a presentation of three major thematic subgenres and three literary currents chosen for text analysis (chapter 2.3, “Subgenres of the Nineteenth-Century Spanish-American Novel”). These are the historical novel, the sentimental novel, and the novel of customs as thematic subgenres, and the romantic novel, the realist novel, and the naturalistic novel as literary currents. Several hypotheses are formulated regarding the textual and stylistic characteristics and coherence expected for these subgenres and currents.

The empirical part of the work has two main parts: chapter 3, “Corpus”, and chapter 4, “Analysis”. The first main goal of this part has been to build up a comprehensive digital bibliography of Argentine, Mexican, and Cuban nineteenth-century novels and a corresponding digital corpus of 256 texts. Both have been elaborated as a prerequisite and basis for the text analysis of subgenres. The selection of novels from the three countries is motivated, and the diachronic limits of the bibliography and the corpus are clarified. General defining characteristics of the novel are discussed as a basis for the selection of works for both digital resources (chapter 3.1, “Selection Criteria”). A special focus is on how the subgenre labels were collected, modeled, and encoded (chapters 3.2.3 and 3.3.4, “Assignment of Subgenre Labels”, for the bibliography and the corpus, respectively). An empirically based adaptation of the semiotic models of Raible (1980) and Schaeffer (1983), which are also presented in the first chapter on genre theory, provides the theoretical foundation for the organization of the subgenre labels in the bibliography and the corpus. The preparation of the bibliography and corpus is explained in detail, including the availability and usage of bibliographical and full-text sources, the treatment of the extracted full-texts, the collection of metadata and text encoding, and the chosen publication strategy. Both resources are published on the web and offered to other scholars for reuse (Henny-Krahmer 2017–2021, 2021a).

The creation of the two collections of data and texts was primarily motivated by the goal of analyzing subgenres of Spanish-American nineteenth-century novels with quantitative methods. Therefore the selection of the materials was guided by the question about the specific subgenres that are the focus of interest here. However, the bibliography and the corpus also aim to provide a foundation for future analysis in other contexts. There are aspects of the corpus that are not employed in the analyses in this dissertation but are nonetheless presented as relevant for

the design of corpora for digital literary genre studies. Examples are chapter structures and paragraphs that were encoded in the corpus but not considered in the analysis. Another example is the separation of direct speech and narrated text, which was realized only for a part of the corpus and was analyzed only on a test basis. Such additional encoding prepares for future analyses beyond the scope of this dissertation. In addition, some structural units of the corpus have already been used for analyses in the CLiGS project, although they are not the focus of the work here.[9] The digital text corpus created here thus claims to go beyond limited, project-specific use. It aims to be a *community data collection* that can be used by different representatives of a research community, is suitable for addressing different questions from a specific research field, is comprehensive, follows discipline-specific standards, and is designed to be archived and reusable in the medium term (Schöch 2017a, 224; National Science Board 2005, 20–21).

Of the two resources, the bibliography constitutes the sampling frame for the novels in the corpus, which means that it represents the larger population of all the novels that were published between 1830 and 1910 in Argentina, Cuba, and Mexico. Of course, the bibliography does not contain information about all these works, as it cannot be known with certainty how many and which novels were published in that time, but it aspires to approximate that amount of novels. It is then possible to compare the novels contained in the bibliography to the ones in the corpus to see how representative the latter is for the novel and its subgenres of the chosen years and countries. This is done in the first part of the analysis chapter, in chapter 4.1, "Metadata Analysis". Not only the question of representativeness is tackled in that chapter, but also which subgenres on which discursive levels were quantitatively relevant. In addition, it is analyzed how the novels can be characterized by other parameters that have a possible impact on the analysis of genre style, for instance, the narrative perspective of the novels or the decades that they were published in. The metadata analysis chapter also provides a general overview of which authors and works are included in the bibliography and corpus and to which subgenres the works are assigned. This informs potential subsequent users of the resources in detail about their content and the distribution of the content in quantitative terms.

The second part of the analysis chapter, chapter 4.2, "Text Analysis", is concerned with the text analysis of the corpus of 256 novels. Two main types of stylistic features are employed in the analysis: most frequent words (MFW) and topics. In the first part of the text analysis chapter (4.2.1, "Features"), both types of features are presented and it is discussed how they relate to literary concepts of style and theme. In the second part of the analysis chapter (4.2.2, "Categorization"), the texts are categorized, first by statistical classification and then with a family resemblance network analysis as an alternative categorization approach. The novels are analyzed on two discursive levels of genre: thematic subgenres and literary currents. Only the subgenres and currents that are most relevant in quantitative terms are analyzed in this part. One goal of the text analysis is to show in empirical experiments how statistical classification and network analysis can be employed to analyze genres on the textual level in terms of different categorical concepts. Another goal is to find out if the conventionally, historically, and theoretically defined

[9] There are two studies based on subparts of the corpus in which the internal structure of the texts was exploited: Schöch, Henny et al. (2016) on the development of topics in different parts of the novels, depending on the subgenres, and Henny-Krahmer (2018) on the connection of sentiments and direct speech versus narrated text in different subgenres.

thematic subgenres and literary currents can be captured at all on the stylistic level of a group of texts, and if yes, how textually coherent the groups of novels associated with these subgenres are. In the classification setting (chapter 4.2.2.1, "Classification"), textual coherence means the degree to which the communicatively established subgenre classifications of the novels can be captured accurately in terms of textually defined classes, and it is measured in terms of classification accuracy. A further question is what can be learned about the subgenres and the individual texts from the errors that the classifier makes.

Besides the statistical classification approach, a family resemblance analysis (chapter 4.2.2.2, "Family Resemblance: Network Analysis") is pursued. While a classificatory approach assumes strict boundaries between the various groups of texts, in a network structure, the focus is on direct and indirect relationships between groups of novels, and the results are more open. In this context, the question of textual coherence refers to the extent to which textually based groups of novels in the network are also related to the same genre or subgenre of novels from a communicative perspective. In this case, coherence cannot simply be measured with an accuracy value but must be assessed by evaluating and interpreting the clusters found in the network. That way, the family resemblance network analysis can also answer questions about the internal structure of subgenres, and it takes into account factors other than the genre that may influence the groupings of texts found in the network.

Just as for the digital bibliography and text corpus, all Python and XSLT scripts used to perform the analyses and all associated data are published on GitHub and Zenodo in script and data repositories (Henny-Krahmer 2021a, 2021b, 2021c, 2021d). From the text of this dissertation, links are always provided to the relevant individual scripts and data in these repositories. Selected result data are also included directly in the text in the form of XML examples, tables, and figures. This book is, therefore, to be understood as an enhanced monograph: the text is a chain of argumentation and a narrative that leads through the data and scripts and becomes complete only with them. In addition, the text of this dissertation itself has been encoded in TEI and is available in a web-based HTML format and as a PDF.[10] Finally, it must be said that this work was submitted as a dissertation in early 2021. Updates could be made only partly, so in essence, the contents reflect the state of research at the time of submission.[11]

[10] The web-based edition of this dissertation can be accessed at https://side17.i-d-e.de/.

[11] In the meantime, for example, the dissertation of my co-doctoral student José Calvo Tello from the CLiGS project has been published (Calvo Tello 2021b), the content of which could not be considered here because the dissertations were prepared at the same time. Due to the joint research project in which the two theses were written, there are, of course, common foundations and references between them.

2 Concepts

A computational stylistic genre analysis of Spanish-American novels builds on terms and concepts from several disciplines. These must be clarified and related to each other, which is the goal of this chapter, in which genre-theoretical aspects, concepts of literary style and literary-historical basics on the Spanish-American novel are discussed. In the first part of this chapter (2.1), concepts of literary genre are approached. First, it is outlined which scholarly disciplines are concerned with genre studies, which ones are relevant for digital genre stylistics, and how they relate (2.1.1). Then three literary theoretical issues about genre, which have caused much debate in literary genre theory, are discussed, namely their ontological status and relevance (2.1.2), the relationship between systems or theories of genres and their history (2.1.3), and three main types of concepts for genres as categories – logical classes, prototype categories, and family resemblance analysis (2.1.4). All of these theoretical issues are related to digital stylistic genre analyses' practices to find out which genre theoretical concepts are useful and applicable in that field and how literary genre theory and computational genre stylistics interact. In the second main part of this chapter (2.2), a working definition of literary style is presented as a basis for analyzing metadata and text in the empirical part of the thesis. In the last part, in section 2.3, literary-historical background information is given for three major thematic subgenres and three literary currents of nineteenth-century Spanish-American novels to formulate hypotheses and establish a basis on which they can be analyzed textually.

2.1 Literary Genres

2.1.1 Disciplinary Locations of Genre Studies

In general language, the term "genre" is used to designate kinds of communicative acts that may be written, spoken, or otherwise represented. Not individual instances of communicative acts are designated by the term "genre", but the characteristics of groups of them. Genres may be of any sort of communication, for example, instruction manuals or podcasts, but in most cases, "genre" refers to forms of art such as kinds of works in the visual arts, performing arts, music, and literature, the latter being at the center of interest here, more precisely in their written form. This investigation thus focuses on literary genres.[12] In a general sense, literary genres can be understood as groups of literary texts that share or can be referred to with a group name because they have something in common. For example, Agatha Christie's "Murder on the Orient Express", Henning Mankell's "Innan frosten", or Mario Vargas Llosa's "Lituma en los Andes" can all be considered *novels* and, more specifically, *crime novels*. There has been much debate in literary studies about what the genre names are or should be, what the commonality of the texts belonging to a genre is, and what role genres play for literary texts at all. The investigation of literary genres is an old but still a central problem of literary studies, whether on a theoretical or

[12] For a general introduction to the notion of genre in literary and cultural studies, see Frow (2015). Genres, in general, are relevant to all the fields in the humanities, for example, linguistics, history, cultural studies, media studies, musicology, and art history. Introductions to genre studies from non-literary backgrounds include Lacey (2000) (media studies) and Bawarshi and Reiff (2010) (rhetorics and applied linguistics).

historical level. The discussion about genres can at least be dated back to antiquity, and often, Aristotle‘s Poetics from c. 335 BCE is cited as one of the initial texts concerned with genre theory.[13] Still today, there is an ongoing debate on the definition of genres both in the sense of general concepts as well as on the level of concrete individual genres, which the vast literature on genre theory and the history of genres shows.[14]

However, literary genres have not only been investigated in literary studies themselves but also within the broader context of textual genres and text classes, for example, in general linguistics, computational linguistics, and information science. While in literary studies, genres are usually understood as kinds of literary works, in linguistics, they tend to be conceived as all sorts of texts, also non-literary ones, and are therefore often referred to as ”text types“.[15] In the field of computational processing of text, there is a tradition, especially in computational linguistics, of describing, detecting, and distinguishing genres and types of text.[16] In computer science, the task of automatically grouping different kinds of texts has been pursued under the labels of “text categorization“ or “text classification“.[17] The term “categorization” is used in different ways in computer science. Sometimes it is understood as equivalent to “classification”, and in other cases, it is only used for unsupervised methods such as clustering.[18] Here, in contrast, the term “categorization” is used in a more general sense to comprise all different kinds of category building. This is the sense of the term that is usually used in literary genre theory (see Müller 2010).

The concern with literary genres, the linguistic characteristics of text types, and the computational processing of text converges in digital literary studies, computational philology, or computational literary studies and more specifically in digital stylistics, or ”stylometry“. The scope of digital literary studies is broad and comprises all points of contact between literature and the computer.[19] The term ”computational philology“ can also be understood as a collective term for all possible uses of the computer in literary studies, with a focus on the creation and use of digital editions, for example, but also on computational text analysis (Jannidis 2007; 2010, 109). Computational literary studies, on the other hand, is a newer term for a subfield of the digital

---

[13] See, amongst others, Fubini (1971, 24–27) and García Berrio and Huerta Calvo (2009, 94). See also Behrens (1940), who examines the history of the traditional classification of literature into lyric, epic, and drama and finds that triadic classifications in themselves have been found since Plato.

[14] Overviews of genre theory in the twentieth century include Dubrow ([1982] 2014), Duff (2010), and Gymnich, Neumann, and Nünning (2007). The latter also focus on the relationship between genre theory and history. On this aspect, see as well the earlier publication by Lamping (1990).

[15] An early discussion of linguistic text types can be found in Gülich and Raible (1972). A recent overview is given in Gansel (2011).

[16] A well-known study of genre variation in English texts based on statistical methods is summarized in Biber (1993b). Another influential study of the automatic detection of text genre from computational linguistics is Kessler, Numberg, and Schütze (1997).

[17] For an introduction to text categorization from the perspective of natural language processing, see Manning and Schütze (1999, 575–608).

[18] Manning and Schütze (1999, 575) use the term as a synonym for “classification”. Oakes, in contrast, differentiates the two terms: “Classification and categorization are distinct concepts. Classification is the assignment of objects to predefined classes, while categorization is the initial identification of these classes, and hence must take place before classification” (Oakes 2003, 95). Oakes bases his view on Thompson and Thompson (1991).

[19] For an overview of the main research areas and scope of digital literary studies, see Siemens and Schreibman (2008).

humanities in which a particular emphasis is placed on quantitative text analysis methods[20]. Digital stylistics, in turn, focuses on studying style with digital methods. Stylistics can be defined as "a sub-discipline of linguistics that is concerned with the systematic analysis of style in language and how this can vary according to such factors as, for example, genre, context, historical period and author" (Jeffries and McIntyre 2010, 12). The paradigmatic case of a digital stylistic study is authorship attribution, i. e. the use of statistical methods to clarify cases of anonymous or disputed authorship. However, quantitative digital methods have also recently been used for genre stylistics.[21] It should be added that stylistics also is a sub-discipline of literary studies when its methods are applied and developed in the context of literary scholarship, especially because style is considered an important characteristic of literary texts (Spillner 2001, 234).

The present study, which aims to create and analyze a corpus of nineteenth-century Spanish-American novels and their subgenres, is situated in the field of quantitative digital literary studies, computational literary studies, or, more precisely, digital genre stylistics. Therefore, the theoretical discussions of genre in general literary studies are only one point of reference. Still, they constitute a central theoretical frame for analyzing literary genres in digital stylistics, and it has to be clarified which aspects of genre can be and usually are analyzed with the text analytical digital approach.

Three issues that have been at the center of genre theoretical discussions in the twentieth and twenty-first centuries are taken up here and related to questions of the design and analysis of digital corpora of literary texts in terms of genre: the question about the ontological status (are they just abstract terms or do they exist?) and the relevance of genres, the debate about the relationship between systematic descriptions and definitions of genres and their historical manifestations, and the question of the type of category that genres can be conceived as.[22] These issues are considered especially relevant for *literary* texts and genres. They are interrelated because they all center around the question of the individuality of texts and the variability of the characteristics of text groups. The following chapters serve to address these essential literary genre theoretical issues and relate them to digital genre stylistics.

2.1.2 Ontological Status and Relevance of Genres

The first of the controversial issues of twentieth-century literary genre theory that is taken up here is the question of whether genres actually exist. Another question related to it is whether genres are a relevant category for literary analysis at all because if they would not

---

[20] See, for example, the Journal of Computational Literary Studies (JCLS; Gius, Schöch, and Trilcke 2022–2023), whose first issue appeared in 2022, and the annual conference on the same topic that has been held since 2022. A proposed definition or description of the field can also be found on the website of the Kompetenzzentrum – Trier Center for Digital Humanities (2023).

[21] For the traditional key issues of stylometry see Holmes (1998). Stylometric studies focusing on literary genre are, for example, Binongo and Smith (1999) and Hettinger et al. (2015, 2016). The former use linguistic features to differentiate between essays and plays written by Wilde, the latter classify various subgenres of the German novel based on stylometric, topic-, and network-based features.

[22] A clear summary of the main concerns of literary genre theory in the twentieth and twenty-first centuries and the principal theoretical positions, particularly in the German-speaking area, can be found in Zymner (2010, 213–219). The three genre theoretical issues addressed here have been selected because they were highlighted by Zymner and are considered relevant for the discussion of genre analysis and corpus design in digital stylistics.

exist, why should they be investigated? Both in the early and late twentieth century, there were theoretical approaches that fundamentally questioned the relevance of genres. According to nominalistic positions, generic terms are just abstract labels to aggregate and subsume similar texts, but genres do not exist. On the other hand, representatives of realistic positions argue that genres exist independently of individual texts, for example, as psychological dispositions or anthropologically basic world views (Zipfel 2010, 213–214). In his book "Gattungstheorie. Information und Synthese", which was published in 1973, Hempfer discusses both kinds of positions in detail by surveying a broad range of approaches that can be subsumed under the labels "nominalistic" versus "realistic". An important early critic of considering art in terms of genre was Croce, who emphasized the uniqueness and individuality of works of art as a result of the aesthetic and creative impetus of human activity. He considers genres useless and views them as intermediate pseudo-concepts between the individual and the universal, unable to capture or describe the individual expressions (Hempfer 1973, 38–41). Genre categories were also questioned later, in particular in post-structuralist theories. For example, Derrida finds that literary texts essentially break rules, while genres start from the opposite idea of a set of normative rules for text production and reception.[23] Still, he indirectly also recognizes the relevance of genre for the production and reception of literary works by stating that texts participate in genres even if they cannot be neatly assigned to them:

> Before going about putting a certain example to the test, I shall attempt to formulate, in a manner as elliptical, economical, and formal as possible, what I shall call the law of the law of genre. It is precisely a principle of contamination, a law of impurity, a parasitical economy. In the code of set theories, if I may use it at least figuratively, I would speak of a sort of participation without belonging—a taking part in without being part of, without having membership in a set. (Derrida 1980, 59)

According to Derrida, texts usually mark their relationship to genres, and for literature, he even sees this characteristic as necessary.[24] This remark can be made consciously or unconsciously, explicitly or implicitly, it can be made relative to several different genres, and it can be made in ways undermining the referenced genre, "mendacious, false, inadequate, or ironic" (Derrida 1980, 64). Frow interprets Derrida's critique of genre as rooted in a very specific concept of it –

---

[23] With the example of the work "La folie du jour", written by Maurice Blanchot, Derrida shows how literary texts resist their categorization in terms of genre: "The genre has always in all genres been able to play the role of order's principle: resemblance, analogy, identity and difference, taxonomic classification, organization and genealogical tree, order of reason, order of reasons, sense of sense, truth of truth, natural light and sense of history. Now, the test of *An Account*? brought to light the madness of genre. Madness has given birth to and thrown light on the genre in the most dazzling, most blinding sense of the word. And in the writing of *An Account*?, in literature, satirically practicing all genres, imbibing them but never allowing herself to be saturated with a catalog of genres, she, madness, has started spinning Peterson's genre-disc like a demented sun. And she does not only do so *in* literature, for in concealing the boundaries that sunder mode and genre, she has also inundated and divided the borders between literature and its others" (Derrida 1980, 81).

[24] "What interests me is that this re-mark—ever possible for every text, for every corpus of traces—is absolutely necessary for and constitutive of what we call art, poetry, or literature. […] Can one identify a work of art, of whatever sort, but especially a work of discursive art, if it does not bear the mark of a genre, if it does not signal or mention it or make it remarkable in any way?" (Derrida 1980, 64).

one that relates genre to prescription and taxonomic endeavors (Frow 2015, 28) – but that is not without alternatives:

> The conception of genre that I have been working towards here represents a shift away from an 'Aristotelian' model of taxonomy in which a relationship of hierarchical belonging between a class and its members predominates, to a more **reflexive** model in which texts are thought to use or to perform the genres by which they are shaped. (Frow 2015, 26–27)

Another direction of the post-structuralist critique of genre is the one developing the concept of *écriture*, which was initially formulated by Barthes. He defines *écriture* as a level between language and style, on which authors can express themselves individually and consciously, engaging in the history of literature and pursuing social intentions. Language, in turn, is naturally given to the writers of a certain period and linguistic context, and it works as a prescriptive and habitual frame. Style, on the other hand, is an individual characteristic of each writer and is just as little controlled as the general language use (Barthes [1953] 2002, 16–18). Compared to genre, the concept of *écriture* focuses more on the singularity of texts, their individual interrelationships, and the writing process. From this viewpoint, genres are seen as mere terms that suggest clear differentiations where in fact, the texts interrelate more freely and openly. In this sense, the idea of *écriture* is linked to recent theories of intertextuality. Nevertheless, as in Derrida's law of genre, the genres remain a point of reference when texts allude to generic terms and conventions, be it to break them (Schmitz-Emans 2010, 107–109).

On the realistic side, there are, amongst others, normative and anthropological conceptions of genre, but also communicative and semiotic approaches, including conceptualist positions.[25] In general, communicative theories assume that genres exist as concepts that influence the production and reception of literary works. In a narrower sense, communicative genre theories are linguistically oriented. In a wider sense, theories that emphasize the social functions of genres can also be subsumed under this term. An influential proposition was Voßkamp's idea to describe genres as "literary-social institutions" that undergo stabilization and dissolution processes and in which socio-historical communicative needs are condensed in a particular time and place. As such, genres are communicative models that are not mere text-internal literary phenomena but determined by a broader societal context (Voßkamp 1977, 30, 32; Zipfel 2010, 215). In semiotically oriented communicative genre theories, texts, genres, and generic terms are all conceived as complex linguistic signs, and genres can be understood as conventionalized models of an intended message or reality (Raible 1980, 324–326). It is assumed that such conventions and models influence authors producing literary texts and that readers, in their turn, use them to categorize and make sense of individual literary works. That way, genres become part of the communicative process and manifest themselves in it without being equated with a particular part of the process. Statements on and expectations about genres are controlled and triggered through generic signals that can accompany literary texts, be inscribed into them, and interpreted from them.

According to Hempfer, genres are only truly communicatively and semiotically determined if they are understood as a precondition for the comprehension of literary texts that authors

[25] For an overview of different genre theories associated with the realistic position, see Hempfer (1973, 56–122).

are forced to take into account and not only as historically possible but not necessary options of communication (Hempfer 1973, 90–92). It follows from this that, communicatively speaking, literary works cannot be without genre. It does not mean, though, that every work needs to be associated with exactly one genre on one specific level. On the contrary, texts can be influenced by several genres and also on different levels of generality. The mentioned "Murder on the Orient Express" and "Lituma en los Andes" can be interpreted as instances of crime novels and, at the same time, novels and, more generally, narrative. However, "Murder on the Orient Express" can also be analyzed more specifically as a "detective novel" and "Lituma en los Andes" has also been assessed as a "novela indigenista" (Martínez Cantón 2008). Then again, other texts are only framed by the genre "novel" but not a specific subtype of it. They are sometimes called "general fiction" or "literary fiction", if the literary merit of the works is stressed.[26] As Raible puts it: "Ein Werk als Exemplar einer Gattung sehen heißt es in eine Reihe von Werken stellen, die analog zu einem Präzedenzfall sind" (Raible 1980, 334). One work alone does not constitute a genre, but when it is produced and received according to communicative models that have formed and have been formed by other works, it becomes part of a system of generic conventions.

If texts that participate in genres – to speak in Derrida's terms – are understood as communicative objects, they should be described both on the level of the communicative situation and on the level of the textual sign itself. This means that both text-external features, for example, the time and place of its publication, and text-internal features, such as certain elements of content or style, determine how a text participates in a genre. Text-external factors can considerably determine a text's form, and they can narrow down the possibilities of a text's interpretation. However, literary works, especially written ones, are functionally less determined than other types of texts (Raible 1980, 334).

An approach reconciling aspects of the nominalistic and realistic conceptions of genre presented so far is Hempfer's position, which he calls "the constructivist synthesis". Following Piaget's theory of knowledge, on the level of scholarly description, he sees genres as structures that emerge from the interaction between the subject that seeks to understand them and the objects to which the structure is applied. These structures constitute a process of approximation between subject and object. As Hempfer formulates it:

> Auf der Ebene der historischen Entwicklung lassen sich die 'Gattungen' nun nicht im gleichen Sinn wie etwa die Geburt Napoleons als 'Faktum' begreifen, sondern es handelt sich, wie in den verschiedensten semiotisch orientierten Gattungstheorien betont wird, um Normen der Kommunikation, die mehr oder weniger interiorisiert sein können. Da diese Normen aber an konkreten Texten ablesbar sind, werden sie für den Analysator zu 'Fakten' und lassen sich demzufolge allgemein als *faits normatifs* verstehen, ein Begriff, den Piaget aus der Soziologie zur Bezeichnung analoger Phänomene in die Psychologie eingeführt hat. Diesen *faits normatifs* wird dann in der wissenschaftlichen Analyse eine bestimmte Beschreibung zugeordnet, die als solche immer ein aus der Interaktion von Erkenntnissubjekt und zu erkennendem Objekt erwachsenes Konstrukt darstellt. (Hempfer 1973, 125)

[26] See for example Cranenburgh and Koolen (2015). The authors analyzed the literariness of general fiction and genre fiction using machine learning based on word bi-grams.

The more interiorized the communicative norms are, the more they approach the status of ahistorical constants (for example, knowledge about what *narrative* is). Hempfer aims to differentiate the ahistorical constants from historical norms that are less interiorized and more subject to open (for example, poetological) discussion and change (Hempfer 1973, 126–127).[27]

This paper follows Hempfer's idea that genres are not to be understood as objective facts, but as communicative phenomena that can, however, leave traces in texts. If genres are understood as norms, then such textual traces can be conceived as normative facts in Hempfer's sense. The connection between genres as communicative norms and the texts on which they have an influence results in turn from the communicatively established assignment of texts to genres. How is it made clear that a text participates in a genre? This can be expressed, for example, through generic signals in the texts but also through signals that accompany the texts (e.g., in subtitles or paratexts). Thus, genre signals and genre names used in connection with literary works have a special significance for establishing genre affiliations. The various references and levels of meaning of such linguistic expressions of genres are broken down, in particular, in semiotic genre theories. Since genre labels are digital genre stylistics' primary approach to communicative genre norms, semiotic genre models are discussed in more detail in the following chapter.

2.1.2.1 Semiotic Models of Genres

One aspect that semiotic models of genres focus on is the multilayered meanings of generic terms, which point to the many communicative levels that genres can be defined on and the complexity of genres as signs. As signs, the generic terms can be understood as models for the even more complex models that the genres themselves are conceived as (Raible 1980, 334). Two semiotic models for generic terms are presented in more detail here. These are used as a basis for an empirically established discursive model of subgenre terms for the corpus of nineteenth-century Spanish-American novels created and analyzed in the context of this dissertation.[28] The first of the two models has been formulated by Raible (1980, 342–345) and involves six dimensions from which generic terms usually draw their meaning and classificatory features:

1. the communicative situation between sender and recipient ("Kommunikationssituation")
2. the object area of the texts involving persons and things ("Objektbereich")
3. the higher order structure of texts ("übergeordnete Ordnungsstruktur")
4. the relationship between text and reality ("Verhältnis zwischen Text und Wirklichkeit")
5. the communicative medium that the text uses ("Medium")
6. the way of linguistic representation ("sprachliche Darstellungsweise").

An example of a generic term that addresses the communicative situation is "children's book". In this case, the genre's name specifies the target group of the texts labeled with it. A term concerning the object area is, for example, "picaresque novel", which refers to the protagonist's social status. According to Raible, instances of the third group of terms, which refers to the higher-order structure of the texts, are relatively rare. He gives jokes as examples, as they involve

---

[27] The former are called "Schreibweisen" by Hempfer and the latter "genres" in a narrower sense (Hempfer 1973, 27).

[28] The corpus-specific model for generic terms is presented in chapter 3.2.3 below.

the expectation of something unexpected in their structure. A generic term that addresses the relationship between the text and reality is, for instance, the fable. Terms that relate to the communicative medium can refer to language, other media (music, mimic, rhythm), and carrier media, for example, the "epistolary novel". Finally, an example for a genre label that alludes to the linguistic representation of the text is "short story", which refers to the length of the form. Raible sees his model as principally open and refinable through applied literary genre analyses (Raible 1980, 342–345). A similar semiotic model of generic terms has been developed by Schaeffer (1983, 64–130). Like Raible, Schaeffer chooses to approach the complex signs that genres are through the also complex but more tangible names of the genres:

> Commençons par une question banale: quel est le statut des *classes générique*? Ou, pour éviter de nous encombrer dès le début d'entités peut-être fantomatiques, demandons-nous plutôt: quel est le statut des *noms de genres*? […] l'identité d'un genre est fondamentalement celle d'un terme général identique appliqué à un certain nombre de textes. […] les noms générique traditionnels sont la seule réalité tangible à partir de laquelle nouns en venons à postuler l'existence des classes génériques […]. (Schaeffer 1983, 65–67)

The names of the genres are significant because they witness that a generic class of texts has a communicative existence, condensed in a name. Furthermore, names applied to texts signal that the texts participate in the genres, and the participating texts, in turn, contribute to the genre's identity. The above quote is intentionally reduced to concentrate on the aspect of relevance of the generic names. In the wider context, Schaeffer explains that the relationship between the generic names and the texts is not at all simple. The names can have different statuses, as analytical ex-post terms, as words in use in a very specific historical situation, or as something between both of these poles. They can be applied collectively to a set of texts at once or to individual texts so that multiple applications of the terms, or their sum, form the genre. The meaning of generic names is not fixed, not synchronically, and especially not over time, and they can be related to other terms. One text can be associated with several generic names, which may involve different levels of significance (Schaeffer 1983, 65–66, 69). Schaeffer emphasizes that he does not want to replace a theory of genre with a lexicological study of generic names but wants to start from them to account for the fluent character of the genres and to understand the kind of phenomena that are covered by the generic names (Schaeffer 1983, 75–76). For Schaeffer, a generic term is any term, "á condition qu'il soit utilisé pour classer des œuvre ou des activités verbales linguistiquement et socialement marquées et encadrées (*framed*), et qui se détachent par là de l'activité langagière courante" (Schaeffer 1983, 77). He thus does not start from a strict separation of generic terms for literary and non-literary works, although his study focuses on the former. A condition for a term to be generic is that it is *used* for classifying works or other linguistically and socially marked and situated verbal activities. This definition again focuses on the communicative function of genres. It is limited to oral and written linguistic acts because it presupposes verbal activities, which excludes, for example, communicative acts in the visual or performing arts or music. However, within the literary-linguistic frame set by Schaeffer, all kinds of generic terms are considered.

A central observation that Schaeffer makes is that generic names do not all refer to one specific dimension of a literary work. Possible dimensions are, for example, the syntactical and semantic chain of a text that expresses a work. Instead, generic terms pick up all kinds of levels of a work as a global discursive act. That way, the generic identity of a literary work is not unique and fixed but depends on the perspective or perspectives taken towards it (Schaeffer 1983, 79–80). Similar to Raible, Schaeffer defines a literary work as a complex semiotic object. He describes dimensions of this object to which generic terms usually refer. Schaeffer's model comprises five principal dimensions:

1. utterance of the discursive act ("énonciation")
2. its destination ("destination")
3. its function ("fonction")
4. its semantic realization ("sémantique")
5. its syntactic realization ("syntaxique").

The first three levels belong to the communicative frame of the discursive act, and the other two concern its textual realization. Each level is further differentiated by reference parameters: for example, a real, fictitious, or simulated instance of utterance, or grammatical, phonetic, metric, or stylistic constraints on the syntactic level. Schaeffer also stresses that the model should not be considered complete but representative or exemplary (Schaeffer 1983, 116). As examples of generic terms pointing to different levels, he mentions, amongst others: a letter or a prayer as instances of a directed utterance, a love poem or ode as examples of an expressive function, a science-fiction story or a western as specific semantic realizations, and a lipogram (forms requiring that specific letters are left out) as a kind of syntactic realization (Schaeffer 1983, 96, 102, 108, 114).

The approaches that view literary genres as complex semiotic objects are characterized by differentiation and openness. An advantage is that they enhance the comparability of different genres by clarifying on which discursive levels generic terms operate without restricting the functioning of genres to a specific communicative level. On the other hand, some genre theoretical aspects are not clarified by these models because they focus on the communicative nature of generic structures. For instance, the semiotic models do not include the generic terms' provenience and their theoretical or historical nature into the core model, nor do they make statements about the kind of categories that genres can be (if they are to be understood as classes of texts, as prototypical structures, or other types of categoric relationships between literary works). Therefore, these two genre theoretical aspects are discussed further in the subchapters 2.1.3 ("System and History") and 2.1.4 ("Categorization").

The focus of the semiotic models on generic names leaves out another aspect of genres: who says that there are no communicative patterns without a name? There may be genres that have not been explicitly discussed or labeled but nonetheless exist as frames for communicative acts. A sign of this is that there are genres that have primarily been labeled by literary scholars in retrospect but that were not explicitly named in their historical peak phase. This does not necessarily mean that the scholars made arbitrary classifications without considering contemporary communicative practices. For example, both the *novela gauchesca* and the *novela indigenista* could not be found as explicit generic labels in the bibliography and corpus of nineteenth-century Spanish-American

novels that were prepared for this dissertation.[29] They are, nonetheless, established subgenres of the novel in the corresponding literary historiography (see, for instance, Ghiano 1957 and Meléndez 1961). Furthermore, generic signals, i.e., text-external or -internal aspects of literary works that indicate in which genres they participate, are not limited to explicit mentions of generic names. They can also be implicit or established through intertextual references (Fowler 1982, 88–105). However, such indirect signals are not directly congruent with a sign-based linguistically oriented approach and thus need to be taken into account in addition.

Up to this point, the ontological status of genres has been discussed, with a particular focus on semiotic theories of genres. In the following chapter, the general, genre-theoretical considerations will be directly related to the approaches of digital genre stylistics. Which possibilities of knowledge about genres arise from digital analyses, if they are carried out starting from certain genre-theoretical foundations? Which methodological aspects of computational genre analyses play a role in this context? Which genre theoretical approaches have been used in digital genre stylistics so far? In the following, these questions will be discussed, focusing on the special role of digital text corpora, genre labels, textual features, and text style in digital genre stylistics.

2.1.2.2 Genres and Digital Genre Stylistics: The Roles of Corpora, Genre Labels, Features, and Text Style

Regarding the question of the ontological status of genres, Hempfer's synthesis can be productively related to approaches pursued in digital genre stylistics. In digital stylistics, the anchoring point between genres as communicative norms and their descriptions in terms of textual features is just the style of the texts. Whenever literary works are associated with specific genres, a basic assumption for a digital stylistic genre analysis is the following: that the text style can be analyzed to assess to what degree there are *normative facts* expressing the generic participations of the works in the genres and of what these facts consist. Obviously, the analysis of text style is limited to the syntactic realization of the discursive act, but this does not mean that digital stylistic concepts of genre are reduced to this level of communication. The level on which digital genre stylistics principally operates is the linguistic, strictly speaking, even the orthographic surface of certain manifestations of literary texts as they are transmitted in a form that is determined by the digital medium. Still, many kinds of discursive aspects can be analyzed on this level. The crucial point is how the participation of the texts in genres is modeled and defined. As several literary genre theorists have pointed out, texts can be classified arbitrarily by any criterion, and this would include any computationally tractable aspect of text style.[30] Such an endeavor is not

[29] See the chapters 3.2.3 and 3.3.4 on the assignment of subgenre labels to the works in the bibliography and the corpus.

[30] In his set of terms, Hempfer, for instance, also includes the term "Sammelbegriff" ("collective term"), which he uses to designate logically disjunct groups of texts established on any characteristic: "Genauso wie man 'Kopf', 'Apfel', 'Platz', 'Tisch' u.ä. mit dem Prädikator 'rund' belegen und somit eine Klasse von Gegenständen bilden kann, der die Eigenschaft 'rund' zukommt, kann man Texte aufgrund ihrer Länge, des Vorhandenseins oder Fehlens eines Erzählers, der Fiktionalität oder Nichtfiktionalität, der Tatsache, ob sie in Vers oder Prosa geschrieben sind, usw., einer bestimmten Textklasse zuordnen. Wie das Beispiel der Klassenbildung mit dem Prädikator 'rund' darlegen sollte, braucht eine solche Klassifizierung keineswegs aufgrund von für die dergestalt klassifizierten Objekte wesentlicher Eigenschaften zu erfolgen, und dieselben Objekte können, je nach der Eigenschaft, die man

the point in itself. The question is to which communicative norms of genre the texts relate and in what way. A good example of taking the relativity and significance of generic assignments into account is a study conducted by Underwood, in which he analyzes different definitions of Gothic novels at different points in time.[31] As Underwood states:

> Distant reading may seem to lend itself, inevitably, to literary scholar's fixation on genre as an attribute of textual artifacts. But the real value of quantitative methods could be that they allow scholars to coordinate textual and social approaches to genre. This essay will draw one tentative connection of that kind. It approaches genre initially as a question about the history of reception — gathering lists of titles that were grouped by particular readers or institutions at particular historical moments. But it also looks beyond those titles to the texts themselves. Contemporary practices of statistical modeling allow us to put different groups of texts into dialogue with each other. (Underwood 2016, 2)

The debate that Underwood engages in is the question of the life cycle of genres. Some critics sustain that genres have relatively short life cycles, roughly corresponding to one generation and about 25 years. Others say that genres can sustain an identity over periods much longer than that, even if there are shifts in the concept of the genre. "Textual analysis won't prove either claim wrong, but it may help us understand how they're compatible" (Underwood 2016, 2). With this assertion, Underwood outlines an important task of digital genre stylistics: not necessarily to refute or confirm claims that are made on other levels of genre investigation (as here in the history of reception), but to look for textual and more specifically stylistic evidence as traces of these other levels. That way, genres are not established *in* style, but *through* style. Underwood aims to investigate how textually coherent the Gothic is over time:

> Evidence of this kind [that only a one-generational linguistic coherence could be found] wouldn't rule out the possibility of longer-term continuity: we don't know, after all, that books need to resemble each other textually in order to belong to the same genre. But if we did find that textual coherence was strongest over short timespans, we might conclude at least that generation-sized genres have a particular *kind* of coherence absent from longer-lived ones. (Underwood 2016, 3)

So digital genre stylistics can help to find out which genres imply stylistic coherence of the texts attributed to them at all and on which levels of text style they do. Underwood finds out

---

wählt, verschiedenen Klassen zugeordnet werden" (Hempfer 1973, 28). Schaeffer too, when discussing the problems of differentiating between theoretical and historical genres, notes that an infinite number of traits can be chosen to compare texts: "En second lieu, et inversement, le nombre de caractéristiques selon lesquelles on peut regrouper deux textes quelconques est indéfini sinon infini. Cela est dû au fait que, lorsqu'on compare deux textes, on ne part pas de leur *identité numérique* (toujours simple), mais de ce que Luis J. Prieto appelle leur *identité spécifique* (défini comme un ensemble de caractéristiques non contradictoires). Or, «comme chaque objet possède un nombre infini de caractéristiques, il peut posséder un nombre infini d'identités spécifiques; et comme n'importe quelle caractéristique que présente un objet donné peut toujours aussi faire partie des caractéristiques d'un autre objet, chaque objet peut partager n'importe laquelle de ses identités spécifiques avec un nombre infini d'autres objets 3»" (Schaeffer 1983, 67–68). The potential arbitrariness of the textual features that describe a text category is thus an important point to consider when text classification as a computational activity is used to determine literary genres.

[31] He differentiates between "detective fiction", "science fiction", and "Gothic" (Underwood 2016, 4).

neither strong evidence for the succession of genre generations nor for a gradual consolidation of the genres over time. The *sensation novel* is short-termed but textually not very coherent, while *detective novels* and *science fiction novels* are textually connected for longer periods (Underwood 2016, 4).[32]

If digital stylistic genre analysis functions as a connector between social and communicative norms of genre and stylistic textual evidence, several aspects in the connective chain need to be defined and selected with care. Usually, a corpus of texts is analyzed, and the generic conventions in question are expressed as genre labels of the texts, which themselves are representatives of literary works. The assignment of genre labels to the texts is the first crucial point. Which kind of genre labels are selected, and how are they assigned to the texts? The semiotic models of generic terms provide a way to differentiate between different discursive levels on which genres can be defined, which can help not to compare "apples with oranges". That would happen if one would, for example, contrast primarily formally defined genres with thematically defined ones. The sources of the genre labels should always be indicated to document which kind of generic convention they represent. Do the genre attributions go back to assignments made to individual texts by different authors, editors, or publishers? Or are they collectively defined, for example, established in a discussion of a set of works by a contemporary critic or poet, by modern institutions, or by a literary historian? Are the assignments made based on explicit generic terms or implicit signals of the texts? Or are they derived from specific theoretical definitions of genres that are applied to the texts? Depending on the answers, quite different kinds of generic conventions can be analyzed. In the worst case, it is not clear which type of genre an analysis aims at, and the goal of addressing a communicative pattern that lies outside of the analysis itself would be missed. Awareness of the kind of analysis target can still be raised in digital genre stylistics.[33] Despite all good advice and intentions to analyze genres or subgenres

---

[32] Underwood's method is predictive modeling with L2-regularized logistic regression based on the top 10,000 word features in the text collection (Underwood 2016, 7). The findings of his article from 2016 have been integrated into his book "Distant Horizons" (Underwood 2019, 34–67).

[33] The case of Underwood is an example of a clear definition and transparent documentation of the target genre convention: "To investigate these questions, I've gathered lists of titles assigned to a genre in eighteen different sites of reception. Some of these lists reflect recent scholarly opinion, some were defined by writers or editors earlier in the twentieth century, others reflect the practices of many different library catalogers (see Appendix A) […] By comparing groups of texts associated with different sites of reception and segments of the timeline, we can ask exactly how stable different categories have been" (Underwood 2016, 4). In their classification of subgenres of the German novel, Hettinger et al. also explain that they analyze genre attributions made by literary scholars: "Literary scholars and common readers use labels like educational novel, crime novel or adventure novel to organize the large domain of fiction. In both discourses the use of these categories is well-established even though they are evolving and tend to be inconsistent. […] Our corpus consists of 628 German novels mainly from the nineteenth century […]. The novels have been manually labeled according to their subgenre after research in literary lexica and handbooks" (Hettinger et al. 2016). Kim et al. also explain the provenience of the genre labels in their investigation of the prototypical emotion developments in literary genres: "We collect 2113 books from Project Gutenberg that belong to five genres found in the Brown corpus […] namely adventure (585 books), romance (383 books), mystery (380 books), science fiction (562 books), and humorous fiction (203 books). […] The selection is based on the Library of Congress Subject Headings in the metadata" (Kim, Padó, and Klinger 2017). A case in which the provenience of the generic assignments to the texts remains implicit is Schöch (2017c). Schöch analyses subgenres of French Classical and Enlightenment Drama by applying topic modeling to a corpus of plays that was initially curated by Paul Fièvre (called the "Théâtre classique" collection). The data which is used in the analysis is presented in detail,

only on a defined discursive level or on the basis of homogeneous sources of subgenre labels, especially large-scale digital analysis using hundreds or even thousands of texts have to face challenges in defining which generic conventions they refer to. Even in qualitatively oriented genre analysis, selecting works for a corpus that aims to cover one or several specific genres is not trivial. A starting point using either certain labels or definitions of the genre(s) has to be found.[34] Beyond that, some strategies of text selection and genre assignment are not viable for very large corpora. It is, for example, not possible to read every text and check it against a genre definition that relies on qualitative textual features, i.e., characteristics of the texts that are not (yet) easily formalized and computationally analyzable. Furthermore, it is very likely that large corpora also cover lesser-known texts which critics have not considered yet. That way, existing critical approaches may only cover part of a text corpus. On the other hand, depending on the kind of genre, explicit labels on historical editions may also not be the norm. These are additional difficulties that quantitative genre analysis faces in defining of its object of investigation through the selection and preparation of the text collection. Such challenges make it even more important to clarify which genre convention is addressed and how this is done.

Besides assigning genre labels to the texts, another fundamental point for a digital genre analysis is the selection of textual features. In the end, the *normative facts* that can be found in the texts, that can be related to genre conventions, and that can be used to establish definitions of genres, depend on which kind of textual material is analyzed. There are different opinions regarding the importance of which kind of features are selected. Underwood, for example, highlights the predictive power of statistical models, which is based on specific features but not directly dependent on them:

> Leo Breiman has emphasized that predictive models depart from familiar statistical methods (and I would add, from traditional critical procedures) by bracketing the quest to identify underlying factors that really cause and explain the phenomenon being studied. Where genre is concerned, this means that our goal is no longer to define a genre, but to find a model that can reproduce the judgements made by particular historical observers. (Underwood 2016, 5–6)

---

and the subgenres are also mentioned, but it is not made explicit where the labels that are finally used come from: "detailed metadata has been added to the texts relating, for instance, to their historical genre label (e.g. *comédie héroique*, *tragédie*, or *opéra-ballet*) as well as the type of thematic and regional inspiration [...]. A large part of this information can fruitfully be used when applying Topic Modeling to this text collection. [...] Finally, all the plays included belong to one of the following subgenres: comedy, tragedy or tragicomedy" (para. 9–10). Looking into the TEI collection on GitHub (e.g., https://github.com/cligs/theatreclassique/blob/master/tei/tc0001.xml, accessed November 28, 2020), it can be noticed that there is a general subgenre assignment in the TEI header (`<term type="genre">Comédie</term>`), but no further information about its provenience is given. As Schöch mentions the historical subgenre labels in his text, it can be assumed that these are the source. The classification of works into dramatic subgenres is probably less debated than into subgenres of the novel. Likely, dramatic subgenre labels are also more often explicitly given on title pages than in the case of novels. Still, it would be better to make the generic convention that is analyzed more explicit because it makes a difference whether labels assigned by librarians or literary historians or genre labels from the historical paratexts of the works are used.

[34] An overview of proposals that have been made for corpus building in literary genre studies is given in chapter 3.3 ("Text Corpus") below.

Taking the example of science fiction, Underwood explains that a very reliable textual clue for this genre are adjectives of size such as "huge" or "tiny" and that a set of some more hundred words would be enough for a statistical model to recognize instances of the genre. Still, he argues, these genre markers do not need to correspond to any definition of the genre that has been formulated so far, and they might not lead to any definition that could be articulated verbally in a useful way (Underwood 2016, 6). Underwood's stance towards selecting textual features is characterized by a reproductive strategy on the one hand and an explorative one on the other. It is a reproductive strategy in the sense that text analysis is used to replicate social constructions of genre to find out about their general relationship to the textual basis. It is explorative in that the kind of features used is not defined in a top-down approach and controlled in advance, but tested as for their reproductive relevance: "To put it more pointedly: computational methods make contemporary genre theory useful. We can dispense with fixed definitions, and base the study of genre only on the shifting practices of particular historical actors – but still produce models of genre substantive enough to compare and contrast. Since no causal power is ascribed to variables in a predictive model, the choice of features is not all-important" (6). Following this approach, the *normative facts* found as traces of social constructs of genres would not lead to descriptions of them in scholarly terms, at least not to definitions focusing on the kinds of facts found.

A different view on the question and relevance of feature selection is formulated by Jannidis, who outlines a set of general methodical working steps for computational text analyses: "1. Thesenbildung, 2. Bestimmung der Indikatoren, 3. Korpuszusammenstellung, 4. Korpusvorbereitung, 5. Suche, 6. Quantitative Erhebung, 7. Überprüfung von Indikatoren und Korpuszusammenstellung sowie Diskussion der These im Licht der Ergebnisse" (Jannidis 2010, 110). In this setup, the choice of indicators is directly linked to the formulation of an initial thesis and is more driven by prior theoretical assumptions than in Underwood's approach.[35] It represents a deductive procedure. In the case of genre analysis, a genre-related thesis would need to be formulated, for instance: "In social novels, there are more different characters than in sentimental novels". To be able to verify or falsify the hypothesis through computational text analysis, it would be necessary to define textual indicators representing the concepts mentioned in the hypothesis. The above example would require an approach to identify characters in the text, for example, by detecting mentions of character names and other linguistic references to characters and resolving to which character they point. It would be necessary to automatically detect the set of different characters in a novel, which is a difficult task. Somewhat easier to formalize and closer to a stylistic analysis would be a hypothesis such as "In social novels, there are more mentions of different character names than in sentimental novels". Like the explorative procedures, also top-down approaches have several advantages and disadvantages. What is good about them is that they start directly from the scholarly field that is also the target context. If the goal is to find out something about literary genres and the hypothesis is formulated in literary scholarly terms, the hypothesis is compatible with the epistemological frame of the investigation. In addition,

[35] However, Jannidis himself comments on the status of the work steps: "Diese hier skizzierte Vorgehensweise ist natürlich stark idealisiert. Nicht selten steht am Anfang nicht die These, sondern ein auffälliger Befund in Texten, der dann als Indikator für eine These gedeutet wird. Doch selbst in diesem Fall einer induktiven Vorgehensweise ergibt sich zuletzt eine ähnliche Forschungsstrategie, wie hier skizziert" (Jannidis 2010, 110).

the selection of textual indicators and features can be motivated theoretically so that meaningful and interpretable results can be expected. The main disadvantage is that the possibility of formalizing the hypotheses depends on the available technical methods. Although research is done in this direction, many literary theoretical concepts still need to be formalizable.[36] Existing text mining methods, for example, topic modeling or sentiment analysis, may also be used to operationalize the hypotheses. Then it must be explained in what way they can be considered formalizations of literary theoretical concepts, such as themes, topoi, or emotions. In any case, in such an approach, the selection of specific textual features is not at all arbitrary or negligible. The features represent the texts and are assumed to cover stylistic aspects that are traces of generic conventions in the texts. The chosen indicators must be suited to check the plausibility of the literary theoretical hypothesis. At the same time, the choice of the indicators themselves is based on assumptions: "Das Verhältnis zwischen Indikatoren und These ist allerdings in vielen Fällen keineswegs selbstverständlich, sondern hat selbst hypothetischen Charakter" (Jannidis 2010, 116). Even in hypothesis-driven digital genre analysis, the suitability of the features needs to be tested empirically to some degree.

In the same way as the kind of generic convention that is analyzed and the selection of corresponding genre labels, the choice of textual features for a digital stylistic genre analysis should also be motivated. How specific the chosen features are and how important it is to clarify their relationship to literary theoretical concepts depends on the kind of strategy that is chosen for the genre analysis: it can be primarily deductive, inductive, or experimental, and it can be theoretically or historically oriented. In principle, digital genre analysis can be used for all kinds of investigations. The goal can be, for instance, to find out if and in what way works with specific historical or critical genre labels are textually coherent, as Underwood did. Another goal can be to test if a specific scholarly definition of a genre holds when it is formalized and applied to a corpus of texts that have been assigned to the genre in question. The results of a stylistic genre analysis could also be used to formulate new, statistically-based definitions of genres. Even if textual variables in statistical models do not necessarily reflect causal relationships, their distribution can be interpreted to reach empirically based genre definitions if genres can be distinguished based on these variables. A sentence in such a definition could be, for example, "the genre X is defined such that the probability of a love topic is significantly higher in texts that participate in the genre X than in texts that do not participate in this genre". In Hempfer's terms, the normative fact that was found is the different probability of a certain kind of topic in two groups of texts that are associated with different genres by convention. The genre is constructed in the interaction of the scholar who decides which textual features to use and which texts to analyze on the one side and the texts themselves on the other. Moreover, the scholar has to interpret the topics and decide that one of them can be described with the term "love". Furthermore, it has to be decided what "significantly higher" means. A definitory phrase such as the one above can itself be used as a hypothesis that can be tested in other empirical settings, for example, with a different corpus of texts. To what extent the found textual characteristics of exemplars of different genres actually correspond to conscious social norms can only be clarified by analyzing contemporary

[36] For an approach to automatically recognize characters in German language novels, see Jannidis et al. (2015). Barth and Viehhauser (2017) made the first attempts to formalize concepts of literary space.

or historical discussions about what the genres in question are. That would not be different in non-computational text analysis. In addition, besides starting from known generic conventions or scholarly definitions of genre, a digital stylistic genre analysis can also start from the texts themselves. For instance, a corpus of texts can be built for a certain period and a specific cultural, geographical, and linguistic context. It can then be analyzed which groups of texts emerge as textually coherent when specific textual features are used. Such an approach would allow for the possibility of detecting *faits normatifs* as signs of communicative patterns that might not have had a high degree of explicitness. They might not have been frequently named or broadly discussed in the historical context, and possibly they have not been described yet in scholarly terms. In such cases, it would as well be possible to complement the quantitative analysis with a qualitative study of intertextual links, of implicit or explicit signals in the texts and paratexts, and of metatextual statements that could substantiate or question the communicative relevance of new findings of text groups.

When stylistic text features are interpreted as signs of generic conventions, a difficult point is how clearly the relationship between both characteristics of texts can be established: having certain features or a specific distribution of them on the one hand and participating in a particular generic convention on the other. In this context, only the causal relationship between the genre label and the textual features is meant, not the question of the kind of social, generic norms and their unconscious or conscious application. In many cases, literary-historical studies of genres focus only on one, positively described genre.[37] Often subtypes of one genre are distinguished as part of the investigation, especially if a major genre is concerned,[38] but explicitly contrastive studies are rare.[39] In corpus-based and empirical digital genre stylistic analysis, it is more usual to directly contrast different genres or subgenres to find distinctive features for each group (see, for instance, Schöch 2018 and Schöch et al. 2018).[40] Approaches based on statistical classification also bring forth features that are decisive for distinguishing different classes of texts.[41] If only the characteristics of one genre are worked out, one cannot say with certainty that there are not other genres that share part of these characteristics. This possibility is avoided in contrastive

---

[37] In the context of subgenres of Spanish-American novels, see, for instance, Gnutzmann (1998) and Schlickers (2003) on the naturalistic novel, Löfquist (1995) and Read (1939) on the historical novel, and Rivas (1990), Rosell (1997), and Sparrow de García Barrío (1977) on the anti-slavery novel.

[38] Suárez-Murias (1963), for instance, is a study of the Spanish-American romantic novel by country, in which subtypes of the romantic novel are presented, such as sentimental novels, historical novels, or novels of customs. A comprehensive overview of thematic subtypes of Spanish-American novels is given in Sánchez (1953). In Molina's (2011) study of the early nineteenth-century Argentine novel, four classes of novels are established ("novela histórica", "novela política", "novela socializadora", "novela sentimental").

[39] The tradition to directly contrast genres is stronger in linguistics than in literary studies. For text linguistic contrastive genre analyses, see, for example, Adamzik (2001), Danneberg and Niederhauser (1998), Gnutzmann (1990), Kaiser (2002, 2008), and Theisen (2016). In literary studies, contrastive analyses are, in particular, used in comparative studies on different cultural and linguistic literary systems. See Lamping (2010) for an overview of comparative genre studies and Jacobs (1986) for a case study on the picaresque and the education novel.

[40] The concept of distinctiveness or keyness, which aims to find words characteristic of one group of text compared to another, is a general one, not limited to analyses of genre. See Burrows (2007), who developed the Zeta-measure for questions of authorship, and Scott (1997) for a general approach to keyword extraction.

[41] For an analysis of the text features that are decisive in detecting the nationality of authors of Spanish language novels, see, for instance, Zehe et al. (2018). Sentiment features that are important in the classification of subgenres of nineteenth-century Spanish-American novels were explored in Henny-Krahmer (2018).

studies. On the other hand, the results of comparative approaches obviously depend on what is compared. A contrastive text analysis aiming to define the sentimental novel, which is based on a corpus of sentimental, historical, and adventure novels, will lead to a definition of the sentimental novel that is relative to the other subgenres. If sentimental novels were instead compared to science fiction and crime novels, different aspects might be decisive in their distinction and recognition. The choice of the corpus that is used to determine a genre in the context of other genres is, therefore, a central aspect of digital stylistic genre analysis. In the case of contrastive analysis, the relationship between the textual features and the genre in question is established *relative* to other genres on which the outcomes depend.

Another challenge in this regard is that text style is not only a function of genre. All kinds of intra- and extra-textual phenomena shape the style of a text. A special awareness of this circumstance has developed in the field of authorship attribution, where it was repeatedly noted that, for example, the period a text was written and published in but also its genre interfere with authorship signal.[42] In the same way, authorship, time period, and other factors can also obscure the link of textual features to genre.[43] It may thus happen that conclusions are drawn about genre that are instead due to unrecognized effects of other variables. Considering the different discursive levels on which generic terms and genres can be defined, it is also likely that these different levels may correlate or interfere with each other. For example, in the bibliography of nineteenth-century Spanish-American novels that was prepared for this dissertation, there are 172 novels with the primary thematic subgenre "novela sentimental". For 85 of these, the literary current to which they belong is unknown. 72 have been associated with the romantic current, 15 with the realist, eight with the naturalistic, and two with the modernist current.[44] Putting aside the cases of unknown literary currents, the numbers suggest a strong correlation between a primarily sentimental theme and the romantic current. Definitions of the sentimental novel derived from this set of novels will therefore be strongly influenced by the period in which the romantic current was the dominant aesthetic program.

If there are undesired factors that influence the target style that is analyzed (e.g., the influence of period on genre style if genre is the primary concern), then several strategies are possible to cope with such factors. A crucial aspect is the construction of the text corpus that is used for the genre analysis. For example, it is possible to include only one text or an equal number of texts per author. This prevents the results from being too much influenced by authors that were very productive writers of texts that can be attributed to specific genres. However, in most cases, creating a balanced corpus means moving away from a text collection that represents the historical proportions of works according to specific criteria. It means that one tries to create a synthetic setting that can be used to get results that are free from unwanted influencing factors in

---

[42] Oakes mentions the following issues that can mask the individual authorial style and make it challenging to attribute texts correctly to a particular author: heterogeneity of authorship over time, genre, gender, variation within a single author, and topic (Oakes 2009, 1072–1073).

[43] An attempt to neutralize authorial signals in genre analysis has, for example, been made by Calvo Tello et al. (2017).

[44] The sum of the numbers related to currents is higher than 172 because the same work can be associated with several literary currents. The bibliographic data containing this information is available at https://raw.githubusercontent.com/cligs/bibacme/master/app/data/works.xml, and the script used to retrieve the numbers of sentimental novels can be viewed at https://github.com/cligs/scripts-nh/blob/master/concepts/subgenre-label-combinations.xsl. Both links accessed November 29, 2020.

order to reach a definition of the subject that is theoretically "clean". This may be very difficult if there are not enough sources for such a corpus. Going back to the Spanish-American sentimental novels, taking an equal number of romantic and realistic sentimental novels would reduce the number of novels to analyze to 30 instead of 172. Besides controlling external factors through corpus building, another possibility is to try to choose the text features in such a way that they are likely to be relevant for specific generic distinctions but not for other kinds of differences. However, it has so far not been possible to isolate features that are only connected to a single extra- or intratextual influencing factor.[45] A third way would be to model the factors that are assumed to influence the target variable, for example, by collecting corresponding metadata and using this information when the analysis results are inspected. In principle, a corpus that is created by random sampling may represent historical imbalances. However, it would still be possible to estimate how much influence other factors have on the stylistic features that are interpreted in terms of the genres that are investigated. The decision for a specific strategy to control different factors that may influence the text style may depend on the aim of the study. The wider the claim of validity is for the results that are reached, i.e., the more independent they should be from contextual determinations and intra-textual correlates, the more important it is to build a corpus and features that are theoretically adequate. The narrower the scope of the genre study, the more it will be acceptable to have a corpus and feature set that is influenced by the specific historical setting, and that leads to a less theoretical but more organic description of the genres in question. Looked at another way, the stronger the theoretical claim on the results of a corpus-based study, the more important it is to control for possible factors influencing the target variable of the study, i.e., literary genres. On the other hand, a primary interest in historical adequacy can be pursued by considering and investigating influencing factors, but not necessarily controlling them, for example, by balancing a corpus of texts. Awareness of influencing factors is still indispensable in both cases to make sure that the assertions made are about genre at all.

It is clear that the ontological status of genres as conventions or norms of communication or social interaction – if they are understood in that way – makes the access that digital genre stylistics has to them one that is mediated by text style. Text style, in turn, is influenced by a number of factors other than genre. These influence factors are never captured as a whole but only on selected levels of textual features that are chosen for analysis. One could say that genre "hangs by a thread" for digital stylistics, but in general, in the debate about the status of genres, the field can be characterized as inclined towards the realistic side. How strong the connection between genres, genre labels, and genre signals, on the one hand, and common features of text groups, on the other hand, is, can be analyzed in detail with digital text analysis. It is to be expected that the results may be quite different depending on the kinds of genres and the literary-historical contexts that are investigated. Quantitative approaches lend themselves very well to analyzing big corpora of formula fiction, that is, popular genre fiction for which uniform patterns of style are expected. In such a research setting, genres are probably more tangible than in an analysis of highly canonized, individualistic works of art where generic references may be

[45] In a presentation about the differentiation of authorship, form, and genre of literary texts, Schöch and Pielström tried to identify statistical components that are clearly attributable to either of these factors. Analyzing French dramatic texts, they found two components that primarily covered differences in authorship but none that were predominantly related to genre (Schöch and Pielström 2014a, 2014b).

weaker.[46] Furthermore, as was pointed out above, digital genre stylistics is a field that is closely linked to computational linguistics, and the existence of linguistic text types is less questioned than the one of literary genres, mainly because it is more evident that the text types constitute communicative norms.

There are further aspects that are as well relevant to the relationship between literary genre theory and digital genre stylistics and that have not been covered yet in this discussion of the role of corpora, genre labels, features, and text style. One aspect is the relationship between genre systems and their history, a field of tension that is also linked to terminological questions. Another aspect is the debate about the type of categories that genres can be conceived as. These issues will be touched upon in the next chapters.

2.1.3 System and History

Besides the ontological status of genres and the question of how genres are to be grasped communicatively and textually, another central point of debate in the theoretical discussions of genre in the twentieth century was about the relationship between a system of genres and their history (Zipfel 2010, 214). In this and the following chapters, major positions on this question are outlined and it is discussed how they relate to approaches of digital genre stylistics. The question involves several issues, among which are:

- the compatibility (or incompatibility) of systematic and historical conceptions of genre and their relationship
- the delimitation of genres in a narrower sense, and other theoretical and historical concepts of text types and discursive practices and conventions
- the theoretical or historical status of generic terms
- the generic identity of individual works and their contextual embedding as texts that are representatives of a certain genre
- the origin, the context, and the historical evolution of genres and genre systems

There is a range of different propositions for defining the relationship between systematic and historical genre concepts. An early proposal was formulated by Todorov, who argued for the necessity to distinguish between theoretical genres that arise from deductive procedures and are based on a theory of literature and historical genres that are captured by observing the historical, literary reality. Todorov sees historical genres as a sub-ensemble of theoretical genres. That the definition of theoretical genres depends on a theory of literature is explained by Todorov as follows: a theory of literature involves a concept of how a literary work is represented. A theory

[46] Jannidis, Konle, and Leinen (2019), for example, analyzed a corpus of 9,000 dime novels in German language, which were published between 2009 and 2017. They aimed to find out how the subgenres of the dime novels can be differentiated and in what way the corpus as a whole is different from high-prestige novels. They used the 8,000 most frequent nouns as features and classified the novels with Logistic Regression. In addition, a clustering was done on the basis of the 2,000 MFW. To analyze the relevant features for the different subgenres, topic modeling and a contrastive analysis with the Zeta measure were performed. They found that the subgenres can be distinguished well both on a stylistic and a thematic level. Regarding the complexity of dime novels compared to high-prestige literature, they found that sentences are shorter in dime novels. However, they did not find any clear differences in the vocabulary richness or length of the words.

of genres then refers to the theoretical concept of the literary work to determine on which levels of this concept genres are defined and which generic characteristics are available on each level. According to Todorov, three aspects must be distinguished for a representation of a work: the verbal, the syntactic, and the semantic aspect. The first one corresponds to concrete phrases of a text that represents a literary work and is connected to questions of register,[47] style, and the instance enunciating or receiving the text. The second level concerns the composition of a literary work, that is, the organization of its different parts (logically, temporally, or spatially). For the third level, the semantic one, the themes or topics of the literary work are relevant (Todorov 1970, 24–25). Possible theoretical genres are deducted from the constellations of characteristics that are available on the different levels that the literary work is defined on. The historical genres are a sub-ensemble of them because not all of the theoretically possible genres may be found in the history of literature. Although Todorov proposes a clear distinction between theoretical and historical genres, he also sees how they are interrelated:

> Les genres que nous déduisons à partir de la théorie doivent être vérifiés sur les textes: si nos déductions ne correspondent à aucune œuvre, nous suivons une fausse piste. D'autre part, les genres que nous rencontrons dans l'histoire littéraire doivent être soumis à l'explication d'une théorie cohérente ; sinon, nous restons prisonniers de préjugés transmis de siècle en siècle […]. La définition des genres sera donc un va-et-vient continuel entre la description des faits et la théorie en son abstraction. (Todorov 1970, 25–26)

Todorov notes that the genres (and it can be assumed that he means theoretical as well as historical genres) do not *exist* in the literary works, but rather that they are *manifested* in them. According to Todorov, if a theory tries to explain it, the relationship of manifestation between genres and works is characterized by *probability* and cannot be seen as absolute (Todorov 1970, 26). An important point can be derived from Todorov's explanations: even historical descriptions of genres depend on theoretical presuppositions, or rather, different genre theories vary in the extent to which they integrate historical observations into their definitions of generic systems and genres. The debate about a system versus a history of genres can then be viewed as one of degree (from the extreme that history is not needed to establish a theory of genres to the other that a theory of genres is only possible in the description of historical circumstances) and terminology (are different kinds of genres, abstract-theoretical and historical ones, to be distinguished and which notion should be called a "genre"?).

As was outlined in the previous chapter, the different genre theories that can be ascribed to the realistic side concentrate on different locations of the *being* of genres. The ones that see it as primarily determined by production would also need to focus on the productive side to describe genres historically, for instance, on the history of the creation and publication of the literary works that participate in the genres. In the same way, if genres are primarily conceptualized as a phenomenon of reception, the history of the reception of literary works that are seen as instances

---

[47] In Todorov's study, the term "register" is probably meant in the linguistic sense of a functionally determined specific way of writing or speaking (for example, formal versus informal) and as a term that is related to that of style.

of the genres becomes a central element of the theory. Such a genre theory is, for example, sketched by the romance philologist Jauß. Croce's rejection of genres as relevant concepts because every work of art is individual and violates genres is taken up by Jauß, who objects that a literary work can only be understood as breaking the rules of genres if there is a previous understanding of these rules: "it still presupposes preliminary information and a trajectory of expectations (*Erwartungsrichtung*) against which to register the originality and novelty" (Jauß 2014, 131). The horizon of what can be expected is conceived as the contemporaneous reader's knowledge of tradition and experience with other literary works. Because such a horizon of expectation is always present, there is no work without a genre. Because the horizon may be shifted with the experience that a reader makes with new works, genres have a "processlike appearance and 'legitimate transitoriness'" (Jauß 2014, 131). Jauß concludes that "literary genres are to be understood not as *genera* (classes) in the logical senses, but rather as *groups* or *historical families*. As such, they cannot be deduced or defined, but only historically determined, delimited, and described" (Jauß 2014, 131). Because the readers' horizon of expectation cannot be determined from a purely theoretical standpoint, Jauß' genre theory is essentially historical. However, it is still a theory, especially when the approach is used to trace the history of one or several genres in broader lines:

> the history of genres in this perspective also presupposes reflection on that which can become visible only to the retrospective observer: the beginning character of the beginnings and the definite character of an end; the norm-founding or norm-breaking role of particular examples; and finally, the historical as well as the aesthetic significance of masterworks, which itself may change with the history of their effects and later interpretations, and thereby may also differently illuminate the coherence of the history of their genre that is to be narrated. (Jauß 2014, 132)

Here, Jauß also recognizes that a specific literary work's role in the history of a genre is not only determined by the contemporaneous context of reception but also depends on how broadly the temporal context is chosen and which perspective the scholar has on it. This view can as well be related to Hempfer's constructivist synthesis: In this case, the *normative facts* are the expressions of horizons of expectations about genres, which must be substantiated through historical sources,[48] and the genres are constructed in the interaction of the person with these facts.[49]

Many more theoretical approaches to genres are concerned with determining trans-temporal basic genres. One example is Goethe's attempt to define the epic, the lyric, and the drama as the three genuine natural forms (Genette 2014, 212). Problems with the definitions of these three

---

[48] How well the horizons of expectations can be captured through the analysis of historical documents is a point of debate (Voßkamp 1977, 29).

[49] Another approach that is strongly dependent on a historical anchoring is Voßkamp's concept of genres as institutions, which are determined by their social and functional history. Voßkamp understands genres as selections among several possible alternatives and argues that prototypical works play an important role in institutionalizing generic conventions and forming generic norms. According to the institutional theory of genres, an important task is to analyze the literary- and socio-historical context and the conditions of the prototypical works' creation, to understand what differentiates them from the alternatives that existed and what their specific social and historical functions were (Voßkamp 1977, 30–31).

natural forms are pointed out by Genette, who introduces a more differentiated terminological system. It aims to clarify which aspects of the three basic forms can be considered trans-historical and which ones are historically bound. For Genette, the lyrical, epical, and the dramatic can be defined as "modes", understood as linguistic categories that describe the mode of enunciation, for example, narration in the case of the epical and dramatic representation for the dramatic. On the other hand, as soon as thematic elements enter the concepts, Genette argues that they become historically variable. Only in this form, in combining formal and thematic elements and in pointing to specific historical conventions, should they be called "genres" (Genette 2014, 210, 213). Even so, to do justice to the importance of the three major genres, lyric, epic, and drama, and the prominent status that they had in the history of literature, Genette calls them "archigenres":

> *Archi-*, because each of them is supposed to overarch and include, ranked by degree of importance, a certain number of empirical genres that – whatever their amplitude, longevity, or potential for recurrence – are apparently phenomena of culture and history; but still (or already) *-genres*, because (as we have seen) their defining criteria always involve a thematic element that eludes purely formal or linguistic description. (Genette 2014, 213)

Genette also attributes a dual status of higher-order categories and specific historically manifested genres to less prominent forms such as the novel or the comedy, which can be subdivided further into "species", a concept which is comparable to subgenres, "with no limit set a priori to this series of inclusions" (Genette 2014, 213). Although Genette separates a level of the trans-historical, linguistically defined *modes*, by admitting a dual status for genres, he maintains a double function of generic terms as theoretical and historical entities, and is inclined towards the theoretical status:

> There is no generic level that can be decreed more 'theoretical', or that can be attained by a more 'deductive' method, than the others: all species and all subgenres, genres, or supergenres are empirical classes, established by observation of the historical facts or, if need be, by extrapolation from those facts – that is, by a deductive activity superimposed on an initial activity that is always inductive and analytical (Genette 2014, 214)

If no generic level can be decreed more theoretical than others, they are all theoretical, although empirically induced. Compared to the production- or reception-aesthetic and the social- and function-historical-oriented approaches on the one side and the essentially literary-theoretical positions on the other, the semiotic models of genres and generic terms as elaborated, for example, by Raible and Schaeffer can be located on a middle position regarding their systematic and historical conception of genres. They are based on a theory of language and on models of the discursive levels that are involved in speech acts, which forms their theoretical core. History enters into this system because the meaning of signs is context-dependent and changes over time. In addition, the speech act is embedded into a situational context. Because there are a speaker and a hearer, or an author and a reader, not only the linguistic but also the extra-linguistic context becomes relevant, although this aspect is not the primary concern of the semiotic approaches.

In its applied form, digital genre stylistics deals with corpora of contemporary or historical texts and has, therefore, a strong empirical foundation. Historical realizations of literary works are at the center of digital text analysis. Para-textual and extra-textual factors are often included in an analysis as metadata. For example, genre labels of different proveniences are included, as well as biographic information about the works' authors, information about how the works have been received and valued by contemporaries or literary critics, or details about the sources of the texts and their publishing (when were the works published, where and by whom?). However, a broader or closer consideration of the literary and extra-literary-historical context is usually not pursued as part of the quantitative digital genre analysis itself. One example is Underwood, who places his analysis of detective, science fiction, and Gothic novels in the context of the history of reception. He uses basic bibliographic metadata to do so, but not entire historical documents, which could serve to reconstruct how the works in question and the genres they are associated with were received in their time (Underwood 2016, 2, 11, 17, 20–21). The approach is reasonable because supervised learning is used, and if genre assignments are the target categories, they need to be formulated as compact terms. However, the idea of relating the analysis of the texts to how they were received in terms of genre historically is there. In most corpus-based analyses, though, no shifts in the relationship between generic terms and texts are analyzed, but one specific synchronic view. The synchronic view may either be based on scholarly and librarian classifications, on those made by contemporary readers, publishing houses, or booksellers, or on labels found on historical editions of the texts.[50] The first group of contemporary labels thus leads to an analysis of how today's genre concepts relate to the style of historical texts and, depending on the kind of textual material that is analyzed – twentieth-century or seventeenth-century texts, for instance – the contexts of production and reception are quite different. In the latter case of using historical genre labels, a specific historical section of the literary field is analyzed, and the contexts of the texts themselves and their generic categorization are congruent. Even if the historical development of genre concepts is not explicitly modeled from the point of view of reception over time, quantitative genre stylistic analyses are likely to involve different relationships between genre labels and texts, especially if the corpora are very large and comprise texts of one long or of several literary-historical periods. In the end, macroanalytic and, thus also, diachronic studies are enabled by the availability of a huge number of literary texts in digital format.[51]

---

[50] See footnote 33 above for examples.

[51] The term "macroanalysis" was coined by Jockers: "The approach to the study of literature that I am calling 'macroanalysis' is in some general ways akin to economics or, more specifically, to macroeconomics. […] There was, however, 'microeconomics,' which studies the economic behavior of individual consumers and individual businesses. As such, microeconomics can be seen as analogous to our study of individual texts via 'close readings.' Macroeconomics, however, is about the study of the entire economy. It tends towards enumeration and quantification and is in this sense similar to bibliographic studies, biographical studies, literary history, philology, and the enumerative, quantitative analysis of text that is the foundation of computing in the humanities" (Jockers 2013, 24). When listing the opportunities of the macroanalytic approach, Jockers mentions several points that are linked to historical contextualization and change: "This approach offers specific insights into literary-historical questions, including insights into: the historical place of individual texts, authors, and genres in relation to a larger literary context; literary production in terms of growth and decline over time or within regions or within demographic groups; literary patterns and lexicons employed over time, across periods, within regions, or within demographic groups; the cultural and societal forces that impact literary style and the evolution of style; the cultural, historical,

Digital genre stylistics can thus refer to both systematic and historical concepts of genre. An important question is whether one can speak of literary genres in the same sense for the text groups constituted or characterized by digital text analyses as is done in traditional literary genre studies. This thesis argues that literary theory's notions of genre are not directly transferable to digital genre stylistics, and that the field needs its own conceptual set with which to meaningfully describe its questions, methods, and results. Such a conceptual set also has the task of clarifying the relationship to existing concepts from other fields, that is, especially literary studies, linguistics, computational linguistics, and computer science. In the following, existing conceptual systems on textual and communicative, theoretical and historical dimensions of genre are discussed, and a conceptual system of our own for digital genre stylistics is proposed.

2.1.3.1 A Conceptual Proposal for Digital Genre Stylistics: Literary Text Types, Conventional Literary Genres, and Textual Literary Genres

In digital genre stylistics, it has sometimes been proposed that one should not speak about the analysis of "genres" in this case, but of "text types". This is because of the focus of digital genre stylistics on the literary texts themselves and more specifically on their linguistic surface and style, and the relatively limited inclusion of information that is related to the literary-historical and social context. At the same time, genre is a concept that appears to be strongly influenced also by extra-textual historical factors. On the one hand, the terminological distinction between genres and text types has its origin in linguistics, where literary genres are differentiated from linguistic text types. On the other hand it has also been proposed in literary genre theory itself as a means of distinguishing between genres that are based purely on textual and linguistic criteria and historical genres. This is similar to Genette's proposal to differentiate between *mode* and *genre*. In the context of text linguistics, text types (in German, "Texttypen") are described as follows:

> Texte werden zu Texttypen zusammengefasst auf der Grundlage linguistischer Kriterien. Texttypen verlaufen quer zu den Textsorten in verschiedenen Kommunikationsbereichen. Als linguistische Kriterien gelten dabei textinterne Merkmale (Merkmale der Textinfrastruktur) wie Stil (Stiltyp, z.B. Ironie, Nominalstil), Medialität (medialer Typ, z.B. digitaler Text, konzeptionell mündlicher Text), Textfunktion auf der Basis sprachlicher Indikatoren (Funktionstyp, z.B. Kontakttext), Themenentfaltung/Vertextung (Vertextungstyp, z.B. explikativer Text). (Gansel 2011, 13)

---

and societal linkages that bind or do not bind individual authors, texts, and genres into an aggregate literary culture; the waxing and waning of literary themes; the tastes and preferences of the literary establishment and whether those preferences correspond to general tastes and preferences" (Jockers 2013, 24). Jockers analyzes novels in English based on metadata and full texts, and he repeatedly points out how they develop historically. For example, he trains models for specific decades and analyzes which texts from other decades are stylistically similar to the initial ones, finding approximately thirty-year generations of style. He also finds correlations between the publication dates of novels and their subgenres and subsequently traces the signals of genre style throughout the nineteenth century (Jockers 2013, 82–89).

Style is directly mentioned as a defining characteristic of text types in the linguistic sense. However, the linguistic term that is used in a way that can be compared to the general conception of literary genres is also "text type" (in German, "Textsorte"):

> Wir definieren Textklasse als das Vorkommen einer Menge von Texten in einem abgegrenzten, durch situativ-funktionale und soziale Merkmale – also textexterne Merkmale – definierten kommunikativen Bereich, in dem sich Textsorten ausdifferenzieren. [Z. B. Textklasse] Religion – [Textsorten] Predigt, Ordensregel, Enzyklika oder [Textklasse] Politik – [Textsorten] Koalitionsvertrag, Parteiprogramm, Regierungserklärung. (Gansel 2011, 12–13)

In the latter sense, which is predominant in text linguistics since the "pragmatic turn", the linguistic text types are also primarily determined by text-external factors that characterize the communicative situation, and they are understood as norms and conventions, which influence how texts are produced and received (Gansel 2011, 8–10; Krieg-Holz and Bülow 2016, 220). These two uses of the term "text type" should thus not be confused. The terminological differentiation between pragmatically determined text types and syntactically and semantically distinguishable ones is also recognized in computational linguistics, where it is labeled as the opposition between genres and text types. It is probably through this influence that the conceptual separation between genres and text types has also been taken up in digital genre stylistics. In his computational study of linguistic genre variation, Biber, for example, declares:

> I have used the term 'genre' (or 'register') for text varieties that are readily recognized and 'named' within a culture (e.g., letters, press editorials, sermons, conversation), while I have used the term 'text type' for varieties that are defined linguistically (rather than perceptually). Both genres and text types can be characterized by reference to co-occuring linguistic features, but text types are further defined quantitatively such that the texts in a type all share frequent use of the same set of co-occurring linguistic features. Because co-occurrence reflects shared function, the resulting types are coherent in their linguistic form and communicative functions. (Biber 1992, 332)

Biber says that also genres are related to the occurrence of specific linguistic features, but in a way that is less consistent than in the case of functionally determined text types. Biber determines the text types in a bottom-up approach by using factor analysis and interpreting the resulting dimensions in terms of linguistically expressed communicative functions. Biber finds five dimensions of variation (informational versus involved, narrative versus non-narrative, elaborated versus situated reference, overt expression versus persuasion, and abstract versus non-abstract style) on which he bases his definitions of text types (Biber 1992, 334–335, 339–340). In their computational linguistic approach to the detection of text genre, Kessler, Numberg, and Schütze, on the other hand, only use the term "genre", which they define broadly as encompassing both literary and non-literary texts: "We will use the term 'genre' here to refer to any widely recognized class of texts defined by some common communicative purpose or other functional traits, provided the function is connected to some formal cues or commonalities and that the class is extensible" (Kessler, Numberg, and Schütze 1997, 33). Both Biber and Kessler, Numberg,

and Schütze recognize that genres are difficult to grasp on the formal linguistic level, but they expect (as Biber) or require (Kessler, Numberg, and Schütze) that some common formal elements can be found for texts that have been associated with the same genre.

In linguistics and computational linguistics, the discussion of the distinction between genres and text types focuses on the differences between conventional and pragmatic characteristics on the one side and structural linguistic features of text groups on the other side. In literary genre theory, in contrast, the point of debate in this terminological question is more oriented towards the systematic versus historical nature of the objects. Fricke, for instance, argues that it is not necessary to abandon neither the function of generic terms as names for historical groups nor their use as classificatory terms but that they should be distinguished. He differentiates a "literary text type" ("literarische Textsorte") from a "genre" and defines the first one as a purely systematic term to categorize literary texts and the latter as a term for historically bound and delimited literary institutions. To determine if a text belongs to a text type, its grammatical, semantic, and textual-pragmatic functions must be analyzed (Fricke 1981, 132–133). According to Fricke, more criteria need to be fulfilled in order to attribute a text to a genre. First, the text needs to conform to a clearly distinguishable literary text type. Second, the literary text type has to be established in the national literature of the text in question when the text is created, so that it corresponds to the expectations that contemporary readers have regarding the characteristics of the literary text type. Third, the text needs to explicitly carry an established name of the literary text type or exhibit other established signals for it. Finally, Fricke also uses the term "Gattung" (German for "genre") as a general term that can be used to both designate literary text types, genres, or any other establishment of groups of literary texts (Fricke 1981, 132). Fricke's proposal to speak of literary text types is fruitful for cases in which literary texts are classified based on syntactic, semantic, and pragmatic linguistic properties or any other textual characteristics without establishing any direct connection to a specific historical genre discourse. This use of the term "text type" is congruent with the definition of "Texttyp" in general linguistics and "text type" as used by Biber. However, Fricke's definition of "genre" is very strict. According to his concept, genres are a subset of the text types that they correspond to, because a common text type is a prerequisite. In addition, the text type has to correspond to contemporary expectations, which means that only generic conventions that are realized as distinguishable text types actually represent genres. Fricke does not say how he would call generic expectations or norms that are not mappable to specific combinations of textual features in a consistent way. Moreover, distinctions of literary genres made after the text's creation, whose participation in a genre is analyzed, are not called genres if the genre concept had not been established before. For example, the *gaucho novel* would not be a genre of nineteenth-century Spanish-American novels if it could not be verified that the term or the concept already existed as a norm that was conscious to the historical authors, publishers, readers, and critics. Furthermore, texts that are similar to others because of linguistic criteria but do not carry explicit or established implicit generic signals would also be considered as not participating in a genre. It makes sense not to assume a common genre simply because of textual similarities – one could, for example, find similar verse structures in the poetry of different centuries and continents by accident, for which no historical relationship could be verified. Nevertheless, if a text was created in the same context as a group of other texts that carry an explicit genre name and they all share textual features, one might argue that

the non-labeled text also participates in the genre. Determining where the same context begins or ends is another question that needs to be solved. One can also not exclude the possibility of cases of generic references spanning different temporal, geographical, or linguistic contexts, even if this is not the usual case. A wider definition of genre would be more appropriate for the purpose of digital genre stylistics. Such a definition should be able to relate conventional or critically established generic groupings of different kinds with findings of groups on the textual level without restricting the kinds or relationships.

Examining Fricke's definition of genre further and following his explanations, even his definition of text types appears too narrow for digital genre stylistics. In requiring the congruence of texts that share the features of a certain text type and that belong to a certain genre convention, Fricke already assumes that the text types are linked to generic expectations. Any arbitrary text classification could not be expected to be congruent with the genre concepts if not by accident. Thus Fricke requires a precise and systematic definition of a text type that can serve as a starting point for an empirical study that aims to verify if the text type was institutionalized as a genre in some historical period (Fricke 1981, 133). The text types should not be defined arbitrarily but according to their scholarly appropriateness: "Um ihrer heuristischen Eignung willen wird man die Textsortenbegriffe folglich in der Regel so festlegen, daß sie aufgrund hypothetischer literaturgeschichtlicher Vorannahmen voraussichtlich auf historisch belegte Texte zutreffen und daß nach Möglichkeit sogar irgendwann einmal ein dieser Textsorte entsprechendes Genre bestanden hat" (Fricke 1981, 134). Defining text types beforehand that probably correspond to historical genres requires a deductive procedure starting from assumed prior definitions of these genres. As was already discussed in the previous chapter, digital genre analyses can be conducted in a number of different ways: they can start with the formulation of theoretically based hypotheses about genres, which are then formalized, but they can also start from genre labels that stand for conventional ideas of genres and directly test if and how they relate to textual characteristics, without following a continuous chain of formalizing steps. The hypotheses about the relevance of specific textual features for genres can be quite loose. A strength of the computational approaches is specifically that series of different hypotheses can be tested and also that unforeseen results can be achieved through experimentation. In many cases, the formalization of literary concepts is not so mature that it would allow for continuous deductive procedures. Consequently, text types found through digital stylistic analyses do not necessarily have to be in line with literary-theoretical text types or historical genres. Rather, what can be looked for are the points of intersection between generic conventions and groups established based on textual features. They can be approximated to each other, but it should not be expected that they depend on each other or that there is a relationship of inclusion between them. I propose to widen the understanding of genre, but in turn to differentiate between two kinds of genres to establish a connection to text types. For the purpose of digital genre stylistics, as it is conducted in this dissertation, the following working definitions are proposed:

- A *literary text type* is an intensional, systematic term used to designate groups of literary texts that are established on the basis of common or similar immanent textual features and feature distributions of any kind and that can be distinguished from other literary text types based on these features and feature distributions.

- A *conventional literary genre* is a term referring to the extension of genres as historically bound literary institutions (in the sense of Voßkamp 1977) and as codifications of discursive expectations towards literary texts that participate in the genres (in the sense of Jauß 2014). The term "conventional" is used in a broad sense here and refers to historical as well as modern conventions. It includes socially and communicatively established genre conventions but also conventions of how texts are to be systematically assigned to genres, e.g., by librarians or literary scholars, because considered on a large scale and across time, systematic approaches are also historically bound.
- A *textual literary genre* is the convergence or intersection of a conventional literary genre and a literary text type. A literary text type and a conventional genre can be congruent to a certain degree, depending on the extent to which the groups of texts participating in them coincide. If the perspective starts from the conventional genre, it may be *textually coherent* to a certain degree. This means that a certain percentage of the texts that are attributed to the conventional genre are also part of a corresponding text type. With a correspondence of more than 50 %, one can speak of a certain textual coherence of the conventional genre. A text can participate in several conventional genres, and it can also belong to several text types. As a consequence, a text can be associated with no, one, or several textual genres.

In what follows, these determinations are explained and justified in more detail. With the differentiation between conventional and textual literary genres, the genres can be both detached from specific textual features or linked to them. To find out about the textual literary genres in which literary text types and conventional literary genres overlap is then a central task of digital genre stylistics. All kinds of relationships between conventional genres and text types can be assumed. For example, an extreme case would be a generic norm that has been discussed in poetological writings and can be described as a conventional literary genre, but that has never been realized on a textual level through a set of literary works. At the other extreme would be a text type that is recognizable in terms of recurrent patterns of feature distributions and that can be distinguished from other text types based on these features, but that has never been named or discussed as a conventional genre. Between these two extreme cases, many other constellations are possible, for example, groups of texts that carry signals of a particular conventional genre but of which only a part is coherent on the textual level. Another possible case is a conventional genre, which is held together by a common name but which relates to several different text types so that it can be described as several different textual genres. Such a result could stimulate further investigation into the conventional genre to see if there are signs of several subtypes.[52] Also possible are cases where two different conventional genres relate to

[52] An example of such a constellation is the German novella, which Schröter described as an instance of "historically discontinuous and heterogeneous genres" (Schröter 2019, 227). Rigorous poetics of the novella existed, but these were not consistent with the texts called novellas. Furthermore, the texts carrying the label "novella" could not be described as a homogeneous group (Schröter 2019, 229). Schröter proposes to define "intersections" between the texts referred to as novellas and classificatory sets of texts so that, for example, texts with the name novella that were published as fictional journal prose and share the characteristics of the latter type of text would form one subtype of the genre novella. Schröter states: "It is important not to summarily define such intersections as novellas, as is often done. Such a definition would result in a classificatory concept of the respective genre, which in turn would no longer be suitable for comprehending the historical use of the generic label and hence the

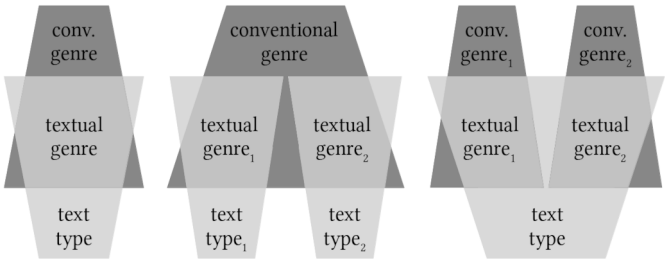


Figure 1. Relationships between text types, conventional genres, and textual genres.

the same text type. For example, it could be analyzed if different terms of shorter narrative prose in the Spanish-American nineteenth-century tradition that do not seem to be limited very clearly as conventional genres ("narración", "relato", "cuento", "novelita") have a common textual basis. Three of the many possible constellations between text types, conventional genres, and textual genres are illustrated schematically in figure 1.

The first case on the left side is one where a conventional genre and a text type overlap to a certain degree so that one textual genre can be identified. The parts of the conventional genre that lie outside the textual genre refer to literary texts that have been marked as being part of the convention but that do not conform to the text type. On the other hand, there are texts whose characteristics correspond to the text type but that have not been assigned to the conventional genre. In the three examples shown in the figure, there is no relationship of inclusion between the conventional genre and the text type, but this would, of course, also be possible. In such a case, all the literary texts that conform to the text type could also be part of the conventional genre, only that some other texts that are associated with the conventional genre are not congruent with the text type (inclusion of the text type in the conventional genre), or all the texts that participate in the conventional genre are in line with the text type, and, in addition, there are texts that fulfill the criteria of the text type but that have not been marked as belonging to the genre by convention (inclusion of the conventional genre in the text type). The second case in the middle of the figure illustrates a relationship of overlap between one conventional genre and two different text types, so that two textual genres are identified. In such a setting, it is conceivable that one genre name is used for different text types that can be distinguished either synchronically or diachronically. The third case to the right shows how two conventional genres, which are, for example, characterized by two different genre names, can be covered by the same text type. Here, too, there result two different textual genres. An example of this constellation would be if the same textual characteristics are found in different historical contexts and different genre names are used in these contexts to refer to them. In the schematic cases shown here,

---

semantics of a genre in literary-historical communication." (Schröter 2019, 228). Using the terminology proposed here, the intersection of texts referred to as novellas and fictional journal prose would be called a "textual literary genre", whereas the semantics of the novella in literary-historical communication would be referred to as the "conventional literary genre". The novella, as a conventional literary genre, would then consist of several different textual genres.

only up to two conventional genres or text types are involved, but there could, of course, also be scenarios where more parts are involved.

What is the status of generic signals that relate literary works to genres in this setup? Explicit genre names, implicit textual genre signals such as conventionalized titles or names, or beginnings of the texts[53] are understood as belonging to the level of conventional genres. The presence of such signals does not necessarily mean that a text conforms to a specific textual genre, and even their absence does not mean that a text cannot be described in terms of the same text type that is connected to a textual genre. An unconventional use of generic signals, for example, a subversive or an ironic one, can be detected when there is a discrepancy between the individual text carrying the signal and other members of a textual genre.

A question that needs to be addressed is how assignments of literary texts to genres that are not established by paratextual or textual signals but text-externally relate to the three terms defined above. Part of this question is how generic terms and definitions of genres, on the one hand, and the text types, conventional, and textual genres, on the other hand, behave in relation to the literary system, literary theory, and literary history. As Schaeffer notes: "les termes génériques on un status bâtard. Ils ne sont pas de purs terms analytiques qu'on appliquerait de l'extérieur à l'histoire des textes, mais font, à des degrés divers, partie de cette histoire même" (Schaeffer 1983, 65). What Schaeffer means is that also terms that are defined externally are part of the history of genres. Very important is his remark that generic terms do so to *different degrees*. Furthermore, the degree also depends on the perspective. If a nineteenth-century Mexican editor labels several books that are published for the first time as "novelas", he applies this term collectively to several works. He thus has an understanding of the genre that already involves a systematic element – he compares the books to his concept of the genre *novela* and decides if they fit into the category or not. However, if we analyze these nineteenth-century novels today, we perceive the editor's decisions as part of the history of the genre and his opinion as formed by the convention of the time. The labels would therefore be signs of the conventional genre. In addition, the understanding that an editor has of a genre is probably not much formed by theoretical considerations. If, instead, a contemporary writer expresses his views on what the *novela* is in poetic terms and lists several works that conform to his definition of the genre, this, too, would represent a systematic and possibly also theoretically motivated approach to the genre.[54] To his or her contemporaries, the initiative could appear as not being part of the genre convention but rather as an attempt to define a textual genre. For the literary historian, again, it would clearly contribute to the genre convention of the time, even if the historical, poetic definition of the genre is by degree more systematic and theoretical than the practice of individual writers or editors who provide the texts that they write or publish with generic labels. Yet another case is that of a modern librarian who classifies historical texts by genre. If one librarian does this, a unique concept of the genre (the one the librarian has) is applied collectively to a set of texts. If several librarians classify the texts together, multiple genre concepts are

[53] For a short description of the usual types of generic signals, see Fricke (1981, 135). A more comprehensive overview is given in Fowler (1982, 88–105).

[54] A Mexican writer who reflected on the novel's role in Mexican national literature was, for example, Ignacio Manuel Altamirano, who expressed his ideas in the essay "Revistas Literarias de México" (Altamirano 1868).

involved.[55] The labeling that the librarians do is not part of the historical genre convention of the context in which the literary texts were initially created and published. Instead, it is part of the modern genre convention, even if that convention has a systematic background and is informed by literary studies. This is the more obvious the more different people participate in applying a specific concept of genre. Finally, literary-theoretical or literary-historical positions *can* also be perceived as conventional if they are analyzed collectively. When the genre assignments of twenty different literary scholars are examined together, it is not very probable that they all share the exact same theoretical understanding of the genres in question. The sum of their judgments constitutes itself a theoretically based genre convention. Conventional genres can thus involve a systematic and theoretical element.[56] The more weight that systematic and theoretical element has, the more probable it is that the genre convention largely correlates with a text type. On the other side, historical poetical genre definitions as well as modern literary theoretical ones could also be understood as theoretical entities and be used as hypotheses about textual genres. They could then be tested by applying them to the texts that have been classified as instances of the genres based on the respective genre definitions.[57] Just as genre conventions are not understood as purely historical, also text types are not conceived as purely systematic or theoretical entities here. Whenever text types are derived from a corpus of literary texts, they involve a historical element because the texts were produced in a specific historical setting. Following Hempfer's assumption that generic norms and expectations enter literary texts and are readable from them as normative facts, the conventional genres influence the text types. Otherwise, there would not be any textual genres at all. More distant to the historical genre conventions are text types that are defined on general theoretical principles without recourse to specific text corpora, for example, text types that are formulated according to a general theory of language or literature. Similar to theoretically defined genres, these could be used as testable hypotheses with regard to literary text types, i.e. to check to what extent they represent actual textual patterns. When such

---

[55] The terms "collectif" versus "individuel" and "unique" versus "multiple" are introduced by Schaeffer when discussing the status of different kinds of generic terms. Furthermore, Schaeffer distinguishes between "noms génériques endogènes", which are used by authors or their public, and "nom génériques exogènes", which are established by literary historians. The generic terms can have a textual (attached to the literary text, for example, as a paratextual element) or a meta-textual status (if they are used as terms to discuss a work but are external to it), and the functions of the labels vary depending on the category they belong to (Schaeffer 1983, 65, 77–78).

[56] Rather than on the question of historical convention and institutionalization, Schaeffer focuses on the place and concept of the genre names in the communicative situation. He uses the property of all texts as speech acts as an argument for characterizing genres as more analytical or historical: "Je pense qu'il faut aller plus loin: les genres théoriques, c'est-à-dire en fait les genres tels qu'ils sont définis par tel ou tel critique, font *eux-mêmes* partie de ce qu'on pourrait appeler la logique pragmatique de la générécité, logique qui est indistinctement un phénomène de production et de réception textuelle. En ce sens on peut dire que l'*Introduction à la littérature fantastique* de Todorov est elle-même un des facteurs de la dynamique générique, à savoir une proposition spécifique pour un regroupement textuel spécifique et donc pour un modèle générique spécifique […]" (Schaeffer 1983, 68).

[57] Schaeffer too argues that there should not be a strict separation of theoretical and historical genres and genre labels, although he recognizes that they follow quite different rules: "le système des genres théoriques, construit à partir d'oppositions différentielles, simples ou multiples, obéit à des contraintes de cohérence qui ne sont pas celles des genres historiques (quelles que soit la réalité de ces genres désignés par les noms de genres traditionells)" (Schaeffer 1983, 67). Schaeffer demonstrates why it is not useful to assume a direct deductive relationship between both. However, if theoretical definitions of genres are used as hypotheses about textual genres, no claim about the integrity of the conventional historical genres is made, so that such a deductive procedure seems viable.

theoretical text types also refer to generic conventions, they are theories about textual genres. That is, theories of literary textual genres can be more inclined towards genre conventions – if they concentrate on historically bound discursive expectations towards literary texts and relate them to textual characteristics – or they can be inclined towards text types if they start from systematic definitions of text groups and relate them to conventional generic terms.

Which consequences does this terminological system have for digital genre stylistics? First: there is room for experimentation. For Moretti, through digitally-based, formal, and quantitative analysis, literary criticism is transformed into an experiment:

> a few years, and we'll be able to search just about all novels that have ever been published, and look for patterns among billions of sentences [...] By sifting through thousands of variations and permutations and approximations, a quantitative stylistics of the digital archive may find some answers. It will be difficult, no doubt, because one cannot study a large archive in the same way one studies as text: texts are designed to 'speak' to us [...] but archives [...] say absolutely nothing until one asks the right question. [...] it turns criticism on its head, and transforms it into an experiment of sorts: 'questions put to nature' is how experiments are often described, and what I'm imagining here are questions—put to culture. Difficult; but too interesting not to give it a try. (Moretti 2008, 114)

By not presupposing theoretically how text types should be constituted to match historical genre conventions in the best way and by not demanding a relationship of inclusion between text types and conventional genres, experiments with different kinds of textual features are possible and can lead to new insights about textual genres. One possible analysis scenario is outlined here:

- generic labels for the texts in a corpus are collected from different modern library catalogs
- a most frequent words-based classification is used to find out how well the texts can be classified by the genre labels
- it is determined which features are most important for the textual distinction between the different genres.

In this scenario, the goal is to work out the degree of textual coherence of the conventional genres and to define the text types underlying the textual genres based on a general set of linguistic features. Texts that are repeatedly misclassified can be assumed to be part of the genre convention to which they were assigned but not of the textual genre to which the genre label in question is primarily attributed. If such texts are assigned to another class several times by the classifier, it can be checked if they are described better by another text type.

However, the distinction between literary text types, conventional genres, and textual genres does not only allow for testing how relevant certain general textual features are to capture the connections between the levels of text and convention. On the other hand, one can as well start from theories of textual genres. Such an analysis scenario is sketched here as an example. For instance, it would be possible to depart from a definition of the historical novel as a novel that is characterized by

- the temporal distance between the writing, publication, and reception of the novel and the past in which the narrated events take place,
- the co-occurrence of invented and historical personages, places, and events,
- and the localization of the narrated events in a precise historical past.[58]

A specific corpus of historical and non-historical novels could be built by using specific kinds of genre labels, for example, indications of the subgenre on book covers. These then function as markers of the conventional genre against which the definition of the textual genre is checked.[59] It would then be necessary to formalize the elements that are part of the literary theoretical definition. In this case, it would not be enough to just use the most frequent words because the deductive chain from the genre definition to the textual features needs to be maintained. Instead, named entity recognition could be used to detect the mention of historical figures in the novels, combined with their identification through authority files. The detection of temporal expressions with temporal taggers could be used to check if the historical novels are characterized by more frequent uses of explicit temporal expressions than other subgenres of the novel that are represented in the corpus. As a result, it could be verified how many of the novels that were marked as historical novels by convention are captured with the text type that is derived from the theoretical definition of the textual genre. In this case, a challenge would be to quantify the boundaries between historical and non-historical novels based on the literary theoretical criteria and to find a way to delimit the text type. This is because the criteria are *positively* formulated and refer to the *presence* of features (for example, "there *is* a temporal distance between the publication date of the novel and the narrated past"). However, it is not said how much of the feature should be present (how large should the temporal distance be? 20 years, or 50, or hundreds of years?) and how the same features are distributed in texts that are part of other genres (are they totally absent there – no temporal distance at all – or just less frequent or less strong – only up to 10 years of temporal distance?). This example shows how difficult it can be to directly map literary theoretical assumptions about genres to a digital quantitative genre analysis. In the above example, possible strategies could be to cluster the texts based on the chosen features and, if several clusters can be clearly distinguished in the data, to interpret them as text types. It could then be checked on which feature ranges and distributions these text types are based and where the boundaries between them are in terms of feature values. In a next step, it could be analyzed how the texts associated with the conventional genres are spread over the different clusters to see if the theoretical assumptions about the relevance of the three text characteristics for the distinction of historical from non-historical novels on a textual level hold. Still, at least two problems would persist in this case. The first problem is the one of underspecification of the assumptions. How can the theory of the historical novel be confirmed or rejected if the decisions about quantifying the definition's different parts are part of the analysis results but not part of its starting point? One could always say that the theory may be valid or invalid for other quantifications. If not to confirm or reject a specific definition of a textual genre, the results of the

[58] For definitions of the historical novel in which the mentioned characteristics play a role, see, among others, Fernández Prieto (1996), Lefere (2013), Lukács (1955), Maxwell (2009) and Spang (1998).

[59] In this case, it would be important *not* to derive the labels *directly from* the theoretical definition of the textual genre in order to avoid circular reasoning.

quantitative analysis could still be used to refine the theory by providing the specification of its assumptions in numerical terms. The second problem that persists is that it would be necessary to define a limit for the ratio of conventional historical or non-historical novels that must be present in the found text types in order to say that the text types correlate with the conventional genres. It is implausible that 100 % of the novels with the conventional label "historical" would end up in the same cluster and none of them in the other(s). Is 90 % enough to say that the theory is adequate for the corpus at hand and describes the textual genre satisfactorily, or is just 60 % enough or only 51 %? The same problem would remain if classification was used instead of clustering. Again, instead of confirming or rejecting the hypotheses entirely, the results could be used to describe the extent of the intersection between the conventional genre and the text type.

There is another consequence of the terminological distinction between literary text types, conventional genres, and textual genres for a corpus-based digital genre analysis aiming to mediate between genre conventions and text characteristics. It is important to clarify and explain several aspects of the data and the analysis:

- the text selection,
- the kind of genre convention and / or genre theory that is addressed,
- the assignment of genres to the works in the text collection,
- and the modeling of factors other than genre that are assumed to have an influence on the distribution of textual features.

These requirements are, in essence, not different for a non-digital, non-quantitative, and non-stylistic analysis of genres, but the ways in which they can be addressed are different. In quantitative text analysis, a genre is determined by contrasting and distinguishing texts associated with it from other texts – otherwise, the analysis would concentrate on the internal structures of the genre but not on its defining characteristics. Consequently, a corpus for the analysis of a specific genre always contains also texts that are not assumed to participate in that genre. This can facilitate the text selection because what has to be defined is the initial broader context of the genre. For example, an analysis aiming to find out about seventeenth-century tragedies can rely on a corpus of seventeenth-century dramatic texts, including comedies, tragicomedies, and other dramatic subgenres. Therefore, it is not all-decisive which texts exactly are assumed to be associated with the genre of interest at the outset of the analysis. A prior definition of the genre that is to be determined by the text analysis is not mandatory. Instead, the extension of the genre can be approached by collecting texts that are similar to the object of interest. In the case of the tragedies, these would be texts that were written and published in the same historical period, and that participate in the same major genre. What is important, though, is to remember that any delimitation of a textual genre that results from such a corpus-based contrastive analysis is relative to the texts that are part of the text collection but are not considered instances of the genre. If the tragedies are only compared to comedies, they are determined as non-comedies. This means that the problem of genre definition is relocated to a more general level, by selecting the wider context of texts in which instances of the textual genre of interest are assumed to be found. *Inside* the contextual text collection, the search for intersections between conventional genres and text types, and thereby the definition of textual genres, is possible, and its results are not predetermined.

Clarifying the target genre convention and making the procedure of assignment of genre labels to the texts in the collection transparent is crucial if the digital stylistic analysis aims to contribute to existing literary-historical research and if it seeks to find out about genres and not only text types. For very large corpora, assigning genre labels by hand may be unfeasible. One option is to use available bibliographic sources in digital format that are processable with scripts and to collect genre labels automatically. Even if labels are assigned individually and manually to the works, it is often necessary to rely on several sources. Single sources usually do not cover all the texts in a corpus, so different kinds of genre conventions may be involved. For example, the genre labels can originate from scholarly handbooks or library catalogs, thus involving two different kinds of systematic approaches to the genres. Even if several sources are used, it is probable that there remain texts of the broader context whose generic status cannot be clarified beforehand, for instance, if they have not been in the focus of scholarly, librarian, or public attention at all. To what extent historical labels can be used depends on the availability of historical editions and book covers. Because the use of large corpora makes it more difficult to pursue a uniform strategy for genre assignments, it is even more important to set out in detail how the task was solved and to discuss the consequences for the textual genre that is determined in the analysis.

As the text types in digital genre analyses are determined through text style, it is important to model factors which influence it, and this is also connected to the question of corpus building. Ultimately, the goal is not just to find arbitrary text types but to delimit textual genres in which textual similarities and communicative generic norms converge. How to find out if the facts found in the texts actually represent normative facts that are due to genre conventions and not traces of other intra- or extra-textual factors? Every genre analysis, a stylistic and also a non-stylistic one, has to make sure that the text corpus does not only consist of texts written by one or a few authors so that the claims built on the corpus analysis are not only about individual authors but about the genre. The issue is, however, more complicated in a stylistic analysis because all characteristics of the texts that are taken into account, modeled, or found lead to the text style or are derived from it. For instance, authorship can be primarily marked by a certain authorial style, and that style can be captured through the analysis of high-frequency words. Yet, the same words can also be properties that give evidence of some higher-order structural concept. For example, the type of narrator that is used in a narrative text can influence the number of specific personal pronouns. There can also be properties that are related to the textual genre, for instance, many general descriptive passages in novels of customs, which lead to a comparatively higher number of undetermined articles. To make sure that it is the genre convention that is captured with the overlap of genre labels and text types, it is, therefore, necessary to control and check possible interfering factors in the text style by modeling them as metadata and considering them in the process of text selection.

2.1.3.2 Text Types, Conventional Genres, and Textual Genres in Semiotic Models of Generic Terms

Several aspects remain that need to be clarified in connection with the systematic or historical status of text types and genres. In chapter 2.1.2.1 above, there was a focus on semiotic theories of genres and on generic terms. How can these models be understood in connection with text types, conventional genres, and textual genres? When they are used to analyze historical generic terms, they approach these with a language-based, communicative theory to find out which levels of discourse the terms address. This means that the norms that a conventional genre involves are described as characteristics of texts as speech acts. Following Raible's model, for example, the subgenre *novela documentaria* would point to the level of the relationship between text and reality (a documenting activity purports that reality is depicted) and the level of the kind of linguistic representation (a documentary style is likely to be neutral and descriptive). These characteristics can be taken as hypotheses about the text types that relate to the conventional genres and thereby also as hypotheses about the features of textual genres, considering that for a stylistic analysis, all these hypotheses would have to be broken down to the level of linguistic representation. Furthermore, the semiotic models of genre constitute a possibility to differentiate between the various discursive levels on which sets of genres are defined, which can help to define more meaningful comparisons of genres. In the case of the novel, for instance, there is a wealth of subgenres, some of which emphasize similar aspects of speech acts and others entirely different ones. The more different the discursive levels are that the generic terms point to, the more probable it is that also the corresponding textual genres are defined on different levels, that the groups of texts associated with the genre conventions overlap, and also that the generic conventions might be different in kind. For example, it seems useful to compare sentimental novels to historical novels if both are interpreted as being defined primarily on a thematic level. On the other hand, it is less fruitful to directly contrast sentimental and romantic novels because the latter often (but not always) have a sentimental theme. In addition, the label "romantic" points to a certain aesthetic movement that was dominant in a specific period of time and not necessarily to thematic elements. This is not to say that sentimental or historical novels are only defined thematically, only that comparing these subgenres makes the most sense on a discursive level that is a prominent reference point for both.

A closer look at Schaeffer's explanations is useful to further understand how the semiotic approach to genre relates to the three terms of text type, conventional genre, and textual genre proposed here. In the last part of his book "Qu'est-ce qu'un genre littéraire?", he discusses generic regimes and logics, by which he means the kinds of relationships that individual texts have with "their genre", and how these regimes are connected to the generic names (Schaeffer 1983, 156). Basically, he distinguishes between the relationships of *exemplification* and *modulation*. By exemplification, he means that texts instantiate genres globally (the whole text as a communicational act represents the genre, and not only parts of it) and without modifying the genre itself (the text is only an example of the genre, the genre stays the same). The discursive level to which generic names refer in such cases is the communicational act. In that case, the genre is not defined in textual (syntactic and semantic) terms but in pragmatic ones, concerned with the discursive attitude and the global intention. The type of convention that is involved in such

cases is a constituting one. That is, one cannot depart from the convention to a certain degree but only fulfill it completely or fail to do so entirely. According to Schaeffer, cases of genres that can be exemplified are, for instance, drama or narration. If a literary work is conceived as a drama, that is the case independently of its exact form. Even if it contains narrative passages, it is still a drama, because the genre is not understood as being dependent on the actual textual realization but on the outer communicative level (Schaeffer 1983, 156–164). Schaeffer's regime of exemplification is a bit difficult to grasp. It seems close to the concepts of *mode* and *Schreibweisen* that Todorov, Hempfer, and Genette develop because it refers to basic discursive attitudes that are linguistically motivated and tend to be historically invariant. However, Schaeffer uses nouns and not adjectives when he refers to this type of genre, and he relates the terms to literary works as a whole, which makes it more challenging to differentiate the *genres by example* from the historical genres. Schaeffer relates this generic regime to a "constituting" convention.

The second main type of generic logic that Schaeffer describes, the *modulation*, is, on the other hand, concerned with the syntactic and semantic entities that texts are. As soon as these discursive levels are addressed, Schaeffer claims, the relationship to the genre cannot be simply exemplifying anymore. Except for the special case of exact copies, each text is individual in this respect and always modifies the genre with which it is associated. In the case of modulation, the determination of the genre depends on the individual works and, thereby, also on the specific historical context in which the works are situated. The same generic names "drama" and "narration" can also be examples of the modulating logic if they refer to specific textual elements such as character constellations, themes, or elements of style. Schaeffer divides this regime further into

- a modulation by the application of rules,
- a hypertextual modulation based on historical genealogies,
- and a modulation by textual resemblance.

The first kind of modulation is related to a regulative convention, the second to a traditional convention, and the third to no convention at all (Schaeffer 1983, 164–180). Schaeffer summarizes his system of generic logics in a table, which is reproduced here in table 1. That way, it can be related graphically to the three terms of text type, conventional genre, and textual genre as they are used in this study.

Three colors are used here to highlight the parts of the table that are related to text types (orange), conventional genres (dark green), and textual genres (light green) – no colors are used in Schaeffer's original table. Text types correspond to the logic of *modulation by resemblance*, which leads to classes of similar texts that are defined statistically. Schaeffer has a double line between the last row of the table and the upper three ones, showing that he sees a clear difference between this generic logic and the other types involving different kinds of conventions. However, for the three generic logics of *global exemplification*, *modulation by application*, and *hypertextual modulation*, the conventional level is not clearly separated from the textual level. They mainly describe how that genre is constituted, for example, as a set of rules that are listed in a poetic formulation of a genre and that have a prescriptive character or as a heuristic description based on works that are traditionally associated with a genre. The two columns "relation" and "écart" (deviation), in contrast, describe to how a specific text instantiates a generic convention and how it may deviate from it. They can be interpreted in terms of the textual genre – as the point of

niveau	référent	relation	définition	description	convention	écart
acte communicationnel	propriété	exemplification globale	en compréhension	contrastive	constituante	échec
texte	règle	modulation par application	prescriptive	énumérative	régulatrice	violation
	classe généalogique	modulation hypertextuelle	heuristique	spécifiante	traditionelle	transformation
	classe analogique	modulation par ressemblance	statistique	typisante	—	variation

Table 1. Generic logics according to Schaeffer.

contact between the text type that the text in question belongs to and the conventional genre that it participates in, even if Schaeffer focuses on divergence (modulation, deviation, violation, transformation) rather than intersection. Nevertheless, in Schaeffer's approach, the levels of conventional genre and textual genre are intertwined. On the other hand, resemblance on the textual level is seen as not being related to the conventions at all, disconnecting the text types from the generic conventions. If one were to transfer this view to digital stylistic genre analysis, which is based on text features that are statistically evaluated, this would mean that it operates in another sphere, which has nothing to say about the literary-historical context of genres. Against this, the initial separation of the textual from the conventional level that is proposed in this study is not meant to isolate both concerns but to provide a more open basis for comparison so that the connections and disconnections of the two levels can be analyzed.

2.1.3.3 Literary Currents, Schools, and Movements

Another question that needs to be addressed is which kind of phenomena that are related to literary-historical conventions can be considered conventional *genres*. Is an aesthetic movement a genre? Literary genre theorists have different views on this topic. Todorov, for example, explicitly excludes literary movements from his concept of genre, but he does so because he presupposes textual coherence for genres, which according to him, cannot be expected with certainty for literary movements:

> Since a genre is the historically attested codification of discursive properties, it is easy to imagine the absence of either of the two components of the definition: historical reality and discursive reality. In the absence of historical reality we would be dealing with the categories of general poetics that are called – depending upon textual level – modes, registers, styles, or even forms, manners, and so on. The 'noble style' or 'first person narration' are indeed discursive realities; but they cannot be pinned down to

> a single moment in time: they are always possible. By the same token, in the absence of discursive reality, we would be dealing with notions that belong to literary history in the broad sense, such as trend, school, movement, or, in another sense of the word, 'style'. (Todorov 2014, 200–201)[60]

With discursive properties, Todorov means textual elements like phonetic or phonological features, thematic elements, and plot structure, but also aspects of the communicative level of texts, such as the factual or fictional status of the utterance (Todorov 2014, 199). Using the example of symbolism, he explains that it is known that this literary movement existed historically but that it is not proven that the works of authors identified with this movement (which would then be, for example, symbolistic poems) are characterized by common discursive properties. Instead, the movement may only have been based on friendships or manifests (Todorov 2014, 202). How can one be sure that there is no discursive reality or textual coherence behind works that are associated with a literary current? I am in favor of the position that this should not be ruled out in advance per definition but that it should be tested empirically, just as the textual coherence of conventional genres in the narrower sense should be examined. Digital genre stylistics provide good opportunities to do so. If there are explicit manifests and poetic writings that define literary schools and movements as frames of expectation for the creation and reception of literary works, this can be taken as proof of their relevance as historical, literary institutions. By locating literary movements on the level of conventional genres, nothing is said in advance about their relationship to text types and their significance as textual genres. However, for Todorov, like for Fricke, purely conventional genres are no genres, or at least they should not be called "genres": "Genres are the meeting place between general poetics [concerned with the theoretical definition of text types] and event-based literary history [concerned with the historical study of the expression of literary conventions]" (Todorov 2014, 201). Both Fricke and Todorov only see textual genres as genres, which shows that their genre concepts are primarily theoretically anchored, despite the references that they make to the relevance of a historical foundation for the definition of textual genres.

In the semiotic models of Raible and Schaeffer, literary movements are not directly included either. Even so, Schaeffer discusses their status in the explanations accompanying his core model of discursive levels to which generic terms point. He explains that there are generic terms that cannot be ascribed to the five levels of the verbal act that he defines, for example, terms that refer to the context, place, and time of the speech act. Schaeffer explains that he does not directly include them in his model but that numerous generic terms exist that refer to these aspects, for example, the "baroque sonnet" or the "Greek epic" (Schaeffer 1983, 117–118). According to Schaeffer, such genre names reinforce how important the historical context is for the determination of genres. Together with the multiple dimensions of the verbal act that are addressed by generic terms, they show that genres cannot be reduced to texts as entire, whole objects (Schaeffer 1983, 119). In the context of this thesis, it was noted that also many generic terms that are associated with nineteenth-century Spanish-American novels refer to time and

---

[60] He thus refers to textual coherence in a broad sense involving all the discursive levels of a speech act. The difference between factual or fictional utterances, for instance, would in fact be difficult to pin down to textual features in a narrower sense.

space. As an extension of Schaeffers model, the levels of temporal and spatial context are therefore included to organize the subgenre labels in the bibliography and corpus presented in chapter 3.[61] Here, references to literary currents are described as labels that relate to the temporal context because literary periods are often named after such movements, and the currents are usually phenomena that are temporally limited.

Fowler utters another critical view on literary movements in relation to genres. In his book on kinds of literature, he mentions them in a section called "Other types", which is part of the chapter devoted to the definition of modes and subgenres: "The system of generic categories is complicated by the existence of several other quasi-generic groupings. These include [...] the collective productions of 'schools' or movements (Metaphysical; Romantic; Georgian). They must be mentioned only to be dismissed" (Fowler 1982, 126–127). Fowler describes them as a "collective œuvre" and says that the features that literary works associated with schools or movements have may be "extensive and coherent enough to suggest genre" but that "they exhibit them independently of the historical kinds" (Fowler 1982, 128). With that, Fowler means that the common features that different works belonging to a particular movement share crisscross the boundaries of the genres that he calls the historical kinds. As an example, he mentions the Metaphysical poets who wrote in different poetic genres. Fowler states that these poems could be more similar to each other than works of the same poetic genre that belonged to various schools, for instance, different love elegies (Fowler 1982, 128). This is, however, only a problem if genre is not analyzed on different discursive levels and varying levels of generality and if it is assumed that there are no overlapping associations of literary works with different kinds of genres. If the love elegy is conceived as being primarily thematically defined (as involving the lament for a tragic love) and a metaphysic poem primarily formally – on the level of representation (as including innovative use of metaphors and allegories), then these two main characteristics do not exclude each other. The metaphysic poem and the love elegy can be understood as two different conventional genres overlapping with different text types. If Fowler says that metaphysical poems are more similar to each other than love elegies of different schools, this could even mean that the conventional genre "metaphysical poem" is textually more coherent than the conventional genre "love elegy". Fowler's argument is, therefore, no obstacle to also considering kinds that are conditioned by literary movements as genres.

A literary scholar who does not see a principle difference between literary genres and movements is Schlickers. In her book about the naturalistic Spanish-American novel, she argues that this distinction is not necessary. First, she confirms that the term "novela naturalista" indeed served as a generic name in the nineteenth century:

> La denominación del objeto de este estudio, sin embargo, no causa mayores problemas, porque la noción de 'novela naturalista' (y sus variantes) funciona(ba) como nombre genérico, indicando, pues, cierta clase de textos literarios y revelando así una conciencia histórica del género, que influía tanto en la producción como en la recepción de las novelas. (Schlickers 2003, 16)[62]

---

[61] See, in particular, chapter 3.2.3.6, where the empirically based discursive model of generic terms is described.

[62] This can be confirmed by the explicit genre labels found in the digital bibliography of nineteenth-century Argentine, Cuban, and Mexican novels created for the present study because the label "novela naturalista" is found in the

Then she discusses the status of the naturalistic novel as a historical and systematic category. According to Schlickers, there is no clear generic model for the naturalistic novel when compared, for example, to the picaresque novel, for which several constant and variable textual characteristics are known: "La novela naturalista, por el contrario, parece constituir más bien una corriente o un movimiento literario" (Schlickers 2003, 16). However, Schlickers says that all of the terms "current", "movement", "genre", and "poetic school" are debatable and have similar extensions, although different intensions. She is in favor of modeling the naturalistic novel as a subgenre because the works that are associated with this convention repeatedly reference the same prototypical works of the French naturalistic tradition. This shows that there was an awareness of the genre and that the works in question constitute a series of naturalistic novels. Schlickers states that conceptualizing the naturalistic novel as a subgenre allows one to analyze its distinctive features, which guarantee the coherence of the texts associated with Naturalism. At the same time, analyzing the naturalistic novel as a subgenre can serve as a constructive and heuristic means by which imprecise and contradictory textual features and purely poetological particularities can be considered (Schlickers 2003, 17). Here she assumes that there is a textual genre "novela naturalista", but that it is not entirely congruent with the conventional genre. The case of the Spanish-American naturalistic novels is only one specific historical example of a convention that is usually described as part of a literary school or movement, but that could also be conceived as a genre. In the present study, the example of the Spanish-American naturalistic novel is taken as an argument in favor of the possibility of understanding literary currents as a type of subgenre. Digital genre stylistics can contribute to expanding the limited knowledge about common textual characteristics of literary currents, and in particular of the naturalistic novels to which Schlickers points.

2.1.3.4 Genre Systems and Hierarchies

Discussing whether literary movements can be understood as genres or not leads over to the issue of theoretically separating genres from other discursive entities and of the place of genres in a system or hierarchy of forms – also beyond the question of the difference between text types and genres. There has been much research in literary genre theory on this topic, and a range of different terminological systems have been proposed. They cannot all be reviewed here, so only selected systems are presented shortly. It will be clarified how the different kinds of theoretical, generic terms that have been proposed relate to the three terms of text type, conventional genre, and textual genre defined above. Furthermore, as the empirical part of this study is concerned with subgenres of the novel, a point of interest is how genres and subgenres relate to each other. A terminological system that is prominent in the German-speaking context is the one

---

subtitles of three historical editions ("¿Inocentes o culpables? Novela naturalista" (1884, AR) by Juan Antonio Argerich, "Los bandidos de Río Frío. Novela naturalista, humorística, de costumbres, de crímenes y de horrores" (1892, MX) by Manuel Payno, and "Conventillo de intelectuales. Novela de índole rebelde y de género naturalista que no deben leer las almas timoratas" (1904, AR) by Francisco Guillo). It is also referred to in the prefaces of two of the naturalistic novels whose full text was examined ("El tipo más original" (1879, AR) by Eduardo Ladislao Holmberg and "Perfiles y medallones" (1886, AR) by Silverio Domínguez).

proposed by Hempfer.[63] On the one hand, he proposes to use the word "genre" (in German, "Gattung") as a meta-theoretical term which may include all other terms used to designate kinds of texts (Hempfer 1973, 16–18), as for example "meta-genre", "subgenre", "form", "kind", "mode", "species", "variety", "text type", or "text class". This general, meta- or pre-theoretical use of the term "Gattung" is what also Fricke (1981, 133) suggests. When no further theoretical distinctions are made, the term "genre" is used in this general sense here, as well. On a second level below the most general meta-term, Hempfer defines several theoretical terms for specific kinds of text groupings. The main components of his terminological system are "Schreibweise" ("diction"), "Typ" ("type"), "Gattung" ("genre"), and "Untergattung" ("subgenre"). "Dictions" are defined as ahistorical constants (e.g., "the narrative", "the dramatic", "the satiric"); "types" as trans-temporal forms of dictions, that is, as the theoretically possible set of types derived from them; "genres" as historical and concrete realizations of the general dictions (e.g., "novel", "epopee", "verse satire"); and "subgenres" as subtypes of single genres (e.g., "picaresque novel", "pathetic verse satire"). According to Hempfer, types and genres can both be derived from dictions via transformations, which means that his system builds on a dynamical and structural concept and not a hierarchical one. Genres and subgenres do not need to be based only on one diction but can be derived from several ones (e.g., a "comic epic"). In addition, Hempfer uses the term "Sammelbegriff" ("collective term") for terms that designate classes in a logical sense. Individual texts can be assigned to such classes on the basis of any characteristic (e.g., "poetry" as texts in verse form) (Hempfer 1973, 27–28). Comparing Hempfer's system to the three terms of text type, conventional genre, and textual genre used here, the following observations can be made. The term that is closest to the text type is Hempfer's *Sammelbegriff* because it designates a purely logical grouping without necessary relationships to a certain theory of genre or a genre convention. However, the *Sammelbegriff* is different in that it presupposes logical classes, which is not done for the text type here. The categorical status of text types and textual genres has to be clarified further, which is done in chapter 2.1.4 below. Not only common but also similar features of texts can be constitutive for text types. Hempfer's *Gattung* corresponds roughly to the textual genre in that it designates groups of texts that share textual features for which there is historical evidence and which were relevant as communicative norms. A slight difference is that the textual features of Hempfer's *Gattungen* are derived from the features of underlying ahistorical constants, whereas the aspect of the origin and motivation of the textual characteristics is not covered by the terminological distinction proposed here. Hempfer's *Untergattungen* can be considered subtypes of textual genres. As far as I can see, there is no equivalence to the purely conventional genre in Hempfer's system of terms. Hempfer's *Schreibweisen* and *Typen* as purely theoretical terms are not covered here, which shows that the three proposed terms of text type, conventional genre, and textual genre focus on the mediation of text-immanent generic features and text-external communicational as well as purely conventional aspects of genres on an empirical level. It is

[63] Hempfer (1973, 14–29) dedicates a whole chapter to the discussion of terminological problems. Although his study was first published in the seventies, it is still influential today in the German-speaking area. Klausnitzer and Naschert discuss his approach as one of the positions in genre theory which sparked some debate in the twentieth century (Klausnitzer and Naschert 2007, 387–404). Neumann and Nünning (2007) also refer to him repeatedly in their overview of problems, tasks, and perspectives of genre theory and history.

a further task to build new theories based on textual genres that are found in this process of mediation or to clarify their relationship to existing literary theories of genre.

Another system of terms is used by Fowler, who distinguishes between "kinds", which are the historical genres (e.g., sonnet, parable), "modes", which are selections or abstractions from kinds (e.g., comic, aphoristic), "subgenres", which are subtypes of kinds (e.g., sea eclogue, historical novel), and "constructional types", which are purely formal patterns (e.g., ring composition, sequence). To distinguish between the different sorts of generic categories, Fowler uses the idea of a "generic repertoire", which means all possible levels on which features that are characteristic of a genre can be chosen. A basic distinction is made between "formal" and "substantive" features. Formal features include structural characteristics and, for example, verse metres. Substantive features are, for instance, related to themes, purpose, and intended audience. The *kinds* usually combine both types of features, whereas the *constructional types* are only based on formal features. *Modes* are more or less unstructured. They have no or only a few formal features but invoke kinds through samples of their substantive features. The *subgenres* are defined as having the same formal features as their corresponding kind, plus additional characteristics related to the content of the texts (Fowler 1982, 55–56). Evaluating the relationship of Fowler's terms to text types, conventional genres, and textual genres, the following observations can be made: *kinds* can be understood as corresponding more or less to the textual genres and subgenres to subtypes of textual genres. What Fowler emphasizes repeatedly is the historical variability of the kinds' features.[64] This raises the question of whether one kind should be conceived as one textual genre allowing for internal variation or eventually several textual genres based on more compact text types. This question will be addressed in the next chapter 2.1.4 on categorization. However, Fowler also states that despite all variation and historical change, kinds are not indeterminate. Preliminarily, he defines them as follows: "a kind is a type of literary work of a definite size, marked by a complex of substantive and formal features that always include a distinctive (though not usually unique) external structure" (Fowler 1982, 74). As Hempfer's *Schreibweisen*, also Fowler's *modes* are not directly covered by the concepts of text type, conventional genre, and textual genre. The latter refer to literary texts as structural units, whereas the former are abstracted and disconnected from external structures: "Modes have always an incomplete repertoire, a selection only of the corresponding kind's features, and one from which overall external structure is absent" (Fowler 1982, 107). [65] When modes are combined with kinds and when they characterize

[64] For instance: "Kinds may in this way give the impression of being fixed, definite things, located in history, whose description is a fairly routine matter. As we shall see, there is something in the idea of definiteness. But describing even a familiar kind is no simple matter. We may think we know what a sonnet is, until we look into the Elizabethan sonnet and are faced with quatorzain stanzas, fourteen-line epigrams, sixteen-line sonnets, and 'sonnet sequences' mixing sonnets with complaints or Anacreontic odes. Besides such historical changes within individual kinds, there are wider changes in the literary model allowed for, with their repercussions on the significance and even categorization of generic features" (Fowler 1982, 57).

[65] A difference between Hempfer's *Schreibweisen* and Fowler's *modes* is that Fowler does not see the modes as ahistorical constants. He describes them as distillations of kinds, i.e. of the features of kinds that seem permanently valuable. In that respect, they are more durable than the historical kinds because they are not linked to external forms that become outdated faster. However, these distillations may also change or become obsolete, as, for example, the heroic mode, which is only conserved in historical or political novels. (Fowler 1982, 111).

the kinds in more detail, the combination of mode and kind is regarded as a subgenre here (for example, a "comic novel").

In this study, no principle difference is made between genres and subgenres. Like genres, also subgenres can be conventional genres, and they can overlap with text types to form textual genres. One difference between both is that the genres usually have a more precise formal delimitation than different subtypes of the same genre. The latter tend to be based on differences in subject matter or style so that different textual characteristics and features become relevant in each case. In addition, subgenres can be very inconsistent because they are formally less fixed than genres. As Fowler remarks, "To determine the features of a subgenre is to trace a diachronic process of imitation, variation, innovation—in fact, to verge on source study. At the level of subgenre, innovation is life" (Fowler 1982, 114). It can therefore be more challenging to find the correspondences between conventional subgenres and text types than between conventional genres and text types, and hence be more challenging to determine the textual subgenres than textual genres. However, the degree of innovation that subgenres undergo depends on the type of subgenre that is investigated and on factors such as canonicity and perceived literariness of the works. Especially in popular literature, it is to be expected that there are works that follow quite schematic patterns of subgenres.

Even if the levels of genre and subgenre are not strictly separated here, it makes sense to consider them in the construction of the corpus to be analyzed to compare genres on a similar level of specificity. If subgenres are the point of interest, it makes sense to build a corpus of instances of the corresponding genre and to include works in it that are associated with different types of subgenres. Then the selection criteria for the corpus as a whole can be based on the formal characteristics that define the genre. That way, determining textual subgenres does not need to be restricted by defining text types beforehand. Different subgenres of the contextualizing major genre can then be contrasted with each other. The relationship of subordination between genres and subgenres is a means of combining a deductive with an inductive procedure in constructing a corpus. Other strategies are conceivable, for example, to combine historical labels as signals for conventional genres with preliminary (sub)genre definitions as proxies to textual genres to build a corpus from which to start the analysis.

2.1.3.5 Genre Identity and Variability

That genre in the narrower sense (*Gattung*, *kind*, conventional, and textual genre) is bound to certain features of formal structure has already been pointed out. Which textual entity is the one that participates in genre? Raible sees an affinity between genres and specific degrees of complexity. He observes that, in general, text is a relative and dynamic notion. A novel such as "Eugenie Grandet" by Balzac is a text, but the series of novels that it belongs to, the "Comédie humaine", or only a subpart of the novel such as a single chapter, are also texts. Usually, genre is bound to the level of the single, whole novel (Raible 1980, 327). Instances of this kind of textual entity are the ones that are collected in a corpus, that are associated with one or several genres, and that are analyzed as to their textual coherence and compliance with generic conventions. However, a novel like "Eugenie Grandet" is a literary work, and as such, it can be realized as text in different forms and contexts. Is genre linked to the literary work as a whole or a specific

realization of it in text form? How stable is the association between the literary work or text and the genre? In principle, it is assumed that the generic identity of a text is identical to the generic identity of the work that the text represents. This means that an English translation of "Eugenie Grandet" would generally be considered as participating in the same genre as the original French text. Different editions of a work in the same language but published in different years or even centuries would also commonly be associated with the same genre. If there is a new work, also the genre may be different. "Eugenie Grandet" as a drama or movie would be a new work and have a different genre. If one looks more closely, this question is, however, not so easy to resolve. First, one can debate whether a new version of a text is another realization of the same work or another work.[66] Second, as genres are conventions that can be described as literary institutions and as horizons of expectation for authors, publishers, readers, critics, etc., which are anchored in specific historical settings, the generic identity of a work can be influenced by the context in which it is realized. This is illustrated clearly by Schaeffer, who discusses the example of the story "Pierre Menard, autor del Quijote", which was published in 1939 by Jorge Luis Borges. The story centers on the idea that the fictional author Pierre Menard publishes parts of the work "Don Quijote" as his own creation in the twentieth century, although formally, they correspond exactly to the text authored by Cervantes and published in the early seventeenth century. As Schaeffer illustrates, the syntactically identical text would have different generic identities because in the twentieth century, it would be considered a historical novel with an archaic style, whereas in the seventeenth century, it was primarily received as a parody on romances of chivalry (Schaeffer 1983, 131–134). In this case, also the author is different, and one could therefore speak of two different works when comparing the original "Don Quijote" to the imagined twentieth-century recreation of it. Nevertheless, the example makes clear that the generic identity of a literary work may depend on its realization as a document in a specific context.

In the digital bibliography of nineteenth-century Spanish-American novels created for this study, there are examples of works that have been associated with different subgenres of the novel in different editions, marked by different subtitles. For instance, the novel "Tomochic" by the Mexican writer Heriberto Frías was first published in 1894 with the subtitles "Episodios de la Campaña de Chihuahua. 1892. Relación escrita por un testigo presencial". In 1899, it was republished without any subtitle, and in 1906, the subtitle was changed to "Novela histórica mexicana". This shows how the novel was initially presented as a testimony and a contemporary

---

[66] In the area of bibliographic modeling, this question is, for example, addressed in the conceptual model FRBR of the International Federation of Library Associations and Institutions (IFLA): "variant texts incorporating revisions or updates to an earlier text are viewed simply as *expressions* of the same *work* […]. Similarly, abridgements or enlargements of an existing text […] are considered to be different *expressions* of the same *work*" (International Federation of Library Associations and Institutions (IFLA) 2009, 17). Translations are also considered as different forms of the same work. On the other side, "when the modification of the *work* involves a significant degree of independent intellectual or artistic effort, the result is viewed […] as a new *work*" (International Federation of Library Associations and Institutions (IFLA) 2009, 17). This includes, for example, paraphrases, summaries, adaptations for children, parodies, and changes from one art form to another. As can be seen, in individual cases, it may be difficult to decide whether changes to a text are a minor revision or a significant modification. Certainly, the question of the definition and unity of a literary work is also central in literary studies. For example, the role of authorship or the prerequisite of completion for a work to be a work can be questioned. For an overview, see Thomé (2007).

documentary novel, and only twelve years later, it was considered a historical novel. Modern critics have interpreted it as a historical, political, social, realist, and naturalistic novel.[67] Such generic variability of individual literary works can be clarified when considering it in relation to concepts of text type, conventional genre, and textual genre. Concerning its text-immanent characteristics, and more specifically, its stylistic features, the work does not change over time because these do not depend on the communicative and historical context in which it is embedded through its various realizations. For this to be true, the simplifying assumption is made that the different published editions do not involve considerable textual adaptations of the work. This means that the different text types that the work can belong to do not change. They depend on the textual level and the kind of textual features that are selected for the analysis (for example, most frequent words or topics). What may have changed in a different historical context is the conventional genre, that is, the concept of the genre that was effective at the time, as well as the specific communicative context in which the work is to be seen, which involves the expectations of publishers and readers. It may then be the case that the work did not fit the conventional criteria for a historical novel when it was first published but that it did with the 1906 edition. The concept of the historical novel may have changed by that time. Furthermore, the novel only falls into the definitory pattern of the historical novel once there is a greater distance between its creation and publication year, which means that the perspective on it has changed. If the generic convention that the literary work is associated with is a different one, also the textual genre becomes another one because it is derived from a different intersection of genre label and text type.

Again, this makes clear that digital stylistics must take conventional genres into account if it aims to produce text analysis results that are historically adequate. In addition, when a corpus is created for digital stylistic analysis, the question of generic identity of the individual texts has to be tackled. For the corpus of nineteenth-century Spanish-American novels created here, it was decided to attribute all subgenre assignments directly to the work level. As laid out in more detail in the chapter on creating the bibliography and corpus, it was encoded if the assignments are contemporary or literary-historical. The boundary between both was drawn based on two criteria: first, the temporal limits of the corpus, and second, the origin of the genre label as endogenous and textual or exogenous and meta-textual. As the corpus covers works that were first published between 1830 and 1910, subgenre assignments that were made after 1910 are considered literary-historical, and the ones before that year as contemporary. In addition, the labels that are derived from paratextual signals are differentiated from the ones that external actors conferred. So in this concrete case, for the purpose of subgenre assignment, a simplification was made by defining the period that is covered by the corpus as broadly homogeneous regarding the historical genre conventions. This was done because the change over time of individual subgenres is not the primary concern of the analysis conducted here. If it was, it would have been more important to consider the changing conventional generic identities.

At the beginning of this chapter, several core issues regarding the relationship between the theory and the history of genres were raised. The last one, referring to the origin and evolution

[67] See the mentions of the novel in Dill (1999), Fernández-Arias Campoamor (1952), Gálvez (1990), Read (1939), Sánchez (1953), and Varela Jácome ([1982] 2000).

of genres, has yet to be discussed. Theories for genre change are an extensive topic of their own, and literary genre theorists have made several different propositions in this regard. For example, Todorov presents a theory about the origin of genres: "Where do genres come from? Quite simply from other genres. A new genre is always the transformation of an earlier one, or of several: by inversion, by displacement, by combination" (Todorov 2014, 197). He thus sees the generic system as one that is in constant transformation, and he approaches the question of the formation of genres not through historical analysis but through systematic considerations. Different types of genres are compared to other kinds of speech acts to which they are related. For example, the prayer as a genre is related to praying as a speech act, the novel to telling, and the sonnet to "sonneting", which does not exist as an institutionalized speech act. Todorov concludes that different kinds of developments are involved, from general simple speech acts to more complex literary genres, but that in general, genres derive from "normal" language: "that makes it possible to see that there is not an abyss between literature and what is not literature, that the literary genres originate, quite simply, in human discourse" (Todorov 2014, 208). In the more historically oriented genre theories, generic change does not need an independent explanation because it is a central part of the genre concepts. For example, of the three generic logics Schaeffer proposes, three are based on the principle of modulation, which entails that each text that participates in a genre modifies the genre characteristics (Schaeffer 1983, 166). The dynamic of genre change is then a question of the extent to which one or several texts alter the genre concept. In Jauß's theory, the history of genres is explained with the variability of the readers' (and authors') horizons of expectations as a "temporal process of the continual founding and altering" (Jauß 2014, 132). Voßkamp's idea of genres as literary-social institutions also involves processes of permanent reductions which lead to stabilizations and destabilizations of the institutions (Voßkamp 1977, 30). A literary scholar who devoted several book chapters to the transformation of genres is Fowler (Fowler 1982, 149–212). He provides an overview of the explanations different genre theorists have developed for the origins and changes of genres and generic systems. Similar to Schaeffer, Jauß, and Voßkamp, Fowler assumes that processes of change are at work constantly, but in his view, these processes are inherently literary and not of a general linguistic or historical nature. Taking into account many literary-historical examples, Fowler describes the main types of transformation processes. Among these, there are topical invention, the combination of generic repertoires, the inclusion of generic repertoires into others, a selection of new repertoire elements from other genres, and their mixture (Fowler 1982, 170).[68]

How are processes of generic change to be understood in the context of text types, conventional genres, and textual genres? In general, also change can be described on these three levels: there can be textual change leading to transformed and new text types, change in conventional genres if there is, for example, a poetic discussion resulting in new elements that are considered necessary for a specific genre, or if texts are associated with new conventions through the use of genre signals or labels. The transformation of the textual genres is then a result of the shifts that are at work on the other two levels: if either the text type or the conventional genre or both at the same time change, then also the textual genre becomes different. A more difficult question is

[68] Fowler not only sets forth transformations of genre but also differentiates them from modal transformations, which he calls "generic modulation", with a different sense than the modulation that Schaeffer describes.

where the boundary between one text type and another, one conventional genre and a different one, and hence also between one textual genre and a new one lies. If genre transformations are assumed to be constantly at work, how can the unity of a text type, a conventional genre, or a textual genre be defined at all? This question is crucial not only for the historical change of genres but also for their description from a synchronic perspective. The boundaries between different text types, various conventional genres, and textual genres are not necessarily clear-cut. The overlap of text types and conventional genres can also take several forms. These problems are discussed in chapter 2.1.4, on categorization.

With access to very large corpora of digital texts that, in theory, can cover whole periods or several periods of literature, digital genre stylistics is enabled to address genre analysis over time. Different approaches to this problem are possible and have been pursued. Although generic change cannot be comprehensively considered in this study, some general remarks on the importance of corpus design in this regard are made. Obviously, a central aspect to capture change over time is the collection of metadata about the creation or publication dates of the literary works that are part of the corpus that is analyzed. Furthermore, decisions have to be made about the status of different editions of the same works. Questions are, for example, if only first historical editions are considered or if modern editions are used, and also if only one edition per work is analyzed or if different versions of works are compared. In most cases in which transformations of genres over time have been analyzed in digital genre stylistics so far, the reference to temporal metadata is the primary strategy to capture change. For example, a corpus can be subdivided into an earlier and a later period, and the texts contained in both partitions can be compared to find the differences between early and late variants of the genres that are covered by the text collection. Such differentiation can also be made in a more granular way, analyzing, for example, differences by decade or by year. Jockers, for example, analyzes the stylistic change of novels in English by decade (Jockers 2013, 82–89). Using publication years as the temporal unit, Underwood proposes to adopt the Foote novelty, an algorithm that was originally developed to locate points of significant change in music. Underwood uses the algorithm to detect "revolutions" in the history of the novel (Underwood 2015a; Foote 2000).[69] In the mentioned cases, the feature set used to analyze the novels over time is the same for all points in time or periods. What is measured is how the feature distributions change in relation to the temporal metadata. One could propose different historical stages of text types if the values and constellations of the textual features vary considerably between one and another period. Another way to model textual change is to create different feature sets for different points in time or features that develop over time. In that case, a challenge is to clarify how the different feature sets or features are related to each other so that it is possible to speak about a change. For example, there are dynamic topic models in which the composition of the topics evolves temporally (Blei and Lafferty 2006). As in the previous cases, the primary concern is the change on the level of the text types. This can be linked to developments of conventional genres if, for instance, metadata about different genre labels that the texts had at different points in time is

[69] The CLiGS group also experimented with this algorithm to detect phases of accelerated literary development in over 300 French twentieth-century novels in a project presented at the conference "Forum Junge Romanistik", using topics and temporal expressions as features (Schöch et al. 2017).

integrated into the analyses, as Underwood did in his account of Gothic novels (Underwood 2016). However, even if the genre labels are constant, changes in the text types produce different intersections with conventional genres and hence altered textual genres.

Several issues that are related to the question of the systematic or historical nature of literary genres were discussed in this subchapter. Different literary genre theories emphasize the dependence of genres from the historical context to different degrees. Several terminological systems have been developed to distinguish between historically constant and more variable generic notions, between linguistically and text-founded types as well as conventional groupings of texts. Digital genre stylistics has been described here as a field of research in which there is a strong focus on the analysis of text types because groups of literary works are formed and analyzed based on stylistic features that are derived from the linguistic surface of the texts. Then again, in its applied form, digital literary stylistics is mainly concerned with empirical analyses of historical text corpora, so that questions of the historicity of the texts inevitably have to be addressed. Such questions are, for example, which literary period to cover with the corpus, how the literary works in the corpus are dated, which types of editions to use, and from which perspective and based on which sources the generic identity of the works is determined. Usually, it is especially the latter aspect of deciding on the type of genre labels that links digital genre stylistic analyses to modern or historical genre conventions. Literary works in very large digital corpora can normally not be labeled according to purely theoretical criteria that have been developed based on close reading of texts and that require close reading for their application. Instead, the genre labeling of works in large text collections usually needs to be based on categorizations that others have made for part of the texts (authors, editors, readers, critics, librarians, scholars, etc.), so that sets of genre labels are the result of collective decisions influenced by the respective generic conventions and theories that apply.

Besides selecting genre labels, selecting works for the corpus also means that a connection to contemporary or historical expectations towards genres is established. If, for example, a corpus of novels is assembled to analyze subtypes of novels, the genre discourse is already involved in the outer delimitation of the corpus, even if the subtypes of the novel are defined by relying exclusively on textual surface features. That is, even a text-centered analysis of literary types addresses questions of genre as soon as this notion is part of the definition of the corpus. Therefore, it is argued that both literary text types as systematic groupings of texts based on textual features and conventional literary genres as historically bound literary institutions are important for digital genre stylistic analyses. However, the potential of digital genre stylistics to enhance the knowledge about the literary history of genres can best be developed if literary text types and conventional literary genres are terminologically separated and defined independently of each other in the first place. That way, insights into textual genres are not limited from the outset by requirements of completeness, integrity, or congruence on either of the other two levels. The initial separation of levels on which genres (in the general sense) are analyzed is the more important as stylistic textual features can be very different from the literary features that have been formulated as defining genres either in historical conventions or literary theories, or both. Defining textual literary genres as convergences or intersections of text types and conventional genres allows for finding new correspondences between both perspectives on the genres.

To consider several levels of analysis for genres and to relate these to each other has also been advocated for in more recent genre theoretical approaches. Gymnich and Neumann (2007), for example, propose a "compact definition" ("Kompaktbegriff") of genre. They consider four main levels of genre analysis (the textual level, the cultural-historical dimension, the individual-cognitive dimension, and the functional dimension) and outline the points of contact between them.[70] The levels of literary text types and conventional literary genres proposed here cover, in particular, the textual and the cultural-historical dimensions. Individual-cognitive and functional aspects can also be connected to both of them, even if they are not the primary focus of digital genre stylistics. Compared to the integration of the different analysis levels in the "compact definition" of genre, the convergence of text types and conventional genres in textual genres described here is more concrete because it refers to intersections of text groups. The terminological differentiation between literary text types, conventional, and textual genres relates directly to the possibilities and characteristics of digital quantitative approaches. They need their own genre theoretical foundation, not because they should be isolated from other directions of genre research, but to have adequate terms that allow for communicating and clarifying what genre stylistics can contribute to genre theory and history in general. In the following section, the question of how text types, conventional, and textual genres can be held together and delimited as categories is addressed.

2.1.4 Categorization

2.1.4.1 Logical Classes

The idea that genre categories can be conceptualized as logical classes prevailed in genre theory for a long time. From the early poetics onwards, attempts were made to systematize the field of literary texts by grouping them on the basis of formal, functional, and content-related similarities, with the goal to provide "as exact a classification of concrete texts into clearly disjunct classes as possible" (Hempfer 2014, 405). Besides one-level sets, also hierarchical systems or taxonomies were proposed, requiring distinct and complementary classes. In some cases, they were conceived in analogy to scientific systemizations, particularly biological ones. Bonheim, for example, developed the "cladistic method of classifying genres", alluding to phylogenetic systematics (Bonheim 1992).[71] Strube summarizes the activity of the classifying literary scholar as follows:

- texts are gathered together in a group based on their similarity as things;
- then the term that serves to designate these texts is defined by formulating the conditions for its use;

---

[70] Gymnich and Neumann synthesize the references between the four levels in a diagram: they interpret the individual-cognitive level as mediating between the textual level and the cultural-historical dimension of genres and describe the functional aspects as superordinate to the other three levels. Their model aims not to provide a homogenizing general theory of genre but an integrative view on the different theoretical approaches to it so that scholarly communication about genres is facilitated (Gymnich and Neumann 2007, 34–35).

[71] Bonheim's method is based on the idea of separating different classes of literary genres by finding necessary ("megafeatures") and optional elements ("microfeatures") for each class. An aspect of his model that is of interest from the point of view of digital genre stylistics is that also "loss features" are considered, i.e., features that are negated or missing in certain genres (Bonheim 1992, 2–3).

- the term is differentiated from other generic terms relying on consistent distinguishing criteria and so that neighboring terms exclude each other intensionally or by content;
- borderline cases are left aside;
- all terms of the classifying system are organized by juxtaposition or hierarchically;
- and the terms are defined form a certain literary theoretical perspective (Strube 1993, 42–43).

Classifying literary texts by genre is a basic activity for literary scholars and is useful for several reasons. For instance, it helps to build corpora of literary texts that cover a delimited area. In addition, classificatory terms can be used to find out to which genre a work belongs at all. Furthermore, knowledge about necessary features of genres is helpful when texts associated with the genres are interpreted. Classification also provides a way to organize the "embroiled world of literature" (Strube 1993, 41)[72] in a clear way.

Nevertheless, when classification – in the narrow sense of defining classes that require the presence of a fixed set of features for its members and that fit into a system of non-overlapping categories – is applied to literary genres, there are also limits to its usefulness. Bonheim describes them when pondering the limits of his cladistic method of classifying genres:

> There are five distinct grounds why cladograms of genres are fraught with such uncertainties. One has been mentioned: our definitions are not sufficiently encyclopedic. A second one is that genres are alive and change from age to age. A third is that genres in France are different from genres in England or Germany; for it is a common observation that genre terms are frequently untranslatable, and we feel forced to speak of a *conte* or a *Bildungsroman* in English and of the *villanelle* and the *short story* in German. So genres and genre terms are marked on grounds of national origin as well as history. The fourth reason is that genre terms are rarely terms in the scientific sense, but semi-terms, made grubby by repeated use and misuse. The fifth reason is that every new critical school, like every new drive into the bottomless pit of linguistic theories, calls our attention to some qualities of texts which, as far as we are aware, had not before been clearly brought into focus. That means that every model of genre is an *open* set of features. No cladogram can be final and valid for all time and users. (Bonheim 1992, 15)

So among the main problems that Bonheim sees are the dependence on a cultural and linguistic context and the historical variability of genres, which a classificatory approach cannot directly cover. Further problems derive from models' theory dependence and selectivity, but these are independent of the categorization procedure. Bonheim concludes that definitions of text kinds are, above all, heuristic tools (Bonheim 1992, 16). That classificatory definitions of genres are not able to cover synchronic variation, including the one caused by the individuality of norm-breaking literary works or historical change, has often been adduced. Alternative concepts of categorization have been proposed.[73] Strube, on the other hand, defends classificatory procedures by arguing that literary scholars do not classify in the same way as linguists or other empirically working scientists:

---

[72] "der gleichsam verwickelten Welt der Literatur".

[73] For example by Hempfer (2014, 416–419), who proposes to use the concept of family resemblance instead to conceptualize historical genres, such as the elegy, or Tophinke (1997, 161–163), who suggests a solution based on prototype theory for official municipal charters and the unofficial ones used by merchants in the Late Middle Ages.

> Die Voraussetzung, die von den Textsortenlinguisten gemacht wird (nämlich daß es auch in der 'Welt der Literatur' scharf abgegrenzte Klassen oder Begriffsextensionen gebe), mag berechtigt sein, solange sie sich auf sogenannte Gebrauchstexte [...] bezieht. Sie auf kompliziertere Arten der Literatur zu beziehen, heißt, in die Irre zu gehen. Wenn überhaupt Begriffe zur Einteilung der (bisherigen) Literatur taugen, dann nicht Textsorten-, sondern Art- und Gattungsbegriffe des vom traditionellen Literaturwissenschaftler gebrauchten Typs. (Strube 1993, 58)

His remarks can be interpreted as a rejection of classification in the logical sense as the only valid procedure and as an advocation for a broader definition of "classification", that includes specific soft classificatory terms, which he calls "univocal", "paronymic", "porous", and "family resemblance terms", of which the univocal terms are closest to classificatory terms in the strict sense (Strube 1993, 13–28).

As digital genre stylistics is a field in which an empirical methodology prevails, including the use of statistical procedures, also logical classification is part of its repertoire and is repeatedly employed for categorizing literary texts by genre. In many literary stylistic papers, classificatory approaches are used with a focus on which features are most suitable to capture differences between genres, including aspects of style and content, but also structural characteristics of the texts such as text order or representation of speech.[74] Digital studies on the classification of genres, in the strict sense of logical classes, are relevant for several reasons:

- they help to find, model, and interpret textual cues that are crucial to recognizing genres,
- they contribute to assessing established methods of text mining, machine learning, and NLP regarding their value for genre categorization,
- when tested empirically on corpora of different languages, periods, cultural contexts, and genres, they expand knowledge about the extent to which the methods are sensitive to the kind of data,
- and they also help to solve practical problems as, for example, to index large collections of texts by genre.

That classification is so frequently used in digital genre stylistics is probably due to the methodological influence of computational linguistics and computer science.[75]

As a supervised method, classification has the advantage that the analyzed text classes are directly linked to the genre labels so that the correspondence (or non-correspondence) of generic terms to specific textual features can be checked straightforwardly. Furthermore, the success or failure of the classification can be easily measured by comparing the actual genre labels, that is, the labels that the texts were associated with at the outset of the analysis, to the ones predicted by a classifier. If the genre labels represent the conventional genres and the texts themselves are the data based on which text types can be defined, then supervised classification is an ideal method to test to what extent both levels overlap. The more successful a classification based

---

[74] Examples of classificatory genre stylistic studies are Calvo Tello (2018), Gianitsos et al. 2019, Henny-Krahmer (2018), Hettinger et al. (2016), Schöch (2017b), Schöch, Henny et al. (2016), and Underwood (2015b). A special focus on the separation of genre from author signals can be found in Calvo Tello et al. (2017) and Schöch (2013).

[75] Classification as a supervised method of machine learning is introduced in more detail in the analysis part of this study. See chapter 4.2.2.1.

on selected genre labels and textual features is, the more probable it is that the conventional genres represented by the labels actually correspond to text types defined on the grounds of the selected features. As a consequence, it is also probable that they constitute textual genres. As at least one positive and one negative class are needed for supervised classification, as a minimum, two conventional genres and two text types are involved in such an analysis. The text types are then differentiated not by the *kind* of features – because the same feature set is used in the whole setting – but by the characteristics of the feature *values* and how the various feature values are distributed in one text type versus the other. Because the classes are clearly separated from each other, it is possible to inspect the feature distributions in each class to learn about its textual characteristics, and it can also be analyzed which features are decisive to separate one class from the other. This means that text types and textual genres are defined contrastively in a classificatory setting. This highlights the importance of deciding which conventional genres and which literary works associated with them should be analyzed together to find the corresponding text types and textual genres.

However, it can also turn out that classification does not yield very accurate results, depending on the kind of conventional genre that is analyzed, the individual literary works that participate in it and that are selected for the text corpus, and the textual level that is examined. Then the reasons for the failure have to be investigated. Low classification rates can indicate problems with the selection of features or the classification parameters and methods, but they may also be due to the underlying category not having the structure that is assumed. It can, for example, be checked which works are repeatedly misclassified to find out which part of the corpus is not compatible with the textual genres that the classifier learned. Are there only a few works that do not fit? A closer look at them can serve to check if they can be seen as outliers, as texts participating in the conventional genre but not in the textual one because they are not compatible with the text type underlying the other works of the genre. What if a bigger group of texts is affected by misclassifications? This might indicate that there is more than one text type that is connected to the conventional genre or that no textual unity can be found for it at all. In such cases, alternative categorization methods can complement the *classic* classification methods to find out how the conventional genre in question is structured internally on the textual level. Considering the observations that have been made in literary theory and history about the individual relationships between literary works and "their" genres and about constant processes of modulation and transformation, it is very probable that there are such cases of conventional genres that cannot be neatly mapped to text types and that are not coherent textual genres.

2.1.4.2 Prototype Categories

Typological approaches to genre categorization are one of the alternatives that already have a tradition in literary genre theory. A typological description of genres starts from individual literary works seen as typical or especially representative of the genre in question. The genre's characteristics are then derived from the exemplary texts by isolating, generalizing, or even exaggerating their properties. That way, an ideal type is constructed, constituting an extreme case when compared to other types (Strube 1993, 60–61). Following the cognitive-psychological and linguistic theory of the prototype effect, the ideal types have been called "prototypes". The

prototype effect describes general perception and cognition principles involved when humans categorize things. Some instances of categories are recognized faster than others because they have some typical characteristics. The typical features do not even need to be necessary for the category. For example, a prototypical table has four legs, but tables do not need to have this exact number of legs. When things are categorized, they are not checked systematically for present or absent features but are compared to different ideal types based on their attributes. The proximity or distance to an ideal type is decisive for perceiving something as a member of the category in question (Rosch and Mervis 1975; Taylor 2003, 41–62). The resulting categories are different from logical classes. Membership is not based on a set of necessary and sufficient conditions that can either be fulfilled or not and is not simply a question of belonging or not belonging. Prototypical categories have a core and an internal structure that runs from the core to the edges of the category in a continuous way. The edges are not sharp but fuzzy because membership is defined in terms of closeness to the prototypical core, which also entails that there are degrees of membership. As Hempfer, who sounded out the potential of prototype theory for literary genre theory, summarizes:

> Given that the difference between categories is constituted by the respective prototypical cores, the problem of 'boundaries' loses its poignancy because categories can indeed overlap at their edges when one entity variously resembles different prototypical cores in roughly equal measure. The key factor in the distinction between class inclusion and prototype theory lies in the terminological shift that relocates distinguishing characteristics of categories from their boundaries to their *cores* and attributes an entity to a category via its relationship of resemblance to the prototypical core. (Hempfer 2014, 412)

Hempfer finds that prototype theory is suitable for capturing the relationship of literary works with modes or transhistorical invariants. As an example, he discusses narrative versus dramatic communication. Paradigmatic (that is, prototypical) forms of narration are the epics of Homer or Cervantes' "Don Quixote", realistic novels in the nineteenth century, or popular twentieth-century fiction. On the other hand, there are also deviations from the prototypical norm, for example, when drama is narrativized, or narrative is performatized. Nevertheless, untypical instances of narrative or dramatic communication can still be recognized as such as long as they have a closer relationship to their respective prototypical core than to the core of the other category, which according to Hempfer, is also a question of "reference" (Hempfer 2014, 414–415). In this application of prototype theory to literary genres (in the form of modes), both the levels of the literary text type and the conventional genre are involved. The implicit or explicit reference to a generic prototype must be viewed as a genre signal indicating to which literary institution a text is assigned, while the amount of narrative or performative passages in a text is located on the textual level. This complicates matters because a literary work then has relationships to prototypical cores on two levels: that of the generic convention and that of the text type. Differentiating these levels, as is done in the present study, can mean that a literary work can be closer to a particular conventional prototype because the work carries the respective genre label or other signals pointing to it. At the same time, on the textual level, a work can be closer to a different prototype. As a result, it can be interpreted as an untypical member of two textual

genres. As an untypical member of the textual genre (or mode) "narrative", a narrativized drama can be closer to the narrative literary text type than to the dramatic one but be more distant to the conventional narrative genre than to the dramatic one. Vice-versa, as an untypical member of the textual genre "drama", it is close to the core of the conventional genre but more distant to the core of the corresponding text type.

For the theory and history of literary genres, the idea of prototypically organized categories is especially attractive when generic terms refer to genealogical categories established on the basis of tradition, and in which the relationship between the literary works and the genre is characterized by hypertextual modulation (in the terms that Schaeffer (1983, 181) uses to describe this type of generic logic).[76] Often, literary works signal their participation in a conventional genre by referencing another work that is viewed as a prototypical representative of the genre, either as a particularly accomplished masterpiece or as the foundational text of the genre.[77] Examples of this are references to Walter Scott's "Waverley" in historical novels or to Goethe's "Wilhelm Meisters Lehrjahre" in education novels. Works that reference their prototype are expected to either imitate it closely or vary it to different degrees. In addition, a literary work can contain references to several different prototypical works. The prototypical genre categories that are grounded on hypertextual relationships are necessarily anchored in time and closely related to the literary-historical context. Literary works can only reference other works that existed before them, and the writers must be aware of the prototypes to which they allude. Furthermore, at least two instances of texts are needed to constitute a genre in this sense: the prototype and a text that references it.

If prototypical structures are established based on hypertextual relationships, they are, above all, defined as conventional genres. With the help of digital stylistics, it can be examined to what extent the conventional genres that are constituted by tradition and intertextual references have correspondences in literary text types. It can, for example, be checked whether traditional prototypical works also function as cores of genres on the textual level or not and how close or distant other works participating in the genre tradition are to them stylistically.

Such questions were analyzed by Henny-Krahmer et al. (2018) in an analysis of nineteenth-century Spanish-American novels. The novels are associated with the subgenres of sentimental, historical, *costumbrista*-, *gaucho*, and anti-slavery novels. For each subgenre, one or several works were predefined as prototypes, following literary-historical indications of their status as ideal types, either as predecessors or as masterpieces and culminations of the subgenres in question. Subsequently, textual similarities of all the novels to the different prototypical works were analyzed based on MFW and topic features. The analysis faced the difficulty that especially the predecessor prototypes can have their origin in other countries (for example, novels from Spain in the case of the *costumbrista* subgenre) and also in other languages (such as the English

---

[76] Schaeffer defines hypertextual relationships as follows: "J'accepte comme relation générique hypertextuelle toute filiation plausible qu'on peut établir entre un texte et un ou plusieurs ensembles textuels antérieurs ou contemporains dont, sur la foi de traits textuel ou d'index divers, il semble licite de postuler qu'ils ont fonctionné comme modèles génériques lors de la confection du texte en question, soit qu'il les imite, soit qu'il s'en écarte, soit qu'il les mélange, soit qu'il les inverse, etc." (Schaeffer 1983, 174).

[77] See, for instance, Schnur-Wellpott's (1983, 149–159) exposition of the two perspectives of a founding text and his followers versus a master and his predecessors.

"Waverley" for the historical novel or the French "La Nouvelle Héloïse" for the sentimental novel). Furthermore, they can also be chronologically distant from the follower works. "La Nouvelle Héloïse", for instance, was first published in the eighteenth century. To be able to compare works that were originally written in other languages, Spanish translations of them were used, but it is clear that such cultural, historical, and linguistic differences significantly complicate comparisons of literary texts on a stylistic level. Therefore, it is not surprising that there was a tendency for the follower works to be more similar to each other than each of them was to the respective prototype. In addition, the stylistic coherence of the works of individual subgenres was not clearly visible in all cases, which means that their integrity as textual genres could not be confirmed with certainty.

The experiment shows that literary-historical generic relationships are not necessarily translatable to a stylistic level. The results of the prototype analysis of the nineteenth-century subgenres of Spanish-American novels also suggest that works which are considered prototypical from a literary-historical point of view might rather be exceptional than typical for a textual genre. Therefore, it could be more useful to look for a prototypical textual core not in the sense of excellent, norm-founding masterpieces but ordinary, average works.

Considering how computational categorization methods in general relate to prototype structures versus logical classes, it can be noted that there is no direct and exclusive relationship between one method and one theoretical concept of genre categories. The two main types of text categorization methods that are used in machine learning and statistical text analysis are classification and clustering. As classification is a supervised method, its aim is to separate the data points in a way that fits best to the class labels provided with the data. As long as only one label is allowed per data point, the results are exclusive classes that do not overlap. So, at a first glance, it seems that statistical text classification is entirely congruent with the idea of treating genres as classes in the logical sense. Looking closer reveals that statistical classification procedures can also be related to prototypical structures. First, many classification algorithms include calculating scores or probabilities for the class assignments, which means that internally, information is collected about the degree to which an instance can be considered as belonging to one class or another, before the final decision for the output is made.[78] This internal data can be interpreted as information about the prototypicality of data points for classes. If classification is used to predict the genre of literary works, the probability with which a certain work is assigned to one genre or the other can also be understood as its proximity or distance to a prototypical core of the genre. Second, the features used for statistical classification are usually not only binary but numerical. When numerical and especially continuous numerical features are involved, decisions are not based on the presence or absence of features but on feature distributions that can be similar or more distant from each other. In the output, the classes have clear boundaries, but they have a complex structure internally. The internal structure of the classes is close to how the prototype effect is described theoretically in linguistics:

> I will assume in the following discussion that entities are categorized on the basis of their **attributes**.[1] These are not the binary constructs of the classical approach. Consider

[78] See, for example, the availability of scores and probabilities for Support Vector Machines (SVM) or Random Forest Classifiers (RF) (Scikit-learn developers 2007–2023m, 2007–2023e).

> the ratio of width to depth. The ratio is a continuous variable. Labov's results show that associated with each of the categories CUP, BOWL, and VASE, there is a certain optimum value, or range of values, for the width-depth ratio. In categorizing an entity, it is not a question of ascertaining whether the entity possesses this attribute or not, but how closely the dimensions of the entity approximate to the optimum value.
>
> [1] From now on I shall restrict the term 'feature' for the abstract features of the classical approach, reserving 'attribute' for alternative, non-classical theories of categorization. (Taylor 2003, 44)[79]

In statistical classification, optimum values and optimal combinations of values are determined quantitatively. If a model is trained with data that shows a concentration of similar values, value ranges, or value combinations for a class, this subspace of features can be interpreted as the prototypical core of the category to which new instances that are to be classified are related. However, the ultimate goal of supervised classification is not to produce models of prototypically structured categories but to define models that separate the data along the lines of the predefined labels that the training data has and to predict membership in distinct classes. Carried over to the definition of genres, it means that what is learned are text types that correspond as closely as possible to the conventional genres – assuming that the labels are expressions of generic conventions. The text types have clear boundaries, but are (usually) not constituted based on of binary features. The model of the text type can be applied to new literary texts to test whether they comply with it or not. True positives can be interpreted as literary works that belong to the text type and the conventional genre and hence to the textual genre. True negatives belong neither to the text type nor to the conventional genre nor the textual genre in question. False positives are part of the text type but not of the conventional genre and, therefore, also not of the textual genre. False negatives are part of the conventional genre but not of the text type and, consequently, also not of the textual genre. Probabilities can be used to check how certain the assignment of a text to the text type is and, thereby, how close it comes to the prototypical core of the text type.[80] However, nothing is said about prototypes of the conventional genre in such a classificatory setting. Prototypicality on the level of convention (and also tradition) is inaccessible to text classification with stylistic features as long as it is not explicitly modeled in the metadata and available in the form of specific genre labels. The cognitive aspect of the prototype model, which involves recognizing instances of categories as a whole and not only as a sum of its parts, is also out of scope for text stylistic categorization: "Finally, it emerges very clearly from Labov's

---

[79] Unlike Taylor, who uses the term "attributes" for non-classical categorization, here the term "features" is used because it is the term that is usually employed in digital text analysis.

[80] Another approach to using statistical classification for analyzing the prototypicality of individual texts with regard to genres is pursued by Konle in his master thesis about word embeddings for literary texts. Using 200-word segments of novels, Konle trained a neural network aiming to find words that are distinctive for different genres. The network is trained for the subgenres of sentimental, crime, science-fiction, and horror novels and is used to classify the segments by genre. When discussing the results, Konle analyses individual segments that were misclassified. He finds out that there are cases in which the misclassifications make sense because the novels are not prototypical for their genres in all parts and contain passages with words that are distinctive for other genres. These can be elements of the plot that are interpreted in terms of another subgenre, for example, passages about the death of characters in sentimental novels that are classified as horror novel segments (Konle 2019, 46–49, 59–64, 69–76).

experiment that no one single attribute, or set of attributes, is essential for distinguishing the one category from the other. Presence of a handle, or function in the drinking of coffee, merely raises the probability that an entity will be categorized as a cup" (Taylor 2003, 44). What can be done is to make clear on which level, textual or non-textual, literary texts are participating in genres or not. The crux with the proposed interpretation of classification results in terms of text types, conventional genres, and textual genres is that it presupposes a good quality of the model. If, for instance, only an overall accuracy of 60 % is reached – is this a sign that the model was not properly trained? Or is it a sign that the conventional genre expressed through the class labels does not correspond to a congruent text type? In such a case, further experimentation and examination of the results are needed, including an analysis of the historical genre conventions that are involved.

The second main type of computational categorization method, which is also often used in digital genre stylistics, is unsupervised clustering.[81] With this approach, the data is partitioned into a certain number of clusters so that data points inside of a single cluster are more similar to each other than to other data points lying outside of it (Müller and Guido 2016, 170). That the method is unsupervised means that no category labels are used to group the data. In the results, there can be different degrees of similarity between the points in a cluster, between points in different clusters, and also between whole clusters.[82] For several clustering algorithms, the cluster centers, called "centroids", are crucial for defining of the clusters. In K-means clustering, for example, the clusters are described by means of the data points contained in them. These means, which are in general not equal to any of the actual data points, are the cluster centroids (Scikit-learn developers 2007–2023b). For many clustering algorithms, the number of clusters to learn has to be set as a parameter value.[83] Some aspects of clustering methods remind us of prototypically organized categories. Each cluster can be interpreted as a category, and their centroids as the ideal prototypes, which represent (but are not equal to) the most typical exemplars of the category. Clusters are formed based on continuous distance relationships between data points and cluster centroids. This corresponds to the idea of prototype categories where some members are central and others marginal. A difference is, however, that clustering results in separate categories which have clear boundaries. However, for single data points, it is possible to calculate the distances to several centroids. That way the cluster boundaries can be crossed when the model is interpreted.

---

[81] For different implementations of clustering algorithms in Python, see Scikit-learn developers (2007–2023a). Practice-oriented introductions to clustering are Müller and Guido (2016, 170–209 and VanderPlas (2017, 462–476). A more theoretically oriented introduction to clustering algorithms is contained in Alpaydin (2016, 143–162).

[82] This becomes especially clear in hierarchical clustering, an approach that divides the data into several levels of clusters, either in a top-down or bottom-up way. The latter is called agglomerative clustering and starts from merging the two most similar points, which again are merged to the next most similar cluster, and so on, either until a certain number of clusters is reached or until all points are merged into the same overall cluster (Müller and Guido 2016, 184). Hierarchical clustering is often used in stylometric analyses that are concerned with authorship attribution, so that it can be inspected on which close or distant level different authorial candidates are grouped together with the text in question (Eder 2017).

[83] This is the case with K-means, spectral clustering, Ward hierarchical clustering, or agglomerative clustering (Scikit-learn developers 2007–2023a).

How could clustering algorithms such as K-means be applied to analyze literary texts and their genres as prototypical structures? The clusters that are found can be interpreted as literary text types that have a prototypical core. The texts that are members of the clusters can be seen as members of the text type with certain proximity or distance to the prototype. In contrast to statistical classification, though, the clustering algorithms determine the text types only through the analysis of feature values and distributions and not by optimizing the relationship between text types and conventional literary genres, because no genre labels are used in the clustering process. Once there is a model of clusters, genre labels can be applied afterward to evaluate to what extent the text types overlap with the conventional genres. The resulting intersections can be interpreted as textual genres. An expectation is that the conventional genres will tend to appear as textually more fragmented when the models are not forced to take into account the traditional genre labels. Therefore, the unsupervised categorization is better suited to detect discrepancies between conventional and textual genres than the supervised approach. However, it is less useful to learn which textual features constitute the common basis for specific conventional genres. Considering the status of the prototypes that are determined through the calculation of cluster centers, they are not to be confounded with literary works that have been described as prototypical and especially influential for a genre by literary historians or that were considered as such by contemporary writers. Instead, these are statistical means of text types and can be viewed as representing the most normal textual characteristics for a literary type. It can be concluded that clustering algorithms are closer to the idea of prototype structures than classification methods, but of course they are also limited to the aspects of categories that are detectable through the textual surface and text style.

2.1.4.3 Family Resemblance Networks

Besides prototypical structures, another alternative approach to conceptualizing genre categories is discussed here: family resemblance networks. The idea to describe categories in analogy to family resemblances goes back to Wittgenstein. In his linguistic-philosophical work, the concept serves to describe linguistic activities involving the use of the same word for phenomena which only have partial and indirect similarities, like the resemblances that can be seen between different members of a family:

> Consider, for example, the activities that we call 'games'. I mean board-games, card-games, ball-games, athletic games, and so on. What is common to them all?—Don't say: 'They *must* have something in common, or they would not be called 'games''—but *look and see* whether there is anything common to all. For if you look at them, you won't see something that is common to *all*, but similarities, affinities, and a whole series of them at that [...]. And the upshot of these considerations is: we see a complicated network of similarities overlapping and criss-crossing [...]. I can think of no better expression to characterize these similarities than 'family resemblance'; for the various resemblances between members of a family—build, features, colour of eyes, gait, temperament, and so on and so forth—overlap and criss-cross in the same way.—And I shall say: 'games' form a family. (Wittgenstein 2009, paras 66–67)

Here, the relationships between members of a group are conceptualized as a network of overlapping – but not in all cases common – similarities. The family resemblance analogy was adopted in literary genre theory already in the 1960s. It became popular because it allowed for more open definitions of genre, able to overcome the necessity that the features relevant to a genre should be present in all literary works attributed to it. Such an open concept of genre categories is especially useful for capturing changes in genres over time. These often entail that the same generic term is applied to literary works with quite different textual characteristics between one period and another. The gradual process of change can be understood as a chain of links between individual works in a network, which can be followed. While immediate neighbors in the network share some features, this must no longer be true for distant relationships. However, the details of applying the notion of family resemblance to literary genre theory are not always laid out, even if the approach is mentioned as a possible option by literary theorists. Often, the term is invoked briefly as a counter-concept to logical classes, which are seen as insufficient for the theoretical representation of historically variable genres, for example by Jauß:

> The continually new 'widening of the genre', in which Croce saw the supposed validity of definitional and normative genre concepts led ad absurdum, describes from another perspective the processlike appearance and 'legitimate transitoriness' of literary genres, as soon as one is prepared to desubstantialize the classical concept of genre. This demands that one ascribe no other generality to literary 'genres' […] than that which manifests itself in the course of its historical appearance. By no means must everything generically general – what allows a group of texts to appear as similar or related – be dismissed […]. Following this line of thought, literary genres are to be understood not as *genera* (classes) in the logical senses, but rather as *groups* or *historical families*. (Jauß 2014, 131)

Jauß's statement shows that the recourse to the family resemblance concept also serves to defend the idea of literary genres per se against the critics who question it entirely. Critics make reference to the constant processes of transformation which are at work when individual works refer to genre conventions and when these individual works contribute to the genre's modification at the same time. More recently, there has been some consensus that literary genres are not to be equated with logical classes. Fowler gives an overview of the genre concepts that prevailed in the history of genre theory. He takes the position that understanding genres as a means of classification is an error. He sees a practical value in using genres for taxonomic efforts but argues that the principal value of genres is functional: "genres have to do with identifying and communicating rather than with defining and classifying. We identify the genre to interpret the exemplar. […] If we see *The Jew of Malta* as a savage farce, our response will not be the same as if we saw it as a tragedy" (Fowler 1982, 38).

Here, a less critical view of the potential textual reality of genres is maintained. To view "The Jew of Malta" either as a savage farce or as a tragedy would mean to view it in light of two different conventional genres connected with different discursive expectations towards the text. In a stylistic analysis, the text would be grouped together with other instances of the conventional genre in question. Depending on the perspective formed by the selection of other texts and features, different textual clues and "normative facts" can be found as traces of

the conventional genres. This does not mean that the chosen generic perspective must lead to the definition of a textual genre based on necessary common features and a definition that fits all exemplars associated with the corresponding conventional genre. However, it is also not excluded from the outset that there can be textual similarities in addition to to the conventional genre's functional and non-textual communicative characteristics. Fowler goes on to say that the missing necessary common features of genres make their definitions impossible: "Either defining characteristics are absent altogether, or they are limited to meager distinctions that do no more than subdivide the genre. In short, genres at all levels are positively resistant to definition" (Fowler 1982, 40). The theory of family resemblance then appears as a solution that allows for describing connections between literary texts participating in the same genre without needing to define what their common characteristics are: "It promises to apply not only to close-knit connections within subgenres [...] but also to far-flung resemblances between widely divergent works [...]. Genres appear to be much more like families than classes" (Fowler 1982, 40). It turns out that Fowler puts definitions on a level with logical classes. He states that descriptions of genres in terms of family resemblances should not be seen as preparations of genre definitions as classes but as descriptions of "a different sort of grouping, not reducible to a class" (Fowler 1982, 42). Although Fowler advocates for conceptualizing genres as families, he sees a need to modify Wittgenstein's concept of family resemblances. According to Fowler, in its original form, it would entirely impede generalizations about literary forms. Fowler criticizes that Wittgenstein focuses on "directly exhibited resemblances" and suggests instead focusing on the traditional links between literary works: "What produces generic resemblances, reflection soon shows, is tradition: a sequence of influence and imitation and inherited codes connecting works in a genre. As kinship makes a family, so literary relations of this sort form a genre" (Fowler 1982, 42). Here Fowler clearly addresses the conventional level of genres. Traditional references between the works are a prerequisite to treating them as members of a family, which limits the possibilities of the network to cover just everything. However, it also means that the conventional genre precedes literary text types. Comparing Fowler's concept to Fricke's definition of genre, the level of convention is not dependent on the textual level but the other way around. Still, it is not entirely clear how Fowler delimits a common tradition (and a common generic convention) because he cautions:

> In its modified form, the theory of family resemblance also suggests that we should be on the lookout for unexhibited, unobvious, underlying connections between the features (and the works) of a genre. As with heredity, with generic tradition too we have to expect quite unconscious processes to be at work, besides those that readers are aware of. It would be strange if genre did not in part operate unconsciously, like other coding systems within language and literature. (Fowler 1982, 43)

His remark on the possibility of unconscious processes that lead to shared features and resemblances between works again points to the textual level because conventional markers of genre are unlikely to be unexhibited. So the common literary tradition must be understood in a wide sense, for example, as a common temporal, geographical, or cultural context. Usefully the common literary tradition can also be understood as a common major genre in which the

works that are part of the family network participate, but then the traditional links are quite loose again.

By analyzing literary works in such a broad frame of common tradition, for instance, by building a corpus that represents it and categorizing the texts based on textual features, digital genre stylistics can contribute to uncovering the underlying connections between the texts. In order to formalize the family resemblance concept, this needs to be done not by presupposing necessary common features or feature values and distributions that are similar for all the texts in a group at once but by allowing indirect and overlapping similarities. This is possible with network analysis. A proposal for such a textual "family resemblance analysis" is made in chapter 4.2.2.2, where it is applied to subgenres of nineteenth-century Spanish-American novels. The families are subgroups in the entire network, and they can be delimited because some regions in the network are interconnected more closely than others. With that, it is possible to define literary text types based on family resemblance relationships. These text types can be related to conventional literary genres so that non-classificatory textual genres become definable. The recourse to stylistic networks as an alternative method for category building solves several problems:

1. the idea of family resemblance networks can be applied in the analysis of textual relationships between literary works, and there is no need to limit the concept to tradition, convention, and non-textual communicative functions (to which it is of course also applicable),
2. the excessive openness of a network that never ends can be remediated by first choosing a corpus of literary works that represents their common broader traditional and/or generic background and second, by identifying subparts of the entire network that represent (sub)families,
3. the inability of classification and clustering algorithms to build categories on top of indirect similarities is overcome.

Using the family resemblance concept not only as an abstract counterpart to logical classification but also to formalize it can help enhancing its reputation in literary genre theory. It has been criticized there for being a handy slogan and also because the boundaries between different categories are not defined sharply.[84] Fishelov, for example, describes the takeover of the family resemblance concept by genre theorists as the attempt to find a philosophical foundation for "the dominant trend in modern critical theory, with its stress on the flexible and dynamic nature of literary genres" (Fishelov 1993, 54). He criticizes the concept as too open because it cannot explain readers' relatively high consensus about boundaries between different genres.

[84] An early critique was formulated by Vivas already in 1968. Vivas discusses several aesthetic, social, and philosophical reasons for the unpopularity of the idea of genres as classes. He criticizes the family resemblance notion as a loophole to avoid the question of the genre's nominalistic or realistic status: "Taken seriously, nominalism involves the notion that structures have no status in being whatever. But how a totally invertebrate world is possible I have never been able to understand. 'There is,' you may say, 'a new solution to this old problem.' 'Yes, I know,' I reply, 'there is a newfangled one: It is the evangel of Saint Ludwig.' According to these glad tidings the members of a class share among themselves, not identities but family resemblances. Obviously I cannot stop to analyze this newfangled solution here. Let me merely lay it down that between two members of a family the resemblance is that of shared identity. We are therefore not farther along than we were before" (Vivas 1968, 101). Nevertheless, Vivas defends the idea of genres as open concepts.

Subsequently, Fishelov again looks for necessary conditions for differentiating genres that others have described in terms of the family resemblance notion, such as the tragedy or the novel (Fishelov 1993, 55–68).[85] Fricke instead proposes to use combinations of necessary and optional features in "flexible" genre definitions. These should overcome both the definitions of genres as combinations of necessary features, which he sees as too rigid to capture historical change, and definitions in terms of family resemblances, which in his opinion, are too loose. Instead, he proposes structures such as, for example, "[1] + [2] + [3] + [4a u/o 4b] + [5a u/o 5b u/o 5c]" (Fricke 2010, 9) for the definition of the anecdote, meaning that the first three features are necessary, plus at least one of the variants of the features 4 and 5, respectively. So Fricke also goes back to necessary features but complements them with additional sufficient or "alternative necessary" features.

Nonetheless, Hempfer emphasizes the potential value of both the prototype categories and the family resemblance networks for genre theory: "I believe that genre theory within literary studies can, on the basis of the concepts of family resemblance and prototypes, manage to realign key questions, especially those arising from the polysemy and historicity of genre concepts" (Hempfer 2014, 414).[86] This optimistic line is followed here. To start with, all three concepts of genres as categories – genres as logical classes, prototype categories, and family resemblance networks – are seen as heuristic tools that can be employed in digital stylistic analyses of genre. They can be used to look for literary text types and their connections to and overlaps with conventional genres in the form of textual genres. In the previous chapter on the system and history of genres, the issue was raised about where to draw a line between different literary text types and different textual genres in relationship to conventional genres. This question can have different answers depending on the chosen categorization method.

With a classificatory approach, text types and textual genres will be more closely attached to the conventional genre expressed by the genre labels so that some internal variation in the category might be covered. With unsupervised clustering, the expectation is that if the number of clusters is optimized and not fixed beforehand, there will be more text types than conventional genres. The text types are expected to represent the genre-internal or cross-genre textual variation more closely but are more distant from conventional genre groupings. The same is assumed for a family resemblance network analysis, only that this approach is even more flexible in allowing for partial similarities, which might lead to text types that can be related to conventional genres in a more meaningful way. However, these assumptions need to be checked and substantiated by empirical work to determine which of the concepts suits which historical genres best on the textual level. Even if it is expected that the more open concepts of prototypes and family

---

[85] Hempfer, in turn, criticizes that the common features that Fishelov finds may be necessary but not sufficient (Hempfer 2014, 409–410).

[86] As an example of the application of the family resemblance concept, Hempfer describes the history of the elegy, a genre that was originally only identifiable metrically and later by a number of other traits, i.a., intertextual references, and motifs (Hempfer 2014, 416–417). Hempfer concludes: "The diachrony of the genre can best be represented as a synchronic network of relations, in which each individual text or epochal version of the genre is linked to other historical versions through common features. […] The genre identity, then, is not produced by a single trait but by the entirety of all relations among their historical versions" (Hempfer 2014, 419). For an application of the family resemblance concept to genre theory, see also Strube, who interprets a definition of the novella set up by Seidler in that way (Strube 1993, 21–25).

resemblances are, in general, better suited to describe the internal structure and variability of historical, literary text types, classification has its value. It can help to find a good feature basis by which the text types can be linked to conventional genres and use that basis as a starting point for the more open procedures. Classification (see chapter 4.2.2.1) and a family resemblance network analysis (chapter 4.2.2.2) are employed in the empirical part of this study to analyze the subgenres of nineteenth-century Spanish-American novels.

All the three techniques that were proposed here as possible implementations of the three genre category concepts – statistical classification, clustering, and network analysis – have been employed in digital genre stylistics. Even other, non-categorizing approaches, for example, Principal Component Analysis (PCA), are common to explore how literary texts appear as groups in features spaces.[87] Even so, the relationship between these methods and the literary theoretical considerations about genre categories has not been analyzed much yet. Thoughts about prototypicality, the historical variability, and social embedding of literary genres have been expressed and connected to statistical text analysis,[88] but the family resemblance concept, for example, has not been explicitly formalized in digital genre stylistics so far. It is important to link key discussions of literary genre theory to digital genre stylistics so that research results of the two areas are confronted more deeply. This can increase the interest of the two disciplines in each other and also challenge the findings achieved in both of them. As with the definitions of text types and genres, also the concepts of genre categorization and the possibilities for their implementation are crucial points of contact between literary genre theory and digital genre stylistics.

2.2 Style

The textual features employed in this investigation are subsumed under the concept of literary style. Similarly to the genre concept, the definition of style varies from one humanities discipline to the other. The subject also has a long tradition within single disciplines where it is still debated.[89] Herrmann, Schöch, and van Dalen-Oskam (2015), who trace the development of the

---

[87] Principal Component Analysis (PCA) is a technique for dimensionality reduction that projects the data points onto so-called "principal components", which aim to preserve as much variation of the data as possible. The number of dimensions that the data has can be reduced by only considering the resulting principal components further. In digital genre stylistics, it has, for example, been used by Schöch to visualize how French classical tragedies, comedies, and tragicomedies distribute over principal components based on topic features (Schöch 2017c, paras 33–41). Schöch groups the PCA analysis under the heading "Clustering", as does Oakes in his general introduction to statistics for corpus linguistics, because the data is grouped based on similarity (or distance) relationships. Oakes also uses the term "categorization" for clustering methods (Oakes 2003, 95). Here, clustering and classification are considered categorization methods (in the general sense of category building). However, PCA is not because the data points as a whole are not assigned to separated text categories. A related method is Factor Analysis, which Biber used to find groups of feature distributions that serve as the basis for defining functional text types (Biber 1993b).

[88] See, for instance, Underwood's approach to genre via the history of reception or Schröter's proposal to apply machine learning methods to reconstruct the historical change of *disordered* genres such as the German *Novelle* (Underwood 2016; Underwood 2019, 34–67; Schröter, forthcoming).

[89] For a general introduction to concepts of style and stylistics, mainly from a linguistic perspective, see Eroms (2008). In their handbook on rhetorics and stylistics, Fix, Gardt, and Knape (2008) give a comprehensive overview of

notion of style in the German, French, and Dutch literary and linguistic tradition since 1945, note six types of definitions of style that re-appear in most approaches to the concept:

> style as
>
> - constituting a higher-order artistic value (assessed through aesthetic experience),
> - a holistic gestalt of single texts,
> - an expression of individuality, subjectivity and/or emotional attitude of an author or speaker,
> - an artifact that presupposes (hypothetical or factual) selection/choice among a set of (more or less synonymous) alternatives,
> - a deviation from some type of norm, involving (quantitative or cognitive) contrast,
> - any property of a text that can be measured computationally.
>
> (Herrmann, Schöch, and van Dalen-Oskam 2015, 30)[90]

Especially the last two of these definitions are relevant for corpus-based computational genre stylistics. The notion of contrast is decisive because the usual procedure to delimit genres as text categories is to compare texts associated with the genre one is interested in with texts that participate in other genres on a similar level (for example, the level of thematic subgenres). Another possibility is to compare texts of one genre with texts that are part of a more general generic context enclosing the genre of interest (for example, comparing crime novels to a set of other novels). The style of the genre or subgenre in question is always defined relative to other elements in a corpus. Suppose the whole literary production covered by the broader context of the corpus is considered representative of a general norm. In that case, the style of a genre can be captured as a subnorm, which deviates from the general one, or as a norm that contrasts with other subnorms. Regarding the relationship between text style and genre, the *norm* is understood as the set of *normative facts* that can be extracted from the texts. These are traces of and represent the conventional and communicative norms of genres. The last notion of style that Herrmann et al. list provides the basis for the kind of style and the kind of normative facts that can be identified in the texts: any property that can be measured with computational means. The target property can be, for example, author style, genre style, or the style of a certain literary period. If it can be measured, it must be possible to capture it through the linguistic surface of the texts. This can be done either directly, when specific words, characters, or syntactic constructions are counted, or indirectly, for example, when textual cues are interpreted in terms of higher order features such as topics, which in turn are interpreted as elements of the target style. Text structure that is interpretable from the surface can also be subsumed under the idea of computationally tractable text elements that can serve to define a target style. This would include, for instance, chapters, headings, paragraphs, lists, tables, typographically marked quotes, or passages of direct speech. The computational treatment of such structures can be facilitated through textual markup. From

---

research on style, addressing a broader spectrum of humanities disciplines. An introduction focusing on style in fiction is Leech and Short (2007).

[90] They focus on definitions at the textual level, including the pragmatic dimension, but do not take into account psychological and cognitive-linguistic theories.

a computational perspective, metadata that summarizes textual aspects is, in the first place, not considered part of the text stylistic features. For instance, if the narrative perspective of a text is captured in a metadata item as the result of a close reading process or based on external literary-historical information, this textual aspect would only be linked to style if it also can be induced from specific textual surface features. The target level of style (author, genre, period, etc.) is usually settled on an extra-textual level of the wider communicative context of the texts. Then again, author styles, genre styles, and the styles of periods can be recognized in texts, which means that such influencing factors can be assumed to be tractable in the texts through consciously used text properties as well as unconsciously left linguistic marks.

In the context of genre, Schaeffer differentiates between three kinds of text properties. In his terms, "generic indexes" are signals that are mainly found in paratexts of the works or the literary context and whose function is metatextual and demonstrative (for example, the label "novela de costumbres"). "Generic traits" are textual properties and have a structural and non-demonstrative function (for instance, verse structure in sonnets). "Generic markers" are textual traces of factors that are part of the communicative level and exemplify an intentional property (such a factor would, for example, be a satiric attitude) (Schaeffer 1983, 174–175). Transferred to the idea of text style, *traits* can be understood as general text structural properties and *stylistic markers* as surface cues that can be linked to pragmatic properties. *Stylistic indexes* can be conceived as stylistic features of the conventional level, which influence the perception of style but are not necessarily congruent with certain traits or markers. A conventional genre label would, for example, be an index for a certain genre style. In his set of terms, Schaeffer brings in different communicative levels: the "structural" level, by which he probably means the syntactic and semantic textual level versus the pragmatic level. He uses the terms to distinguish between properties that are somehow inherent and constitutive of texts (the traits) and others that are present as traces of communicative functions of the texts (the markers). However, he does not differentiate between the linguistic surface level and higher-order textual characteristics, at least not explicitly. This means that a structural-semantic text property, such as a certain number of characters in a dramatic play, could be called a trait in the same way as the use of certain types of nouns, for example, a scientific vocabulary in a naturalistic novel. For a computational analysis concerned with stylistic properties, such a distinction would be useful to differentiate between properties of texts in general and properties of text style. Furthermore, on the linguistic surface, everything is the same. It is difficult to explain or decide which features are due to the conscious structuring and design of texts and which ones are unconsciously employed or stemming from pragmatic intentions or other text-external factors. Because of these difficulties in describing text style, Schaeffer's distinction between *traits* and *markers* is not directly employed here.

A terminology that covers similar aspects of describing the connection between textual patterns and literary theoretical or conventional conceptions of genres has been proposed by Kessler, Numberg, and Schütze:

> The traditional literature on genre is rich with classificatory schemes and systems, some of which might in retrospect be analyzed as simple attribute systems. […] We will refer here to the attributes used in classifying genres as GENERIC FACETS. A facet is simply a property which distinguishes a class of texts that answers to certain practical interests,

> and which is moreover associated with a characteristic set of computable structural or linguistic properties, whether categorical or statistical, which we will describe as 'generic cues.' In principle, a given text can be described in terms of an indefinitely large number of facets. For example, a newspaper story about a Balkan peace initiative is an example of a BROADCAST as opposed to DIRECTED communication, a property that correlates formally with certain uses of the pronoun *you*. It is also an example of a NARRATIVE, as opposed to a DIRECTIVE (e.g., in a manual), SUASIVE (as in an editorial), or DESCRIPTIVE (as in a market survey) communication: and this facet correlates, among other things, with a high incidence of preterite verb forms. (Kessler, Numberg, and Schütze 1997, 33)

Kessler, Numberg, and Schütze differentiate between *generic facets* and *generic cues*. They use the first term to refer to higher-order attributes that are used to distinguish texts of a certain class from those belonging to a different class. If one interprets the examples that Kessler, Numberg, and Schütze give, one concludes that they understand facets as applying to the text as a whole and not to specific parts of it. Furthermore, the facets that Kessler, Numberg, and Schütze mention are themselves functional linguistic text types ("broadcast" versus "directed communication"; "narrative", "directive", "suasive", "descriptive") or, from the point of view of literary genre, modes which are understood as attributes of genres ("newspaper story"). Facets are characteristics that are not to be equated with specific linguistic traits of the text but are connected to them. The latter are the *generic cues*. The authors assume that generic facets are computationally and linguistically tractable through their cues. The advantage of Kessler, Numberg, and Schütze's terminology for a text stylistic analysis is that they provide an own term for features of the structural-linguistic surface. Their term "cues" will also be employed here to designate low-level stylistic features. The meaning of the term "facets", though, should be described in a more differentiated and comprehensive manner. Here it is preferred to call the examples that Kessler, Numberg, and Schütze give (narrative, directive, descriptive) either functional text types – from the linguistic point of view – or modes – from the point of view of literary genre theory. The term "facets" can include these if they are meant as pragmatic textual properties of genres, but there can also be other types of facets. At the beginning of their article, Kessler, Numberg, and Schütze describe facets as "attributes used in classifying genres" (Kessler, Numberg, and Schütze 1997, 33) and these can, in principle, be of any type. Returning to the examples of the number of characters in a dramatic play or the use of scientific vocabulary in a naturalistic novel, these two aspects can both be understood as facets if they are attributes that are ascribed to genre categories and if they can be linked to surface cues. However, "the use of scientific vocabulary" can commonly be considered an element of style, and there is a direct relationship of this facet to cues (the specific words that are part of scientific vocabulary). In contrast, the facet "a high number of characters in a comedy" would commonly not be treated as style and is much further away from surface cues (it would need to be mapped to mentions of character names and the use of personal pronouns, for instance). To give another example, the narrative perspective of a text, for instance, can be perceived as part of a text's style but is also primarily situated on a pragmatic level: it is signaled that the whole narrative text or major parts of it are designed to be narrated in a certain perspective. This has effects on the syntagmatic realization of the text, for instance, the use of

pronouns and verb forms in a certain person in the narrated text. Therefore, the relationship of the "narrative perspective" facet to surface cues is also more mediated than in the case of "the use of scientific vocabulary". Here, facets are considered as attributes of the target category of style (e.g., genre, author, period, etc.) but not themselves elements of style. The term "stylistic traits" is introduced to designate elements of style that can be more abstract than surface cues but are still attributes of how the text is represented syntagmatically (that is, structurally, linguistically, syntactically, and by surface semantics). A certain narrative perspective can then be conceived as a facet when it is used to characterize a genre or some other category and as a stylistic trait that can be linked to stylistic cues. A high number of characters in a comedy can also be a facet, but it is not considered a stylistic trait. However, if it can be linked to stylistic cues, it can influence style. A high number of character mentions in a comedy, in contrast, can be a stylistic trait because it is situated on the level of the linguistic representation of the text. In the following, the terms "facet", "stylistic trait", and "stylistic cue" are used here.

At the end of their article, Herrmann, Schöch, and van Dalen-Oskam formulate a broad definition of style, aiming to provide a generally useful concept for computational, quantitative, and empirical studies of style in literary texts: "Style is a property of texts constituted by an ensemble of formal features which can be observed quantitatively or qualitatively" (Herrmann, Schöch, and van Dalen-Oskam 2015, 44). Here, the definition is narrowed down a bit and also differentiated and elaborated for the purpose of computational and quantitative stylistic genre analysis:

- Literary genre style is a property of literary texts constituted by an ensemble of observable features, feature values, and distributions, which is distinctive for the literary text type that the texts belong to and, by extension, widely congruent with the conventional literary genre that the texts participate in.
- The features can be directly described as formal low-level stylistic cues, which are measurable as structural, syntactic, functional, or semantic linguistic surface features, or as higher-level stylistic traits, which can be induced from the low-level cues. Both stylistic cues and stylistic traits can be linked to facets, which are attributes of textual genres.
- In a wide sense, all kinds of linguistic features are considered stylistic cues; in a narrow sense, purely semantic features are excepted.

Being less general, the definition directly addresses literary genre style and defines it in terms of the textual genre as the intersection of text type and conventional genre. Furthermore, feature values and distributions are added to make clear that presence or absence is not the only possible form of features and also that specific feature combinations can be decisive. That the ensemble of features should be distinctive means that it can be used to distinguish between different genre categories on the textual level. To be able to speak about the style of a certain literary genre, it is required that there is considerable overlap between the corresponding text type and conventional genre. This is formulated vaguely as "widely", so that the necessary degree of intersection needs to be discussed in each case. Nevertheless, an overlap of more than 50 % can be taken as a minimum. For the features, a difference is made between formal, low-level stylistic cues and higher-level stylistic traits to make clear that there can be different degrees of abstraction from the textual surface. Linguistic low-level features (*stylistic cues*) are, for example,

the number of pronouns, n-grams, or collocations of a specific type. A linguistic higher-order feature (*stylistic trait*) is, for instance, the number of descriptive or argumentative passages in a text. For the stylistic traits, a link to the stylistic cues must be established to make sure that they are observable and computationally measurable. Typical literary features can also correspond to cues, for example, rhyme structures or other rhetoric means. However, it is more common for literary features to be defined on a higher level, stylistic or non-stylistic (for instance, narrative perspective, character constellations, or plot elements).[91] For the higher-order features, it is not easy to draw a strict line between features that can still be considered (indirectly) observable and those that are mainly interpretable. In this respect, too, it will be necessary in each case to discuss the nature of the features: if they are stylistic traits or not and if they can be linked to stylistic cues or not. Finally, the third part of the definition splits the concept of literary genre style into a wide one, including all kinds of linguistic cues, and a narrower one excluding semantic cues. This distinction is made to be able to characterize features more precisely as being primarily structural and functional or as being related to the content of the texts as well. To what extent can semantic cues be understood as elements of a text's style? The concept of style would be overstretched considerably if elements of the content of a literary text, such as the type and number of characters involved, the basic theme, setting, or the underlying sequence of events of the story, would directly be declared stylistic features. In the first place, these are considered textual facets here that can be part of genre definitions. However, if such elements are assessed and described as they occur on the textual surface (via stylistic traits or directly linked to stylistic cues), it seems legitimate to include them in a broad definition of formal textual style because the same story involving the same content-related elements can be told differently. It may vary, for example, how often the characters are mentioned explicitly by their name or through which textual topics or motives the theme of a text is expressed. One may decide to include such elements in using a wide concept of style or to exclude them by adhering to a narrower style definition.

Some observations can be made concerning the relationship of the style definition to different types of categorization. Following the class concept, the above definition of literary genre style entails that a clear line can be drawn between one group of texts and others based on how an ensemble of formal features is constituted in these texts. In the prototype setup, the genre style would be defined based on a prototypical constitution of an ensemble of formal features. The other texts belonging to the same or to other textual genres would be situated relatively to this central feature constitution. From a family resemblance perspective, genre style can be

[91] The field of "literary" features is understood here as including aspects of the texts that are constitutive, typical, or relevant for them according to literary theory. Usually, the linguistic level is an intermediary between the literary features and their formal expression on the textual surface. Literary features are thus more difficult to formalize than linguistic features because their expression in linguistic surface features is not necessarily straightforward. Regarding the question of what constitutes a literary and what a linguistic feature, there is, of course, some area of overlap. Rhetorical figures, for example, can be conceived as elements characteristic of literary style, but their definition is often directly based on linguistic concepts. For genre analyses, specifically literary features are relevant because in definitions of literary genres, usually literary concepts are used and not linguistic ones. So if the results of digital genre stylistics should be linked to literary theoretical discussions of genre, the question of how to formalize literary features and how to link them to surface text style is a prerequisite.

understood as a set of related ensembles of formal features linked via parts of the various feature ensembles.

2.3 Subgenres of the Nineteenth-Century Spanish-American Novel

The empirical part of this study is concerned with creating a bibliography and a corpus of nineteenth-century Spanish-American novels to analyze subgenres of those novels on the level of metadata and text. The empirical work and findings are presented in chapter 3, where the bibliography and corpus are described, and in chapter 4, which contains the metadata and text analysis of the novels and their subgenres. Novels from three countries, Argentina, Cuba, and Mexico, were selected for the analysis, and diachronically, the bibliography and the corpus are limited to 1830–1910. Both the literary-historical developments of the time and the historical, political, and geographical context are discussed in the corpus chapter, where the choice of the countries and the time period are explained. In addition, the general defining characteristics of the novels in question and the assignments of subgenre labels to the novels are laid out in the empirical part.

The current chapter presents a selection of major subgenres of nineteenth-century Spanish-American novels as they are defined, described and discussed in literary-historical research, independently of and preceding the findings of the text analysis conducted in the present study.[92] First, three thematic subgenres are treated: the historical novel (*novela histórica*), the novel of manners or customs (*novela de costumbres*), and the sentimental novel (*novela sentimental*). Subsequently, three subgenres related to literary currents are introduced: the romantic novel (*novela romántica*), the realist novel (*novela realista*), and the naturalistic novel (*novela naturalista*). As argued before, literary currents can be described as conventional literary genres, as well, and potentially even as textual literary genres.[93] Of course, these are not the only subgenres of Spanish-American novels in the nineteenth century, and the metadata analysis of the corpus and bibliography shows a great variety of them. However, they are the most important ones on the two discursive levels of theme and literary current from a quantitative perspective and have therefore been chosen for text analysis. Furthermore, they have repeatedly been treated in literary histories and other literary-historical research about nineteenth-century Spanish-American literature, which is not the case for all the other different subgenres of the novel found in the database that was created for this study. The correlation between their quantitative importance and the critical interest in them is probably not accidental.

As the chosen subgenres also existed and exist outside of the Spanish-American tradition, especially in the European countries, there is a wealth of research literature on each one of them. It cannot be taken into consideration in its entirety, so the presentation of the subgenres concentrates on the three chosen countries. General histories of Latin- and Spanish-American literature and novels are also taken into account but primarily for general statements about

---

[92] The results of the metadata analysis are taken into account, though, because they provide relevant information about the general status of the subgenres, for example, how often they were explicitly mentioned in subtitles or the narrative perspective in which the novels are written. The first aspect describes them in terms of genre conventions, and the second one can be considered a stylistic trait that characterizes or subdivides individual subgenres.

[93] See chapter 2.1.3.3 above on the status of literary currents.

the subgenres in question and not to analyze them for all the countries of the region. There are monographs on some of the subgenres that are specialized in the nineteenth century and either cover Spanish America as a whole or one of the three countries Argentina, Cuba, and Mexico. Most of them are concerned with the literary currents.[94] Of the thematic subgenres, the historical novel has been the focus of literary-historical work. However, no monographic studies could be found for the *novela de costumbres* and the *novela sentimental* in Spanish America in the nineteenth century.[95]

For each subgenre, it is analyzed whether or not its literary-historical characterization allows to formulate hypotheses about its textual coherence, that is, whether or not it can be expected that the novels participating in the subgenre are held together on a conventional *and* textual level, and if so, which textual characteristics are significant. On the other side, it is also possible that the literary-historical knowledge about a subgenre suggests that textual subgroups or textual heterogeneity can be expected. Yet another possiblity is that the literary-historical accounts may not be suitable for formulating such hypotheses at all for some of the subgenres. In case they can be developed, the hypotheses about the relationships between the subgenres' conventional and textual aspects are considered when the novels are analyzed in chapter 4.2. Before the individual subgenres are discussed, some general considerations about the novel as a genre and the Spanish-American nineteenth-century novel are made.

Novels, in general, can hardly be defined uniquely and in formal terms. Necessary formal conditions are not sufficient to distinguish novels from other fictional narrative prose texts of considerable length, nor can additional formal or content-related criteria serve to capture all types of novels (Fludernik 2009, 627; Hempfer 2014, 410). Fowler summarizes the status of the novel as a genre (or historical kind, in his terms):

> A genre so comprehensive can have but a weak unitary force. Indeed the novel has largely ceased to function as a kind in the ordinary way. Its minimal specification has even been stated as 'an extended piece of prose fiction'—a specification in which external form appears, but only as 'extended' and 'prose.' Within this enormous field, the novel in a stronger sense—the verisimilar novel of Austen and Thackeray, which many would consider the central tradition—is now only one of several equipollent forms. (Fowler 1982, 118)

The difficulties in defining the novel formally are also present in the case of nineteenth-century Spanish-American novels. There are some early narrative prose texts that were written in the colonial era and can be considered forerunners of the novel in Spanish America. However, the genre only took off considerably during the nineteenth century (Gálvez 1990, 15–25; Lindstrom 2004). So especially in the early decades of the century, the Spanish-American novel is characterized by literary experimentation, for instance, by fluctuations between very short prose forms

---

[94] For the romantic novel in Spanish America, see Suárez-Murias (1963). The Mexican realist novel is treated in Navarro (1955). The naturalistic novel in Spanish America is covered by Prendes (2003) and Schlickers (2003). For the Argentine naturalistic novel see Gnutzmann (1998).

[95] A monograph about the nineteenth-century Spanish-American Mexican historical novel was published by Read (1939). The Latin-American sentimental novel in general (not limited to the nineteenth century) is covered by Zó (2015).

that are close to stories and excessively long serial novels published in newspapers. Besides the varying length of the texts, their status as fictional works is marginal in some cases as well. Up to the end of the century, many novels served the need to document and discuss contemporary or past political and historical events, sometimes in the form of "memories" or as concealed social critique, and the novel had an important function in the process of nation building after the wars of independence. This aspect is particularly relevant for Argentina and Mexico, but even in the Cuban case, the novels already served identity functions in the colony's last decades. Because of the closeness of many works to non-fictional discourses, the early Spanish-American novels are special, and in some cases, it is difficult to draw a line between fictional and non-fictional texts.[96]

Nonetheless, many novels were published in Argentina, Cuba, and Mexico in the nineteenth century, and their number increased considerably, especially after 1880.[97] Even before, there were more novels than the few ones usually cited as representative of the nineteenth century in literary histories. Molina, for example, analyses the early Argentine novel between 1838 and 1872 and finds that the novels "popped up like mushrooms" (her book having the title "como crecen los hongos", referring to the commentary that an Argentina journalist made in 1860 about the increasing number of novels that were published). She analyzes 86 novels for a period considered unproductive for the Argentine novel (Molina 2011, 13–17).

Even though the novel was a genre that was relatively new as a broad phenomenon, there were specific types of subgenres that were practiced by the Spanish-American writers. For one thing, early Spanish-American novels were influenced by European models, including types of romantic novels such as sentimental novels and historical novels. At the same time, the subgenres were varied to serve the needs of expression in the Spanish-American context in a better way.[98] To analyze these subgenres quantitatively and examine to what extent and in what way the genre labels used by writers and publishers and the ones assigned to the works by literary scholars actually correspond to distinguishable text types is thus of interest from several perspectives.

If the novel itself is so manifold and difficult to define, how about its subgenres? Fowler concludes that "the novel is still a kind, even if one badly in need of subdivision" (Fowler 1982, 120). The novel is rooted in many other earlier fictional and non-fictional genres, such as epic, romance, history, or letter, which influence its more mature forms, "giving rise in some instances to distinct subgenres" (Fowler 1982, 120). So it may be the case that some of the novel's subgenres are textually more coherent than the novel itself is on a general level. On the other side, subgenres of the novel can be specified from a whole range of perspectives, which makes them "extremely volatile. To determine the features of a subgenre is to trace a diachronic process of imitation, variation, innovation—in fact, to verge on source study" (Fowler 1982, 114). It is thus an open question whether Spanish-American nineteenth-century novels can be easily

---

[96] For details about the criteria used to select novels for the bibliography and corpus, see chapter 3.1. The role of the novels in the process of nation-building is, for instance, discussed in Brushwood (1966) for Mexico, Ferrer (2018) for the Cuban context, and Sommer (1993) for Latin America as a whole.

[97] See an overview of the number of novels per decade in chapter 4.1.3. The number of works that were recorded for the bibliography increased from around 20 works that were first published in the 1840s to over 80 works that were published in the 1870s. From the 1870s to the 1880s, the number doubled to about 180 works and remained on that level in the following decades.

[98] For a comprehensive overview of content-related subgenres of the Spanish-American novel, see Sánchez (1953).

separated into textual classes corresponding to their conventional subgenres or whether they resist such classification. The mentioned characteristics of the novel in general and of the early Spanish-American novels, in particular, suggest that at least for some of the subgenres, alternative categorization approaches such as the family resemblance analysis can be helpful to examine where the levels of genre convention, text types, and textual genres diverge.[99] In the next two chapters on thematic subgenres and subgenres related to literary currents, the subgenres that were chosen on each of these two levels are presented to provide background information for the text analysis that is conducted in chapter 4.2.

2.3.1 Thematic Subgenres

2.3.1.1 *Novela histórica*

Characteristical of the historical novel is the crossing-over of fiction and history. Historical events, places, persons, and conditions of a past epoch are represented and arranged narratively (Álamo Felices 2011, 84). While factual history aims to reconstruct a past period by identifying and recording significant personalities and events and by tracing the development of social institutions, historical novels serve to revitalize historical characters and the world they lived in (Read 1939, ix). As a literary subgenre, the historical novel gained considerable popularity during the nineteenth century through the novels of Walter Scott, also in the sphere of the Romance languages and cultures and in the Spanish-American countries (Dill 1999, 131; Janik 2008, 64–67; Lefere 2013, 17; Maxwell 2009). However, in the Spanish tradition, historical fiction also has remote origins in the historical narratives of the Spanish conquests in the Americas (Read 1939, 1–28). In general, several characteristics have been pointed out to be constitutive of the historical novel (Fernández Prieto 1996; Lefere 2013, 17–62; Spang 1998):

- temporal distance between the writing, publication, and reception of the novel and the past in which the narrated events take place
- co-occurence of invented and historical personages, places, and events
- localization of the narrated events in a precise historical past

According to the bibliographical data that was collected for this study, the historical novel was the most frequent thematic subgenre in the nineteenth century. Furthermore, many historical novels were explicitly marked as such in series or subtitles, which shows that the subgenre was well established on a conventional level.[100] In the Spanish-American historical novels, typical topics are the period of conquest, the colonial era, and the wars of independence. Especially the latter are not part of a remote historical past, but in the case of Mexico and Argentina, they took place in the early nineteenth century itself, and for Cuba towards the end of the century. In some historical novels published in Argentina, Cuba, and Mexico, a European setting is chosen, for

[99] First experiments with the corpus of nineteenth-century Spanish-American novels have been conducted by the author of this study in cooperation with members of the CLiGS project. They have been presented at the German and international DH conferences. For a prototype analysis based on MFW and topics and a classification of subgenres with sentiment features, see Henny-Krahmer et al. (2018) and Henny-Krahmer (2018), respectively.

[100] See chapter 4.1.5.3.1 for an overview of the proportions of thematic subgenres in the bibliography and the corpus and chapter 4.1.5.1 for a list of the most frequent explicit subgenre labels.

example, from Antiquity or the Middle Ages, but most are concerned with local history, which is temporally more immediate. This makes the Spanish-American historical novels special with regard to the first general defining aspect of temporal distance because that distance is not given in some cases. However, the novels are often explicitly declared historical, even if the events are contemporary or took place in a very recent past. Even if the more distant past of the conquest or colonial period is treated, this is done differently than in European historical novels: "Scott, for instance, tried to escape from his century and return to a spirit of the past; Mexican historical novelists who dealt with the distant past attempted to interpret that past in terms of their own nineteenth-century [liberal] thought" (Read 1939, 58). Nevertheless, nineteenth-century Spanish-American historical novels can be divided into two groups: romantic historical novels dealing with the conquest and colonial times on the one side and the ones dealing with contemporary historical events on the other (ix). Read calls the latter "novels of contemporary history" (x) and Molina "prospectively historical novels" ("novelas prospectivamente históricas", Molina 2011, 285–312).

Another typical element of romantic historical novels is a sentimental plot. The most prominent example is the novel "Amalia" (1851–1855) by the Argentine José Mármol, which tells the story of a group of resistance fighters against the dictatorship of Rosas[101] and of the protagonist's tragic love relationship (Dill 1999, 127). As Dill states: "Amputierte man den genannten Romantypen [politischer Roman und historischer Roman] ihre politisch-sozialen Teile, würde der private Part samt Gefühlswelt der bürgerlichen Protagonistinnen, d. h. der sentimentale Roman, übrigbleiben" (Dill 1999, 138). However, the historical novel was not only present in the romantic period. It continued to be practiced towards the end of the nineteenth century and also in the twentieth century. It was, for example, influenced by Spanish realist and French naturalistic authors. The Mexican novel "Los bandidos de Río Frío" (1892) which was written by Manuel Payno, for instance, carries the subtitle "Novela naturalista, humorística, de costumbres, de crímenes y de horrores". That it has been classified as a historical novel despite five other explicit subgenre labels in its subtitle marks it as one of the cases with a clear discrepancy between the generic convention, the textual form, and the critical tradition that the works have been seen in.[102] Another example of realist historical novels are the works belonging to the series "Episodios Nacionales Mexicanos" (1903) authored by the Mexican Victoriano Salado Álvarez (Fernández-Arias Campoamor 1952, 84–85; Read 1939, 293–303). Read characterizes them as follows:

> In the collection of material for his work Salado Álvarez exhausted accessible periodicals of the period, and gleaned carefully the memoirs of acquaintances and little known books and documents, in search of human aspects of the period and its chief characters, attempting to bring to light fresh information instead of revamping the threadbare stories that constituted the patriotic equipment of various partisan groups and that had been the chief source of of such writers as [the late romantic author] Juan A. Mateos. (Read 1939, 295)

---

[101] Juan Manuel de Rosas (1793–1877) was a governor of the province of Buenos Aires who established a dictatorial system marked by repressive measures that lasted between 1829 and 1852 and that enforced a political and economic hegemony of Buenos Aires over the other provinces.

[102] Read (1939, 260) calls it "the best historical novel of the nineteenth century."

In sum, several aspects can be noted regarding the expected textual coherence of nineteenth-century Spanish-American historical novels. First, the subgenre is long-lasting, there are numerous works attributed to it and it was recognized as a subgenre by contemporaries. These factors suggest that the subgenre is also stylistically distinguishable from other subgenres. What might make its textual classification difficult, though, is that it is often mixed with a sentimental plot and thus has characteristics in common with the sentimental novels, as well. Furthermore, the novels of contemporary history treat similar political and social subjects to other types of novels that are not designed and communicated as historical ones. In addition, stylistic differences can be expected between the early romantic versus the later realist and naturalistic historical novels. Moreover, the range of different special topics (conquest, colonial era, wars of independence) might lead to a subdivision of the textual genre into several subtypes, at least if semantic features are used to categorize the texts.

2.3.1.2 *Novela de costumbres*

The *novelas de costumbres* (also called *costumbrista* novels, novels of customs, or novels of manners) derived from short pieces of prose, the so-called "cuadros de costumbres", which had a French origin and were very popular in Spain before they reached Spanish America.[103] These prose texts were mainly published in journals and periodicals. In them, life in urban or rural settings is described, and the male and female types of different social strata are portrayed with their habits. Furthermore, traditional festivities are a common topic in the texts. The *cuadros* evolved to full novels retaining the same characteristics. Observing the particular, the traditional, and the vernacular were important concerns of the novels of customs. They were, therefore, well suited to take up elements of the Spanish-American reality of life and contribute to the development of a national novel. Stylistically, they are characterized as creating a local color, for instance, by including colloquial speech and creating a picturesque effect, which links them to picaresque novels. Moreover, a humorous or critical-ironic attitude is typical for the novels of customs in their romantic form. In Spanish America, they were also used as a vehicle for political purposes (Janik 2008, 60–64).

As in the case of the historical novels, also the novels of customs were often explicitly designated as "novelas de costumbres".[104] The label was used by convention but also with specific, explicitly formulated purposes. The Mexican author José Tomás de Cuéllar, for example, was a prolific writer of novels of customs. He produced a whole series called "La linterna mágica" (published between 1871 and 1892). The first novel of the series is preceded by an introduction explaining Cuéllar's literary program. The magical lantern is a symbol for the purpose of the novels of customs. Cuéllar claims that he first happened upon the expression in a small corner shop of some Mexican village. This is already indicative of the aim to represent local and rural customs.

---

[103] In the Mexican case, a different view on the *costumbrista* tradition sees the origin of the novels of customs not in Spanish models but in the early works of the Mexican author Fernández de Lizardi (Calderón 2005, 316–317).

[104] "Novela de costumbres" is the second most frequent explicit label for the thematic subgenres. Independently of the explicit label, the novels of customs on the sixth rank of the most frequent primary thematic subgenres in the bibliography and on the third rank in the corpus. See chapters 4.1.5.1 and 4.1.5.3.1 for the corresponding overviews.

The lantern illuminates all kinds of spots the author visits, showing the multitude of ordinary people he wants to portray. It enlarges the view on vices and defects and, at the same time, minimizes the size of each person so that they come to the fore above all as social groups. The reader is invited to follow the lantern's light which accepts the insufficiencies of the people with humor and also presents alternative models of virtue (Cuéllar 1890, xii–x). The consciousness and the frequency with which the term "novela de costumbres" was employed show the strength of the historical generic convention. However, the novels of customs are not detached from other prominent thematic subgenres. Novels can be at the same time historical and of customs, as for example "Calvario y Tabor. Novela histórica y de costumbres" (1868, MX) by Vicente Riva Palacio or "Julia" (1868, MX) by Manuel Martínez de Castro, which had the subtitle "novela de costumbres mexicanas" in the first edition and "novela histórica y de costumbres" in the second one. There is also a close relationship between novels of customs and sentimental novels, as Janik points out:

> Die literarische Darstellung von Natur, Landschaft, landwirtschaftlicher Arbeit, von Gutshäusern und Gutsverwaltung, ebenso von städtischem Leben und städtischen Sitten sollte die gekannte Wirklichkeit ansichtig machen, doch eingebunden in eine tragende Handlungsstruktur. […] Die Hauptaufgabe, die sich jedem Autor stellte, war die erzählerische Entwicklung eines komplexen Geschehens mit dramatischen Peripetien. Im Gefolge der genannten französischen Autoren [Balzac und Stendhal] und ihrer 'romantischen' Wegbegleiter (Rousseau und Chateaubriand) bildet in einer ganzen Reihe von Werken die Entstehung einer tiefen existentiellen Liebesbindung […] die Grundstruktur. (Janik 2008, 67–68)

A prominent example of such a combination of the *novela de costumbres* and the sentimental novel is the Cuban work "Cecilia Valdés o La Loma del Ángel" (1839/1882) by Cirilo Villaverde, which contains the opposing tendencies of a romantic love story and a political stance focusing on the Cuban reality in the 1830s (Janik 2008, 75–77). Although they form mixed types with other romantic subgenres, the novels of customs still represent early forms of realistic texts. In the later nineteenth century, they entered into similar relationships with the realist and naturalistic novels. As Kohut notes, the delimitation of Romanticism, Realism, and Naturalism is difficult in Spanish America for several reasons, one of them being the *Costumbrismo* as an element between the different literary currents:

> Zum Realismus gehört die Zuwendung zur Gesellschaft, zur Romantik die häufig idyllisierende Perspektive. […] Wichtiger als der *Costumbrismo* als eigenständige literarische Richtung ist die entsprechende Einfärbung zahlreicher realistischer bzw. naturalistischer Romane. So gab der Chilene Alberto Blest Gana seinem Roman *Martín Rivas* (1862) den Untertitel *Novela de costumbres político-sociales*, der Argentinier Lucio Vicente López seinem Roman *La gran aldea* (1884) den Untertitel *Costumbres bonarenses*. (Kohut 2016, 196)

Like the historical novel, also the *novela de costumbres* is a primarily thematic subgenre that was vital in the whole nineteenth century, although it was most popular in the Romantic period.

Because one of its aims is to depict local habits realistically and faithfully, including linguistic peculiarities, there may be stylistic traits that help to distinguish this type of novel from the other thematic subgenres. However, in the literary-historical characterizations, this subgenre appears as one that enters into fusions with quite distinct other subgenres, which might make it difficult to classify it on the basis of features related to themes.

2.3.1.3 *Novela sentimental*

The third thematic subgenre that was chosen for text analysis is the sentimental novel proper. Contrary to the historical novel and the *novela de costumbres*, sentimental novels are usually not marked with explicit subgenre labels by authors or editors. In the whole bibliography and corpus that were compiled for this study, there is just one work carrying the subtitle "novela sentimental" – the novel "El canto del cisne (Novela sentimental)" (1910, AR) by Roque C. Otamendi. From the point of view of historical genre convention, the status of sentimental novels is, therefore, different from that of historical novels and novels of customs. It was a subgenre that was signaled in more subtle ways, for example, by using female first names as titles, e.g., "Soledad" (1847, AR) by Mitre, "Esther" (1851, AR, Cané), or "Clemencia" (1869, MX) by Altamirano. Furthermore, a sentimental plot is present in the majority of the nineteenth-century Spanish-American novels, which makes this subgenre one that is part of a general generic repertoire of the time. It can appear in a pure form or as a basic element to provide a narrative structure for novels that have other, additional, or superordinate thematic objectives. It can, for example, motivate descriptions of historical and contemporary conflicts and settings. Because of the omnipresence of sentimental elements in different kinds of novels over time, the sentimental novel has also been called a "metagenre", which might challenge the classification of novels into subgenres (Varela Jácome [1982] 2000, sec. 1.4; Zó 2015, 14–16).

In pure form, the plot of a sentimental novel is centered on a highly personal conflict with exceptionally sensitive protagonists whose development of emotions is shown in the novel. Usually, the conflict is resolved in either a tragic or a happy end (Molina 2011, 375–386). Typical structural elements of sentimental novels are letters or passages of diaries included in the text, which give insight into personal communication and the protagonists' mental world and sensitivities. As the historical novel, also the sentimental novel has European, especially French models that the Spanish-American authors followed, for instance, the epistolary novel "Julie ou la Nouvelle Héloïse" (1761, FR) by Rousseau, "René" (1802) by Chateaubriand, or the autobiographic novel "Graziella" (1852, FR) by Lamartine, but also the German epistolary novel "Die Leiden des jungen Werhers" (1774) by Goethe. In the sentimental novels proper, the protagonists are described in idealizing terms, emphasizing their moral superiority and tending to a flowery and metaphorical style. The heroine of Mitre's novel "Soledad", for example, is described as "imagen escapada de las telas de Rafael", "un serafín bajado del trono del Señor", and "la estatua de la castidad meditando". Not only the protagonists but also the settings are described from subjective perspectives (Varela Jácome [1982] 2000, sec. 1.4). The novels are often written in the first person.[105] If they do not have a European setting, the action is commonly placed in the

[105] See the overview of thematic subgenres by narrative perspective in chapter 4.1.5.3.1.

bourgeois milieu of the Creole middle or upper class and thematically focused on private life, concerned with the home, love, and family (Dill 1999, 138).

The prototypical sentimental novel should have a recognizable textual coherence, as it has its own set of structural elements, theme development, and stylistic means. The assumption is that it forms a separate text type besides the historical novels and the novels of customs that have a more realistic style. Moreover, the historical novels almost exclusively have a third-person narrator. This can have a significant influence on the novels' style, at least if the textual features used for categorization include personal pronouns and declined verb forms. The *novelas de costumbres* that are part of the corpus, in contrast, are half narrated in the first and half in the third person, which supports the observations made by Janik and Kohut about their intermediate status. As for the sentimental novel in a wider sense, that is, novels with a primary sentimental plot, but, for example, a historical setting or social and political concerns, the textual distinction of the various subgenres is assumed to be more difficult.

2.3.2 Subgenres Related to Literary Currents

2.3.2.1 *Novela romántica*

In Spanish America, Romanticism was the dominant literary current in the second third of the nineteenth century and continued to influence the production of novels until the end of the century. Compared to the development of the romantic current in the European countries and North America, its acclimatization in Spanish America was delayed because of the wars of independence and subsequent political struggles such as civil wars and oligarchic regimes. The cultural sector was not fully developed yet, local cultural models were missing, and also the ideological climate was different. Nevertheless, foreign models of various kinds entered the cultural area of the Spanish-American countries discontinuously and were soon adopted, for example, French sentimental novels, or the historical novels of Walter Scott and James Fenimore Cooper (Dill 1999, 120; Lichtblau 1959, 65; Varela Jácome [1982] 2000, sec. 1.1.3). In Spanish America, the romantic current is connected to political ideas of liberalism and nationalism. The writers are "civic poets" and not primarily literati, and they use the novel to denounce social and political grievances, promote ideologies, and suggest new models of society (Dill 1999, 121).

Formally, romantic novels are characterized by looseness, unconventionality, and the rejection of formal rigor. The protagonists are idealized heroes, who are often solitary, rootless and driven by instincts. Emotions and motives of the characters tend to be represented in stereotypical ways, and there is a tendency to reduce conflicts and rivalries to simple contrasts between good and bad characters, whose role is underlined by employing a corresponding oppositional descriptive vocabulary (Dill 1999, 120, 127–128). In the novels, verisimilar elements can be combined with fantastic ones. Typical elements of the plot are turbulent and violent actions, deception, fraud, and surprising effects, for example, through disguises and discoveries (Varela Jácome [1982] 2000, sec. 1.4). Even passages of pure description, for example, those concerned with representing the local landscape, are characterized by the use of a subjective and idealizing style (Lichtblau 1959, 66).

The romantic current put forth several distinct types of subgenres of the novel, among them sentimental, historical, political, social, adventure, indianist novels, and also the novels of customs. That different subgenres of the novel played an important role in the period in which the romantic current was dominant is readily apparent from the structure of many literary histories concerned with Spanish-American nineteenth-century literature. Usually, the description of the romantic novel is subdivided into separate sections concerned with the different subgenres, which is not always the case for the later currents.[106]

Regarding the textual coherence of the romantic novels, several different kinds of hypotheses can be formulated. Thematically, the inclusion of a sentimental story is a typical element of all kinds of romantic novels and might facilitate its classification against realist and naturalistic novels. Even if a sentimental plot also occurs in novels belonging to the other currents, it is less typical. Then again, the clearly distinguishable thematic subtypes of romantic novels lead to the assumption that the conventional romantic subgenre might fall into several textual subgroups if semantic features are used. In the narrow sense of linguistic style, though, it is probable that the romantic novel is coherent. A hypothesis is that specific parts of speech, vocabulary, and punctuation marks are used to express subjectivity and emotionality, for instance, many personal and possessive pronouns, qualifying adjectives and adverbs, interjections, or exclamation and question marks.

What makes a clear distinction of works by current difficult is that there are individual works that combine elements of several currents. Especially romantic and realist elements are often combined, which is an effect of the chronologically discontinuous integration, adaptation, and further development of foreign currents in the Spanish-American context.[107] Furthermore, the presence of romantic works up to the end of the nineteenth century raises the question of whether there is a contrast between an early and a late romantic style which could become visible in the categorization of the texts.

---

[106] Dill, for example, structures the chapter on the romantic novel into the subsections "Der politische Roman", "Der historische Roman", "Der indianistische Roman", "Der kubanische negristische Roman", and "Der sentimental Roman". The chapter on the realist novel is not subdivided, and the one on the naturalistic novel only has a subchapter concerned with the city novel ("Der Großstadtroman") as a special type of naturalistic novel (Dill 1999, 125–139, 159–166, 168–176). In the introduction to her book on the Spanish-American romantic novel, Suárez-Murias lists the sentimental novel, the indianist novel, the historical novel, the *costumbrista* novel, the *Roman à thèse* (*novela de tesis*), and the dime novel (*novela de folletín*) as subgenres of the romantic current (Suárez-Murias 1963, 12–13). Gálvez, who studies the Spanish-American novel up to 1940, structures the chapter on the novel of the romantic period into subchapters on the historical and the political, the indianist, and the sentimental novel. She dedicates another subchapter to the novel of the transition to realism. That chapter includes parts on the historical, social and *costumbrista* novel, the latter including the *gaucho*, *indio*, and antislavery novel. As she is concentrating on the novel alone, Gálvez's account is more differentiated than the one of Dill. She takes up several subgenres again in the chapters on the later realist, modernist, and regionalist currents, for example, the historical novel and the novel of customs, showing that they did not cease to exist, but continued to be practiced under the influence of different literary currents (Gálvez 1990). Nevertheless, the later currents did not produce the same range of new, own, distinguishable, and widely recognized thematic subgenres as the romantic current did.

[107] On the simultaneous presence of romantic and realist elements, see Lichtblau (1959, 66) and Varela Jácome ([1982] 2000, sec. 2).

2.3.2.2 *Novela realista*

In its artistic meaning, the term "Realism" emerged in the eighteenth century to designate an aesthetic model opposed to the idealism and individualism propagated by Romanticism (Álamo Felices 2011, 118). In Europe, it began to spread in the first half of the nineteenth century and dominated from the middle of the century onwards. In Spanish America, in contrast, numerous realist novels were only published after 1880 (Varela Jácome [1982] 2000, sec. 3). The realist movement was linked to the formulation of positivist philosophical theories and the development of the natural sciences. According to the positivist philosophy, new insights should derive from positive findings that are sensually graspable, actually observable, and verifiable. For the arts, this meant that their primary objective was to reflect the social reality aesthetically by accurately depicting the characters and describing the environments meticulously. In realist novels, dialogues are an important means to develop the action. They are represented in detail but without the need to put aside the omniscient perspective. In sum, the goal was to create an effect of verisimilitude, an *effet de réel*, which aimed to make the reader believe that what is described is the real world. By convention, a relationship of identity is assumed between the contemporary world and the mimetic fictional world. To what extent the author is understood as contributing to the creation of the fictional world with his constructive and imaginative work depends on how closely the ideal to reflect the external world directly and mimetically is followed (Álamo Felices 2011, 118–119).

That the depicted world is not shown through an idealizing filter in the realist novels entails that not only the bourgeois milieu is represented but also lower social strata. Social and political ills, such as corruption, intrigues, lobbyism, or poor education, are disclosed, for example, in the novels of the Mexican writer Emilio Rabasa. The portrayal of personal defects, such as greed, excessive materialism, or drive for recognition, is as well part of the realist repertoire, for example, in the Cuban novels "Un hombre de negocios" (1883) by Nicolás Heredia or "Mi tío el empleado" (1887) by Ramón Meza (Instituto de Literatura y Lingüística de la Academia de Ciencias de Cuba 1999, sec. Realismo). On the one hand, the Spanish-American realist novels described the rapid economic and social development and life in the big cities, for instance, in the novel "La gran aldea. Costumbres bonaerenses" (1884) written by the Argentine Lucio Vicente López. On the other hand, rural environments were also a topic, for example, in the novel "La parcela" (1898) by the Mexican author López Portillo y Rojas (Rössner 2007, 148, 188).[108]

Realism as a literary current prevailed in the second half of the nineteenth century. Besides the novels that are directly associated with this current, realistic elements are already present in earlier works. For instance, the *novelas de costumbres* also aspired to represent the local world and society in all its facets, but with a different motivation than the realist novels in the proper sense. Furthermore, especially Cuban novels are marked by a realistic orientation from early on and persistently in the nineteenth century because many authors were concerned with the social ills caused by the system of slavery. In terms of literary currents, this leads to works that mix romantic and realist(ic) elements, for instance, the novel "Francisco" (1839/1880) by Anselmo Suárez Romero:

[108] For the defining characteristics and topics of the realist novel, see also Navarro (1955, 20–24).

> la idílica presentación de los desgraciados amores de Francisco y Dorotea contrasta con las escenas de la penosa vida de los esclavos en los barracones y los castigos inhumanos que les eran infligidos por parte de sus mayorales, descritas con gran crudeza. Esta coexistencia de elementos de ambas normas estéticas – la romántica y la realista –, que tan tempranamente se inicia, caracteriza buena parte de nuestra narrativa decimonónica y perdura hasta los inicios del presente siglo (Instituto de Literatura y Lingüística de la Academia de Ciencias de Cuba 1999, sec. Realismo)

The aesthetic closeness of the early Cuban novels with the later realist and naturalistic novels has led Molina to consider some nineteenth-century novels as instances of Naturalism before its time. Her study covers works written and published between 1830 and 1927 (Molina 2001).

The presence of realistic aspects in some of the novels of the first half of the nineteenth century is expected to complicate the textual separation of romantic and realist novels. On the other hand, the simultaneity of romantic, realist, and naturalistic works in Spanish America after 1880 and the existence of novels that included elements of several currents at once is another factor that could render the classification of the works by literary current more difficult.[109] However, the assumption can also be made that there are prototypical cores of romantic, realist, and naturalistic novels in the sense of distinctive genres following aesthetic models that primarily have a European origin. In the case of the realist novel, elements that could lead to a coherent text type are specific topics such as life in the city, business and economic speculation, and the portrayal of different social classes. On a stylistic level, the hypothesis can be formulated that there are many and long descriptive passages and dialogues which have specific linguistic characteristics, for example, regarding the use of many different adjectives and nouns. These could contribute to the identification of a realist text type. A second hypothesis is that besides a realist text type in the narrow sense, another text type could become visible, in which realist and romantic elements are combined and in which the many novels that have conventionally been described as mixtures of romantic and realist(ic) elements take part.

2.3.2.3 *Novela naturalista*

Like the first Spanish-American romantic and realist novels, also the naturalistic ones started from a foreign model that was successively adapted to the particular needs of expression in Argentina, Cuba, and Mexico. In the case of the naturalistic novel, the model was primarily supplied by the French author Émile Zola, who propagated a scientific conception of man and a biological and medical way of thinking. In the naturalistic novel, realist techniques were carried forward to utilize the literary work as a means of experimentation and and for case studies of

---

[109] In literary-historical accounts of the nineteenth-century Spanish-American novel, some works are described as transitional between the romantic and realist currents, for example, "La Calandria" (1890, MX) by Rafael Delgado, or the historical novels of the Mexican writer Juan Antonio Mateos (Varela Jácome [1982] 2000, sec. 2.1.3; Gálvez 1990, 96, 105). Furthermore, literary historians come to different conclusions regarding the status of some works in relationship to literary currents. The novel "Cecilia Valdés o La Loma del Ángel" (1839/1882, CU) by Villaverde, for instance, is described as primarily romantic with realistic elements by Suárez-Murias and Varela Jácome, as realist by Dill, and as transitional between both currents by Gálvez (Dill 1999, 160–161; Gálvez 1990, 115–117; Suárez-Murias 1963, 36–40; Varela Jácome [1982] 2000, sec. 1.3).

milieu, anthropology, genetics, or pathologies (Dill 1999, 168–169). Compared to the realist novel, the importance of the author's creativity for producing an effect of reality is replaced by the aspiration for a pure documentary style. Nature shall be discovered "as is". The development of the characters is a logical one, and inner reactions are subjected to mechanical processes. Reality is shown in its raw form, and no moral limits are set to the kinds of environment or human abysses that are studied. The goal is to stimulate the readers' reflection by reproducing things as they are (Varela Jácome [1982] 2000, sec. 4). In terms of characters, the broad ensemble of the realist novel tends to be reduced, and the focus is on the protagonists whose psychological development is examined. The characters expose themselves through their way of speaking, either in dialogue or indirect speech, showing the regional or social determination of their means of expression. The influence of the milieu on the life of the characters is worked out (Dill1999,169).

In France, the naturalistic novel spread from the 1870s onwards, and in Spanish America it became popular in the 1880s. This shows that the adoption of this literary current happened with less delay than in the case of the romantic and realist novel. However, this led to the already mentioned superimposition of all three currents in the Spanish-American countries (Varela Jácome 2000, sec. 4). In Spanish America, naturalistic works were published in several different countries and in all of the three countries that are the focus here. The first and most naturalistic novels appeared in the Río de la Plata region, which includes Argentina.[110] Authors whose works have been associated with the naturalistic current are, for example, the Argentine Eugenio Cambaceres, the Mexican Federico Gamboa, and the Cuban Martín Morua Delgado. For the authors, the naturalistic aesthetic provided an appropriate means to reflect the rapid economic upswing, as well as the technological and social processes of modernization that took place in the Spanish-American countries in the last decades of the nineteenth century, with all the challenges and difficulties that these brought for individual social groups or people (Schlickers 2003, 9–10). The range of topics covered in naturalistic novels includes aspects of biological and environmental determinism, immigration, mental ills, materialist neuroses, sexual rage, problems of alcoholism, the destructive force of prostitution milieus, and the representation of the urban Moloch. For example, the working conditions in factories and mines, but also the tedium of idle members of the upper class, are taken into account (Dill 1999, 169; Schlickers 2003).

Naturalistic novels share some features with realist novels because of the ambition to depict reality closely, and the different literary currents concurred temporally in the last decades of the nineteenth century. There are, however, some elements of naturalistic novels on the basis of which it can be hypothesized that they are distinguishable as textual genres. The crude style, usage of scientific or other specialized vocabularies, and the specific topics of this type of novel, such as the topics of immigration, mental and physical problems, and processes of social disintegration, are expected to facilitate their classification against realist and especially romantic novels. It is assumed that the novels that are conventionally labeled as naturalistic ones are also textually coherent to a relatively high degree. However, a difficulty might arise from the confusion and flexibility of the terms "realist" and "naturalistic" in the corresponding literary critical discourse

[110] The great number of naturalistic novels that were written in Spanish America becomes evident in the comprehensive study that Schlickers published on the Spanish-American naturalistic novel. She includes 63 novels in her book and discusses almost every novel in its own chapter (Schlickers 2003).

and thus from the conventional class labels. If these terms are used with an overlapping or even identical scope, they are no longer a sign for different subgenres.[111]

[111] On the difficulty to delimit the terms and concepts of *costumbrismo*, *realismo*, *regionalismo*, and *naturalismo*, see Navarro (1955, 12–19). Sánchez, for instance, only uses the term "novela naturalista" for works of both the realist and naturalistic types (Sánchez 1953, 257–259).

3 Corpus

The corpus used for the analysis of subgenres in this dissertation is presented in this chapter. Besides the text corpus itself, a bibliographical database of nineteenth-century Spanish-American novels was created. On the one hand, it had the purpose of serving as an information pool from which to retrieve data about authors, works, and editions during the process of corpus creation. On the other hand, it approximates the population from which the actual text corpus was sampled so that eventual particularities of the corpus can be assessed. Furthermore, the digital bibliography and corpus, which were created in the context of this thesis, constitute general databases for digital text or metadata analysis on nineteenth-century novels from Argentina, Cuba, and Mexico. In this chapter, all the aspects of these two resources that are relevant for their use in digital genre analysis are presented so as to provide a thorough documentation of both databases and to encourage reuse, even if not every aspect of the metadata and text encoding is used in the text analysis part of this dissertation.

The chapter is organized as follows: In chapter 3.1, the criteria used for the selection of texts for the bibliography and the corpus are discussed. The creation of the bibliographical database and the corpus itself – their sources, data model, text treatment, metadata, and text encoding – are outlined in chapters 3.2 and 3.3. Overviews of the contents in the bibliography and the corpus are given in the chapter following this one: In chapter 4.1, the authors, works, editions and subgenres contained in both resources are analyzed and compared regarding their distribution by selected metadata and text parameters (for instance, by country and time period). At some points, the discussion of the selection criteria in chapter 3.1 already refers to digital bibliographical information and full texts as bases for decision-making because the processes of defining the selection criteria and building the databases went hand in hand: an initially broad data basis was analyzed and successively cut to satisfy stricter criteria.

3.1 Selection Criteria

Unless otherwise stated, the selection criteria that are discussed in this subchapter apply both to the bibliographical database and to the text corpus. As the subject of this study are subgenres of the novel, a definition of the novel itself as the higher-level genre is necessary to be able to select the texts. Texts of all kinds of subgenres are included, even though the analysis focuses on some of them: determining the subgenres is a topic in itself and the corpus serves as a background foil for individual subgenres. The *boundaries of the novel* are discussed in chapter 3.1.1. Although this dissertation aims to analyze subgenres of Spanish-American novels, not all of the countries belonging to the region are taken into account simply because it would be too challenging to regard all the individual literary-historical contexts of the new nations and old colonies. Instead, it was decided to concentrate on three countries: Mexico, Cuba, and Argentina. In chapter 3.1.2, it is explained why these three countries were chosen and how it was decided which novels are associated with each of them. Chapter 3.1.3 explains which limits of the nineteenth century were used here to select the texts.

To facilitate an understanding of the examples, also in the cases of lesser-known works, whenever individual works are mentioned, the year of their first publication and a country code is given in parentheses after the title. For all the selection criteria, it was an objective to find ways to decide that are suitable for a quantitative study, in that the amount of necessary close reading of the texts is kept as low as possible, with the goal to make the selection criteria in principle applicable to a corpus of any size.

3.1.1 Boundaries of the Novel

The bibliography and corpus are intended to include literary texts that belong to the genre *novela*. In general, a novel can be defined as a longer fictional narration in prose that is usually published as one or sometimes several independent books (Fludernik 2009, 627; Steinecke 2007, 317). Besides the general characteristics of the form, manifestations of the novel are very varied, for example, regarding the content of the texts and the kinds of characters or elements of the plot. Most of the criteria that go beyond the broad formal characterization of the genre are only valid for one or several subgenres, excluding others.[112] Because no subgenres or types of novels are excluded here from the outset, the general definition of the novel is followed. However, even the above-mentioned formal elements need to be clarified further because they depend on the cultural and historical context under consideration.[113] In the following, the individual elements of the above definition of the novel (fictionality, narrativity, prose, length, independent publication) are discussed for the Spanish-American context in the nineteenth century. The methods used to assess these properties for the texts in question are outlined, with a special focus on borderline cases, in order to exemplify where the boundaries of the novel were drawn. Finally, additional criteria complementing the formal aspects are explained, and the various factors are summarized in a working definition of the novel.

3.1.1.1 Fictionality

In a pretheoretical understanding, fictionality describes the property of a text (or other medium) to involve fiction, which means that it is about something imagined and invented. A novel, for example, is about events that did not actually take place, even if the author was inspired by the reality he or she knows and even if the author alludes to this reality in the text. Even so, theoretical considerations of fictionality show that it is not enough to assume that a text is fictional if it is about imaginary worlds (Klauk and Köppe 2014, 3).[114] Recent approaches focus

[112] For example, the concentration on one principal character in the *Bildungsroman* versus a broad picture with several important characters in historical novels (Fludernik 2009, 628).

[113] Up to the seventeenth century, for example, also works in verse form could be called *Romane* (novels). (Steinecke 2007, 317).

[114] Nevertheless, earlier attempts to define fictionality that focused on reference semantics took this line. According to these views, fictional and factual texts are characterized by a respective specific mode of referentiality. Whereas a factual text could be defined as a text that references the empirical reality directly and is conceived and perceived as such by the author and its readers, a fictional text predominantly references invented places, characters, or events. Elements existing in the real world can also be part of a fictional text, but there should be elements in a fictional text that do not have any counterpart in reality. This means that these elements cannot be referred to anything existing before and outside of their linguistic formulation and creation in the text. However, it has been

on pragmatic aspects to determine the fictionality of a text. According to the "institutional" theory of fictionality, for example, certain texts are considered fictional because of a coordinated and conventional social practice (an institution). A text is produced with the intention to be received according to the conventions of the fictionality institution. The sender and recipient of a fictional text enter into a contract establishing that questions of empirical referentiality and truth are not posed within the confines of the fictional text. The reader accepts the existence of the entities presupposed in the text and engages with them imaginatively if he or she recognizes the intention of the author to write a fictional text. For this, the authorial intention needs to be manifest in the text in some way, but ultimately, it is a pragmatic attribution to determine the fictional intention of a text (Köppe 2014, 35; Weidacher 2017, 378–381).

In accordance with this view, the fictionality of the texts to be included in the bibliography and the text corpus was assessed as follows. Statements of authors and readers regarding the fictionality of a text were taken into account. If it was indicated clearly that the text was conceived and received as fictional at the time and place of its publication, these signals were highly rated. In addition to explicit statements concerning the fictionality of the text, other paratextual and textual signals were evaluated. A comprehensive overview of potential signals of fictionality is given by Zipfel (Zipfel 2014, 97–119), who organizes them as follows:

- textual signals
- peritextual signals: place of publication, for example a specific collection or journal; book format; publisher; series; author; title and subtitle, possibly with genre labels; cover text; dedication; preface; etc.
- epitextual signals: publisher's statements, reviews, interviews with the author; etc.

Of the various potential signals of fictionality, peritextual signals were especially useful to evaluate whether texts are to be considered fictional and if they should become part of the bibliography and the text corpus because they are very accessible.[115] Details such as author, title and subtitle, place of publication, publisher, and series are usually included in bibliographic descriptions of work editions and can, therefore, also be taken into account when the texts themselves are not available.[116] A good indicator is a genre label in the title or subtitle of a work that refers to a fictional text type. Examples of such titles for Spanish-American narrative texts in the nineteenth century are: "novela", "relato", "narración", "leyenda", "romance", "cuento", or "drama". There are other labels that are also common but less clear regarding the fictional status of the texts, for example: "historia", "crónica", "estudio", "esbozo", "cuadro", "escenas", "episodio", "memorias", "apuntamientos", "anécdotas". Sometimes labels refer to subgenres, such

shown that the assumption of different modes of referentiality for fictional and factual texts is problematic, at least in the outlined narrow conception of the term (Weidacher 2017, 375–378).

[115] The peritext includes paratexts that are published as part of a work, e.g., prefaces and dedications, whereas the epitext involves paratexts that are published outside of the immediate context of the work (Genette 1987).

[116] For some of the works listed in bibliographies of the Spanish-American novel in the nineteenth century, no edition could be found neither in the WorldCat (a union catalog currently containing more than 430 million bibliographic entries from libraries worldwide, see OCLC 2023) nor in individual relevant library catalogs. Even if editions could be located, it was not in every case feasible to see them, especially if there were only print editions distributed over different American libraries. The issue of accessibility of the texts will be discussed in more detail in chapter 3.3.1 ("Selection of Novels and Sources").

as “aventuras” or “costumbres”. To be able to decide whether a text is to be considered fictional or not in cases where labels are ambiguous, or where there are no explicit labels at all, other kinds of information were used. Where editions of a work were accessible, prefaces, introductions, and headings were consulted to see whether they clear up the issue of fictionality. Textual signals on the level of the story and on the level of the narration were also taken into account, but only in cases of doubt. A textual signal that is easy to recognize typographically and is typical for fictional narrative texts, though it is neither a necessary nor a sufficient criterion, is the reproduction of direct speech. Words or phrases that mark the end of a story or text can also be easily identified. Epitextual signals were not systematically researched. Especially for the bibliographical database, decisions were also based on information from existing bibliographies of fictional texts, literary histories, and other critical research literature.

In the case of Spanish-American novels, there are several factual text types that share characteristics with certain subtypes of the novel in terms of content or narrative mode. These are historiographic works versus historical novels, (auto)biographies versus (auto)biographical novels, travelogues versus travel novels, philosophical treatises versus philosophical novels, political treatises versus political novels, etc. That the boundaries between some kinds of fictional and factual texts are not always clear is influenced by several factors. Many of the authors in the nineteenth century who wrote novels were also authors of historiographic, political, journalistic, or philosophical works because there were still very few professional literary writers. Furthermore, many Spanish-American countries reached their political independence in the early nineteenth century, and there was a need to justify it and to contribute to the creation of a national identity not only through historiography but also by means of literary works (Kohut 2016, 171–172; Lindstrom 2004, 76–77; Sommer 1993). In his essay “Revistas literarias de México” from 1868, the Mexican author Ignacio Manuel Altamirano explains the ever more important role of the novel in this process:

> La novela es indudablemente la producción literaria que se ve con más gusto por el público, y cuya lectura se hace hoy más popular. Pudiérase decir que es el género de literatura más cultivado en el siglo XIX y el artificio con que los hombres pensadores de nuestra época han logrado hacer descender a las masas doctrinas y opiniones que de otro modo habría sido difícil que aceptasen. […] la novela hoy ocupa un rango superior, y aunque revestida con las galas y atractivos de la fantasía, es necesario no confundirla con la leyenda antigua, es necesario apartar sus disfrazes y buscar en el fondo de ella el hecho histórico, el estudio moral, la doctrina política, el estudio social, la predicación de un partido o de una secta religiosa: en fin, una intención profundamente filosófica y trascendental en las sociedades modernas (Altamirano 1868, 17–18)

As long as they are either designated directly or indirectly as fictional in their paratexts or exhibit characteristics that are typical for fictional texts, these works were included in the bibliography and the corpus, even if they resemble factual texts because of their content or because of the way the narration is organized.

For example, the Mexican author Ireneo Paz wrote several historical novels that he labeled as such, but also a series of “leyendas históricas”. They are all centered on historical figures, as their titles suggest: “El Lic. Verdad”, “La Corregidora”, “Hidalgo”, “Morelos”, “Mina”, “Guerrero”,

"Antonio Rojas", "Manuel Lozada", "Su Alteza Serenísima", "Maximiliano", "¡Juárez!", "Porfirio Díaz", and "Madero" (Pi-Suñer Llorens 2005, 386). They could also be interpreted as historical biographies, but because they are labeled as "legends" and contain direct speech, detailed descriptions of situations (e.g., weather conditions) and characters (e.g., behavior and appearance in specific situations), they are considered fictional texts here.

A work that is sometimes mentioned in critical works on the Spanish-American novel is "Vida de Juan Facundo Quiroga" (1845, AR) by Domingo Faustino Sarmiento.[117] In the first part of the work, the country, its inhabitants and their customs are described, followed by a biography of the Argentine caudillo Juan Facundo Quiroga. The last part contains considerations about Argentina's political and economic future (Lichtblau 1959, 39–40). In a preface, the author refers to reactions by readers who missed certain details in the descriptions of historical events. Sarmiento defends himself by explaining how difficult the coordination of events that occurred in so many different places and at so many different points in time was challenging with the limited means he had (some reports of eyewitnesses, some simple manuscripts, some aspects recalled from his memory). He ends with the intention to improve his work in these aspects if time allows:

> Quizá haya un momento en que, desembarazado de las preocupaciones que han precipitado la redacción de esta obrita, vuelva a refundirla en un plan nuevo, desnudándola de toda digresión accidental, y apoyándola en numerosos documentos oficiales, a que sólo hago ahora una ligera referencia. (Sarmiento [1845] 2000, sec. Advertencia del autor)

In this authorial statement, there cannot be recognized any intention to write a fictional text. Moreover, the different parts of the work are not unified, and there are very few passages where direct speech is reported. "Vida de Juan Facundo Quiroga" is therefore considered a non-fictional text and excluded from the bibliography and the corpus.

Other borderline cases are descriptions of travels, for example, "La tierra natal" (1889, AR) by Juana Manuela Gorriti, "Mis montañas" (1893, AR) by Joaquín Víctor González, and "Una excursión a los indios ranqueles" (1870, AR) by Lucio Victorio Mansilla. All three texts also include autobiographical elements. For a factual travel narrative, three conceptual aspects are essential:

1. the discourse is organized around a journey, for example, according to its itinerary or its chronology,
2. the narrator can be identified as the author who recounts his or her experiences, which presupposes that there was an actual journey that took place before the narration,
3. there is a general tendency towards description and objectivity.

The travel narrative can furthermore be identified on the basis of paratextual signals (Alburquerque-García 2015, 19–29; Chávez and Urdapilleta 2015, 11). The discourse of a fictional travel narrative is equally organized around a journey, but the narrator can usually not be identified with the author, and no actual journey is needed as a basis for the narration. In addition, in a fictional travel narrative, narration tends to prevail over description and subjectivity

[117] In the portal "Novela hispanoamericana del siglo XIX", which is part of the "Biblioteca Virtual Miguel de Cervantes", for example, "Facundo" is classified as "novela histórica argentina" but also as a biography of Juan Facundo Quiroga and as an autobiography of the author Sarmiento (Sarmiento [1845] 2000).

over objectivity. It is likely that there will be paratextual signals pointing to the fictional status of the text. The second aspect (identification of the narrator as the author) is also relevant to distinguish an autobiography from an autobiographical novel. In the latter, the narrator and protagonist are not to be identified with the author.

When the three examples are examined, the following characteristics can be determined. In "La tierra natal", the framing story is a railway trip from Buenos Aires to Salta. The text is structured into chapters that roughly correspond to stops of the journey. The traveler and first-person narrator gives an account of the journey and inserts conversations of fellow passengers, but also memories of her hometown. In a preface, Gorriti calls her work "páginas de lejanas memorias" (Gorriti [1889] 2001, 1). The end of the narration is marked with the word "Fin".

In "Mis montañas", the first-person narrator gives a report of a trip to the Sierra de Velazco in the Argentine province of La Rioja. The text is divided into 21 chapters which consist of landscape descriptions and impressions, historical background information and the imagination of historical events, the portrayal of local customs, the evocation of local characters and episodes, and personal memories. The work is prefaced by the Argentine writer Rafael Obligado, who gives several intertextual references. For example, he compares "Mis montañas" to the epic poem "La cautiva" by Esteban Echeverría. However, he does so not to stress its fictionality but the literary treatment of the Argentine landscape: "La propiedad artística de la cordillera argentina pertenece a Vd. de hoy para siempre, como la de la llanura al poeta de La Cautiva" (González [1905] 2001, X).[118]

"Una excursión a los indios ranqueles" begins with a letter written by the narrator, identified as "Lucio" and "coronel Mansilla", just like the author, to his friend Santiago, in which he explains the circumstances of his expedition to the province of Córdoba where the *indios ranqueles* live. In 68 chapters, the narrator recounts his experiences in the form of letters to his friend. The work contains descriptive passages concerned with sociological, zoological, botanic, philological, and folkloristic facts, but also an intercalated novella and novelistic amatory and military scenes (García 1952, 132; cited by Lichtblau 1997, 609; Rössner 2007, 186–187).

In all three works, the discourse is organized around a journey that actually took place. All the texts are narrated in the first person, and the narrator can be identified with the author, either because of an explicit mention in the text ("Una excursión a los indios ranqueles") or because of implicit formulations in the prefaces ("La tierra natal" and "Mis montañas"). In the paratexts, there is no clear evidence that the three travelogues were conceived or perceived as fictional. As to the third defining aspect of a factual travel narrative, the nature of the three works under consideration is less clear. All of them combine descriptive with narrative passages and objective representations with subjective perceptions to different degrees. One indicator for a narrative style, and hence a fictional text, that can be evaluated quantitatively is the amount of direct speech in the three texts. In figure 2, the travelogues are compared to other novels in the corpus regarding the proportion of paragraphs containing direct speech.[119] As can be seen, the amount

[118] Other similar references are made to the authors Alexander von Humboldt, Jacques Bernardin Henri de Saint-Pierre, William Wordsworth, Gregorio Gutiérrez González, François-René de Chateaubriand, Henry Wadsworth Longfellow, Domingo Faustino Sarmiento, and Olegario Víctor Andrade (González [1905] 2001, XII, XX).

[119] The script that produced the box plot is available at https://github.com/cligs/scripts-nh/blob/master/corpus/direct-speech-travelogues.xsl and the result at https://github.com/cligs/data-nh/blob/master/corpus/direct-speech-

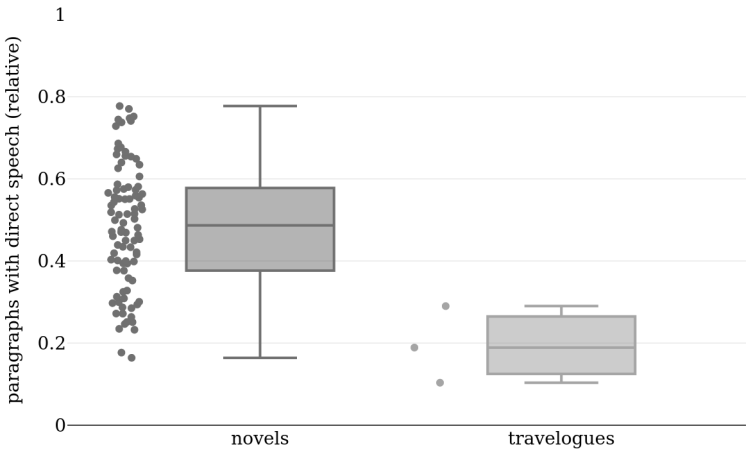


Figure 2. Proportion of paragraphs containing direct speech, travelogues versus novels.

of direct speech in the three travelogues is less than in 75 % of the novels, so even if there are novels with an equal proportion of paragraphs containing direct speech, they do not represent the typical novel.

To conclude, even though the three travelogues resemble novels in some aspects (narrative, subjective, and probably also fictional passages), they also share essential characteristics with factual travel narratives, and there are no indications that they were intended and read as fictional texts in their time. As a consequence, they were excluded from the bibliography and the corpus, even though they exhibit a certain generic ambiguity.[120]

In contrast to the examples that were discussed in detail above, in the majority of cases, the fictional status of the texts that were candidates for the bibliography and the text corpus could be determined easily based on paratextual information, bibliographical and literary-historical sources. In the unclear cases, a reasoned decision was made, as exemplified above, whereby textual and paratextual information was preferred over critical discussions as far as possible.

---

travelogues.html. Accessed January 24, 2020. The group of novels that the travelogues were compared to is a subset of the whole corpus consisting of 92 novels in which direct speech has been marked up.

[120] The presence of narrative, subjective, and fictitious elements in travelogues has a long tradition in Spanish-American writings, going back to some chronicles of the Conquista (Anderson Imbert 1995, 17–48). This contributes to the literary character of these works but does not justify considering them plain fiction. The generic ambiguity of the three travel narratives discussed here is also evidenced by their inclusion or exclusion in other text collections and bibliographies. In the "Biblioteca Virtual Miguel de Cervantes", for example, all three texts are part of the portal "Novela hispanoamericana del siglo XIX". While "La tierra natal" and "Mis montañas" are also classified as novels in the general keyword system of the virtual library, "Una excursión a los indios ranqueles" is not. It is labeled with "Descripciones y viajes" (Gorriti [1889] 2001; González [1905] 2001; Mansilla [1870] 2001). In the bibliography of the Argentine novel authored by Lichtblau, "Una excursión a los indios ranqueles" is included as a borderline case, while "La tierra natal" and "Mis montañas" are not mentioned (Lichtblau 1997).

3.1.1.2 Narrativity

According to Weber, narration is "[1] adressierte, [2] serielle, [3] entfaltete berichtende Rede [4] mit zwei Orientierungszentren [5] über nicht-aktuelle (meist: vergangene), [2] zeitlich bestimmte Sachverhalte (besonders: Ereignisse in zeitlicher Folge) [6] von seiten eines Außenstehenden" (Weber 1998, 63; cited by Zymner 2017, 365).[121] The various elements of this definition will be briefly explained here. While Weber's definition also holds for oral narration, it will only be applied to written narration in this context.

1. That narration is addressed means that there is someone (a narrator) narrating and addressing someone else (a reader).
2. It is the serial exposition of chronologically determined circumstances, facts, or events, which means that something is told in a specific order, which does not have to correspond to the underlying chronological order of the events. This may typically be an exposition, followed by the complication of events and a subsequent clarifying conclusion, but can, of course, also take other forms. As to the underlying circumstances or events, a narrow definition of narration presupposes that there are at least two propositions that involve development or shift. Narration can then be understood as the representation of a situational change. A broader definition would also include a series of discrete propositions, which do not necessarily have to be connected in the form of succession.
3. "Entfaltete berichtende Rede" refers to the relationship between a report and a narration. The latter can be conceived as a detailed, stylistically evolved report.
4. A narration is centered on two points of orientation because, on the one hand, there are the narrated circumstances, facts, or events that did already take place and the people involved in them (the first system of orientation). On the other hand, there is the moment of reporting the events including the presence of the narrator (the second system of orientation).
5. Usually, what is narrated is past, at least from the point of view of the narrator. Because it is also possible that something imagined is narrated, either imagined as past or as possible in the future, it is more appropriate to say that what is narrated is not present.
6. A narration is presented by someone external and distanced who can report the events without the necessity to stick to their immediacy, succession, chronology, or unity (Weber 1998, 11–63).

Although this definition was very useful for the decision to include texts into or exclude them from the corpus, provided that they were, in principle, available, its usefulness for the selection of entries for the bibliography was limited in the same way as for fictionality. Where editions of the texts could not be accessed, it was necessary to rely only on available metadata and on third-party information. In terms of metadata, mentions of narrative genres in book titles and subtitles or in titles of book series were especially helpful. Regarding third-party information, it had to be taken into account how narrativity was defined in each context (if it was defined at all). For example, Lichtblau discusses the selection criteria for his bibliography as follows:

> The problem of identifying those works that clearly belong in the classification 'novela argentina' beset me at every stage in the preparation of this bibliography. But I have

[121] The following clarifications of the definition of narration are a summary of Weber's more detailed explanations.

> attempted, within a certain arbitraryness inherent in all literary categorization, to be consistent in the selection or omission of the works cited. […] In addition, I have included a few celebrated works of Argentina literature that, although not novels, retain many of the characteristics of that genre and are associated with its development and artistic expression. We may thus say that Echeverría's *El matadero*, Cané's *Juvenilia*, and Mansilla's *Una excursión a los indios ranqueles* have been recruited for this bibliography without having the proper credentials as 'novel'. I did leave out, however, Sarmiento's *Facundo*, not wishing to stretch the point too much. (Lichtblau 1997, XV–XVI)

He does not provide an explicit definition of the novel and does not refer to the concept of narrativity. His criteria could only be inferred from the examples that he mentions.[122] Therefore, wherever full texts were available, the information obtained from other bibliographies was checked before a work was included in the current bibliography. An example of a text that is included in Lichtblau's bibliography (Lichtblau 1997, 309), but excluded here, is "La flor de las tumbas" (1866, AR), written by Santiago Estrada because the text has the form of a dramatic text instead of a narrative text. It starts with a cast list, is divided into acts and scenes, contains stage directions, and consists entirely of character speech. This does not fulfill the criteria established by Weber, especially that a narration should be addressed, reported by someone external, not be immediate, and have two centers of orientation. In the preface, the author explains how he conceived his work generically:

> Este trabajo no es un *drama* en la acepción literaria de la palabra. Moriría en el teatro, para el cual no está dedicado. El artista puede revestir sus concepciones en la forma que mejor se avenga a su expresión espontánea.—Este trabajo es un romance. Dibujar los cuadros o pintarlos, eso queda al arbitrio del artista. ¿Quién me obligaría a prestarle el empaste de la *narración*?
>
> ¿Puedo esperar que una lágrima escapada del alma del lector, le de el colorido que yo le niego, dejándolo en la simplicidad elemental de sus líneas?… No lo sé.—Escribo para sentir, y nada más.
>
> Su forma no carece de precedentes. Sin traer a recuerdo magistrales producciones literarias, que tomando la división y sencillez del drama, no han aspirado a la exhibición viva de la escena, citaré solamente los conocidos romances que un poeta francés ha llamado: *comedias de sillón*,—y las que el marqués de Varennes ha denominado: *proverbios*.
>
> Esto por lo que respecta a la forma. (Estrada 1866, 5)

Estrada thus says that his work is not a drama because it is not intended to be presented on stage. Instead, he calls it "romance". However, he also clearly says that it does not have the form of a narration. It is kept "simple" and "rudimentary", without coloring, drawn, but not painted, which a narration in the sense of a detailed, stylistically evolved report would be.

In general, however, it was easier to determine the narrativity of the texts eligible for the bibliography and the text corpus than their fictionality. As to the borderline cases for fictionality,

[122] He is explicit in two aspects, though: What he considers an Argentine novel and how he distinguishes novels and short novels from short stories (Lichtblau 1997, XV–XVI).

the historical biographies and the travelogues are, for the most part, narrative. Only Sarmiento's "Vida de Juan Facundo Quiroga" is not predominantly narrative, but it would still have to be discussed how much narrativity a text needs in order to be interpreted as a narration. As Weber states, when he elaborates his definition further, normally, a narration does not consist entirely of narrative text. It can also contain other forms of presentation, for example, the report of direct speech, descriptions, argumentative passages, or comments (Weber 1998, 64–70). An example of a text containing scenic presentation is the historical novel "La loca de la guardia" (1896, AR), written by Vicente Fidel López. In chapter 40, the conversation between a judge and an accused person in a trial has the form of dramatic speech. Nevertheless, this passage amounts only to about 5,300 words, and the entire novel has a length of approximately 97,500 words, so it can still be considered a narrative text.

3.1.1.3 Prose

"Prose" can be defined as a form of text that is metrically not bound, as opposed to text in verse form (see, for instance, Kleinschmidt 2003, 168). This criterion concerns primarily the distinction between narrative prose and poetry. Many of the Spanish-American novels in the nineteenth century contain inserted poems. They may be quotations at the beginning of individual chapters or part of the narration, for example, if they are recited in public by a character or are part of a love letter that is represented in the text. In general, these insertions only make up a small part of the entire text and do not question that a work is written predominantly in prose. As for the selection of texts for the bibliography, caution is required when works carry the generic label "romance" or "leyenda" because they can either be novels written in prose (for example, "El romance de un médico" (1905, AR) by Cupertino del Campo and "Un santuario en el desierto. Leyenda original" (1890, MX) by José Francisco Sotomayor) or epic texts written in verse (e.g., "Perfiles de la conquista. Romance histórico. 1521–1887" (1887, MX) by Juan Antonio Mateos and "Un ángel desterrado del cielo. Leyenda religiosa" (1855, MX) by Niceto de Zamacois). The latter were excluded from both the bibliography and the corpus.[123] There are also many texts without generic labels, which can be of any genre (novels, collections of short stories or poems, plays, other types of literary or non-literary texts) and be written in prose or verse. In these cases, the recourse to existing bibliographies of the novel and to library catalogs that include information about the genre was indispensable to finding the relevant texts.

3.1.1.4 Length

The length of the text is one of the criteria that serve to distinguish the novel from other forms of fictional narration in prose, especially shorter ones such as the novella and the short story. However, usually, these genres are also differentiated according to other criteria because there may be exceptions, for example, very short novels and very long novellas, so that a novella

[123] Sometimes novels have also been versified by other authors, for instance, the novels of the Argentine Eduardo Gutiérrez, which have been reworked by Bartolomé R. Aprile, Silverio Manco, and Apolinario Sierra. See, for example, Gutiérrez and Aprile ([1944] 2015), Gutiérrez and Manco ([1948] 2015), and Gutiérrez and Sierra ([1944] 2015).

might be longer than a novel in individual cases. Moreover, there is no consensus on the exact or approximate lower boundary of the length of a novel. Traditionally, the length of a fictional narration is expressed in page numbers which can only be a rough indicator because of differences in book format, layout, and typography from one edition to another.[124] It is more precise to measure the length of a text independently of the design of a print edition, for example, in the number of words or characters, but this is only feasible for texts which are available in electronic form and machine-readable.

In "Aspects of the novel", a collection of literary lectures about the English language novel held in 1927, Forster claims: "Any ficticious prose work over 50,000 words will be a novel for the purposes of these lectures" (Forster 1927, 17), but without motivating the number. In the context of a German handbook on literary genres, Fludernik mentions the following page limits: She sets an upper limit of 40 to 50 pages for the short story and the novella and a lower limit of 80 pages for the novel, leaving a corridor of about 30 pages for unclear cases (Fludernik 2009, 632). Unfortunately, she also does not explain how she arrives at these numbers. A more detailed discussion about the extension of the short story, novella, and novel can be found in "La novela corta mexicana en el siglo XIX" by Mata, who is looking for pragmatic criteria allowing him to define the scope of his object of study. He points out that every proposal of an exact number can, at best, apply to a specific historical context but not to the novel in general. As to Forster's suggestion, Mata states that the number of 50,000 words seems appropriate for the typical, extensive novels of the nineteenth century but not for many of the paradigmatic novels of the twentieth century, which are shorter (Mata 1999, 16). It should be added that also the geographical and the cultural context determine the characteristics of a historical genre. In the nineteenth century, the novel had a longer tradition in Europe than in Spanish America and was more stabilized as a genre (Fludernik 2009, 638–645),[125] so it can be assumed that more works complied with the established model of the time. The range of the texts considered novels in the nineteenth century in Spanish America was broad. In the early century, many of the novelistic narrative texts in prose were quite short,[126] while European models – extensive historical, realist, and naturalistic novels – gained more ground towards the middle and end of the century.[127] Towards the turn of the century and in the twentieth century, many novels were shorter again, in correspondence, interrelation, confrontation, and also independence from European developments.[128] Using the limit set by Forster, many texts that can be assigned to

[124] Especially editions with a layout in two columns lead to more words per page than single-column layouts. Some of Eduardo Gutiérrez' novels were published with a two-column layout. See, for instance, Gutiérrez ([1893] 2016); Gutiérrez ([1880] 2016a); Gutiérrez ([1880] 2016b).

[125] Fludernik traces the history of the novel from early modern precursors up to the twentieth century and states that the novel is a European genre which spread internationally in particular in the nineteenth and twentieth centuries. Nevertheless, the novel as a genre was not unknown in the Spanish-American colonies in earlier centuries. The circulation and reception of European novels in the colonies and the existence of precursors of the Spanish-American novel are set out, for example, in Sánchez (1953, 67–127). See also Lindstrom (2004, 47–77).

[126] For example, the short novels written by José Joaquín Pesado, Ignacio Rodríguez Galván, Ramón de Palma y Romay, Félix Tanco y Bosmeniel, and Juana Manuela Gorriti.

[127] For instance, the historical novels of Ireneo Paz and Juan Antonio Mateos, the realist novels of Carlos María Ocantos, and the naturalistic novels of Eugenio Cambaceres and Federico Gamboa.

[128] Like the modernist novels written by Amado Nervo and Efrén Rebolledo, for example.

the genre *novela* would be excluded from analysis. The strategy followed by Mata is to consult calls for literary competitions to see which limits they pose for the length of texts belonging to different narrative genres. On that basis, he arrives at the following numbers: a maximum of 5,000 words for short stories, a minimum of 5,000 words and a maximum of 35,000 words for short novels, and more than 35,000 words for novels (Mata 1999, 16–17). Despite his remark on the historicity of genre lengths, Mata relies on modern literary competitions in order to establish the length of novellas or short novels in the nineteenth century, which he analyses. It can only be speculated why he did not use information about literary competitions in the nineteenth century – maybe because of the scarcity of sources?

An important question is whether it would be more appropriate to distinguish the novel from other, shorter forms of narrative prose not on the basis of text length but using structural and content-related criteria. Usually, the novel is described as a complex form of narration, while the shorter text types are characterized as simpler, single-stranded forms. According to general definitions, the novella, for example, is said to present an exemplary story with one central event, with a closed structure and only a minor elaboration of the characters' life. The short story is characterized by a relative unity of place, time, and plot. The latter is usually limited to the representation of single events and has an abrupt ending. The characters tend to be typified. In the novel, in contrast, several parallel storylines and subplots, changes of place and time, and fully elaborated characterizations are more common. These structural and content-related aspects are, of course, also induced by the extent of the form (Fludernik 2009, 632; Strube 1993, 21; Zymner 2017, 371–380). Ultimately, the complex interplay of the different factors would have to be taken into account to determine to which genre a narrative prose text belongs because none of the criteria is in itself sufficient. The use of general generic definitions is problematic, though, because they do not take into account the cultural and historical context.

It is questionable whether the novella, for example, was a common genre in literary production in Spanish America in the nineteenth century at all, and even if it was, it is doubtful whether the above-mentioned characteristics would have applied. While novels and short stories can often be distinguished based on the works' subtitles ("novela" versus "cuento")[129], there is no distinctive term for short novels in Spanish. They are often called "novela", as well, and sometimes "novelita" or "novela corta" (Mata 1999, 32–33).[130] Many short novels were produced in Argentina, Mexico, and Cuba in the nineteenth century. Some were published independently in book form[131], some as part of collections of several shorter narrative texts[132] and the majority in journals (Mata

[129] There are exceptions here, too. "La loca de la guardia" (1896, AR) by Vicente Fidel López has the subtitle "Cuento histórico" and is a work published as a book with almost 500 pages. There are also some texts of intermediate length which are called "cuento", for example, "María, la perla de la Diaria. Cuento cubano" (1866, CU) by Rafael Otero, published independently with 119 pages, and "El hogar en la pampa (Cuento)" (1866, AR) by Santiago Estrada with 133 pages. They can all be considered novels.

[130] Mata mentions a whole range of denominations for the short novel in nineteenth-century Mexico.

[131] For example, "María del Consuelo. Novela" (1894, MX) by Alberto Leduc with 39 pages, "Comunidad de nombres y apellidos. Novela original" (1845, CU), and "Teresa. Novela original" (1839, CU) by Cirilo Villaverde with 63 and 93 pages, or "Un ángel y un demonio, o el valor de un juramento (Novela original)" (1857, AR) by Margarita Ochagavia with 104 pages.

[132] The "Panoramas de la vida. Colección de novelas, fantasías, leyendas y descripciones americanas" (1876, AR) by Juana Manuela Gorriti, "Tardes nubladas. Colección de novelas" (1871, MX) by Manuel Payno, and "Mesa revuelta.

1999, 29; Molina 2011, 58–59). In his account of the nineteenth-century short novel in Mexico, Mata states that short novels were among the first kind of narrative texts which were published a lot in journals shortly after the country's independence. He characterizes them as generally not having much literary value and not having been designated with the term "novela corta", which was practically unknown in the early nineteenth century. Many of the terms that were used in the titles of the texts point to the preliminary character of the works: "pequeña novela", "esbozo de novela", "proyecto de novela", "esquema de novela", "tentativa de novela", "ensayo de novela", "apuntes para una novela", etc. (Mata 1999, 32–33). Mata relates these titles, as well as the fact that many shorter novels were simply called "novela", to the problem of the missing term for the intermediate narrative genre, which on the other hand, already existed in other languages. According to him, the term "novela corta" only became common in the Iberian Peninsula and Mexico towards the end of the nineteenth and the beginning of the twentieth century, an observation which can be confirmed by analyzing the works consulted for the bibliographic database (Mata 1999, 33).[133] Towards the end of the century, short novels gained prestige, especially in the context of the *Modernismo* current (Mata 1999, 143). Mata argues that all these texts of intermediate length should be treated as "novelas cortas", understood as a genre between the short story and the novel, which existed from the early nineteenth century on but has been neglected by literary critics and historians (Mata 1999, 139). When defining this short novel in the first chapter of his book, he refers to Walter Pabst's study "Novellentheorie und Novellendichtung", an account of the origins of the European novella in Romance languages (Mata 1999, 11–12). From a taxonomic perspective, this may make sense, as all of these narrative texts are of intermediate length, but if genre is understood as an historico-cultural phenomenon, it would have to be analyzed if there is a direct relation between the early "novelitas" and the European novellas at all. Mata's argument that the early short novels were the protagonist of the initial period of the Mexican (national) narrative (Mata 1999, 141) – in their capacity as first attempts towards the genre "novela", fostered and popularized by the press – seems more likely. Nevertheless, it would have to be examined in detail to what extent authors, readers, editors, and critics of the time understood the early short novels as representatives of the genre novella. For the later short novels, this link would equally have to be discussed, although there is certainly more awareness for the "novela corta" because the term is used more often. Even so, novels,

---

Colección de artículos de amena literatura, opúsculos, juicios críticos, historietas, novelas, folletines, revistas viejas y otras muchas cosas" (1860, CU) by Francisco Calcagno, for instance.

[133] Assessing materials for the bibliographic database, the following works and collections with the subtitle "novela(s) corta(s)" were found: "La manigua sentimental. Novela corta" (1910, CU) by Jesús Castellanos, "Otras vidas. Novelas cortas" (1909, MX) by Amado Nervo, "Gil Luna, artista. Novelas cortas" (1908, CU) by Luis Rodríguez Émbil, "El enemigo. Novela corta" (1908, MX) by Efrén Rebolledo, "Thespis (Novelas cortas y cuentos)" (1907, AR) by Carlos Octavio Bunge, "Voces perdidas (Novelas cortas y cuentos)" (c. 1907, AR) by Jorge Lavalle Cobo, "Sucesos y novelas cortas" (1903, MX) by José López-Portillo y Rojas, "Novelas cortas de varios autores" (1901, MX) – a compilation of earlier short novels – and "La capilla de los álamos. Colección de novelas cortas" (1892, MX) by Manuel Covarrubias y Acevedo. Furthermore, there were volumes of collected works entitled "novelas cortas": "Obras del Sr. D. J. María Roa Bárcena. Novelas cortas" (1910, MX), "Obras de Don Florencio M. del Castillo. Novelas cortas" (1902, MX), "Obras de Don Manuel Payno. Novelas cortas" (1901, MX), "Obras del Lic. D. J. López-Portillo y Rojas. Novelas cortas" (1900, MX). They were all published in the late nineteenth and early twentieth century. An exception is the earlier collection "Horas de tristeza. Novelas cortas" (1849, MX) by Florencio M. del Castillo.

in general, tended to be shorter again, making it difficult to differentiate between "novela" and "novela corta".[134]

To conclude, the short novel is not easily recognizable as an independent genre with a certain coherence in Argentina, Cuba, and Mexico in the nineteenth century. Furthermore, there are reasons to consider many of the shorter novels as novels, as well.[135] Therefore, in this dissertation, neither the lower limits for the novel set by Forster (50,000 words) nor by Mata (35,000 words) are used. Instead, an own limit of words was deduced from bibliographic descriptions of novels, taking into account the extent of the texts in conjunction with historical subgenre labels in order to approximate the minimum and the typical length of a novel for contemporary authors and editors. Of course, not all the novels were labeled as such, but the majority were, which makes it possible to arrive at a better understanding of the extent of the texts belonging to the genre in their time. The term "novela" is understood as designating novels, not novellas, despite exceptional cases where it is clearly used in the latter sense.[136] Works with the subtitle "novela corta" or "novelita" were excluded from the calculation.

In principle, it would have been possible to also use structural and content-related criteria to select texts for the corpus, but this would not have been very efficient because an application of these criteria would have presupposed either access to detailed summaries of the texts or a close-reading of all the texts. To be able to decide upon the inclusion of texts into the bibliography, again, either detailed summaries or the full texts of all eligible works would have had to be accessible, which was not the case. Furthermore, the use of structural and content-related criteria would have presupposed established definitions of the various narrative genres, which, especially for the Spanish-American short novel, are not available. The extent of the text, in contrast, is usually part of bibliographic descriptions of the works and is a piece of information that is easy

---

[134] There are other approaches to the Mexican short novel besides Mata's. In particular, the portal "La novela corta. Una biblioteca virtual" has been developed by a research project hosted at the Universidad Nacional Autónoma de México (2008–2023). The portal is accompanied by critical approaches to the short novel published in five volumes, among them Chaves (2011). Like Mata, Chaves (115–119) links the history of the short novel in Mexico to European traditions (German, French, and English). In a compilation of Mexican romantic short novels, Ruedas de la Serna concludes: "Sin embargo su interés radica precisamente en que fueron los primeros ensayos narrativos de nuestros escritores en que surge una clara conciencia de la expresión literaria. Cierto que estos avanzaban penosamente en el dominio de esta nueva técnica de representación de la realidad, de la que, como de tantas otras cosas, se nos había privado. Cuánto, sin embargo, no habrían contribuido estas obritas en la batalla de nuestros intelectuales del siglo pasado por transformar su sociedad, y cuánto no deben a estas primicias los novelistas posteriores" (Cárabes and Ruedas de la Serna 1998, 71–72), thus evaluating the early short novels as first narrative attempts, the view preferred here. For Argentina and Cuba, no comprehensive studies of the short novel in the nineteenth century could be found.

[135] In accounts of the nineteenth-century Spanish-American novel, shorter novels are often included, especially the early ones published in the 1830's and onwards, which are mentioned as first novels of their kind, for example, "Netzula" (1837, MX) by José María Lacunza as the first indianist novel (Brushwood 1966, 71; Sánchez 1953, 546; Varela Jácome [1982] 2000, 4). Suárez-Murias mentions various "novelitas" when tracing the development of the Cuban romantic novel (Suárez-Murias 1963, 22–27). Molina includes several short novels in her book about the early Argentine nineteenth-century novel (Molina 2011, 405–489).

[136] For example, in the title "Fru Jenny. Seis novelas danesas" (1915, AR) by Carlos María Ocantos, a cycle of six novellas that share the same geographic setting and are published together in one volume. This work is out of the scope of this dissertation anyway because of its publication after 1910.

to access. It is therefore used as a proxy here to distinguish between novels and other shorter types of narrative prose texts.

The unit chosen here to measure the extent of the texts is the number of words. For each eligible text that is accessible in a full-text format of good quality,[137] this number was accessed with a simple regular expression counting all the tokens separated by non-word characters (such as white space or punctuation marks).[138] With this approach, complex linguistic structures like compounds or words with clitics are not assessed, but this is acceptable because the focus is on the comparability of text length and not on the linguistic characteristics of the texts. For the entries in the bibliography, the number of pages was used and converted to an estimated number of words. One hundred pages were selected randomly from 50 different nineteenth-century Spanish-American novels to identify an average number of words per page and to balance out differences in layout, typesetting, and font. The words on these pages were then counted.[139] Figure 3 shows the distribution of the number of words per page for the random sample.[140] The number of words per page ranges from 50 to 475, with a median of 191 words. In the following, this median is used to estimate the number of words of a text with a known number of pages.

To examine the range of lengths of nineteenth-century Spanish-American novels, 129 full texts and 252 bibliographic entries of works carrying the label "novela" either directly in the title or subtitle or in the title or subtitle of a series to which the work belongs were analyzed.[141] In

---

[137] The quality of texts that have been digitized using OCR without being corrected afterward is usually not sufficient. See, for example, the full-text versions of texts uploaded to the Internet Archive (e.g., Ramírez [1868] 2008).

[138] `tokens = re.split(r"W+", text, flags=re.MULTILINE)`

[139] If a page turned out to be a blank page or a page containing an index or only an image, it was replaced by the next (or preceding) regular text page because these are considered exceptional page types. On the other hand, chapter beginnings and endings with fewer words than full text pages were kept because they appear regularly in the novels to a certain extent. Parts of words occurring at the beginning or the end of a page were counted as whole words.

[140] The script used to generate the list of random pages and the box plot are available at https://github.com/cligs/scripts-nh/blob/master/corpus/words-per-page.py. Lists of the random pages and the corresponding editions of novels can be accessed at https://github.com/cligs/data-nh/blob/master/corpus/random-pages.csv and https://github.com/cligs/data-nh/blob/master/corpus/pages-novels.xml, respectively. The full texts of the selected pages are collected in the file https://github.com/cligs/data-nh/blob/master/corpus/pages-text.xml, and the resulting box plot file can be found at https://github.com/cligs/data-nh/blob/master/corpus/words-per-page.html. Accessed January 25, 2020.

[141] The data that was used here were preliminary corpus and bibliography, which were successively refined. The works chosen were written by Argentine, Cuban, and Mexican authors (see chapter 3.1.2 for an explanation of how the country to which an author belongs was determined) and published between 1830 and 1910 (see chapter 3.1.3 explaining the chronological limits used here). See chapters 3.2.1 and 3.3.1 for details about the sources of the full-text corpus and the bibliographic database, respectively. All bibliographic references of novels missing page numbers were left out. Furthermore, works contained in the following collections were not considered for the calculation of the typical length of a novel: "Mesa revuelta. Colección de artículos de amena literatura, opúsculos, juicios críticos, historietas, novelas, folletines, revistas viejas y otras muchas cosas" (1860, CU) by Francisco Calcagno, "Leyendas, novelas y artículos literarios" (1877, CU) by Gertrudis Gómez de Avellaneda, and "Panoramas de la vida. Colección de novelas, fantasías, leyendas y descripciones" (1876, AR) by Juana Manuela Gorriti because it is unclear which of the genres mentioned in the titles apply to which texts; "Horas de tristeza. Colección de novelas" (1850, MX) by Florencio María del Castillo and "Tardes nubladas. Colección de novelas" (1871, MX) by Manuel Payno because they were published in other editions with the title "Novelas cortas"; and "Novelas en germen" (1900, MX) by Emilio Bobadilla because the texts are all shorter than 50 pages and the title "Novelas en germen" can be interpreted as designating short novels.

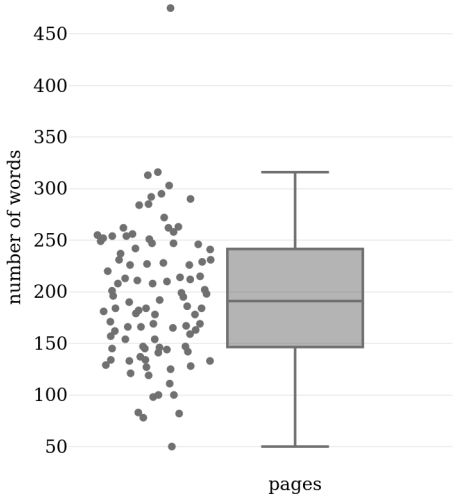


Figure 3. Number of words per page for a sample of 100 pages.

the case of the full texts, the words were counted. For the bibliographic entries, the number of pages was converted to a number of words using the median number of words per page.[142] The results for the full texts, the bibliographic entries, and both combined are displayed in figures 4, 5, and 6, respectively.[143] All the distributions have a pyramidal form which means that they are right-skewed: the higher the number of words, the fewer works carrying the label "novela" there are, or, in other words, most of the "novelas" are rather short.[144] Looking at the numbers, the shortest novel in figure 5 has 3,438 words, and the longest one 334,441, which is almost a hundred times as long, so the spectrum of lengths is very large. The median is at 44,000 words, the first quartile at 25,000 words, and the third quartile at 73,000 words.[145] With a lower limit of

[142] Obviously every work can have multiple editions. In the case of several different editions, the mean of their respective number of pages was used to balance out differences regarding the number of words per page. A work with several editions was considered eligible as a novel if at least one of the various editions published between 1830 and 1910 carried the label "novela". For the full texts, only one available edition was used to count the words. Another factor of uncertainty when using page numbers of bibliographic entries is that they usually refer to the pagination of the books and not to the number of pages of the work, so prefaces, indexes, appendices, etc. might be included, which means that the works themselves are possibly shorter than calculated here.

[143] The script that was used to select the works from the preliminary corpus and bibliography, to calculate the numbers of pages and words, and to create the box plots for figures 4, 5, 6, and 7 is available at https://github.com/cligs/scripts-nh/blob/master/corpus/words-novelas.xsl. The corresponding data and results can be viewed at https://github.com/cligs/data-nh/tree/master/corpus/words-novelas. A list of all the works that were used for the calculation of the word (and page) limit is given at https://github.com/cligs/data-nh/blob/master/corpus/words-novelas/novelas-length.csv. Accessed January 27, 2020.

[144] A right-skewed distribution is one with many low and few high values and with a mean that is higher than the median.

[145] The medians and quartiles are rounded to the nearest thousand. For the full texts alone, the median is at 61,000 words, and for the bibliographic entries alone, at 36,000 words. That the numbers are higher for the full texts is probably due to the selection of the texts: many of the digitized nineteenth-century novels available in full-text format are long novels, which tend to be considered paradigmatic. Furthermore, many historical novels were

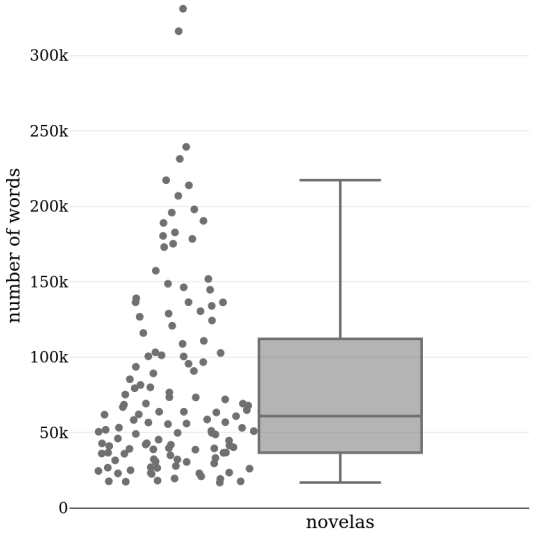


Figure 4. Number of words for the full texts of 129 works carrying the label "novela".

50,000 words as proposed by Forster, more than half of the "novelas" would be left out, and with Mata's limit of 35,000, still more than one-fourth of them would be considered short novels.

Based on these results, the question remained where to make a cut-off. It did not seem reasonable to include all the texts with the same length as the shortest "novelas", as these are only about 20 pages long, so they clearly overlap with novellas and longer short stories.[146] In these cases, a recourse to structural and content-related criteria would have been indispensable to be able to differentiate between the genres. It was helpful to look at the length of texts explicitly labeled as "novela corta" to define a lower word limit. Figure 7 shows the distribution of word lengths of 65 "novelas cortas".[147] Again, the shorter texts dominate, with a few outliers of greater length. The median for the short novels is around 7,300 words, the first quartile at 4,900, the third quartile at 10,400, and the upper fence at 16,800 words.[148]

Cutting off the "novelas" at the first decile – meaning that the shortest 10 % are left out – leads to a value of 16,000 words as a minimum[149], which is very close to the upper fence of the "novelas

chosen for the corpus, and these have a tendency to be longer than novels of other subgenres. In addition, very short novels were avoided in the collection of the full texts, but in the bibliographic entries, they were included.

[146] For example, the novels "Dos niñas hechiceras (novela original)" (1874, AR), published independently under the pseudonym "Guindilla", and "Una sanjuanina, o sea Carolina. Novela de costumbres" (1881, MX) by Guillermo Quiroga, which are both only 18 pages long.

[147] The word length for the "novelas cortas" was determined in the same way as for the "novelas". Three works were available as full texts so that their words were counted with a regular expression. For the other 62 works, the number of pages was converted to a number of words using the mean number of words per page calculated above.

[148] The values were rounded to the nearest hundred. The upper fence defines a limit between values that can still be considered typical in a distribution and those that can be considered outliers. It is set at the largest sample that is larger than the third quartile (Q3) but still lower than Q3 + 1.5 * IQR (the interquartile range, which is Q3 - Q1).

[149] Rounded from 16,044.

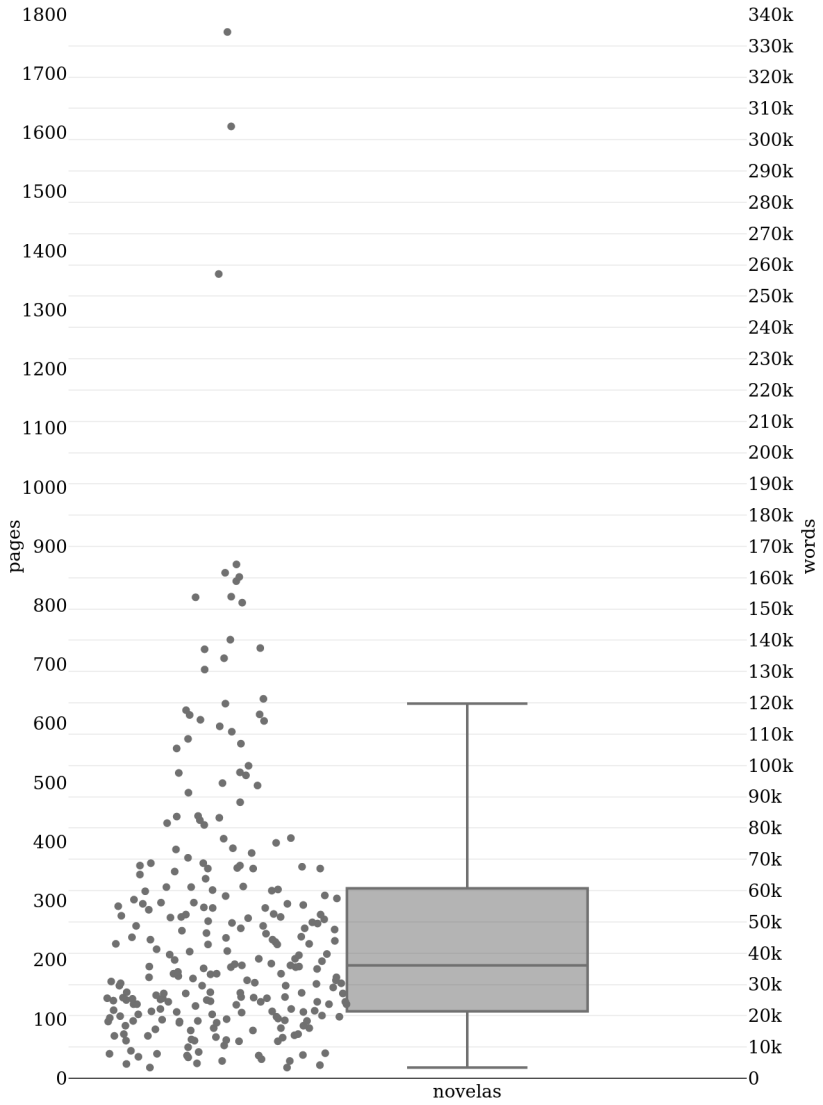


Figure 5. Number of pages and words for the bibliographic entries of 252 works carrying the label “novela”.

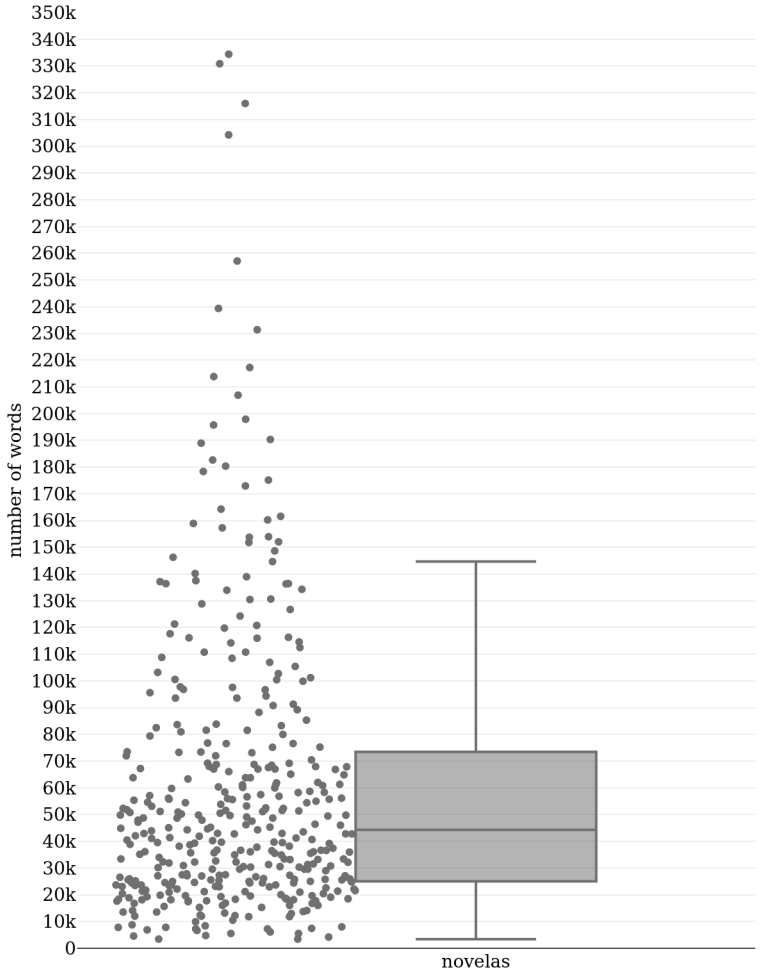


Figure 6. Number of words for 381 works carrying the label “novela”.

cortas”. That way, exceptionally long “novelas cortas” are included, while “novelas” of the same length as typical “novelas cortas” are excluded. Reformulated in page numbers, the limit amounts to 84 pages.[150] In this dissertation, the word limit was used to select texts for the corpus, and the page limit for the selection of entries for the bibliographic database.[151] To be independent of the naming conventions again, all fictional narrative texts in prose of this length were included.

Of course, this cut-off is still arbitrary to a certain extent – why should a “novela” with 15,000 words or 79 pages be excluded, but a “novela corta” with 16,000 words or 84 pages be included? It is nevertheless a limit deduced on the basis of empirical data from the same cultural-historical context as the works to be analyzed, which makes it probable that it approximates the generic

[150] Rounded from 83.8. Interestingly, this limit is very close to the 80 pages assumed by Fludernik (2009, 632).
[151] The number of words in the full texts was rounded to the next thousand before the limit was applied.

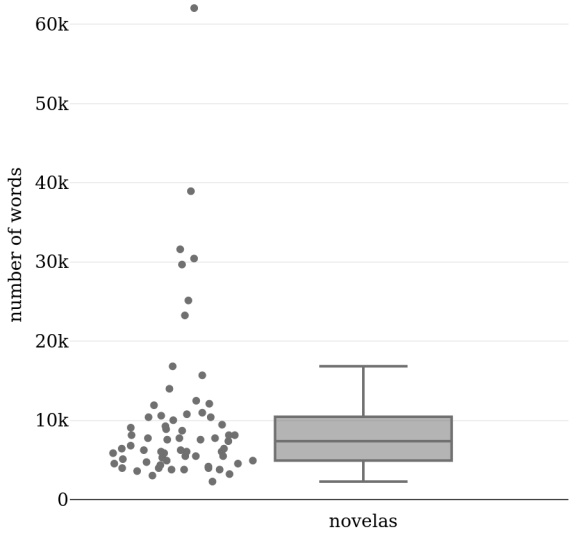


Figure 7. Number of words for 65 works carrying the label "novela corta".

conventions of the time. Furthermore, no clear cut could be seen in the data, the transition from very short to longer novels being rather fluent so that every other limit would have led to a similar arbitrary split. In addition, a numeric criterion is directly usable in a quantitative study without the need for extensive close reading.

3.1.1.5 Independent Publication

In some definitions of the novel, an independent publication as one or sometimes several books is mentioned as one of the characteristic traits of the texts belonging to the genre (Fludernik 2009, 627; Steinecke 2007, 317). However, an independent publication will not be required here in order to select texts for the bibliography and the corpus for several reasons. First, the publication of a work as one or several independent books depends to a certain extent on the length of the text. As discussed in the previous subchapter, many of the nineteenth-century Argentine, Cuban, and Mexican novels were quite short and were sometimes published in a volume together with other works, especially when the authors wrote a whole series of novels, for example, the "Entretenimientos literarios" (1843–1844, CU) by Virginia Felicia Auber de Noya or the "Episodios nacionales mexicanos" (1902–1903, MX) by Victoriano Salado Álvarez. Shorter novels were also published in collections of works of various narrative genres, such as the "Panoramas de la vida" (1876, AR) by Juana Manuela Gorriti. Second, the publication in book form corresponds to a particular model of distribution for literary works, which was not the only one in nineteenth-century Spanish America. A large part of the novels was published in journals and literary magazines, many of them in serial form.[152] Not all of these novels were also

[152] Two forms of serial publication were common: the *novela por entregas*, where parts of the novel were delivered loosely, as a booklet accompanying a newspaper, or included in a literary magazine, and the *novela de folletín*,

published in book form afterward. Whether a contemporary or modern monographic publication exists also depends on the degree of canonization of a work. As the present study aims to include as many novels as possible so as to broaden the empirical basis for the description and analysis of subgenres of nineteenth-century Spanish-American novels, no restrictions are made regarding the form of publication of a work.[153]

However, an independent publication in book form is also not just a practical matter related to text length and modes of distribution. Although the question of a novel's unity and delimitation is not easily answered by requiring it to be published independently, this still emphasizes its autonomy as a work of art. As discussed in the section on length above, very short novels published in book form existed. On the other hand, there are also novelistic works which are so long that they do not fit into one physical volume. These are often published in several books called "tomos", for example, the first book editions of "El fistol del diablo" (1859–1860, MX) by Manuel Payno with four or "Amalia" (1855, AR) by José Mármol with eight volumes. In the case of sequels and cycles published as several books, it is less obvious if each part should be considered its own novel or if they form one novel altogether. Often, the connection between the texts is indicated in titles and subtitles, as the following examples illustrate:

- "Libro extraño", "Libro extraño. Genaro. Tomo II", "Libro extraño. Don Manuel de Paloche. Tomo III", "Libro extraño. Méndez. Tomo IV", "Libro extraño. Hacia la justicia. Tomo V" (1894, 1895, 1899, 1897, 1902, AR) by Francisco Sicardi[154]
- "Dramas militares. El Chacho", "Dramas militares. Los montoneros. Continuación del Chacho", "Dramas militares. El rastreador (Continuación de Los montoneros)", "Dramas militares. La muerte de un héroe. Continuación y fin de El Chacho, Los montoneros y El rastreador" (all 1886, AR) by Eduardo Gutiérrez[155]
- "Entre dos luces" and "El candidato. Segunda parte de Entre dos luces" (1892, 1893, AR) by Carlos María Ocantos[156]
- "Las dos tragedias. Primera parte de Pepa Larrica", "La confesión de un médico. Segunda parte de Pepa Larrica", "Religión o muerte. Tercera parte de Pepa Larrica" (all 1899, AR) by Rafael Barredo

In the first case, some aspects point to the unity of the work (that the first volume has the same title as the whole cycle, "Libro extraño", and that the volumes are called "tomo" like different physical volumes of the same novel in other cases). In contrast, others emphasize the independence of the different parts (that the parts have their own title from the second volume

---

where the novel was published subsequently in specific columns of a daily newspaper (Molina 2011, 27; Villegas Cedillo 1984, 12–15).

[153] In spite of this, a bias towards the selection of the more canonized works can hardly be avoided because they are the ones that are better transmitted and more often digitized, especially as full texts. Moreover, bibliographies of the novel refer mainly to monographic publications. See chapters 3.2.1 and 3.3.1 on the selection of texts for the bibliography and the corpus for details.

[154] In this case, it is strange that the third part has a later publication date than the fourth part. There must be an earlier edition of "Don Manuel de Paloche", but this could not be verified.

[155] Gutiérrez was a very productive writer who wrote 34 novels, many of them organized in cycles. Besides the "Dramas militares", he also wrote "Dramas cómicos", "Dramas policiales", and "Dramas del terror".

[156] Ocantos wrote a whole series of 20 novels called "Novelas argentinas", published between 1888 and 1929, to which also "Entre dos luces" and "El candidato" belong (Ianes 2018, 19).

on and that they were all published, and thus probably written and finished, in different years). In the second case, all the parts have a common "supertitle", "Dramas militares", they are all published in the same year, and each sequel refers to the previous part(s). Even so, all the parts also have their individual title. In the third case, the title of the first novel does not convey any information about a superordinate work, but the subtitle of the second novel indicates that it is a sequel to the first one. These two works were published in subsequent years. In the last case, all the books are numbered parts of the common superordinate title "Pepa Larrica", and they were all published in the same year, suggesting a united work. A factor complicating the decision in all of these cases is that none of them includes the label "novela".

As a rule of thumb, a work is considered an independent novel here if it has its own title (and optionally a subtitle indicating the genre) that is not a subtitle of a part (such as "Primera parte: El prólogo de un gran libro", "Segunda parte: La víspera de un gran día", etc.), if it has its own structure starting with a first chapter and optionally ending with a trailer indicating the end of the work (e.g., "Fin", "Fin de la obra"), and if it is optionally published in one or several independent books. These parameters are easy to determine not only for texts that are eligible for the corpus but also for bibliographic entries because viewing the table of contents is enough to decide, and no close reading of the full text is needed.[157]

Following this rule, the parts of the first three cases above are all considered individual novels, while the fourth case as well as the different parts of a work published in several volumes but all carrying the same title, such as "El fistol del diablo" o "Amalia", are considered one novel. Thereby, the decision of an author (or editor) to publish a novel with its own title in an independent book is, by and large, respected. The relationship between different parts of a novelistic cycle should, however, not be ignored because it can be expected that there are similarities in content and style that influence the results of an analysis of a whole corpus of novels: it is very probable that these works are closer to each other when compared to other independent works. It can also be assumed that the degree of similarity varies according to the closeness of the parts. The books of "Libro extraño" probably have a stronger stylistic relationship than the different parts of a more extensive and looser series such as the ten novels of "La linterna mágica. Colección de pequeñas novelas / Colección de novelas de costumbres mexicanas" (published between 1871 and 1892, MX) by José Tomás de Cuéllar or the thirteen "Leyendas históricas de la independencia" (published between 1886 and 1913, MX) by Ireneo Paz. The existence of cycles and series of novels with different degrees of connectivity is another factor contributing to the great variance of the genre novel in terms of extent which also a quantitative analysis has to deal with. With the decisions made here, a short novel of around 15,000 words is compared to a novel of several hundreds of thousands of words and both to individual parts of sequels of varying length. If text length is not taken into account in the calculations, several shorter parts of a sequel have more influence on the results than a very long novel considered as one. This must be remembered when analyzing the results of the stylistic analysis.

---

[157] If content-related criteria would be considered, they could be: Is the set of main characters identical in the different parts? Is the setting the same? Is the plot a direct continuation or predecessor of another part's plot? The relationships between the various parts of "Libro extraño", for example, are discussed by Gnutzmann (1998, 183–185).

Applied to texts not published independently, the rule of thumb leads to the following decisions: a novel published in a journal, possibly in serial form, is considered one work if it has its own title and structure. Such a work is considered finished if all the existent parts are included, and if there is no obvious interruption of the structure.[158] Likewise, shorter novels included in collections are treated as individual works if they fulfill the above criteria.[159] On the other hand, collections of short stories published independently are excluded because each work contained in them has its own title and, eventually, its own structure.[160] Generally, only novels published for the first time between 1830 and 1910 are included.[161]

3.1.1.6 Additional Criteria

So far, only the very general formal criteria of fictionality, narrativity, prose, length, and form of publication were discussed to select texts for the corpus of novels. Although it is intended not to restrict the definition of the novel much further so as not to exclude texts of certain novelistic subgenres from the beginning, two additional criteria going beyond the form are discussed here. The first one refers to the target readership of the novels. In the bibliography and corpus used in this dissertation, only novels written for adults are included. There are also some novels written especially for children which were published between 1830 and 1910 in the three countries of interest here.[162] Although small in number, these are not considered because it is assumed that the target readership influences the writing style, and if they were included, children's literature would be another influencing factor that would have to be taken into account.

The second additional criterion is a realistic representation of characters and setting, which has been adduced as an important factor in the definition of the novel in order to distinguish it from epic narrative texts and romances. The latter are characterized by mythical heroes and

---

[158] That way, clearly unfinished and unpublished works are excluded (for example, "Beatriz" (MX) by Ignacio Manuel Altamirano) but works where only a self-contained part was realized and published or transmitted are included (for example, "Ambarina. Historia doméstica cubana. Tomo I" (1858, CU) by Virginia Felicia Auber de Noya).

[159] For example, "El pozo del Yocci" (1876, AR) by Juana Manuela Gorriti, which was published as part of the collection "Panoramas de la vida", or "Las ranas pidiendo rey. Confesiones de una afrancesada (1861–1862)" and "La corte de Maximiliano. Nuevas confesiones de una afrancesada (1863–1867)" (1903, MX) by Victoriano Salado Álvarez, published as parts of two different volumes of the "Episodios Nacionales Mexicanos".

[160] Usually, novels can be easily distinguished from collections of short stories regarding their structure because, typically, novels have numbered chapters. Cases that need a closer look are books containing narrative, fictional text, a title not mentioning the genre, and various parts with headings but without numbering because they could either be novels or collections of short stories. A special case in this regard is the work "Pago Chico" (1908, AR) by Roberto Payró. Originally, its parts were published individually and at different times in a literary magazine. In the monographic form, the parts are connected as numbered chapters. The work is characterized as follows by literary historians: "*Pago Chico*, loser Kranz von Erzählungen mit gemeinsamem Protagonisten, der Stadt Pago Chico" (Dill 1999, 210); "Payró, der in der Provinzstadt Bahía Blanca selbst Opfer politischer Repression geworden war, zeichnete im Verlauf seiner journalistischen Karriere das satirische Bild dieses Systems sowohl in Artikeln als auch in einer Serie von Erzählungen, die 1908 und 1928 unter den Titeln *Pago Chico* und *Nuevos cuentos de Pago Chico* in Buchform zusammengefasst wurden" (Rössner 2007, 347–348). Even though the monograph was published during the lifetime of the author and in the time frame of this study, this work is excluded from the bibliography and the corpus because it was not primarily conceived as a novel.

[161] See chapter 3.1.3 ("Limits of the Nineteenth Century") explaining the chronological limits used here.

[162] E.g., the novel "El manantial" (1908, AR) by Emma de la Barra and the series of didactic and popular scientific novels "La ciencia recreativa" (1871–1879, MX) by Alberto F. Arriaga.

vague and exotic mythical sceneries (Fludernik 2009, 628–629). This criterion does not necessarily hold for all subtypes of the novel, for example, historical, fantastic, and science fiction novels. Nonetheless, it is helpful to exclude some texts which are very far away from the prototypical realistic novel. In this dissertation, texts with non-realistic elements are included as long as these do not dominate the text and as long as the other selection criteria for novels are fulfilled. Two texts that are sometimes included in bibliographies and representations of the nineteenth-century Spanish-American novel are excluded here: "Peregrinación de Luz del Día o Viaje y aventuras de la Verdad en el Nuevo Mundo" (1871, AR) by Juan Bautista Alberdi and "Los dioses de la Pampa" (1902, AR) by Godofredo Daireaux.[163] The protagonist of "Peregrinación de Luz del Día" is the allegorical figure "Verdad" who travels to America to flee from the political and social conditions in Europe. This work has been characterized as a satire, a philosophical dialogue, a novelized allegory, or an allegorical novel (Lichtblau 1997, 16; Molina 2011, 403). It is excluded here because the protagonist is not realistic. In "Los dioses de la Pampa", Apollo and the Muses travel to Buenos Aires hoping to find the "new Athens". Disappointed because the arts are disregarded in this big city, they return to Greece. Before leaving, they only catch a glimpse of the Pampa, whose unbeknown, natural gods are presented in the main part of the book and are affiliated with the birth of the Argentine Republic. Because also this text has allegorical traits and, furthermore, no coherent plot, it is excluded, as well.

3.1.1.7 A Working Definition of the Novel

If one summarizes the selection criteria outlined in the previous sections, the following working definition of the novel can be set up for the present study:

A text is considered a novel if:

- it was conceived and received as fictional at the time and place of its publication
- and it is predominantly narrative
- and it is predominantly written in prose
- and it is at least 16,000 words or 84 pages long
- and it is published with an own title and structure, either independently, as part of a monographic collection of works, in a journal or a magazine
- and it is written for an adult readership
- and its characters and setting are predominantly realistic.

This definition of the novel is, on the one hand, general, because some of its elements (fictionality, narrativity, prose, realistic representation) correspond to characteristics mentioned in other general definitions of the novel, as well. On the other hand, it is context-specific because the length and publication criteria were derived from the pool of historical texts considered here. The adult readership criterion is one that is probably not critical in general definitions of the novel but that is included here to avoid stylistic outliers. However, as could be seen in the previous sections, even the general criteria need to be interpreted and broken down into specific paratextual and

[163] "Peregrinación de Luz del Día" is included in Lichtblau's bibliography of the Argentine novel (Lichtblau 1997, 15–16), whereas "Los dioses de la Pampa" is not, but it is labeled as a novel in the "Biblioteca Virtual Miguel de Cervantes" and in "Wikisource" (see Daireaux [1945] 2001; Daireaux 2006).

textual markers in order to be applicable to individual texts in a specific historical and cultural setting.

This definition is conceived as classificatory, which means that all the conditions should be met by a text to be considered a novel. That way, clear decisions can be made to include texts into a general corpus of novels, which in turn sets the frame for the analysis of subgenres. Inside this classificatorily defined corpus, alternative definitory concepts of (sub)genre(s) are examined.

3.1.2 Borders of Argentina, Cuba, and Mexico

This study aims to contribute to the research of subgenres of the novel in Spanish America beyond one specific regional and national context. Therefore, novels from three countries were chosen: Argentina, Cuba, and Mexico. There is a tradition of scholarship concerned with the literature of Latin America or Spanish America as a whole. Usually, "Latin America" includes the countries where the Spanish and Portuguese languages dominate[164] while "Spanish America" concentrates on the predominantly Spanish-speaking countries. Several histories of literature and research on the novel exist for these regions.[165] However, it can be discussed to what extent it makes sense to speak of "the Spanish-American novel" in the nineteenth century. In general, the literary histories and books on the subject present the nineteenth-century Spanish-American literature (and novel) as a comparison or juxtaposition of the developments in the different countries or regions of neighboring countries such as the Caribbean or Andean countries.[166] The differentiated expositions indicate that the common denominator "Spanish-American" is, above all, a retrospective label summarizing individual histories of national or regional literatures and that it does not reflect a coeval self-conception and common literary system. Indeed, literature and especially the novel, had an important function in the consolidation of the nations (Brushwood 1966; Sommer 1993). It was only towards the end of the nineteenth century, with the advent of the *Modernismo* current, that the awareness of a common literature evolved clearly:

> Zu einem der entscheidenden Merkmale des hispanoamerikanischen Modernismo wird, daß er von Anbeginn ein kontinentales Selbstverständnis entwickelt. Seit den Jahren der Unabhängigkeitskämpfe zu Beginn des 19. Jhs., als Andrés Bello in seinem Londoner Exil mit dem nie vollendeten Gedicht *América* eine eigene hispanoamerikanische Literatur

[164] In that sense, it is synonymous to "Ibero-America". In a broader understanding, "Latin America" can also include French-speaking Caribbean and South-American countries or the whole geographical region south of the United States of America (Ardao 1980, 13–27).

[165] On Latin-American literature, e.g., Dill (1999), Rössner (2007), Smith (1997), and Sommer (1993). On Spanish-American literature, see, for example, Anderson Imbert (1954), Janik (2008), and Zum Felde (1954). On the Spanish-American novel, for instance, Alegría (1959), Gálvez (1990), Goić (1980), Meléndez (1961), Meyer-Minnemann (1979), Phillips-López (1996), Sánchez (1953), Schlickers (2003), Suárez-Murias (1963), and Varela Jácome ([1982] 2000). There are also academic journals dedicated to the literature of the region, e.g., the "Anales de Literatura Hispanoamericana" and the "Cuadernos de Literatura del Caribe e Hispanoamérica".

[166] See, for example, Rössner (2007, 130–199). The literature from 1820 up to 1900 is presented in chapters on different regions: Mexico, Central America, the Caribbean, Columbia and Venezuela, the Andean countries, the Cono Sur, and Brazil. The Cono Sur designates the southern area of South America, comprising Chile, Argentina, Uruguay, and Paraguay. A book organized by country is Suárez-Murias (1963). Examples of approaches presenting the developments in each chapter (e.g., on the historical novel or the naturalistic novel) by country are Dill (1999) and Sánchez (1953).

> begründen wollte, hatte es ein solches Selbstverständnis nicht mehr gegeben. Nun trat in Hispanoamerika erneut eine Literatur auf, die beanspruchte, eine Literatur des ganzen Kontinents zu sein. Damit fügte sie sich in ein wachsendes Interesse für Iberoamerika bzw. Lateinamerika, wie es seit der Mitte des Jahrhunderts zunehmend genannt wurde, als Ganzes ein, das die kultur- und geschichtsphilosophische Diskussion des Kontinents bestimmte. (Rössner 2007, 207)

From a comparative perspective, it is nevertheless productive to analyze the subgenres of the novel in several nineteenth-century Spanish-American countries together. Even if there is no shared self-conception of literature throughout the whole century and even if there are no direct historical links in the literary communication and the formation and practice of the subgenres between all the countries and regions, there are still similar historical conditions and indirect connections triggering parallels. As Olea Franco, who examines a series of Spanish-American narrative texts from different countries from the early nineteenth up to the early twentieth century, states: "Creo que mi propia exposición, si bien discontinua, mostrará que en nuestra literatura se produce un diálogo cultural que propicia una unidad de sentido global, tanto en la generación de los textos como en su recepción crítica" (Olea Franco 2011, 25). For Olea Franco, a central aspect of the Spanish-American identity lies in the cultural and, in particular, the linguistic Spanish heritage. Through their language, narrative texts make aesthetic proposals that constitute an implicit or active reflection on cultural identity. In addition, by choosing a topic and a genre for their texts, authors propose in which cultural tradition they expect them to be read (25–26). In the context of the Spanish-American independence movements, the Creole elites had the common task of liberating themselves from the colonial heritage in their search for autonomy. A way to achieve an independent literature was to integrate modes of expression coming from the diverse American realities (28–29).[167] The choice of topics and genres also contributed to this goal, for example, the description of regional settings, customs, and types and of local and national (contemporary) historical events in the *novelas de costumbres* and the *novelas históricas*, the two subgenres most frequently mentioned explicitly in the subtitles of the novels in the three countries considered here.[168] On the other hand, the emerging Spanish-American national literatures all integrated European models (genres, topics, and also stylistic preferences) into their repertoire, so they had similar points of reference, for example, for the romantic sentimental novel, the realist, and naturalistic novels (Cárrega 1986, 49–69; Navarro 1955, 9–12; Schlickers 2003, 27–51; Varela Jácome [1982] 2000, 12). So for most of the nineteenth century, the "Spanish-American novel" can be conceived as a frame of a common colonial historical background, similar strategies to develop national novels and related literary influences until a supranational Spanish-American literature begins to emerge. The interest in comparing subgenres of the novels from different countries and regions lies in the possibility to examine the structure of trans-regional similarities and local differences and to analyze it as a pre-phase to a continental literature.

---

[167] The same arguments – a common linguistic heritage and the need to overcome a Spanish past – are invoked by Rojas Mix as part of a first cultural *Hispanoamericanismo* in the spirit of Simón Bolívar while a later, second *Hispanoamericanismo* (the one expressed by the modernist writers) involves reconciliation between Spain and the Americas in favor of a common Hispanic identity (Rojas Mix 1987, 60–64).

[168] See chapter 4.1.5 below for overviews of the subgenres in the bibliography and the corpus.

The countries Argentina, Cuba, and Mexico were chosen because, within the common frame of their colonial heritage, they represent different regions of Spanish America with different geographical and cultural backgrounds and economic, historical, and political developments, which are reflected in the novelistic production, including the different subgenres of the novel. A second reason for the choice of these countries is that their capitals already were or evolved into important cultural centers during the nineteenth century, leading to a great number of novels published there.[169] In addition, there were also novels written by Argentine, Cuban, and Mexican writers and published elsewhere.[170] In the following, the three countries are characterized briefly regarding historical and socio-economic aspects that had an effect on the number and kinds of novels written in them during the nineteenth century.

---

[169] Before deciding on the selection of the countries, several digital catalogs and libraries were checked to see if the number of novels would be enough to create a digital corpus of considerable size and suitable for quantitative analyses. The first search was performed in the WorldCat, a union catalog containing items of print and digital media alike (see OCLC 2001–2023). Searching for items published between 1830 and 1910 with the keywords "novela" and the names of Spanish-American capitals gave the following results: México (926), Buenos Aires (352), Habana (240), Bogotá (121), Santiago de Chile (120), Lima (87), La Paz (61), Montevideo (61), Caracas (46), Guatemala (36), Quito (19), Asunción (0). All searches were performed lastly on October 21, 2019. In the advanced search of the WorldCat, there is no field for the place of publication, so the place names were entered as general keywords. This leads to some false positives in the results because the keyword might also be part of a title or of a name. In this and the following searches, of the Middle American countries and capitals, only Guatemala was searched for because, in the other countries, the establishment of national literatures was thwarted by a long process of disintegration after the cease of the Viceroyalty of New Spain (Rössner 2007, 149). The second search was performed in the HathiTrust Digital Library (see HathiTrust 2008–2023). A search for catalog items including the word "novela" which were published between 1830 and 1910 in different Spanish-American capitals yielded the following numbers of results: México (178), Habana (72), Buenos Aires (66), Bogotá (31), Santiago de Chile (26), Montevideo (23), Caracas (18), Lima (15), La Paz (15), Guatemala (7), Quito (5), Asunción (0). In HathiTrust's advanced search, the language and publication year can be searched explicitly, but the place of publication cannot. It was therefore added as a general search term. In the "Biblioteca Virtual Miguel de Cervantes" (see Centro Biblioteca Virtual Miguel de Cervantes 2023), searches for "novela argentina", "novela mexicana", etc. and "Siglo 19°" resulted in: novela mexicana (49), novela argentina (42), novela colombiana (21), novela cubana (21), novela chilena (13), novela uruguaya (12), novela peruana (9), novela ecuatoriana (5), novela venezolana (3), novela boliviana (1), novela guatemalteca (1), novela paraguaya (0). A search for places of publication is not possible in the "Biblioteca Virtual Miguel de Cervantes". A search for a range of publication dates is also hardly possible. In the advanced search, there is no specific search field for the year of publication. There are several subject areas related to chronology, but they overlap (e.g., "Narrativa argentina – Siglo 19°", "Novela argentina – Siglo 19°", "Novela histórica argentina – Siglo 19°" where the latter are not necessarily contained in the former) and the search for the specification of the subject area (the part after "–") does not work. Therefore the result lists were checked manually for nineteenth-century novels. Of course, these searches only approximate the number of novels published in the different countries, but they show that Argentina, Mexico, and Cuba were comparatively rich in novels in the nineteenth century, followed by Columbia.

[170] There were several reasons for writers to publish their novels in other countries, for example, political exile or residence in another country for professional or personal reasons (especially in the case of Cuba, which was still a Spanish colony up to 1898). Numbers are available from the bibliography that was created for this dissertation. Of the Argentine works, 90 % of the editions appeared in Argentina, 6 % in Spain, 4 % in France, and 2 % in other countries. Of the Mexican works, 90 % of the editions appeared in Mexico, 9 % in Spain, 4 % in France, and 5 % in other countries. Of the Cuban works, 60 % of the editions appeared in Cuba, 28 % in Spain, 3 % in the USA, and 8 % in other countries. The sums are not exactly at 100 % because the numbers were rounded and also because some publishing houses had branches in several countries. See also chapter 4.1 on metadata analysis for details about the number of works and editions.

Argentina belonged to the Viceroyalty of Peru until 1776 when the Viceroyalty of the Río de la Plata was founded, and Buenos Aires became its capital. At that time, Buenos Aires was still a small town but strategically important because of its position at the mouth of the Río de la Plata. However, because of the lack of precious metals, the region was rather neglected and only sparsely settled. The economy remained primarily agrarian during the colonial period. Moreover, the territory belonging to the Río de la Plata region was vast and included extensive rural and unexplored areas such as the Pampa and Patagonia (Lichtblau 1959, 13–21). The contrast between the backcountry and Buenos Aires, which evolved into a big city and a political, economic, and cultural center in the course of the nineteenth century, influenced the types of novels written by Argentine writers. On the one hand, the economic and social life of the capital was a main topic in many realist and naturalistic novels written towards the end of the century. For example, the role of immigrants in the metropolitan society was discussed because, unlike in many other Spanish-American countries, Argentina's population was predominantly of a European background. On the other hand, rural life was depicted in gaucho novels (136–184, 19, 121–135). The nation's political development was also taken up in the novels. Not long after Argentina's declaration of independence in 1816 and successive disputes between unitarians and federalists about the organization of the country[171], the federalist Juan Manuel de Rosas became the governor of the province of Buenos Aires and established a dictatorial system that persisted until 1852. The Rosas era was the topic in a whole series of novels that depicted its cruelties (Molina 2011, 285–312; Lichtblau 1959, 15–16 and 43–54).

Just like Mexico, during colonial times, Cuba belonged to the viceroyalty of New Spain, which was the first administrative region that Spain established in Latin America and which existed from 1535 to 1821. However, Cuba did not become independent with the end of the viceroyalty. It remained a Spanish colony until 1898 (Kahle 1993, 55, 84–85, 95–96). This makes Cuba a special case because its literature is more closely related to the Spanish literature during the nineteenth century than that of the other independent countries. Depending on the point of view, Cuban-Spanish authors are sometimes claimed to be Spanish authors and sometimes Cuban.[172] But even before the existence of a Cuban nation-state, there was a Cuban literature, and it contributed to the emergence of a national identity.[173] The capital Havana played an important

[171] The unitarians advocated for a centralized government favoring Buenos Aires, while the federalists wanted a federation of autonomous provinces.

[172] The most prominent case is Gertrudis Gómez de Avellaneda. She was born in Puerto Príncipe in Cuba in 1814 and died in Madrid in 1873. She lived both in Cuba and in Spain, where she remained after 1840. Her novels were published partly in Cuban and in Spain, and also the settings and topics of her works cover American as well as European spheres (Instituto de Literatura y Lingüística de la Academia de Ciencias de Cuba 1999, sec. Gómez de Avellaneda, Gertrudis; Remos y Rubio 1945, 148–157). Gómez de Avellaneda is mentioned in Spanish literary-historical works (see, for instance, Neuschäfer 2001, 269; Wolfzettel 1999, 48) but is a more prominent figure in Cuban literary histories, especially because of the significance of her novel "Sab" (1841) with a Cuban theme (Mitjans [1918] 2010, 355–367; Remos y Rubio 1945, 148–152 and 227–243). There are many cases of authors who were either born or died in Cuba or Spain, changed their residence from the colony to the mother country or vice versa, and unfolded their literary activities in one or both places. How such cases are treated regarding the bibliography and corpus created here is explained further below.

[173] In a study on the Cuban novel and nation, Ferrer explains that the process of developing a Cuban national consciousness before the country's political independence is unquestioned and that it can just be debated how early this awareness matured. He maintains that already the first Cuban novelistic production between 1837 and

role in this process. The city was founded by the conquerors in the early sixteenth century and became an important trading post from early on. Important cultural institutions such as the colony's first printing press and the university of Havana were founded there in the eighteenth century (Armas 1997, 235; Zeuske 2002, 20, 28). For the formation of the novel, private literary gatherings that took place in the houses of *habaneros* from the early nineteenth century onwards were significant.[174] In addition, the Cuban literature was also brought forward by emigrated intellectuals (Armas 1997, 235). Social topics and critique were important for the Cuban novel from the beginning onwards as a means for expressing on the cultural level what was not possible on the political one. The *novela de costumbres*, describing local customs and expressing civic concerns, was a subgenre suitable to this end. A specifically Cuban topic was the problem of slavery. The economy of the country, characterized above all by sugar mills, coffee plantations, and tobacco farming, depended heavily on it. In the *novelas abolicionistas*, the system of slavery was documented critically in all of its components.[175]

When Mexico was conquered by the Spaniards, it was a region populated by many different indigenous people and dominated by the Aztecs, the *mexica*, whose capital Tenochtitlan was an urban center reflecting the power and cultural development of their civilization. Before, the Maya had had their flowering period in the southern areas of today's Mexico. The colonial era was characterized by the establishment and maintenance of an administrative system guaranteeing the Spanish hegemony over the vast territory of the viceroyalty of New Spain. This involved missionary work aimed at christianising the indigenous population and also the economic exploitation of the land, especially the mining of silver and agricultural use (Ruhl and Ibarra García 2000, 22–28, 50–55, 66–97). After Mexico's independence in 1821, the country struggled for its political consolidation, with alternating periods of opportunistic, liberal, and conservative government. Together with social and economic problems, the political difficulties culminated in the Mexican Revolution, which broke out in 1910 (Ruhl and Ibarra García 2000, 130–131). The process of political emancipation was closely related to the development of a literary self-conception, which was also reflected in the novels written in the nineteenth century, which took up the cultural, social, and political past and present. The *novela indigenista* contributed to a revaluation of Mexico's indigenous past. The historical novels served to denounce abuses of the Spanish colonial power and to highlight the merits of heroes of the independence. Furthemore, contemporary history was thematized and judged with partiality. Types and customs of the middle and lower social strata were sketched in *novelas de costumbres*. Towards the end of the century, in particular, the currents of realism and naturalism influenced the novelistic production (Rössner 2007, 140–148).

1846 played a major role in the consolidation of a coherent image of a Cuban nation (Ferrer 2018, 11–19). Sáinz de Medrano also sees a connection between the nineteenth-century Cuban novels and the development of a national consciousness: "Cuba conoció en esa centuria [el siglo XIX] un extraordinario desarrollo del relato en prosa, que parecía querer compensar la anterior penura literaria. Coincide este auge con movimientos sociales y políticos de notable intensidad, determinados en gran parte por la crisis abierta en torno a las relaciones de dependencia con España [...] y la sedimentación de una conciencia nacional. Un factor de marcada incidencia en este contexto será el problema de la esclavitud, que dará lugar a todo un ciclo de novelas" (Sáinz de Medrano 1987, 145).

[174] Especially the meetings in the house of Domingo Delmonte (1804–1853) (Suárez-Murias 1963, 20–21).

[175] On the system of slavery and the economic conditions in the nineteenth century, see Zeuske (2002, 69–89). On the novel of customs and the antislavery novel in Cuba, see Rivas (1990) and Suárez-Murias (1963, 23–24 and 25–40).

As can be seen from the above overviews, the three countries chosen for the corpus and analyses of novels here represent different political, economic, and cultural systems with local historical developments. The kinds of novels written in Argentina, Cuba, and Mexico in the nineteenth century are a result of these varying circumstances, but at the same time, they are an expression of a common cultural-linguistic colonial heritage, emancipatory concerns, and similar literary influences. The analysis of the various subgenres of the novel intends to examine how these references are reflected stylistically in the texts.

In order to select texts for the bibliography and the corpus, it is necessary to decide which novels are associated with which country. The strategy followed here is inclusive and based on two criteria: the first one is the place of publication of a novel, and the second one, the nationality of an author. If the first edition of a novel was published in one of the three selected countries, it is considered to belong to that country. That means that also novels written by authors of another nationality can be included. The place of publication of the first edition is interpreted as a sign that the author is somehow connected to that place. On the other hand, novels whose first edition is published in another country but whose author is Argentine, Cuban, or Mexican are also included. It is assumed that the birth of an author in a country entails that she or he identifies her- or himself with that country in some way. However, also authors who emigrated from another country and became Argentine, Cuban, or Mexican are considered. The content of the texts, in contrast, is not regarded as decisive.[176] With this strategy, the Argentine, Cuban, and Mexican literatures are defined geographically as well as culturally. It has the advantage that many special cases are covered, for instance, authors living in exile[177], or authors residing abroad for personal or professional reasons.[178] In addition, if the first edition of a work is published in one of the countries, it is not necessary to have full biographical information about the authors, which makes it possible to extend the bibliography and the corpus beyond the well-known canon and also to select works written by anonymous authors. Applying the criterion of nationality to Cuban authors during the country's colonial period requires an explanation. Here, authors are considered Cuban if they were born in the colony or if they spent a considerable lifetime on the island, were involved in its cultural life, and published their works there. In the latter

[176] A similar strategy is followed by Molina (2011, 395), who includes works written by Argentine authors or published in Argentina. Lichtblau considers nationality, residence, and cultural identification but does not explicitly include all novels published in the country: "The problem of identifying those works that clearly belong in the classification 'novela argentina' beset me at every stage in the preparation of this bibliography. But I have attempted, within a certain necessary arbitrariness inherent in all literary categorization, to be consistent in the selection or omission of the works cited. As used in this bibliography, an Argentine novel is understood to be any novel written by an Argentine or by a person residing in Argentina and culturally identified with that country" (Lichtblau 1997, xv). In other monographs and bibliographies, the question is not treated explicitly, e.g., in Fernández-Arias Campoamor (1952) or Torres-Rioseco (1933). In the "Diccionario de la literatura cubana", the inclusion of an author is explained in each unclear case (Instituto de Literatura y Lingüística de la Academia de Ciencias de Cuba 1999).

[177] Some Cuban authors had to leave the country because they openly opposed the colonial regime, for example, José Martí, who was a leading figure in the struggle for political and cultural independence of the country. Many Argentine writers also left the country during the time of the Rosas regime, for example, José Mármol, who published the first part of his novel "Amalia" (1855, AR) in Montevideo (Lichtblau 1959, 43; Rössner 2007, 207–208).

[178] For example, the Argentine Carlos María Ocantos, who worked in Spain as a diplomat, or the Cuban Gertrudis Gómez de Avellaneda, who moved to Spain with her family as a young woman (Cárrega 1986, 27–30; Instituto de Literatura y Lingüística de la Academia de Ciencias de Cuba 1999).

case, the decision is made for each author individually. Finally, it was decided to only treat novels written in the Spanish language and also to omit translations. Works primarily written in another language would have been difficult to process and compare stylistically to the other works. Moreover, another primary language implies that the work is, in the first place, associated with another cultural context, at least linguistically.[179]

3.1.3 Limits of the Nineteenth Century

The chronological limits of this study are set to 1830 and 1910, defining a long nineteenth century, which starts late. The lower limit marks the period of the upcoming national literatures after the wars of independence in the Argentine and Mexican cases and the beginning of the development of national conscience in the Cuban case. The 1820s were not considered because of the scarcity of novels published during that decade.[180] 1910 was chosen as the last year because it marked the beginning of the Mexican revolution, which gave rise to an own new type of novel. Furthermore, several new literary currents emerged around that date, such as the *mundonovismo*, involving a counter-movement to Modernism's cosmopolitanism and avantgardistic movements oriented towards contemporary European art movements (Janik 2008, 109–134; Meyer-Minnemann 1979, 2–4; Rössner 2007, 236–238, 263). Most Spanish-American general literary histories and histories of the novel make a caesura around this date.[181] Because the development of the novel in nineteenth-century Argentina, Cuba, and Mexico is closely related to contemporary historical events in that it was influenced by them and in that the events were, in turn, reflected in the novels, the political history between 1830 and 1910 is briefly sketched here for the three countries, following existing presentations in literary-historical works.[182]

After the end of the River Plate viceroyalty in 1810, Argentina suffered a period of internal conflicts characterized by the dispute between federalists, who favored a system of equally entitled provinces, and unitarians, who sought to establish a hegemonic position of the capital Buenos Aires. The period between 1829 and 1852 was marked by the dictatorship of the federalist Juan Manuel de Rosas, who enforced a political and economic hegemony of the province of Buenos Aires, governed by him, over the other provinces. After the end of the Rosas regime, the country had to be politically reorganized in order to overcome the conflicts between the

---

[179] For example, the novel "Pablo ou la vie dans les pampas" (1869, AR) by Eduarda Mansilla, which was published in Spanish as "Pablo o el hombre de las pampas" one year later, is excluded. In chapter 3.2.2 on the data model and text encoding it is explained how the authors' places of birth and death, their nationalities, the places of publication of the novels' editions, and the assignment of works to a country are encoded in the bibliography.

[180] Argentina's independence was declared officially on July 9, 1816, after the overthrow of the River Plate viceroyalty in 1810 (Lichtblau 1959, 15). Mexico became independent on February 24, 1821, when the catholic church and the creoles opted for a constitutional monarchy (Rössner 2007, 137). On the first cultural expressions of a beginning awareness for the own country in Cuba, see Ferrer (2018, 225–281) and Rössner (2007, 152–153). In Mexico, Fernández de Lizardi published several works before 1830, especially the novel "El Periquillo Sarniento" (1816, MX). However, these are not considered here because of their exceptional status. They are often described as forerunners of the nineteenth-century novel (Alegría 1959, 18–26; Janik 2008, 34–36; Sánchez 1953, 111, 115–123).

[181] The year 1910 is also chosen by Anderson Imbert (1954). Some set the limit a bit earlier at the turn of the century or later, e.g., in 1920 (Alegría 1959; Ertler 2002; Rössner 2007).

[182] For more detailed overviews of the political, economic, and social history, see Bernecker (1992, vol. 2: Lateinamerika von 1760 bis 1900).

provinces and to make a unified nation possible. In 1852, Argentina became a federation under the unitarian Justo José de Urquiza, with a constitution adopted in 1853. Yet Buenos Aires joined the federation only in 1860. A civil war broke out, ending in the victory of the forces of Buenos Aires under the command of Bartolomé Mitre, who became the president of the united republic in 1862. This moment initiated a phase of political and social stabilization and economic growth (Lichtblau 1959, 15–21). Between 1865 and 1870, Argentina was involved in the War of the Triple Alliance between Paraguay and the alliance of Argentina, Brazil, and Uruguay, which ended with the defeat of Paraguay. In a military campaign between 1878 and 1884 known as the "Conquista del Desierto", indigenous people were fought in the Pampa, Patagonia, and the Chaco region with the objective of securing the Argentinian-European dominance in the remote regions. In 1880, Buenos Aires was officially declared the capital of the republic, and the liberal Julio Argentino Roca was elected as president (Kahle 1993, 113–114). Liberal governments stayed in power until 1916, promoting immigration, foreign commerce, and a general economic upswing, interrupted by a severe financial crisis in 1889 and 1890 (Lichtblau 1959, 138–142).

After the wars of independence, Cuba became the most important Spanish colony. Havana was the most important city of the remaining Spanish empire, and Cuba's plantation economy satisfied the European demand for sugar, coffee, and other colonial goods. In the first decades of the nineteenth century, the Spanish crown benefited the loyal oligarchy with a reform of restoration. On the other hand, a group of intellectuals and literates advocated for the development of a Cuban national identity and criticized the system of slavery supporting the plantation economy. Furthermore, because of unstable political conditions in the mother country, a new group of annexationists emerged who envisaged the attachment of Cuba to the United States. The fear of a slave revolt was another factor leading to an approximation to the US-American southern states. In the 1840s, different ideas between loyalty, autonomy, annexation, or separation existed for the future of the country (Zeuske 2002, 90–99). In 1868, an attempt by the Cuban bourgeoisie to obtain more political and economic autonomy from Spain failed. This initiated a period of internal wars of independence, lasting until 1898 when the United States provoked the Spanish-American War and intervened in the Cuban struggle for autonomy. Cuba became independent from Spain but remained under the control of the USA. Even the Cuban constitution from 1902 did not bring about true sovereignty because it guaranteed the United States the right to intervene should their interests be at risk. In the following years, Cuba suffered several military interventions by its superior (124–162).

Like Argentina, also Mexico experienced a period of political agitation after its independence was declared in 1821. The first government was a constitutional monarchy led by Agustín de Iturbide, which was overthrown by the military under the leadership of General Antonio López de Santa Anna in 1823. In the same year, the provinces of Central America (present-day Costa Rica, El Salvador, Guatemala, Honduras, and Nicaragua) declared themselves independent from Mexico. In 1824, Mexico became a republic with a federal constitution, which was replaced by a centralistic organization introduced by conservative forces in 1835. Subsequently, several provinces strove for autonomy, among them the English-speaking colonists in Texas. After the Mexican-American war from 1846 to 1848, Mexico lost considerable territory to the United States of America. In 1855, an era of reform began when the liberals defeated the military strongman Santa Anna, who had dominated the political events since the 1820s. It was intended to lead to economic

growth and political strength, but anticlerical and -military actions triggered the resistance of the conservatives. A civil war between 1858 and 1861, which was won by the liberals, led to further measures against the Church. Moreover, a planned moratorium on foreign debt provoked a French intervention at the end of 1861, which in turn resulted in the establishment of an empire governed by the Austrian archduke Maximilian von Habsburg. However, this monarchical system lasted only until 1867 when it was ended by the liberal troops under Benito Juárez. The presidency of Juárez marked the beginning of a period of modernization and reconstruction of the society and the economic system. It was continued by Porfirio Díaz, but his measures of domestic and foreign policy neglected the middle class and rural population, leading to social protest that culminated in the Mexican Revolution breaking out in 1910 (Rössner 2007, 137–140; Ruhl and Ibarra García 2000, 130–166).

The historical developments in the nineteenth century in Argentina, Cuba, and Mexico show that all three countries had to go through a longer period of political turbulences, economic stagnation, and social problems before a consolidation of the nations was reached. For Argentina and Mexico, relative stability was achieved from the middle of the century onwards, while a Cuban nation-state was not yet fulfilled. The respective historical circumstances affected the cultural life and, thereby, also the production of novels. When one looks at the numbers of novels included in the bibliography, connections to the historical developments in the countries can be assumed. In Argentina, the number of novels written increased moderately after 1851 and considerably after 1880, coinciding with the beginning of the liberal government of Roca. A slight decrease can be noted in the 1890s and 1900s. This might be related to the financial crisis of 1889 and 1890 but also to the prevalence of the *Modernismo* current that focused on other genres, especially poetry and short prose texts. In Mexico, the production of novels took off in the 1860s, increasing almost steadily until the 1900s. Apparently, the French intervention in the 1860s did not have a negative impact on the publication of novels, and the presidencies of Juárez and Díaz provided conditions that were favorable for it. The development of the number of Cuban novels is not that clear. Most novels were published in the 1850s. Beyond that, there are slight ups and downs, but no clear increase over time is visible, and the overall number of novels is lower than in Argentina and Mexico. This suggests that Cuba's status as a colony and the struggle for independence breaking out openly in 1868 held back the development of the novel in that country.[183] Besides influencing the number of novels published, the contemporary political-historical events and social, economic, and political issues of the time supplied thematic material for many novels and contributed to the formation and adaptation of some subgenres of the novel, for example, historical novels treating contemporary issues or the anti-slavery novel (Brushwood 1966; Lichtblau 1959, 43–54, 121–135, 138–143; Molina 2011, 285–375; Rivas 1990).

[183] The number of novels published in the three countries were calculated based on the bibliography described in chapter 3.2. How the kinds of sources of the bibliography might have influenced the numbers is discussed in chapter 3.2.1. For each novel, the decade of the first known edition was determined. The following numbers resulted: Argentina: 1830–1840 (2 novels), 1841–1850 (2), 1851–1860 (25), 1861–1870 (17), 1871–1880 (23), 1881–1890 (87), 1891–1900 (75), 1901–1910 (72); Mexico: 1830–1840 (3), 1841–1850 (5), 1851–1860 (9), 1861–1870 (55), 1871–1880 (64), 1881–1890 (83), 1891–1900 (74), 1901–1910 (102); Cuba: 1830–1840 (8), 1841–1850 (19), 1851–1860 (31), 1861–1870 (12), 1871–1880 (10), 1881–1890 (15), 1891–1900 (22), 1901–1910 (14). Charts displaying these numbers are presented in chapter 4.1.3.

After deciding upon the temporal limits of the investigation, it was necessary to develop criteria to be able to assign the novels to the chronological frame. In general, the publication date of the first known edition is decisive. Works that are listed in bibliographies of the Argentine, Cuban, and Mexican novel but for which no publication date could be verified are not considered. Novels published posthumously are taken into account as long as they were first published between 1830 and 1910. Works that are clearly unfinished are not included.[184] Two of the Cuban novels were treated in an exceptional way. Both were published much later than they were written because of their political topicality. The novel "Francisco" (1839, CU) by Anselmo Suárez y Romero was written in 1839 but only published in 1880, and "Cecilia Valdés" (1839, CU) by Cirilo Villaverde was also written in 1939 but first published in its entirety in 1882 (Rössner 2007, 156–157). It is assumed that the style of the texts is mainly characterized by their time of creation, and because they were written so much earlier than they were published, in these cases, the date of creation is taken and not the date of the first publication.

Regarding the full-text corpus, another question to consider is which editions of the novels to select. There are novels that changed considerably over time when authors reworked them for subsequent editions. For example, the novel "El fistol del diablo" (1859–1860, MX) by Manuel Payno was published in book form in four volumes, the first time from 1859–1860, again in 1871, and then as two volumes in 1877 and in 1906. It was extended several times. The first edition, for example, contains 49 chapters, and the second 86 chapters. On the other hand, most of the novels were only published once between 1830 and 1910, so cases with divergent versions of novels are the exception rather than the norm.[185] The strategy that would be most appropriate from a historical point of view would be to only include first editions, considering that also the dates of the novels are derived from their first editions. Unfortunately, the state of digitization did not allow for such a stringent methodology, and different types of editions had to be selected for the corpus.[186]

Summing up the selection criteria used for the bibliography and the corpus, it can be noted that a general definition of the novel is followed that allows for including a broad range of subgenres. On the other hand, the general criteria are strictly applied because the size of the bibliography and the corpus make it difficult to make case-by-case decisions. As a consequence, some texts that are considered novels in other contexts are excluded here, while others that are neglected elsewhere, are included because the usual canon of texts is not taken as the general basis. Novels from three countries that represent different regions of Spanish America were chosen. On the one hand, the selection of novels was made based on the place of publication, capturing the local production of literature in the countries. On the other hand, the national and cultural identity of the authors was used as a criterion. That way, the literatures of the countries are defined broadly

[184] Works are considered clearly unfinished if it is obvious from the structure or content that parts of the work are missing, for example, several chapters. On the other hand, if a series of several novels was envisaged by an author but not finished, individual novels that form part of it are still included. See also chapter 3.1.1.5 above on questions of the unit of the "novel".

[185] The data about the number of editions per novel are taken from the bibliography created for this investigation. Of the 829 novels, 70 % had only one edition that was published between 1830 and 1910, 19 % had two editions, and 10 % more than two editions. The numbers are rounded. See also chapter 4.1.4 for overviews of the data about editions contained in the bibliography and the corpus.

[186] See chapter 3.3.1 for an overview of the types of editions used.

as cultural-geographical units. The subgenres of the Argentine, Cuban, and Mexican novels are meant to be analyzed comparatively from the phase of the struggle for and the achievement of political independence up to a political and economic stabilization throughout the nineteenth and early twentieth century, involving the literary currents of Romanticism, Realism, Naturalism, and *Modernismo*. In the next sections, the creation of the bibliographical database and the corpus are described based on the selection criteria outlined so far.

3.2 Bibliographical Database

The bibliographical database, which is also called Bib-ACMé ("Bibliografía digital de novelas argentinas, cubanas y mexicanas, 1830–1910") in the following, was created with the goal of getting an overview of all the Argentine, Cuban, and Mexican novels published between 1830 and 1910.[187] The main motivation for creating the database was to have a pool from which to select novels for the digital corpus and to get a sense of the dimension of the resulting corpus when compared to the overall novelistic production of the time. Unfortunately, the goal of creating a complete bibliography cannot be reached because not all the novels were documented bibliographically, and it is very probable that many texts are not preserved anywhere in libraries, archives, or private collections, especially those not published in book form but only in journals and magazines. Nevertheless, the size of a digital full-text corpus is limited by more factors than that of a bibliographical database, so that it is still worthwhile to undertake the effort to get a picture of the field which is as complete as possible. Furthermore, in comparison with printed and digitized bibliographical works, a truly digital bibliography has the advantage that the information contained in it is programmatically analyzable. How many novels were written by which authors, and how often, when, and where were they published? How long were the novels, and to which subgenres can they be assigned? In what follows, it is explained how the bibliographical database was prepared to be able to answer these questions. In chapter 3.2.1, the sources used to collect the bibliographical entries are accounted for, and it is set out how the selection criteria for novels defined in chapter 3.1 above were applied to choose entries from the sources. Usually, bibliographic entries of literary works include several levels of information: details about authors, editors, publishers, the work itself, the time and place of its publication, etc. To be able to analyze the various information levels contained in such entries, a special data model was developed for the database to which the entries were mapped. This model and its application in the form of text encoding are presented in chapter 3.2.2. In the last part of this chapter, in 3.2.3, the assignment of subgenre labels to the works contained in the bibliographical database is described.

[187] An earlier version of the bibliographic database was presented at the conference HDH2017 at the University of Málaga (Henny-Krahmer 2017). The work on the bibliography is managed on the version control platform GitHub. The first version of BibACMé from 2017 can be accessed at https://github.com/cligs/bibacme/releases/tag/v1.0. Accessed October 30, 2019. Ongoing work is available at: https://github.com/cligs/bibacme. Accessed October 30, 2019. For an online publication of the database with background information, basic search functionality, and some synoptical charts, see Henny-Krahmer (2017–2021).

3.2.1 Sources

Three main sources were chosen for the creation of Bib-ACMé, one for each of the three countries covered: for Argentine novels, the work "The Argentine novel: an annotated bibliography" created by Myron Lichtblau was used, for Cuban novels the "Diccionario de la literatura cubana" (DLC) edited by the "Instituto de Literatura y Lingüística de la Academia de Ciencias de Cuba", and for Mexican novels the "Bibliografía de la novela mejicana" by Arturo Torres-Rioseco (Lichtblau 1997; Instituto de Literatura y Lingüística de la Academia de Ciencias de Cuba 1999; Torres-Rioseco 1933). These sources were preferred over national bibliographies for several reasons. In the case of Argentina, to date, there is no national bibliography.[188] On the website of the "Biblioteca Nacional de Cuba José Martí", the work of several bibliographers over the centuries is presented as the national bibliography (Biblioteca Nacional de Cuba José Martí 2011). Of these bibliographic endeavors, the "Bibliografía Cubana del Siglo XIX" by Carlos Manuel de Trelles, which is available for download as PDF files with images on the website of the Cuban National Library, is relevant here (Trelles 1911). However, in the eight volumes of this bibliography, works of all kinds are registered and presented by year of publication, so it would be necessary to go through all the years between 1830 and 1910 and look for novels. Although it would be desirable to evaluate Trelles' bibliography in this regard, this could not be accomplished within this dissertation. In the "Diccionario de la literatura cubana", on the other hand, primarily literary works are listed, making it much easier to find relevant novels. Furthermore, the dictionary is organized into articles about literary currents, genres, institutions, journals and magazines, and biographical entries, including bibliographic information. The biographical entries are helpful in deciding which authors can be considered Cuban writers because the authors' relation to Cuba is described.[189] For Mexico, the "Instituto de Investigaciones Bibliográficas" is responsible for the publication of the national bibliography "Bibliografía Mexicana".[190] Its digital products include the electronic catalog and search system "Bibliografía Mexicana del Siglo XIX" (Instituto de Investigaciones Bibliográficas n.d.). In order to find relevant novels, one would, for instance, have to know the authors' names beforehand and search for the works published by them or look for entries including the term "novela" in the title, which would only yield part of the results. Another possibility would be to search year by year. In comparison, it is more expedient to use Torres-Rioseco's work which focuses on the novel.[191] Furthermore, the national bibliographies usually register works published in the respective countries, but as works written by Argentine, Cuban, and Mexican authors which were published elsewhere are also included here, specialized bibliographical works which consider them as well are advantageous.[192]

Other sources were used to complement the information extracted from the main sources. Information about authors (names and life data) was gathered from the Virtual International Authority File (VIAF) (OCLC 2010–2021b). Further information about works and editions was added primarily from the following digital sources: "Biblioteca Digital Hispánica" (BDH), "Enci-

---

[188] For a discussion of the problem up to the year 2004, see Romanos de Tiratel (2004).

[189] Because Cuba was a Spanish colony until 1898, there are many authors who were born in Cuba but moved to Spain or vice-versa.

[190] For an overview of the history of and the current bibliographic work in Mexico, see Escalona Rios (2006).

[191] It is important to note that the bibliography of Torres-Rioseco builds mainly on the earlier work by Iguiniz (1926).

[192] See chapter 3.1.2 on the "Borders of Argentina, Cuba, and Mexico" above.

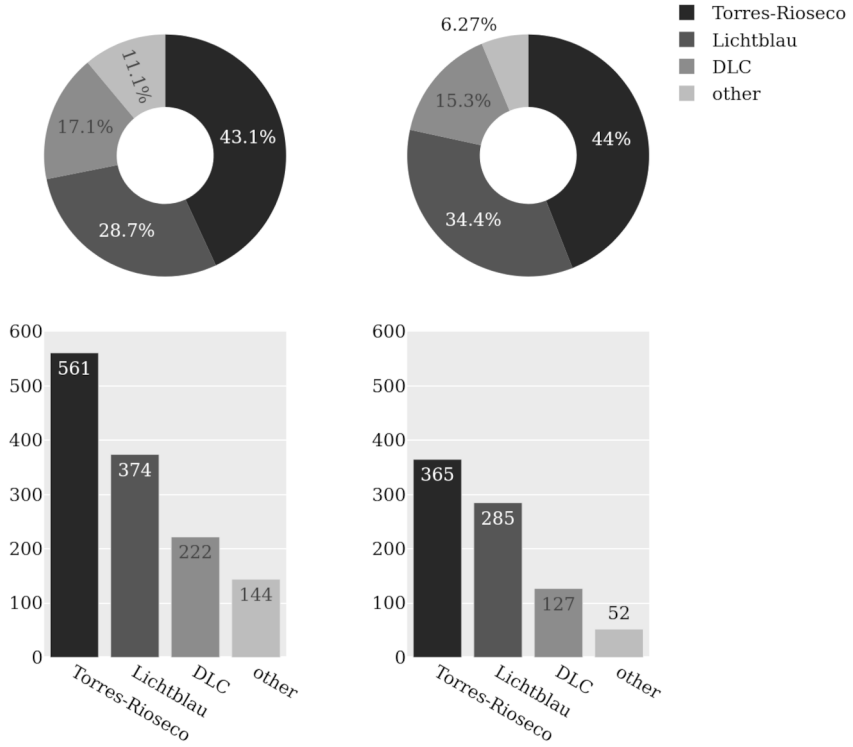


Figure 8. Works by source. Left: candidates, right: entries in the bibliography.

clopedia de la literatura en México" (elem.mx), "HathiTrust Digital Library", "Internet Archive", "Wikimedia Commons", and the "WorldCat" (Biblioteca Nacional de España 2023; Fundación para las Letras Mexicanas A.C. 2018; HathiTrust 2008–2023; Internet Archive n.d.; Wikimedia Commons 2023; OCLC 2001–2023).

By using the different sources, 1,301 candidates for novels were identified. The selection criteria defined in chapter 3.1 above were applied to the candidates, resulting in 829 works that were included in BibACMé. Figure 8 shows from which sources the works were compiled.[193] The candidates are shown on the left side, and the remaining entries of the right side. As can be seen, almost one-third of the candidates were sorted out after the application of the selection criteria. Of the three main sources, most novels come from the Mexican bibliography, and the fewest from the Cuban dictionary.

Several factors may have caused these varying amounts. First, it is probable that the number of novels published between 1830 and 1910 in Argentina, Cuba, and Mexico and by Argentine, Cuban, and Mexican writers differs per se. It may well be the case that most novels were Mexican as the country's cultural institutions were more developed than Argentina's in the early nineteenth

[193] See the script at https://github.com/cligs/scripts-nh/blob/master/corpus/bibacme-sources.py. The resulting chart "sources shares" can be downloaded at https://github.com/cligs/data-nh/tree/master/corpus/bibliography-sources. Accessed January 27, 2020.

century and that there were much lesser Cuban novels because of Cuba's colonial status until the end of the century. Other political, economic, cultural, and demographic factors may also play a role.[194] Nevertheless, it is also very likely that the kind of bibliographic sources that were used here influence this result because the DLC is a general dictionary of literature. It is not specialized in novels and does, therefore, probably not reach the same degree of comprehensiveness as the other two main sources.

The numbers of the remaining entries are, of course, also influenced by the extent to which the selection criteria led to the omission of works from the different sources. In the DLC, many novels, especially those published in journals and magazines, are mentioned in the biographic articles but not listed in the corresponding bibliographical lists. These were only integrated into Bib-ACMé when the time and place of publication could be verified, and when the length of the text could be estimated. Likewise, Lichtblau includes many novels in his bibliography that were only published in journals, but because there is usually no indication of the extent of the text, these entries were neglected. On the other hand, in Torres-Rioseco, the works listed were almost exclusively published as independent books, balancing out the differences because of missing information to a certain degree.

When deciding upon the inclusion of the bibliographic references into Bib-ACMé, the selection criteria for novels defined in chapter 3.1 above were applied as follows. It was generally assumed that the works mentioned in bibliographies of the novel are fictional, narrative texts in prose and that works carrying the label "novela" also meet these criteria. In cases of doubt, often triggered by the works' titles, digital editions[195] were checked whenever they were available. When no edition was accessible, doubtful cases were sorted out rather than included.[196] The criteria of a publication with its own title and structure, an adult readership, and predominantly realistic characters and setting were checked in a similar manner. The titles of the works were interpreted with regard to the selection criteria, and, wherever possible, the works were checked by consulting editions. Doubtful cases that could not be cleared up in this way were left aside.[197] In Lichtblau's bibliography, the entries are made on the level of editions of the individual literary work, meaning that shorter works published in a collection are listed separately. In the DLC and Torres-Rioseco's bibliography, in contrast, the entries correspond to publications and not

---

[194] See chapter 3.1.2, which touches upon the historical backgrounds of the three countries.

[195] In this context, "digital editions" does not necessarily refer to digital critical scholarly editions but also to digitized editions of all kinds (e.g., published in full text, HTML format, or as PDF or image files). Print editions of the novels were also checked, but not comprehensively, for reasons of time and cost. In part, novels could be obtained through the German interlibrary loan, especially from the "Ibero-Amerikanisches Institut" in Berlin, but many editions can only be consulted in American libraries.

[196] This applied, for example, to the work "Doce episodios de la vida de Bernabé Loyola, escritos por él mismo y dedicados a sus queridos hijos" (1876, MX) by Bernabé Loyola, which is listed in Torres-Rioseco's bibliography of the Mexican novel, but whose fictional status is unclear because the name of the author is the same as the name mentioned in the title.

[197] Two works that are mentioned in the "Bibliografía de la novela mejicana" but which were excluded here are, for example, "Narraciones humorísticas y cuentos infantiles" (1885, MX) by Manuel Covarrubias y Acevedo because it is a collection of short stories written for children, and "Staurófila. Precioso cuento alegórico. Parábola en que se simboliza los amores de Jesucristo con el alma devota" (1903, MX) by María Nestora Téllez Rendón because it is probably not predominantly realistic.

necessarily individual works, so collections are listed as one entry.[198] These were checked to extract novels contained in them. When insight into the table of content of a collection was not possible, it was disregarded.

However, most of the entries from the sources that were dropped here were excluded because of the length criterion. Whereas Lichtblau explicitly includes short novels (Lichtblau 1997, xvi), Torres-Rioseco does not explain his selection criteria regarding the extent of the texts. Although the bibliography is entitled "Bibliografía de la novela mejicana", it is rather a bibliography of fictional narrative texts of all kinds and lengths or a bibliography following a definition of the novel that is broader than the one used here. Where digital full-texts were available, the number of words was checked. Otherwise, the number of pages was decisive. The extent of the text is not always indicated in the bibliographies, and in the DLC, no page numbers are given at all. In many of these cases, the page numbers could be added through the WorldCat, but not always. It was decided to exclude novels without page numbers that were exclusively published dependently (in journals, magazines, or books). There are, of course, novels only published in a journal that are longer than 84 pages, especially serial novels, but many of the novels that were not published in book form are short novels. On the other hand, novels published independently are usually longer than 84 pages.[199] In order not to omit too many relevant works, it was decided to keep monographic works even if no page numbers were available.

As for the assignment of the novels to the three countries, only those works were excluded where the author could neither be associated with the country[200] nor the work was first published there.[201] For some bibliographic entries in the sources, the publication date was not given. When no edition of the work was found that could be dated to the period from 1830 to 1910, the work was not included in Bib-ACMé. Figure 9 summarizes how many of the candidates were kept and why the others were excluded.[202] The chart shows that only a few entries did not comply with the criteria of fictionality, narrativity, prose, an adult readership, and a realistic representation. Most had to be dropped because they were too short or because the bibliographic information was not complete enough to decide. For details about individual works, a tabular overview showing the application of the selection criteria to the entries from the bibliographic sources is available on GitHub.[203]

---

[198] Of course, a collection of short novels written and published by the same author can also be considered a literary work; an anthology of works by different authors compiled for a secondary publication would usually not be considered as such. Here, "individual literary work" refers to works in the smallest sense, i.e., individual novels.

[199] In those cases, there are also exceptions, but they are relatively few in number. See chapter 3.1.1.4 above.

[200] By nationality, or, in the case of Cuba, by birth or predominant place of activity. See chapter 3.1.2 above.

[201] This was the case for the novel "El dios del siglo. Novela original de costumbres contemporáneas" (1848, ES) by Jacinto de Salas y Quiroga. In Torres-Rioseco's bibliography it is included with an edition of 1853 published in Mexico, but the work was published first in 1848 in Madrid, and the author is Spanish.

[202] See the script at https://github.com/cligs/scripts-nh/blob/master/corpus/bibacme-sources.py. The resulting chart "sources inclusion" can be downloaded at https://github.com/cligs/data-nh/tree/master/corpus/bibliography-sources. Accessed January 27, 2020.

[203] See https://github.com/cligs/bibacme/blob/master/app/data/entries-sources.csv. Accessed January 27, 2020. For the DLC, only possible candidates are listed, but not all the bibliographic entries, because the source is not a bibliography of the Cuban novel but a general literature dictionary, so no other definition of the novel applies. For Lichtblau and Torres-Rioseco, which are general bibliographies of the novel, only entries referring to 1830–1910 and those with unclear publication dates are listed because they were candidates for Bib-ACMé. All other entries of

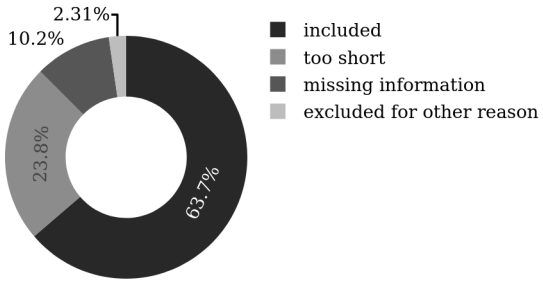


Figure 9. Inclusion and reasons for exclusion of works.

To conclude the discussion of Bib-ACMé's sources, it must be said that the contribution of this digital bibliography lies primarily in the compilation, restructuring, integration, and enrichment of existing bibliographies of nineteenth-century Argentine, Cuban, and Mexican novels. The selection criteria were applied in a way that favors a high precision, meaning that all the novels contained in the bibliography should meet the criteria of the working definition formulated in chapter 3.1.1.7 above. That way, the full-text corpus of novels can be compared to a relevant population. Other bibliographic works aim at a higher recall, including many candidates for their subject, so as to be as comprehensive as possible. Moreover, a definition of the novel different from the one advocated for here would obviously lead to a different bibliography. Furthermore, this bibliography could still be completed further using more sources.[204] In any case, the modeling and preparation of the bibliographic information in digital format enhance the usability of the data, as outlined in the next section, and facilitate future reuse also in other contexts.

3.2.2 Data Model and Text Encoding

The data model of Bib-ACMé is centered around the three notions of *author*, *work*, and *edition*. These three entities are defined in accordance with the Functional Requirements for Bibliographic Records (FRBR), a conceptual model developed by the International Federation of Library Associations and Institutions (IFLA) (2009). In FRBR, four basic entities have been defined for the products of intellectual endeavors that are described in bibliographic records: work, expression, manifestation, and item. A second group comprises entities responsible for the intellectual content: person and corporate body (International Federation of Library Associations and Institutions (IFLA) 2009, 13).[205] Of these entities, "work", "expression", "manifestation", and "person" are relevant to explain the data model of Bib-ACMé. According to the FRBR model, a "work", as opposed to an expression of a work or a manifestation of an expression, is defined as "a distinct intellectual or

works published before 1830 or after 1910 were disregarded from the beginning. In the bibliographies of Lichtblau and Torres-Rioseco, explicit or implicit definition criteria for the novel apply, which are not entirely congruent with the ones developed in this dissertation, so the table shows where decisions to include a work as a novel differ between the bibliographies and Bib-ACMé. The table also contains works that were not included in the three main sources but were added from other sources. The entries are made on the work level, not by publication or edition.

[204] Especially archival, print, and hemerographic sources.

[205] A further type of entity defined in FRBR (subjects of intellectual endeavors: concept, object, event, and place) is not relevant here.

artistic creation" and as an "abstract entity; there is no single material object one can point to as the *work*" (International Federation of Library Associations and Institutions (IFLA) 2009, 17). A work is recognizable through its individual realizations, i.e., expressions, but they are not to be identified with the work. An "expression" is thus "the intellectual or artistic realization of a work", and a "manifestation" is "the physical embodiment of an expression of a work" (International Federation of Library Associations and Institutions (IFLA) 2009, 13).[206] A "person" is responsible for the creation and the intellectual or artistic content of a work (International Federation of Library Associations and Institutions (IFLA) 2009, 25).

The idea of a work as an abstract entity is useful for this study because the goal is to analyze the novels as literary works and not as specific expressions of it. Ultimately, a full-text version of a work in the corpus is an individual expression, such as a particular edition. However, it functions as a representative which points to the work and does not stand for itself because the interest is not, for example, in the study and comparison of different expressions of the same work. Furthermore, the generic signals of the work that occur in titles and paratexts were interpreted across different editions. Genre assignments made by other literary historians are usually also not bound to a specific realization of a work.[207] In the FRBR report, it is stated that the boundary between one work and another is not easily drawn and is also culturally determined, but that the "modification of a *work* involves a significant degree of independent intellectual or artistic effort" and that, inter alia, "adaptations of a work from one literary or art form to another (e.g., dramatizations, adaptions from one medium of the graphic arts to another, etc.) are considered to represent new *works*" (International Federation of Library Associations and Institutions (IFLA) 2009, 18). It is assumed here that the generic identity is determined at the work level.[208] A dramatized or versified version of a novel is considered a new work,[209] whereas a new edition of a novel with some additional chapters or a new title is not.[210] Regarding the treatment of bibliographical information, the abstract notion of a work serves to group various publications of the same work. A novel might be published in a journal, in several subsequent monographic editions, as part of an anthology, or as part of the complete works of an author. These are all different manifestations in FRBR terms. However, such a novel has only one work entry in Bib-ACMé. In the bibliography, the levels of expression and manifestation are combined in the notion of *edition*. That way, every new realization of a novel that is published, for example, a new version with changes in the text, is registered as a new edition, but every new reprint

---

[206] In the case of a novel, the expression means "the specific words, sentences, paragraphs, etc. that result from the realization of a work in the form of a text" (International Federation of Library Associations and Institutions (IFLA) 2009, 19).

[207] See chapters 3.2.3 and 3.3.4 below, where it is explained how the assignment of subgenre labels to the bibliographic entries and texts in the corpus was made.

[208] This point could be discussed further. On the level of genres that are at least in part determined formally, the issue is quite clear, but on the level of (sub)genres that are mainly determined thematically and by content, it is more complicated. How much change is needed to affect to which subgenre a novel belongs? Would it then be considered a new work? However, this discussion is out of the scope of this dissertation because the evolution of individual works is not traced here.

[209] For example, the versified versions of Eduardo Gutiérrez's novels. See footnote 123 above.

[210] "Un año en California" (1869, AR) and "Un viaje al país del oro" (1876, AR) by Juana Manuela, Gorriti, for instance, are considered as one work here because it is just the title of the novel that changed, but not its content.

is also considered a new edition. Together, the number of new realizations and manifestations indicate how successful and popular a novel was. The level of single exemplars is not considered here, although the circulation (the number of printed items of a manifestation) would also convey information about the popularity of the novels. The FRBR concept of *person* is narrowed down to *author* in Bib-ACMé to designate the individuals responsible for the creation and content of the novels. In bibliographic descriptions of novels, an author may appear under different names. Whenever pseudonyms could be associated with the same person, these were grouped together in one author entry in Bib-ACMé.[211]

The information in Bib-ACMé is encoded in XML, following the standard of the Text Encoding Initiative (TEI) in version P5.[212] Compared to full text editions, bibliographic information is highly structured. Therefore, one could also opt for a relational database system to model bibliographic information. However, the use of XML and TEI has some advantages here. For the encoding of historical bibliographical entries, it is very useful to be able to indicate the degree of certainty of information anywhere in the data model because publication dates and places, life dates of authors, etc., are not always well evidenced. In addition, it is reasonable to document the sources of information on several levels, such as the mention of a work in general, the person responsible for a note on a particular edition, and so on. The TEI offers general attributes for this purpose (Text Encoding Initiative Consortium 2023a), which can be added to many different elements. Furthermore, the level of detail is not the same for all pieces of information. For example, sometimes only the year of an author's birth is known, and in other cases, also the month and day. The same applies to novels published serially in a journal: in some cases, the exact dates of the first and last published part are known, and in others, only a year is indicated. For this, it makes sense to have a flexible data model.

Bib-ACMé consists of the following TEI files: "authors.xml", containing all the information about the authors of the novels; "works.xml", where the works are listed with their author, their main title, and additional information such as the subgenre of the novel; and "editions.xml", including information about different editions of the works. The three main files are complemented by "nationalities.xml", "countries.xml", and "sources.xml", which contain controlled values that are referenced from the main files.[213] Example 1 shows one entry from "authors.xml":

```
<person xml:id="A367">
  <persName>
    <surname>Iglesia</surname>
    <forename>Álvaro de la</forename>
    <addName type="pseudonym">Pedro Madruga</addName>
    <addName type="pseudonym">Eligio Aldao y Varela</addName>
    <addName type="pseudonym">Artemio</addName>
    <addName type="pseudonym">A. L. Baró</addName>
    <addName type="pseudonym">Vetusto</addName>
  </persName>
  <birth>
    <date when="1859-04-03">3 de abril de 1859</date>
    <placeName>La Coruña</placeName>
```

[211] Some authors had a whole range of pseudonyms, for example, the Cuban author Teodoro Guerrero y Pallarés, who wrote under seven other names (Instituto de Literatura y Lingüística de la Academia de Ciencias de Cuba 1999, sec. 'G').

[212] For an overview, see Burnard (2014).

[213] The data is accessible at https://github.com/cligs/bibacme/tree/master/app/data. Accessed November 7, 2019.

```
    <placeName>España</placeName>
  </birth>
  <death>
    <date when="1940">1940</date>
    <placeName>La Habana</placeName>
    <placeName>Cuba</placeName>
  </death>
  <sex>masculino</sex>
  <nationality>cubana/o</nationality>
  <idno type="viaf">120788045</idno>
  <note source="#DLC_1999">Llegó a Cuba en 1874. Se estableció en Matanzas.</note>
  <note type="country">
    <country>Cuba</country>
  </note>
</person>
```

Example 1. An entry from "authors.xml".

Each author has a unique identifier used to reference her or him from the works and editions. The author's name is encoded, differentiating between surname, forename, and eventually additional names. When an author's real name is unknown, and the pseudonym does not have the form of forename plus surname, only one name is given. Following the name, information about the birth and death of the author is encoded. The dates are given either only as years or as full dates, depending on the availability of the information. Further information that is given is the sex of the author, the nationality, a note about the country an author is associated with, and optionally a VIAF identifier and a general note. It is important to note that the element `<nationality>` is used in a wide sense here because authors born in Cuba or otherwise assigned to that country before its independence are also listed as "cubana/o".[214] The note indicating the country serves to clearly assign all the authors to one of the three countries Argentina, Cuba, and Mexico. This assignment may correspond to the nationality or country of birth of an author but is not necessarily bound to either of them. For example, an author can have another nationality but be associated with one of the three countries represented in the bibliography because he or she first published his or her works there.

```
<bibl xml:id="W925">
  <author key="A367">Iglesia, Álvaro de la</author>
  <title>La alondra</title>
  <term type="subgenre.title.explicit">novela original</term>
  <term type="subgenre.title.explicit.norm" resp="#uhk">novela original</term>
  <term type="subgenre.title.implicit" resp="#uhk">–</term>
  <term type="subgenre.title.interp" resp="#uhk">novela general</term>
  <term type="subgenre.summary" subtype="signal" resp="#uhk">novela realista</term>
  <term type="subgenre.summary" subtype="theme" resp="#uhk">novela social</term>
  <term type="subgenre.summary" subtype="current" resp="#uhk">novela realista</term>
  <idno type="cligs">nh0221</idno>
  <country>Cuba</country>
  <note type="source">
    <ptr target="#DLC_1999"></ptr>
  </note>
</bibl>
```

Example 2. An entry from "works.xml".

[214] The description of the element in the TEI Guidelines (Text Encoding Initiative Consortium 2013h) allows interpreting it in this wide sense because it is said to contain an informal description of a person's citizenship, and the type of nationality can be characterized further (by birth, naturalized, or self-assigned).

Example 2 shows an entry in "works.xml". A work is encoded as a simple bibliographic citation with a unique identifier. Only the author and main title of the work are given here because information about the publication, i.e., the publication date, publication place, concrete titles and subtitles of an edition, etc., does not correspond to the abstract work level. The author's name in the work entry is connected to the person in "authors.xml" with a key corresponding to the author ID. Further information given in a work entry are terms indicating the subgenre,[215], an optional CLiGS identifier for works that are included in the corpus, the country the work is associated with, and a note pointing to the bibliographic source of the entry. Here, the country is not to be equated with the publication place because a work is also included if an author belongs to the country, even if it was never published there. On the other hand, there are works first published in a country but written by foreign authors.[216] The affiliation of a novel to a country is instead made on the work level. Example 3 below shows an entry from "editions.xml" which corresponds to the above work entry.

```
<biblStruct corresp="works.xml#W925" xml:id="E1284">
  <monogr>
    <author key="A367">Iglesia, Álvaro de la</author>
    <title level="m" type="main">La alondra</title>
    <title level="m" type="sub">(El secreto de Estrovo)</title>
    <title level="m" type="sub">Novela original</title>
    <imprint>
      <publisher>Biblioteca de Follas Novas</publisher>
      <pubPlace corresp="countries.xml#CU">La Habana</pubPlace>
      <date when="1897">1897</date>
    </imprint>
    <extent n="315">315 pp.</extent>
  </monogr>
  <ref>https://catalog.hathitrust.org/Record/100345975</ref>
  <ref>https://archive.org/details/laalondraelsecr00iglegoog</ref>
</biblStruct>
```

Example 3. An entry from "editions.xml".

Editions are encoded with a structured bibliographic citation. Each edition is connected to the work it realizes via the `@corresp` attribute, which points to the "works.xml" file and the work ID. The edition itself also has a unique ID given in `@xml:id`. Depending on the type of publication, details are either given in a combination of the elements `<analytic>` (for dependent publications) and `<monogr>` (for independent publications) or simply the latter. Information about a series a book belongs to may also be given in a `<series>` element.[217] As in "works.xml", the author's name is associated with the person in "authors.xml" via the author key. Another piece of information that is mapped to a controlled list of values is the publication place. In the `@corresp` attribute of the element `<pubPlace>`, the file "countries.xml" and a country key are given. This was made to be able to analyze in which countries the works were published without having to interpret the names of cities on the fly. Part of the information included in an edition entry is also the extent of the publication in page numbers. Finally, when digital versions of the edition were found, links to them were referenced at the end of the entry.

---

[215] See the next chapter 3.2.3, for details about how the subgenres were assigned to the entries in Bib-ACMé.
[216] See chapter 3.1.2 on the "Borders of Argentina, Cuba, and Mexico" above, where this decision is explained.
[217] For details about how bibliographic references are encoded in TEI, see Text Encoding Initiative Consortium (2023b).

The different TEI files are each controlled by their own schemas. It was decided to use different schemas and not one for all the files to be able to keep the data model as strict as possible. The kinds of elements allowed in a file and their order are regulated in RELAX NG schemas.[218] In addition, Schematron files are used to control the content of selected elements and attributes.[219] Example 4 shows one of the rules contained in the Schematron file for "works.xml".

```
<sch:rule context="tei:listBibl/tei:bibl">
  <sch:let name="work-id" value="@xml:id"/>
  <sch:assert test="matches($work-id,'^W\d+$')">The id of a work should have the form "W
       + number"</sch:assert>
  <sch:assert test="doc('../data/editions.xml')//tei:biblStruct[@corresp = concat('works
      .xml#',$work-id)]">There is no corresponding work-id in editions.xml</sch:assert>
</sch:rule>
```

Example 4. A rule in the Schematron file "works.sch".

The rule applies to the context of an individual bibliographic entry and tests two assertions. The first assertion checks the form of the work identifier, and the second assertion tests if there is an edition in "editions.xml" that corresponds to the work in question. The example shows that Schematron can be used to make validations across several XML documents, which is important for Bib-ACMé because it is organized in separate TEI files that contain references to each other. That way, it can be assured that the identifiers used for authors, works, and editions, are consistent throughout the database and that there are no superfluous or missing entries. In addition, checks that involve the comparison of values are not possible with the general schema language RELAX NG. Other aspects that are controlled with the Schematron files are the correspondence of author names between the different files, the structure of CLiGS identifiers, and that source and country codes are referenced correctly.

The preparation of the entries from the bibliographic sources so that they conform to the data model of Bib-ACMé makes a wide range of analytical approaches possible. The data can be evaluated on the three main levels of authors, works, and editions and regarding more detailed information encoded in the TEI files. Overviews of the information contained in Bib-ACMé are given in chapter 4.1 below, where the bibliography of novels is compared to the corpus.

---

[218] RELAX NG is a schema language for XML. It can either be expressed as an XML document or in a compact, non-XML syntax. See Murata (2014). For Bib-ACMé, the compact syntax is used because it gives a quick overview of the documents' structure.

[219] Schematron is a rule-based schema language using the query language XPath to validate XML documents. It allows to define rules that consider the context of elements and attributes so that very specific constraints can be formulated (Siegel 2022). All the schema files for Bib-ACMé are available at https://github.com/cligs/bibacme/tree/master/app/schemas (the RELAX NG files ending in ".rnc" and the Schematron files in ".sch"). Accessed November 8, 2019. Although the TEI offers the possibility to create a meta schema called ODD ("One document does it all", Text Encoding Initiative Consortium n.d.b) from which schemas in different syntaxes can be derived, it was decided not to use one or several ODDs for Bib-ACMé. An ODD is very useful for general TEI models created for mixed content scenarios, where the same element can be used in different contexts and with a variety of attributes, but it is quite complex to narrow an ODD down to a strict data model for highly structured data. In the case of Bib-ACMé, this effort would not have been compensated by the additional benefits of the ODD.

3.2.3 Assignment of Subgenre Labels

Several kinds of subgenre labels were assigned to the works in Bib-ACMé to get an overview of the subgenres to which the novels in the bibliography belong. The labels fall into three principal groups: The first group is derived from main titles, subtitles and series titles of the novels ("subgenre.title") and includes explicit as well as implicit genre signals, the second is taken from literary-historical sources ("subgenre.litHist"), and the third group summarizes and categorizes the subgenre values collected in the other two groups ("subgenre.summary"). In this chapter, first, an example is presented to illustrate how the subgenre labels were assigned to the bibliographic entries of the novels and how they were encoded in TEI. More general considerations regarding the assignment of subgenre labels to the novels are made when discussing the first example but also in the sections following it. Different levels of subgenre terms that are used in the encoding are presented in chapter 3.2.3.2. On the one hand, the subgenre labels are differentiated by the type of source from which they were collected. Labels can be explicit historical labels or be derived from implicit historical signals, or they can be collected from literary-historical sources. The differences between these kinds of labels are discussed in chapters 3.2.3.3 ("Explicit and Implicit Subgenre Signals"), 3.2.3.4 ("Interpretive Subgenre Labels"), and 3.2.3.5 ("Literary-Historical Subgenre Labels"). On the other hand, the subgenre labels are sorted according to discursive aspects. It is assumed that a literary work is a complex discursive and semiotic object to which generic terms refer on different levels. A model summarizing the discursive levels that are relevant for the bibliography and corpus of novels at hand is presented in chapter 3.2.3.6 ("A Discursive Model of Generic Terms").

3.2.3.1 An Example

Example 5 shows the work entry of the novel "Los casamientos del diablo" (1889, AR) by Enrique Ortega in the digital bibliography, which includes several subgenre labels.

```
<bibl xml:id="W563">
 <author key="A227">Ortega, Enrique</author>
 <title>Los casamientos del diablo</title>
 <term type="subgenre.title.explicit">novela histórica americana</term>
 <term type="subgenre.title.explicit.norm" resp="#uhk">novela histórica</term>
 <term type="subgenre.title.explicit.norm" resp="#uhk">novela americana</term>
 <term type="subgenre.title.explicit.norm" resp="#uhk">novela</term>
 <term type="subgenre.title.implicit" resp="#uhk">novela sentimental</term>
 <term type="subgenre.title.implicit" resp="#uhk">novela romántica</term>
 <term type="subgenre.summary.signal.explicit" resp="#uhk">novela histórica</term>
 <term type="subgenre.summary.signal.explicit" resp="#uhk">novela americana</term>
 <term type="subgenre.summary.signal.explicit" resp="#uhk">novela</term>
 <term type="subgenre.summary.signal.implicit" resp="#uhk">novela sentimental</term>
 <term type="subgenre.summary.signal.implicit" resp="#uhk">novela romántica</term>
 <term type="subgenre.summary.theme.explicit" resp="#uhk" cligs:importance="2">novela
     histórica</term>
 <term type="subgenre.summary.theme.implicit" resp="#uhk">novela sentimental</term>
 <term type="subgenre.summary.current.implicit" resp="#uhk">novela romántica</term>
 <term type="subgenre.summary.identity.explicit" resp="#uhk">novela americana</term>
 <term type="subgenre.summary.mode.reality.explicit" resp="#uhk">novela histórica</term>
 <term type="subgenre.summary.mode.representation.explicit" resp="#uhk">novela</term>
 <country>Argentina</country>
```

```
 <note type="source">
   <ptr target="#Lichtblau_1997"/>
 </note>
</bibl>
```

Example 5. Subgenre labels for the work “Los casamientos del diablo”.

The subgenre labels are encoded in `<term>` elements that are characterized further by the attribute `@type`. First, explicit generic labels that occur in the main title, subtitle, or series title, are marked as `"subgenre.title.explicit"`. In the above example, there is an edition of the novel with the subtitle “novela histórica americana”. Because the generic identity is determined on the work level, information about the subgenre is taken from all the work’s editions. Therefore, if there are several editions and only one carries the explicit subgenre label, it is nonetheless included here. If there are several editions with differing subgenre labels, all of them are considered. The second type of label is called `"subgenre.title.explicit.norm"` and contains a normalized version of the explicit subgenre label. The `@resp` attribute indicates by whom the normalization was done. For the novel at hand, the label “novela histórica americana” is normalized to several individual subgenre labels: “novela histórica”, “novela americana”, and “novela”. The primary purpose of the normalization step is to make the explicit subgenre labels comparable in computational analysis. In the current case, the first label “novela histórica” refers to a subgenre predominantly determined by the theme of the novel. The second label “novela americana” points to the cultural-geographical and linguistic origin and identity of the novel. In the bibliography, there are also novels with the subtitle “novela argentina”, “novela cubana”, “novela mexicana”, “novela original”, etc. These kind of labels either refer to the continent (“americana”), to the country (“argentina”, “cubana”, “mexicana”), or to the fact that the novel was originally written in Spanish and not translated (“original”).[220] The third label extracted from the example, “novela”, refers to the genre of the text. It is encoded as a subgenre label here, as well, because in the bibliography, only about half of the works carry the explicit label “novela”.[221] Because all the works that are included follow the selection criteria for novels defined in chapter 3.1.1 above, it is assumed that the explicit label “novela” points to a subtype of all the texts that can be considered as novels formally. Other labels usually designating genres, such as “cuento”, “drama”, “ensayo”, or “leyenda”, are treated in the same way.

[220] The usage of the label “novela original” in the Spanish context is described by Botrel as follows: “Las normas/formas tipográficas bibliográficas permiten también observar cómo después de un período en el que se precisa el origen de la novela («novela escrita en francés por Mr.» o «Madama…», «en inglés por Mistress…» o «Sr…» y «traducida al castellano por…» iniciales) al preeminencia del título unida con la hispanización casi sistemática de los nombres de los autores traducidos (Javier de Montepín, Pablo Feval, etc. y la importancia numérica de las traducciones, con la desaparición de la mención del traductor, al menos en las referencias bibliográficas, hace que el género novela venga disociado de una por otra parte deseada hispanidad y asociado con una patronímica y toponimia extranjerizante, como producto extranjero o, más probablemente, asimilado. La mención «novela original» o «española» introducirá durante cierto tiempo una distinción poco decisiva, estadísticamente al menos” (Botrel 2001, paras 12, 13).

[221] Of 829 works in the bibliography, only 403 carry the explicit label “novela”.

After the explicit generic signals of the titles, implicit signals are evaluated and captured in terms of the type `"subgenre.title.implicit"`. Here again, the attribute `@resp` serves to indicate who made the interpretation. In this case, two implicit labels, "novela sentimental" and "novela romántica", are recorded. The word "casamientos" in the title is interpreted as a reference to a sentimental plot, and the whole title "los casamientos del diablo" is interpreted as a sign of a novel of the romantic current.[222]

The above example does not contain terms of the second group (`"subgenre.litHist"`) because for the novel "Los casamientos del diablo", no statements about its subgenre were found that were made by literary historians. Therefore, the terms of the third group (`"subgenre.summary"`) only take up the values that were inferred from the title. The summary at the end of the different subgenre terms has the function of organizing the previous data into categories of generic information in order to enhance the comparability of the terms throughout the bibliography for further analysis. What kind of generic information is given, is indicated in the part of the `@type` attribute after `"subgenre.summary"`. The summary values have five subtypes: `"signal"`, `"theme"`, `"current"`, `"identity"`, and `"mode"`. Terms of the type `"subgenre.summary.signal"` contain all the subgenre labels that were signaled by the work title either explicitly (marked as `"subgenre.summary.signal.explicit"`) or implicitly (`subgenre.summary.signal.implicit`). In the above example, all the subgenre labels are derived from signals of the text. However, in other cases, there are further subgenre labels that were assigned to the work by critics, but that cannot be deduced from the work's title. The second subtype of the summary values is `"theme"`. Terms of this type contain all the labels that refer to subgenres defined primarily or in part by the theme of the text. In the current example, there are two thematic labels: "novela histórica" and "novela sentimental". The first one was given explicitly and is therefore encoded as a term of the type `"subgenre.summary.theme.explicit"` whereas the second one was deduced from the title and is marked as `"subgenre.summary.theme.implicit"`.

The first of the "theme" terms carries the attribute `@cligs:importance` with the value `"2"`. With this attribute, an order of priority is given for cases with several subgenre terms of the same type. It was decided to use this attribute only for "theme" and "current", i.e., for subgenre labels belonging to these two categories. These are the types of subgenre labels that are at the center of interest of this dissertation. Furthermore, most literary histories and critical studies refer to novelistic subgenres of this kind. As to the priorities, in general, only one high priority (`"2"`) is assigned, while the other terms without this attribute are interpreted as low-priority terms. Just as the normalization of explicit titles serves to enhance comparability, this prioritization has the pragmatic function of being able to select one value for each subgenre term of the types "theme" and "current" for cases where unique values are needed in an analysis. However, it is a simplification because it is ultimately not possible to map different subgenre assignments to a discrete numerical system as they usually represent different perspectives on the literary work.[223] As rules of thumb, terms deduced from explicit signals are rated higher than those going back to implicit ones. Furthermore, signals that are stronger are valued higher, for example, if there are

---

[222] In romantic novels, it is common that issues and people are portrayed in black and white (e.g., the good and the bad), exaggerated, or dramatized, and the word "diablo" points in that direction.

[223] In the case at hand, "novela sentimental" refers primarily to the content and plot of the novel, while "novela romántica" points to the literary current. Nevertheless, there are overlaps between both terms.

several signals pointing to a certain subgenre and only one signal points to another. In addition, terms that are mentioned by literary critics are valued higher than those that are not.

After the thematic terms, those referring to literary currents are listed (`"subgenre.summary.current"`). In the above example, there is only one term of this kind, the "novela romántica". The term "novela americana" is encoded as a term of the type `"subgenre.summary.identity"`. Finally, there are two subgenre labels grouped into the "mode" category: "novela histórica" belongs to the category `"subgenre.summary.mode.reality"` and "novela" to `"subgenre.summary.mode.representation"`. The "mode" group contains labels that are not thematic and do not refer to literary currents or the cultural or linguistic identity of the works. Instead, these are labels indicating how the works relate to extratextual circumstances or to the way the text is organized and presented. In the example, `"mode.reality"` designates labels that involve the relationship of the text to reality. Usually, a historical novel intends to present settings and events of the past, but not the present reality. `"mode.representation"` includes labels that indicate how the novel is organized and presented linguistically. The term "novela" means that the text is presented in the narrative mode and not, for example, as a dramatic text. As can be seen in the example, some subgenre labels are repeated in the summary, in this case, "novela histórica", which falls into the two categories `"subgenre.summary.theme"` and `"subgenre.summary.mode.reality"`. On the other hand, each novel can have several subgenre labels of the same kind, as the two thematic labels "novela histórica" and "novela sentimental" of this example show. Finally, because all the values in the summaries are normalized, the summary terms also carry a `@resp` attribute that shows who entered the values.

3.2.3.2 Levels of Subgenre Terms

The system of the summary values needs to be explained further. Which categories were chosen, for which reasons, and which values can they take? The subgenre categories chosen ("theme", "current", "identity", and "mode" with further subtypes) are not generally exhaustive from a genre theoretical perspective and not congruent to one specific theoretical model of genre. Instead, they reflect the generic signals that occur in the collection of novels represented in the bibliography and the corpus, as well as the terms with which the subgenres of these novels are described by literary historians. There are general models for describing the different levels to which generic labels might refer. Some of these models include more categories than the ones chosen here, and others have fewer or different categories. The categories chosen here are, for the most part, derived from a model developed by Wolfgang Raible (Raible 1980). In table 2, the different subgenre categories used in the present model are listed, exemplified, and commented on, and the levels of Raible's model that correspond to the ones here or are similar to them are given.

Kind of subgenre label	Value of type	Examples	Explanation	Level in Raible's model
signal	subgenre.summary.signal	*novela histórica*, *novela naturalista*, *novela original*, *memorias*	subgenre labels that are derived from explicit or implicit signals of the novel	-
theme	subgenre.summary.theme	*novela gauchesca*, *novela histórica*, *novela sentimental*	subgenre labels that refer to a main theme of the novel	*Objektbereich*
current	subgenre.summary.current	*novela romántica*, *novela realista*, *novela naturalista*	subgenre labels that refer to the literary current of the novel	-
identity	subgenre.summary.identity	*novela americana*, *novela mexicana*, *novela original*	subgenre labels that refer to the cultural-geographical and linguistic identity of the novel	-
mode	subgenre.summary.mode	*novela epistolar*, *novela fantástica*, *novela humorística*, *cuadros*, *drama*, *memorias*	subgenre labels that refer to the mode the novel is narrated in / the form it is presented in	*Kommunikationssituation*, *Verhältnis zwischen Text und Wirklichkeit*, *Medium*, *sprachliche Darstellungsweise*
intention	subgenre.summary.mode.intention	*novela cómica*, *novela moralista*, *novela de propaganda*	subgenre labels that refer to the aim the author/narrator pursues with the novel	*Kommunikationssituation*

attitude	subgenre.summary.mode.attitude	*novela política*, *novela satírica*	subgenre labels that refer to the attitude the author/narrator has towards what is represented in the novel	*Kommunikationssituation*
reality	subgenre.summary.mode.reality	*novela científica*, *novela fantástica*, *novela histórica*, *leyenda*	subgenre labels that refer to the relationship between the novel and reality	*Verhältnis zwischen Text und Wirklichkeit*
medium	subgenre.summary.mode.medium	*novela epistolar*, *croquis*, *cuadros*, *páginas*, *panorama*	subgenre labels that refer to the medium that the novel uses (also in a figurative sense)	*Medium*
representation	subgenre.summary.mode.representation	*cuento*, *drama*, *ensayo*, *episodios*, *novela*	subgenre labels that refer to the mode the novel is represented in linguistically (or narratively)	*sprachliche Darstellungsweise*

Table 2: Types of summarizing subgenre labels.

3.2.3.3 Explicit and Implicit Subgenre Signals

Different types of subgenre labels were already introduced with the example "Los casamientos del diablo" above. In what follows, some more general considerations regarding the system of subgenre labels developed here are made, beginning with the category "signal". It comprises subgenre labels that are derived from explicit or implicit signals of the novel, either in paratextual elements (in a title, subtitle, series title, preface, epigraph, etc.) or in the opening of the texts. These labels can be of any of the following kinds of labels (thematic labels, labels referring to literary currents, or other types of labels). For the bibliography, they were only derived from the titles because this is the only paratextual information directly available in the bibliographic records. For the corpus, other signals were evaluated as well. Apart from that, the approach to assigning the subgenres is the same for the bibliography and the corpus. Therefore, the general points are explained in this section, while only the additional corpus-specific aspects are explained below in chapter 3.3.4.

It is important to note that signals can be explicit subgenre labels, for example, the subtitle "novela histórica americana" above. Besides that, they can also be aspects of the title (and other paratextual and textual elements) that can be interpreted in terms of subgenre labels, for example, "casamiento" as pointing to a sentimental novel and "diablo" to a romantic novel. The evaluation of the signals thus involves a significant interpretive step, and it presupposes knowledge about possible subgenres. The knowledge is, on the one hand, derived from the bibliography and the corpus itself (which subgenres occur frequently and what are their characteristics?) and, on the other hand, from representations of the subgenres in literary-historical works. By encoding many steps of this interpretation process (starting from explicit labels, going on to normalized values, mentioning implicit signals, summarizing all in the categorized labels, and keeping their origin as "explicit" or "implicit"), it should be possible to follow the decisions made here closely for each of the novels in the bibliography. Nevertheless, another encoder might have reached other results. The position adopted here is that the genre or subgenre of a text cannot be determined unequivocally without presuppositions. To avoid the influence of the own previous knowledge or the necessity of previous definitions, one could opt for only referring to explicit generic labels. However, in the case of the bibliography and corpus at hand, this would have led to a very reduced setup because only some kinds of explicit subgenre labels are very frequent. Table 3 lists the top most frequent explicit labels, ordered by the frequencies of the normalized versions.[224]

---

[224] The counts are based on the final bibliography. Novels carrying several different labels were counted twice, and the percentages were rounded. See also chapter 4.1.5 for an overview of the subgenres in the bibliography and the corpus. The normalized frequencies are higher because some novels carry explicit labels, e.g., "histórico" or "costumbres", but not the whole label involving the word "novela", i.e., "novela histórica" or "novela de costumbres", or they contain additional elements. For example: "novela de carácter histórico", "episodio histórico", "leyenda histórica", "historia novelada"; "novela original de costumbres", "cuadro de costumbres", "boceto de costumbres", "ensayo de costumbres", etc. All the subgenre labels that occur at least ten times in the normalized version are included. In addition, the number of novels without any explicit label is also given. The script used to determine the subgenre label counts given in this and the following tables in this section is available at https://github.com/cligs/scripts-nh/blob/master/corpus/frequencies-subgenre-labels.xsl and the full table of frequencies can be viewed at https://github.com/cligs/data-nh/blob/master/corpus/bibliography-subgenre-labels/frequencies-explicit-labels.csv. Accessed March 7, 2020.

Subgenre label	Frequency explicit		Frequency explicit normalized	
novela	398	48 %	403	49 %
novela histórica	73	9 %	133	16 %
novela original	97	12 %	113	14 %
novela mexicana	6	1 %	67	8 %
novela de costumbres	25	3 %	57	7 %
episodios	64	8 %	54	7 %
memorias	49	6 %	54	7 %
leyenda	42	5 %	44	5 %
novela cubana	18	2 %	35	4 %
drama	25	3 %	28	3 %
novela nacional	0	0 %	26	3 %
historia	22	3 %	25	3 %
cuento	15	2 %	15	2 %
novela argentina	5	1 %	15	2 %
novela social	2	0 %	13	2 %
novela americana	4	0 %	12	1 %
escenas	11	1 %	12	1 %
novela policial	0	0 %	11	1 %
novels without any explicit label	207	25 %	-	-

Table 3. Top most frequent explicit subgenre labels in the bibliography.

As can be seen, only a few of the top most frequent explicit labels refer to subgenres of the novel in common sense, i.e., labels related to the themes of the novels: 16 % of the works carry the label “novela histórica”, 7 % the label “novela de costumbres”, 2 % the label “novela social”, and 1 % the label “novela policial”. The other frequent labels are either of a very general nature (“novela”, “leyenda”, “drama”, “historia”, “cuento”) or they refer to aspects of the novels that are usually not focused on in subgenre studies, such as the identity of the texts (“novela original”, “novela mexicana”, “novela cubana”, “novela nacional”, “novela argentina”, “novela americana”) or the way the text is structured and presented linguistically (“episodios”, “memorias”, “escenas”). Furthermore, it can be noted that even the topmost frequencies decrease sharply. Finally, not all the novels have explicit labels: 25 % of the novels in the bibliography do not convey any generic information explicitly.

For some of the subgenre labels that entered the top list, an *author and series bias* can be noted.[225] Most of the occurrences of the terms “drama” and “novela policial”, for example, stem from the numerous novels written by the Argentine author Eduardo Gutiérrez, which are organized in series and carry subtitles of the form “dramas policiales”, “dramas militares”, “dramas cómicos”, etc. The many “episodios”, “memorias”, “leyendas”, “novelas nacionales”, and “historias” are connected to series of historical novels, some of which are called “episodios nacionales” or “leyendas históricas”. It was decided not to keep the combined labels in the

[225] Strictly speaking, this is no bias, though, because the novelistic production of the time is as it is: some authors wrote more novels than others, and some wrote whole series of novels of a certain subgenre.

Subgenre label	Frequency absolute	Frequency relative
novela	404	49 %
novela romántica	269	32 %
novela sentimental	252	30 %
novela histórica	244	29 %
novela social	177	21 %
novela de costumbres	133	16 %
novela realista	122	15 %
novela original	113	14 %
novela naturalista	81	14 %
novela mexicana	67	8 %
novels withouth any subgenre assignment	51	6 %

Table 4. Top most frequent subgenres in the bibliography.

normalized form, though, because even if some of the combinations occur several times and lead to correlations in the frequencies of the labels, their components are also part of other kinds of subtitles. Furthermore, the combinations of individual subgenre labels in the subtitles are so varied and often individual that it would be impossible to compare them without any normalization step. The original combinations can still be reproduced because they are encoded in terms of the type `"subgenre.title.explicit"`.

3.2.3.4 Interpretive Subgenre Labels

If only the thematically oriented explicit subgenre labels would be regarded, most of the 829 novels in the bibliography would have had to be considered general fiction because only 36 % of the novels have such labels.[226] However, many of the novels have been interpreted as belonging to certain subgenres, and many signal their subgenre(s) implicitly. Some well-known and also relatively frequent subgenres of the novel are rarely indicated explicitly, for example, sentimental novels. In the whole bibliography, there is only one novel with the explicit subtitle "novela sentimental", but many more novels can be assigned to this subgenre. When also implicit signals and literary-historical assignments are included, the picture of the top most frequent subgenres changes, as table 4 shows.[227]

Many more novels in the bibliography are covered with this approach. So for the reasons given, it was decided to include interpretive subgenre labels, as well. Both implicit signals evaluated by the author of this dissertation and assignments made by other literary historians are considered as such. The difference between both is that for the other literary-historical labels, it is not known

[226] See https://github.com/cligs/data-nh/blob/master/corpus/bibliography-subgenre-labels/frequencies-explicit-thematic-labels.csv for a full table of frequencies of explicit thematic labels in the bibliography. Accessed March 7, 2020.

[227] The table shows the top ten subgenre assignments and the number of novels without any assignment. See https://github.com/cligs/data-nh/blob/master/corpus/bibliography-subgenre-labels/frequencies-subgenre-labels.csv for the full table. Accessed March 7, 2020.

Kind of subgenre	Subgenre labels
theme	*Künstlerroman*, *novela abolicionista*, *novela biográfica*, *novela científica*, *novela contemporánea*, *novela criminal*, *novela de aventuras*, *novela de costumbres*, *novela de familia*, *novela de la ciudad*, *novela de misterio*, *novela de viajes*, *novela didáctica*, *novela doméstica*, *novela filosófica*, *novela gauchesca*, *novela histórica*, *novela humorística*, *novela indigenista*, *novela militar*, *novela moralista*, *novela picaresca*, *novela política*, *novela psicológica*, *novela regional*, *novela sentimental*, *novela social*
current	*novela romántica*, *novela realista*, *novela naturalista*
identity	*novela regional*
mode.intention	*novela didáctica*, *novela humorística*, *novela moralista*
mode.attitude	*novela abolicionista*
mode.reality	*novela científica*, *novela contemporánea*, *novela de misterio*, *novela histórica*
mode.representation	*novela filosófica*, *novela psicológica*

Table 5. Set of subgenres used as a basis for the interpretation of implicit signals.

in detail on what bases they were assigned.[228] Literary-historical labels are discussed in more detail below. The interpretive labels worked out here are derived from specific textual signals: the titles (in the case of the bibliography) and additional paratextual elements (in the case of the corpus). The decisions rest on a certain set of subgenres taken as the basis for interpreting the implicit signals. This set does not comprise all of the existing explicit labels, though. Instead, the focus is on subgenres related to themes and literary currents, as these are the kinds of subgenres most often referred to in literary histories and also because there are known concepts of these subgenres that can be used.[229] In addition, the set contains some subgenres that repeatedly occur as explicit labels in the bibliography, and that can be inferred from textual signals in other cases, even if they are not part of the critical subgenre canon, for example, the "novela contemporánea". Table 5 contains the set of subgenres used to interpret implicit signals. Like the list of kinds of subgenres, this set is also not exhaustive from a general perspective on the subgenres of the novel. Instead, it is based on the relevance of the subgenres for the bibliography and the corpus.

Some of the subgenres of this set that are included in the thematic group also belong to other levels of the model defined above.[230] They are listed again in the lower part of the table for the sake of completeness. Nevertheless, when this set of subgenres was applied to interpret implicit signals, the focus was on the thematic aspects. Furthermore, the thematic subgenres are placed on different levels of generality. The types *novela de familia*, *novela de la ciudad*, and *novela doméstica* are more specific than, for example, *novela social*, and they could also be subsumed under the latter term. That terms of different levels of generality occur in the list is because there are signals in the bibliography and the corpus that can best be interpreted with these labels. For

[228] This is usually only outlined explicitly in specific and comprehensive studies of certain subgenres, e.g., in Rivas (1990) for the anti-slavery novel or Schlickers (2003) for the naturalist novel. In contrast, in overview works such as general literary histories, the criteria for the assignment of novels to subgenres are normally not explained.

[229] See chapter 2.3 for a presentation of the subgenres related to themes and literary currents.

[230] Which subgenre labels are grouped in which subgenre category here is explained in detail below (see table 10).

Subgenre label	Frequency absolute	Frequency relative
novela sentimental	252	30 %
novela histórica	244	29 %
novela social	177	21 %
novela de costumbres	133	16 %
novela política	51	6 %
leyenda	44	5 %
novela criminal	37	4 %
novela de la ciudad	27	3 %
novela indigenista	27	3 %
novela gauchesca	21	3 %
novels withouth thematic label	134	16 %

Table 6. Top most frequent thematic subgenre labels in the bibliography.

Subgenre label	Frequency absolute	Frequency relative
novela romántica	269	32 %
novela realista	122	15 %
novela naturalista	81	10 %
novela modernista	8	1 %
novela verista	5	1 %
novela clasicista	3	0 %
novels without label of literary current	424	51 %

Table 7. Frequencies of subgenre labels related to literary currents in the bibliography.

example, the novel "La familia de Sconner" (1858, AR) by Miguel Cané (father) is interpreted as a *novela de familia* and a *novela social* and the novel "La sociedad y sus víctimas. Escenas bonaerenses" (1902, AR) by Matías Calandrelli both as a *novela de la ciudad* and a *novela social*.[231]

Following up on the question of how many novels are covered when also interpretive subgenre labels are included, tables 6 and 7 show the most frequent subgenre labels related to themes and literary currents, including explicit as well as implicit signals and literary-historical assignments.[232]

The four biggest thematic groups are sentimental, historical, social novels, and novels of manners (*novela de costumbres*). For 16 % of the novels, no thematic label could be assigned. The literary current most frequently assigned are romantic novels, followed by realist and naturalist novels. In the case of the literary currents, more than half of the novels in the bibliography do not have any label of this kind (51 %). One reason for this is that the literary current is usually

[231] Which subgenre labels are related to others is explained below in table 10.

[232] The table shows the top ten plus the number of entries in the bibliography without this type of subgenre label. Full tables are available at https://github.com/cligs/data-nh/blob/master/corpus/bibliography-subgenre-labels/frequencies-thematic-labels.csv and https://github.com/cligs/data-nh/blob/master/corpus/bibliography-subgenre-labels/frequencies-labels-currents.csv. Accessed March 8, 2020.

not given explicitly: there are only five novels in the whole bibliography with explicit signals for the “novela naturalista” and six for the “novela realista”. The term “novela romántica” does not occur at all. The second reason is that literary currents are mainly a concern of literary historians, and for 48 % of the novels in the bibliography, no assignments made by literary historians could be found. An important point to consider when looking at the numbers is that they do not mean that the novels that do not have a certain subgenre label do not possibly belong to that subgenre. The distribution of subgenre labels only indicates that these are the cases where information (explicit, implicit, literary-historical) is available.

3.2.3.5 Literary-Historical Subgenre Labels

Going on to the discussion of literary-historical labels, example 6 shows the entry of the work “Santa” (1903, MX) by Federico Gamboa. For this work, there are no labels from the first group (explicit or implicit labels inferred from the title) but many from the second (labels taken from literary-historical sources).

```
<bibl xml:id="W838">
  <author key="A326">Gamboa, Federico</author>
  <title>Santa</title>
  <term type="subgenre.litHist" resp="#Schlickers_2003">novela naturalista</term>
  <term type="subgenre.litHist" resp="#Dill_1999">naturalistischer Roman</term>
  <term type="subgenre.litHist" resp="#Dill_1999">Großstadtroman</term>
  <term type="subgenre.litHist" resp="#Varela-Jacome_1982">novela naturalista</term>
  <term type="subgenre.litHist" resp="#Sanchez_1953">novela de tendencia objetiva</term>
  <term type="subgenre.litHist" resp="#Sanchez_1953">novela naturalista</term>
  <term type="subgenre.litHist" resp="#Sanchez_1953">novela de tendencia mixta</term>
  <term type="subgenre.litHist" resp="#Sanchez_1953">novela social</term>
  <term type="subgenre.litHist" resp="#Galvez_1990">novela de tendencia naturalista</
      term>
  <term type="subgenre.litHist" resp="#Galvez_1990">novela del período realista</term>
  <term type="subgenre.litHist" resp="#Lichtblau_1959">Naturalism</term>
  <term type="subgenre.litHist" resp="#Roessner_2007">Naturalismus</term>
  <term type="subgenre.litHist" resp="#FernandezAriasCampoamor_1952">Naturalismo</term>
  <term type="subgenre.litHist.interp" resp="#uhk">novela naturalista</term>
  <term type="subgenre.litHist.interp" resp="#uhk">novela social</term>
  <term type="subgenre.litHist.interp" resp="#uhk">novela realista</term>
  <term type="subgenre.summary.signal.implicit" resp="#uhk">novela naturalista</term>
  <term type="subgenre.summary.theme.litHist" resp="#uhk">novela social</term>
  <term type="subgenre.summary.current.implicit" resp="#uhk" cligs:importance="2">novela
       naturalista</term>
  <term type="subgenre.summary.current.litHist" resp="#uhk">novela naturalista</term>
  <term type="subgenre.summary.current.litHist" resp="#uhk">novela realista</term>
  <idno type="cligs">nh0080</idno>
  <country>México</country>
  <note type="source">
    <ptr target="#Torres-Rioseco_1933"/>
  </note>
</bibl>
```

Example 6. Subgenre labels for the work “Santa”.

In various literary-historical works, “Santa” is classified as a naturalistic novel. The literary-historical labels are collected in terms of the type `"subgenre.litHist"`, and the respective source is given in the attribute `@resp`. All the different literary-historical assignments are summarized

Kind of subgenre	Subgenre labels
theme	*Bildungsroman*, *crónica*, *Künstlerroman*, *memorias*, *novela abolicionista*, *novela científica*, *novela criminal*, *novela de aventuras*, *novela de costumbres*, *novela de familia*, *novela de la ciudad*, *novela didáctica*, *novela documentaria*, *novela de misterio*, *novela de viajes*, *novela fantástica*, *novela gauchesca*, *novela histórica*, *novela indigenista*, *novela moralista*, *novela picaresca*, *novela política*, *novela psicológica*, *novela regional*, *novela sentimental*, *novela social*
current	*novela clasicista*, *novela modernista*, *novela naturalista*, *novela realista*, *novela romántica*, *novela verista*
identity	*novela regional*
mode.intention	*novela didáctica*, *novela moralista*
mode.attitude	*novela abolicionista*, *novela política*, *novela satírica*
mode.reality	*novela científica*, *novela de misterio*, *novela fantástica*, *novela histórica*
mode.medium	*novela epistolar*
mode.representation	*crónica*, *memorias*, *novela documentaria*, *novela epistolar*, *novela psicológica*

Table 8. Set of subgenres used as a basis for the interpretation of literary historical subgenre labels.

in terms of the type `"subgenre.listHist.interp"`. In the case of "Santa", the subgenre assignments made by critics are quite unanimous. Besides being classified as a naturalistic novel, "Santa" is also described as a realist and a social novel. Like explicit and implicit signals, literary-historical labels are also summarized and categorized further in terms of the type `"subgenre.summary"`, following the procedures explained with the first example above. For "Santa", the literary-historical labels result in three summary terms: "novela social" is encoded as a thematic label (`"subgenre.summary.theme.litHist"`), and "novela naturalista" and "novela realista" are grouped as labels referring to literary currents (`"subgenre.summary.current.litHist"`). The label "novela naturalista" is weighted higher than "novela realista" because it is mentioned more often by literary critics (`@cligs:importance="2"`) and also because it occurs as an implicit signal in the paratext of the novel, as indicated with the terms `"subgenre.summary.signal.implicit"` and `"subgenre.summary.current.implicit"`. This implicit signal is not derived from the title of the novel, though, but from other paratextual elements. In the case of "Santa", this is possible because the novel is part of the corpus and was analyzed in more detail.[233]

In the same way that the explicit generic information occurring in the titles and other paratexts is normalized, also the subgenre labels collected from literary-historical works are interpreted and standardized because not all literary historians use the same terminology. Table 8 lists the different interpretive values contained in terms of the type `"subgenre.litHist.interp"` as well as the kinds of subgenres with which these values can be associated.

As for the set of subgenre labels used to interpret titles and other paratexts of the novels, the set of labels interpreted from literary-historical subgenre labels also focuses on thematic labels and labels referring to literary currents. The other kinds of subgenre labels are only of secondary importance in the interpretation process. A subgenre that is often mentioned in literary-historical

[233] See chapter 3.3.4 about the assignment of subgenre labels for the texts in the corpus.

works is the *novela costumbrista* (Gálvez 1990, 100–101; Remos y Rubio 1935, 57–109; Sánchez 1953, 227–256). It was decided to interpret this label as *novela de costumbres* because, historically, the novels carried the latter label. *Costumbrismo* is also often described as a literary current (Dill 1999, 155–157; Rössner 2007, 146–147), but this aspect is not highlighted here because the *novelas de costumbres* were written and published throughout the whole the nineteenth century. Some of them can be attributed to the romantic current, others to the realist current, and even some naturalistic novels carry labels including the word "costumbres".[234] Other standardizations were made. For example, novels related to gauchos were all normalized to *novela gauchesca*, novels related to cities to *novela de la ciudad*, novels about indigenous people to *novela indigenista*, novels about the system of slavery in Cuba to *novela abolicionista*, and novels related to crimes to *novela criminal*.[235]

In Bib-ACMé, only a selection of literary-historical sources was used for the assignment of subgenre labels. The critical literature on Spanish-American novels is vast, so a choice had to be made. Works of different scopes were selected, preferably those where the assignment of a novel to a subgenre is explicit. The sources used are listed in table 9.

In literary-historical works, assignments to subgenres sometimes occur in the text and also often through the structure and organization of a literary history if a novel is mentioned in a section carrying the title of a subgenre.[236] However, clear assignments are not always made

---

[234] Examples for romantic *novelas de costumbres* are "Ironías de la vida" (1851, MX) and "La hora de Dios" (1865, MX) by Pantaleón Tovar and the series of "novelas de costumbres" written by José Tomás de Cuéllar (Fernández-Arias Campoamor 1952, 63–65e been characterized as). Novels of this type that havrealist are "La familia Quillango" (1880, AR) by José María Cantilo and "Antón Pérez" (1903, MX) by Manuel Sánchez Mármol (Fernández-Arias Campoamor 1952, 86; Gálvez 1990, 126–127). Naturalistic novels carrying the label "costumbres" in their title are "Quimera. Boceto de costumbres" (1899, AR) by José Luis Cantilo and "Fruto vedado (Costumbres argentinas)" (1884, AR) by Paul Groussac. On *Costumbrismo* as a longer lasting phenomenon, Fernández-Arias Campoamor writes: "Los novelistas románticos que fueron costumbristas constituyen el puente tendido entre el romanticismo y el realismo. Costumbrismo cultivado ocasionalmente, en realidad, lo hubo siempre en todas las literaturas […] Pero el costumbrismo como inclinación extensa y generalizada se inicia en el romanticismo" (Fernández-Arias Campoamor 1952, 56). Kohut, too, points to the significance of *Costumbrismo* for several other literary currents: "Die Abgrenzung zwischen Romantik, Realismus und Naturalismus gestaltet sich schwierig. […] Die Problematik wird durch den sogenannten *Costumbrismo* zusätzlich kompliziert, der wie in Spanien zwischen Romantik und Realismus steht. Zum Realismus gehört die Zuwendung zur Gesellschaft, zur Romantik die häufig idyllisierende Perspektive. […] Wichtiger als der *Costumbrismo* als eigenständige literarische Richtung ist die entsprechende Einfärbung zahlreicher realistischer bzw. Naturalistischer Romane. So gab der Chilene Alberto Blest Gana seinem Roman *Martín Rivas* (1862) den Untertitel *Novela de costumbres político-sociales*, der Argentinier Lucio Vicente López seinem Roman *La gran aldea* (1884) den Untertitel *Costumbres bonaerenses*" (Kohut 2016, 196).

[235] Alternative terms are *novela del gaucho*, *novela urbana*, *novela indianista*, *novela antiesclavista*, and *novela policial*. In some cases, the alternative formulations do not involve differences in the meaning of the terms (this is assumed for *novela gauchesca* and *novela del gaucho* and for *novela de la ciudad* and *novela urbana*). In other cases, there are slight differences in the meaning. *Novela criminal*, for example, is more general than *novela policial*, and was therefore preferred here. *Novela abolicionista* was preferred over *novela antiesclavista* because it is the term that was in use historically. The term *novela indigenista* was preferred over *novela indianista* because it is more neutral. The *novela indianista* refers to romantic works that re-evaluate the historical past before the conquest of the Americas (Meléndez 1961; Rössner 2007, 144).

[236] An example of an assignment made in the running text is: "De este episodio tomó apuntes el joven subteniente [Heriberto Frías], dándole base para construir una novela histórica a la que dió el nombre de *Tomochic*" (Fernández-Arias Campoamor 1952, 83). In Dill's literary history of the Spanish-American novel, several subchapters are entitled with subgenre labels so that the texts mentioned in them can be attributed to these subgenres. The chapter

Scope	Title	Editor / Author	Year
Spanish-American literature	Geschichte der lateinamerikanischen Literatur im Überblick	Dill, Hans-Otto	1999
Spanish-American novel	La novela hispanoamericana (hasta 1940)	Gálvez, Marina	1990
Spanish-American novel	Proceso y contenido de la novela hispano-americana	Sánchez, Luis Alberto	1953
Nineteenth century Spanish-American novel	Evolución de la novela hispanoamericana en el siglo XIX	Varela Jácome, Benito	2000
Spanish-American romantic novel	La novela romántica en Hispanoamérica	Suárez-Murias, Marguerite C.	1963
Nineteenth century Argentine novel	The Argentine novel in the nineteenth century	Lichtblau, Myron I.	1959
Argentine novel (1838–1872)	Como crecen los hongos. La novela argentina entre 1838 y 1872	Molina, Hebe Beatriz	2011
Cuban novel	Tendencias de la narración imaginativa en Cuba	Remos y Rubio, Juan J.	1935
Mexican novel	Novelistas de Mejico. Esquema de la historia de la novela mejicana (De Lizardi a 1950)	Fernández-Arias, Campoamor, José	1952
Nineteenth century Mexican historical novel	The Mexican historical novel. 1826–1910	Read, John Lloyd	1939
Spanish-American naturalistic novel	El lado oscuro de la modernización: estudios sobre la novela naturalista hispanoamericana	Schlickers, Sabine	2003
Cuban naturalistic novel	El Naturalismo en la novela cubana	Molina, Sintia	2001

Table 9. Literary historical sources for the assignment of subgenres.

because many novels are rather described in their relationship to a certain subgenre and are also evaluated as mixtures or deviations.[237] Regarding the discussion of novelistic subgenres, many literary-historical works tend to focus on the individuality of the works. Subgenres provide a frame for the description of groups of works. However, they are rarely understood as strict classes and more often as anchor points that help to analyze and represent a complex overall novelistic production in an ordered way.[238]

The subgenre labels assigned to the novels by literary historians have a different status than those occurring explicitly in the titles and other paratexts of the novels and those that are or signaled implicitly in the texts. Literary historical labels do not represent a contemporary perspective, and the agents who decided on the labels are different. They are scholars of the twentieth and twenty-first centuries aiming to provide systematic perspectives on the novelistic production of the nineteenth century and not authors, editors, or contemporary critics. Nevertheless, the labels do not behave differently as a whole. Like every author, editor, or contemporary might use the labels in a slightly different manner, also the approaches of scholars can differ. Definitions of subgenres and criteria for the composition of the corpus are more often given in studies that concentrate on one subgenre of the novel.[239] In general literary histories that are dedicated to a whole range of genres and subgenres, it is usually not explicitly discussed how the works were assigned to the subgenres. Comparative studies concerned with several types of subgenres are

---

"Der Roman der Romantik", for example, is subdivided into "Der politische Roman", "Der historische Roman", "Der indianistische Roman", "Der kubanische negristische Roman", and "Der sentimentale Roman". Works mentioned in the chapter "Der politische Roman" are, for example, "Amalia" (1855, AR) by José Mármol or "Clemencia" (1869, MX) and "El Zarco" (1901, MX) by Ignacio Manuel Altamirano (Dill 1999, 125–139).

[237] For example, Dill mentions Emilio Rabasa's novels in the chapter on the naturalistic novel but designates them as anti-naturalistic (Dill 1999, 170). In her work on the Spanish-American naturalistic novel, Schlickers dedicates her own subchapter to each of the novels that she included in her corpus. In these detailed discussions of the works, she reasons about how each of the novels is in accordance with the criteria that she set up for a novel to be naturalistic and, in some cases, comes to the conclusion that they are not, e.g., for the novel "León Zaldívar" (1888, AR) by Carlos María Ocantos: "Resulta que *León Zaldívar* no es una novela naturalista, sino una mezcla entre novela rosa/folletinesca y costumbrista; Lichtblau […] califica la novela de 'happy combination of romantic and realistic elements'. […] A nivel de la expresión, la distancia respecto a la poética naturalista se marca por los frecuentes comentarios del narrador que marca su *hic et nunc* y coincide ideológicamente tanto con el autor implícito como con el protagonista idealizado […]. A pesar de una escritura por lo general 'realista', no se concretan ni el tiempo de la historia […], ni se citan nombres –por ejemplo de los políticos que se critican. Así, la novela gana en dimensión alegórica lo que pierde en valor referencial, facilitanto así la transmisión y recepción masiva de la intención de sentido: la crítica del materialismo, la reivindicación de la sincera práctica de la religión católica y la idealización de la mujer abnegada y sumisa, para terminar con una lección moralizante: […]" (Schlickers 2003, 200). In this way, Schlickers checks the novels that can be provisionally assigned to the naturalistic current because of their theme or certain generic signals against the more strict formal criteria that she set up for the subgenre.

[238] This can be seen, for example, in the monograph "The Mexican historical Novel" authored by Read. In order to give a comprehensive overview of the Mexican nineteenth-century fiction that can be considered historical, he also includes cases that are only historical on certain levels of the narration, e.g.: "López Portillo y Rojas pronounced Altamirano's *Clemencia* the best Mexican novel of its time. […] The setting of the story is the region around Guadalajara in December, 1863, the year of the French occupation of Mexico. The historical material serves only as a frame for the actions of two officers in the army of the republic […]. A conflict developed over the affections of a beautiful woman, with a none too pleasant result" (Read 1939, 164–166) or "Martínez de Castro's novel *Eva*, published in 1885, though not purely historical in nature, has enough of historical background to justify its inclusion in this study" (Read 1939, 256).

[239] See, for example, Rivas (1990, 121–154) and Schlickers (2003, 27–46).

rare.[240] When no definitions of the subgenres are given, it can only be hypothesized how the assignments come about: they might be based on explicit historical labels, on previous assignments made by other literary historians, or on background knowledge and reading experience. As a result, the focus here is not on how the subgenre terms are defined in each case but the fact that they are signaled by literary historians. Together with the explicit and implicit signals found in the texts of the novels themselves, the subgenres emerge as categories that are collectively defined, and this includes a certain fuzziness.

3.2.3.6 A Discursive Model of Generic Terms

Returning to the levels into which the subgenre labels – explicit, implicit, and literary-historical ones – are sorted in the summary, some more remarks are to be made. Regarding the relationship to the model proposed by Raible, it is evident that his model is a semiotic one in the linguistic sense of the term. Raible's model, designed for literary and also non-literary genres, covers general aspects of the communication situation, the content and structure of the message, the medium, and the linguistic representation. With the level concerning the relationship between the text and reality, he addresses a point specifically relevant for literary texts. However, the aspects of the literary currents and cultural and linguistic identities of the texts are not covered by him.[241] A model similar to Raible's is the one developed by Jean-Marie Schaeffer, who also starts from the assumption that a literary work is a complex semiotic object and that generic terms can refer to different levels of this object. Broadly, Schaeffer distinguishes between the communicative act ("L'acte communicationnel") and the realized discursive act ("L'acte discursif réalisé"). Raible's level "Kommunikationssituation" overlaps with Schaeffer's "L'acte communicationnel" and the "Objektbereich" with the "L'acte discursif réalisé". The other levels defined by Raible can also be associated with Schaeffer's two main levels. That is, also in Schaeffer's semiotic approach, the levels that are named "identity" and "current" here are not included in the core model. Nevertheless, Schaeffer discusses these aspects as an aside:

> Parmi les noms de genres que j'ai collectés, certains se réfèrent cependant à des déterminations qui sont irréductibles aux cinq niveaux de l'acte verbal que je viens de distinguer. J'ai indiqué plus haut que le modèle de la communication dont je me servais ne tenait pas compte du contexte, du lieu et du temps. Or, il existe de nombreux noms de genres qui sont composés à l'aide de déterminants de lieu ou de temps. Ainsi des termes comme tragédie élisabéthaine, tragédie classique, roman antique, sonnet baroque, etc., délimitent des traditions dans le temps, c'est-à-dire se réfèrent à des genres historiques au sense le plut fort du terme. […]

---

[240] An example of a study where several different subgenres of the novel are defined is Molina (2011). It is not a comparative study in a strict sense because rather than comparing the different subgenres, her goal is to describe and systematize the whole novelistic production in Argentina between 1838 and 1872. However, Molina finds that all the novels can be classified into one or several of four main types: "novelas históricas", "novelas políticas", "novelas socializadoras", and "novelas sentimentales" (246–386).

[241] In the description of the level "Objektbereich", he mentions the Old French terms "matière de Bretagne" and "matière de Rome" that designate legendary material concerned with the history of Brittany and Rome, but there is no own level for identity-related (culturally specific or national) texts (Raible 1980, 343).

> La modification selon le lieu se rencontre sous deux formes. La première est celle de la spécification d'un genre selon les communautés linguistiques, mais à l'intérieur d'une sphère culturelle historiquement plus ou moins solidaire. Le phénomène est trés répandu en Occident: nous parlons ainsi de *l'épopée grecque* et de *l'épopée romaine*, du *roman français* et du *roman anglais* [...]. (Schaeffer 1983, 117–118)

Following Schaeffer's explanations, generic terms referring to literary currents could be subsumed under the temporal context, and terms related to the linguistic and cultural identity under the spatial context. Schaeffer is aware that these aspects are not covered by his model, but also that most of the generic terms are not reducible to a single discursive level, neither regarding aspects of the verbal message ("the text") nor contextual factors:

> L'existence de se modifications temporelles et spatiales des noms de genres pose la question de la contextualization historiques des déterminations génériques, question que le schéma communicationnel que j'ai retenu occulte [...] elle ne peut évidemment que renforcer la conclusion qu'imposait déjà la prise en compte de la multidimensionnalité du message verbal, à savoir que les noms de genres, loin de déterminer tous un même objet qui serait « le texte » ou même un ou plusieurs niveaux invariants de ce texte, sont liés, selon les noms, aux aspects les plus divers des faits discursifs. (Schaeffer 1983, 119)

This is not only true for the terms associated with the levels of *theme* and *mode* here, as could be seen in table 5 above, but also for the groups of *current* and *identity*. Generic terms referring to literary currents do not only localize the subgenre temporally and historically. They also entail preferences regarding the themes of the novels as well as stylistic properties. Similarly, the terms subsumed under *identity* do not only relate to the cultural, geographical, and linguistic localization of the novel as a discursive object but can also point to thematic aspects of its content when American, Cuban, Argentinian, or Mexican matters are treated. The assignment of terms to the levels thus sets focuses for their analysis but is not to be understood as exhaustive or exclusive. This is also the reason for the reduced modal level here compared to the models of Raible and Schaeffer. With the modal subtypes of *intention*, *attitude*, *reality*, *medium*, and *representation*, the model used here contains aspects of the communicative situation and the textual message that revealed themselves to be relevant for the nineteenth-century Spanish-American novels analyzed here because they are implied by explicit subgenre labels. Figure 10 situates the categories of subgenre labels used in the encoding model of the bibliography here in a more general communicative model mainly based on the one proposed by Schaeffer.

The figure shows that the subtypes of *intention*, *attitude*, *reality*, *medium*, and *representation* can be grouped under the aspect of how the literary text is communicated and presented (*mode*). The category of *theme* stands for what is communicated, and the catetories of *identity* and *current* point to the context in which something is communicated.[242] On the one hand, the model used for the bibliography and the corpus here has to be understood as an application, adaptation,

[242] In the figure, the levels of *medium* and *syntactic* are connected because, in the context of novels, generic terms that refer to a medium are usually to be understood as indications of how the text is represented structurally, e.g., in the form of letters in epistolary novels. Many of the labels relating to medial aspects are more vague, for example, those originally referring to painting and drawing: "cuadros", "bocetos", "esbozos", and "impresiones" describe figuratively how the text is represented.

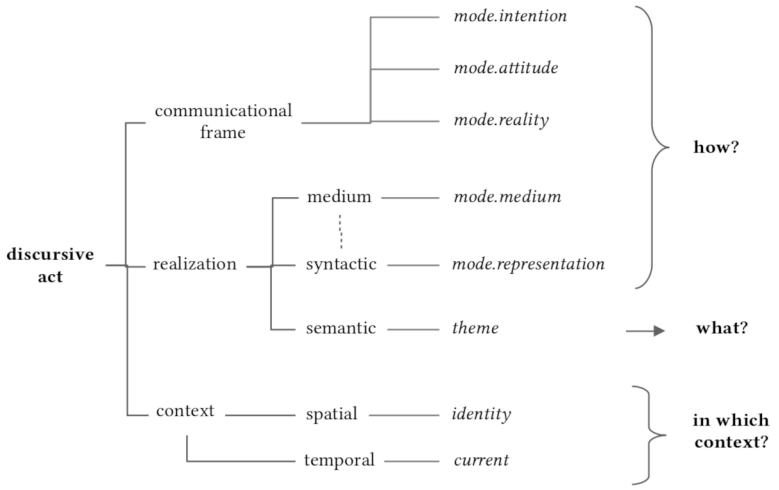


Figure 10. Kinds of subgenres in the context of a discursive model.

and selection of the general semiotic models of (literary) genres. On the other hand, it is a bottom-up approach. Only those aspects of the general models that occur in the generic signals and literary-historical assignments of genres to the works in the bibliography and corpus of nineteenth-century Cuban, Argentine, and Mexican novels were selected. It is thus an empirically driven discursive model for generic terms of the novel in a specific cultural-geographical and historical context. It would be interesting to see which levels of the general models are activated for other corpora of the novel in order to find out which kinds of subgenre labels are typical for the novel in general and which ones are determined by contextual factors. When one looks at the relevance of the different levels for the whole set of novels, it can be observed that the levels of *theme* and *current* are the two main levels used in literary-historical approaches. Themes are also frequently involved in explicit historical subgenre labels, but regarding the quantitative relevance, only the *novela histórica* and the *novela de costumbres* stick out. The level of *identity* is very present among the top most frequent explicit labels but is only indirectly discussed in the critical literature.[243] The level of *mode.representation* gains quantitative relevance because of the distinction between novels explicitly labeled as "novela" and those that are not. More specifically, there are also many "episodios" and "memorias". The level of *mode.reality* plays an important role because many of the terms that are thematic also point to the relationship between text and reality ("novela histórica" and "leyenda", for instance). The other modal categories (*mode.intention*, *mode.attitude*, *mode.medium*) are less important in terms of numbers but are also present. Intention and attitude play a role in terms that are also thematic, for example, in political novels. Terms related to the medium are creatively used by the authors in a range of different generic labels for the novels.[244] Table 10 contains an alphabetically ordered list of all the different subgenre labels found for the novels in the bibliography and the corpus. In the table, it is indicated to which levels of the model the subgenre labels were assigned.[245]

[243] Of course, there is research on *the Cuban/Argentine/Mexican/Spanish-American* novels, and also "national novels" (Sommer 1993) are discussed, but these approaches are usually not related to explicit generic terms.

[244] See also chapter 4.1.5, where a series of charts that display the distribution of subgenre labels is included.

[245] The whole table is also available at https://github.com/cligs/data-nh/blob/master/corpus/bibliography-subgenre-labels/overview-subgenres.csv. Accessed March 28, 2020.

Table 10. Set of subgenres occurring explicitly or implicitly in the bibliography.

Subgenre label	Kind(s)	Supplement	Explicit occurrence
apuntamientos	mode.medium, mode.representation	-	yes
apuntes	mode.medium, mode.representation	-	yes
auto-novela	mode.reality	-	yes
Bildungsroman	theme	-	no
boceto	mode.medium, mode.representation	-	yes
bosquejo	mode.medium, mode.representation	-	yes
capricho	mode.representation	-	yes
cinematógrafo	mode.medium, mode.representation	-	yes
comedia de carácter	theme, mode.intention	comedia	yes
confesiones	theme, mode.representation	-	yes
contornos	mode.representation	-	yes
croquis	mode.medium, mode.representation	-	yes
crónica	theme, mode.representation	novela histórica	yes
cuadros	mode.medium, mode.representation	novela de costumbres	yes
cuento	mode.representation	-	yes
drama	mode.representation	novela romántica	yes
elegía	theme, mode.medium, mode.attitude	-	yes
ensayo	mode.representation	-	yes
entretenimientos	mode.intention	-	yes
episodios	mode.representation	novela histórica	yes
epopeya	theme, mode.representation	novela histórica	yes
esbozos	mode.representation, mode.medium	-	yes
escenas	mode.medium, mode.representation	-	yes
estudio	mode.representation, mode.intention	novela social, novela realista, novela naturalista	yes
fragmentos	mode.representation	-	yes
historia	mode.representation	-	yes
impresiones	mode.representation	-	yes
juguete	mode.intention	-	yes
Künstlerroman	theme	-	no
lecturas	mode.intention	novela didáctica	yes

Subgenre label	Kind(s)	Supplement	Explicit occurrence
leyenda	theme, mode.reality	novela histórica, novela romántica	yes
medallones	mode.representation	-	yes
memorias	mode.representation (theme)	-	yes
narración	mode.representation	-	yes
notas	mode.medium, mode.representation	-	yes
novela	mode.representation	-	yes
novela abolicionista	theme (mode.attitude)	novela social	yes
novela americana	identity (theme)	-	yes
novela analítica	mode.representation (mode.intention)	-	yes
novela andaluza	identity (theme)	-	yes
novela anecdótica	mode.representation	-	yes
novela argentina	identity (theme)	-	yes
novela azteca	identity (theme)	(novela mexicana)	yes
novela biográfica	theme	-	no
novela bonaerense	identity (theme)	(novela argentina)	yes
novela camagüeyana	identity (theme)	(novela cubana)	yes
novela científica	theme, mode.reality	-	(yes)
novela clasicista	current (theme)	-	no
novela cómica	mode.intention (mode.attitude)	novela humorística	(yes)
novela contemporánea	theme, mode.reality	novela social and/or novela política	yes
novela corta	mode.representation	-	yes
novela criminal	theme	-	(yes)
novela criolla	identity (theme)	(novela americana)	yes
novela cubana	identity (theme)	-	yes
novela curiosa	mode.intention	-	yes
novela de actualidad	theme, mode.reality	novela contemporánea, novela social and/or novela política	yes
novela de aventuras	theme	-	(yes)
novela de costumbres	theme (current)	(novela social)	yes
novela de crímenes	theme	novela criminal	yes
novela de familia	theme	novela social	no
novela habanera	identity (theme)	(novela cubana)	yes
novela de horrores	mode.intention	-	yes
novela de la ciudad	theme	novela social	no
novela de misterio	theme, mode.reality	-	no
novela de propaganda	theme, mode.intention	novela política and/or novela social	yes
novela de Tabasco	identity (theme)	(novela mexicana)	yes

Subgenre label	Kind(s)	Supplement	Explicit occurrence
novela de viajes	theme	-	(yes)
novela didáctica	theme, mode.intention	novela social	yes
novela documentaria	theme, mode.representation	novela social and/or novela política	no
novela doméstica	theme	novela social	yes
novela en acción	theme	novela de aventuras	yes
novela enciclopédica	theme, mode.intention	novela didáctica	yes
novela epistolar	mode.medium, mode.representation	-	yes
novela espiritista	theme, mode.reality	novela científica	yes
novela fantástica	theme, mode.reality	-	yes
novela festiva	mode.attitude	-	yes
novela filosófica	theme, mode.representation	-	yes
novela franco-argentina	identity (theme)	(novela argentina)	yes
novela gauchesca	theme	-	no
novela histórica	theme, mode.reality	-	yes
novela humorística	mode.intention (mode.attitude)	-	yes
novela india	identity (theme)	-	yes
novela indigenista	theme	-	no
novela jurídica	theme	novela criminal	yes
novela kantabro-americana	identity (theme)	(novela americana)	yes
novela mexicana	identity (theme)	-	yes
novela militar	theme	novela histórica	yes
novela mixteca	identity (theme)	(novela mexicana)	yes
novela modernista	current (theme)	-	no
novela moralista	theme, mode.intention	novela social	no
novela nacional	identity (theme)	-	yes
novela naturalista	current (theme)	novela realista, novela social	yes
novela original	identity	-	yes
novela patriótica	theme, identity	-	yes
novela picaresca	theme	-	no
novela policial	theme	novela criminal	yes
novela política	theme, mode.attitude	-	yes
novela popular	theme	novela social	yes
novela porteña	identity (theme)	(novela argentina)	yes
novela psicológica	theme, mode.representation	-	no
novela realista	current (theme)	novela social	yes
novela regional	theme, identity	-	yes
novela romana	identity (theme)	-	yes
novela romántica	current (theme)	-	no
novela satírica	mode.attitude	-	yes

Subgenre label	Kind(s)	Supplement	Explicit occurrence
novela sentimental	theme	-	yes
novela siciliana	identity (theme)	-	yes
novela social	theme (mode.intention)	-	yes
novela suriana	identity (theme)	-	yes
novela tapatía	identity (theme)	(novela mexicana)	yes
novela verista	current (theme)	novela realista, novela social	no
novela yucateca	identity (theme)	(novela mexicana)	yes
panorama	mode.medium, mode.representation	-	yes
perfiles	mode.representation	-	yes
páginas	mode.medium, mode.representation	-	yes
recuerdos	mode.representation, theme	-	yes
reflexiones	mode.representation	novela filosófica	yes
relación	mode.representation	-	yes
relato	mode.representation	-	yes
reseña	mode.attitude, mode.representation	-	yes
romance	theme, mode.representation	novela sentimental, novela histórica	yes
silueta	mode.representation	-	yes
tradición	theme, mode.reality	-	yes
tragedia	theme, mode.representation	-	yes

The generic terms consist of nouns (“Bildungsroman”, “croquis”, “episodios”) or nouns that are characterized further by attributes (“novela argentina”, “novela de costumbres”, “novela histórica”). One or several kinds of subgenre labels that are considered the most important are given for each term. Kinds in parentheses mean that these are also possible assignments but that they were not encoded in the bibliography because they were not considered crucial and in order not to mix the categories too much. In general, it is assumed here that generic terms are complex signs and that many of the terms refer to different levels of discourse. For example, the term “novela naturalista” refers to the literary current of Naturalism, but also to the themes preferred by that current, i.e., social topics, including the account of the situation of outsiders and lower classes, and tabooed subjects such as adultery or prostitution. Furthermore, it refers to certain representational techniques used in naturalist novels and so on. The generic terms are loaded semantically through the characteristics of the works that carry the terms over time. However, it is also assumed here that there are differences in how relevant the different levels are for the terms and that it is possible to determine primary levels. For example, the term “memorias” is assigned to the level of *mode.representation* here and only secondary to the level of *theme*. Even though memories can imply certain themes (a life story, for example), the themes are not very

specific and the presentation of the text in the form of memories, looking back on what was experienced and is remembered, is more important. Other examples are the terms referring to literary currents, for instance, "novela realista" or "novela romántica". Although they are also connected to certain themes, this aspect is considered subordinate here. The same was decided for terms referring to a certain city, region, nation, or people, such as "novela habanera", "novela de Tabasco", "novela mexicana", or "novela azteca", for which the level of *identity* is taken as the primary level and thematic aspects as secondary. That way these subgenres are differentiated from the ones that are primarily thematic, for example, "novela sentimental", "novela histórica", or "novela de costumbres".

The third column in the table indicates subgenres that are implied by the term in the first column. For example, the term "crónica" often includes a historical theme, and the terms "novela naturalista" and "novela realista" often include social themes. The supplements can serve to normalize terms, to generalize them, and to make other levels that are implied by such terms explicit. The supplements are not automatically assigned, though. They are assigned depending on whether they make sense in the individual case. Some of the possible supplements are given in parentheses in the table. These are not assigned in the bibliography and the corpus to make the subgenres more distinguishable. The "novela de costumbre", for example, could also be understood as a form of social novel but would then have the same thematic label as naturalistic and realist novels because the term "social" has so many facets.

The last column of the table indicates whether the generic term occurs explicitly in the bibliography or not. A "yes" in parentheses means that the term occurs, but only in a normalized explicit form and not verbatim. Examples of terms that never occur explicitly in the bibliography but are often used by critics to characterize the novels are "novela gauchesca", "novela indigenista", and "novela romántica". Others that occur primarily in the explicit form are, for example, "novela jurídica" or "novela contemporánea".

The goal of the systematization of subgenre labels by normalizing terms and assigning them to different levels of discourse is to put the whole range of generic terms associated with the novels through explicit or implicit paratextual elements and critical assignments in a certain order to be able to analyze groups of the novels quantitatively. Effectively, the set of generic terms with which novels are designated is open and endless, full of variants and nuances, as already the terms in the table show. If one would take the terms as they are, in many cases, it would be difficult to form groups. For instance, there are "novelas jurídicas", "novelas policiales", and "novelas de crímenes", which are subsumed under the term "novela criminal" here because it is assumed that these terms can be interpreted as referring to the same subgenre. On the other hand, it would be challenging to analyze subgenres that are defined on entirely different levels of discourse together, for example, a "novela romántica" compared to a "novela mexicana" or a "novela fantástica". The whole structure of subgenres is not organized hierarchically and it does not consider historical change.

To summarize, the assignment of subgenres to the novels discussed in this section follows a strategy that is, for one thing, historically oriented, because all the explicit labels occurring in the titles of the novels in the bibliography (and also in other paratexts for the novels in the corpus) are collected, and furthermore, also implicit signals are evaluated. Explicit signals are normalized in order to make them comparable, and the interpretation of implicit signals requires

prior knowledge about the subgenres, but all the steps are documented to make the process transparent. In general, historically adequate terms are preferred over ahistorical critical ones. Then again, the information available from historical signals is complemented by subgenre assignments made by literary critics in order to also open up those novels to an analysis of subgenres that would otherwise have to be considered *general fiction* because there are no clear signals available from the paratexts. However, because the sources of the generic information are encoded in detail for each entry in the bibliography (and corpus), it is possible to conduct analyses only on one or the other kind of information or to backtrack the statements about subgenres when the results of a combined analysis are interpreted.

No previous selection of certain types of subgenres was made. Instead, within the frame of the general, formal working definition of the novel provided in chapter 3.1.1.7 above, all kinds of generic information were collected. The information was then systematized in summarizing terms, following a discursive model based on other general, semiotic models of generic names. That way, it was captured which (types of) subgenre labels are the most frequent in the bibliography. Regarding explicit labels, the distinction between works carrying the label "novela" and those that do not is quantitatively relevant. Regarding thematic aspects, "novelas históricas" and "novelas de costumbres" are frequent. Furthermore, there is a considerable group of novels that have a label related to the cultural and linguistic identity of the text ("novela original", "novela mexicana", etc.) as opposed to a bigger group of novels that does not have such a generic signal. When also implicit signals and literary-historical labels are considered, the most frequent kinds of subgenre labels are thematic or refer to the literary current(s) of the texts. Even if other kinds of subgenre labels are not so frequent, for example, labels concerning the representational mode of the text (other than the general term "novela") or labels pertaining to the groups of *intention*, *attitude*, *reality*, or *medium*, this generic information is still valuable as background information when other subgenres are analyzed, and when the results need to be interpreted. The great variety of subgenre labels found for the works in the bibliography shows how open the genre *novela* is in general and how extensive the network of generic references is. Even labels referring to other major genres (e.g., "drama", "tragedia", "comedia", "epopeya", etc.) are used to mark novels. On the other hand, only a few subgenres were very frequent in the period and countries examined in this study.

3.3 Text Corpus

Based on the general information about the Argentine, Cuban, and Mexican novels published between 1830 and 1910 that was collected for the digital bibliography Bib-ACMé, a corpus of digital full texts called Conha19 ("Corpus de novelas hispanoamericanas del siglo XIX") was prepared. The resulting text collection is aimed to be used for digital, quantitative literary analysis. While there is a long tradition of preparing and using digital corpora for linguistics,[246]

[246] The creation and usage of (digital) corpora in linguistics is the subject of a whole subdiscipline, corpus linguistics. See Andresen and Zinsmeister (2019) for a recent textbook on the topic. A comprehensive handbook including information about the history of the discipline, the compilation and different types of corpora, preprocessing, their use and exploitation, including statistical and computational methods, is Lüdeling and Kytö (2008). Some of the differences between linguistic and literary corpora are that the former also focus significantly on spoken language

the development of best practices for creating digital literary corpora is still underway. Of course, the use of corpora for literary scholarship also has its history. However, it is traditionally more closely related to scholarly textual editing and the preparation of smaller datasets as a basis for qualitative interpretation.[247] Recently, also the creation of bigger corpora of digital literary texts suitable for quantitative analyses has been reflected.[248] Hoover, Culpeper, and O'Halloran (2014), for example, emphasize how valuable the methods developed in corpus linguistics are for the digital study of literary texts, as well. They build and analyze corpora of character speech from dramatic texts, novels, and lyric poems. In the project CLiGS, the context in which this dissertation is elaborated, we have developed small prototypical digital collections of literary texts in Romance languages (French, Spanish, Italian, and Portuguese). We concentrated on practical aspects, including the compilation of texts, the collection of metadata, text encoding, publishing, archiving, and how to encourage reuse (Schöch et al. 2019). Another example of a project practicing and reflecting the creation of corpora for digital literary analysis was the COST Action "Distant Reading for European Literary History", which involved creating a diachronic, multilingual corpus of novels from 1840–1919 called "The European Literary Text Collection" (ELTeC) (Odebrecht et al. 2021).

In this chapter, these developments are taken into account. The chapter serves to clarify questions of text selection, text treatment, metadata, and text encoding, the assignment of subgenre labels, the creation of derivative corpus formats, and its publication. The chapter is organized as follows: How the novels were selected for the corpus and which sources were used is described in chapter 3.3.1 below. In chapter 3.3.2, it is explained how the full texts were obtained from digitized images and how the quality of the resulting texts was checked. This step is also important for full texts that were directly included from other sources. Just as for the bibliography, it was decided to encode the texts in XML, according to the standard of the

---

and that the principles of representativeness and sampling play a major role. Usually, the goal of linguistic corpora is to represent language (or a specific subsystem of language) as a whole in order to be able to get generalizable results when the corpora are analyzed. At the same time, except in extreme cases, it is not possible to record all the linguistic utterances that are relevant to a certain domain. Literary corpora, on the other hand, put their emphasis on written texts. In the literary domain, it is also often not possible to build a corpus comprising the whole population of textual production (because sources are lost or unavailable for other reasons). However, the gap between a corpus and the target domain is usually smaller. It is, for example, possible to compile the complete works of an author. As a consequence, regarding the creation and exploitation of corpora, literary studies can build on many of the research findings achieved in corpus linguistics, but there is still a need to adapt them.

[247] One reason for the close relationship between literary corpora and scholarly editing is that a literary work is an abstract notion: a work is not necessarily manifest in a single document. There may be subsequent versions of works that can be viewed and compared to get a critical version serving as a base text, or a decision can be made for a specific historical version that is prepared with scholarly editorial methods. For the history of editorial scholarship in the philologies and different types of literary scholarly editions, see, for example, Sahle (2013, 107–224). Principles for creating corpora for the study of literary genres are discussed in Hempfer (1973, 128–136) and Zymner (2003, 122–139; 2010, 23–25).

[248] From a more general perspective, the creation of *data collections* (including collections of spoken language and text, but also of movies, pictures, musical pieces, or other artifacts) for digital analysis is examined by Schöch (2017a). Criteria for reviewing digital text collections in general have been developed by the Institute for Documentology and Scholarly Editing (IDE) (Henny and Neuber 2017). Nevertheless, Gius, Krüger, and Sökefeld (2019, 165) point out that the preparation of corpora specifically for literary studies is still hardly reflected upon and that this entails certain risks, especially in the case of larger, digital corpora (e.g., unwanted correlations in the data).

Text Encoding Initiative (TEI). That way, structural information such as chapter divisions or headings could be added to the texts, and the metadata for the novels could be included in each file. The kind of metadata that was collected and the model of text encoding that was applied are presented in chapter 3.3.3. A special focus is given to the assignment of the subgenre labels, which is discussed in chapter 3.3.4. Chapter 3.3.5 serves to present two derivative corpus formats (plain text and a linguistically annotated version) and to explain the publication strategy for the corpus. The contents of the corpus are presented in chapter 4.1 on "Metadata Analysis", where they are related to those of the whole bibliography of novels.

3.3.1 Selection of Novels and Sources

A special challenge in creating a corpus for genre analysis is the so-called chicken-and-egg problem: a genre can only be analyzed in the form of individual texts attributed to it, but a previous definition of the genre is needed for selecting these texts. From a genre theoretical perspective, two approaches are proposed to handle the problem: following the inductive approach, a corpus is built without a previous working definition of the genre, and texts are, for example, chosen because of historical labels, with the drawback that not all of the relevant texts necessarily carry the same label and that the meaning of the labels is subject to change. Another possibility is a deductive approach starting from a general definition of the genre, which serves to organize the historical material. It has the advantage that the definition is clear, but the disadvantage is that it is not necessarily historically adequate. According to Zymner, both procedures lead to contradictions, and in practice, mixed approaches are more common (Zymner 2010, 23–24). The strategy for creating the corpus at hand can also be characterized as a mixed one. It starts from a general, formal working definition of the novel (see chapter 3.1.1.7 above) which is the deductive aspect, but the subgenres of the novel are not previously defined formally. Rather, they are established based on explicit historical labels and implicit signals as well as a range of literary-historical assignments. For a large digital corpus, this strategy has the advantage that the elements of a general definition of the novel are easier to check than aspects of specific definitions of subgenres of the novel. In addition, that way, the analysis of subgenres is not predetermined theoretically.

Based on the data collected for Bib-ACMé, a corpus of 256 novels was created. The novels that are included in the bibliography were, in principle, all eligible for the text corpus, as well, because the selection criteria as outlined in chapter 3.1 above were applied to the bibliographic entries. Nevertheless, the novels were checked again before they were taken over from a source because insight into the full texts allowed for a stricter application of the selection criteria. For example, the extent of the text could be measured in the number of words instead of pages. Furthermore, the text and not just the work title could be checked for signals of fictionality.

Two main factors determined which texts from the bibliography were selected to be included in the text corpus: first, characteristics of the texts influencing the balance of subgroups in the corpus, and second, practical matters regarding the availability of the texts. As to the first factor, the corpus seeks to assemble several large groups of novels belonging to certain subgenres to be able to analyze these quantitatively. Therefore, texts pertaining to subgenres that were common in the nineteenth century, namely the thematically oriented subgenres of *novelas históricas*, *novelas*

de costumbres, and *novelas sentimentales*, and subgenres related to different literary currents, i.e., *novelas románticas*, *novelas realistas*, and *novelas naturalistas* were preferred. On the other hand, the aim was to create a corpus reasonably balanced by country, publication date, and author. In this context, *balanced* does not mean achieving an equal number of texts from all the authors, the three countries, and the years between 1840 and 1910. Such a corpus would be an artificial construct and hardly possible to realize because the bibliographic data shows that the number of novels published in Argentina, Cuba, and Mexico in the different phases of the nineteenth century differs significantly. The number of novels published by individual authors also varies greatly. Rather, the aim was to build a corpus that is balanced under the conditions of the population. For example, when having the choice to include either three novels written by the same author in the same decade or three novels written by different authors and published in different decades, the second option was preferred.[249] The second factor influencing the shape of the corpus – the availability of the texts – also affected the first one. The overall availability and the state of digitization of the texts varies for the three countries, the different points in time, and for the individual authors. In general, older texts are rarer, Cuban novels harder to obtain than Argentine and Mexican ones, and works of less canonized authors more difficult to procure. Moreover, the novels that were already available in a digital full-text format belong to a broad range of subgenres. They were all included so that the corpus does not exclusively contain novels of the subgenres that are analyzed in more detail. The selection of texts was prioritized according to the following practical availabilities to keep the creation of the corpus feasible:

- novels obtainable in a digital full-text format (either plain text or text with markup),
- novels obtainable as digitized images (in PDF format or image files),
- novels obtainable as print editions suitable for digitization.

Whereas the number of texts in the first group is rather low, hundreds of novels are available as digitized images and even more as print editions. The size of the corpus was, therefore, mainly limited for reasons of time and cost. With more resources, the digital full texts of more novels could be extracted, and it is to be hoped that this task will be embraced at the institutional level.[250]

The majority of the novels in the corpus (81.3 % or 208 texts) was collected from digital sources and only about one fifth (18.8 % or 48 texts) from print sources.[251] The print sources were mainly used to complement the corpus in terms of subgenres. Regarding the file formats of the sources for the texts in the corpus, one sees that only about one-third of the novels (32 % or 82 texts) was available in a full-text format whereas more than three-thirds (68 % or 174 texts) were obtained from image files. Obviously, all the print sources were converted to digital images, but also

[249] The relationship between the population of novels, as approximated with the bibliographical database and the texts in the corpus, is outlined in chapter 4.1 below.

[250] A better state of digitization would also allow using specific editions of the novels, e.g., only first or last editions. For this dissertation, no specific strategy for selecting editions could be followed because the main issue was having access to at least one edition of a novel in digital format.

[251] In the script and data repositories, there are statistical charts related to these data, created with the following script, which was also used to create the following figures in this chapter: https://github.com/cligs/scripts-nh/blob/master/corpus/corpus-sources.py. The metadata about the sources of the novels is available at https://github.com/cligs/data-nh/blob/master/corpus/metadata_sources.csv. The resulting charts can be downloaded from the folder https://github.com/cligs/data-nh/tree/master/corpus/corpus-sources. Accessed January 29, 2023.

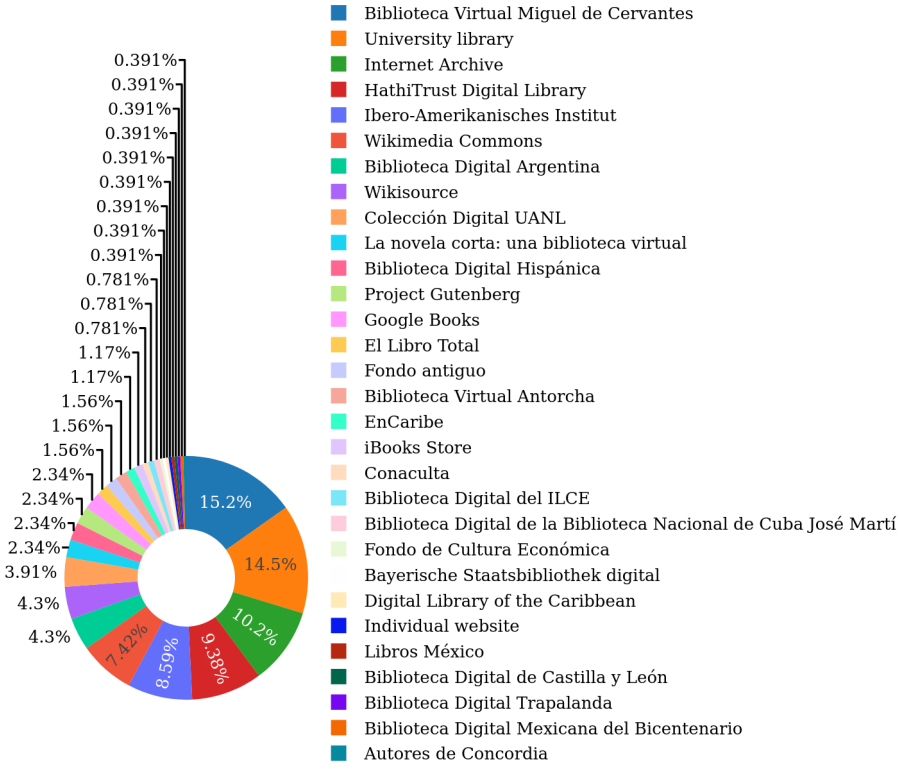


Figure 11. Sources by institution.

the majority of the digital sources were only accessible as image files. All the novels that were available in a full-text format were included, so the proportions of file types underline the need for more full-text digitization. Without the work of extracting text from digital images, this dissertation would not have been possible because the corpus would have been too small and unbalanced.

Figure 11 gives an overview of the different institutional sources used to obtain the text of the novels. More than 30 different sources were used. In the chart, the sources are ordered by the number of texts taken from them. Some of the sources were grouped: individual websites, for example, about single authors, and university libraries from which printed books were loaned are not listed separately. Around 65 % of the novels were obtained from six main sources:[252] the "Biblioteca Virtual Miguel de Cervantes" (15.2 % or 39 texts), university libraries (14.5 % or 37 texts), the "Internet Archive" (10.2 % or 26 texts), the "HathiTrust Digital Library" (9.38 % or 24 texts), the "Ibero-Amerikanisches Institut" (8.59 % or 22 texts), and "Wikimedia Commons" (7.42 %

[252] In this chapter, *main sources* and *major sources* mean sources that were especially important for the corpus at hand, whereas *minor sources* provided fewer relevant texts. These formulations do not characterize the sources in themselves. Wikisource, for example, is a minor source for this corpus but a major source for digital texts in general.

or 19 texts).[253] Already the main sources show that the corpus was gathered from a broad range of sources because there is no general, comprehensive digital repository for Spanish-American nineteenth-century novels yet. The "Biblioteca Virtual Miguel de Cervantes" is a very important source because it contains many texts in HTML format and novels from many Spanish-American countries.[254] University libraries constitute the most important kind of source for printed editions of novels that were scanned and OCRed for this corpus.[255] Two of the other main sources are general repositories of multimedia content: the "Internet Archive" and "Wikimedia Commons". The "HathiTrust Digital Library", a collaborative platform of academic and research libraries based in the USA, and the "Ibero-Amerikanisches Institut", a German library specialized on Ibero American literature, were also significant sources.[256] Of course, several novels are available in more than one repository or library, so the overview given here is also the result of the text collection strategy pursued for this dissertation.[257] Because the "Biblioteca Virtual Miguel de Cervantes" was consulted first to obtain the digital full-texts, it is the most prominent source. However, it is also important to note that some of the other sources have great potential: the "Internet Archive" and the "HathiTrust Digital Library" contain many more Spanish-American nineteenth-century novels than the ones included in this corpus.[258] With more resources, the full text of these novels could be extracted, as well, to build a more extensive corpus for future research in the area. The digital library of the "Ibero-Amerikanisches Institut" is also constantly expanding, so it can be expected that this institution will play a major role as a source of Spanish-American literature in digital format in the future.

Figure 12 demonstrates which file types were obtained from which sources. The upper chart shows the institutions from which image files were obtained (68 % of the novels in the corpus), and the lower chart the ones for text files (32 % of the novels). It becomes clear that the "Biblioteca Virtual Miguel de Cervantes" is the only major source offering digital full texts and that more than half of the full texts were collected from minor sources. On the other hand, many of the institutions that offer Argentine, Cuban, and Mexican nineteenth-century novels only publish digital images.[259] The number and variety of sources used for this corpus shows that there is still

---

[253] Links to the websites of the digital libraries, repositories, and institutions are given in table 46 ("Sources of the novels in the corpus") in the appendix.

[254] It can be inferred from the HTML tags that the underlying data format of the library is TEI, but unfortunately, the texts are currently not offered in that format to the public. Besides texts in HTML, this virtual library also contains texts downloadable as PDF files with images.

[255] See chapter 3.3.2 for details about the text treatment. OCR stands for optical character recognition and is an umbrella term for procedures of the automatic conversion of images of text into machine-readable text.

[256] Thanks to cooperation with the "Ibero-Amerikanisches Institut" and financial support from the project CLiGS, several novels were scanned by the library and added to the digital collections of the institute.

[257] Links to digital versions of the novels' editions are given in Bib-ACMé. See https://github.com/cligs/bibacme/blob/master/app/data/editions.xml. Accessed December 8, 2019. The collection of links is not exhaustive, though, also because the availability of digital texts and images changes over time.

[258] The novels contained in these two repositories overlap to a great extent, probably because many of the scans were made by Google and uploaded to both. The "Internet Archive" is more permissive in that all the image files are downloadable also from outside the USA. The digitization work done by Google is impressive and of utmost importance for this dissertation.

[259] There are, of course, also novels in the form of printed books, which were loaned from university libraries, scanned, and converted to digital images.

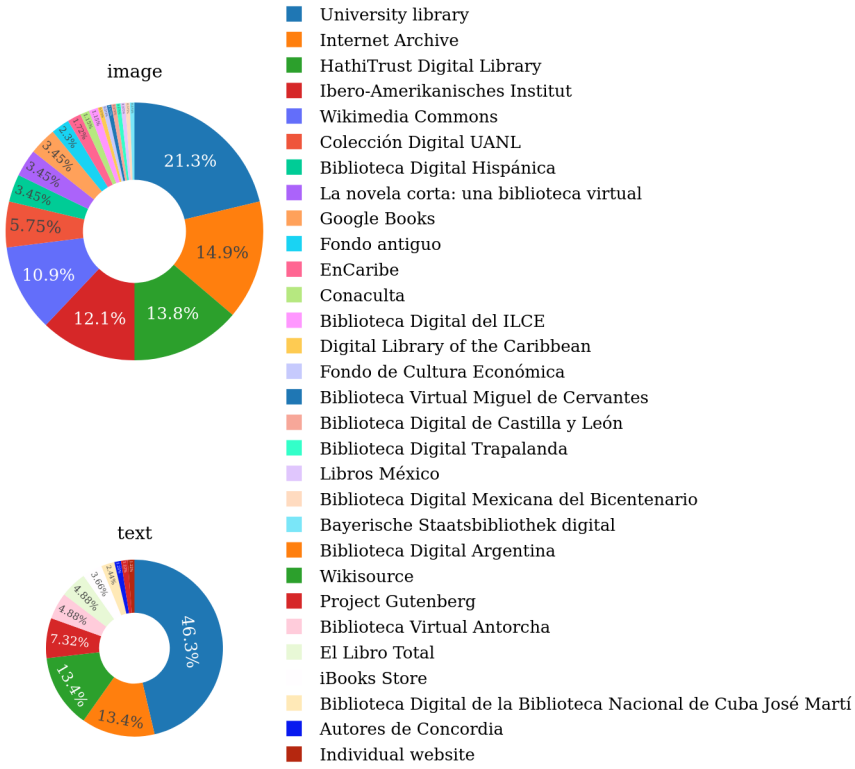


Figure 12. Sources by file type and institution.

much work to be done to facilitate future research on digital text analysis of nineteenth-century Spanish-American novels. For example, a supra-institutional portal gathering and pointing to different sources would be very helpful. Some of the above sources are designed as such portals, but they are still quite specialized or selective regarding the kind of information they provide. Furthermore, much more full-text digitization is needed to free future research projects from the necessity to invest considerable time and effort in full-text digitization before being able to analyze their corpus.

Because of the number and kind of different sources, it was indispensable for this corpus to control the quality of the incoming texts.[260] One question is how many novels were obtained from scholarly sources and how many from general ones.[261] Only two-thirds of the texts (64.1 %) come from repositories that can be associated with scholarly undertakings. Having in mind that

[260] How the quality control was done is outlined in the next chapter 3.3.2 on text treatment.

[261] Digital libraries and portals connected to libraries, universities, and governmental institutions were counted as scholarly. Individual, personal websites were subsumed under the general category. The texts taken from "Wikimedia Commons" are included in the general group here, but ultimately they have a scholarly background, as well, because they are all digital reproductions of novels held by the "Academia Argentina de Letras" (Wikimedia Commons 2019). That the Argentine Academy decided to upload the PDF files of the novels (and other texts) to "Wikimedia Commons" for the benefit of the general public is exemplary.

the text analysis done for this dissertation aims to be scholarly, the share of general sources is large. Usually, the scholarly sources are more reliable regarding the provision of metadata about the texts and in terms of long-term stability, but not principally. The stability and accuracy are also connected to the general scope, relevance, and functioning of the (digital) institutions. The platforms of the Wikimedia Foundation ("Wikisource" and "Wikimedia Commons"), for example, are stable, or rather, changes and enhancements are well documented. At the same time, minor institutional websites are more prone to be altered or to disappear.[262]

Another relevant question about the corpus sources is which kind of editions underlie the digital or print sources of the novels. The kinds of editions were grouped into four categories:

- first editions,
- historical editions other than the first but within the period up to 1910 analyzed here,
- modern editions published after 1910,
- and novels where the kind of underlying edition is unknown.

The corpus is built upon a mix of editions: The majority of the editions are historical (86.8 %), with 64.1 % of first editions and 22.7 % of historical editions other than the first one. 30.5 % are modern editions, and the underlying editions of 20 novels (7.81 %) are unknown. These results stress the need for a quality check and harmonization of the textual basis.[263] With a better state of digitization or with better access to modern digital editions, for example, ebooks, some of which are protected by copyright, it would have been possible to pursue a more consistent strategy regarding the kinds of editions, e.g., to only consider first editions or only modern ones. However, not all the novels, especially the lesser-known ones, have been re-edited. The first editions are also not preserved for all the novels so a strict approach would not have been fully realizable even under better conditions of digitization and access. In figure 13, the information about the type of edition and type of institution is combined, showing that cases in which the underlying edition is unknown are more frequent in general sources than in scholarly ones. Nevertheless, there are also scholarly resources in which the source edition of the novels' digital version is not indicated. Historical editions are frequent both in scholarly and general sources, and the modern editions in the corpus are mainly from scholarly sources.[264]

In view of the above, it becomes clear that the selection of novels for the corpus was guided by criteria related to the genre and other factors influencing the style of the texts but that the possibilities to compile a representative and balanced corpus were also limited by practical aspects concerning the availability of the texts from different sources and in different formats. One-third of the texts were already available in a full-text format, one-fifth was digitized, and the rest was extracted from image files. In total, more than 30 different sources were used, and different types of editions (first editions, other historical editions and modern ones) had to be employed. In light of the non-uniform composition of the corpus regarding its sources, the text of the novels had to

[262] Details about which sources have changed during the work on this dissertation can be found in table 46 in the appendix.

[263] For details, see the next chapter 3.3.2 on text treatment.

[264] Most of the modern editions are from scholarly sources due to the books loaned from the university libraries, which were almost exclusively modern editions. Very old editions can usually not be loaned in the German library system. In addition, digital repositories tend to offer older editions for which no copyright issues are to be expected.

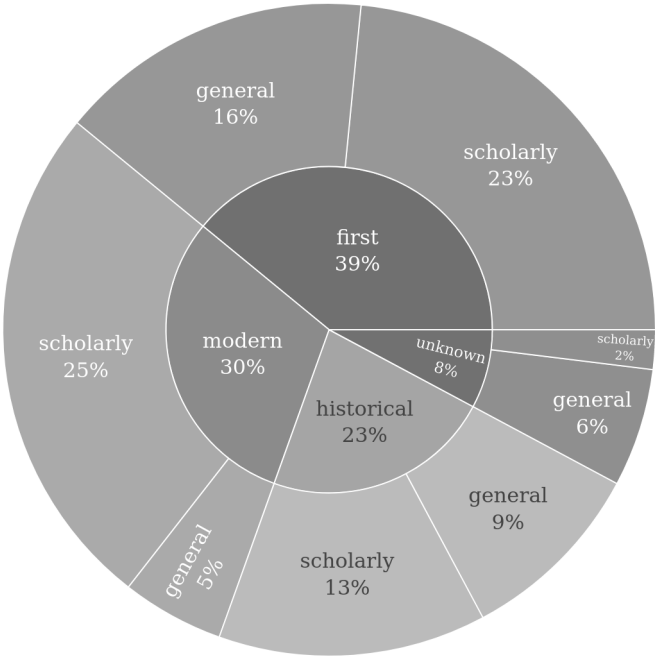


Figure 13. Sources by type of edition and type of institution.

be treated in different ways and to be checked to homogenize the collection, which is described in the following chapter.

3.3.2 Text Treatment

Depending on the type of source, the text of the novels for the corpus had to be prepared differently. The further away a source text was from a high-quality digital full text, the more steps were necessary. Table 11 lists the different processing steps that were followed. In the case of a printed book, all the steps had to be undertaken. If, in contrast, the source was an HTML file containing the full text of a novel, only the last two steps were carried out. The other types of sources required a number of processing steps between the two extremes. The preparation of the full text also included the addition of basic structural information in the form of markup, because adding this kind of information was intended anyway, and the goal was not to lose existing relevant information.

The first step, scanning, was necessary for novels that were only accessible as printed books. A selection of books was scanned by the "Ibero-Amerikanisches Institut" (IAI) in Berlin and added to their digital library. The books can be viewed online and downloaded as a PDF file. In the digital library of the IAI, the books are enriched with general, administrative, and structural metadata, including the assignment of persistent identifiers.[265] The library also holds high-quality images of the scans. The remainder of the books that needed to be scanned were treated by the

[265] See, for example, Paz ([1883] 2017).

	Step	Type of source
1	Scanning	Printed books
2	OCR	Image files and image-based PDF files
3	Correction of OCR results	OCR-output
4	Conversion and/or addition of structural information	Corrected OCR-output, HTML-Files, plain text files
5	Spell check	Full text

Table 11. Steps for the preparation of structured full text.

author of this dissertation. The scans were done at the University of Würzburg with ordinary scanners in the library. They were done in an ad-hoc manner with the goal of being able to extract the text of the novels and not of keeping the image files. The scans were mainly done of modern editions, and the development of a professional digitization workflow like the one that the IAI established was not part of the CLiGS project. The bibliographical metadata added to the resulting files can still be inspected to check which editions were used as a basis for the texts.

The second step involved the conversion of the digital images of printed text into machine-readable text with the help of optical character recognition (OCR). This applied to scans of printed books and also to novels already available in the form of image files or image-based PDF files. The software used to perform the OCR was ABBYY Finereader 12 Professional because it proved to achieve good results for nineteenth-century texts in the Spanish language.[266] All the novels processed with ABBYY Finereader were checked page by page to correct the results of the OCR. General mistakes were corrected with the help of the Find/Replace routines of the software, and individual mistakes were corrected on the pages themselves.

Because the source editions were historical as well as modern ones,[267] it was decided to unify the orthography to a modern one as far as possible. There are several reasons for this decision. First, the aim of this study is not detailed historical linguistic analyses of the texts but stylistic analyses focusing on general linguistic and semantic aspects of the novels.[268] Second, it would not have been possible to only use source editions from a certain historical point, so a unifying strategy was necessary anyway, and modern editions can hardly be converted back to a historical spelling. Third, most natural language processing (NLP) tools that support Spanish as a language expect a modern spelling, so historical spellings would have led to additional problems in the analyses of the texts. Fourth, with the standard setting for the Spanish language, ABBYY Finereader automatically corrects many words to a modern spelling. Instead of considering this a drawback, it was taken advantage of. The most frequent words that were corrected were conjunctions and prepositions (á/é/ó/ú → a/e/o/u), the adverb "mas" (→ "más"), and verb forms in the preterite imperfect (e.g., "hacia" → "hacía", "sabia" → "sabía", "venia" → "venía"). A

[266] Tests were also done with the free software Tesseract, but the results were not as satisfying as the ones achieved with ABBYY Finereader.

[267] For an overview of the shares of the different types of source editions in the corpus, see the previous chapter 3.3.1 above.

[268] See chapter 4.2.1 about the textual features used.

problem that persisted were verb forms that included enclitic pronouns such as "decíale" (instead of "le decía"), "olvidábasenos" (instead of "se nos olvidaba"), "viose" (instead of "se vio"), and so on because they cannot easily be changed automatically. As a result, in some of the texts in the corpus, the old verb forms are included, whereas others only have modern forms. It has to be kept in mind that stylistic analyses involving, for example, the examination of the usage of archaic forms as a typical sign of certain subgenres, for instance, historical novels, are not possible here because of the different types of source editions. Furthermore, the composite verb forms might cause problems for some NLP tools.

Corrections in the text (neither corrections of obvious errors in the OCR results nor orthographic modernizations) were not encoded in detail. This decision was made because the focus of the CLiGS project was not on the creation of scholarly historical editions but on large-scale stylistic analyses of digital text. Of course, the basic full texts produced in this project could be used as a starting point for the creation of critical editions, but to undertake these encoding steps for hundreds of novels would not have been neither plausible nor manageable here.[269]

The next processing step was adding structural information, or, if such information was already present in the source files, its conversion. This step was applied to the corrected OCR output, but also if the sources were HTML or plain text files. The goal was to create a basic structural markup for the novels, including the encoding of headings, paragraphs, chapters, and parts of the novels. The target format chosen is the encoding standard of the Text Encoding Initiative (TEI). As this is the general data format used for the corpus of novels, it is described in more detail in the following chapter 3.3.3 on metadata and text encoding. In the case of plain text files, blank lines indicating paragraph boundaries were exploited with regular expressions. From HTML files, all relevant structures were extracted either with the help of XSLT scripts or with Python scripts using regular expressions, depending on whether the HTML files could be processed as well-formed XML files (a requirement for the XSLT processor) or not. Depending on the kind of web source, in some cases, the download of the HTML files also involved the scraping of individual pages (e.g., chapters) that belonged to the same novel before the files could be processed further.[270] Because there is an option to export the OCR output from ABBYY Finereader as HTML, the files processed with this software could also be transformed to basic TEI with the help of an XSLT script.[271] Most of the scripts used for crawling web pages and for converting or adding structural information were written in an ad-hoc manner and changed from source to source. In some cases, the HTML structure was inconsistent from novel to novel, even within the same source repository. All the resulting basic structures (parts, chapters, headings, and paragraphs) were checked manually.[272] Some contents and structures were not taken over: In

[269] An example of a large-scale project aiming at the creation of digital editions suitable for historical linguistic analyses is the "German Text Archive" ("Deutsches Textarchiv") (BBAW 2022).

[270] In "Project Gutenberg", for example, all the text of a novel is presented on a single page. On the platforms "Wikisource" and "Biblioteca Virtual Miguel de Cervantes", in contrast, a novel is presented on several pages (usually by chapter on "Wikisource" and apparently arbitrary divisions in the "Biblioteca Virtual Miguel de Cervantes"). As examples, see Cambaceres (2008, [1885] 2000), and Ocantos ([1913] 2007).

[271] The script is available at https://github.com/cligs/scripts-nh/blob/master/corpus/text_treatment/clean.xsl. Accessed March 17, 2020.

[272] For parts and chapters of the novels, it was checked if all the units were there, for example, by ensuring successive chapter numbers. For the paragraphs, the check was done in a rough manner. For instance, it was tested whether

the case of modern editions, prefaces, introductions, and appendices written by the editors were left aside, primarily to prevent copyright issues when publishing the corpus. On the other hand, historical title pages, dedications, and prefaces were kept because they were checked for generic signals.[273] Some novels contain pictures illustrating selected scenes of the plot. These were dropped because the analysis of illustrations is not intended here. Notes by authors as well as editors were not kept. Even though authorial notes tend to be more frequent in some subgenres of the novel (historical novels and science fiction novels, for instance) it was not possible to distinguish between authorial and editorial notes in all cases. Even though the goal of this project was not to create critical editions of the texts, one phenomenon was nevertheless documented: gaps in the text. Reasons for gaps are:

- missing pages, either in the originals or in the digital reproductions of the novels,
- unreadable or missing passages because of aging signs or damage to the books,
- or missing words or characters because of print errors.

This problem occurred only in historical editions. Wherever possible, other historical editions of the texts were checked to see if the gaps could be filled that way. Nevertheless, some gaps remained, for example, in cases where the text was missing in all the available editions or where the edition used was the only one that could be obtained. In total, 96 gaps were detected in 30 texts.[274] Most gaps consist of individual or several illegible words (106), followed by missing pages (32), illegible lines (16), and characters (18).[275] In view of the overall size of the corpus, the number of gaps is considered acceptable.

The last step in the pipeline of text treatment was a spell check, which was applied to the full texts resulting from the previous processing steps. With the final spell check, it was intended to find errors remaining after the OCR correction of texts that were obtained from digital images. Texts obtained from existing plain text and HTML files were also checked because of the great variety of sources. A Python module was written to perform the spell check with the library PyEnchant.[276] One of the backends used by the underlying Enchant library is MySpell, a project also used in OpenOffice (or LibreOffice) to perform spell checks. Via MySpell, dictionaries for several languages are available, including Spanish. The spell check can be performed for individual files or a whole collection of text files. It is possible to indicate files with exception words containing, for example, proper names of people and places or words from foreign languages. For

---

there were paragraphs starting with a lower-case letter which was often a sign that there was a superfluous paragraph boundary.

[273] Other historical paratexts such as tables of content, lists of errata, or advertisements of other publications were dropped. See chapter 3.3.4 below about the assignment of subgenre labels to the novels in the corpus.

[274] A bar chart analyzing the gaps was produced with the script https://github.com/cligs/scripts-nh/blob/master/corpus/text_treatment/check-gaps.xsl. The resulting data can be downloaded at https://github.com/cligs/data-nh/tree/master/corpus/text-treatment. Accessed January 29, 2023.

[275] The number of missing items was estimated depending on the extent of the missing text on the page.

[276] The script is available at https://github.com/cligs/scripts-nh/blob/master/corpus/text_treatment/spellchecking.py. It was also used to create the charts about spelling errors included below. Accessed March 17, 2020. The core function for the spell check and the functions for the visualization of errors were written by the author of this dissertation. An additional function in the module for the automatic correction of errors was written by Christof Schöch. The idea behind the spell check and results from a preliminary corpus of Spanish-American novels and from a corpus of French novels are presented in Henny-Krahmer and Schöch (2016).

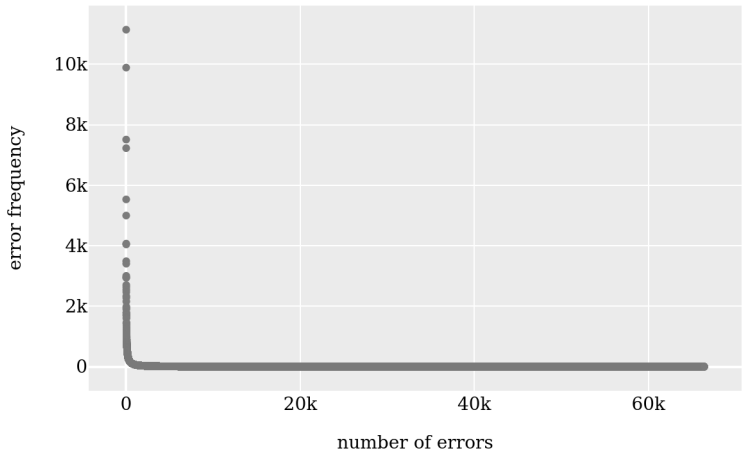


Figure 14. Distribution of spelling errors without exception words.

the corpus of Spanish-American novels, the spell check was performed for each file individually. The lists were then checked for genuine errors, including errors resulting from the OCR process, orthographic errors contained in full-text files from external sources, or errors resulting from historical spellings. All the genuine errors that occurred more than once in a file were corrected. That way, the most frequent and typical errors were solved. However, for reasons of time, it was not possible to also correct all the errors occurring only once because of their sheer number in some cases. Even if the resulting full texts are not perfect, the spell check helped to get an impression of the orthographic quality of all the texts in the collection, and it was helpful to align the level of correctness of the files obtained from different sources. In general, the full text extracted from modern editions or obtained from portals that themselves checked the texts has a higher quality than text extracted from historical editions or collected from sites without their own quality control. It is important to note, though, that the quality was checked against a dictionary of modern Spanish here.[277] In figures 14 and 15, the distribution of the spelling errors that remained after the correction of the individual texts is displayed. No lists of exception words were included.[278]

The figures show that the frequency of the errors drops quite sharply, but also that many different errors remain. The total number of different errors is 66,399, which is 33,6 % of the whole vocabulary of the collection,[279] which is quite a lot. Nevertheless, 62.5 % (i.e., 21 % of the vocabulary) of the different errors occur only once, and only 7.6 % (2.5 % of the vocabulary) occur more than ten times. Consequently, regarding not the types but the tokens, the proportions are

[277] As an aside, the results of the spell check can also be used for purposes other than controlling the orthography. The words that occur in the check files but are not real errors indicate how many and how many different proper names the novels contain, how many foreign words, how many words from special areas of vocabulary, and so on, so they are also interesting from a stylistic point of view.

[278] The CSV file containing the results of the spell check is available at https://github.com/cligs/data-nh/blob/master/corpus/text-treatment/spellcheck.csv. Accessed March 22, 2020.

[279] The size of the whole vocabulary is 197,520. This number was determined with the script https://github.com/cligs/scripts-nh/blob/master/features/bow.py. Accessed March 19, 2020.

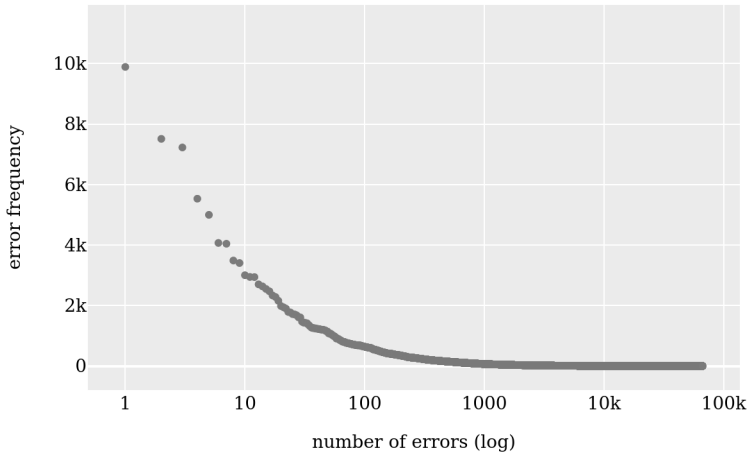


Figure 15. Distribution of spelling errors without exception words (logarithmic scale).

different: The total number of errors is 543,693, which is 3.2 % of all the tokens in the collection.[280] 41,481 (0.2 %) of the tokens are errors that occur only once, and 430,284 (2.5 %) are mistakes with more than ten occurrences. What follows from these numbers regarding the analysis of the texts? First, measures of statistical similarity will probably not be influenced too much by the errors because most of them are so infrequent. On the other hand, an analysis of the hapax legomena[281] in the corpus is not advisable.[282] However, the above numbers represent all of the words that were not recognized by the spell checker, but many of them and especially the frequent errors that were not corrected in the individual files during the preparation of the texts, are not genuine errors. The most frequent error in the whole collection, for example, is the word "vd", an abbreviation for the personal pronoun "usted", with 11,145 occurrences. Figure 16 shows the top 30 most frequent spelling errors in the corpus.[283]

Among the most frequent errors, there are, for example, forms of address ("vd", "v", "ud" → "Vd.", "V.", "Ud." → "usted"; "V. A. → "Vuestra Alteza"; "V. E." → "Vuestra Excelencia"; "V. R." → "Vuestra Reverencia"; "V. S." → "Vuestra Señoría; "d" → "D.", "D.ª", → "Don", "Doña"; "s" → "S. M." → "Su Majestad", etc.). The individual letters "v" and "s" can also stand for other words, for example, the number five ("V") or the word "San" (as in "S. Fernando", "S. Juan de Dios", etc.). Apart from that, most of the top errors are proper names ("María", "Juan", "Pedro", etc.) and place names ("México", "España"). A possibility to exclude these errors from the results of the spell check is to create lists with exception words.

Several strategies were followed to generate exception lists for the spell check of the corpus. First, free lists of exception words available on the web were used to see which of the items contained in them also occur in the error list resulting from the previous spell check round. The matching items were then stored in corpus-specific exception lists, which can be further adapted

[280] The whole corpus contains 17,104,856 tokens (see the script in the previous footnote).
[281] I.e., tokens that occur only once in the whole corpus.
[282] These conclusions were already drawn in Henny-Krahmer and Schöch (2016).
[283] For the spell check, all the tokens were converted to lowercase.

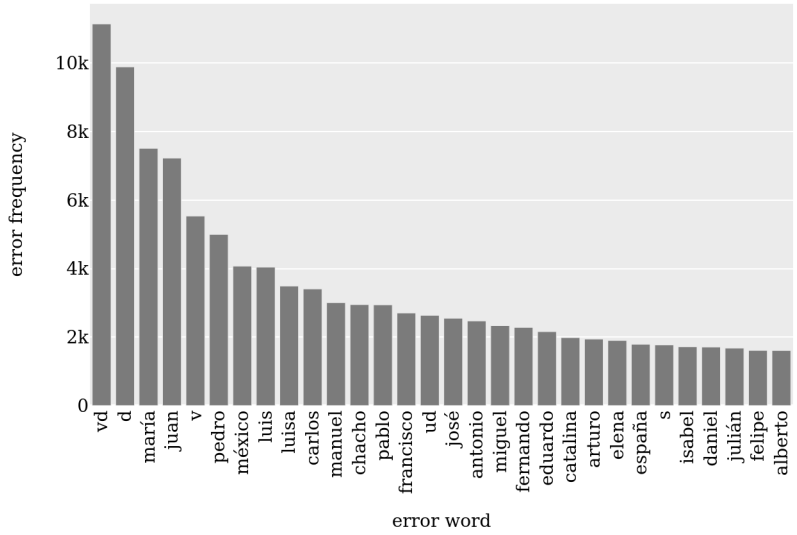


Figure 16. Top 30 spelling errors.

Noun type	Word number in list	Number of error types covered		Number of error tokens covered	
proper names	455	282	0.42 %	125,906	23.2 %
surnames	103	42	0.06 %	12,772	2.3 %
countries	193	60	0.09 %	9,274	1.7 %
capitals	182	33	0.05 %	2,315	0.4 %
Sum:	933	417	0.62 %	150,267	27.6 %

Table 12. Error words mapped with general lists of proper nouns.

manually. This strategy was followed for proper names, surnames, names of countries, and capitals.[284] Table 12 summarizes how many supposed error words could be mapped that way.[285]

The table shows that from the four external word lists, the one with proper names was most useful because more than half of the names it contains occur in the spell check results, and the total amount of error tokens could be reduced by more than one-fifth using this list. The other three lists with surnames, countries, and capitals did only have a minor effect.

[284] The web source for the lists of proper names and surnames was Olea (2021). From that source, the files "nombres-propios-es.txt" and "apellidos-es.txt" were used. The names of the countries were obtained from Wikipedia (2022). The list of capitals was retrieved from Frech (1998–2017). To compare these lists to the spell check error list, the function `generate_exception_list()` in the script "spellchecking.py" was used. See footnote 276. The resulting exception lists ("exceptions-proper-names.txt", "exceptions-surnames.txt", "exceptions-countries.txt", "exceptions-capitals.txt") are contained in https://github.com/cligs/data-nh/tree/master/corpus/text-treatment/exception-words. Accessed March 28, 2020.

[285] The numbers were calculated with the function `interprete_exception_list()` in the module "spellchecking.py". In the table, the relative numbers were rounded.

The second strategy that was pursued to generate exception lists was the usage of word patterns expressed as regular expressions. Looking at the spell check results, many false errors from specific word classes stood out, among them words with diminutive suffixes (e.g., “abuelito”), superlatives (e.g., “interesantísimo”), adverbs ending in “mente” (e.g., “aceptablemente”), and verb forms with pronoun suffixes, of which many are archaic (e.g., “diósele”). In all these cases, the range of possible words is so extensive that it is hardly possible to match them individually. Even with the use of a dictionary, productive word formations would not be covered. However, it is possible to match these kinds of words fairly accurately with patterns. The regular expression “`.*i(t|ll)(a|o)s?\b`”, for example, matches all the diminutive words ending in “-ito”, “-itos”, “-ita”, “-itas”, “-illo”, “-illos”, “-illa”, and “-illas”, such as, for instance, “abuelita”, “caminillo”, or “milloncitos”. Compared to word lists, patterns have the advantage that many more forms can be matched without the need to anticipate their exact construction. A slight disadvantage of the patterns is that they can also cover false positives. In the case of the diminutives, for example, also proper names and misspelled general nouns were matched: “Antillas”, “álito” (which should be “hálito”), “exito” (which should be “éxito”), and “estrepito” (which should be “estrepíto”). Furthermore, the use of patterns is only reasonable if the morphology of the language allows it to match specific word classes quite unambiguously. Fortunately, this is possible for Spanish diminutives, superlatives, adverbs, and verb forms with pronoun suffixes.

The patterns were used in the same way as the word lists. They were applied to the error list resulting from the spell check to generate a corpus-specific list of exception words, which can then be used in the next spell check round. To have such corpus-specific lists is not only useful for the spell check process. It can also be interesting to analyze them from a stylistic point of view, to find out which texts or groups of texts contain many non-standard words of a certain kind. For example, they could be used to see how frequent the diminutives are in novels of a particular genre, from certain countries, or authors. Furthermore, the exception lists can help to improve the results of natural language processing tools that do not recognize certain non-standard word forms, for example, if they are not based on a model of historical Spanish.[286]

It is more complex to map the verb forms with pronoun suffixes than the diminutives, superlatives, and adverbs because many more combinations of forms are possible. These are verb forms to which reflexive, passive, and personal pronouns are directly suffixed, for example, “ofrecióselas” instead of “se las ofreció”, “oíasele” instead of “se le oía”, or “urgíame” instead of “me urgía”. In table 13, regular expressions to map such forms are displayed.

Table 13. Regular expressions for verb forms with pronoun suffixes.

Pattern	**Kind of verb forms matched**	**Examples from the corpus**
`.*[aei]rse\b`	*infinitivo* with a reflexive pronoun	*celebrarse*, *apetecerse*, *percibirse*
`.*[aeiáéí]r(se)?l[eao]s?\b`	*infinitivo* with a reflexive pronoun and with a personal pronoun in third person singular or plural in dative or accusative	*estarle*, *caerle*, *irle*, *oírlo*, *serles*, *mostrarselo*, *torcérselas*

[286] See chapter 3.3.5, where the linguistic annotation of the corpus files is described.

Pattern	Kind of verb forms matched	Examples from the corpus
`.*[áéíóú]r[ae]n?se\b`	*presente*, *pretérito indefindo* in third person singular or plural with a reflexive or passive pronoun	*hubiérase*, *érase*, *ignórase*, *asegúrase*, *refiérese*, *palpáranse*
`.*[éóo][mts]e\b`	*presente*, *gerundio*, *futuro simple*, *pretérito indefinido*, *pretérito (pluscuam)perfecto* in first or third person singular, with a reflexive or passive pronoun, or with a personal pronoun in 1st or second person singular	*encaminéme*, *miréte*, *parecióme*, *ruégote*, *levantóse*, *detúvose*, *irguiéndose*, *decorádose*, *hubiése*, *diréte*
`.*[éó]l[eao]s?\b`	*pretérito indefinido* in first or third person singular with a pronoun in third person singular or plural in dative or accusative	*contéle*, *alarguéla*, *preguntóle*, *tomóla*, *parecióles*, *chingólos*
`.*[áé]ndol[eao]s?\b`	gerundio with a personal pronoun in third person singular or plural in dative or accusative	*temblándole*, *siéndole*, *reflexionándolo*, *faltándoles*
`.*[éóo][mts]el[eao]s?\b`	*presente*, *gerundio*, *pretérito indefinido* in first or third person singular, with a reflexive pronoun or a personal pronoun in first, second or third person singular in dative or accusative	*entreguéselo*, *diómela*, *avisándotelo*, *acercándosele*, *ocurriósele*, *pelándoselas*, *hubiésele*
`.*[éóo]n?os\b`	*presente*, *imperativo*, *gerundio*, *pretérito indefinido* in first or second person singular or plural, or third person singular, with a personal pronoun in first and second person plural in dative or accusative	*detenéos*, *noticiándoos*, *suplicoos*, *sucediéndonos*, *vímonos*, *proporcionóos*
`.*[óo]n?osl[eao]s?\b`	*presente*, *gerundio*, *pretérito indefinido* in first person singular or plural, or third person singular, with personal pronouns in first and second person plural in dative or accusative, and in third person singular or plural in dative or accusative	*conociéndonosla*
`.*[mt]el[eao]s?\b`	*infinitivo*, *imperativo* in singular or plural, with personal pronouns in first or second person singular, and in third person singular or plural in dative or accusative	*conquistármelo*, *amarrartela*, *consagrártelos*, *créanmelo*
`.*[mn]?osl[eao]s?\b`	*infinitivo*, *imperativo*, *presente*, *futuro simple*, *pretérito indefinido*, *pretérito imperfecto* in first or second person plural, with a personal pronoun in third person singular or plural in dative or accusative	*bebémosla*, *hicímoslo*, *recordábamosle*, *llamarémosla*, *enderezémosle*, *atajárnosla*, *arrojarnoslos*, *anunciároslo*
`.*[áé]isme\b`	*presente*, *imperativo* in second person plural with a personal pronoun in first person singular in dative or accusative	*prometéisme*, *ordenáisme*

Pattern	Kind of verb forms matched	Examples from the corpus
`.*[áé]isl[eao]s?\b`	*presente*, *futuro simple* in second person plural with a personal pronoun in third person singular or plural in dative or accusative	*veréisle*, *habéislo*, *conocéisla*
`.*é(se)?l[eao]s\b`	*presente*, *imperativo*, *pretérito indefinido* in first, second or third person singular, with a reflexive pronoun and with a personal pronoun in third person singular or plural in dative or accusative	*firmélos*, *quitélas*, *hélos*, *délas*, *fuéselos*, *véseles*
`.*í[jz]ol[eao]s?\b`	*presente*, *pretérito indefinido* of certain irregular verbs in first or third person singular with a pronoun in third person singular or plural in dative or accusative	*hízole*, *exíjoles*, *díjola*, *díjoles*, *bendíjolas*
`.*(á|é)ron[mts]e\b`	*pretérito indefinido* in third person plural with a reflexive pronoun or a personal pronoun in first or second person singular in dative or accusative	*dijéronme*, *hospedáronme*, *guardáronse*, *humedeciéronse*
`.*(á|é)ronn?os\b`	*pretérito indefinido* in third person plural with a personal pronoun in first or second person plural in dative or accusative	*hiciéronnos*, *mejoráronos*
`.*(á|é)ronse[mt]e\b`	*pretérito indefinido* in third person plural with a reflexive pronoun and a personal pronoun in first or second person singular in dative or accusative	*antojáronseme*
`.*(á|é)ron([mts]e)?l[eao]s?\b`	*pretérito indefinido* in third person plural with a reflexive pronoun or personal pronouns in first or second person singular and third person singular or plural in dative or accusative	*trajéronle*, *justificáronlos*, *encendiéronle*, *erizáronsele*, *reveláronmele*
`.*(ába|ía)n?[mts]e\b`	*condicional*, *pretérito indefinido* in third person singular or plural with a reflexive pronoun or a pronoun in first or second person singular in dative or accusative	*congratulábame*, *habríate*, *reuniríase*, *citábase*, *concedíanse*
`.*(ába|ía)n?n?os\b`	*condicional*, *pretérito indefinido* in third person singular or plural with a personal pronoun in first or second person plural in dative or accusative	*llamábannos*, *hallábaos*, *habíanos*, *habríaos*
`.*(ába|ía)n?se[mt]e\b`	*condicional*, *pretérito indefinido* in third person singular or plural with a reflexive pronoun and a personal pronoun in first or second person singular in dative or accusative	*habíaseme*, *olvidábaseme*

Pattern	Kind of verb forms matched	Examples from the corpus
`.*(ába|ía)n?([mts]e)?l[eao]s?\b`	*condicional*, *pretérito indefinido* in third person singular or plural with a reflexive pronoun or a personal pronoun in first or second person singular and a personal pronoun in third person singular or plural in dative or accusative	*impedíamelo*, *anudábansele anuncíale*, *acogeríanlo*, *acogíalos*

The regular expressions illustrate the complexity of the Spanish verbal and pronominal system. The verb form patterns are determined by the verb class (verbs ending in "-ar", "-er", or "-ir", regular and irregular verbs), tense, person, and number of the verbs, as well as by the person, number, gender, case (dative or accusative), and mode (passive, reflexive, indicative, subjunctive) of the attached pronouns, and finally by spelling variants (e.g., "dárselos" versus "darselos"). The regular expressions displayed here aim to cover most of the usual cases, and they are quite compact in that they cover several types of verb forms at once. It would also be possible to create individual regular expressions for each theoretically possible type of verb form with pronoun suffixes, but this would result in several hundred different expressions because all the verb forms would have to be combined with one or more pronouns in all possible forms.[287] Here, a mix of systematic and heuristic approaches was preferred to match many cases occurring in the corpus. Many of the verb forms with pronoun suffixes are historical, e.g., the forms in the past tense (for example, "diómela" or "decorádose"). On the other hand, infinitive, gerund, and imperative forms with attached pronouns (e.g., "conquistármelo", "pelándoselas", or "créanmelo") are still in use in modern Spanish, but they were not recognized by the spell checker, either. Table 14 contains the results of the diminutive, superlative, adverb, and verb form mappings.[288]

When compared to the false errors matched with the word lists, the results for the word ending patterns show that many more error types are covered this way – more than one-fourth of all the error types – but not necessarily more error tokens. Especially for the verb forms with suffixed pronouns, the generation of patterns is quite laborious and is only worthwhile because it also helps to improve NLP results.

[287] A version of the above list of verb form patterns with disassembled expressions is available at https://github.com/cligs/data-nh/blob/master/corpus/text-treatment/exception-words/source-lists/verb-form-patterns-detail-es.txt. Accessed March 31, 2020. Already this list comprises 442 different expressions. In reality, though, not all theoretically possible combinations necessarily occur in the language and in language use, so it would be even more work to create a list of verb forms with pronoun suffixes that is, on the one hand, complete and, on the other hand, adequate for the linguistic reality. The one created here is only an approximation of such a list.

[288] The table is sorted by the number of error tokens covered. The files with the corresponding patterns are named "verb-form-patterns-es.txt", "diminutive-patterns-es.txt", "superlative-patterns-es.txt", and "adverb-patterns-es.txt". They are contained in https://github.com/cligs/data-nh/tree/master/corpus/text-treatment/exception-words/source-lists. To create the exception lists from the patterns, the function `generate_exception_list()` in the module "spellchecking.py" was used. The resulting exception lists can be viewed at https://github.com/cligs/data-nh/tree/master/corpus/text-treatment/exception-words. The function `interprete_exception_list()` in the same module served to evaluate how many errors are covered by each exception list. The links were accessed on March 31, 2020.

Pattern type	Number of error types covered		Number of error tokens covered	
verb form endings	10,591	16 %	39,217	7.2 %
diminutive endings	4,582	7 %	34,927	6.4 %
superlative endings	1,286	2 %	6,739	1.2 %
adverbs	698	1 %	2,134	0.4 %
Sum:	17,157	26 %	83,017	15.2 %

Table 14. Error words mapped with word patterns.

The third part of the strategy to generate exception lists is manual editing. As could be seen in figure 16 above, some forms of address are among the most top frequent errors. These can best be covered with a simple list created as needed when looking at the top errors in the spell-check results. Other types of words for which it is not easily possible to obtain ready lists or generate them on the basis of patterns are, for example, foreign words, specialized vocabulary, or forms of oral speech. Manual editing is also a good strategy to adapt lists obtained elsewhere to the needs of the corpus, such as the lists of proper and place names. When creating exception lists, it is advisable to proceed with caution and also look into the texts in some cases because there are words that can both be an exception word or a real error (e.g., the entry "nina", which in the corpus referred to the proper name "Nina" but also was a misspelled version of "niña"). Moreover, words can belong to several kinds of exception words at once. This is often the case for surnames and place names (e.g., "villaclara" or "villanueve"). Table 15 summarizes how many false errors could be detected with the help of manually created and manually enhanced lists.[289]

As can be seen, manual lists can be very effective if they cover high-frequency errors, as is the case of the "other" list, and if frequent corpus-specific exception words are added to external lists, for example, special proper names such as "Moctezuma" or "Chacho". Although the exception lists in themselves do not help to improve the quality of the texts, they allow us to evaluate the amount of real errors in a better way. However, the process of creating and refining exception lists, as well as correcting remaining errors, can be carried forward infinitely – or rather, until everything is cleaned up – but there is a point when this is not effective anymore. For the corpus, all the errors that occurred more than 50 times in the whole collection were checked, and the words were either added to exception lists or corrected. All the remaining entries in the spell check result list were left as they are. So the texts are not entirely free of errors but corrected as far as possible. Having a look at the remaining errors, at the top of the list, there are still predominantly exception words, while there are more real errors with decreasing frequency. Figures 17 to 23 summarize the effect of all the exception word lists and show how many and

[289] The table is sorted by the number of error tokens covered. The files with the corresponding exception lists are named "exceptions-proper-names_ext.txt", "exceptions-surnames_ext.txt", "exceptions-other.txt" and "exceptions-places.txt", "exceptions-countries_ext.txt", "exceptions-foreign.txt", "exceptions-special.txt", "exceptions-oral.txt", and "exceptions-archaic.txt". They are contained in https://github.com/cligs/data-nh/tree/master/corpus/text-treatment/exception-words. The function `interprete_exception_list()` in the same module served to evaluate how many errors are covered by each exception list. The links were accessed on March 31, 2020.

Type of list	Number of error types covered		Number of error tokens covered	
proper names (enhanced)	574	0.86 %	178,610	32.9 %
surnames (enhanced)	378	0.57 %	60,418	11.1 %
other (containing e. g. individual forms of address)	26	0.04 %	34,275	6.3 %
places[290]	108	0.16 %	13,885	2.6 %
countries (enhanced)	62	0.09 %	9,993	1.8 %
foreign words	47	0.07 %	4,334	0.8 %
specialized vocabulary	34	0.05 %	2,911	0.5 %
oral speech	9	0.01 %	1,666	0.3 %
archaic vocabulary	105	0.16 %	561	0.1 %
Sum:	1,343	2.01 %	306,653	56.4 %

Table 15. Error words mapped with manually edited exception lists.

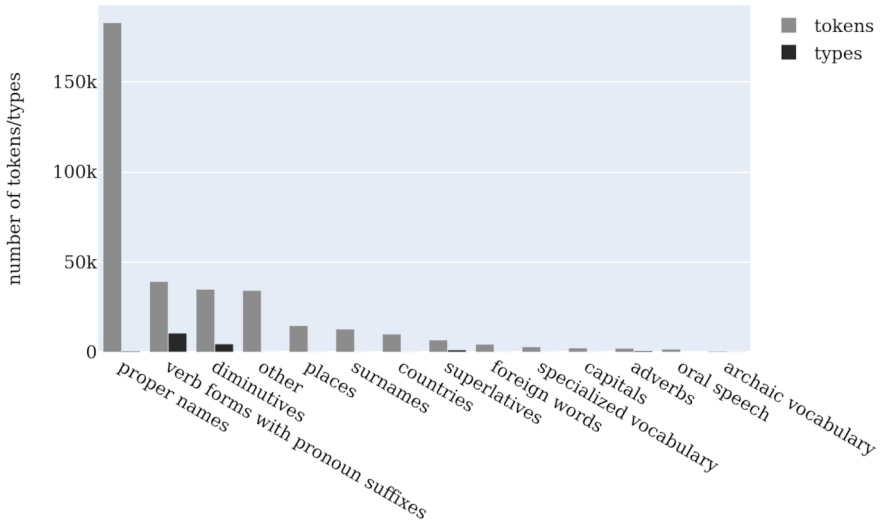


Figure 17. Number of error tokens and types covered by exception lists.

what kind of errors remain after their application and after further correcting errors that were frequent in the whole collection.[291]

[290] The manually edited list of places covers place names other than countries and capitals.

[291] The list of errors that remained after including exception words is available at https://github.com/cligs/data-nh/blob/master/corpus/text-treatment/spellcheck_exc.csv. The charts were produced with the module "spellchecking.py" and are available as HTML files at https://github.com/cligs/data-nh/tree/master/corpus/text-treatment ("coverage-exception-lists.html", "distribution-spelling-errors-exc.html", "distribution-spelling-errors-files.html", "distribution-spelling-errors-files-relative.html", "distribution-spelling-errors-files-editiontype.html", "distribution-spelling-errors-files-filetype.html", "distribution-spelling-errors-files-institution.html"). Accessed April 4, 2020.

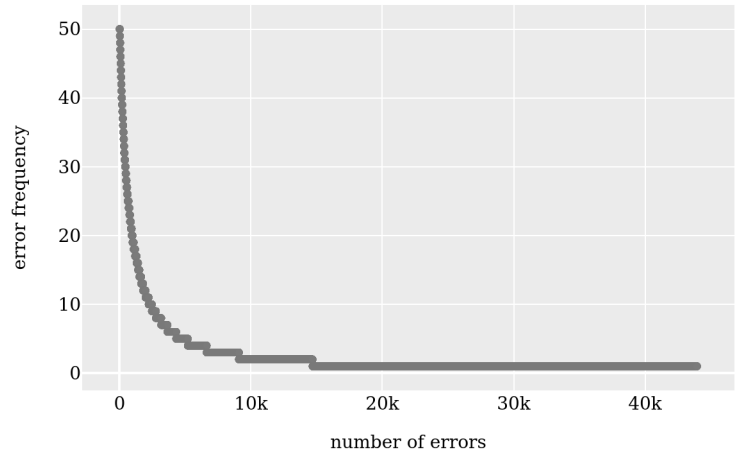


Figure 18. Distribution of spelling errors with exception words.

Looking at the results for all the exception lists together displayed in figure 17, it becomes clear that proper names are by far the most frequent *false error* tokens, but that also certain morphological constructions, in particular the verb forms with pronoun suffixes and diminutives, play an important role. On the other hand, some types of words that one could have expected to be more significant are stylistically marked words such as foreign words, specialized or archaic vocabulary, and words representing oral speech. As the figure shows, at least among the most frequent errors, they are not decisive. In sum, the exception lists cover 344,339 tokens (63 % of all the error tokens) and 18,197 types (27 % of all the error types), so they helped to clean the spell check results considerably.

The number of errors that remain is 121,442 tokens, which is 0.7 % of all the tokens in the corpus, and 43,955 types, which is 22 % of the whole corpus vocabulary. Of these, 29,266 (15 % of the vocabulary and 0.2 % of the tokens) occur only once, and 2,212 types and 47,670 tokens occur more than ten times (i.e., 1 % of the vocabulary and 0.3 % of the tokens). Figure 18 shows the distribution of the remaining errors.

Compared to the previous error distribution, the curve is not so steep anymore, but still, relatively few errors are frequent. To clean up the remaining individual errors would be far too time-consuming, but also the other residual errors comprise several thousand entries. A final aspect worth considering is how many misspelled words there are per novel in the corpus. This is summarized in figure 19.

Both for error tokens and types, the mean (474 and 181) is higher than the median (351 and 204, respectively), meaning that there are several outliers with many errors. Indeed, the ranges go from 19 to 2,437 error tokens and from 18 to 1,664 error types. As the novels are of different lengths, and it is probable that this influences the number of errors, the same distribution is shown again in relative numbers in figure 20.

Now the mean error rate for tokens is at 0.7 % and for types at 2.4 %, and the medians are at 0.6 % and 2.2 %, respectively. That the number for types is higher follows from the above observation that most of the remaining errors are individual ones. The figure shows that the

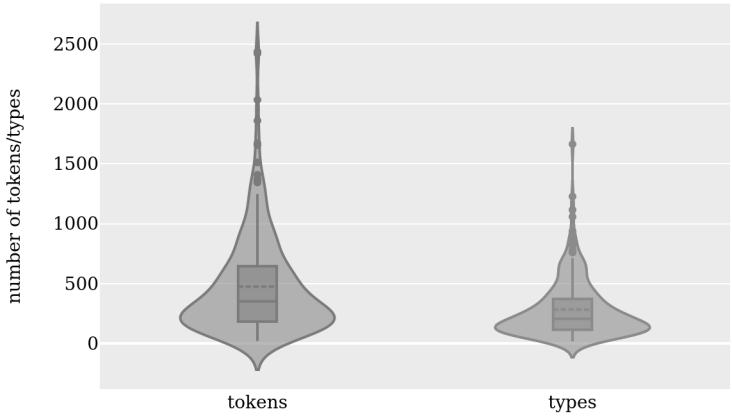


Figure 19. Distribution of error tokens and types for the corpus files (absolute).

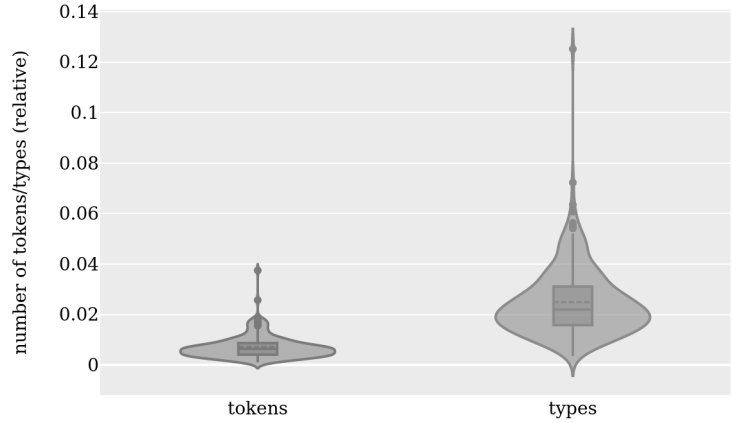


Figure 20. Distribution of error tokens and types for the corpus files (relative).

spread is much smaller for tokens than for types, meaning that the correction of the most frequent errors contained in the texts that were included in the corpus from various sources helped to level the token error rate. Nevertheless, because individual errors were not corrected systematically, the range of the error type rate is more extensive, going from 0.4 % to 12.5 %. A way to look for factors that might have influenced the text quality is to combine the information about errors with the metadata about the sources of the texts. In figures 21 to 23, the distributions of error tokens and types are charted distinguished by the type of edition used (first, historical, modern, or unknown), by the source file type (image versus text), and by the different source institutions.[292]

A look at the type of source editions also confirms that the token error rates could be reduced to a similar level by the text treatment procedure, independently of the type of edition used. In contrast, the type error rates differ slightly. Their median is highest for texts where the kind of source edition is unknown (2.5 %) and lowest for modern editions (2.0 %), while first and historical

[292] For all of these figures, values relative to text length were used.

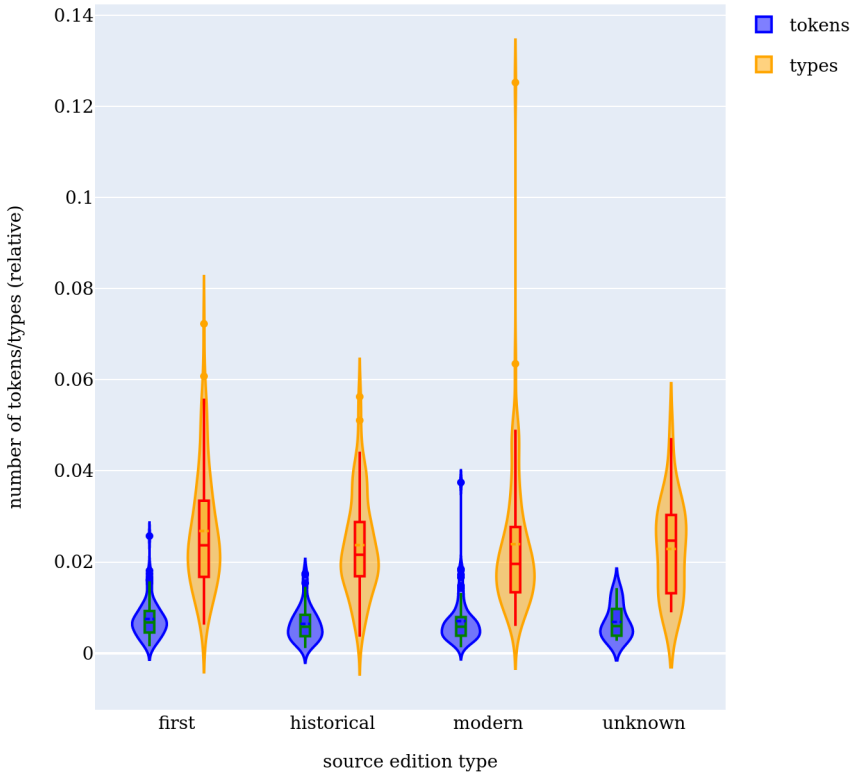


Figure 21. Distribution of error tokens and types for the corpus files (by type of source edition).

source editions lie in between. There is a notable outlier of a modern edition with 12.5 % of error types. This is the science fiction novel “En busca del eslabón. Historia de monos” (1888, CU) by Francisco Calcagno. It contains 1,664 different errors, but most of them are exception words: proper names, foreign words, scientific and other special, also invented vocabulary that was not covered by the exception lists created above, e.g., “Blumenbach”, “Goethe”, “link”, “chimp”, “gibones”, “hisquiáticas”, “niamsniams”, “Ibizapitanga”, or “Sinonimolandia”.[293] This example shows to what extent checks of the text quality can be obstructed by special vocabulary.

In figure 22, the error rate distributions are distinguished by the source file type. Here the median of the type rates is a bit higher for texts that were extracted from image files (2.3 %) than those that were collected from text files (2.0 %), showing on the one hand that the OCR process entails that a certain amount of spelling errors is introduced into the texts, but also that existing full-text files are usually not free from errors.

Finally, the distribution of error rates is differentiated by the source institutions in figure 23. Again, as a result of the text treatment process, the medians of the token error rates are quite similar throughout the different institutions. Regarding the type error rates, there is a bit more

[293] See the spell-check result file for this novel at https://github.com/cligs/data-nh/blob/master/corpus/text-treatment/spellcheck_nh0215.csv. Accessed April 5, 2020.

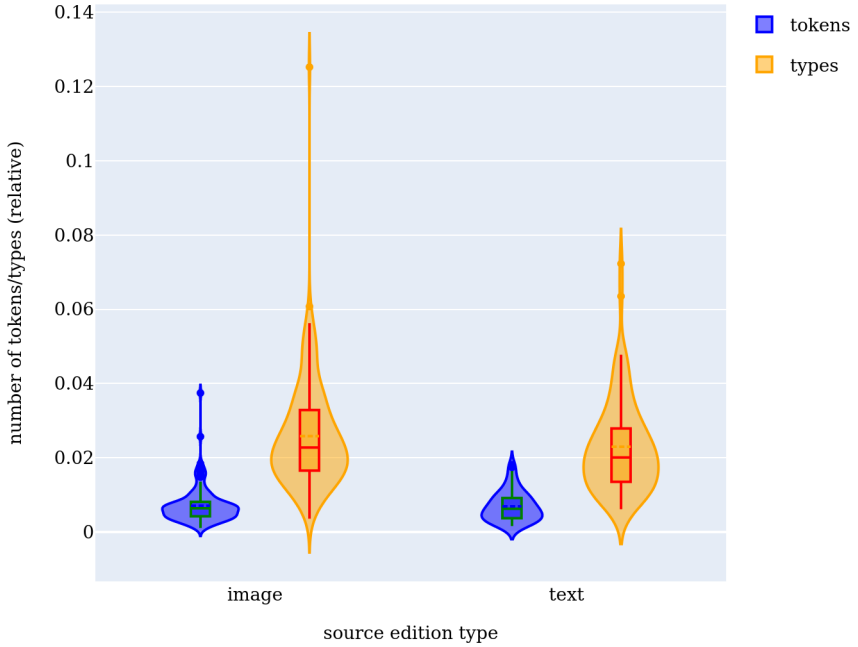


Figure 22. Distribution of error tokens and types for the corpus files (by source file type).

variation from institution to institution. Sources with very good rates are, for example, the "Biblioteca Digital Argentina" (BDA) and the digital library "La novela corta" with a median of 1.5 % each. There are higher rates, for example, for the iBooks Store (4.4 %) and Conaculta (4.1 %). Interestingly, the files from the BDA and the iBooks Store were processed as text, while the ones from "La novela corta" and Conaculta went through the OCR process.

Summing up, the process of text treatment that was necessary for the creation of the corpus at hand involved different steps ranging from rather simple structural conversions of marked-up files to a whole pipeline of digitization in other cases, because many different sources had to be used in order to gather a corpus of Argentine, Cuban, and Mexican novels of considerable size. When so many different types of sources are used, it is especially important to check the quality of the incoming texts to make sure that errors in the texts do not skew the results of later analyses too much. For this corpus, a spell check was performed using a standard dictionary for modern Spanish, and the results were refined through the creation of corpus-specific exception lists. That way, a certain quality of the texts could be assured and achieved. Furthermore, the spell check revealed some peculiarities of the corpus vocabulary, such as the existence of many verb forms with pronoun suffixes. Knowledge about them is helpful when the texts are further processed. However, the analysis of the "false" and real spelling errors also revealed that it is hardly possible to create a corpus of perfect text quality, at least when the range of source edition types, file types, and institutions is broad. It also became clear that spelling exceptions and errors are influenced by a lot of factors: the mentioned kinds of sources, but also the kinds of novels.

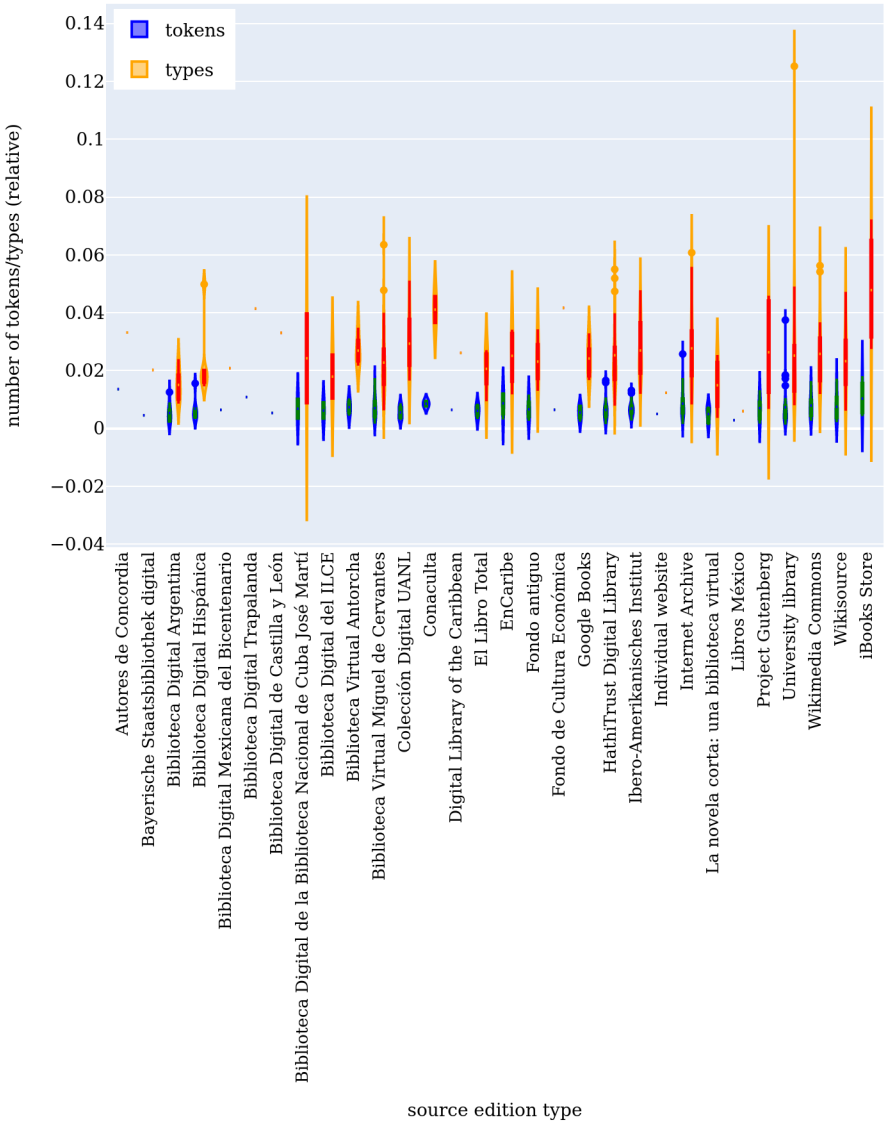


Figure 23. Distribution of error tokens and types for the corpus files (by source institution).

3.3.3 Metadata and Text Encoding

Starting from the basic structured full texts that were prepared according to the processing steps described in the previous section, each novel in the corpus was enriched with metadata and further structural markup. In CLiGS, we decided to use a common data model for all the text collections produced in the context of the project based on the text encoding standard of the TEI in version P5. It is not so common for large-scale text analysis projects to use XML-based markup, though. In most cases, large corpora consisting of simple plain text files are used together with metadata indicated directly in the file names or stored in tabular format.[294] The decision for the TEI standard was made here because the analysis of genres and subgenres rests on detailed metadata about the texts that cannot easily be represented in simple tables. As is the case for the digital bibliography presented above, also the metadata for the full-text corpus of novels (the corpus at hand, but also the other corpora of narrated and dramatic texts produced in the CLiGS project) is best recorded in a model that allows indicating responsibilities (who entered the information?) and degrees of certainty (how sure was the person who entered the information that it is correct?). Furthermore, it is important that the metadata can be structured further (e.g., through the addition of markup in bibliographic information or the indication of levels of metadata). The main text also profits from the possibilities of markup. It would also be possible to infer paragraph or chapter boundaries from plain text files (for example, via the use of blank lines), but a structure of hierarchical markup allows to differentiate between main parts, chapters, subchapters, headings, and paragraphs, and inserted texts, such as letters, verse lines, or dramatic speech. All this kind of structural information can then be used in the analyses of the texts. Moreover, because the TEI is an encoding standard widely used in the digital humanities, the reuse of the files produced in the CLiGS project in other contexts is facilitated, so the usage of this standard can be considered a sustainable solution.[295] In this chapter, the TEI-based data model developed for the corpus is presented, starting with the elements and attributes used to encode the metadata collected for the novels (in chapter 3.3.3.1) and going on with how the structures of the textual body were encoded (in chapter 3.3.3.2). XML snippets, mainly from one novel, are included as examples. Where aspects of the text encoding need to be clarified further for the whole corpus, they are discussed in connection with the individual examples, e.g., the declaration of rights for the TEI files.

In the corpus, each text was stored as an individual TEI file. The file names consist of a shortcut for the corpus, in this case, “nh” (“novelas hispanoamericanas”) plus four digits for a serial number, so the first file in the corpus has the file name “nh0001.xml” and the last one “nh0256.xml”.[296] Because the file names are unique, they are, at the same time, the identifiers of the novels in

---

[294] See, for example, the “Corpus of German-Language Fiction”, consisting of almost 3,000 prose works in plain text format or the corpora prepared by the Computational Stylistics Group (Fischer and Strötgen 2017, Computational Stylistics Group 2023).

[295] The overall approach that was used in the CLiGS project to encode literary corpora in TEI is described in Calvo Tello, Henny-Krahmer, and Schöch (2018) and Schöch et al. (2019).

[296] Obviously, for the corpus at hand, three digits would have been enough, but it was decided to use four because other corpora in the CLiGS project are more extensive. In addition, future extensions of the corpora should be possible.

the corpus (the so-called “CLiGS identifiers”). That way, they can be referenced elsewhere, for example, in the digital bibliography, and they can also used to identify the texts in analyses.

3.3.3.1 TEI Header

In general, in TEI, the metadata is encoded in the TEI header, which contains descriptive and declarative metadata associated with the digital resource. Of the five principal components that are available for the TEI header, four were used in the TEI model here:

- the file description, which contains bibliographic information,
- the encoding description, in which it is documented which kind of information was encoded when the digital file was created based on one or several other source files,
- the profile description, which includes information about non-bibliographic aspects of the texts,
- and the revision description, in which the revision history of the file is given.[297]

In the following, each of these parts is presented, using the TEI file of the novel “Adoración” (1894, CU) by Álvaro de la Iglesia as an example. Particular focuses are the declaration of rights for the corpus files in chapter 3.3.3.1.2 and the text classification with keywords in chapter 3.3.3.1.6.

3.3.3.1.1 Title and Publication Statements

The file description is primarily used for the encoding of bibliographic information about the digital file itself, but also about its sources. Example 7 shows the first part of this section of the TEI header, the title statement.

```
<titleStmt>
  <title type="main">Adoración</title>
  <title type="short">Adoracion</title>
  <title type="sub">Novela original</title>
  <title type="idno">
    <idno type="viaf">-</idno>
    <idno type="bibacme">W923</idno>
  </title>
  <author>
    <name type="full">Iglesia, Álvaro de la</name>
    <name type="short">IglesiaA</name>
    <idno type="viaf">120788045</idno>
    <idno type="bibacme">A367</idno>
  </author>
  <principal xml:id="uhk">Ulrike Henny-Krahmer</principal>
</titleStmt>
```

Example 7. Title statement of the novel “Adoración”.

It contains the different parts of the work’s title. In the example, there are a main title (“Adoración”) and a subtitle (“Novela original”). In addition, a short title without blank spaces and accents is given that can be used as a shortcut, for example, in the visualization of results (“Adoracion”). The shortcut is especially useful if the title of the novel is longer than the one in

[297] For general documentation of the TEI header, see the corresponding chapter in the TEI guidelines (Text Encoding Initiative Consortium 2023c).

this example. Other possible elements of the title, which are not present in this example, are the title of a series the novel belongs to (<title type="series">), an alternative title (<title type="alt">), and title parts (<title type="part">).[298] Where a novel is registered as a work in the "Virtual International Authority File" (VIAF), this number is given in a title element of the type "idno" (<idno type="viaf">). In the present example, no such identifier is available. Another identifier is added to connect the corpus with the digital bibliography: for each novel, its work ID in Bib-ACMé is encoded (<idno type="bibacme">). That way, additional information can be retrieved both ways, from the bibliography to the corpus and vice versa. The second part of the title statement consists of information about the author. Like the work's title, also the author's name is given in a full version (<name type="full">) and a short version (<name type="short">). For some authors, also pseudonyms are given (<name type="pseudonym">) if they published novels under that name. If available, the authors are identified with a VIAF number, as well (<idno type="viaf">), and also their ID in Bib-ACMé is indicated (<idno type="bibacme">).[299] Finally, the responsibilities of the people involved in the creation of the TEI file of a novel are indicated as part of the title statement. In the case at hand, the file was created and edited just by one principal investigator. In other cases, further responsibility statements are included.

```
<extent>
  <measure unit="words">44670</measure>
</extent>
<publicationStmt>
  <publisher>
    <ref target="http://cligs.hypotheses.org/">CLiGS</ref>
  </publisher>
  <availability status="free">
    <p>This work is in the public domain. It is provided here with the <ref target="
        https://creativecommons.org/publicdomain/mark/1.0/deed.de">Public Domain Mark
        Declaration</ref> and can be re-used without restrictions. The XML-TEI markup is
         also considered to be free of any copyright and is provided with the same
        declaration.</p>
  </availability>
  <date>2020</date>
  <idno type="cligs">nh0018</idno>
  <idno type="url">https://github.com/cligs/conha19/blob/master/tei/nh0018.xml</idno>
</publicationStmt>
```

Example 8. Extent and publication statement of the novel "Adoración".

[298] An example of a series title is "Dramas militares", because there are several novels by the Argentine writer Eduardo Gutiérrez associated with this label, e.g., "El Chacho" (1884, AR) and its sequels. An alternative title means that the novel has been published under different titles. Sometimes, the title of a novel changes from the first edition to subsequent ones. For example, the novel "Amar al vuelo" (1884, AR) by Enrique E. Rivarola was first published with the title "El arma de Werther". Furthermore, different editions of the novels often have different subtitles, but these are encoded as several titles of the type <title type="sub"> and not as alternative main titles. The tag <title type="part"> is only used for one special case in the corpus: the novel "Pepa Larrica" (1884, AR) by Rafael Barreda was interpreted as one work consisting of three parts that were published separately with their own title ("Las dos tragedias", "La confesión de un médico", and "Religión o muerte"). See chapter 3.1.1.5 above, where this decision is explained. Where novels are published under one main title but with several parts ("Parte primera: …", "Parte segunda:…", etc.), the titles of the parts are only encoded as headings in the main body of the text and not in the title statement of the TEI header.

[299] The VIAF entry of Álvaro de la Iglesia is available at OCLC (2010–2021b).

After the title statement, the file description continues with a part on the extent of the novel. It contains an element documenting the number of words in the novel (`<measure unit="words">`, see example 8 above). Words are understood as tokens here, and their number is counted with a simple regular expression in Python applied to the main body of the novel's text, excluding headings and notes (`tokens = re.split(r"\W+", text, flags=re.MULTILINE)`). Many other measures could be included in the TEI header, for example, the number of chapters, paragraphs, sentences, characters, and so on. However, because all of these measures can be determined programmatically and are not adjusted manually here, it was decided only to note the number of words because this measure is basic to characterize the files in the corpus and is used very often. Other measures can be calculated ad hoc when needed. Next, information concerning the publication of the TEI file is given. This includes the indication of the publisher, in this case, the project CLiGS. Furthermore, details about the availability of the text are encoded. The question of access to the TEI files needs some more discussion and is explained further below. Additional parts of the publication statement are the year in which the TEI file was first published (`<date>`), the CLiGS identifier (`<idno type="cligs">`), which is also used for the file names, and a URL pointing to the repository where the file is published (`<idno type="url">`).[300]

3.3.3.1.2 Declaration of Rights

Regarding the availability of the TEI files, their status can be either "free" or "restricted". The TEI files of all the free texts are published with the Public Domain Mark Declaration, allowing the reuse of the files without restrictions (see Creative Commons n.d.). Almost all the texts of the corpus are in the open domain according to German copyright laws. In Germany, a work becomes free from copyright 70 years after the author's death (Bundesamt für Justiz n.d.a). An overview of the authors' death years is given in figure 24.[301]

If one takes the year of 2022 as a reference point, there is only one author of novels in the corpus who died after 1953: the Argentine writer Enrique Larreta (1875–1960). There is one novel written by Larreta in the corpus, so the TEI file of this novel can only be published in 2030.[302] In addition, there are 13 authors whose years of death are unknown. In such cases, the German rule is that the copyright expires 70 after the first publication of the work.[303] Because all the works in the corpus were first published at the latest in 1910, the novels of these authors are all in the open domain.[304] In figure 25, the years of the novels' first editions are displayed.

---

[300] See chapter 3.3.5 below for further information about the publication of the corpus.

[301] Figures 24 to 27 were produced with the following Python module: https://github.com/cligs/scripts-nh/blob/master/corpus/metadata_encoding/corpus_copyright.py. The resulting charts "authors-death-years.html", "first-publication-years.html", "base-publication-years.html", and "copyright-status.html" can be downloaded at https://github.com/cligs/data-nh/tree/master/corpus/metadata-encoding. Accessed April 6, 2020.

[302] Until then, the TEI file of this novel is kept in a private repository that is part of the GitHub space of the CLiGS project: https://github.com/cligs/novelashispanoamericanas. Accessed March 20, 2020. The novel in question is "La gloria de Don Ramiro" (1908, AR) by Enrique Larreta.

[303] In the law, the rule is formulated as applying to anonymous and pseudonymous works for which the author's name cannot be verified. Here, the names of the authors are known, but their dates of death are not, so the regular law cannot be applied (Bundesamt für Justiz n.d.b).

[304] The authors concerned are Ventura Aguilar (?–?, AR), C. M. Blanco (?–?, AR), Rodolfo Díaz Olazábal (?–?, AR), Silverio Domínguez (1852–?, AR), José Rafael Guadalajara (1863–?, MX), Ramón Machali (?–?, AR), Vicente Morales

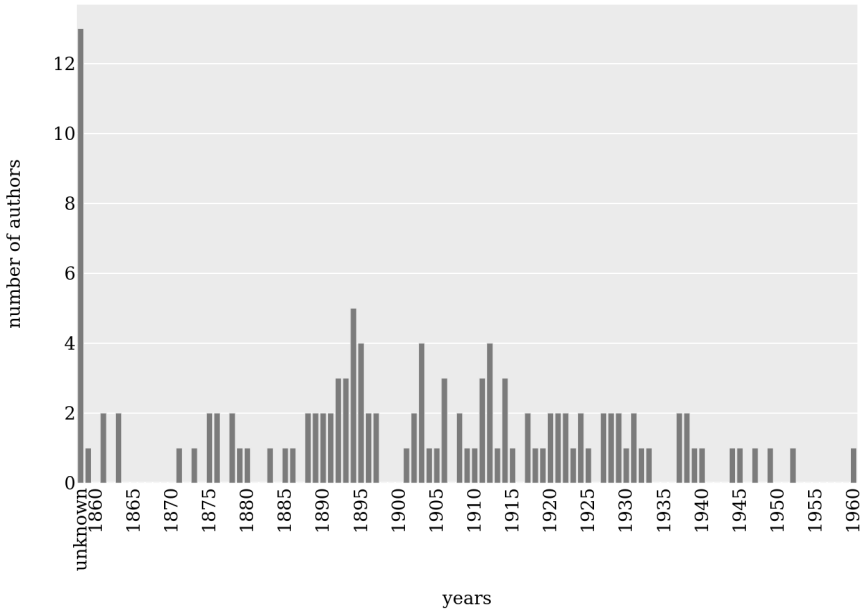


Figure 24. Death years of authors.

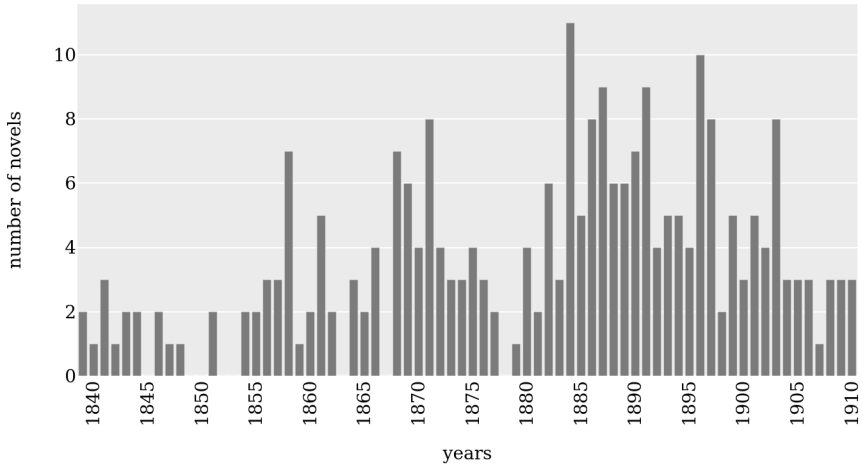


Figure 25. Years of the novels' first publications.

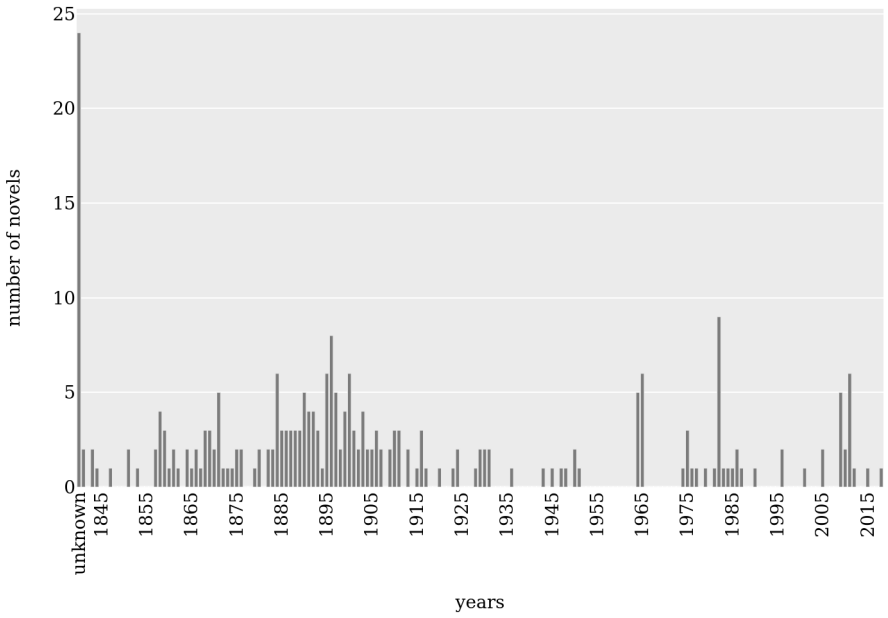


Figure 26. Publication years of basis editions.

Part of the German copyright law is also the ancillary copyright protecting, for example, scholarly editions of works that are, in principle, free. This protection ends 25 years after the publication of the edition (Bundesamt für Justiz n.d.c). This law is relevant for the corpus because also modern editions were used to extract the texts of the novels. Figure 26 shows the publication years of the editions that were used as a basis for the TEI files in the corpus. These publication years refer to print editions when these were used directly, to print editions underlying a digital reproduction, or to digital editions that form a new textual basis and are not considered simple reproductions.[305]

Among the novels, there are twelve novels whose text was extracted from print editions that were published after 1997 and for which access is also restricted here, as indicated in example 9.[306]

(?–?, MX), Pedro G. Morante (?–?, AR), Margarita Rufina Ochagavia (1840–?, AR), Andrés Portillo (?–?, MX), Pedro Robles (?–?, MX), Mercedes Rosas de Rivera (?–?, AR), and Victorio Sylva (?–?, AR). Some of the dates could possibly be found out with more rigorous historical research, but checking VIAF, the bibliographies mentioning the works of these authors as well as general searches on the web, did not lead to any results.

[305] The difference between the latter two is not always easily identified. Besides the characteristics of the digital edition itself, it was also taken into account here if copyright is claimed by the editors of the digital edition or not. This is discussed in more detail below for the novels in question.

[306] The works in question are the following (the dates in parentheses indicate the year of the first edition/year of the digital or print edition used/year of the expiration of the protection): “Astucia” (1866/2005/2030, MX) by Luis Gonzaga Inclán, “Dos partidos en lucha” (1875/2005/2030, AR) and “El tipo más original” (1879/2001/2026, AR) by Eduardo Ladislao Holmberg, “El espejo de Amarilis” (1902/2011/2036, MX) by Laura Méndez de Cuenca, “Antón Pérez” (1903/2011/2036), “Juanita Sousa” (1890/2011/2036, MX), “Pocahontas” (1882/2011/2036) and “Previvida” (1906/2011/2036, MX) by Manuel Sánchez Mármol, “Clemencia” (1877/2012/2037, AR) and “La huella del crimen”

```
<availability status="restricted">
  <p>This file is prepared for personal research use only and not for publication
      because the ancillary copyright of the underlying print edition has not yet
      expired according to German law.</p>
</availability>
```

Example 9. Restricted access for the novel “María de Montiel”.

Other cases that need to be clarified are novels where digital editions are available, but the underlying print editions are unknown. As long as the works themselves are in the open domain and no special rights are declared for the digital editions, it is assumed here that these editions are not considered new scholarly revisions of older editions but reproductions of existing historical editions. In consequence, the publication of the corresponding TEI files should be unproblematic.[307] Next, there are some cases of digital editions for which copyright is claimed because they constitute new scholarly preparations of old texts that are themselves out of copyright. All of these novels were retrieved from the portal “La novela corta: una biblioteca virtual” (Universidad Nacional Autónoma de México 2008–2023).[308] In one case, the underlying print edition is unknown, and in five cases, it is known but is itself not affected by the ancillary copyright. Nevertheless, because these digital editions can be considered scholarly editions and copyright is claimed for them, they are interpreted as falling under the ancillary copyright and are therefore classified as “restricted” here. Finally, there are two more cases that are not very clear. Two novels were downloaded from the “Biblioteca Digital del Instituto Latinoamericano de la Comunicación Educativa” (ILCE), “La Rumba” (1891, MX) by Ángel de Campo y Valle and “El diablo en México” (1858, MX) by Juan Díaz Covarrubias. Both novels can be downloaded as PDF files. In the first case, the edition only contains the base text but no introduction, notes, or other scholarly commentary, and it is not indicated on what print edition the digital one is based. However, an organizational editor and a publication year are indicated and the following claim is made: “Las particularidades de esta edición están protegidas por derechos de autor” (Campo y Valle 2009). In the second case, the underlying print edition is also unknown. In addition, the publication date of the digital edition is not given, no indication of an individual person responsible for the creation of the edition is made, there is no introduction, and there are no notes. However,

(1877/2009/2034, AR) by Luis Vicente Varela, “María de Montiel” (1861/2010/2035), AR) by Mercedes Rosas de Rivera, and “Stella” (1905/2011/2036, AR) by Emma de la Barra. In principle, it would be possible to examine in detail to what extent the used print editions are scholarly editions that differ significantly from previous editions. However, for the sake of simplicity and to avoid legal ambiguities, all the affected TEI files are kept unpublished until the ancillary copyright expires.

[307] This applies to 22 novels obtained from the following sources: “Wikisource” (6 novels), “Biblioteca Digital Argentina” (4), “Biblioteca Virtual Antorcha” (4), “El Libro Total” (3), “Project Gutenberg” (2), “Autores de Concordia” (1), “EnCaribe” (1), “Individual website” (1).

[308] The following novels are concerned: “Antonia” (1872, MX) by Ignacio Manuel Altamirano, “Confesiones de un pianista” (1873, MX) by Justo Sierra Méndez, “Historia vulgar” (1904, MX) by Rafael Delgado, “Los fuereños” (1883, MX) by José Tomás de Cuéllar, “Los maduros” (1882, MX) by Pedro Castera, and “¡Vendía cerrillos!” (1889, MX) by Federico Gamboa. Four were edited in 2009, one in 2010, and one in 2018, so the ancillary copyright will cease in 2034, 2035, and 2043, respectively. In the virtual library, five of these novels are contained in the collection “Novelas en tránsito – Primera serie”. Unfortunately, these digital editions are not retrievable anymore. Another related collection containing the sixth novel is still accessible: “Novelas en tránsito – Segunda serie”. There, it can be seen that the editions of the portal are prepared according to scholarly standards and that copyright is claimed by the “Universidad Nacional Autónoma de México” in the PDF versions of the editions (see, for instance, Sierra 2018).

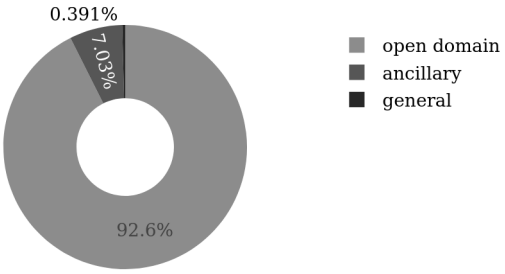


Figure 27. Copyright statuses of the novels in the corpus.

at the end of the PDF file, the following advice is given: "Material autorizado sólo para consulta con fines educativos invariablemente como fuente de la información la expresión 'Edición, culturales y no lucrativos, con obligación de citar digital. Derechos Reservados. Biblioteca Digital © Instituto Latinoamericano de la Comunicación Educativa ILCE'" (Díaz Covarrubias n.d.). Although copyrights are declared, these two editions are not considered as falling under the ancillary copyright because no added scholarly value is visible. They are therefore classified as "free" here.

So, in total, there is one novel that is still protected by the general copyright and 18 by the ancillary copyright. As a consequence, there are, in total, 19 of the 256 TEI files of the corpus that cannot be published immediately.[309] This information is illustrated in figure 27.

The discussion of copyrights shows that preparing a digital full text and TEI corpus of novels poses some challenges in this regard. Whereas the determination of the general copyright is relatively clear because it depends on the authors' death dates and the dates of the first publication of the works, the German ancillary copyright is often more difficult to assess. First, existing source editions can be of very different kinds: print sources, images, PDF files, plain texts, or web pages. The relationship between originals, reproductions, and edited versions is not always clear because it is not always explained, and in some cases relevant information is missing. Moreover, the legal status of source editions can be difficult to determine when no publication dates or responsibilities for their creation are given. On the other hand, some claims for copyright on material in the open domain are exaggerated. Another problem is that web resources are not necessarily stable, not even if they are published by a scholarly institute. They may cease to be accessible after some years so that information that is relevant to the editions' legal status cannot be retrieved anymore. In other cases, updates of contents that were produced earlier postpone the publication date and thereby also the end of the ancillary copyright.

3.3.3.1.3 Source Description

Apart from the title statement, information about the extent of the novel, and the publication statement, the file description in the TEI header also contains the source description, in which

[309] A tabular overview of the corpus metadata that is relevant for copyright questions is available at https://github.com/cligs/data-nh/blob/master/corpus/metadata_copyright.csv. Accessed April 6, 2020.

details about the sources that the digital text was derived from are encoded in the form of bibliographic references (see example 10).

```
<sourceDesc xml:base="https://raw.githubusercontent.com/cligs/bibacme/master/app/data/
    editions.xml">
  <bibl type="digital-source" xml:id="DS">Iglesia, Álvaro de la. "Adoración. Novela
      original." <seg rend="italic">HathiTrust Digital Library</seg>, <ref target="
      https://catalog.hathitrust.org/Record/009049820">https://catalog.hathitrust.org/
      Record/009049820</ref>. Accessed 31 April 2018.</bibl>
  <bibl type="print-source" n="222" xml:id="PS" corresp="#E1786">Iglesia, Álvaro de la.
      <seg rend="italic">Adoración. Novela original.</seg> Barcelona: Ed. F. Granada, <
      date when="1906">1906</date>. 222 p.</bibl>
  <bibl type="edition-first" xml:id="E1" corresp="#E1280">Iglesia, Álvaro de la. <seg
      rend="italic">Adoración. Novela original.</seg> Matanzas: Imprenta de la
      Propaganda, <date when="1894">1894</date>.</bibl>
</sourceDesc>
```

Example 10. Source description of the novel "Adoración".

Three main types of bibliographic references are included in the source description: the first one documents which digital source was used, the second reference describes the print source underlying the digital source edition, and the third one documents the first known edition of the novel. The date of the first edition is the one generally referred to when the novels are mentioned in this dissertation and also when they are analyzed. In the case of the novel "Adoración", digital images were retrieved from the "Hathi Trust Digital Library" and were used to extract the full text. Here, the underlying print edition is a historical one from 1906, but not the first one, which was published in 1894. In other cases, the used print edition may correspond to the first known edition so that the entries "PS" and "E1" reference the same edition. For some novels in the corpus, there is no digital source (when print editions were used directly), and for others, the print source of the digital source edition is unknown, so there may also be just two levels of sources. On the other hand, more than three sources may be listed in cases where different front matters of historical editions were transcribed to extract genre labels occurring on them. In these cases, further bibliographic entries of the type "edition" are added. The attribute `@corresp` is used on the bibliographic entries to indicate to which edition they correspond in the bibliography Bib-ACMé, in which more structured bibliographic descriptions of the editions can be found. The identifiers pointed to in this attribute can be resolved using the base URI indicated in `@xml:base` on the element `<sourceDesc>`.

3.3.3.1.4 Encoding Description

After the file description, the TEI header continues with the encoding description. A short general description of the text treatment and text encoding is given in each file of the corpus, as example 11 shows.

```
<encodingDesc>
  <p>The source PDF file was processed with OCR. The software used was ABBYY Finereader
      12 Professional, with Spanish as recognition language. The result of the OCR
      process was checked, but due to temporal restrictions, corrections were only made
      in a rough manner and remaining errors cannot be excluded.</p>
  <p>The spelling was checked and corrected where appropriate.</p>
  <p>The following phenomena were marked up: front matter (where available, e.g. title
      page, dedication, preface, introduction), part and chapter divisions, headings,
```

```
      paragraphs, inserted texts (e.g. letters or newspaper articles), direct speech or
      thought, verse lines, dramatic text, quotations (e.g. epigraphs), notes by the
      author, and gaps.</p>
</encodingDesc>
```

Example 11. Encoding description of the novel “Adoración”.

The phenomena that were marked up in the texts are explained further below. The encoding description is followed by the profile description, where non-bibliographic metadata about the texts is documented. For the corpus at hand, two sections of the profile description are used: abstracts and text classification.

3.3.3.1.5 Abstracts

If available, abstracts summarizing the content of the novels or containing comments on the novels made by literary historians are given. For the novel “Adoración”, a description of the plot coming from the preface of the novel itself is quoted. A section of the abstract is reproduced here in example 12.

```
<abstract source="#Iglesia_1906">
  <p>
    <quote><p>Es el caso de un joven, casi adolescente, que está enamorado a la vez de
        dos jovencitas. El mismo narra el desarrollo de ese complejo estado de
        conciencia, a que asiste en cierto modo, a veces aturdido, a veces espantado,
        sin acertar a explicárselo [...]</p></quote>
    <bibl>Varona, Enrique José. "Prólogo." In: Iglesia, Álvaro de la. <seg rend="italic"
        >Adoración. Novela original.</seg> Barcelona: Ed. F. Granada, <date when="1906">
        1906</date>.</bibl>
  </p>
</abstract>
```

Example 12. Abstract of the novel “Adoración”.

The source of the abstract is encoded as a bibliographic citation (`<bibl>`), and, in addition, it is indicated in the attribute `@source` with a pointer to an external list of bibliographic references.[310] The abstract itself is encoded as a quotation (`<quote>`) that is structured further with paragraph elements if needed. Each TEI file can contain none, one, or several abstracts. The abstracts are helpful in getting an overview of the content of the novels when the results of the genre analyses are interpreted.

3.3.3.1.6 Text Classification with Keywords

Besides the abstract, the profile description also contains the element `<textClass>` (“text classification”). In general, this element is used to “group[…] information which describes the nature or topic of a text in terms of a standard classification scheme, thesaurus, etc.” (Text Encoding Initiative Consortium 2023j). Inside `<textClass>`, the `<keywords>` element is used to group a list of keywords describing the nature of the text from various perspectives, for example, the

[310] The external list is a file named “bibliography.xml”. In it, the bibliographic references cited for abstracts and subgenre assignments throughout the corpus are collected. The filename is not used as part of the pointer in order to keep the references short. See https://github.com/cligs/conha19/blob/master/bib/bibliography.xml. Accessed April 8, 2020.

genre or the setting of the novel. To illustrate the usage of this taxonomic system for the corpus, the list of keywords for the novel “Adoración” is represented in example 13.

```
<textClass>
  <keywords scheme="../schema/keywords.xml">
    <term type="author.continent">America</term>
    <term type="author.country">Cuba</term>
    <term type="author.country.birth">Spain</term>
    <term type="author.country.death">Cuba</term>
    <term type="author.country.nationality">Cuba</term>
    <term type="author.gender">male</term>
    <term type="text.source.medium">digital</term>
    <term type="text.source.filetype">image</term>
    <term type="text.source.institution">HathiTrust Digital Library</term>
    <term type="text.source.edition">historical</term>
    <term type="text.publication.first.country">Cuba</term>
    <term type="text.publication.first.medium" cert="medium">book</term>
    <term type="text.publication.first.type" cert="medium">independent</term>
    <term type="text.publication.type.independent" cert="medium">yes</term>
    <term type="text.language">Spanish</term>
    <term type="text.form">prose</term>
    <term type="text.genre.supergenre">narrative</term>
    <term type="text.genre">novel</term>
    <term type="text.title" n="1894">Adoración. Novela original</term>
    <term type="text.title" n="1901">Adoración. Novela original</term>
    <term type="text.title" n="1906">Adoración. Novela original</term>
    <term type="text.genre.subgenre.title.explicit">novela original</term>
    <term type="text.genre.subgenre.title.implicit" resp="#uhk">novela sentimental</term
        >
    [...]
    <term type="text.narration.narrator">autodiegetic</term>
    <term type="text.narration.narrator.person">first person</term>
    <term type="text.speech.sign">—</term>
    <term type="text.speech.sign.type">single</term>
    <term type="text.setting.continent">America</term>
    <term type="text.setting.country">Cuba</term>
    <term type="text.time.period">unknown</term>
    <term type="text.time.period.author">contemporary</term>
    <term type="text.time.period.publication">contemporary</term>
    <term type="text.prestige">low</term>
  </keywords>
</textClass>
```

Example 13. Keywords for the novel “Adoración”.

The types of keywords are encoded in `<term>` elements that are specified further by the attribute `@type`. The keyword values are given as the content of the `<term>` elements (e.g., `<term type="text.genre">novel</term>`). The whole system of keywords is regulated by an external taxonomy referenced from the `<keywords>` element (`<keywords scheme="../schema/keywords.xml">`). The taxonomy, explained further below, defines which types of keywords exist and which values they can take. In general, the keyword types are organized hierarchically with the goal of systematizing the different kinds of metadata. In the values of the terms’ `@type` attribute, the different levels of the hierarchy are separated by a dot, so the type `"text.genre"`, for example, refers to the keyword level “text” and to the sublevel “genre”. For some keyword types, the list of possible values is closed, meaning that only certain specific values are allowed, and for others, it is open, depending on the kind of information. For instance, there is a keyword

type referring to the narrative perspective of the text (<term type="text.narration.narrator">). Only three keyword values are possible for this type: "autodiegetic", "homodiegetic", and "heterodiegetic". On the other hand, the values of the keywords concerning explicit mentions of the subgenre of the text (e.g., <term type="text.genre.subgenre.title.explicit">) are not previously defined.

In the above example, all the types of keywords used for the corpus are included, except the terms used to record the subgenre of the novels, because these are described in more detail in the next chapter 3.3.4, in which the assignment of subgenre labels to the corpus is explained. As can be seen, there are two general groups of keywords: on the one hand, keywords about the author of the text and, on the other hand, keywords about the text itself. Information about the author is already present in the title statement (her or his name and the VIAF and Bib-ACMé identifiers) as well as in the digital bibliography. Furthermore, an author can be the creator of several novels in the corpus, so that the information is eventually repeated in the metadata of several texts. Nevertheless, some authorial metadata is encoded in the keyword system because it is especially relevant for the analysis of the novels. That way, it is not necessary to retrieve this information from external files every time that the novels are analyzed. Furthermore, even though all the TEI files of the corpus are embedded in a *corpus ecosystem*, including the digital bibliography, a keyword taxonomy, and schema files, it should be possible to reuse a subset or individual files of the corpus without the necessity to rebuild the whole system. Therefore, some metadata that is considered essential for the stylistic analysis of the novels is repeated to make the TEI files more self-contained.

The authorial keywords concern the gender (<term type="author.gender">) as well as the geographic, cultural, and national belonging of the author ("author.continent", "author.country", "author.country.birth", "author.country.death", "author.country.nationality"). The values for the continent and country correspond to the general assignment of an author to one of the three countries covered with the corpus (Argentina, Cuba, and Mexico).[311] With the additional terms, the assignment to a country is differentiated further because authors can have a different country of birth, death, or nationality than the one to which they are generally assigned. The author of the novel "Adoración", Álvaro de la Iglesia, for example, is considered a Cuban author because he moved to Cuba as a young adult and was active and naturalized there, but he was born in Spain.

The first group of keywords related to the text itself is about its sources: the medium of the source (<term type="text.source.medium">), its filetype ("text.source.filetype"), the institution it was retrieved from ("text.source.institution"), and the kind of edition of the source ("text.source.edition"). The medium can be either "digital" or "print", the filetype "image" or "text", the type of edition "first", "historical", or "modern", and the keyword about the

[311] See chapter 3.1.2 above for details about how the authors were assigned to the countries.

source institution can take any value from an open list of institutions.[312] This kind of metadata is important to document from what sources the corpus was constructed.[313]

Next, keywords about the publication of the novel are included: in which country was it published first (`<term type="text.publication.first.country">`), in which medium (`"text.publication.first.medium"`), in what type of publication (`"text.publication.first.type"`), and has it been published independently (`"text.publication.first.independent"`)? The term concerned with the medium can take the values "book", "journal", "magazine", or "unknown", and the type of publication can be either "independent" (e.g., in book form), "dependent" (e.g., in a journal, magazine, or as part of a book), "collection" (dependent, but together with other items of the same kind, e.g., an anthology or the œuvre of an author), or "unknown". In many cases, the information about the medium and type of publication of the novel cannot be given with high certainty because it depends on the knowledge of all the (historical) editions of the work. Here, the attribute `@cert` serves to indicate the degree of certainty about these metadata values. Information about how the novel was published historically is of interest from various perspectives: it is related to the question of the generic identity of the work[314] and also its canonicity.[315] The historical conditions of the production and reception of novels can also be investigated by analyzing how they were (first) published.

Some general keywords about the text follow: in what language it is written (`<term type="text.language">`), in what form it is composed (`"text.form"`), and to which major genres it belongs (`"text.genre.supergenre"` and `"text.genre"`). The values of these keywords are the same for all the works in the corpus: they are all written in Spanish, composed in prose, and they are all narrative texts as well as novels. So these keywords are not used to distinguish the texts inside of the corpus from each other but to give some general information about them, which can be useful when this corpus is reused in other contexts, for example, in a multilingual setup or in a study contrasting different major genres.

The next terms in the example contain the title of the novel as it appears in different editions, including series titles and subtitles. In the attribute `@n`, the year of the edition is indicated (e.g., `<term type="text.title" n="1894">`). This information is more fully documented in the digital bibliography but is repeated in a compact form in the individual corpus files because the titles of the novels' editions are analyzed when their subgenres are determined. That way, all the information necessary to reproduce the subgenre assignment is available directly in the respective

---

[312] The filetype is only categorized roughly into "image" or "text" instead of more detailed information such as "HTML", "PDF", etc. because the distinction between image- and text-based filetypes is the most relevant one affecting the way the texts had to be prepared to be included in the corpus (see the previous chapter 3.3.2 on text treatment). Filetypes such as HTML or PDF are not necessarily informative in this regard because both can contain the novel as text or images. See, for example, the short novel "La baronesa de Joux" by Gertrudis Gómez de Avellaneda, which is offered in an HTML format that consists of structural information with embedded images in the "Biblioteca Virtual Miguel de Cervantes" (Avellaneda [1871] 2008). As for the type of edition, "historical" refers to editions published in the period covered by the bibliography and the corpus (1830–1910), and "modern" refers to editions published after 1910. The list of institutions is, in principle, open and is only controlled to make sure that the institutional names are not differently spelled when they are used.

[313] This metadata was already analyzed above in chapter 3.3.1 on the selection of novels and sources for the corpus.

[314] In some definitions of the novel, an independent publication of the text is a necessary feature. See chapter 3.1.1.5 above.

[315] It is assumed that it is less likely for novels that were not published independently to enter the literary canon.

TEI file. In the example, there are three editions from 1894, 1901, and 1906, but the title and subtitle of the novel do not change from one edition to the other. The keywords following the text's titles all relate to the subgenre of the novel. The first type of term concerning the subgenre serves to record explicit subgenre labels that occur in the title of the novel (`<term type="text.genre.subgenre.title.explicit">`). The novel "Adoración" has the explicit label "novela original". The second subgenre term indicates a subgenre that is signaled implicitly in the title (`<term type="text.genre.subgenre.title.implicit">`). In this case, this is a "novela sentimental" because the main title "Adoración" means "admiration". Because the inference of implicit signals is an interpretive process, this `<term>` element carries a `@resp` attribute documenting who entered the value. Here, only two of these terms are illustrated because the assignment of subgenres is discussed more fully in the next chapter 3.3.4.

Next, there are three groups of keywords related to the content of the novel: the narrative perspective of the text, the kind of speech sign used in it, its setting, and the time period covered by the plot of the novel. The narrative perspective is given in two variants: first, as the kind of narrator (`<term type="text.narration.narrator">`), which can be "autodiegetic", "homodiegetic", or "heterodiegetic" and second, indicating the person in which the text is narrated (`<term type="text.narration.narrator.person">`), with the possible values "first person" and "third person". The narrative perspective is an important metadata item in the context of a stylistic analysis because it significantly influences the language of the text. For example, a novel that is written in the first person contains many more verbs in the first person than a novel narrated in the third person, where the first-person verbs only occur in direct speech or thoughts. Of course, the narrative perspective can change throughout the novel. The perspective encoded here is the one dominating the text because, from a statistical point of view, this affects the linguistic material of the text the most. Minor shifts are neglected. Literary-historical characterizations of the texts were consulted to determine the narrative perspective. The openings of the novels were read, and other parts of the novels were checked randomly.

After the keywords describing the narrative perspective, two terms defining the type of speech sign used in the novel follow (`<term type="text.speech.sign">` and `<term type="text.speech.sign.type">`). The first of these terms has the purpose to indicate which typographical sign is predominantly employed to mark direct speech, and the second term classifies the speech sign as "single" or "double". A speech sign of the type "single" functions as a marker for the beginning and eventually also for the end of a speech. It is a single sign indicating a change in the narrative mode. A speech sign of the type "double", in contrast, serves to enclose passages of direct speech and usually consists of two different signs, an opening and a closing one (e.g., the double angle brackets « and »). The metadata about speech signs is collected to enable a rule-based automatic detection of direct speech using this typographic information (see chapter 3.3.3.2.8 below). In the novel "Adoración", the main speech sign is a long hyphen (—), which is a speech sign of the type "single".

The setting of the novel is described in two keywords stating the continent (`<term type="text.setting.continent">`) and the country (`<term type="text.setting.country">`) in which the plot takes place. Here, too, only one principal value is given, although the setting can involve several continents or countries, for example, in travel novels. The main setting is understood to be the primary place of action and, if there are several ones and no predominant place can

be determined, the place where the action starts and the characters come from. The setting was taken up in the metadata because it is related to the question of how *American* or *national* the Argentine, Cuban, and Mexican novels were in terms of content. The same strategy as for the narrative perspective was followed to find out the setting of the novels.

The third group of content-related keywords covers the time period in which the action of the novel takes place. The first keyword of this kind serves to hold a concrete time span, if available (`<term type="text.time.period">`). A regular expression was used to locate years explicitly mentioned in the text (`"\d{4}"`) to find out the time period. The found years were checked to see if they were only mentioned or if they referred to the action and which span of years they covered. Furthermore, summaries of the novels and first chapters were consulted to find information about the time period of the plot. In the novel "Adoración", there is no explicit temporal localization, so the corresponding term takes the value "unknown". In other cases, the values are statements such as "1827", "1539–1541", or "~1700".[316] Even when dates are mentioned, the time period cannot always be determined exactly. The novel "María Luisa" (1896, MX) by Andrés Portillo, for example, begins with the following statements:

> Era joven aún este siglo XIX que hoy contemplamos anciano y moribundo, tan lleno de glorias y cargado de responsabilidades.
>
> México había derramado su oro y su sangre por espacio de once años para librarse de la dominación española y lanzábase a la vida independiente con la vaguedad del hombre que acaba de tener un sueño penoso.
>
> Se ensayaban todas las formas de gobierno, se convocaban congresos nacionales, se defendían principios y contraprincipios y había de una parte, quienes suspiraban por el régimen colonial, y de otra, quienes aplaudían las doctrinas más atrevidas de la revolución francesa. (Portillo [1896] 2020)

The action is located temporarily somewhere in the early nineteenth century. It is said that Mexico is already independent, so it must be after 1821, and that several forms of government have been tried out, so some years must have passed since the declaration of independence. This is encoded as `<term type="text.time.period" n="1830">`. The main purpose of this metadata is to find out if the novels are set in the present, in a recent or more distant past, or even in the future because the time period is an important feature related to the subgenres of the novels: contemporary, different kinds of historical, and science fiction novels. Therefore, the values encoded in the first term of this type are set in relation to the life dates of the author (`"text.time.period.author"`) and to the year of publication of the novel (`"text.time.period.publication"`) in the subsequent keyword terms. These terms can take the following values: "contemporary", "recent past", "past", and "future". When the time period is not marked in the text, it is assumed that the time frame of the action can be considered contemporary, as in the current example "Adoración". Table 16 summarizes how the values for these keyword types are determined.

---

[316] See the corresponding novels "El guajiro" (1842, CU) by Cirilo Villaverde, "La cruz y la espada" (1866, MX), and "El filibustero" (1864, MX) by Eligio Ancona at https://github.com/cligs/conha19/blob/master/tei/nh0001.xml, https://github.com/cligs/conha19/blob/master/tei/nh0026.xml, and https://github.com/cligs/conha19/blob/master/tei/nh0180.xml respectively. Accessed March 24, 2020.

Type of keyword	Value	Explanation
text.time.period.author	contemporary	If the narrated time is contemporary to the author (during the author's lifetime) or if it is not marked at all.
text.time.period.author	recent past	If the narrated time is within 30 years before the author's birth date.
text.time.period.author	past	If the narrated time is more than 30 years before the author's birth date.
text.time.period.author	future	If the narrated time is more than 100 years after the author's birth date.
text.time.period.publication	contemporary	If the narrated time is contemporary to the publication date (within 30 years before and after) or if it is not marked at all.
text.time.period.publication	recent past	If the narrated time is between 30 and 60 years before the publication date.
text.time.period.publication	past	If the narrated time is more than 60 years before the publication date.
text.time.period.publication	future	If the narrated time is more than 30 years after the publication date.

Table 16. Values for the time period covered by a novel.

Regarding the author, novels that take place during her or his lifetime are classified as contemporary. They are categorized as belonging to the recent past if the narrated time is within 30 years before the author's birth date and as past if it is more than 30 years away from it. A novel set in the future is one where the narrated time is located more than 100 years after the author's birth date. The temporal limits were chosen based on the assumption that 30 years approximately mark a generation and that an author who placed the action of the novel more than 100 years away from his birth date did not expect to live in that future anymore. The time spans were chosen slightly differenty to decide upon the temporality of the novel in relation to its publication date, but they were also based on generational changes. A novel is marked as contemporary if the narrated time is within 30 years before or after its publication date, as recent past if the narrated time lies within 30 to 60 years before its publication, as past if it is more than 60 years ago, and as future if it is located more than 30 years after the appearance of the novel. Obviously, the time spans are narrower for the publication because it is a point in time and wider for the author because his or her life dates are a period of time.

The last type of keyword included in the text classification section of the TEI header serves to classify the novels in terms of prestige (`<term type="text.prestige">`) as either "high" or "low". This metadata value is useful to assess the composition of the corpus regarding the canonicity of the texts. High or low literary prestige can be described and measured in many different ways, for example, considering literary prizes that the works have won, the number of editions and copies of the texts that were produced, the number and kind of critical and scholarly engagements with them, assessing the prestige of the authors or subgenres of the novels, etc. For this corpus, it was decided to use a measure that is simple to capture and that reflects how the

texts have been valued by scholars and the public in the second half of the twentieth up to the twenty-first century. To this end, the union catalog WorldCat was used to check which novels were republished between 1860 and 2020 as new editions or reprints of historical editions. All the novels that were republished at least once during this period are classified as high prestige, the others as low.[317] This measure results in many novels being classified as high prestige without differentiating further between those that were only reprinted or reedited once and others that received much more attention. On the other hand, it clearly points out which works have been largely forgotten. As the measure applies to works and not authors, there are cases where some novels of an author are classified as "high" and others as "low". In the corpus, 174 novels have high, and 82 have low prestige.[318]

Many more kinds of metadata could be collected for the novels, especially regarding their content. For example, information about the characters could be included. Some of this metadata can be created automatically or semi-automatically, but many kinds need manual checks of selective or full reading. The selection of metadata encoded for this corpus was made to gain insight into some principal parameters and contents of the novels, but as this dissertation focuses on the analysis of subgenres of the novel, more attention was put on metadata related to this aspect. Nevertheless, besides their overview function, the metadata about the settings and time periods covered by the novels can also be used as control values for characteristics of the texts determined automatically with text mining and NLP methods.

As stated above, the keyword system is controlled by an external taxonomy stored in the file "keywords.xml".[319] It serves to describe and order the possible types of keywords and their values and is itself also formulated in TEI. Example 14 shows an excerpt from the taxonomy.

```
<taxonomy xml:id="keywords">
  <category xml:id="author">[...]</category>
  <category xml:id="text">[...]
    <category xml:id="text.narration">
      <catDesc>text.narration</catDesc>
      <category xml:id="text.narration.narrator">
```

[317] The ways in which literary prestige can be measured have been reflected in the context of a joint research project on "Computational Approaches to Complexity in Literary Texts" between the Universities of Osaka (led by Tomoji Tabata) and Würzburg (led by Fotis Jannidis) and funded by the German Academic Exchange Service (DAAD) and the Japan Society for the Promotion of Science (JSPS) from 2017 to 2019. The way to measure the prestige of the Spanish-American nineteenth-century novels here results from what was reflected in that project, in which the author of this dissertation participated. In detail, the following rules were set up for the search in the WorldCat: (1) all kinds of republications were counted, whether scholarly or general, printed or digital; (2) the only exception being digital editions of the IAI in Berlin dated to 2017 because these are scans of the novels that were commissioned by myself, and they do therefore not reflect the general prestige that the novels have gained; (3) the complete works of an author were neglected, meaning that a novel that was only republished as part of complete works is still considered as low prestige. The assumption behind this decision is that complete works show an interest in an author and in his or her work as a whole but do not necessarily imply that all the individual works are valued highly; (4) for sequels, it was considered enough to find a reprint of (the title of) one (often the first) part because works published in several parts originally are often published together in later editions; (5) for works that were originally published dependently it was also looked up if they were republished that way (for example, as part of a collection of selected works). The search in the WorldCat was performed on June 4, 2020.

[318] See also chapter 4.1.3.2, where an overview of the novels in the corpus is given.

[319] The taxonomy file is available at https://github.com/cligs/conha19/blob/master/schema/keywords.xml. Accessed March 24, 2020.

```
        <catDesc>text.narration.narrator</catDesc>
        <category xml:id="text.narration.narrator_1">
          <catDesc>autodiegetic</catDesc>
        </category>
        <category xml:id="text.narration.narrator_2">
          <catDesc>homodiegetic</catDesc>
        </category>
        <category xml:id="text.narration.narrator_3">
          <catDesc>heterodiegetic</catDesc>
        </category>
        <category xml:id="text.narration.narrator.person">
          <catDesc>text.narration.narrator.person</catDesc>
          <category xml:id="text.narration.narrator.person_1">
            <catDesc>first person</catDesc>
          </category>
          <category xml:id="text.narration.narrator.person_2">
            <catDesc>third person</catDesc>
          </category>
        </category>
      </category>
      [...]
    </category>
  </category>
</taxonomy>
```

Example 14. A section of the TEI taxonomy of keywords.

In the example, the keywords about the narrative perspective of the novel are listed. Each keyword level and type is encoded in a `<category>` element whose attribute `@xml:id` serves as a unique identifier for the category in question. The system of categories is organized hierarchically, which is expressed by the XML element structure. In the identifiers, which are used in the corpus files to reference the keyword types, this hierarchy is mapped to a string separated by dots. Categories on the lowest level correspond to the values that the keyword type can take. On the different levels, `<catDesc>` elements are used to either indicate the name of the category, a description of it, or its possible values. In the example, the possible values for both `"text.narration.narrator"` and `"text.narration.narrator.person"` consist of closed lists, meaning that these keywords can only take one of the values listed in the taxonomy. In other cases, for example, the authors' countries of birth, lists of values mean that these are the countries that appear in the corpus, but the list is, in principle, open for more entries. Open lists have the function to ensure that the values of the keywords are spelled identically each time that they are used. At the same time, they document the range of values occurring in the corpus.

The external taxonomy in itself does not guarantee that the keyword types and values are used in the intended way in the TEI files of the novels. A Schematron file was created and is referenced from each corpus file to make sure that the usage of the keywords is consistent throughout the corpus.[320] This file is not only used to check the keywords but also the other metadata contained in the TEI header, as example 15 shows.

```
<sch:pattern>
  <sch:rule context="tei:titleStmt">
```

[320] For more information about Schematron, see chapter 3.2.2 above, where the data model of the digital bibliography is explained. The Schematron file checking the corpus metadata can be viewed at https://github.com/cligs/conha19/blob/master/schema/keywords.sch. Accessed March 24, 2020.

```
    <sch:assert test="tei:title[@type = 'short'][. != '']">
      <sch:value-of select="$cligs-idno"/>: TEI header error: Short title is missing.</
          sch:assert>
  </sch:rule>
</sch:pattern>
<sch:let name="keywords-file" value="document('keywords.xml')"/>
<sch:pattern>
  <sch:let name="cat-narration" value="$keywords-file//tei:category[@xml:id='text.
      narration']"/>
  <sch:rule context="tei:term[@type'=text.narration.'narrator]">
    <sch:assert test="normalize-space(.) = $cat-narration/tei:category[@xml:id = 'text.
        narration.narrator']/tei:category/tei:catDesc">
        <sch:value-of select="$cligs-idno"/>: Metadata error: text.narration.narrator</
            sch:assert>
  </sch:rule>
</sch:pattern>
```

Example 15. Schematron file to control the metadata.

The first rule applies to the title statement. It contains an assertion testing whether there is a `<title>` element of the type "short". If this is not the case, an error message is displayed. The context of the second rule is a keyword term of the type "text.narration.narrator". The external keywords file and the definition of the keyword type to check ("text.narration") are stored in Schematron variables. Then it is tested whether the term of the type "text.narration.narrator" contains one of the possible values listed in the external taxonomy. If not, a metadata error is raised. The Schematron file is a good way to complement the general schema controlling the TEI structure of the corpus because it allows to check the content of the attributes and elements depending on the XML structure and on the external taxonomy.[321] That way, it allows the definition of more detailed and rigorous rules, which is useful to ensure that the metadata is consistent throughout the corpus.

3.3.3.1.7 Revision Description

After the profile description, including the keywords list, the last part of the TEI header is the revision description, a section holding information about the revision history of the TEI file. It is useful to document changes made between different versions of the files, especially when many different files are updated manually and when several people work together. For the current project, the revision description was not essential because the corpus was prepared by one person and because it does not have a long public history yet. Therefore, up to now, in most cases, the revision descriptions of the corpus files only contain one entry indicating when the TEI file was first created (see example 16). However, the encoding of changes made to the files might become more important in the future when this corpus is possibly reused by other researchers.

```
<revisionDesc>
 <change when="2015-03-31" who="#uhk">Initial TEI version.</change>
</revisionDesc>
```

Example 16. Revision description of the novel "Adoración".

[321] The general corpus schema is commented on further below after the discussion of the TEI encoding of the textual body. There it is also explained how the schema files are processed to check the whole corpus and how errors are reported.

To sum up, the encoding of the corpus metadata in the TEI header is kept simple for general administrative and bibliographic information and is more elaborated in the keywords part, where different aspects of the novels that are considered relevant for their stylistic analysis are described. Some of the metadata that is encoded as part of the taxonomic keyword system could as well be placed elsewhere in the TEI file, but it was decided to keep this kind of metadata in one place and in an analogous structure to facilitate the analysis of the texts.

3.3.3.2 TEI Body

Besides the TEI header, the second main part of each corpus file is the transcription and encoding of the novel in the `<text>` element. It is further subdivided into three parts: `<front>`, `<body>`, and `<back>`. While the body is present in all the TEI files of the corpus, the other two parts are optional. The front part may contain "any prefatory matter (headers, abstracts, title page, prefaces, dedications, etc.) found at the start of the document, before the main body" (Text Encoding Initiative Consortium 2023g). The back part can contain appendices of any kind (Text Encoding Initiative Consortium 2023e). In the corpus, the front part was used to encode title pages, dedications, and prefaces of available historical editions of the novels because they often provide information about the subgenres of the texts. Such front matters were included in 231 files of the corpus. For the other 25 novels, no historical editions could be accessed, so no front matter is available.[322] Front matters of modern editions were not transcribed. In example 17, an excerpt of the front matter for the novel "Adoración" is shown.

```
<front>
  <div source="#PS" n="1906">
    <div type="titlepage">
      <ab>Biblioteca de Autores Americanos</ab>
      <ab>Alvaro de la Iglesia</ab>
      <ab>Adoración</ab>
      <ab>Novela original</ab>
      <ab>Tercera edición</ab>
      <ab>Barcelona</ab>
      <ab>F. Granada C.ª, Editores</ab>
      <ab>Calle de Escudillers, 20</ab>
      <ab>Buenos-Aires</ab>
      <ab>Serafín Ponzinibbio, Editor</ab>
      <ab>B. Mitre, 1.100</ab>
      <ab>1906</ab>
    </div>
    <div type="dedication">
      <p><seg rend="italic">A Antonio Herrera</seg></p>
      <p><seg rend="italic">en El Mundo</seg></p>
      <p><seg rend="italic">testimonio sincerísimo de afecto.</seg></p>
      <ab>
        <seg rend="italic">El Autor.</seg>
      </ab>
    </div>
    <div type="preface">
      <head>Prólogo</head>
      <p>Si los hombres no fueran tan dados a vaticinar y tan reacios a escarmentar, no
          obstante la facilidad con que se viene abajo la fábrica de sus pronósticos, no
```

[322] Editions are considered "historical" here if they were published within the chronological scope of the bibliography and corpus (1830–1910).

```
          oyéramos con tanta frecuencia los horóscopos que anuncian casi para día fijo
         la muerte de la poesía. [...]</p>
      <ab>Enrique José Varona.</ab>
      <ab>Habana (Cuba).</ab>
    </div>
  </div>
</front>
```

Example 17. Front matter of the novel “Adoración”.

In the example, the front matter of one historical edition of the novel from 1906 is transcribed. It includes a title page, a short dedication, and a longer preface, of which only a part is shown here. In other cases, there may be several front matters. Each front matter is enclosed by a division element, indicating its source edition in the attribute `@source` (`source="#PS"`). This attribute contains a reference to the edition’s bibliographic description in the source description in the TEI header. The year of the corresponding edition is encoded in the attribute `@n` (`n="1906"`) on the `<div>` element. Inside the main division for each front matter, its different parts are encoded in further subdivisions (e.g., `<div type="titlepage">`, `<div type="dedication">`, `<div type="preface">`). Although the TEI offers specialized elements for the encoding of front matter, e.g., `<titlePage>`, `<byline>`, `<docImprint>`, etc. (Text Encoding Initiative Consortium 2023k), only the general elements `<div>`, `<ab>`, `<head>`, `<p>`, and `<seg>` are used here to keep the overall TEI model for the corpus simple and because there is no special interest in the semantics of the front matter structure here. Instead, the front matters are transcribed with the primary goal of interpreting their contents with regard to the subgenres of the novels.[323]

A back matter is included in 140 of the 256 TEI files. In general, back matters are less relevant for the subgenre assignment. In most cases, they only contain a phrase marking the end of the novel (“Fin”) or a dateline documenting where and when the novel was written (e.g., “Buenos Aires, Agosto 27 de 1858”). Only rarely notes or comments by the authors are appended, such as, for example the following remarks made by Ignacio Manuel Altamirano about the length of his novel “Clemencia” (1869, MX), as shown in example 18.

```
<back>
  <div>
    <head>Nota</head>
      <p>El menor de los defectos de esta pobre novelita es que para cuento parece
          demasiado larga. Pero no hay que tomar formalmente la ficción de que el doctor
           relate esto en una noche. Es un artificio literario, como otro cualquiera,
          pues necesitaba yo que el doctor narrara, como testigo de los hechos, y no
          creí que debía tener en cuenta el tamaño de la narración. Además, a pesar de
          mi pequeñez me amparan, para hacer perdonable lo largo del cuento, los
          ejemplos de Víctor Hugo en Bug-Jargal, de Dickens en varios de sus Cuentos de
          Navidad, de Erkmann Chatrian en sus Cuentos populares, de Enrique Zschokke en
          sus Cuentos suizos, y de Hoffman en muchos de los suyos. En lo que si no tengo
           amparo es en lo demás, y no me queda más recurso que apelar a la bondad de
          los lectores.</p>
    <ab type="signed">EL AUTOR</ab>
  </div>
</back>
```

Example 18. Back matter of the novel “Clemencia”.

[323] See the next chapter 3.3.4 on the assignment of subgenre labels to the novels in the corpus for details.

The encoding of the main body of the novel's text is kept simple, as well. Above all, the markup is used to represent how a novel is structured into parts and chapters to be able to use this structural information in the analysis of the texts. Example 19 shows how the beginning of the novel "Adoración" is encoded.

```
<body>
  <div type="chapter">
    <head>I</head>
    <head>El escenario</head>
    <p>Allí donde se rompe sobre el acantilado granítico el inmenso empuje de dos mares
        y el movimiento formidable del Océano levanta al aire blancas trombas de
        rugiente espuma manteniendo en un constante clamoreo las aguas de la costa, la
        labor eterna de las olas ha abierto una ensenada en el abrupto litoral en que va
         a morir la resaca como en un remanso, cual si cansada de su fatigoso golpeo se
        tendiera perezosa en las brillantes arenas de la playa.</p> [...]
  </div>
</body>
```

Example 19. Beginning of the novel "Adoración".

In general, divisions are marked with the element `<div>`, using the attribute `@type` to characterize the kind of division further into `"part"`, `"subpart"`, `"chapter"`, or `"subchapter"`. Headings and paragraphs are also encoded. In general, no difference is made between the main and subheadings. Only longer descriptions of the content of a following chapter are marked as `<head type="argument">`. Regarding the structure inside of the main textual divisions, it was decided to generally encode blocks separated by line breaks or blank lines with the element `<p>`, following a typographic definition of a paragraph. The only exceptions made are for verse lines, which are encoded with `<l>`, and dramatic speech, encoded with `<sp>` because these are considered important distinctions from the point of view of genre analysis. It follows from this that the content of a `<p>` element does not always correspond to the structural linguistic definition of a paragraph as a sequence of semantically related sentences or as a thematic building block of a written text.[324] A ubiquitous phenomenon in the novels, for example, is blocks of direct speech. These are also marked with `<p>` and additionally with `<said>`, as explained further below.

The TEI standard includes many different elements for the encoding of text blocks, for example, the neutral element `<ab>`[325] or special elements for structures like openers and closers in letters, for list or table entries, etc. In principle, such alternative elements would be a better choice to encode blocks in the novels that are not paragraphs in the linguistic-semantic sense. However, a detailed analysis of the text bodies would be required to identify such structures. In addition, more specialized markup would require advanced scripts for querying the XML structure. Furthermore, it can be estimated that non-paragraph blocks that are not verse lines, dramatic speech, or direct speech are few in number in the novels. The `<p>` element was therefore preferred here as a general solution for the markup of typographic blocks in the text body.

Besides this general structure of divisions, headings, and paragraphs, some more phenomena were encoded, as summarized in table 17.

---

[324] For a characterization of the paragraph from a linguistic point of view, see Rinas (2015).

[325] The anonymous block element is described as "any arbitrary component-level unit of text, acting as an anonymous container for phrase or inter level elements analogous to, but without the semantic baggage of, a paragraph" (Text Encoding Initiative Consortium 2023d).

Type of phenomenon	TEI element(s) used
Typographically marked subdivisions of the text (e. g., with a line or asterisks)	`<milestone>` with `@unit` and `@rend`
Typographically highlighted words or phrases	`<seg>` with `@rend`
Gaps	`<gap>`
Verse lines	`<lg>`, `<l>`
Dramatic text	`<castItem>`, `<castList>`, `<sp>`, `<speaker>`, `<stage>`, `<said>`
Representations of written text	`<writing>` with `@type`
Quotations	`<quote>`
Direct speech or thought	`<said>`
Text contained in quotation marks that is not a representation of written text, not a quotation, and not direct speech or thought	`<q>`
Embedded texts interrupting the surrounding text	`<floatingText>`

Table 17. Encoding of textual phenomena in the main body of the novels.

The encoding of the first two phenomena (typographically marked subdivisions of the text and typographically highlighted words or phrases) aims to preserve minor structural information that was already contained in a structured way in the editions used as sources for the corpus. These typographic details may be interesting when individual sections of texts are analyzed, but they can hardly be used for comparative analyses of all the texts in the corpus. They depend highly on how a specific source edition of a novel was typeset and also on how much of possibly existing typographic information in the sources was kept when editions were digitized. Gaps were encoded to get a quantitative overview of how much text is missing in the novels. The other phenomena that were encoded in the body focus on how the narrated text is presented in terms of genre (prose versus poetry versus drama), medium (written versus spoken), voice, and perspective (quotations, narration, and the representation of speech and thought). In the remainder of this subchapter, examples of the different phenomena that were encoded in the main body of the novels are given, with a special focus on the detection of direct speech and thought.

3.3.3.2.1 Typographically Marked Subdivisions of the Text

Regarding the structure of the novels, sometimes chapters are divided further into sections. Such subdivisions are marked with different typographic means in the editions, for example, using a line between two paragraphs, one or more asterisks or other symbols, or just more blank lines than between paragraphs of the same section. Wherever this information was contained in the digital editions used as sources for the corpus or where it could be marked in the process of text treatment, it was kept and encoded with the element `<milestone>`, as example 20 from the novel "A fuego lento" (1903, CU) by Emilio Bobadilla illustrates.

```
<p>La vida, durante la noche, se concentraba en la plaza de la Catedral, donde estaba,
    de un lado, el <seg rend="italic">Círculo del Comercio,</seg> y del otro, <seg rend=
    "italic">El Café Americano.</seg> Las familias tertuliaban en las aceras o en medio
    del arroyo hasta las once. En el silencio sofocante de la noche, la salmodia de las
```

```
    ranas alternaba con el rodar de las bolas cascadas sobre el paño de los billares y
    el ruido de las fichas sobre el mármol de las mesas. La calma era profunda y
    bochornosa. El cielo, a pedazos de tinta, anunciaba el aguacero de la madrugada o
    tal vez el de la media noche.</p>
<milestone unit="section" rend="asterisks"/>
<p>La casa de don Olimpio andaba manga por hombro. Misia Tecla, su mujer, gritaba a los
    sirvientes, que iban y venían atolondrados como hormiguero que ha perdido el rumbo.
    Una <seg rend="italic">marimonda</seg>, que estaba en el patio, atada por la cintura
     con una cuerda, chillaba y saltaba que era un gusto enseñando los dientes y
    moviendo el cuero cabelludo.</p>
```

Example 20. Encoding of a subdivision inside a chapter in the novel “A fuego lento”.

In the edition, several asterisks mark the transition from one section to the other. The paragraph before the section boundary contains the description of a scene on a public square. In the following paragraph, the setting switches to the house of Don Olimpio, so the section boundary coincides with a content-related change inside of a chapter of the novel. However, because it is hard to verify if section boundaries inside of chapters, if present, are represented reliably throughout the different editions of a novel, the corresponding milestones will not be used systematically to analyze the structure of the novels. They were primarily encoded to not lose the existing structural information and because they can still be useful when individual passages of the novels are inspected.

3.3.3.2.2 Typographically Highlighted Words or Phrases

In the editions of some novels, individual words or phrases are highlighted using italics. A number of reasons can be identified for such highlighting, for example:

- the inclusion of foreign languages into the novel,
- the representation of oral speech that does not fulfill grammatical or orthographic rules,
- the use of special vocabulary,
- the general emphasis of or distancing from a term used by the narrator or a character, among other reasons.

Such emphases are stylistically relevant, but unfortunately, their usage varies substantially from edition to edition, and they are not reliably included in the different digital source editions. Furthermore, in some novels, the same aspects are highlighted with quotation marks, which in turn are also one possible means of marking direct speech, making it very difficult to single out the different types of stylistic emphases automatically. Highlights in italics are, therefore, only kept with the aim of preserving existing typographic information in this corpus. The TEI element used to mark them is `<seg>` with the attribute `@rend` indicating how the emphasis is rendered typographically, as shown in example 21.

```
<p>—¿Cómo se llama Vd.?</p>
<p>—<seg rend="italic">Me ñama Ginoveve Santa Crú. Mi marío e Tribusio Polanca. Elle
    tien uno sijo ñamao Malanga que ha sacao mala cabesa. ¡Ha matao ma branco!... Tondá
    lo coge como ratón con quesa le dominga depué de Niño perdío, cuando diba nel
    entierre de ña Chepa Alarcó.</seg></p>
<p>—¿Chepilla Alarcón? repitió preguntando María de Regla.</p>
<p>—<seg rend="italic">Sí, sí, agrego Genoveva. Le meme. Asín se ñamaba. Ha perdío un
    güen caserite.</seg></p>
<p>—¿Tenía una nieta?</p>
```

```
<p>—<seg rend="italic">Sí, tube un. ¡Ma linde! ¡Ah! ¡qué bunite! No la ha vito ma bunite
     en la vía.</seg></p>
```

Example 21. Encoding of non-standard oral speech highlighted in italics in the novel "Cecilia Valdés o la Loma del Ángel".

3.3.3.2.3 Gaps

In the case of incomplete, partly damaged, or illegible source editions, there may be gaps in the text body. Wherever they became apparent in the process of text treatment, these gaps were marked up with the element <gap>, with the goal to be able to quantify the overall amount of missing text in the corpus. The encoding of a gap is illustrated in example 22.

```
<p><said>—Sí, pero yo quiero romper esas ligaduras; sí, quiero romperlas, porque un día
    miro cercano en que no pudiendo ya mi corazón hacerse más violencia, romperá las
    cadenas de su deber, atro-</said></p>
<gap unit="page" extent="2" reason="missing"/>
<p>luchar hasta morir al frente de vuestros ejércitos, y si llegáis a vencer un día,
    acordaos de que soñó con vuestra independencia el infortunado nieto de Cárlos I.</p>
```

Example 22. Encoding of a gap in the novel "El tálamo y la horca".

Three attributes are added to the <gap> element, characterizing it further. In the case at hand, the gap consists of two (`extent="2"`) missing (`reason="missing"`) pages (`unit="page"`). Possible values for `@unit` are `"page"`, `"line"`, `"word"`, and `"char"`. The number of missing units is given in `@extent`. Sometimes it cannot be known exactly how many items, for example, words or characters, are missing. In such cases, the number is estimated. For the purpose of this text collection, the attribute `@reason` may take the values `"missing"` or `"illegible"`.[326]

3.3.3.2.4 Verse Lines

Verse lines were encoded using the elements <lg> for groups of verse lines and <l> for single verse lines. The main interest in encoding verse lines in the novels lies in the ability to calculate the proportion of poetry contained in the prose texts. Verse lines were detected in the process of text conversion from the source editions to the TEI files and also searched for with a simple XPath expression in the resulting XML files: `//p[count(tokenize(.," "))< 10][ not(contains—(.,""))]`. This expression finds blocks that are encoded as paragraphs, that are shorter than ten tokens separated by whitespace, and that do not contain a long hyphen, which is a conventional speech sign. The expression assumes that verse lines are usually short. It also returns short prose paragraphs but helps to scan through possible candidates for verse lines quickly. Poems are typically included in the novels as part of quotations, for example, at the beginning of chapters, as part of the representation of written materials, for instance, love letters, or as songs sung by characters, as in example 23 below.

```
<p>Boca-lobo guardó su moneda.</p>
<p>El jefe se levantó de su asiento, y los demás le siguieron.</p>
<p>Al retirarse, el jefe, para demostrar su complacencia,
   se puso a cantar en voz baja:</p>
```

[326] See also chapter 3.3.2 on text treatment above on this topic.

```
<said>
  <lg>
    <l>"El que pasa una noche</l>
    <l>en rumbantela,</l>
    <l>si está triste se alegra</l>
    <l>y se consuela;</l>
    <l>que está probado</l>
    <l>que, el que de rumba anda</l>
    <l>nunca está <seg rend="italic">triste</seg>."</l>
  </lg>
</said>
```

Example 23. Encoding of verse lines in the novel “Los Hermanos del Silencio”.

3.3.3.2.5 Dramatic Text

Dramatic text was encoded for the same reason as verse lines – to get an overview of how much of this structure that is characteristic of another major genre, drama, is included in the novels. In the CLiGS TEI schema, all the typical elements for encoding dramatic text are available because the schema also covers collections of drama. In this corpus of novels, mainly the elements `<sp>` for speech in a performance text, `<speaker>` for labels giving the name of a speaker, `<p>` for the structure of the speech, and `<stage>` for stage directions are used, as illustrated in example 24. In the excerpt taken from the novel “Pot-pourri (Silbidos de un vago)” (1882, AR) by Eugenio Cambaceres, a whole chapter is presented as a dramatic scene. In this case, the narrator uses this generic shift as a stylistic means to caricature the behavior and personality of other characters. As can be seen, even though elements of drama are used, in this case, they are mixed with prose paragraphs in which the narrator comments on the dialogue.

```
<div type="chapter">
  <head>X</head>
  <sp>
    <speaker><seg rend="italic">Juan</seg>.</speaker>
    <p>— ¡Una y mil veces malditos los negocios!</p>
    <p>¡Quién pudiera nutrirse de ambrosía como los habitantes del Olimpo!</p>
    <p>Ved aquí a un hombre joven, sano, alegre, dispuesto, que no ambicionaría otra
        cosa, sino que lo dejaran vivir eternamente mano a mano con su mujercita a quien
         adora, siendo a su vez adorado por ella... <stage><seg rend="italic">(le da un
        beso)</seg></stage>.</p>
  </sp>
  <sp>
    <speaker><seg rend="italic">María</seg>.</speaker>
    <p>— ¡Juan, por Dios, qué dirá este caballero! <stage><seg rend="italic">(poniéndose
         colorada hasta la punta de la nariz con incomparable modestia)</seg>.</stage></
        p>
  </sp>
  [...]
</div>
```

Example 24. Encoding of dramatic speech in the novel “Pot-pourri”.

To find passages of dramatic text contained in the novels, again, they were checked during the process of text conversion. Furthermore, the XPath expression `//p[tokenize(.," ")[1][ends-with(.,":")or ends-with(.,".")]]` was used on the TEI files to detect paragraphs beginning with the pattern `NAME:` or `NAME.`

3.3.3.2.6 Representations of Written Text

A phenomenon that occurs in many of the novels is that some kind of written text that forms part of their fictional world is presented by the narrator or by characters. This can be, for example, a diary entry, a letter, a newspaper article, a short note, a historical document, an inscription, for example, on a tombstone, or one of many other types of writings. The inclusion of written texts into the novels ranges from pure mentions, for instance, that somebody received a letter, to selective citations of their content and full representations of the documents. In some cases, the written texts are shown by the narrator, in others, they are read by characters. Representations of written text are often easy to detect in the novels because they are usually typographically differentiated from surrounding text in the source editions and are often introduced with angular or curved quotation marks («…» or "…"). The encoding of inserted written texts is of interest for stylistic analyses of novels and their subgenres for two main reasons. First, it appears that certain types of writings are typically included in novels of a certain subgenre. Letters, for example, are often found in romantic and sentimental novels, and source documents are often cited in historical novels. Being able to analyze the amount of different types of written text represented in the novels allows us to examine such hypotheses. Second, when written texts are represented directly, they often entail a change of perspective in the novels, for example, from a third-person to a first-person narrator or vice-versa, which also affects the style of the novels. The element `<writing>` was used to encode representations of written text, as shown in example 25, which contains a newspaper advertisement included in the novel "La virgen del Niágara" (1871, MX) by José Rivera y Río.

```
<p>Felipe mostró a su amigo un aviso que el día anterior había hecho insertar en el <seg
     rend="italic">Herald</seg> y que decía poco más o menos lo siguiente:</p>
<p>
  <writing type="newspaper">“Los dos caballeros que entraron en el ómnibus de la
      estación de Fulton, se considerarán muy agradecidos si las dos hermosas señoritas
      vestidas de luto a cuyo frente se sentaron y con quienes se sonrieron, les
      conceden el honor de una entrevista. Dirigirse al despacho del <seg rend="italic">
      Herald,</seg> a F. y <seg rend="italic">M.</seg>”</writing>
</p>
<p>—Esta aventura—,dijo Felipe—,nos va a indemnizar del tedio de este domingo.</p>
```

Example 25. Encoding of a newspaper ad in the novel "La virgen del Niágara".

The element can be used inside paragraphs to mark short stretches of written text, but it can also contain entire embedded documents. Table 18 lists the types of written texts that were differentiated in this corpus, as indicated in the attribute `@type` on the `<writing>` element.

Although the overall range of types of written texts represented in the novels is broad, it was decided to focus on a few recurring types and to define these types broadly. From a systematic point of view, the different kinds of writing may overlap. A letter, for example, can be published in a newspaper, or a poem can be part of a diary entry. The most obvious and prominent type was chosen for each writing, also depending on how it is announced in the novel. Some of the types of writings are usually connected to changes in the narrative perspective, for example, letters, diary entries, and speeches. The others primarily entail a style and type of language use that differs from the surrounding narrated or spoken text.

type	Description
letter	letters and any other kind of notice directed to someone
newspaper	newspaper articles of any kind
diary	diary entries and other kinds of written monologues (e. g., memoirs)
document	other kinds of written documents (e. g., notes, reports, historical sources, inscriptions)
book	parts of printed books
poem	written poems
speech	written speeches directed to someone
unknown	if it is just known that something is written but the kind of writing cannot be specified

Table 18. Encoding of types of written texts represented in the novels.

3.3.3.2.7 Quotations

According to the TEI Guidelines, a quotation is "a phrase or passage attributed by the narrator or author to some agency external to the text" (Text Encoding Initiative Consortium 2023i). The element <quote> is used in that sense to mark passages that are clearly attributed to other authors, like the two quotations of Balzac and Milanés in example 26 below.

```
<div type="part">
  <head>Libro Primero.</head>
  <quote>
    <p xml:lang="fr">La paix profonde et sereine imprimée par les sculpteurs aux visages
         des figures vierges destinées a representer la justice, l'innocence, toutes les
         divinités qui ne savent rien des agitations terrestres; ce calme est le plus
        grand charme d'une fille, il est le signe de sa pureté; rien encore ne l'a emuée
        ; aucune passion brisée, aucun intérét trahi n'a nuancé la plácide expression de
         son visage; est il joné, la jeune fille n'est plus.</p>
  </quote>
  <p>De Balzac.</p>
  <div type="chapter">
    <head>I.</head>
    <quote>
      <lg>
        <l>Necio, y digno de mil quejas</l>
        <l>el que ronca sin decoro</l>
        <l>cuando el sol con rayo de oro</l>
        <l>da en las domésticas tejas.</l>
      </lg>
    </quote>
    <p>J. J. Milanes.</p>
    <p>Acababa de amanecer, y el tibio sol de invierno a principios de febrero del año
        pasado, derramaba mansas olas de luz sobre los techos y campanarios de la ciudad
        , que comenzaba e despertar de un delicioso sueño. [...]</p>
  </div>
</div>
```

Example 26. Encoding of quotations in the novel "La joven de la flecha de oro".

In the novels, such quotations are often found at the beginning of parts and chapters, as in the example, but do also occur inside of chapters, where they are usually highlighted with quotation marks. Representations of written texts that are part of the fiction are understood to be internal

to the text and are not treated as quotations in this corpus. Direct speech reported on whatever level is also not interpreted as quotation.

3.3.3.2.8 Direct Speech and Thought

Regarding structural elements of the texts below the chapter level, it was also decided to encode direct speech and thought expressed by characters of the novels. The difference between the narrator text and the representation of character speech and thought is a fundamental aspect of the various possible narrative strategies used in novels to present the plot, characters, and setting, and it can be considered a stylistic choice (Leech and Short 2007, 255–281). Therefore, it is also of interest for a stylistic analysis of the subgenres of the novels. For example, it can be asked to what extent the amount of direct speech in a novel differs depending on the subgenre to which the novel belongs, or in what way narrated text and direct speech differ stylistically from subgenre to subgenre. Because also indirect forms are possible, direct speech or thought is only one of the variants of character speech and thought representation, but it is the one that is easiest to detect because it is often introduced by speech signs. To simplify, by differentiating the direct forms from the surrounding text, indirect variants are considered as part of the narrated text here. The encoding of direct speech and thought in the TEI files prepares its use as a feature for textual analysis. Hence, the topic is covered in this section with regards to text encoding, but the generation of features for genre analysis that are based on the distinction between direct speech and narrated text is a future task that can be carried out starting from the encoding of the texts in this corpus.

To manually markup all direct speech in novels would be a very time-consuming task. Therefore, an automatic approach relying on the usage of typographic speech signs and using regular expressions was pursued here for a part of the corpus. However, given that the typographic signs were not reliable enough, expensive manual checks were indispensable. The problem of unreliable signs for the detection of direct speech is, on the one hand, a general one. It is caused by the overall tradition of how character speech is signaled typographically in Spanish language novels. On the other hand, the issue is complicated if a corpus considerably relies on historical editions, such as the one described here, because speech signs are handled less consistently in historical than in modern editions. Therefore, it was decided to only encode the direct speech in a subset of the corpus. Little more than one-third of the novels (92 texts) were prepared with the mentioned semi-automatic approach. These texts were selected randomly from the corpus, with no special focus on certain kinds of editions, authors, genres, or narrative perspectives.[327] The direct speech encoding in this part of the corpus constitutes a gold standard that can be used as a training set to build machine learning models for detecting direct speech in novels.[328]

---

[327] In the following metadata file, it is listed in which novels direct speech and thought were encoded: https://github.com/cligs/data-nh/blob/master/corpus/metadata_direct-speech.csv. The metadata file was produced with the script https://github.com/cligs/scripts-nh/blob/master/corpus/metadata_encoding/direct-speech-metadata.xsl. Accessed April 20, 2020.

[328] For such machine learning approaches, see, for instance, Brunner (2013, 2015), Byszuk et al. (2020), Jannidis et al. (2018), and Schöch, Schlör, et al. (2016). An alternative to a machine learning approach would be to apply the simple regular expression approach to the other two-thirds of the novels without manual correction and to accept the resulting error rate.

In what follows, the encoding of direct speech in the TEI files is outlined. First, some general problems of defining what direct speech is are discussed, which arise independently of the question of how to detect it, but have consequences for it. Then the difficulties of using typographic speech signs as indicators in nineteenth-century Spanish-American novels are explained. Finally, the subset of the corpus with encoded direct speech is used to estimate the loss of information that would result from only relying on typographical signs. This is done by comparing the checked annotations with results that would have been obtained by applying the pure regular expression approach to the same files. The score can be compared to results usually obtained with machine learning approaches based on other features and indicate if and to what extent a learning-based approach would be advantageous over a simple regular expression approach.

For the encoding of direct speech and thought, the TEI element `<said>` was used, as example 27 from the novel "Adoración" illustrates.

```
<p><said>—¿Pero ya estás bien, Daniel?</said>—me preguntó solícita.</p>
<p><said>—Bien del todo, Adoración. ¡Si no valió nada! Lo que siento solamente es los
    días que me privó de verte.</said></p>
<p>Adoración lanzó un suspiro.</p>
<p><said>—¡He pensado tanto en ti! Mira; hasta me atreví a suplicarle a Estrovo que
    averiguase la causa de tu ausencia.</said></p>
<p><said>—Sí; él acaba de decírmelo.</said></p>
<p><said>—¿Te lo contó?</said>—exclamó Adoración poniéndose muy colorada.<said>—Yo
    siento que haya sido una indiscreción de mi parte, pero estaba tan cuidadosa......</
    said></p>
<p>¡Qué remordimientos más atroces sentí entonces! Aquella pobre niña, con su afecto
    tierno y sincero tan ingenuamente expresado, era el mayor castigo que pudiera yo
    recibir por mi defección. La tomé una mano [...]</p>
```

Example 27. Encoding of direct speech in the novel "Adoración".

In "Adoración", the beginning of direct speech is marked with a long hyphen. If not otherwise indicated, the speech ends with the end of the paragraph (e.g., "—Sí; él acaba de decírmelo."). Alternatively, there may be a phrase closing the speech (e.g., "—me preguntó solícita.") which is itself introduced by a speech sign. In addition, there are insertions leading over from one speech act to the next inside of paragraphs, which are also marked with hyphens (e.g., "—¿Te lo contó?—exclamó Adoración poniéndose muy colorada.—Yo siento […]"). In the example, the direct speech corresponds to speech acts expressed by the characters in dialogue and is easy to distinguish from the surrounding narrative text, also because the speech signs are used in a consistent way in this edition of the novel "Adoración".

In general, however, the report of direct speech in novels is not limited to the representation of character dialogues. For example, the transition between direct and indirect speech and thought reported by the narrator can be smooth, as example 28 from the novel "S. Y." (1895, CU) by Francisco Calcagno shows.

```
<p>¿Qué infame suerte era aquella suya que lo condenaba siempre a ser deudor del hombre
    a quien ansiaba poder odiar? ¡Ah! ¡el viejo rey Priamo tuvo, que besar la mano del
    matador de su hijo; y él, Milanés, lleno de rencor y de odio, tenía que dar las
    gracias a quien le había de arrebatar a la mujer que amaba! ¡Y no poder ni vengarse!
     Un puñal en aquel alevoso pecho... no, ¡él no era hombre de puñal! pero frente a
    frente, cuerpo a cuerpo, con armas iguales, escarnecerlo, atacarlo, estrecharlo,
    confundirlo y atravesarle de parte a parte el pecho...</p>
<p><said>—¡Oh!... ¡si llegara ese momento</said>—,pensaba—, <said>aunque cayera sobre su
     cadáver el mío! Miserable que atrincherado en su colosal riqueza, me confunde a
```

```
    beneficios, y tal vez ríe de mi impotencia; me arrebata la felicidad, y seguro de su
     triunfo, tan impotente juzga a su rival que lo protege y lo enriquece y paga sus
    deudas y hasta le cede el puesto al lado de su novia. ¡Y no puedo batirme! ¡y no
    puedo matarlo! ¡y Jacinta me prohíbe que lo afrente! y tengo,... ¡ah! no; yo le
    arrojaré sus favores a la cara; mi odio será más potente que sus beneficios, yo le
    haré ver que prefiero la indigencia a la riqueza que proceda de su infame mano... yo
     le... pero ¡ay!... ¿y mi pobre madre?</said></p>
<p>¡Así suele cegarnos la pasión! así, ofuscado por el lóbrego porvenir de su amor
    desgraciado, se empeñaba en acriminar a aquel don Cristóbal cuya meritoria conducta
    mal su grado se veía obligado a reconocer y admirar.</p>
```

Example 28. Encoding of direct thought in the novel “S. Y.”.

Here, the thoughts of a character are represented, switching between free indirect and direct thought. Only the direct thought is encoded with the <said> element. In other cases, the narrator uses quotation marks to highlight individual words or passages, only some of which can be understood as citations of character speech, as in examples 29 and 30, taken from the novels “Los precursores” (1909, MX) by José López Portillo y Rojas and “La Ginesa” (1894, AR) by Carlos María Ocantos, respectively.

```
<p>Y para que nada faltase a sus inocentes hechizos, había recibido de Dios la índole
    más mansa y cariñosa que se ha visto. Nunca se oponía a nada, a todo estaba
    constantemente dispuesta; su complacencia era perpetua e intuitiva. La primera
    palabra que aprendió a decir, después de <q>"mamá"</q>, fue <q>"sí"</q>. A todo
    cuanto se le decía, contestaba que <said>"sí"</said>.</p>
```

Example 29. Encoding of direct speech in the novel “Los precursores”.

```
<p>Y en esto llegó una carta de la capital para Lía, y Lía la guardó sin abrirla,
    trémula, leyendo con los ojos del alma los garrapatos del niño adorado: <writing
    type="letter">«Ven, Ginesita; ¡si no vienes, me muero! ¡no puedo más! cruel,
    perversa, ingrata...»</writing> Llegó otra y también la guardó sin abrirla, pero
    adivinando cuanto decía: <writing type="letter">«Si no vienes, creeré que te has
    marchado con otro y que eso de haberte refugiado en Las Piedras, al lado de tu madre
    , es una papa...»</writing> Y tres más, cuyos sobres dejó intactos, aunque
    figurábaselas cariñosas, coléricas, o desesperadas; pero, cuando a la visita matinal
     del cartero sucedió la del telegrafista y Cándido la entregó, con ligera sonrisa
    irónica, el feo sobre color de ladrillo, diciéndole al oído: <said>«¿Será del hijo
    del patrón, que se impacienta?...»</said> no hubo más remedio que enterarse del
    despacho y que el niño dispuesto estaba <q>«a venir, si ella no iba»</q>. Tal susto
    llevó, que la fiebrecilla, que remitido había días atrás, la retentó nuevamente.</p>
```

Example 30. Encoding of direct speech in the novel “La Ginesa”.

In such passages, of the stretches in quotation marks, only those that are announced as speech by a reporting clause or that are recognizable as direct speech by the form of the pronouns and verbs are marked as direct speech. The others are interpreted as another form of emphasis by the narrator, which is marked up with the general element <q> for quoted material here.[329] In the first example, the words “mamá” and “sí” are first mentioned as general linguistic units and are therefore encoded with <q>. In the next sentence, the word “sí” is cited as an answer that the little girl Berta mentioned in the passage gives regularly and is therefore marked with <said>. In the second example, the quotation marks are used with several different functions. First, the

[329] This element is only used in the files in which also direct speech is annotated with the element <said> to be able to evaluate how many tokens contained in quotation marks are mistaken as direct speech.

content of two letters is cited («Ven, Ginesita [...] and «Si no vienes [...]) and marked as written text with the element <writing> here. Next, a question directly uttered by the character Cándido, which is marked as direct speech, follows: «¿Será del hijo del patrón [...]?». Third, the content of a telegram is cited in quotation marks but in indirect form («a venir, si ella no iba»). Therefore, it is not interpreted as direct speech but instead marked with <q>. In the above examples, the degree of mediation of character speech by the narrator varies inside of the same and between subsequent paragraphs. These examples show that beyond the classic character dialogue, there are cases where detailed decisions are required to draw the line between what is considered direct speech and what is not.

Another aspect that needs to be considered is that the thought and speech of characters, even if it is clearly represented directly, can take the form of a monologue or a longer argumentative or narrative passage. Without knowing the context of the utterances, it can be difficult to recognize that such passages are direct speech. Furthermore, direct speech can be reported on several levels in a novel. If a character speaks and becomes the narrator, he or she can cite the speech of other characters directly. The question is whether all character speech, independently of its functional text type,[330] should be treated as direct speech. Here it was decided to mark up direct speech on several levels so that nested structures are possible and to rely on the outer structure to decide if a token is part of direct speech or not. Nevertheless, if the speech of a character is announced as narration inside of the novel and extends over many paragraphs or whole chapters without being explicitly marked by speech signs, it is typified as "narration" with the attribute @ana, as example 31 shows.

```
<p><said>—¿Duermes, bella Cheherazada?</said> —dije a Laura cuando le hube contado seis
    horas de —sueño. <said>Pues si estás despierta, refiéreme, te ruego, esa interesante
     historia.</said></p>
<p><said>«De cómo Laura moribunda recobró la salud y la hermosura por la ciencia
    maravillosa de un médico homeópata».</said></p>
<p><said ana="#narration">Un día, uno de los peores de mi dolencia, en su interminable
    charla sobre las excelencias de la homeopatía, recordó la insigne calaverada de un
    joven cliente suyo, tísico en tercer grado, que apartándose del método por él
    prescrito, impuso a su arruinado pulmón la fatiga de interminables viajes.</said></p
    >
<p><said ana="#narration"><said>—Y, extraña aberración de la naturaleza</said> ——añadió,
     <said>aquel prolongado sacudimiento, aquel largo cansancio, lo salvaron; sanó...
    Pero son esos, casos aislados, excepcionales, que no pueden reproducirse. Aplíquese
    el tal remedio aquí, donde ya no hay sujeto; y en la primera etapa todo habrá
    acabado.</said></said> [...]</p>
```

Example 31. Encoding of direct speech in the novel "Peregrinaciones de un alma triste".

The example includes selected paragraphs of the first chapter of the novel "Peregrinaciones de un alma triste" (1876, AR) by Juana Manuela Gorriti. The novel is narrated in the first person. However, the principal narrator mainly cedes the word to her friend Laura who narrates her travels through Chile, Argentina, Paraguay, and Brazil, so that almost the whole novel could be interpreted as direct character speech. Even so, as Laura becomes the narrator, her speech is marked as *narration* here and is excluded from the direct speech analysis. On the other hand, the character speech cited by her, such as the words of the doctor in the example, are counted in.

[330] In linguistics, several functional text types have been distinguished, for example, descriptive, narrative, expository, argumentative, and instructive text (Werlich 1975, 30–34).

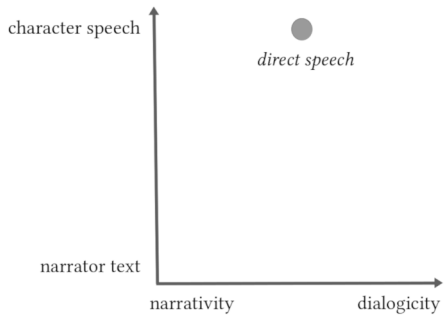
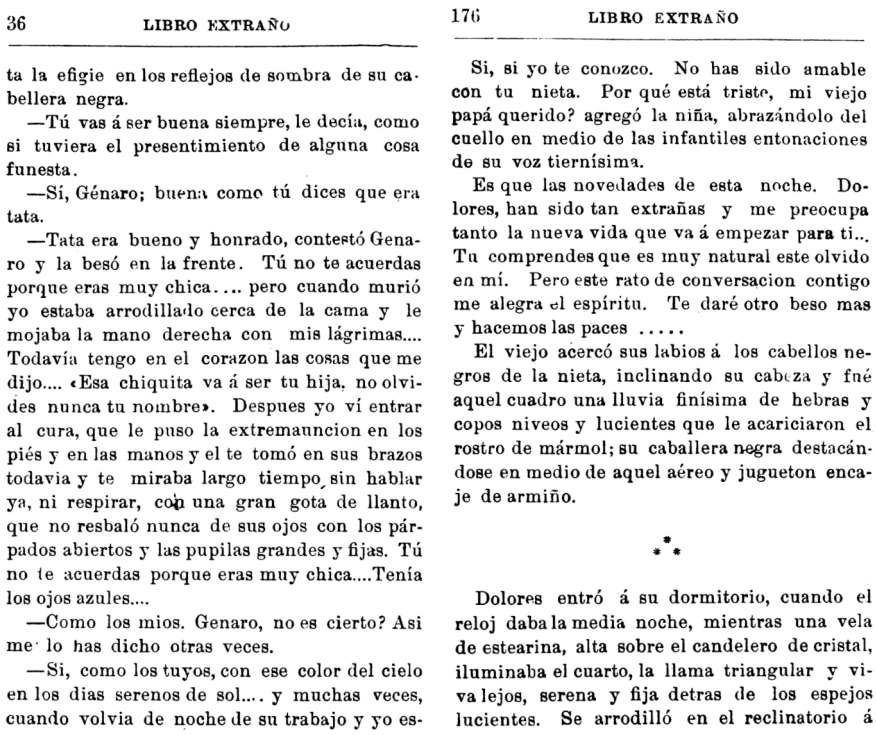


Figure 28. Characterization of the direct speech annotated in the corpus.

In figure 28, a sketch of the kind of direct speech annotated in this corpus is given along the two axes of narrator text versus character speech and narrativity versus dialogicity. Here it was decided to mark up character speech that can be part of a dialogue, but that can also be a monologue, and that can be narrative to a certain degree. Nevertheless, when passages are formally character speech (considering the overall structure of the novel) but linguistically and typographically not distinguishable from narrator text (considering the local context), they are marked as *narrative speech*. On the other hand, only speech that is contrasted with the narrator text through speech signs and/or linguistically recognizable as such is marked up as direct speech. Ambiguous forms are associated with the narrator text. Another axis that is not displayed in the figure is the one between written and spoken language. Here it was decided not to mark up representations of written text (e.g., a letter inserted into a novel) as direct speech unless they contain cited spoken language (e.g., a character dialogue cited in a letter inserted into a novel). However, written language can be close to oral speech (e.g., diary entries, notes, or letters) and is often also marked with the same signs as speech in the novels. The choices made here focus on the local context of the utterances and favor clear typographic and linguistic signs. This probably meets the characteristics of automatic recognition of direct speech quite well. Still, the passages marked up as direct speech are not limited to simple character dialogue. When comparing the results of direct speech recognition in fictional narrative prose texts, it should be kept in mind that the definition of direct speech underlying the analysis influences the results.

Turning to the question of how to capture direct speech technically, a rule-based approach can be used in the case of consistent use of speech signs, as in the edition of "Adoración" cited in the first example above. To this end, an XSLT script was created,[331] which marked all paragraphs beginning with a speech sign as direct speech or thought. Subsequently, the encoding was refined by also transforming insertions and closing phrases inside of paragraphs. Regular expressions were used inside the XSLT to detect relevant cases. The script differentiates between different types of speech signs (e.g., dashes versus angular or curved quotation marks) and speech sign types (single versus double). It focuses on one primary speech sign type per novel. The aim of the script is to detect as much direct speech as possible with rather simple rules.

[331] The script is available at https://github.com/cligs/scripts-nh/blob/master/corpus/metadata_encoding/copy-all-but-said.xsl. Accessed April 16, 2020.



36 LIBRO EXTRAÑO

ta la efigie en los reflejos de sombra de su cabellera negra.

—Tú vas á ser buena siempre, le decía, como si tuviera el presentimiento de alguna cosa funesta.

—Sí, Génaro; buena como tú dices que era tata.

—Tata era bueno y honrado, contestó Genaro y la besó en la frente. Tú no te acuerdas porque eras muy chica.... pero cuando murió yo estaba arrodillado cerca de la cama y le mojaba la mano derecha con mis lágrimas.... Todavía tengo en el corazon las cosas que me dijo.... «Esa chiquita va á ser tu hija, no olvides nunca tu nombre». Despues yo ví entrar al cura, que le puso la extremauncion en los piés y en las manos y el te tomó en sus brazos todavia y te miraba largo tiempo, sin hablar ya, ni respirar, con una gran gota de llanto, que no resbaló nunca de sus ojos con los párpados abiertos y las pupilas grandes y fijas. Tú no te acuerdas porque eras muy chica....Tenía los ojos azules....

—Como los mios. Genaro, no es cierto? Asi me lo has dicho otras veces.

—Si, como los tuyos, con ese color del cielo en los dias serenos de sol.... y muchas veces, cuando volvia de noche de su trabajo y yo es-

176 LIBRO EXTRAÑO

Si, si yo te conozco. No has sido amable con tu nieta. Por qué está triste, mi viejo papá querido? agregó la niña, abrazándolo del cuello en medio de las infantiles entonaciones de su voz tiernísima.

Es que las novedades de esta noche. Dolores, han sido tan extrañas y me preocupa tanto la nueva vida que va á empezar para ti... Tu comprendes que es muy natural este olvido en mí. Pero este rato de conversacion contigo me alegra el espíritu. Te daré otro beso mas y hacemos las paces .....

El viejo acercó sus labios á los cabellos negros de la nieta, inclinando su cabeza y fué aquel cuadro una lluvia finísima de hebras y copos niveos y lucientes que le acariciaron el rostro de mármol; su caballera negra destacándose en medio de aquel aéreo y jugueton encaje de armiño.

***

Dolores entró á su dormitorio, cuando el reloj daba la media noche, mientras una vela de estearina, alta sobre el candelero de cristal, iluminaba el cuarto, la llama triangular y viva lejos, serena y fija detras de los espejos lucientes. Se arrodilló en el reclinatorio á

Figure 29. Pages with direct speech from "Libro extraño" by Francisco Sicardi, with initial speech signs (left page) and without speech signs (right page).

Unfortunately, this strategy is not suitable to detect direct speech and thought reliably across all kinds of editions because there is much variation in the presence and absence of speech signs. In many cases, no signs are used at all. Sometimes only the narrative insertions and closing phrases are not marked, but there are also cases where even the beginning of the direct speech is not indicated with a special sign. Furthermore, when a character speaks over several paragraphs, usually, only the first one is marked with a speech sign, and the reader has to infer from other indicators that the speech continues. An automatic approach that is only based on speech signs can miss longer passages of direct speech in such cases. Even more complicated are editions where no consistent usage of speech signs can be recognized at all. An example is shown in figure 29, displaying two pages of the first edition of the novel "Libro extraño" (1894, AR) by Francisco Sicardi.[332]

On the first of the two example pages, speech signs are used, but only to mark the beginning of direct speech. Insertions in the middle of the speech are not marked typographically (for example: "—Tata era bueno y honrado, contestó Genaro y la besó en la frente. Tú no te acuerdas porque eras muy chica...."). Speech inside of speech is highlighted with angular quotation marks ("[...] las cosas que me dijo.... «Esa chiquita va á ser tu hija, no olvides nunca tu nombre»"). On

[332] The pages were obtained from the 1894 edition of the novel available at Sicardi 2016.

the second example page, in contrast, no speech signs are used at all: “Si, si yo te conozco […] Por qué está triste, mi viejo papá querido? agregó la niña […]”. In such editions, direct speech can hardly be captured with simple rules relying on punctuation and speech marks.

A case where the direct speech of a character continues in subsequent paragraphs without being indicated typographically is illustrated in example 32, taken from the novel “Puebla” (1903, MX), which is part of the work “La Intervención y el Imperio (1861–1867)” by Victoriano Salado Álvarez.

```
<p>Miró el mexicano a su compañero con cara de espanto, y el otro, sin esperar a que le
    pidieran explicaciones, habló así:</p>
<p><said>—Me llamo Nicolás Chardon, soy originario de París y mi padre es normando, de
    tierra de Rouen. Profesor de latín en las Universidades de provincia, no ha cesado
    un punto desde que se entronizó el Imperio, de hacerle la guerra mediante la
    propaganda republicana más activa...</said></p>
<p><said>El ministro Duruy, que atribuyó a mi padre los famosos <seg rend="italic">
    Propos de Labbiennus,</seg> le persiguió con durísima saña, pues aseguraba que
    ninguno de los profesores de Francia podía escribir una sátira tan erudita y tan
    mordaz... [...] </said></p>
<p>Refirió Miguel su vida y sus andanzas; y cariñoso el otro le ofreció su amistad y su
    afecto.</p>
```

Example 32. Encoding of direct speech in the novel “Puebla”.

In the example, the speech is introduced with the phrase “habló así:” and an opening speech sign. The character Nicolás Chardon talks about his origin and career over several paragraphs. The only way for the reader to know that the direct speech ends is to note the change of perspective that is signaled by the person of the verb forms and pronouns and by the mention of the characters involved: “Refirió Miguel su vida y sus andanzas; y cariñoso el otro le ofreció su amistad y su afecto”.

An additional factor complicating the automatic capture of direct speech indicated by hyphens is that the same sign may also be used as a marker for explanatory, meditative, or other kinds of parenthesis that are not direct speech, as depicted in example 33.

```
<p>Nunca me habló de su familia: yo creo que jamás la tuvo. Vivía sólo como un hongo,
    allá en su vivienda—,que como he dicho ya, no conocí nunca—,y solo también entre los
     arrecifes en que buscaba el sustento, ya con la caña, ya con el disparo en que era
    diestrísimo. Algunas veces me veía cerca y me saludaba con respetuoso afecto.</p>
```

Example 33. A parenthesis introduced with hyphens in the novel “Adoración”.

Angle and curved quotation marks, too, are not only used to mark direct speech but also for representations of written text, for quotations, for highlighting foreign words, or for other types of emphasis, as was shown above.

Because of the limitations of the regular expression-based approach using typographic speech signs, it was only applied to a subset of 92 novels in the corpus, which were then checked manually. To be able to estimate the loss of information caused by only using typographical indicators, the checked annotations were compared with the results obtained by the pure regular expression approach. To this end, tokenized versions of the 92 novels in TEI were created, to which stand-off markup with direct speech annotation was added. The first set of stand-off annotations is for the manually checked direct speech gold standard (`DS_gold`), and the second set is for the speech annotation based on regular expressions (`DS_reg`). In example 34, an excerpt of this derivative format is given for the novel “El guajiro” (1842, CU) by Cirilo Villaverde.

```
<w xml:id="p367.w135">las</w>
<w xml:id="p367.w136">palabras</w>
<w xml:id="p367.w137">no</w>
<w xml:id="p367.w138">le</w>
<w xml:id="p367.w139">salieron</w>
<w xml:id="p367.w140">enteras</w>
<w xml:id="p367.w141">:</w>
<w xml:id="p368.w1">—</w>
<w xml:id="p368.w2">¡</w>
<w xml:id="p368.w3">Señores</w>
<w xml:id="p368.w4">,</w>
<w xml:id="p368.w5">fuera</w>
<w xml:id="p368.w6">de</w>
<w xml:id="p368.w7">la</w>
<w xml:id="p368.w8">valla</w>
<w xml:id="p368.w9">!</w>
<w xml:id="p369.w1">Despejada</w>
<w xml:id="p369.w2">enteramente</w>
<w xml:id="p369.w3">la</w>
<w xml:id="p369.w4">valla</w>
<linkGrp type="DS_gold">
  [...]
  <link target="#p367.w135 #NARR"/>
  <link target="#p367.w136 #NARR"/>
  <link target="#p367.w137 #NARR"/>
  <link target="#p367.w138 #NARR"/>
  <link target="#p367.w139 #NARR"/>
  <link target="#p367.w140 #NARR"/>
  <link target="#p367.w141 #NARR"/>
  <link target="#p368.w1 #DS"/>
  <link target="#p368.w2 #DS"/>
  <link target="#p368.w3 #DS"/>
  <link target="#p368.w4 #DS"/>
  <link target="#p368.w5 #DS"/>
  <link target="#p368.w6 #DS"/>
  <link target="#p368.w7 #DS"/>
  <link target="#p368.w8 #DS"/>
  <link target="#p368.w9 #DS"/>
  <link target="#p369.w1 #NARR"/>
  <link target="#p369.w2 #NARR"/>
  <link target="#p369.w3 #NARR"/>
  <link target="#p369.w4 #NARR"/>
  [...]
</linkGrp>
<linkGrp type="DS_reg">
  <link target="#p367.w135 #NARR"/>
  <link target="#p367.w136 #NARR"/>
  <link target="#p367.w137 #NARR"/>
  [...]
</linkGrp>
```

Example 34. Excerpt from the tokenized version of the novel “El guajiro”, with stand-off direct speech annotation.

As the example shows, the direct speech annotation is made per word token. Here, the words with the identifiers p367.w135 up to p357.w141 are marked as narrated text (`#NARR`) in the gold standard, followed by direct speech (`#DS`) up to word p368.w9, continuing with narrated text. The structure of the second annotation set DS_reg is the same as for the gold standard so that it

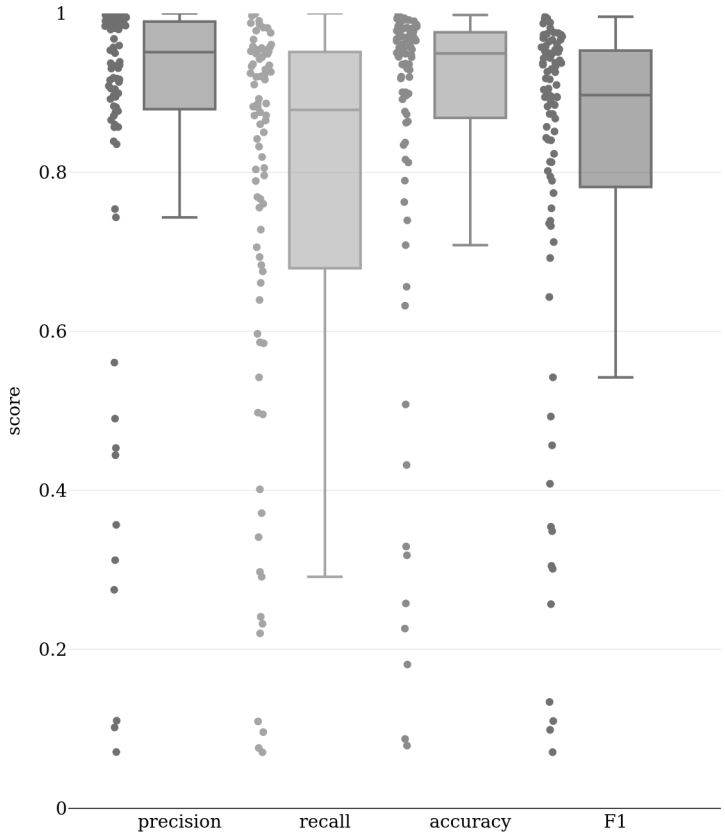


Figure 30. Scores for direct speech recognition (gold standard versus regular expression approach).

is possible to compare directly whether there are differences in the two approaches.[333] This was done by calculating the precision, recall, accuracy, and F1 scores for all the novels, and comparing the `DS_gold` annotation with `DS_reg`.[334] The resulting scores are displayed in figure 30.

The median F1 score is at 90 %, which is quite a good result for a regular expression-based approach. It is comparable to the results achieved with machine learning approaches in other

[333] The script with which the tokenized TEI files and direct speech stand-off mark-up were produced is available at: https://github.com/cligs/scripts-nh/blob/master/corpus/metadata_encoding/evaluation-direct-speech.xsl. As a basis for this script, on the one hand, the TEI master files of the corpus that contain checked direct speech mark-up were used (as available at https://github.com/cligs/conha19/tree/master/tei). On the other hand, versions of the same TEI files without direct speech markup (available at https://github.com/cligs/conha19/tree/master/tei_ns) were treated with the pure regular expression approach producing a second version with direct speech markup (available at https://github.com/cligs/conha19/tree/master/tei_ds). These two versions of TEI files were then evaluated with the mentioned script. All links were accessed on August 16, 2020.

[334] These calculations, as well as figures 30 to 32, were produced with the same script (see footnote 333). An overview of the scores in CSV format is available at https://github.com/cligs/data-nh/blob/master/corpus/metadata-encoding/direct-speech-evaluation-F1.csv. Accessed August 16, 2020.

studies.[335] For the second and third quartile, the scores range from 80 % to 95 %, which also seems acceptable, but when also outliers are considered, the dispersion of values is broad, and there are some cases with very low scores. This means that the regular expression-based approach is very successful in many cases, but apparently, it fails in some cases, so it is not very reliable. Considering not only F1 but also other types of scores, the differences between them point out some strengths and weaknesses of the regular expression approach. The precision and accuracy scores are higher and vary less than the recall scores, which means that there are more false negatives (i.e., actual direct speech tokens that were not recognized) than false positives (i.e., actual tokens of narrated text that were mistaken as direct speech). So apparently, whole paragraphs of direct speech that are missed because there is no initial speech sign weigh more in quantitative terms than individual tokens of narrated text that are contained in paragraphs with initial speech signs but not marked explicitly by further signs, at least if the speech signs are single dashes and not double marks. In figures 31 and 32, the F1 scores are differentiated by the kind of source edition (modern, historical, or unknown)[336] and by the type of speech sign (single or double) to see if these factors have an influence on the results.

Contrary to what one would expect, the median F1 score is similar for the three kinds of editions: 91 % for historical editions, and 90 % for both modern and unknown editions, so historical editions are not more problematic than other kinds of editions. However, the comparison of types of editions relies on different group sizes. In the corpus, there are 158 historical editions, but only 78 modern ones, and 20 cases where the kind of source edition is unknown. The results might be different in a dataset that is more balanced in this aspect. Above, several factors complicating the use of speech signs for direct speech recognition were discussed. The findings for the different kinds of source editions suggest that inconsistencies in the usage of speech signs, which were recognized more often in historical editions than in modern ones, are not decisive.

To look at the type of speech sign is not very instructive, either, because, in the whole corpus, there are only three novels based on source editions with double speech signs.[337] To get a better sense of the factors influencing the results, it would be necessary to inspect the passages and tokens that were misclassified, which is considered a future work. Furthermore, this semi-automatically edited gold standard can be used to develop a machine-learning workflow to see if even better results can be achieved with it. Moreover, such a workflow could be reused in other contexts. Research into the automatic detection of direct speech in narrative texts is not abundant

---

[335] An F1 score of 0.939 has been reported for the recognition of direct speech in nineteenth-century French novels (Schöch, Schlör, et al. 2016), and an accuracy of 0.9 for German novels (Jannidis et al. 2018). Brunner achieved an F1 score of 0.87 for the recognition of direct speech, thought, and writing representation in German narrative texts (Brunner 2013). In their approach to a multilingual collection of nineteenth-century novels, Byszuk et al. report F1 scores ranging between 0.65 and 0.98 for the different languages when comparing the results of a regular expression approach with manually annotated samples. In their multilingual deep learning-based approach, they achieve a general F1 score of 0.873 (Byszuk et al. 2020).

[336] In the metadata of the corpus, four kinds of source editions are differentiated: “first”, (other) “historical”, “modern”, and “unknown”. Here, the two categories, “first” and “historical”, are joined.

[337] These are “Lucía Miranda” (1860, AR) by Eduarda Mansilla de García, “Pot-pourri” (1882, AR) by Eugenio Cambaceres, and “El fatalista” (1866, CU) by Estebán Pichardo y Tapia. In the corpus, the text of the first novel is based on an edition published in Buenos Aires in 1882, the second one on an edition published in Argentina in 1984, and the third one on the first edition published in Havanna. In all three, double angular quotation marks are used instead of the usual single hyphen.

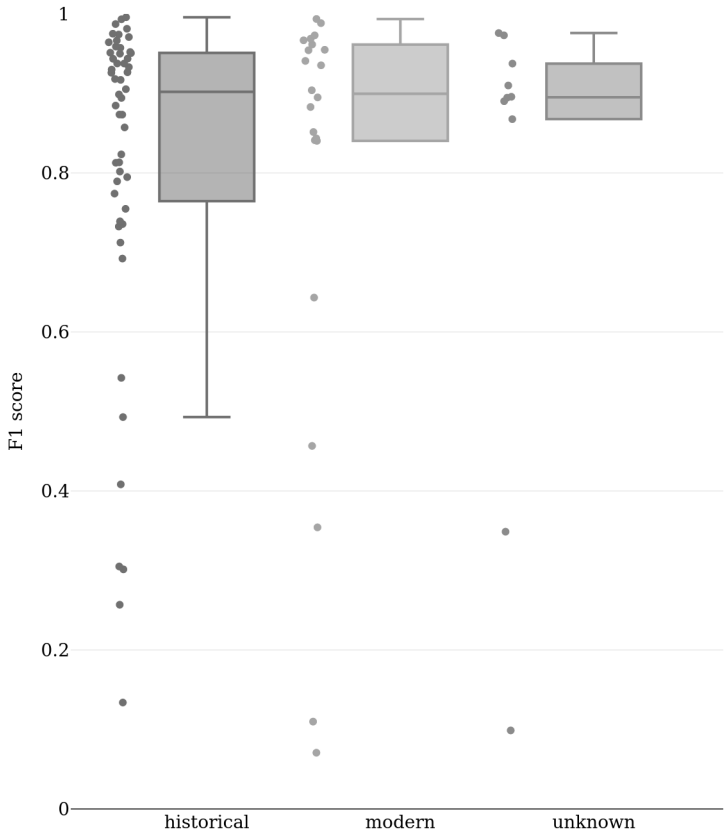


Figure 31. F1 scores for direct speech recognition by kind of edition.

and has not been conducted based on a corpus of Spanish-American novels yet. Developing such a workflow would, therefore, also be of interest from a methodological point of view.

3.3.3.2.9 Embedded Texts

To conclude the presentation of TEI elements that were used to mark up the text body of the novels, the element `<floatingText>`, which serves to encode embedded texts, needs to be introduced. It can contain an entire textual body with divisions, paragraphs, etc. In the TEI guidelines, this element is defined as follows: "`<floatingText>` contains a single text of any kind, whether unitary or composite, which interrupts the text containing it at any point and after which the surrounding text resumes" (Text Encoding Initiative Consortium 2023f). It is thus a useful element to encode passages in novels that occur inside of chapters but have their own structure, for example, an own title, an own heading, or own chapters.[338] The encoding of floating texts is shown in examples 35 and 36.

[338] It would not be possible to use a simple division element (`<div>`) for that purpose because, following the rules of the TEI, any divisional structure opened inside another one has to be continued. That is, it would cause an error to continue with paragraphs of the running chapter after the insertion of a division for an embedded text.

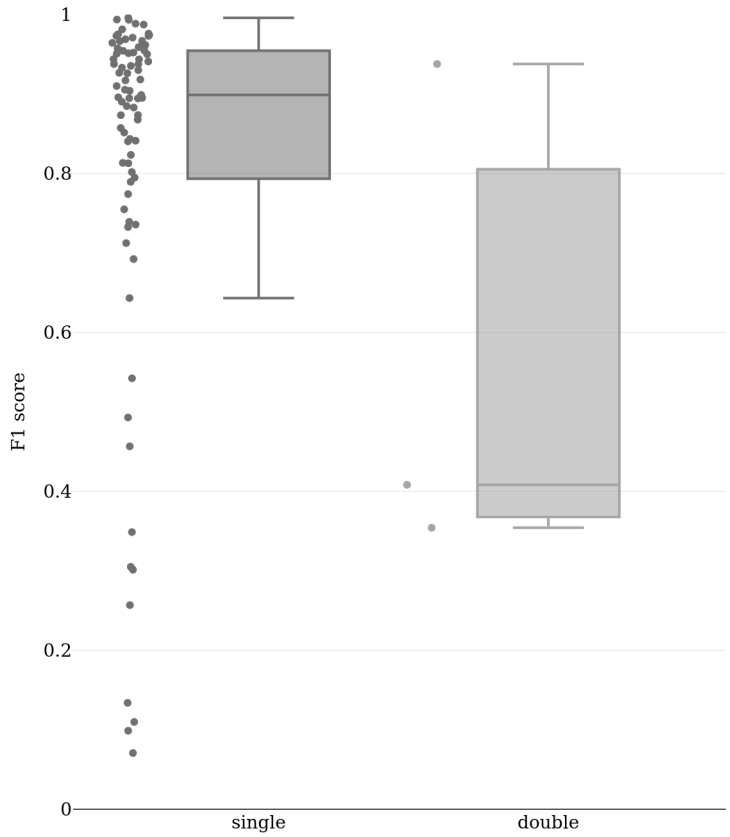


Figure 32. F1 scores for direct speech recognition by type of speech sign.

```
<p>D. Luis empezó a leer la siguiente carta:</p>
<said>
  <writing type="letter">
    <floatingText>
      <body>
        <div>
          <p>Parroquia de Santa Maña, Febrero 2 de 1798.</p>
          <p><seg rend="italic">Señor D. Luis Ferri.</seg></p>
          <p>«Mi querido padre:</p>
          <p>Después de largo tiempo que ha trascurrido sin tener noticias de vd. a
              pesar de las repetidas cartas que le he escrito, vuelvo otra vez a
              dirigirle esta para anunciarle que dentro de poco, si termino unos asuntos
               que tengo entre manos, iré a su lado para recibir su bendición y
              abrazarlo con el cariño que sabe le profeso.</p>
        </div>
      </body>
    </floatingText>
  </writing>
</said>
```

Example 35. Encoding of an embedded letter in the novel “Amelia de Floriani o el castillo del diablo”.

The first example is taken from the novel “Amelia de Floriani o el castillo del diablo” (1887, AR) by José Victoriano Cabral and shows a letter that is read aloud by the character D. Luis. The element <floatingText> is used to mark that the various paragraphs of the letter, i.e., the dateline, address, salutation, and the text itself, belong together. In addition, the letter is marked as written text (<writing type="letter">) and as direct speech (<said>) because it is read by a character.

```
<p>No. No podía ser un contemporáneo, porque sintetizaba demasiado. Uno de mis camaradas
     hubiera entrado en mayores detalles, no hubiera visto las cosas a bulto, hubiera
    cometido menos errores. Vean ustedes: aquí tengo el recorte, con su título y todo:</
    p>
<writing type="newspaper">
  <floatingText>
    <body>
      <div>
        <head><seg rend="capital">Divertidas aventuras del nieto de Juan Moreira</seg></
            head>
        <p>«Tan ignorante y tan dominador como el abuelo, nació en un rincón de
            provincia, y creció en él sin aprender otra cosa que el amor de su persona y
             la adoración de sus propios vicios.</p>
        [...]
      </div>
    </body>
  </floatingText>
</writing>
<p>Y seguía una larga serie de anécdotas, casi todas falsas —entre ellas el <q>'
    envenenamiento'</q> de —Camino, pero tras de cuyas líneas se transparentaba
    claramente mi persona, para terminar diciendo:</p>
<writing type="newspaper">
  <floatingText>
    <body>
      <div>
        <p>»El que esto escribe no quiere mal al nieto de Juan Moreira, ni a don
            Mauricio Gómez Herrera, ni a... ¡tantos otros! [...]</p>
        <p>»¡Que el nieto de Juan Moreira nos represente en Europa! [...]</p>
      </div>
    </body>
  </floatingText>
</writing>
<p>Y firmaba <writing type="newspaper">«Mauricio Rivas»</writing>.</p>
```

Example 36. Encoding of an embedded newspaper article in the novel “Divertidas aventuras del nieto de Juan Moreira”.

The second example includes sections of a newspaper article that are cited in the novel “Divertidas aventuras del nieto de Juan Moreira” (1910, AR) by Roberto Payró. Here the text is not read but shown to the reader by the narrator (“Vean ustedes: aquí tengo el recorte”). Not the whole newspaper article is represented, but some excerpts that the narrator comments on. All the parts are marked up as representations of written text (<writing type="newspaper">), but only those with a cohesive structure that is more complex as a single paragraph are marked additionally as floating texts.

In this corpus, the element <floatingText> is only used for structural reasons. The classification of the embedded text as a representation of written text, as direct speech, as a quote, etc. is expressed through elements that are defined more narrowly semantically, such as <writing>, <said>, <quote>, and so on, which can all be wrapped around a floating text or be used independently of it inside of individual paragraphs. Furthermore, the examples show that the floating

texts can both be texts that are embedded as a whole, as the letter in the novel “Amelia de Floriani”, or partially and subsequently, as the newspaper article in the novel “Divertidas aventuras del nieto de Juan Moreira”.

Overall, encoding the novels in the TEI body served several purposes. Some phenomena were only marked up to keep structural information that existed in the texts’ source files (typographically marked subdivisions of the text and typographically highlighted words or phrases) and to document missing information (gaps). Other structures were marked up because they are of interest for the analysis of subgenres of the novel (verse lines, dramatic text, representations of written text, quotations, and direct speech or thought), and finally, floating texts were marked up to achieve a valid TEI structure. The TEI offers more elements to encode information in literary narrative texts, and more levels of information than the ones chosen here could be useful for genre analysis, so the choices made for this corpus are a selection focusing on the insertion of non-narrative generic forms and on the representation of writing, speech, and thought.

3.3.3.3 TEI Schema

The encoding of the novels is controlled by a RELAX NG schema, which in turn is based on a more abstract ODD file.[339] The RELAX NG schema makes sure that the general TEI vocabulary and structure of the corpus are consistent. It is complemented by a Schematron file that serves to check the structure and content of the metadata in the TEI header in a more detailed way (see chapter 3.3.3.1.6 on text classification in the TEI header above). Links to both schema files are included as processing instructions in each of the corpus files, as shown in example 37.

```
<?xml-model href="../schema/keywords.sch" type="application/xml" schematypens="http://
    purl.oclc.org/dsdl/schematron"?>
<?xml-model href="https://raw.githubusercontent.com/cligs/reference/master/tei/cligs.rng
    " type="application/xml" schematypens="http://relaxng.org/ns/structure/1.0"?>
```

Example 37. Processing instructions in a TEI corpus file.

The Schematron file (“keywords.sch”) is corpus specific and is therefore kept inside of the same repository.[340] The RELAX NG schema (“clings.rng”) and the underlying ODD file, on the other hand, are designed more generally for all the corpora developed in the CLiGS project and are therefore stored in a separate repository called “reference”.[341]

The CLiGS TEI schema includes elements that are basic for the encoding of literary narrative, dramatic, and poetic texts. However, it avoids other specialized block-level and inline elements. Its definition was kept as restrictive as possible, only allowing for elements and attributes that are actually in use in the different corpora. Compared to other established TEI customizations such as TEI Lite, TEI Simple, or the DTA-Basisformat (DTABf), the CLiGS schema is more restrictive,

---

[339] The TEI standard offers many different modules for encoding all types of texts. Normally, projects dealing with a specific type of text define a subset of the TEI’s modules to work with. Such a subset can be documented in a so-called ODD (“One document does it all”) file, a format that is used to express TEI customizations and that is independent of other formal schema languages. Schemas in different languages can be generated from the ODD in a second step. RELAX NG is one schema language that is suitable for XML (Murata 2014; Text Encoding Initiative Consortium n.d.b)

[340] See https://github.com/cligs/conha19/blob/master/schema/keywords.sch on GitHub. Accessed August 19, 2020.

[341] See https://github.com/cligs/reference. Accessed August 20, 2020.

although it is not an exact subset of either of them. On the other hand, a few attributes have been added to the schema in the project-specific namespace "https://cligs.hypotheses.org/ns/cligs" (Schöch et al. 2019, paras 14–18). In the main TEI corpus files, the only additional attribute is `@cligs:importance`, used to assign degrees of importance to metadata category values, for example, of different subgenre assignments.[342] Second, several custom attributes have been added to the schema to hold linguistic annotations produced with the NLP package FreeLing.[343]

A Python script is used to test the validity of the corpus files against the RELAX NG schema. It produces a log file reporting the success or failure of the validation process for each TEI file.[344] Validating the TEI files against the Schematron file requires a different strategy. In principle, Schematron validation is possible with the Python module "lxml", but only if the queries used in the Schematron file conform to the XSLT 1.0 standard (Behnel 2022). To check that the metadata keywords in the TEI header conform to the keyword taxonomy, however, it was necessary to also use XSLT 2.0 expressions in the Schematron file. An alternative way for validation without Python is to compile the Schematron file as XSLT and apply this transformation script to all the TEI files in the corpus using Saxon directly from the command line. The error output of this transformation process is stored in a log file.[345]

To summarize this chapter on the encoding of the text corpus, it can be said that this thesis focuses on the encoding of detailed metadata about the novels rather than on a very detailed encoding of the texts themselves. This is due to the kind of resource that the corpus is intended to be: it is an edited text collection aimed to serve as the basis for quantitative genre analysis where metadata about the authors, source editions, the form and content of the texts and, above all, about the subgenres that the novels have been assigned to plays an important role. For the encoding of the textual body, a special emphasis was put on the markup of direct speech in a subset of the novels. In the next section, the assignment of subgenre labels to the novels, which was bypassed in this general chapter about metadata and text encoding, is set out in more detail.

3.3.4 Assignment of Subgenre Labels

In principle, the assignment of subgenre labels to the novels in the corpus Conha19 follows the same criteria as the assignment of subgenre labels to the novels contained in the digital

---

[342] See the examples in chapter 3.2.3 above.

[343] The output of the FreeLing tagger is integrated into the TEI structure of the main corpus files, and the results are stored as a derivative TEI format, as explained further in chapter 3.3.5 below.

[344] The Python file for validation is accessible at https://github.com/cligs/scripts-nh/blob/master/corpus/metadata_encoding/validate_tei.py. The resulting log file can be viewed at https://github.com/cligs/conha19/blob/master/schema/log-rng.txt. Accessed August 20, 2020.

[345] Here, the compilation of the Schematron file as XSLT was done with the software Oxygen by choosing the preconfigured Transformation Scenario "ISO Schematron to XSLT (compile)". The resulting XSLT file can be viewed at https://github.com/cligs/conha19/blob/master/schema/keywords-compiled.xsl. The subsequent command line call for saxon is: `java -jar /home/ulrike/Programme/saxon/saxon9he.jar -s:/home/ulrike/Git/conha19/tei/ -o:/home/ulrike/Git/conha19/tei-checked/ -xsl:/home/ulrike/Git/conha19/schema/keywords-compiled.xsl > /home/ulrike/Git/conha19/schema/log-schematron.txt`. The resulting Schematron log file is accessible at https://github.com/cligs/conha19/blob/master/schema/log-schematron.txt. If the file is empty, now errors were found. Accessed August 20, 2020.

bibliography Bib-ACMé, as presented in chapter 3.2.3 above. The same literary-historical sources and bibliographic information were used to collect subgenre labels, the same discursive model to organize them, and the same encoding strategies to express them. In contrast to the novels in the bibliography, however, more information that is relevant to the subgenre assignment is available from the full-text editions of the novels in the corpus. This chapter briefly summarizes the overall encoding of subgenre labels in the corpus files. It focuses on the kind of labels that were only added to the novels in the corpus but not in the bibliography.

As in Bib-ACMé, also in Conha19, subgenre labels were collected from a selection of literary-historical sources. On the other hand, explicit and implicit indications of subgenres in the titles of the novels' editions were evaluated. To recapitulate, example 38 shows the subgenre labels that were added to the bibliographic entry of the novel "Rastaquouère" (1890, ARG) by Alberto del Solar.

```
<bibl xml:id="W1234">
  <author key="A464">Solar, Alberto del</author>
  <title>Rastaquouère</title>
  <term type="subgenre.title.explicit">Ilusiones y desengaños sudamericanos en París</
      term>
  <term type="subgenre.title.implicit" resp="#uhk">novela naturalista</term>
  <term type="subgenre.title.implicit" resp="#uhk">novela realista</term>
  <term type="subgenre.litHist" resp="#Schlickers_2003">novela naturalista</term>
  <term type="subgenre.litHist" resp="#Sanchez_1953">novela de tendencia mixta</term>
  <term type="subgenre.litHist" resp="#Sanchez_1953">novela social</term>
  <term type="subgenre.litHist.interp" resp="#uhk">novela naturalista</term>
  <term type="subgenre.litHist.interp" resp="#uhk">novela realista</term>
  <term type="subgenre.litHist.interp" resp="#uhk">novela social</term>
  <term type="subgenre.summary.signal.explicit" resp="#uhk">novela social</term>
  <term type="subgenre.summary.signal.explicit" resp="#uhk">novela de costumbres</term>
  <term type="subgenre.summary.signal.explicit" resp="#uhk">estudio</term>
  <term type="subgenre.summary.signal.explicit" resp="#uhk">cuadros</term>
  <term type="subgenre.summary.signal.implicit" resp="#uhk">novela naturalista</term>
  <term type="subgenre.summary.signal.implicit" resp="#uhk">novela realista</term>
  <term type="subgenre.summary.theme.explicit" resp="#uhk" cligs:importance="2">novela
      social</term>
  <term type="subgenre.summary.theme.explicit" resp="#uhk">novela de costumbres</term>
  <term type="subgenre.summary.theme.litHist" resp="#uhk">novela social</term>
  <term type="subgenre.summary.current.implicit" resp="#uhk" cligs:importance="2">novela
       naturalista</term>
  <term type="subgenre.summary.current.implicit" resp="#uhk">novela realista</term>
  <term type="subgenre.summary.current.litHist" resp="#uhk">novela naturalista</term>
  <term type="subgenre.summary.current.litHist" resp="#uhk">novela realista</term>
  <term type="subgenre.summary.mode.intention.explicit" resp="#uhk">estudio</term>
  <term type="subgenre.summary.mode.medium.explicit" resp="#uhk">cuadros</term>
  <term type="subgenre.summary.mode.representation.explicit" resp="#uhk">cuadros</term>
  <term type="subgenre.summary.mode.representation.explicit" resp="#uhk">estudio</term>
  <idno type="cligs">nh0255</idno>
  [...]
</bibl>
```

Example 38. Subgenre labels for the novel "Rastaquouère" in the work list of Bib-ACMé.

The novel "Rastaquouère" has the explicit subtitle "Ilusiones y desengaños sudamericanos en París" (as encoded in the term `"subgenre.title.explicit"`), which is interpreted as a sign for a naturalistic and realist novel (`"subgenre.title.implicit"`). In literary-historical works, the novel has been classified as *novela naturalista*, *novela de tendencia mixta*, and *novela social*

("subgenre.litHist"). The literary-historical assignments are interpreted and normalized in terms of the type "subgenre.litHist.interp". Following this, the different subgenre label values are summarized to capture values that are signaled in the text explicitly or implicitly ("subgenre.summary.signal.explicit" and "subgenre.summary.signal.implicit"). In addition, the values are summarized to sort them according to the discursive model developed in chapter 3.2.3 above ("subgenre.summary.theme", "subgenre.summary.current", "subgenre.summary.mode", etc.). In the summarizing part, all subgenre labels are included, not only the ones derived directly from the title of the work and from literary histories (as for all bibliographic entries in Bib-ACMé) but also values that were only collected for the texts in the corpus. The values "estudio" and "cuadros", for instance, are added here as a result of a further examination of generic signals for the novel "Rastaquouère" because it is part of the text corpus. The origin of these additional subgenre labels will now be explained.

For the novels in the corpus, beyond the work title, also other paratextual elements were assessed, including further information on title pages, in dedications, prefaces, headings, or tables of content. All these elements are part of the peritext, i.e., the paratexts that are published together with the work itself. Exceptionally, also information from the epitext, i.e. paratexts outside of the immediate context of the work, was considered, for example, statements about the subgenre of a novel made by contemporaries and published elsewhere. However, this kind of information was not researched systematically. Finally, in cases where no subgenre signals were available from the paratexts, the opening of the novels, typically the first chapter, was evaluated.[346] In the TEI files of the corpus, the explicit and implicit generic signals are collected in the keyword section of the TEI header.[347] Besides the terms that were already used in the work list of the bibliography Bib-ACMé for the assignment of subgenres, some additional terms are available in the corpus, as listed in table 19.

The genre signals that occur in the wider paratext of the work, i.e., beyond the work's title, are collected in terms of the type "text.genre.subgenre.paratext", differentiating between explicit and implicit signals. Statements made by contemporaries about the subgenre of a novel are encoded in a term of the type "text.genre.subgenre.contemp.explicit". Subgenre signals that are interpreted from the opening of a novel are given in the keyword type "text.genre.subgenre.opening.interp". In addition, three keywords of the type "text.genre.subgenre.historical" serve to summarize all previous explicit and implicit values. All of these terms may occur several times in the keyword section of a novel's TEI file. Example 39 represents the encoding of the entirety of subgenre labels in the corpus file of the novel "Rastaquouère".

```
<keywords scheme="../schema/keywords.xml">
  [...]
  <term type="text.title" n="1890">Rastaquouère. Ilusiones y desengaños sudamericanos en
       París</term>
  <term type="text.genre.subgenre.title.explicit">Ilusiones y desengaños sudamericanos
      en París</term>
```

[346] Generic signals can occur in the entire text of literary works, but the opening is the most prominent place for them: "The generic markers that cluster at the beginning of a work have a strategic role in guiding the reader. They help to establish, as soon as possible, an appropriate mental 'set' that allows the work's generic codes to be read. One might call them the key words of the code, although they may serve this purpose at an unconscious level, or at least beneath the level of attention" (Fowler 1982, 88).

[347] See chapter 3.3.3.1.6 for a general explanation of the encoding of metadata in the keyword section.

Keyword type	Description
text.genre.subgenre.paratext.explicit	the subgenre as given explicitly and literally in the paratext of the work (beyond the title)
text.genre.subgenre.paratext.implicit	the subgenre as indicated by implicit genre signals in the paratext of the work (beyond the title)
text.genre.subgenre.contemp.explicit	the subgenre as given explicitly and literally in statements made by contemporaries
text.genre.subgenre.opening.interp	the subgenre as interpreted from genre signals in the opening of the text (e. g., in the first chapter)
text.genre.subgenre.historical.explicit	a summary of text.genre.subgenre.title.explicit, text.genre.paratext.explicit, and text.genre.subgenre.contemp.explicit
text.genre.subgenre.historical.explicit.norm	a normalized version of the historical subgenre label
text.genre.subgenre.historical.implicit	a summary of text.genre.subgenre.title.implicit, text.genre.subgenre.paratext.implicit, and text.genre.subgenre.opening.interp

Table 19. Additional keyword terms for subgenre signals in the text corpus.

```
<term type="text.genre.subgenre.title.implicit" resp="#uhk">novela naturalista</term>
<term type="text.genre.subgenre.paratext.explicit">estudio de crítica social</term>
<term type="text.genre.subgenre.paratext.explicit">escritor de costumbres</term>
<term type="text.genre.subgenre.paratext.explicit">pintor de cuadros de circunstancias
    </term>
<term type="text.genre.subgenre.paratext.explicit">estudio de las costumbres</term>
<term type="text.genre.subgenre.paratext.explicit">Balzac</term>
<term type="text.genre.subgenre.paratext.explicit">Comedia Humana</term>
<term type="text.genre.subgenre.paratext.implicit" resp="#uhk">novela realista</term>
<term type="text.genre.subgenre.paratext.implicit" resp="#uhk">novela de costumbres</
    term>
<term type="text.genre.subgenre.paratext.implicit" resp="#uhk">novela naturalista</
    term>
<term type="text.genre.subgenre.paratext.implicit" resp="#uhk">novela social</term>
<term type="text.genre.subgenre.historical.explicit">Ilusiones y desengaños
    sudamericanos en París</term>
<term type="text.genre.subgenre.historical.explicit">estudio de crítica social</term>
<term type="text.genre.subgenre.historical.explicit">estudio de las costumbres</term>
<term type="text.genre.subgenre.historical.explicit.norm" resp="#uhk">estudio</term>
<term type="text.genre.subgenre.historical.explicit.norm" resp="#uhk">novela social</
    term>
<term type="text.genre.subgenre.historical.explicit.norm" resp="#uhk">novela de
    costumbres</term>
<term type="text.genre.subgenre.historical.explicit.norm" resp="#uhk">cuadros</term>
<term type="text.genre.subgenre.historical.implicit" resp="#uhk">novela naturalista</
    term>
<term type="text.genre.subgenre.historical.implicit" resp="#uhk">novela de costumbres<
    /term>
<term type="text.genre.subgenre.historical.implicit" resp="#uhk">novela realista</term
    >
<term type="text.genre.subgenre.historical.implicit" resp="#uhk">novela social</term>
<term type="text.genre.subgenre.litHist" resp="#Schlickers_2003">novela naturalista</
    term>
```

```
  <term type="text.genre.subgenre.litHist" resp="#Sanchez_1953">novela de tendencia
      mixta</term>
  <term type="text.genre.subgenre.litHist" resp="#Sanchez_1953">novela social</term>
  <term type="text.genre.subgenre.litHist.interp" resp="#uhk">novela naturalista</term>
  <term type="text.genre.subgenre.litHist.interp" resp="#uhk">novela realista</term>
  <term type="text.genre.subgenre.litHist.interp" resp="#uhk">novela social</term>
  <term type="text.genre.subgenre.summary.signal.explicit" resp="#uhk">novela social</
      term>
  <term type="text.genre.subgenre.summary.signal.explicit" resp="#uhk">novela de
      costumbres</term>
  <term type="text.genre.subgenre.summary.signal.explicit" resp="#uhk">estudio</term>
  <term type="text.genre.subgenre.summary.signal.explicit" resp="#uhk">cuadros</term>
  <term type="text.genre.subgenre.summary.signal.implicit" resp="#uhk">novela
      naturalista</term>
  <term type="text.genre.subgenre.summary.signal.implicit" resp="#uhk">novela realista</
      term>
  <term type="text.genre.subgenre.summary.theme.explicit" resp="#uhk" cligs:importance="
      2">novela social</term>
  <term type="text.genre.subgenre.summary.theme.explicit" resp="#uhk">novela de
      costumbres</term>
  <term type="text.genre.subgenre.summary.theme.litHist" resp="#uhk">novela social</term
      >
  <term type="text.genre.subgenre.summary.current.implicit" resp="#uhk" cligs:importance
      ="2">novela naturalista</term>
  <term type="text.genre.subgenre.summary.current.implicit" resp="#uhk">novela realista<
      /term>
  <term type="text.genre.subgenre.summary.current.litHist" resp="#uhk">novela
      naturalista</term>
  <term type="text.genre.subgenre.summary.current.litHist" resp="#uhk">novela realista</
      term>
  <term type="text.genre.subgenre.summary.mode.intention.explicit" resp="#uhk">estudio</
      term>
  <term type="text.genre.subgenre.summary.mode.medium.explicit" resp="#uhk">cuadros</
      term>
  <term type="text.genre.subgenre.summary.mode.representation.explicit" resp="#uhk">
      cuadros</term>
  <term type="text.genre.subgenre.summary.mode.representation.explicit" resp="#uhk">
      estudio</term>
  [...]
</keywords>
```

Example 39. Encoding of subgenre labels in the novel “Rastaquouère” in the corpus file.

As can be seen, several explicit and implicit subgenre labels stem from the paratext of the novel and are added to the ones derived from the work’s title and from literary-historical works: “estudio de crítica social”, “escritor de costumbres”, “pintor de cuadros de circunstancias”, “estudio de las costumbres”, “Balzac”, “Comedia Humana” as explicit terms and “novela realista”, “novela de costumbres”, “novela naturalista”, and “novela social” as implicit ones interpreted from the paratexts. The explicit values may be all kinds of terms or phrases that carry generic meanings. In this example, there are not only classifications of the work itself (“estudio de crítica social”, “estudio de las costumbres”), but also characterizations of the work’s author (“escritor de costumbres”, “pintor de cuadros de circunstancias”) that imply the subgenre of the novel as well as intertextual references pointing to another author and work that served as a generic model for the novel at hand (“Balzac”, “Comedia Humana”). The values that are interpreted from the explicit terms and phrases correspond to a closed set of subgenre labels, which is based on the overall set of empirical

historical subgenre terms found in the bibliography and corpus, as well as on literary-historical knowledge, as documented in chapter 3.2.3 above.

In the case of “Rastaquouère”, there is an introduction to the novel in the first edition of 1890, which contains several hints to the generic frame in which the author sees his work. This introduction is included in the front matter of the TEI corpus file. Some excerpts containing the generic signals evaluated in the TEI header keyword section are given in example 40. The signals are highlighted in curly brackets.

```
<text>
  <front>
    [...]
    <div type="introduction">
      <head>Introducción</head>
      <head>El por qué de este libro y su propósito</head>
      [...]
      <p>¿Qué somos los americanos del sud para una gran parte de los europeos que nos
          juzgan?</p>
      [...]
      <p>Unidos por vínculos de raza y por sentimientos naturales de confraternidad,
          forman nuestras colonias sud-americanas en Europa una familia numerosa y
          compuesta en su mayor parte de gente conspicua y respetable, que se esfuerza,
          con patriótico empeño, en exhibir allí las prendas y cualidades que más
          tiendan a hacer estimables en el extranjero nuestros hábitos, nuestra manera
          de ser y nuestras condiciones de sociabilidad y cultura. Pero sucede a veces
          que dichas personas tropiezan con el inconveniente de tener que luchar en el
          sentido de destruir o borrar el mal efecto producido por las debilidades, los
          candores, las inconveniencias de otros determinados compatriotas, salidos de
          algún rincón cualquiera de esta América lejana, y convertidos, allá en el
          Viejo Mundo, por virtud de la expatriación y por las ventajas que les
          proporcionan la independencia y la libertad con que viven, en personajes de
          valía, en pseudo-notabilidades de su tierra.</p>
      [...]
      <p>¿Se prestará, por ventura, el examen de las costumbres y modos de ser de esas
          gentes a conclusiones tan claras y precisas que alcancen a darnos tema para un
           {estudio de crítica social} tan completo como el que desearíamos ofrecer a
          nuestros lectores?...</p>
      <p>He aquí las preguntas que nos hicimos cuando se nos ocurrió, por vez primera,
          la idea de emprender la composición de este volumen.</p>
      <p>La tarea, sobre ser de suyo ardua, se nos presentaba, por entonces, como
          escabrosa y compromitente. Todo lo que se parezca a alusión personal directa,
          nos decíamos, debe ser rechazado en absoluto por el {escritor de costumbres},
          llamado únicamente a censurar lo que crea censurable, a la manera del {pintor
          de cuadros de circunstancias}, que, al hacer el dibujo de las siluetas que
          juzga conveniente explotar, se cuida, ante todo, de no reproducir
          satíricamente en su lela la fisonomía de algún prójimo viviente determinado.</
          p>
      [...]
      <p>Al intentar llevar a cabo el {estudio de las costumbres} de una mínima fracción
           de ese inmenso todo que se llama la sociedad — conjunto que tan
          magistralmente trató, observándolo en detalle, analizándolo y definiéndolo con
           criterio sin igual el ilustre {Balzac} — hemos pensado que debíamos seguir,
          por nuestra parte, las doctrinas del maestro, y buscar, a nuestra vez, el tema
          , el medio ambiente y los personajes de nuestra fábula dentro del gran
          escenario del mundo, dentro de la misma vida real, aunque manteniéndonos
          forzosamente en una esfera estrecha, que nos obligaba a no salir de los casos
          concretos y de las colectividades sueltas; ya que en el orden social
          particularísimo a que estos apuntes se refieren, la verdadera especie, tal
          como {Balzac} la comprendió en su inmortal {Comedia Humana}, no existe todavía
           entre nosotros.</p>
```

```
      [...]
    </div>
</front>
</text>
```

Example 40. Excerpts from the introduction to the novel "Rastaquouère".

In the introduction, the author presents the motivation, aim, and theme of the novel and refers to several generic models. He starts with the question of how South Americans are seen and judged by Europeans when they travel to European countries. From his point of view, his compatriots are, in general, respectable and sociable persons. However, their reputation suffers from a small group of people who give themselves airs as celebrities without being honorable ("personajes de valía", "pseudo-notabilidades de su tierra"). The title of the novel, "Rastaquouère", refers to this group of newly rich South Americans who resided in Paris at the end of the nineteenth century.[348] The novel aims at studying the customs of this special social group in a detailed critical analysis ("estudio de crítica social tan completo"). Subsequently, the author elaborates on his concept of a novel of customs: in painting his pictures of circumstances ("pintor de cuadros de circunstancias", "dibujo de las siluetas") the writer should avoid direct references to his personal surroundings in order to formulate a general critique and not a particular satire. Furthermore, he bases his novel on the model of Balzac's "La Comédie humaine", contributing one piece to the superordinate goal of creating a total picture of contemporary society, a project not yet realized in his socio-cultural context. So, on the one hand, the author refers to the Hispanic tradition of the *novela de costumbres* and, on the other hand, inscribes his novel in the realist and naturalistic ("estudio de crítica social", "Balzac", "Comedia Humana") movements of French origin.

The example shows that paratextual information can contribute considerably to assessing what subgenres the novels were assigned to historically by their authors, editors, and other contemporaries (in the case of dedications and prefaces written by others). For the whole corpus, it was intended to add at least the front matter of one historical edition to each novel, including the title page and possibly other existing prefatory matters. This was achieved for 231 of the 256 novels.[349] In 42 cases, front matters of several different historical editions were added and evaluated as to their generic signals. The front matters need not correspond to the source editions used to extract the full texts of the corpus because the subgenre assignments to the novels are made on the work level and not on the level of the work expression and manifestation.[350]

The additional information about the subgenre of a novel that is available through its paratexts varies from case to case. At the one extreme are novels that carry their generic program directly with them. In the edition of 1890, the novel "Ensalada de pollos" (1871, MX) by José Tomás de Cuéllar, for example, is preceded by a prologue sketching the design and purpose of the whole

[348] See the etymological information on "rastaquouère" in the lexical portal of the Centre National de Ressources Textuelles et Lexicales: "personne méprisable [...] tanneur, grossiste en peaux, en cuirs [...] Le sens péj. du fr. est prob. dû au fait que beaucoup de Sud-américains à l'élégance tapageuse qui séjournaient à Paris à la fin du xixes. devaient leur fortune récente au commerce des cuirs et peaux" (Centre National de Ressources Textuelles et Lexicales (CNRTL) 2012).

[349] See also chapter 3.3.3.1 on the TEI encoding of front matters.

[350] See chapter 3.2.2 on the question of generic identity and entities of intellectual creations.

series of "novelas de costumbres mexicanas" called "La linterna mágica", of which "Ensalada de pollos" is the first part:

> QUÉ linterna es esa? […]
>
> Este título, que bien puede servirle á una tienda mestiza, ¿es una palabra de programa, altisonante y llamativa para anunciar el parto de los montes, ó encierra algo provechoso para el lector? […]
>
> Yo he copiado á mis personajes á la luz de mi linterna, no en drama fantástico y descomunal, sino en plena comedia humana, en la vida real, sorprendiéndoles en el hogar, en la familia, en el taller, en el campo, en la cárcel, en todas partes […] he tenido especial cuidado de la corrección en los perfiles del vicio y la virtud: de la manera que cuando el lector, á la luz de mi linterna, ría conmigo, y encuentre el ridículo en los vicios, y en las malas costumbres, ó goce con los modelos de la virtud, habré conquistado un nuevo prosélito de la moral y de la justicia.
>
> Esta es la linterna mágica: no trae costumbres de ultramar, ni brevete de invención; todo es mexicano, todo es nuestro, que es lo que nos importa, y dejando á las princesas rusas, á los dandies y á los reyes en Europa, nos entretendremos con la *china*, con el *lépero*, con la *polla*, con la *cómica*, con el *indio*, con el *chinaco*, con el *tendero* y con todo lo de acá. (Cuéllar 1890, vii–x)

The "magic lantern" illuminates the characters and the living spaces that the author wants to represent. He aims to "copy" them from real life, avoiding dramatic, fantastic, and incredible effects. At the same time, he sees it as his task to clearly point out vices and virtues, ridiculing the former and elevating the latter, guiding the reader to internalize morality and justice. An important aspect of his program is to bring to light Mexican and not foreign customs and to focus on characters that are social outsiders or belong to the lower classes of society ("china", "lépero", "polla", "cómica", "indio", "chinaco", "tendero").[351] In consequence, all the novels of the series "La linterna mágica" can be assigned the label "novela de costumbre mexicana".

On the other hand, there are novels that do not exhibit any clear subgeneric signals in their paratexts. In these cases, the opening of the novels was checked for signs pointing to a certain subgenre. The novel "La Mestiza" (1891, MX) by Eligio Ancona, for example, only carries the subtitle "Novela original" and is not preceded by any preface or introduction. Nevertheless, the beginning of the novel is typical for a romantic and sentimental novel, as the excerpts from the first chapter given in example 41 show.

```
<div type="chapter">
  <head>Capítulo I</head>
  <head>La callejuela de San Sebastián</head>
  <p>Eran las siete y media de la mañana de un hermoso día de primavera. La atmósfera
      estaba limpia de vapores y el bellísimo azul de los cielos se ostentaba entonces
      con toda su imponente majestad y hermosura. [...]</p>
```

[351] A "china" is a person with indigenous traits or of a different race; a "lépero" is an indecent, ordinary person; "polla" is probably a colloquial designation for a young woman; a "chinaco" is a pejorative name for a liberal guerrilla fighter; a "tendero" is the owner of a grocery store. See the definitions of these terms in the Spanish Royal Academy's "Diccionario de la lengua española" (Real Academia Española (RAE) 2023a).

```
  <p>No dejaba de participar de estos beneficios una angosta callejuela del barrio de
      San Sebastián, formada por dos hileras de rústicas albarradas, cuya mala
      construcción desaparecía en parte bajo un tapiz de silvestre enredadera. [...]</p>
  <p>Acababa de presentarse en la callejuela un joven de veinte a veinticinco años, de
      una figura bastante recomendable y simpática, que venía tarareando distraídamente
      una canción, como el que trae demasiado ocupado el pensamiento.</p>
  [...]
  <p>Ya partía con los dientes la ciruela más tierna qué había encontrado, cuando
      dirigiendo la vista por la centésima vez al lugar que hemos indicado, vio aparecer
       al extremo izquierdo de la callejuela el blanco vestido de una mestiza.</p>
  [...]
  <p>—¡Dolores! —exclamó el joven con alegría.</p>
  <p>—¡Señor Pablo! —respondió la mestiza.</p>
  <p>Y después de este reconocimiento, porque no creemos que merezca otro nombre, ambos
      guardaron silencio; el joven contemplaba ávidamente a Dolores, y Dolores inmóvil
      junto a él, tenía los ojos fijos en tierra porque sentía clavadas sobre su
      semblante las miradas de fuego de Pablo.</p>
  <p>Y Pablo tenía razón en devorar con sus miradas a la mestiza, porque Dolores era una
       bellísima criatura.</p>
  [...]
</div>
```

Example 41. Excerpts from the first chapter of the novel "La Mestiza".

The first chapter is entitled "The alley of San Sebastián" and begins with a detailed description of the setting, emphasizing the impression that the weather and surroundings, including the vegetation and buildings, have on the observer. Soon the main topic becomes a meeting between the young man Pablo and the *mestiza* Dolores in the said alley. Both are described as pleasant and beautiful ("un joven de veinticinco años, de una figura bastante recomendable y simpática", "Dolores era una bellísima criatura"), and the romantic relationship between them is clearly suggested ("el joven contemplaba ávidamente a Dolores", "Dolores [...] tenía los ojos fijos en tierra porque sentía clavadas sobre su semblante las miradas de fuego de Pablo"). When one reads the first chapter of the novel, a sentimental theme and a romantic style are expected. Here implicit subgenre signals can be located at the beginning of the text, but because of the missing explicit signals, the novel should also be considered as representing the general narrative fiction of its time.

In this section, the assignment of subgenre labels to the novels in the corpus Conha19 was explained, starting from the strategies that were already used for the assignment of subgenre labels to the novels contained in the bibliography Bib-ACMé. There, series titles, work titles, and subtitles were evaluated regarding explicit mentions and implicit signals. This was done to cover the historical characterization of the novels as representatives of particular subgenres. In addition, literary-historical descriptions of the novels' subgenres were assessed. For the corpus, further historical textual elements were analyzed for subgenre signals, including paratexts beyond titles and openings of the texts. As a result, a bundle of explicit, implicit, historical, and critical subgenre assignments to the novels is available. It is organized into several levels of an empirical, discursive model of subgenre terms which serves as the basis for analyzing the subgenres in the corpus.

3.3.5 Derivative Formats and Publication

In the previous sections, the text corpus has been presented in terms of the sources and selection of novels, the treatment of the full texts, the encoding of metadata about the novels and the texts themselves in TEI, as well as the assignment of subgenre labels to the novels. In this final chapter about the corpus, two further aspects are covered: the creation of other corpus formats derived from the TEI and the organization and strategy for the publication of the corpus. Several derivative formats were created to prepare the analysis of the corpus with different tools. One of them, a tokenized version of a subset of the corpus with direct speech annotation, was already presented in chapter 3.3.3.2.8 about the encoding of direct speech and thought. Two other basic derivative formats are a plain text version of the corpus files and a linguistically annotated version. Plain text files are required as an input format for many natural language processing and text analysis tools, and a prepared linguistically annotated version of the corpus allows the use of lexical and grammatical categories in further analyses. More derivative formats can be created in an ad-hoc manner, but it was decided to prepare these two fundamental corpus versions so that they are ready for use in a variety of contexts.

The corpus created for this dissertation is published for several reasons. Most of the texts are in the public domain,[352] which makes it possible to redistribute this part of the corpus freely. Moreover, considerable preparatory work was invested to create this corpus of novels for subgenre analysis, and it is desirable to share it with the research community and general public for reuse in other contexts, not least because also this work builds on previous efforts made by others to edit, digitize and curate the works in question. As the corpus covers a broad time period in the nineteenth and up to the beginning of the twentieth century, works from three different Spanish-American countries, written by many different authors and attributable to a whole range of subgenres of the novel, it can be hoped that there will be other scenarios to use it. Therefore, the TEI master files, schemas, and main derivative formats of the corpus were prepared for research data publication. This subsection serves to first document the creation of the two main derivative corpus formats, followed by an overview of the corpus publication.

The plain text format is derived from the TEI files with an XSLT script designed to process a single file. It can be applied to the whole corpus using the Saxon XSLT processor from the command line (Saxonica n.d.).[353] For the plain text version of a corpus file, the TEI header, front and back parts are ignored. Also headings of book parts and chapters are skipped. In case of dramatic speech inside of the novels, the speaker names are omitted, as well. The text of paragraphs is copied and separated by blank lines. Groups of verse lines are also separated by blank lines, but individual verses are only copied with a newline. A snippet of the plain text version of the first novel in the corpus, "El guajiro" (1842, CU) by Cirilo Villaverde, is shown in example 42.

```
[...]
Un ronquido profundo, como el estertor de un agonizante, fue la única respuesta que
    obtuvo la enamorada muchacha. Continuaba en sacudir y pellizcar a la negra; pero la
    misma voz volvió a dejarse oír con esta otra décima:
```

[352] See chapter 3.3.3.1 above for details about the novels' copyright status.

[353] The XSLT file is available at https://github.com/cligs/scripts-nh/blob/master/corpus/derivative_formats/get-plaintext.xsl. Accessed August 23, 2020.

```
Dices que no hay ocasión
para que hablemos aquí,
donde me temes a mí
y teme tu corazón.—

¡Mentira, mentira! —dijo precipitadamente, sin ser dueña a contenerse, y como si él
    pudiera —oirla, yo no te temo a ti, Tatao mío, ni temo por mí, sino a mi padre, que
    es duro y tiene el sueño más ligero que un pájaro. ¡Ay, si te oyese! Si yo pudiera.

El canto la obligó a interrumpirse.
[...]
```

Example 42. Excerpt from the plain text version of the novel "El guajiro".

For the linguistically annotated version, the tool FreeLing was used. It is a suite of open-source language analysis tools based on C++ and was chosen because it includes a comprehensive morphological dictionary for Spanish, containing over 555,000 forms and over 76,000 lemma-PoS combinations (Padró and Stanislovsky 2012; Padró n.d.a). FreeLing was used in version 4.0. FreeLing has a front-end called "analyzer", which is its main program and was used in client/server mode to annotate the corpus files (Padró n.d.e.).[354] Each call of this program serves to process one file. A sample command line call to process the first chapter of the first novel in the corpus is given in example 43.

```
analyze -f es.cfg --server on --port 50005 --workers 1 --outlv tagged --sense ukb --nec
    --output xml & analyzer_client 50005 < /home/ulrike/Git/conha19/txt/nh0001_d1.txt >
    /home/ulrike/Git/conha19/tei_annotated/nh0001.xml
```

Example 43. Command line call of the FreeLing analyzer program.

The first line of the call serves to set the default configuration file for Spanish (`-f es.cfg`), to establish the client/server mode (`--server on --port 50005 --workers 1`), and to set the options for the linguistic annotation. Here, part-of-speech annotation (`--outlv tagged`), sense annotation (`--sense ukb`), and named entity classification (`--nec`) are performed. Finally, the output format is set to a FreeLing-specific XML format (`--output xml`). The second line of the call specifies the input file to be processed and the path to the output file. An excerpt of the annotation result in the FreeLing format is shown in example 44.

```
<sentence id="1">
  <token id="t1.1" form="Más_allá_de" lemma="más_allá_de" tag="SP" ctag="SP" pos="
      adposition" type="preposition"/>
  <token id="t1.2" form="el" lemma="el" tag="DA0MS0" ctag="DA" pos="determiner" type="
      article" gen="masculine" num="singular"/>
  <token id="t1.3" form="pueblo" lemma="pueblo" tag="NCMS000" ctag="NC" pos="noun" type=
      "common" gen="masculine" num="singular" wn="07942152-n"/>
  <token id="t1.4" form="de" lemma="de" tag="SP" ctag="SP" pos="adposition" type="
      preposition"/>
  <token id="t1.5" form="San_Diego_de_Núñez" lemma="san_diego_de_núñez" tag="NP00G00"
      ctag="NP" pos="noun" type="proper" neclass="location" nec="LOC"/>
  <token id="t1.6" form="," lemma="," tag="Fc" ctag="Fc" pos="punctuation" type="comma"/
      >
  <token id="t1.7" form="en" lemma="en" tag="SP" ctag="SP" pos="adposition" type="
      preposition"/>
```

[354] The client/server mode is advantageous if many small files are processed. It was used here because the novels were treated per paragraph, resulting in 531,006 plain text snippets to be analyzed for the whole corpus.

```
  <token id="t1.8" form="la" lemma="el" tag="DA0FS0" ctag="DA" pos="determiner" type="
      article" gen="feminine" num="singular"/>
  <token id="t1.9" form="isla" lemma="isla" tag="NCFS000" ctag="NC" pos="noun" type="
      common" gen="feminine" num="singular" wn="09316454-n"/>
  <token id="t1.10" form="de" lemma="de" tag="SP" ctag="SP" pos="adposition" type="
      preposition"/>
  <token id="t1.11" form="Cuba" lemma="cuba" tag="NP00G00" ctag="NP" pos="noun" type="
      proper" neclass="location" nec="LOC" wn="02795169-n"/>
  <token id="t1.12" form="," lemma="," tag="Fc" ctag="Fc" pos="punctuation" type="comma"
      />
  <token id="t1.13" form="camino" lemma="camino" tag="NCMS000" ctag="NC" pos="noun" type
      ="common" gen="masculine" num="singular" wn="00172710-n"/>
  <token id="t1.14" form="de" lemma="de" tag="SP" ctag="SP" pos="adposition" type="
      preposition"/>
  <token id="t1.15" form="Bahía_Honda" lemma="bahía_honda" tag="NP00G00" ctag="NP" pos="
      noun" type="proper" neclass="location" nec="LOC"/>
  [...]
</sentence>
```

Example 44. Annotation result in the FreeLing XML format.

Here, the first sentence of the novel's first chapter is annotated, starting with the phrases "Más allá del pueblo de San Diego de Núñez, en la isla de Cuba, camino de Bahía Honda [...]". FreeLing marks sentence and token boundaries and attaches the linguistic annotations to the tokens. The tagset for the part-of-speech annotation is based on the EAGLES Recommendations (e.g., "NC" for "common noun" and "NCMNS000" for "common masculine noun in nominative singular") (Expert Advisory Group on Language Engineering Standards (EAGLES) 1996; Padró n.d.d). The sense annotation is based on WordNet and results in sense identifiers (e.g., "00172710-n" for the noun "camino") (Fellbaum 1998; Miller 1995; Padró n.d.b). The named entity classification differentiates between persons, geographical locations, organizations, and others (e.g., "LOC" for "San Diego de Núñez"). A very useful feature of FreeLing is that it is able to recognize words consisting of several tokens, such as the preposition "más allá de" and the place names "San Diego de Núñez" and "Bahía Honda" in the example (Padró n.d.c).

To annotate the whole corpus, the functionality of the FreeLing analyzer program was integrated into an annotation workflow, aiming to produce derivative TEI files keeping the TEI header and basic text structure (parts, chapters, and paragraph-like structures[355]) of the TEI master files, but replacing the contents with the linguistically annotated text. That way, the structures that were marked up in the texts are still available for analysis in conjunction with linguistic information. On the other hand, if the linguistic annotation had been applied to the entire plain text files of the novels, the structural information would have been lost in the process. Integrating the annotation output directly into the TEI structure required adapting the FreeLing XML output a bit in order to conform to the TEI standard. Furthermore, the FreeLing sense annotation output was enhanced by adding WordNet lexnames to the synset identifiers that were

[355] In the corpus, paragraph-like structures include paragraphs, verse lines, and headings other than part and chapter headings.

produced by FreeLing itself.[356] The annotation workflow was written in Python, including XPath and XSLT calls, and comprises the following steps:[357]

- preparation: store the TEI structure for each novel, identify paragraph-like structures, extract the plain text of each paragraph-like unit
- annotation: calls to the FreeLing analyzer and WordNet for each paragraph-like unit
- post-processing: reintegrate the annotated paragraph-like units into the prepared TEI structure, adapt the FreeLing XML output to the TEI standard

The result of this process is shown in example 45.

```
<div type="chapter">
  <p xml:id="nh0001_p1">
    <s>
      <w cligs:form="Más_allá_de" lemma="más_allá_de" cligs:tag="SP" cligs:ctag="SP" pos
          ="adposition" type="preposition" cligs:wnsyn="xxx" cligs:wnlex="xxx">
          Más_allá_de</w>
      <w cligs:form="el" lemma="el" cligs:tag="DA0MS0" cligs:ctag="DA" pos="determiner"
          type="article" cligs:gen="masculine" cligs:num="singular" cligs:wnsyn="xxx"
          cligs:wnlex="xxx">el</w>
      <w cligs:form="pueblo" lemma="pueblo" cligs:tag="NCMS000" cligs:ctag="NC" pos="
          noun" type="common" cligs:gen="masculine" cligs:num="singular" cligs:wnsyn="
          07942152-n" cligs:wnlex="noun.group">pueblo</w>
      <w cligs:form="de" lemma="de" cligs:tag="SP" cligs:ctag="SP" pos="adposition" type
          ="preposition" cligs:wnsyn="xxx" cligs:wnlex="xxx">de</w>
      <w cligs:form="San_Diego_de_Núñez" lemma="san_diego_de_núñez" cligs:tag="NP00G00"
          cligs:ctag="NP" pos="noun" type="proper" cligs:neclass="location" cligs:nec="
          LOC" cligs:wnsyn="xxx" cligs:wnlex="xxx">San_Diego_de_Núñez</w>
      [...]
    </s>
  </p>
</div>
```

Example 45. Annotation result in the TEI format.

The example shows that the TEI chapter and paragraph structures are preserved. Inside paragraphs, `<s>` elements were produced, which in turn contain the individual `<w>` elements carrying the linguistic information in different attributes. Of the attributes produced by FreeLing, `@lemma`, `@pos`, and `@type` conform to the TEI standard, but the others (e.g., `@tag`, `@gen`, or `@nec`) are not available in TEI and were therefore attributed to a custom CLiGS namespace, to which also the WordNet-related attributes `@wnsyn` and `@wnlex` were added.

As a result, the linguistically annotated derivative format of the corpus can be directly used for analytic purposes, for example, by querying them to calculate the frequencies of specific word

[356] WordNet lexnames are lexicographer files into which the synsets are organized and consist of syntactic categories (nouns, verbs, adjectives, adverbs) and logical groupings (e.g., nouns denoting animals versus body parts). They add more lexical information to the synsets (Princeton University 2023).

[357] The workflow for the linguistic annotation was designed as part of the work in the CLiGS project. Various members of the group were involved in the programming of its different parts, as indicated in the Python files. The workflow consists of three Python files: "workflow_teifw.py" (for settings and to start the process) and the two modules "prepare_tei.py" (for pre-processing and post-processing of the FreeLing annotation) and "annotate_fw.py" (for the FreeLing annotation itself and WordNet calls). The scripts are available at https://github.com/cligs/scripts-nh/tree/master/corpus/derivative_formats. Previous versions can be viewed at the CLiGS group's toolbox repository: https://github.com/cligs/toolbox/tree/master/annotate. Accessed August 24, 2020.

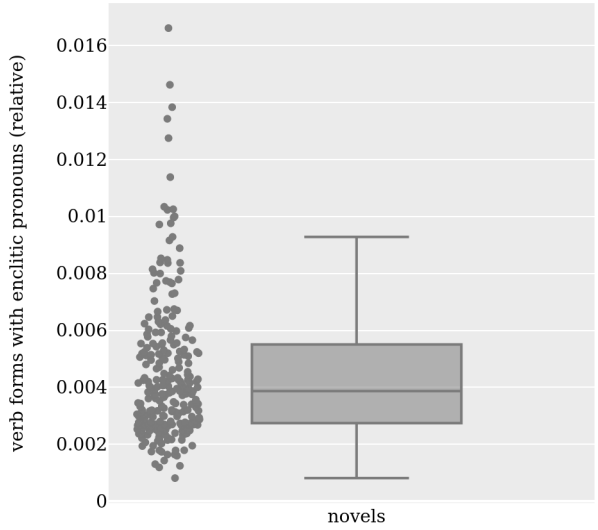


Figure 33. Verb forms with enclitic pronouns in the novels of the corpus.

categories, lemmas, etc. in the novels. It can also be used to produce other formats of the texts as starting points for further analyses, such as, for example, a text version consisting only of lemma nouns which would be suitable for topic modeling.

The quality of the part-of-speech (POS) annotation was checked in one aspect that had been noted during the text treatment and the spell-checking process as a specific characteristic of Spanish historical texts: the frequency of verb forms with enclitic pronouns that were not recognized by the spell-checker, such as, for example, "habíase" instead of "se había" or "diósela" instead of "se la dio".[358] As a first step, the list of regular expressions that was prepared to capture such word forms as exceptions in the spell-checking process was used to detect how many of these forms occur in the texts of the corpus.[359] The results are summarized in figure 33.

In the figure, the counts of the verb forms with enclitic pronouns are given relative to the novels' text length in the number of tokens. The median is at 0.3 %. The novels with the maximum relative amount have about 1 % of verb forms with enclitic pronouns. There is no clear separation of the values into two groups which would suggest that the texts of novels with historical source editions are completely different in this aspect than the ones of novels based on modern editions. It has to be reminded, though, that not all of the verb forms with enclitic pronouns are out of use

[358] See chapter 3.3.2 above for details.

[359] The list of regular expressions that was used is available at https://github.com/cligs/data-nh/blob/main/corpus/derivative-formats/verb-form-patterns-es-detail.txt. A list of exception words was prepared to cover cases where words that are not verb forms with enclitic pronouns are matched by the regular expressions. The exception list can be viewed at https://github.com/cligs/data-nh/blob/main/corpus/derivative-formats/verbs-enclitics-exceptions.txt. The exception list is not exhaustive but covers the most frequent cases. The Python script used to evaluate the corpus with regards to verb forms with enclitic pronouns is published at https://github.com/cligs/scripts-nh/blob/master/corpus/derivative_formats/verbs_enclitics.py. The resulting counts are accessible at https://github.com/cligs/data-nh/blob/main/corpus/derivative-formats/verbs_enclitics_in_files.csv. Accessed November 20, 2020.

today. They are still used with infinitive or imperative forms, for example (“hablarnos”, “dáselo”). As the infinitive forms can be matched unequivocally by single regular expressions, they were ignored for this analysis, but the imperative forms are not that easily separated and were kept. In a second step, it was checked how many verb forms with enclitic pronouns persist as entire tokens in the FreeLing output. With the standard settings, FreeLing separates the enclitic pronouns from the verbs and returns two tokens, as shown in example 46, which shows the annotated phrase “comenzó a hablarnos”.

```
<w cligs:form="comenzó" lemma="comenzar" cligs:tag="VMIS3S0" cligs:ctag="VMI" pos="verb"
     type="main" cligs:mood="indicative" cligs:tense="past" cligs:person="3" cligs:num="
    singular" cligs:wnsyn="00348746-v" cligs:wnlex="verb.change">comenzó</w>
<w cligs:form="a" lemma="a" cligs:tag="SP" cligs:ctag="SP" pos="adposition" type="
    preposition" cligs:wnsyn="xxx" cligs:wnlex="xxx">a</w>
<w cligs:form="hablar" lemma="hablar" cligs:tag="VMN0000" cligs:ctag="VMN" pos="verb"
    type="main" cligs:mood="infinitive" cligs:wnsyn="xxx" cligs:wnlex="xxx">hablar</w>
<w cligs:form="nos" lemma="nos" cligs:tag="PP1CP00" cligs:ctag="PP" pos="pronoun" type="
    personal" cligs:person="1" cligs:gen="common" cligs:num="plural" cligs:wnsyn="xxx"
    cligs:wnlex="xxx">nos</w>
```

Example 46. FreeLing output for verb forms with enclitic pronouns (in CLiGS TEI-format).

However, verb forms that are not recognized because they are no longer in use are not separated and tend to be misclassified, as the following example 47 of the phrase “diósela a Bruno” shows.

```
<w cligs:form="diósela" lemma="diósela" cligs:tag="NCFS000" cligs:ctag="NC" pos="noun"
    type="common" cligs:gen="feminine" cligs:num="singular" cligs:wnsyn="xxx"
    cligs:wnlex="xxx">diósela</w>
<w cligs:form="a" lemma="a" cligs:tag="SP" cligs:ctag="SP" pos="adposition" type="
    preposition" cligs:wnsyn="xxx" cligs:wnlex="xxx">a</w>
<w cligs:form="Bruno" lemma="bruno" cligs:tag="NP00G00" cligs:ctag="NP" pos="noun" type=
    "proper" cligs:neclass="location" cligs:nec="LOC" cligs:wnsyn="xxx" cligs:wnlex="xxx
    ">Bruno</w>
```

Example 47. FreeLing output for verb forms with enclitic pronouns (in CLiGS TEI-format).

In the example, the verb form “dió” and the two pronouns “se” and “la” attached to it are interpreted as a common noun. Because of the way in which verb forms with enclitic pronouns are usually treated by FreeLing (separation of verb and pronouns), it was concluded that tokens in the FreeLing output that still match the regular expressions for verb forms with enclitic pronouns are misclassifications. These forms were collected, and it was analyzed to which part of speech they were assigned, as visualized in figure 34.[360]

In total, 24,131 forms were found, compared to 80,694 forms that were found in the non-annotated plain text files of the corpus, which means that the morphological structure of 70 % of the forms was probably analyzed correctly by FreeLing. Of the forms that remained, 26 % were analyzed as verbs and the others as other parts of speech. 56 % were marked as nouns, more than half of them as proper nouns, and the other part as common nouns. 17 % were analyzed as adjectives, only 1 % as adverbs, and only one instance each as a number and an interjection. If almost one-third of the remaining forms were classified as verbs, why were they not separated into verbs and pronouns morphologically? A look into the verb matches shows that more than

[360] A summary of the counts is available at https://github.com/cligs/data-nh/blob/main/corpus/derivative-formats/verbs-enclitics-freeling.csv. Accessed 20 November 2020.

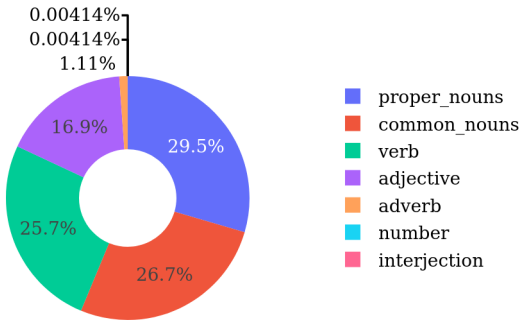


Figure 34. FreeLing POS of verb forms with enclitic pronouns.

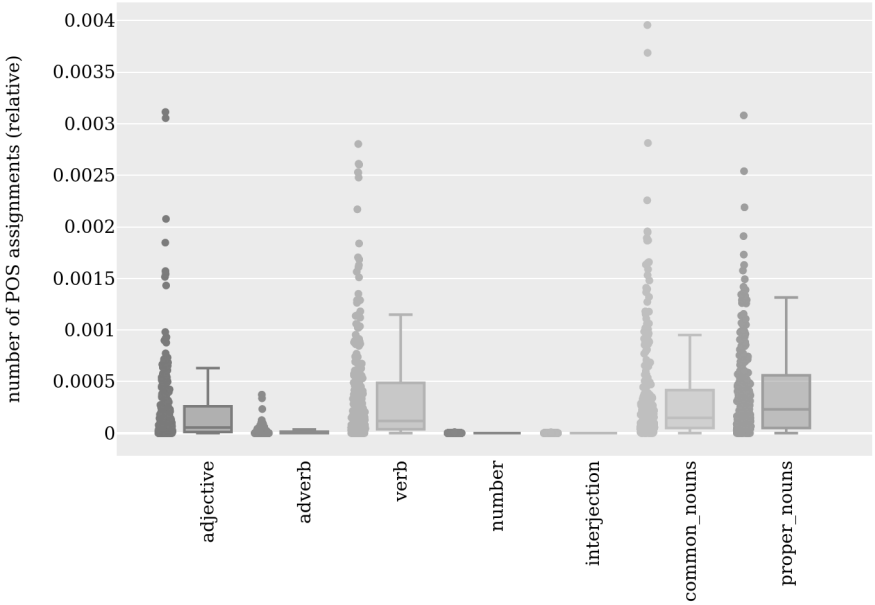


Figure 35. FreeLing POS of verb forms with enclitic pronouns in the texts of the corpus.

half of them were recognized as subjunctive forms. In Spanish, imperfect subjunctive forms can have the same structure as verb forms with the enclitic pronoun “se”. For example, “hablase” can be used in a context like “no quería que hablase” (“I did not want him to speak”, verb form in imperfect subjunctive) or “Hablase de intrigas” (“There is talk of intrigues”, (historical) verb form in present tense with the enclitic passive pronoun “se”). In the other cases, the tense of verb forms was not recognized correctly. For example, preterit, imperfect, and conditional forms with enclitic pronouns were mistaken as indicative present tense forms (“salióle”, “parecíale”, “faltábale”, “bastaríame”). For the misclassified verb forms with enclitic pronouns, it was also analyzed how they are distributed in the novels of the corpus relative to text length in tokens, as represented in figure 35.

Here it becomes clear that it is not the number of verb forms with enclitic pronouns, in general, that is very unequally distributed in the novels, but the number of misclassified forms of this type, for which it can be assumed that they are no longer in use. As can be seen, the boxes in the plot have a much smaller variance in the first and second quartiles than in the third and fourth ones. This means that there are many novels with zero or very low misclassifications and another half with higher, varying proportions of them. Such an imbalance in the quality of part-of-speech assignments can potentially have distorting effects on the results of stylistic analyses. For example, verb forms and enclitic pronouns that are not separated are not counted as individual tokens in a bag-of-words approach. Instead, they end up as new items in the vocabulary. The influence of the misclassifications also depends on which kind of word forms are used in an analysis. If one wants to analyze named entities, the verbs with enclitic pronouns classified as proper nouns will permeate the set of entities found. Alternatively, if only nouns are selected, as is often done for topic modeling, again, the verb forms with enclitic pronouns that were classified as common nouns will affect the results.

As a provisional solution, the set of regular expressions was used to split the misclassified forms into verbs and pronouns and to correct the main part-of-speech assignment.[361] In the corrected form, the above-mentioned phrase "diósela a Bruno" looks as shown in example 48.

```
<w cligs:form="dió" lemma="dió" pos="verb">dió</w>
<w cligs:form="se" lemma="se" pos="pronoun">se</w>
<w cligs:form="la" lemma="la" pos="pronoun">la</w>
<w cligs:form="a" lemma="a" cligs:tag="SP" cligs:ctag="SP" pos="adposition" type="
    preposition" cligs:wnsyn="xxx" cligs:wnlex="xxx">a</w>
<w cligs:form="Bruno" lemma="bruno" cligs:tag="NP00G00" cligs:ctag="NP" pos="noun" type=
    "proper" cligs:neclass="location" cligs:nec="LOC" cligs:wnsyn="xxx" cligs:wnlex="xxx
    ">Bruno</w>
```

Example 48. Phrase with corrected morphological analysis and POS assignment.

As the regular expressions cannot match the verb forms with enclitic pronouns unequivocally in all cases, there can be false positives in this approach. To prevent this as much as possible, a list with exception words was created. To identify the exception words, all the matches of supposedly misclassified verb forms with enclitic pronouns that occurred five times or more often were checked and false positives were added to the exception list.[362] Linguistic knowledge is indispensable to find a sustainable and more precise solution. A lexicon of verb forms and rules for the recognition of historical enclitic constructions could be used to improve the linguistic annotation in the first place instead of correcting the output afterward. Nevertheless, the regular expression-based solution works as a first approach to improve the linguistic annotation as a basis for further text analysis.

The text corpus Conha19 ("Corpus de novelas hispanoamericanas del siglo XIX") is published in a GitHub repository at https://github.com/cligs/conha19. GitHub is a commercially driven,

[361] For unambiguous cases, also the accents were corrected with the help of substitution rules. For example, "dábanle" is transformed to "daban" and "le", and "hízose" is transformed to "hizo" and "se" (the accent is not needed anymore and would be incorrect on the freestanding verb form). The list of accent replacements in verb form endings is available at https://github.com/cligs/data-nh/blob/main/corpus/derivative-formats/verb-form-endings-accents.txt. Accessed November 20, 2020.

[362] See footnote 359 above.

Directory / file name	Description of contents
metadata.csv	selected, basic corpus metadata in CSV format[363]
tei	TEI master files
schema	Taxonomy for metadata keywords, Schematron file for keyword control, validation log files
bib	Bibliography file (in TEI) holding full bibliographic references of literary historical works cited in the corpus files
txt	plain text version
annotated	linguistically annotated version (in TEI)
annotated_corr	linguistically annotated version (in TEI) with corrected POS annotation for verb forms with enclitic pronouns
tei_ns	subset of 92 files without direct speech mark-up (in TEI)
tei_ds	subset of 92 files with direct speech mark-up only based on regular expressions (in TEI)
tei_tokenized_ds	subset of 92 files as tokenized text with two stand-off direct speech annotations (DS_gold vs. DS_reg; in TEI)
spellcheck	lists with exception words and results of the spell check in CSV format, for the whole corpus and per novel

Table 20. Elements of the corpus published on GitHub.

web-based open platform for source code management and collaborative version control. Because it is a working environment, the corpus can be continued to be curated in the repository and be published in subsequent stable and referenceable releases. The collaborative features of GitHub facilitate other researchers to reuse the corpus by cloning or forking the repository. Comments and suggestions on the corpus can be created as issues. Because this environment alone is not suitable for long-term archiving, the stable corpus releases are additionally stored on Zenodo.org, an archiving service for researchers that is managed by the European OpenAire program and operated by CERN (Nielsen 2013).[364] Publications on Zenodo.org receive Digital Object Identifiers (DOI) so that the corpus releases are identifiable and reachable in the long term. The different components of the corpus publication are listed in table 20.

Although the text corpus has been designed specifically for the study of subgenres of nineteenth-century Argentine, Cuban, and Mexican novels, its open publication aims to encourage the reuse of the data in other contexts. As the creation of richly annotated and curated collections of historical, literary texts is labor-intensive, it should be a goal to share the results of this work as far as the legal conditions allow. The corpus at hand could, for example, also be useful for studies concentrating on one of the countries or on individual authors. It could also be integrated into

[363] This metadata was generated with the script "metadata.xsl" available at https://github.com/cligs/scripts-nh/blob/master/corpus/metadata.xsl. Accessed September 24, 2020.

[364] The decision to rely on the two infrastructures of GitHub and Zenodo.org for publishing the corpus is the result of the work in the junior research group CLiGS (Schöch et al. 2019, paras 36–38).

more extensive corpora comprising different genres or a wider chronological range. In addition, the TEI files could serve as starting points for creating digital critical editions of individual novels.

From the point of view of quantitative digital literary studies, with its 256 novels, Conha19 can be considered a corpus of medium size, lying somewhere between small-scale text collections for stylometric studies and the "million volumes" analyzed by Underwood (Underwood 2015b, 2–3). The medium size of the corpus made it possible to add detailed metadata and structural markup to the texts. On the other hand, the size of the corpus made it necessary to rely on an automatic orthography check to assess the quality of the full texts. Moreover, in this medium-sized corpus, not only canonical works are included but also lesser-known ones. Furthermore, the corpus is new in this composition and was not retrieved from one source but from a whole range of different source institutions. It was also built from different types of source editions (historical and modern, scholarly as well as general ones). Finally, also the range of subgenres included in the corpus is broad, and the number of different authors is considerably high. In the following section, overviews of the corpus' contents are given from various perspectives. They are compared to the works included in the digital bibliography Bib-ACMé to estimate how the distribution of novels in the corpus relates to the overall production of novels of the time in the three countries in question.

4 Analysis

4.1 Metadata Analysis

Before the novels are analyzed by subgenre textually, the data contained in the digital bibliography and the text corpus are analyzed on a metadata level in this chapter. One goal of this analysis is to provide a general overview of the contents in both databases: how many novels are there in the reference bibliography (of which subgenre, written by which authors, published in which of the three countries and when)? How many novels are there in the corpus, and is the corpus similarly structured in quantitative terms, or are there differences between both resources? Furthermore, the quantities of novels by subgenre in the corpus are assessed to find out which groups are big enough to carry out a quantitative text analysis. A choice is then made for two discursive levels (thematic subgenres and literary currents) and a specific set of subgenre labels (the primary labels on these two levels) to analyze the novels further in the text analysis part. When the numbers of novels in the bibliography and corpus are compared, and differences are observed, these are mainly described in qualitative terms, which means that the numbers are interpreted and set in relation to each other. However, statistical tests for significance are only done in the case of the novels' text length. On the level of the metadata categories, most of the groups are quite small, so most differences are not expected to be significant in a statistical sense. Nonetheless, they show how specific subsets of the bibliography and corpus proportionally vary.

4.1.1 On Representativeness

When analyzing the subgenres in a corpus of novels, an important question is to what extent the results can be interpreted as statements about the subgenres in question and not only about the selected novels in the corpus, that is, how far they are generalizable. The search for an answer to this question involves determining the representativeness of the corpus. Assuming that the corpus does not consist of the entire literary production, to which degree does it represent it? How to capture "the entire literary production"?

For linguistic corpora, questions of representativeness in corpus design have been addressed by Douglas Biber, in particular (Biber 1993a). As he formulates, "[r]epresentativeness refers to the extent to which a sample includes the full range of variability in a population" (Biber 1993a, 243). Two important terms are introduced here: the *population* as a whole, such as the entire production of a spoken or written language or the entire literary production, and the *sample* as a selected section of the population. Biber states that the assessment of the representativeness of a sample depends first on a prior definition of the population and second on the sampling technique used to make selections from it. He mentions two important aspects for the definition of the population: "(1) the boundaries of the population—what texts are included and excluded from the population; (2) hierarchical organization within the population—what text categories are included in the population, and what are their definitions" (Biber 1993a, 243).

The population of this corpus has been defined in chapter 3.1 on selection criteria, making use specifically of the first aspect. The "Boundaries of the Novel" (3.1.1) specified what kind of texts are included (novels) and how this generic kind is defined and delimited in the current

context. Furthermore, the population was situated cultural-geographically and chronologically by discussing the "Borders of Argentina, Cuba, and Mexico" (3.1.2) and the "Limits of the Nineteenth Century" (3.1.3). On the other hand, regarding the second aspect, no restrictions were made for the internal organization of the population in terms of types of subgenres. No specific subgenres were set or excluded. However, the population is internally organized into works from the three countries.

The definition of a population is, first of all, theoretical work because it does not mean it would be possible to have complete access to it. An operational definition of the population is needed, which is called "sampling frame" by Biber: "an itemized listing of population members from which a representative sample can be chosen" (Biber 1993a, 244). The *sampling frame* for the corpus Conha19 is the bibliography Bib-ACMé (see chapter 3.2), to which the population's selection criteria were applied and the sources of which were presented in chapter 3.2.1.

Biber describes several sampling strategies. Probabilistic sampling relies on random selection and can, for example, be realized as a simple random sampling, where all items have the same chance to be selected. Another variant of probabilistic sampling is stratified sampling, which makes use of subgroups in the population and applies random sampling to each subgroup in a second step (Biber 1993a, 244). For the creation of the corpus Conha19, no formal random sampling was applied, neither general nor stratified, because the selection of texts from the bibliography was restricted to the texts actually available in digital format or a format suitable for digitization. So in this case the availability of the sources had a strong influence on the resulting sample. Nevertheless, in an informal procedure, the texts were selected in a way to ensure a balance of countries, authors, and major subgenres as far as possible.

One way to evaluate representativeness is by looking at the sample size. Which overall proportion of the population is contained in the sample? However, the aspect of how much of the population's variability is included is considered even more important by Biber. In the context of linguistic corpora, he finds that

> variability can be considered from situational and from linguistic perspectives, and both of these are important in determining representativeness. Thus a corpus design can be evaluated for the extent to which it includes: (1) the range of text types in a language, and (2) the range of linguistic distributions in a language. (Biber 1993a, 243)

It is very clear that this view on variability is specifically linguistic: text types are bound to communicative situations, and the relevant text-internal features are linguistic distributions. For a literary corpus, corresponding requirements could be formulated in the following way:

1. the range of genres in a given literature
2. the range of textual distributions in a given literature
3. the set of authors in a given literature (because the authorship of texts plays a much more important role in literary corpora and influences the style of the texts considerably)
4. periods in a given literature (because there are eras for which a distinctive literary style can be identified)[365]

[365] In digital literary stylistics, the separation of different stylistic signals that correspond to literary categories (authorship, genre, epoch, etc.) on the textual level has repeatedly been a concern (see, for instance, Burrows 2002;

Of course, linguistic periods have similar relevance for the construction of linguistic corpora. It can be assumed that Biber set a focus on text types for questions of variability inside a corpus because the linguistic period is only at issue in historical and especially diachronic corpora.

If genres are understood as external attributions to the texts in question, then the first, third and fourth factors are external, while the second one depends on the internal characteristics of the texts. However, using the second criterion to determine the internal variability of a literary corpus is not straightforward. More research is available on the range of linguistic distributions in languages than on textual distributions in literature.[366] First of all, it would be necessary to determine what kind of textual distributions are relevant. Distributions of linguistic features in literary texts? Or distributions of specifically literary features? If the latter, which kind of features would these be? If the "literature" in this case is "the novel", it would be necessary to have general knowledge about textual distributions in novels. To give examples, this could be knowledge about the typical range in the amount of direct speech in novels or knowledge about the typical distribution of topics in novels.[367] As things now stand, though, knowledge about such textual distributions in literary texts is rather still the aim of digital literary studies than a fund of basic knowledge to which one could refer. Therefore, the second point is not used here to evaluate the representativeness of the corpus. Instead, the sampling frame Bib-ACMé and the

---

Schöch 2013). For an attempt to "neutralize" the authorial signal to be able to analyze genre style, see Calvo Tello et al. (2017). On the other hand, differences between linguistic and literary styles are usually not emphasized in linguistic textbooks on the topic, where a general coverage of stylistic phenomena is pursued. This is also due to revised definitions of style. As Sowinski notes: "In älteren Arbeiten zur Stilistik wird Stil ausschließlich literarischen Werken zugesprochen […]. Erst in neuerer Zeit ist diese Einschränkung gefallen: Stil wird heute allen Texten zugesprochen, wenn auch in unterschiedlicher Art und Weise. […] Auch die Gebrauchstexte pragmatischer Natur müssen nicht arm an Stilelementen sein, wenn hier auch oft andere Stilzüge (z.B. Ökonomie, Präzision) den Verzicht auf bestimmte Stilelemente, z.B. des affektiven oder bildhaften Bereichs bedingen können" (Sowinski 1999, 73). However, Sowinski does not discuss how the different elements of style relate to authorship or period. In their "revisited" definition of style, the digital humanists Herrmann, Schöch, and van Dalen-Oskam also strive to offer a generally applicable notion: "In our approach, style is not something unique to literary works; rather, every text has a certain kind of style." They also have a general view on the issue of the relationship between style and other categories, but with a focus on literary ones: "In our definition, style can be associated with categories such as genre, epoch, author, and many more. In many cases, correlations between specific style markers or groups of style markers with these categories may be observed. What is more, even in the absence of conscious intentions, causal relationships may be hypothesized: genre can cause style (e.g., by means of conventions: form and themes), authors can cause style (e.g., by means of idiosyncrasies), theme and topic can cause style. The interpretability of style relative to categories such as authorship, literary genre, or literary period, is hence paramount. This means that any stylistic phenomenon can ultimately be considered the trace of or the index towards such categories […]" (Herrmann, Schöch, and van Dalen-Oskam 2015, 46). Nevertheless, in empirical digital literary studies on genre, authorship is usually the factor that interferes most. The importance of considering literary periods for corpus building in literary studies is stressed by Gemeinböck (2016, 37). Therefore, the position taken here is that although general principles for corpus building and the assessment of representativeness are the same for linguistic and literary corpora (such as determining a population and a sampling frame and following certain sampling strategies), there are differences in the kind and in the relevance of categories to assess internal variability.

[366] For instance, in Biber's text on representativeness itself, studies of distributions of linguistic features are presented (Biber 1993a, 248–255).

[367] Some of the first studies in this direction are, for example, Jannidis et al. (2018), Schöch, Schlör, et al. (2016), and Jockers (2013, 118–153). The first and second are concerned with the detection and distribution of direct speech in German and French nineteenth-century novels, respectively, and the last with the distribution of topics in nineteenth-century English language novels.

sample Conha19 are compared on several levels that are derived from the metadata encoded for both: the authors (in chapter 4.1.2), works (4.1.3), editions (4.1.4), and subgenres (4.1.5) covered. Both the sample size and its relative variability in relation to the sampling frame are assessed.

In addition, specific overviews are given for the works in Conha19 (in 4.1.3.2) for features available for the corpus but not for the whole bibliography. These are specific metadata such as the novels' narrative perspective or status as high- or low-prestige literature, but also characteristics derived from the full texts themselves, such as their length. These overviews thus offer a more descriptive perspective on the corpus. They are nonetheless important because they highlight specific characteristics of the corpus that influence later analyses. This allows for assessing what is typical for the whole corpus and what is a specific result in a particular analysis. Nevertheless, when interpreting the corpus-specific overviews, it must be remembered that these distributions have not been checked against a sampling frame. They are properties of the corpus and do not necessarily allow for generalizations about the novel that the corpus aims to represent.

Besides the factors of authorship and genre and the work and edition levels, in the overviews, also the chronological aspect is covered. On the one hand, distributions over the years and decades are used to get a sense of the overall production of novels over time and how it is proportionally reflected in the corpus. On the other hand, the question arises of how Bib-ACMé and Conha19 are organized in terms of literary periods. In chapter 3.1.3 above, the chronological limits for the whole bibliography and corpus were set to 1830 and 1910, including the first national literary productions and delimiting the corpus from new avantgardistic literary currents arising in the twentieth century. However, during this long nineteenth century, several different literary currents played a major role in Spanish-American novels, particularly Romanticism, Realism, Naturalism, and Modernism. A non-trivial question is how to map these different currents to chronological periods that would allow comparing their relative coverage in the bibliography and the corpus. As Varela Jácome explains, different literary currents of European provenance reached Spanish America with delay and also simultaneously. As a consequence, there are chronological overlaps of works that can be attributed to the different currents and also works that draw their aesthetic influence from several currents at once (Varela Jácome [1982] 2000, sec. 1.1.3, 1.4.1., 2). Nonetheless, he sees a clear breakthrough of Realism in the 1880s (sec. 3). Rössner also describes the year 1880 as the beginning of a phase that was marked by significant changes in the social and economic life of all the Spanish-American countries, which led to the development of the *Modernismo* current (Rössner 2007, 200). Without deciding on clear chronological limits for the various literary currents of the nineteenth century and without needing to establish a temporal sequence of currents, the year 1880 will be used as a cutting point here to see how many works were published before and after that year both in the bibliography and the corpus.

An aspect that is not relevant for the sampling procedure here and hence also not for the evaluation of representativeness is sampling within the texts themselves. Only whole novels are included in the corpus and not, for example, selected subchapters or randomly selected text snippets of a certain size. As Gemeinböck points out in her study on "Representativeness in Corpora of Literary Texts", to use extracts of prose fiction is not advisable because, for example, beginnings and endings of the texts would differ considerably. The loss of information about entire text sections is to be considered more problematic than texts of different length (Gemeinböck 2016, 36).

Technically, the overviews in the following sections entirely draw on information that is encoded in XML-TEI. Therefore, an XSLT script was used for the calculations and to generate the visualizations.[368] With XSLT, also complex structures can be easily assessed, for example, exact publication dates versus ranges or relationships between publication dates of editions and the biographical data of authors.

Three points are important when interpreting the numbers in the following overviews. First, only authors, works, and editions contained in Bib-ACMé and Conha19 are considered without claiming completeness. The sources of both the bibliography the corpus were presented in chapters 3.2.1 and 3.3.1 above, respectively. There are probably more authors, works, and above all, more editions that would have been eligible according to the selection criteria. They are, however, not captured because they were not included in the sources selected to create the databases.

Second, there are more authors, works, and editions in the context of the bibliography and the corpus that were not included because they did not correspond to the selection criteria. For example, an author that is part of the corpus may have published more works after 1910 that are not represented here. So the following overviews concentrate on the authors, works, and editions selected for the purpose of this study but they do not represent the entire literary field.

Third, in the TEI files of the bibliography and the corpus, there are metadata values that are marked with degrees of certainty. The nationality of an author, for example, may have been assigned with low certainty if it was implied from source information but could not be verified, for instance, by an entry in an authority file. Such values with lesser degrees of certainty are not reflected in the overviews but are counted in as if they were certain. Then again, there are also completely unknown metadata values. Their number in turn is mentioned or included in the overview figures.

The analysis of the metadata on the bibliography and the corpus has produced many overview graphs since the two resources have been studied from many different perspectives to paint as comprehensive a picture as possible of their characteristics. In order to still keep the text in this chapter readable, it was decided to describe the results in the text but to outsource the actual graphs, which can be found in the appendix to this dissertation ("Appendix of Figures").

4.1.2 Authors

In the bibliography, 829 works by 383 different authors are included. The corpus contains 256 works by 121 different authors, which corresponds to 31 % of the overall number of works and 32 % of the overall number of authors.[369] The mean number of works per author is 2.2 in the bibliography and 2.1 in the corpus. The numbers show that little less than one-third of the novels and authors in the bibliography are part of the text corpus.[370]

---

[368] The script "overview.xsl" is available at https://github.com/cligs/scripts-nh/blob/master/corpus/overview.xsl. The various resulting files can be viewed at https://github.com/cligs/data-nh/tree/master/corpus/overview. Accessed September 1, 2020.

[369] Here and in the following, the percentages are rounded numbers.

[370] The number of works per author is shown in more detail in figure 94 in the appendix of figures.

The majority of authors (230, 60 %) in the bibliography wrote just one novel. 130 authors (34 %) wrote two to five novels, and 23 authors (6 %) more than five novels. In the corpus, most authors are also represented with just one novel (67 authors, i.e., 55 %). There are 47 authors (39 %) with two to five novels in the corpus and 7 authors (6 %) with more than five novels. Comparing the bibliography and the corpus, the number of authors with only one novel is a little bit lower in the corpus, whereas the number of authors with two to five novels is a bit higher. However, all in all, the proportions of the number of works per author are similar. The most productive authors in the bibliography and the authors with the most novels included in the corpus are listed in table 21.[371] Authors who only occur at the top ranks of the bibliography but not of the corpus are marked in blue, and those who only appear at the top of the corpus but not the bibliography are in orange.

In the bibliography, the most productive author is the Argentine Eduardo Gutiérrez, who is responsible for 34 novels and who wrote many popular crime and gaucho novels. He is followed by the Mexican author of historical novels Enrique Olavarría y Ferrari with 22 works, and Ireneo Paz with 17 works, a Mexican author who wrote historical as well as sentimental novels. More authors mainly dedicated to historical novels are part of the top positions in the bibliography: Juan Antonio Mateos with 14, Victoriano Salado Álvarez with 10, Francisco Guillo and Vicente Riva Palacio with 7, and Eligio Ancona with 6 novels. Among the other authors in the top list of Bib-ACMé are the Argentine writer of Realist novels Carlos María Ocantos with 13 works, the Mexican Naturalist Federico Gamboa with 8 novels, and the Mexican writer of *novelas de costumbres* José Tomás de Cuéllar with 10 novels, who are all well-known. On the other hand, there are some lesser-known authors of mainly sentimental and romantic novels who were very productive in their time: the Cuban authors Virginia Felicia Auber de Noya, Francisco Puig y de la Puente (11 novels each), Teodoro Guerrero y Pallarés (8 novels), José Güell y Renté, and Álvaro de la Iglesia (7 novels each), as well as the Mexican writer José Rivera y Río (10 novels). Together with both lesser-known authors of historical novels Enrique Olavarría y Ferrari and Victoriano Salado Álvarez, all of them are highlighted in blue and thus not part of the top list of the corpus.

In Conha19, the authors who are represented with the most novels (9 each) are again Eduardo Gutiérrez and the Mexican writer of *novelas de costumbres* José Tomás de Cuéllar. They are directly followed by Federico Gamboa and Carlos María Ocantos with 8 novels each, and the famous Cuban writer Gertrudis Gómez de Avellaneda with 7 novels, who ranges a bit lower in the top list of Bib-ACMé. Authors that entered the top list of the corpus but not of the bibliography are the Mexicans Ignacio Manuel Altamirano (5 novels), Pedro Castera, Rafael Delgado, Emilio Rabasa, and Manuel Sánchez Mármol (4 novels each), the Argentine authors Francisco Sicardi (5 novels) and Eugenio Cambaceres (4 novels), and the Cuban authors Cirilo Villaverde (5 novels) and Ramón Meza (4 novels). Except for Sicardi and Sánchez Mármol, these are all well-known authors who entered the general literary-historical canon of the countries in question. Sicardi reaches a top place because he wrote a cycle of five novels called “Libro extraño”, which is

[371] Here the top 21 were chosen because they include all the authors with 6 or more novels from the bibliography and with 4 or more novels in the corpus without having to make a cut inside of a group of authors with the same amount of novels. Authors with the same amount of novels are ordered alphabetically by surname, so the top positions can only be compared roughly. To list all authors who share a position together would have resulted in too many names at the top, especially for the corpus.

	Bib-ACMé			Conha19		
	Author name	**Country**	**Novels**	**Author name**	**Country**	**Novels**
1	Gutiérrez, Eduardo	AR	34	Cuéllar, José Tomás de	MX	9
2	Olavarría y Ferrari, Enrique	MX	22	Gutiérrez, Eduardo	AR	
3	Paz, Ireneo	MX	17	Gamboa, Federico	MX	8
4	Mateos, Juan Antonio	MX	14	Ocantos, Carlos María	AR	
5	Ocantos, Carlos María	AR	13	Gómez de Avellaneda, Gertrudis	CU	7
6	Auber de Noya, Virginia Felicia	CU	11	Calcagno, Francisco	CU	6
7	Puig y de la Puente, Francisco	CU		Paz, Ireneo	MX	
8	Cuéllar, José Tomás de	MX	10	Altamirano, Ignacio Manuel	MX	5
9	Rivera y Río, José	MX		Ancona, Eligio	MX	
10	Salado Álvarez, Victoriano	MX		Holmberg, Eduardo Ladislao	AR	
11	Gamboa, Federico	MX	8	Sicardi, Francisco	AR	
12	Gómez de Avellaneda, Gertrudis	CU		Villaverde, Cirilo	CU	
13	Guerrero y Pallarés, Teodoro	CU		Cambaceres, Eugenio	AR	4
14	Calcagno, Francisco	CU	7	Castera Cortés, Pedro	MX	
15	Guillo, Francisco	AR		Delgado, Rafael	MX	
16	Güell y Renté, José	CU		Gorriti, Juana Manuela	AR	
17	Iglesia, Álvaro de la	CU		Mateos, Juan Antonio	MX	
18	Riva Palacio, Vicente	MX		Meza, Ramón	CU	
19	Ancona, Eligio	MX	6	Rabasa, Emilio	MX	
20	Gorriti, Juana Manuela	AR		Riva Palacio, Vicente	MX	
21	Holmberg, Eduardo Ladislao	AR		Sánchez Mármol, Manuel	MX	

Table 21. Authors with most novels in BibACMé and Conha19.

completely included in the corpus, and Sánchez Mármol because there is a recent edition of his complete works, including novels, from 2011 (Sánchez Mármol 2011).

The comparison of the top productive authors in the bibliography and the corpus shows that 12 of the 21 authors occur in both lists, which is a bit more than half of them. Moreover, some differences become visible: in the bibliography, many of the writers who wrote much did so in specific subgenres of the novel. In addition, some lesser-known authors are prolific writers. On the other side, in the top list of the corpus, well-known canonical authors play a more important role, and specific subgenres are a little less important. This has very practical reasons: the corpus is built as much as possible on novels that were available in a digital full-text format and to date, there are more such digital editions of works written by the more prominent authors.

How does the picture change if not the number of works but the number of editions per author is considered? The number of historical editions[372] that have been published of an author's works is not so much a sign of productivity but of success, be it because the works were valued highly by contemporaries or read much.[373]

Most authors in the bibliography have only published one edition (191 authors, i.e., 50 %). 148 authors (39 %) published two to five editions, and 44 (11 %) authors more than five editions. In the corpus, 41 authors (34 % of all the authors in the corpus) are represented with one edition, 54 authors (45 %) with two to five, and 26 authors (21 %) with more than five editions. If one compares the proportion of authors represented with a certain number of editions in Conha19 and Bib-ACMé, the numbers show that the corpus contains fewer authors with only one edition, a bit more with two to five editions, and considerably more with more than five editions. The numbers of editions indicate that the works contained in the corpus were, on average, republished more often than the works in the bibliography. All in all, the authors in the corpus were more popular, more successful, or had more prestige than the average author in general. This observation is in line with the above finding that the authors represented with most works in the corpus are the ones that are more known and more canonized. The same picture emerges when looking at the list of authors with most editions in Bib-ACMé and Conha19, represented in table 22.[374]

Compared to the list of authors with the most works, the top lists of the authors with the most editions differ less between the bibliography and the corpus. Only five authors instead of nine are not contained in the other list, respectively. This is because, by the number of editions, more of the well-known and successful authors enter the bibliography list, although they are not the ones that wrote most works. These are the Mexicans Ignacio Manuel Altamirano and Rafael Delgado and the Cuban author Cirilo Villaverde. Authors that newly enter the corpus top list are the Argentine José Mármol, the Mexican Manuel Payno, and the Cuban author Teodoro Guerrero y Pallaŕes. José Mármol is famous for just one novel, "Amalia", which was very successful.

---

[372] I.e., inside the chronological limits of Bib-ACMé: 1830–1910.

[373] In figure 95 in the appendix of figures, the number of editions per author is shown for both Bib-ACMé and Conha19. For Conha19, only editions of the works that are included in the corpus are counted.

[374] As in the previous table for the authors with the most works, also here, the top positions are ordered first by the number of editions and then alphabetically so that the positions cannot be interpreted in a strict sequence. Here, 20 positions were chosen because they include all the authors with 8 or more novels from the corpus. For the bibliography, a cut-off had to be made inside the group of authors with 10 editions, leaving only Juana Manuela Gorriti in the table.

	Bib-ACMé			Conha19		
	Author name	**Country**	**Novels**	**Author name**	**Country**	**Novels**
1	Gutiérrez, Eduardo	AR	89	Gutiérrez, Eduardo	AR	29
2	Olavarría y Ferrari, Enrique	MX	41	Gómez de Avellaneda, Gertrudis	CU	24
3	Gómez de Avellaneda, Gertrudis	CU	25	Altamirano, Ignacio Manuel	MX	17
4	Mateos, Juan Antonio	MX		Cuéllar, José Tomás de	MX	16
5	Paz, Ireneo	MX		Gamboa, Federico	MX	15
6	Puig y de la Puente, Francisco	CU	20	Mateos, Juan Antonio	MX	13
7	Cuéllar, José Tomás de	MX	18	Riva Palacio, Vicente	MX	12
8	Altamirano, Ignacio Manuel	MX	17	Villaverde, Cirilo	CU	
9	Ocantos, Carlos María	AR	16	Díaz Covarrubias, Juan	MX	11
10	Riva Palacio, Vicente	MX		Calcagno, Francisco	CU	10
11	Gamboa, Federico	MX	15	Delgado, Rafael	MX	
12	Guerrero y Pallarés, Teodoro	CU		Mármol, José	AR	
13	Rivera y Río, José	MX	14	Ocantos, Carlos María	AR	
14	Villaverde, Cirilo	CU	13	Cambaceres, Eugenio	AR	9
15	Auber de Noya, Virginia Felicia	CU	11	Paz, Ireneo	MX	9
16	Calcagno, Francisco	CU		Sicardi, Francisco	AR	9
17	Delgado, Rafael	MX		Castera Cortés, Pedro	MX	8
18	Díaz Covarrubias, Juan	MX		Guerrero y Pallarés, Teodoro	CU	8
19	Holmberg, Eduardo Ladislao	AR		Holmberg, Eduardo Ladislao	AR	8
20	Gorriti, Juana Manuela	AR	10	Payno, Manuel	MX	8

Table 22. Authors with most editions in BibACMé and Conha19.

Manuel Payno published three works, two of which were successes (“El fistol del diablo” and “Los bandidos de Río Frío”) and re-edited in his time. Teodoro Guerrero y Pallarés enters the list because he wrote many novels, of which several were published with more than one edition, especially “Anatomía del corazón”. New to both top lists is the Mexican Juan Díaz Covarrubias, author of three novels that were all re-edited. So in terms of quantity, the field of top authors shifts when considering the number of editions instead of the number of works, bringing the bibliography and the corpus closer together.

Other important points to present about the authors in Bib-ACMé and Conha19 are their provenance, nationality, and belonging to a certain country because both resources include authors from Argentina, Cuba, and Mexico.[375]

[375] In figure 96 in the appendix of figures, the proportions of authors by country are displayed for both the bibliography and the corpus.

In Bib-ACMé, most authors are associated with Mexico, followed by Argentina. In Conha19, in contrast, there are more authors belonging to Argentina than to Mexico. However, the numbers for these two countries range between 37 and 46 %, so the difference is not too big. In both cases, there are fewer authors that are connected to Cuba, 14 % in the bibliography and 20 % in the corpus, meaning that Cuban authors are a bit overrepresented in the latter.

The division into three countries is a simplification because authors were assigned to the countries based on several different criteria. They can, for example, have the nationality of the country, either because they were born there or naturalized at some point, or they are considered as belonging to the country because it was their primary place of residence and work and they published their novels there. A closer look into the nationalities, birth, and death places shows that also other countries beyond Argentina, Cuba, and Mexico are involved.[376]

By nationality, most authors are Mexican, Argentina, and Cuban in the bibliography, and Argentine, Mexican, and Cuban in the corpus. Besides that, also authors with Spanish nationality play a role in both contexts. In the corpus, further nationalities are only represented by one author each (Chilean, Dominican, French, and Uruguayan). The nationality of one author in the corpus is unknown (C. M. Blanco, whose novel "Salvaje. Novela argentina" was published in Barcelona and Buenos Aires in 1891). In the bibliography, there are seven authors with Uruguayan nationality, two each with Chilean, Dominican, and French nationality, and further nationalities represented with just one author. In Bib-ACMé, the nationality of seven authors is unknown. Altogether, twelve different nationalities are involved.

The picture is different when the authors' countries of birth are considered. For most authors in the bibliography (44 %), the place of birth could not be verified. Otherwise, most authors included in Bib-ACMé are born in Mexico, followed by Argentina, Cuba, and then Spain and Uruguay. Interestingly, the proportion of authors born in Argentina (12 %) is only slightly higher than the proportion of authors born in Cuba (10 %). However, more authors were associated with Argentina because they were included in the source bibliographies, especially the comprehensive bibliography of the Argentine novel by Lichtblau, and because their works were published in Argentina, but there is not much knowledge about many of these authors. In the corpus, the order of countries of birth is the same as in the bibliography (Mexico, Argentina, Cuba, Spain, Uruguay), but the proportion of authors with unknown countries of birth is much less (13 %). This again illustrates that the corpus authors are mainly well-known writers or at least that their share is bigger than in the bibliography.

The country of death is also unknown for most authors in the bibliography (48 %), followed by Mexico, Argentina, Cuba, and Spain. A country that gains a bit more relevance as a place of death is the USA, where six authors died. These are authors born in Cuba, Mexico, and the Dominican Republic. In the corpus, again, the place of death is known for many more authors (it is unknown for only 14 %). Besides that, the proportion of authors who died in Argentina and Mexico is equal, followed by Cuba, Spain, and the USA, where four of the authors died. The overviews of the relationships between authors and countries make clear that the Argentine, Cuban, and Mexican literatures, as understood in the context of this study, are not fixed and closed spaces, but that connections to other countries exist, as is probably the case for all "national" literatures.

---

[376] See the corresponding figures 97 to 99 in the appendix of figures.

Another topic is the gender of the authors. In the bibliography, the great majority is male (353 authors, i.e., 92 %), and there are only 23 (6 %) female authors. In 7 cases (2 %), the gender of the author is unknown. In the corpus, the proportion of female authors is a bit higher (11 authors, i.e., 9 %), and there is only one author whose gender is unknown (the author of the novel "Salvaje", called "C. M. Blanco").[377]

It is not only of interest to know how many authors of a particular gender there are but also for how many of the works they are responsible. In the bibliography, 756 works (91 %) are written by male authors, 58 works (7 %) by female authors, and 15 works (2 %) by authors of unknown gender. In the corpus, 229 novels (89 %) are written by male authors, 26 novels (10 %) by female authors, and one novel by an author of unknown gender. If one compares these numbers to the number of authors in general, it can be noted that, on average, female authors are slightly more productive than male authors.

Finally, also the life dates are of interest to get a sense of which authors are included in the bibliography and the corpus. Unfortunately, they could only be verified for a subset of the authors.[378]

The complete life dates, i.e., the years of birth and death, are only known for 63 % of the authors in the bibliography and for 88 % of the authors in the corpus. No life dates at all are known for 33 % of the authors in the bibliography and 8 % in the corpus. For 4 % of the authors in Bib-ACMé and 3 % in Conha19, only the year of birth or death is known. That much more is known about the life dates of the authors in the corpus than in the bibliography again shows that the latter covers more of the less canonized literary production. This state of knowledge has to be kept in mind for the following overviews, in which life dates are used to calculate how many authors were alive or active at a certain point in time and how old they were when they published their works.[379]

In the bibliography, the first author was born in 1776. This is the Mexican José Joaquín Fernández de Lizardi. He was also the first author who died (in 1827). Lizardi is often considered the author of the first Mexican or even Spanish-American novel "El Periquillo Sarniento" (1816, MX). On the other hand, his novels are also described as forerunners of the nineteenth-century Spanish-American novel proper (Alegría 1959, 18–26; Janik 2008, 34–36; Sánchez 1953, 111, 115–123). Because its publication date lies outside the scope of this study, the novel "El Periquillo Sarniento" is not included here. In the bibliography, Lizardi is only represented with the novel "Don Catrín de la Fachenda", which was first published posthumously in 1832 but is also not included in the corpus. The next author, included in both the bibliography and the corpus, is the Cuban Esteban Pichardo y Tapia, born in 1799. In the following decades, the number of births increases considerably. More than half of the authors in the bibliography and the corpus whose birth dates are known were born between the 1830s and the 1860s. The last authors were born in the 1880s. One of them is part of the corpus: the Argentine Enrique García Velloso, who was born in 1880. Considering the years of death, after Lizardi, the first authors died in the 1850s, and the last ones in the 1960s. The first author in the corpus who died was Juan Díaz Covarrubias, a

[377] The author gender proportions in Bib-ACMé and Conha19 are visualized in figure 100 in the appendix of figures.
[378] See figure 101 in the appendix of figures.
[379] In figure 102 in the appendix of figures, the number of births and deaths of the authors are visualized by decade.

Mexican writer who died in 1859 at the age of 21 in the civil war of the Reform (Yin 1992, 195), and the last one was the Argentine Enrique Larreta, who died in 1960. Most authors died in the 1890s, 1910s, and 1920s. Without detailed biographic research, it cannot be said with certainty why there were fewer deaths in the 1900s than in the preceding decade and the following two decades. It may have had an influence that the 1900s were a politically and economically more stable decade than the others. All in all, the life dates of the authors comprise 190 years, from the 1770s to the 1960s, for a bibliography and corpus that is limited to 80 years. In such a broad range, several generations of authors are involved, and not all the authors experienced the same historical times. Nevertheless, there is a core of contemporaneity. Between the 1850s and the 1910s, more than half of the authors for whom birth and death dates are known were alive.[380]

Another question is when these authors were not only alive but also active. "Activity" is interpreted here as the phase when the authors published new works, i.e., the years in which they actually published or in which they already had and still were to publish more works.[381] Compared to the top period of authors alive, the most authors that were active at the same time are to be found later, in the 1880s and the 1890s. For the bibliography, the top is reached in the years 1886 and 1887, when 53 authors (22 % of all the authors with known life dates) were active at the same time. In the corpus, the top year is 1884, with 33 authors (31 %). It becomes clear that the bibliography and corpus are closer together in the early decades, meaning that the coverage of authors (at least of the ones with known life dates) is better in this phase. Although the corpus includes more authors that were active in the later decades, there are even more in the bibliography, showing that the overall number of active authors and works published increased considerably towards the end of the century.

How old were the authors when they published works? This question brings the two perspectives of "authors alive" and "authors active" together.[382] The median age of the authors when they published a novel is the same for the bibliography and the corpus and lies at 37 years. Considering that most authors were born in the 1850s, it makes sense that most of them were active in the 1880s. The youngest author at publication was Carlos María Ocantos, whose novel "El esclavo" was supposedly published when he was 14 years old (Lichtblau 1997, 744). The oldest author was Vicente Fidel López, who published "La Gran Semana de 1810" and "La loca de la guardia" at 81 years. The average life expectancy of the authors in the bibliography was 66 years, and in the corpus, 65 years.[383]

All in all, not many differences were found between the authors contained in the bibliography and those included in the corpus. About one-third of the authors in Bib-ACMé are also represented in Conha19. Most of the authors only published one work between 1830 and 1910, of which, in most cases, also only one edition was produced. This is probably not the impression one gets when reading literary histories, where the center of interest is often on the minority of more productive, well-known authors. These are a bit overrepresented in the corpus when the number

[380] See figure 103 in the appendix of figures. The figure displays the sums of authors who were not yet born, alive, and dead over the years. The death curves level off towards the end, probably a sign that the authors got older over time.
[381] The sums of active authors per year are shown in figure 104 in the appendix of figures.
[382] See figure 105 in the appendix of figures.
[383] See figure 106 in the appendix of figures.

of works and also editions are considered. However, there are some also lesser-known authors who wrote much, which are more present in the bibliography than in the corpus. Regarding the distribution of authors by country, there are no big differences between the bibliography and the corpus. In the latter, there are relatively more Argentine authors and fewer Mexican authors than in the bibliography. Furthermore, Cuban authors are a bit overrepresented in the corpus, although they are the smallest group. Additional countries play a role in the nationalities of the authors and as countries of birth and death, especially Spain, but they range below 10 % of the authors. Regarding gender, there are relatively more female authors in the corpus than in the bibliography, but the difference is only about 3 %. The life dates of the authors were also evaluated and not much difference was found between authors in Bib-ACMé and Conha19. Most authors lived between the 1850s and the 1910s, and most were active in the 1880s and 1890s. The average age of an author when publishing a work is the same in the bibliography and the corpus, and the age of death of the authors also only differs by one year.

4.1.3 Works

4.1.3.1 Comparison of Bib-ACMé and Conha19

829 works are registered in the bibliography, of which 256 (31 %) are contained in the corpus. The previous chapter discussed how many works were published per author. In this chapter, the first focus is on the number of works published over time, using the publication years of the first known editions of the works.[384]

First, it is analyzed how many works were published per year between 1830 and 1910. The first work in the bibliography is from the year 1832, and in the corpus, there are two works first published in 1839. The last works in both Bib-ACMé and Conha19 were published in 1910. Apart from the 1830s, when only a few works were published, almost all the years are covered in the bibliography. Exceptions are the years 1849, 1852, and 1867. As the numbers were generally low in the 1840s and the early 1850s, it is possible that no novels at all were published in 1849 and 1852. In the year 1867, however, it is surprising. It may be the case that the political situation in the three countries made it difficult for authors to write or publish novels in that year. During the time, Argentina was involved in the War of the Triple Alliance, Cuba stood at the beginning of a period of internal wars, and in Mexico, the emperor Maximilian was overthrown by liberal troops. However, verifying that this had an effect on the number of novels that were published would require more research into the personal circumstances of the authors and the history of the publishing sector. Other years that are not represented in the corpus are some of the years in the 1830s, 1845, 1850, 1853, 1863, and 1878. Apart from the 1830s, for which it is more difficult to access the few novels that were published, this is interpreted as the effect of random selection. The number of works published increased considerably towards the end of the century. From 1880 on, at least ten works were published per year.

[384] See the figures 107 to 109 in the appendix of figures.

Summarizing the values for decades, the coverage of works in the corpus in relationship to the bibliography becomes clearer.[385] From the 1850s to the 1890s, the share of works in the corpus is about one-third, which corresponds to the overall proportion of works included in the corpus. The 1860s are slightly overrepresented with 39 %. In the margins, i.e., the early and late decades, the numbers deviate more. The 1830s, 1900s, and 1910s are underrepresented in the corpus, and the 1840s are strongly overrepresented. Apart from the 1900s, such deviations are more likely in these decades because the overall number of works is much lower than in the central decades.

Summarizing even more and comparing the period before 1880 to the period in and after that year,[386] this results in a better representation of the earlier period. 37 % of the works in the bibliography that were published before 1880 are also contained in the corpus. In the later period, 28 % of the works in the bibliography are also part of the corpus. This means that the corpus contains proportionally fewer works in the period after 1880, although in total, more works were published in the later decades of the nineteenth century.

Another perspective on the number of works over time is obtained by differentiating by country.[387] When the development of the number of works published over the decades is observed that way, different patterns become visible. In Argentina, the number of works exploded in the 1880s. According to the bibliography, 90 works were published in that decade, compared to around 20 works each in the three decades before. Apart from the 1830s, for which only one work is included in the bibliography, all the decades are also represented in the corpus. For Argentina, of the central decades, the 1870s are overrepresented and the 1900s underrepresented. In Mexico, the number of works published rose earlier and not so sharply. Considerably more works were published from the 1860s onwards, and their number increased towards the end of the century, whereas in Argentina, the number decreased again after the 1880s. Regarding the decades with many works, the 1860s are overrepresented for Mexico in the corpus. As for Argentina, no work from the 1830s is included in the corpus. In contrast to Argentina and Mexico, the number of Cuban works published between 1830 and 1910 does not show significant growth. Actually, most works that are included in the bibliography were published in the 1850s (28 works), followed by the 1890s (21 works). Compared to Argentina and Mexico, more Cuban works were published in the early decades, from the 1830s to the 1850s. It is very probable that this different development of the number of published novels in the course of the nineteenth century is due to Cuba's status as a colony, which only ceased in 1898, and which prevented the growth of the literature marked. On the other hand, Cuba was colonized early and had a close connection to its motherland Spain, which could explain why relatively more novels were published in the early decades by Cuban-Spanish than by Argentine and Mexican authors. However, the quality of the bibliographic sources used can also play a role, as discussed above in chapter 3.2.1. Comparing the overall number of works by country, most novels in Bib-ACMé and also in Conha19 are Mexican, followed closely by Argentine novels. The Cuban novels make up the smallest part, with 16 and 19 %, respectively.[388]

---

[385] In figure 108 in the appendix of figures, the percentages indicate the proportion of works contained in the corpus for each decade.

[386] See figure 109 in the appendix of figures.

[387] See figure 110 in the appendix of figures.

[388] See figure 111 in the appendix of figures.

Looking not at which countries the novels are generally associated with in the bibliography and the corpus but in which countries they were first published, the role of Spain becomes visible: 7 % of the novels in the bibliography and 9 % of the novels in the corpus were first published in that country.[389] Apart from the lowest numbers, there is no difference between the bibliography and the corpus concerning the ranks of the countries where the novels' first editions were published. Most novels were first published in Mexico, followed by Argentina, Cuba, Spain, France, and the USA. Comparing the numbers of the publication places to the general numbers by country, it becomes clear that not only part of the Cuban novels were first published in Spain, but also Mexican and Argentine novels.

Comparing the works in Bib-ACMé and Conha19 revealed that both are proportionately quite congruent when the distribution of works over time and the share of works by country are considered. Nevertheless, on a level of detail, also some differences became visible. In the corpus, especially the 1860s are overrepresented, and the 1900s are underrepresented. As a result, the period before 1880 is covered to a higher degree in the corpus than the one after this year. Regarding the countries, there are relatively more Argentine and Cuban and fewer Mexican works in the corpus than in the bibliography, but these differences range only between 2 to 6 %. As to the overall distribution of works over time, almost all the years between 1830 and 1910 are covered in the bibliography and the corpus, with the exception of some early years from the 1830s to the early 1850s, plus the exceptional year 1867, in which no works were published. Especially from the 1880s onwards, the number of works published increased considerably. However, this development is not the same in all three countries. In Argentina, most works were published in the 1880s; in Mexico, the number of works grew already in the 1860s; and in Cuba, no significant growth over time can be recognized at all.

4.1.3.2 Corpus-specific Overviews

Besides the metadata that can be evaluated for both the bibliography and the corpus, some informative aspects about the novels are only available for the corpus. They depend on more specific metadata that has only been collected for Conha19 or on the full texts of the novels that are only available in the corpus. Such aspects are analyzed in this chapter. Part of the metadata that was gathered for the corpus refers to technical and administrative aspects, such as the type of source medium, the kind of source edition, and the institution that held the source. Summaries of these data were already given in chapter 3.3.1 ("Selection of Novels and Sources") above and are not discussed here any further.

One metadata item that was only collected for the novels in the corpus is their status as high- or low-prestige novels.[390] In Conha19, 174 novels (68 %) are classified as high prestige and 82 novels (32 %) as low prestige. There is no difference in the proportion of high- and low-prestige novels from Cuba, but from Mexico there are more high-prestige novels, and from Argentina more low-prestige ones.[391] There are several probable reasons for this. Surely, the quality of

[389] See figure 112 in the appendix of figures.
[390] See chapter 3.3.3.1.6 ("Text Classification with Keywords") above for details on how prestige was modeled.
[391] See figure 113 in the appendix of figures for an overview of the proportions of high- and low-prestige novels by country.

the bibliographic sources used as a basis for selecting novels for the corpus is an influencing factor. The bibliography of the Argentine novel authored by Lichtblau is very comprehensive and also includes many authors and works that are not well-known. Furthermore, the state of digitization and access to digital sources plays a role. The collection of Argentine novels published on Wikimedia Commons by the Argentina Academy also contains many works written by lesser-known authors. Independently of the reasons, the corpus has a certain bias towards low-prestige Argentine and high-prestige Mexican novels.

The analysis is now deepened by considering the distribution of novels by prestige over time.[392] Over the decades, most low-prestige novels were included in the 1890s, 1880s, and 1860s. On the other hand, low-prestige novels are underrepresented in the 1840s, 1870s, and the 1900s. The decades 1830 and 1910 are not really informative because the number of works in them is so small. It makes sense that in the decades in which the overall production of novels increased considerably, more low-prestige novels were produced and were also selected for the corpus. Regarding the 1900s, they are, in general, underrepresented in Conha19[393], so the probability of selecting high-prestige works is higher. The 1840s, in contrast, are generally represented quite well in Conha19[394], so it can be assumed that there were not many works in that decade that are considered low-prestige today, or if there were, they are not known. Why low-prestige works are underrepresented in the 1870s is not clear. Summarized to the two periods before and in or after 1880, the proportion of low-prestige works is higher in and after 1880.

Another metadata item that is only available for the corpus is the narrative perspective of the novels. In general, there are 44 novels (17 %) with a first-person narrator and 212 novels (83 %) with a third-person narrator, so the latter clearly prevails. Regarding the distribution of the narrative perspective by country, it is interesting that the proportion of Cuban novels is much lower for the novels narrated in the first person than for those narrated in the third person. A hypothesis is that the individual, personal perspective was not so suitable for novels published in the colony. The first-person novels are mainly Argentine, closely followed by the Mexican novels, and the third-person novels are above all Mexican.[395]

When analyzed over time[396], it becomes visible that the proportion of novels written in the first person was highest in the 1870s, 1890s, and 1880s. No first-person novel is included from the 1850s, and also the 1840s and 1900s are mostly represented with novels narrated in the third person. The drop of first-person novels in the 1900s is surprising because otherwise, they became more frequent after the middle of the nineteenth century. Again, this decade is generally underrepresented in the corpus, which might be a reason. Comparing the period before 1880 to the one in and after 1880 shows that narrations in the first person are relatively and absolutely more frequent in the latter period, although the difference between the two periods only amounts to 6 %.

---

[392] See figures 114 (by decade) and 115 (before versus in or after 1880) in the appendix of figures.
[393] The mean proportion of novels in the corpus compared to the bibliography is 31 %, and the 1900s are only represented with 20 %. See the overviews in the previous chapter.
[394] Comparing the corpus to the bibliography, in the 1840s, 54 % of the works are covered.
[395] See figure 116 in the appendix of figures.
[396] See figures 117 and 118 in the appendix of figures.

Further information that was collected for the novels in the corpus is the continent and country of the setting. 9 % of the novels are primarily set in Europe, and just one novel is set on another continent[397], so the great majority of 90 % is set in America. Looking at the country of the setting, Mexico, Argentina, and Cuba are most frequent, corresponding to the countries of origin of the novels. Other places the novels are set in are the European countries Spain, Italy, France, Greece, Switzerland, the USA, and other South-American countries (Peru, Chile, Bolivia, Brazil, and Uruguay). Together, American countries other than Mexico, Argentina, and Cuba make up for 4 % of the cases.

An evident question is if the preference for a European or American setting was influenced by the country of origin of the novels, i.e., if there is a difference between the Argentine, Mexican, and Cuban novels in this aspect. Analysis of the metadata shows that this is indeed the case.[398] While the proportions of Mexican, Argentine, and Cuban novels set in America correspond largely to the general significance of these countries in the corpus, the numbers are quite different for the novels set in Europe. Here, the majority is Cuban and the minority Mexican, suggesting that Cuba's status as a colony during most of the nineteenth century had an influence on the setting of the novels. In addition, the relationship between Argentina and Europe was closer than that between Mexico and Europe as regards the choice of setting for the fictional texts. However, in absolute numbers, only 24 novels are set in Europe, so these trends should also not be overinterpreted. Following the proportions of works set in Europe over the decades and comparing their share in the period before 1880 and in or after that year makes clear that the number of works set in Europe decreased over time. Moreover, at least in the corpus, the 1860s were already a decade in which an American setting was clearly preferred.[399]

Besides the continent and country of the setting, also the time period covered by the novels is registered in the metadata. Three time periods are distinguished: contemporary, recent past, and past. For each novel, two values were encoded: the time period relative to the author's birth year and relative to the year of the first known publication of the novel.[400] From both points of view, a contemporary setting is the most frequent one: related to the authors' birth years in 82 % and relative to the publication date in 73 % of the cases. Novels set in the past are the second most frequent group, with 13 % and 16 %, respectively. The recent past is treated in 4 % of the novels when the authors' birth years are concerned and in 11 % of the works when the publication year is decisive. The differences between the two approaches show that more novels treat a period that is past in relation to the publication date but still part of the contemporary experience of the authors. Fewer novels treat a period that lies in a more distant past. In the following, only the approach of comparing the publication date to the time period covered by the novels is considered further.

---

[397] See figure 119 in the appendix of figures. The novel with another setting is the science fiction novel "Viaje maravilloso del Señor Nic-Nac" (1875, AR) by Eduardo Holmberg, which tells an imagined trip of the protagonist to the planet Mars.

[398] See figure 120 in the appendix of figures.

[399] See figures 121 and 122 in the appendix of figures.

[400] See chapter 3.3.3.1.6 ("Text Classification With Keywords") above for details. The proportions of works set in the different time periods are visualized in figure 123 in the appendix of figures.

How do the proportions of the three time periods covered in the novels relate to the three different countries that the novels are associated with in the corpus?[401] As can be seen, the Argentine and Mexican novels cover most of the contemporary perspective. As the general proportions of works by country were 42 % Mexican works, 39 % Argentine works, and 19 % Cuban works,[402] a contemporary setting is a bit overrepresented in the Argentine novels and slightly underrepresented in the Mexican and Cuban novels. A setting in the past is overrepresented in Mexican and Cuban novels and underrepresented in Argentine novels. Here the differences range between 3 and 6 % of the novels. The preferences are most striking concerning a setting in the recent past. Here, Mexican novels are overrepresented by 8 % and Cuban novels by 10 %, while Argentine novels are underrepresented by 18 %. All in all, the past is more a topic in the Mexican and Cuban novels, and the recent past is relatively most important in the latter ones, while the contemporary perspective is favored in the Argentine novels. This might be explained by the fact that the colonial history of Cuba and Mexico is longer than that of the Argentine region. In the case of Cuba, another factor is the difficulty of broaching contemporary issues in a country that was still under the control of the mother country, which may have led to a preference for representing the recent past. On the other side, Argentine society and economy developed rapidly in the nineteenth century, which supplied much material for the novels treating the contemporary period.

Did the preference for setting the novels in a certain time period change in the course of the nineteenth century? An analysis of the distribution of time periods per decade and a comparison of the period before 1880 to the one after that year suggests that there is no clear chronological trend but that there are some intermittent preferences instead.[403] The contemporary period was always dominating as a setting for the novels. Interestingly, the past was favored more in the 1860s and then again in the 1900s. Leaving out the first and last decades with very few works, also the recent past reaches the top positions in these two decades. The lowest proportions of novels treating the past and recent past can be seen in the 1880s and 1890s. So a first trend of treating historical issues came up shortly after independence was reached in most Spanish-American countries, probably as a way of contributing to writing their own history in literary terms. Then, in the decades of significant social and economic development, contemporary issues were more relevant. A return to a greater interest in the recent and further past at the beginning of the new century marks a new phase.[404] Condensing this development to the phase before and after 1880

---

[401] See figure 124 in the appendix of figures.

[402] See figure 111 on the works by country in the appendix of figures, which was discussed in previous chapter.

[403] See figures 125 and 126 in the appendix of figures.

[404] The occupation with events that belong to the distant or recent past or are contemporary is not necessarily to be equated with the novels being historical novels or not. As Read states in his study of the Mexican historical novel: "It will be apparent as this study progresses that the works involved fall naturally into two groups, the romantic historical novels that deal with the conquest period and colonial times, and the novels that deal with historical events of the nineteenth century. The first group is essentially romantic, corresponding to the type developed by Walter Scott but with a distinctly local 'middle ages'. Instead of turning to medieval Europe for exotic material, Mexican writers of this type of fiction sought out characters and institutions of their own dim past. Their hostility to the Spanish regime was still fresh enough to inspire them with a feeling of spiritual kinship with the Amerinds who had been the traditional enemies of the Europeans. […] In this same group of romantic historical novels belong those fictional works that deal with colonial times. The Inquisition […] is naturally the center of

results in relatively more novels treating the contemporary period in the latter and some form of the past in the former.

A characteristic of the novels in the corpus that goes beyond metadata is the length of the texts. In the context of the definition of boundaries of the novel, the minimum length of novels was discussed in detail in chapter 3.1.1.4 above. Therefore, regarding the novels that were included in the corpus, the question remains how long these actually are in terms of the number of tokens.[405] The shortest novel in the corpus has about 16,000 tokens, the longest one has about 331,000 tokens, and the median length is approximately 53,000 tokens.[406] It is interesting that the medium length of the Spanish-American novels in this corpus corresponds almost to the minimum length for novels set by Forster to 50,000 words (Forster 2016, 17), so the nineteenth-century Argentine, Mexican, and Cuban novels, as defined here, tend to be shorter than the typical English novel that Forster had in mind. 25 % of the novels are between 16,000 and 35,000 tokens long, the next 25 % between 35,000 and 53,000 tokens, the third quarter is between 53,000 and 96,000 tokens, and the last one between 96,000 and 331,000 tokens, so the spread of lengths increases considerably for the upper 50 % of the novels and the longest novels are clearly outliers.

---

interest of these poetic interpretations of life in the colony. The second class of Mexican historical novels is that which finds its material in the history of the nineteenth century, the epoch in which the writers themselves had been actors in the dramas they presented. Such works may properly be called novels of contemporary history. Many of them were patterned after the *Episodios nacionales* of Pérez Galdós and the various historical romances of Erckmann-Chatrian. But though these two groups of works deal with materials from widely separated periods, they have much in common. Whatever the period involved, it was interpreted in terms of the ideals of the nineteenth century when Mexico was attempting to constitute itself a new nation […]. Patriotism, a new sense of national identity and zeal for liberty and justice were the emotive forces that determined the trend of interpretation in both groups of historical novels to which attention has been called" (Read 1939, ix–xi). Both types of novels also existed in Argentina: historical novels in the strict sense, which broached the issues of the Conquest and colonial times, and novels that treated contemporary historical events or those of the very recent past. Many of the early novels of the latter kind had the Rosas regime as their subject. Molina subsumes them under the group of political novels and calls them "novelas prospectivamente históricas" (Molina 2011, 246–249, 285–311). A contemporary setting was also predominant in the realist and naturalistic novels of the later nineteenth century: "With reference to Argentina, it is significant that the development of the realistic novel should coincide with the extraordinary growth and progress which the Republic manifested during the years 1880 to 1900. As this economic and material transformation took place, greatly affecting every facet of the nation's life […] eager writers sought to mirror that rapid change and portray the new society that was surging forth" (Lichtblau 1959, 138). Regarding the Argentine naturalistic novel, Lichtblau remarks: "Not only did Argentina produce the first naturalistic novelist in Hispanic America in the figure of Cambaceres, but that country displayed as well the greatest over-all development of the naturalistic current in the nineteenth century. The tremendous material advancement, the great influx of immigrants, the changing social pattern, and the growing industrialization of the Republic–all these things writers used to advantage in applying the tenets of Zola to Argentine fiction." (176–177).

[405] This is summarized in the box plot in figure 127 in the appendix of figures.

[406] The numbers are rounded to the next thousand. The shortest novel is "Gubi Amaya" (1865, AR) by Juana Manuela Gorriti, and the longest one is the historical novel "El mendigo de San Ángel" (1865, MX) by Niceto de Zamacois. Surprisingly, the shortest and the longest novel of the corpus were first published in the same year. Despite the lower limit being 16,000 tokens, two novels with approximately 15,800 tokens, among them "Gubi Amaya", were included because the number of tokens changed in the course of the preparation of the corpus, amongst other things, due to the correction of spelling errors. So in a strict sense, these two novels only fulfill the minimum length criterion when the number of tokens is rounded to the next thousand.

Analyzed by country, the distribution of lengths is very similar for the Argentine and Cuban novels but different for the Mexican novels.[407] The median Argentine novel is 48,000 tokens long, and the median Cuban novel has 50,000 tokens, so they are both shorter than the overall median novel. The longest Argentine novel has 231,000 tokens, and the longest Cuban novel has 198,000 tokens. Compared to that, Mexican novels are longer. The medium Mexican novel is 67,000 tokens long, and the three longest novels with over 300,000 tokens are also Mexican. The three longest novels are historical novels, so it should be examined if there is a correlation between the length of the novels and their subgenre, which is done for thematic subgenres and literary currents in chapter 4.1.5.3 below. Testing for statistical significance, it turns out that the difference in length between the Mexican and the Argentine, as well as between the Mexican and the Cuban novels, is indeed significant.[408]

How does the novels' length develop over the decades?[409] First, in the decades 1830 to 1850, the median length drops from 110,000 to 36,000 tokens. In the 1860s, it jumps to 87,000 tokens, and after that, it raises from 47,000 in the 1870s to 99,000 in the 1910s. The works in the 1830s and 1910s are very few, though. Regarding the median, it is especially interesting to see the exceptional decade of the 1860s. As was found out in the evaluation of the time periods of the novels' settings, in this decade, representations of the past were relatively favored. In addition, they were preferred in Mexican novels, and these were more numerous in the 1860s than Argentine and Cuban novels.[410] Together with the observation that the longest novels of the corpus are historical novels, this might explain why there were more long novels in this decade than in the others. A test for statistical significance reveals that the text lengths can be considered significantly different in the following constellations of decades: 1860s versus 1870s, 1860s versus 1880s, and 1880s versus 1900s.[411] It is also noteworthy to see that the variability of the texts' length (in terms of the spread of length in the two central quartiles) is lower from the 1870s to the 1890s than before and after that decade. In the last three decades of the nineteenth century, very long novels are the exception.[412]

---

[407] See figure 128 in the appendix of figures. Again, in the following, the numbers are rounded to the next thousand.

[408] The script used for the significance tests is available at https://github.com/cligs/scripts-nh/blob/master/analysis/sign.py. Accessed January 3, 2021. Because the data is not normally distributed, the Mann-Whitney U test was used (instead of a t-test, for instance). The p-value that was calculated for the text lengths of Mexican versus Argentine novels is 0.001, for Mexican versus Cuban novels 0.04, and for Argentine versus Cuban novels 0.2, which means that there is no significant difference in the latter case.

[409] See figure 129 in the appendix of figures.

[410] See figure 110 ("Works by decade and country") in the appendix of figures.

[411] The script used for the calculation of significances and variance ratios is available at https://github.com/cligs/scripts-nh/blob/master/analysis/sign.py. Accessed January 3, 2021. The data is not normally distributed, so the Mann-Whitney U test was used. The following p-values resulted: 0.02 for 1860s versus 1870s, 0.003 for 1860s versus 1880s, and 0.045 for 1880s versus 1900s. The other constellations had p-values above 0.05. The 1830s and 1910s were not included in the calculations because there are only 2 and 3 works for these decades, respectively.

[412] Calculating the ratio of text length variances for different pairs of decades shows that the differences in variance are biggest for the 1860s versus 1890s, the 1860s versus 1880s, and the 1860s versus 1890s. The difference in variance can be considered significant for the 1850s versus 1880s-1900s, the 1840s versus 1880s-1900s, and for all the constellations of the 1860s versus later decades. The differences in text length variance between the 1870s and the 1880s as well as the 1870s and 1890s are also significant, but the remaining ones are not. Variance ratios between 0.5 and 2.0 are considered similar, and values below 0.5 or above 2.0 as significantly different. The three biggest ratios of variance are 5.4 for the 1860s versus 1890s, 4.8 for the 1860s versus 1880s, and 4.0 for the 1850s

4.1.4 Editions

As already stated in the overview section on authors, an evaluation of the number of editions emphasizes the role that the works played in the (literary) society of their time and also how the works were anchored in time and place. As "expressions" and "manifestations" of the works, realizing and embodying their intellectual content (International Federation of Library Associations and Institutions (IFLA) 2009, 13), editions link the works to their socio-cultural, historical, and geographical background. In the section on works above, the first known editions served as placeholders to look at where and when the works were published. Obviouesly, editions also play a role beyond the first appearance of a work. In this chapter, all the editions that were collected in the bibliography Bib-ACMé are analyzed together.

In total, 1,220 editions that were published between 1830 and 1910 are included in Bib-ACMé. All the editions of the works contained in the corpus were considered, even though the full texts usually rely only on one specific edition. However, as explained above in the sections on the assignment of subgenre labels (see chapter 3.2.3 for the bibliography and 3.3.4 for the corpus), all available editions were evaluated for generic signals in order to determine the subgenre of a work. This was done in terms of metadata and paratexts of the editions. As a result, the corpus covers 498 editions, which is 41 % of the editions in the bibliography. Assessing the number of editions is especially interesting when they are compared to the number of works. What changes with this other perspective?

The number of editions per author was already shown in the overview chapter on authors above (chapter 4.1.2). Here, the number of editions per work is analyzed.[413] In Bib-ACMé, most novels were only published in one edition (582 works, i.e., 48 %), followed by 161 works (13 %) with two, 54 works (4 %) with three, 17 works (1 %) with four and 15 works (also 1 %) with five or more editions. The work with the most editions (10) is "Amalia" (1855, AR) by José Mármol, followed by "Clemencia" (1869, MX) by Ignacio Manuel Altamirano and "Anatomía del corazón" (1856, CU) by Teodoro Guerrero y Pallarés with seven editions each. These three works were all first published early. The first two are famous representatives of the nineteenth-century novel of their respective countries, while the third one is rather nameless from today's point of view. In the corpus, in comparison, works with just one edition are underrepresented (22 % of the works with only one edition in the bibliography), and works with more than one edition are overrepresented (42 %, 69 %, and 65 % of the works with two, three, and four editions in the bibliography, respectively). The numbers of editions show that the sample size of the corpus is larger in terms of editions than in terms of works, where it was about one-third, reconfirming that the corpus contains relatively more popular or successful works than the bibliography as a whole.

In the following, the distribution of editions over time is analyzed from three perspectives: by years, decades, and the period before or in and after 1880.[414] Some of the early years are not represented at all (1830, 1831, 1833, 1834, 1835, 1849, and 1852). These are the same years as in

versus 1890s. The 1830s and 1920s were not included in the calculations because of the low number of works in these two decades.

[413] See the corresponding figure 130 in the appendix of figures.

[414] See figures 131 to 133 in the appendix of figures.

the case of works, except for the year 1867, which now has one edition of the work "Anatomía del corazón". This work was first published in 1856 in Madrid and republished in La Habana in 1867, inter alia. The three years with the most editions are 1886, 1887, and 1903, which corresponds to the years with the most works.

The distribution of editions over the decades is comparable to the development of the number of works, only that the absolute numbers are higher in the case of the editions. Their number increases steadily from the 1830s to the 1870s and then sharply in the 1880s, where it reaches the top and then remains high in the next decades. Apart from the 1830s with very low numbers, the share of editions in the corpus is a bit above average in the early decades up to the 1870s and below average in the 1880s, 1900s, and 1910 (when compared to the bibliography). This results in a higher representation of the period before 1880 in the corpus.

These conditions are similar to the distribution of works over the two periods, which is unsurprising if almost 50 % of the novels only had one edition between 1830 and 1910. However, for the number of works, the difference between the period before 1880 and in or after 1880 amounted to 9 % and for editions only to 5 %, meaning that relatively more works with several editions published in or after 1880 are included in the corpus.

Another point of interest is to see how many editions were published by country and also in which cities the editions appeared. In these analyses, editions for which several places of publication are given on the title pages are counted several times.[415] In the bibliography, most editions are published in Mexico, followed by Argentina, and in the corpus, it is the other way around. In contrast to the corresponding overview for works, where only the places of the first publication were considered, for all the editions, the third most important country of publication was Spain and not Cuba, both in the bibliography and the corpus. This means that many works that were first published in Mexico, Argentina, or Cuba, were republished in Spain.

The most important cities of publication were the three capitals Mexico (34 % of all the editions), Buenos Aires (33 %), and Havana (8 %). In the corpus, Buenos Aires outranks Mexico. Given that the corpus contains more works associated with Mexico (42 %) than with Argentina (39 %), this means that the Mexican works contained in the corpus were more often published elsewhere than Argentine works, be it in another Mexican city or in another country. Right after the three capitals, the Spanish cities Barcelona (6 % of the editions) and Madrid (4 %) follow, and Paris (3 %) occupies the sixth rank. These numbers and also the whole list of cities illustrate that the publishing of the novels was centralized to a high degree and that European metropolises also played a role in the distribution of the novels. On the other hand, there is also a long list of individual publication places, showing a greater diversity of local and foreign publishing activity, if not quantitatively, at least qualitatively. There are, for instance, 32 different Mexican, 11 Argentine, and 10 Cuban publication places.

To summarize, comparing the number of editions in the corpus and the bibliography to the number of works contained in both, the corpus involves relatively more editions, meaning that the works in the corpus were republished more often than the average work in the whole

[415] See figure 134 for the proportions of editions by country and figure 135 for the number of editions published in different cities in Bib-ACMé and Conha19. Both figures can be found in the appendix of figures. In the figure on cities, only those that appear at least twice in the bibliography are included.

bibliography. Relatively, the numbers of editions over time are comparable to the numbers of works. Regarding the number of editions, the period before 1880 is a bit better represented in the corpus than the period after that year, but the difference between the two periods is smaller than in the case of the works. Considering the countries and places of publication of the editions, Spain plays a bigger role when all the editions are considered and not only the first editions of the works. The main places of publication are the three capitals of the countries selected for the bibliography and corpus, followed by European cities and a whole range of other publication places of minor importance.

4.1.5 Subgenres

This chapter gives overviews of the subgenres to which the novels in the bibliography and the corpus are assigned. According to the model of subgenre terms developed in chapter 3.2.3 above, a distinction is made between explicit subgenre signals that are directly mentioned in titles and other paratexts of the novels and implicit signals that were inferred from them. Furthermore, labels that are signaled (explicitly or implicitly) are differentiated from labels that were assigned to the novels by literary historians. In addition and cross to the above distinctions, the subgenre labels are organized into several semiotically justified levels (theme, current, identity, and several modes of the medial and syntactic realization and the communicational frame). If not otherwise stated, multiple assignments of subgenre labels are all counted in.

4.1.5.1 Explicit Signals, Implicit Signals, and Literary-Historical Labels

In the bibliography, 622 novels (i.e., 75 % of all the novels) carry an explicit (sub)generic signal of any kind, while 207 novels do not carry any explicit signal at all. In the corpus, 204 novels (80 % of all the novels in the corpus) have an explicit signal. The explicit label "novela" is carried by 404 (49 %) of the novels in the bibliography and by 134 (52 %) of the novels in the corpus. How can this be interpreted? Either the novel, as defined here, is a genre that is so self-evident that its representatives do not need the explicit denomination to be recognized, or it is so vaguely defined that as many other texts are covered by it. However, as information about almost all the works in Bib-ACMé and Conha19 was retrieved from relevant bibliographies and literary histories of the novel, the former aspect is more plausible.[416] In what follows, the proportions of works in the bibliography with and without the explicit label "novela" are analyzed by decade.[417] Up to the 1870s, more than half of the works in the bibliography carry the label "novela". From the 1880s on, this label becomes rarer, suggesting a change in the conventions of labeling the

[416] In a future analysis, it could be interesting to analyze if the presence or absence of the label "novela" corresponds to different subtypes of the genre.

[417] See the corresponding figure 136, in which a series of donut charts is given, in the appendix of figures.

works over time.[418] However, both types of works, those with and without the explicit label "novela", were present in all the decades.

In total, 108 different explicit subgenre labels are found in the bibliography. Although these labels are called "explicit" here, they do not correspond exactly to the historical denominations used to mark the novels because the values were normalized in order to be comparable. Part of this normalization is that compound labels were split up, and each part was marked up separately.[419] The top 20 of these regularized explicit labels in the bibliography are analyzed here and compared to the corresponding labels in the corpus.[420] The general label "novela" is the most frequent one. The other top explicit labels are of different kinds. Some labels are directly related to the themes of the novels and are recognizable as subgenres of the novel: "novela histórica" (on rank 2 in the bibliography), "novela de costumbres" (rank 5), "novela social" (rank 13), and "novela policial" (rank 17). Labels referring to the linguistic and cultural identity of the novels also recur in this top list: "novela original" (rank 3), "novela mexicana" (rank 4), "novela cubana" (rank 9), "novela nacional" (rank 11), "novela argentina" (rank 15), and "novela americana" (rank 16). Of the remaining labels, several are (not exclusively, but often) related to different kinds of historical novels: "episodios" (rank 6), "memorias" (rank 7), "leyenda" (rank 8), and "historia" (rank 12). The labels "escenas" (rank 18) and "cuadros" (rank 19) are often connected to novels of customs. Interestingly, also labels designating other genres, such as "drama" (rank 10), "cuento" (rank 14), and "ensayo" (rank 20), are among the top labels for the novel.

For the corpus, the ranks of explicit labels are similar, but there are also a few differences.[421] Labels that are in the top 20 for the corpus but not for the bibliography are "estudio" (rank 15), "novela realista" (rank 18), "crónica" (rank 19), and "novela militar" (rank 20). On the other hand, labels that are in the bibliography top 20 list but not in the corresponding corpus list are "novela nacional", "novela policial", "escenas", and "ensayo". When the first ranks of the bibliography and corpus lists are compared, differences are, for example, that the number of "novelas históricas" is almost equal to the "novelas de costumbres" in the corpus, whereas it is around twice as high in the bibliography. Furthermore, "novela cubana" is on rank 6 in the corpus compared to rank 9 in the bibliography.

Analyzing the top explicit subgenre labels brings to light several characteristics of the novels in the bibliography and the corpus. First, the most prominent explicitly marked subgenre is the historical novel, both according to the number of occurrences of the literal label and also based on several other subgenre labels related to it, which leads to the conclusion that the most prominent subgenres are historical novels. Second, there was an evident need to explicitly mark the linguistic, cultural, or national identity of the novels. As there are so many different kinds of *identity labels*, it is of interest to check how many novels carried such labels. In the bibliography,

---

[418] The analysis is based on the information if the works first published in the respective decades ever carried the label "novela" between 1830 and 1910 because the works are dated according to their first known edition, but their subgenre labels are collected for all the editions that were published in the chronological frame of this study. This introduces a certain fuzziness concerning the anchoring of subgenre labels in time, so the effect of change might even be stronger.

[419] See chapter 3.2.3 above for details.

[420] See figure 137 in the appendix of figures. The top 20 positions were calculated from the bibliography's point of view.

[421] See figure 138 in the appendix of figures.

272 novels (33 %) had an identity label, and in the corpus, 100 novels (39 %). As with the general label "novela", also here the question arises if the use of identity labels depends on the period of publication of the novels.[422] A trend becomes visible over the decades, as the number of novels carrying identity labels decreases continuously. A third point that can be drawn from the top frequent explicit labels is which subgenres are more important in the corpus than in the bibliography. These are the novels of customs, realist, and naturalistic novels linked to the labels "novela realista" and "estudio" and also Cuban novels.

If implicit signals are included in the evaluation, the range of subgenres broadens because some subgenres are never marked explicitly. This also means that the assignment of subgenre labels gets more interpretive. For 511 works (62 %) in the bibliography and 207 works (81 %) in the corpus, implicit signals were found.[423] If this is added to the explicit information, subgenre signals were recognized for 738 novels (89 %) in the bibliography and 254 novels (99 %) in the corpus. In the following, the top 20 subgenre labels for Bib-ACMé and Conha19 are analzyed, taking explicit and implicit signals into account together.[424]

When also implicit signals are included, more subgenres related to the themes of the novels and to literary currents enter the top positions: in the bibliography, rank 2 is still occupied by the "novela histórica" followed by the primarily thematic labels "novela sentimental" (rank 3), "novela social" (rank 6), "novela de costumbres" (rank 7), "novela política" (rank 14), "novela criminal" (rank 17), and the "novela de la ciudad" (rank 20). Labels relating to literary currents are "novela romántica" (rank 4), "novela naturalista" (rank 12), and "novela realista" (rank 15). This shift could be expected because the assessment of implicit signals requires an interpretation frame, and subgenres focusing on theme and literary currents are the dominant perspectives in literary histories.

Comparing the top 20 signals of Bib-ACMé to Conha19, again, some subgenres gain more weight in the corpus: the "novela sentimental" moves from rank 3 to 2, the "novela de costumbres" from rank 7 to 4, the "novela gauchesca" enters the top list on rank 20, the "novela naturalista" from rank 12 to 8, and the "novela realista" from rank 15 to rank 10. The "novela romántica", on the other hand, gets less important and moves from rank 4 to 6, the "novela histórica" switches from rank 2 to rank 3, and the "novela de la ciudad" is not part of the top list anymore.

When only statements made by literary historians are evaluated, the picture changes even more. Literary-historical assignments were recorded for 433 works (52 %) in the bibliography and 224 works (88 %) in the corpus. That the proportion of works with literary-historical labels is much higher in the corpus is certainly because the corpus contains more works that are better known and researched. There are 34 different literary-historical labels in the bibliography and 32 different ones in the corpus. As with the explicit paratextual signals, also here, the labels were homogenized to be comparable and do not correspond to literal statements in every case.

In the bibliography, the label most often assigned is "novela romántica", followed by "novela histórica" and "novela realista".[425] Labels that were not included in the top ranks of generic signals,

---

[422] See figure 139 in the appendix of figures.

[423] That more implicit subgenre signals were found for the novels in the corpus also results from the fact that more paratextual information was evaluated for them (title pages, prefaces, etc.).

[424] See the figures 140 and 141 in the appendix of figures.

[425] See figures 142 and 143, which display the top 20 subgenre labels assigned to the works by literary historians.

but are top literary-historical labels, are “novela indigenista”, “novela abolicionista”, “novela modernista”, “novela de aventuras”, “novela verista”, “crónica”, “novela científica”, and “novela satírica”. Here, different critical perspectives on nineteenth-century Spanish-American novels are introduced, for example, specific topics and socio-cultural concerns (“novela indigenista”, “novela abolicionista”) or particular literary currents (“novela modernista”, “novela verista”), but also general generic subcategories that are not specified culturally (“novela de aventuras”, “crónica”, “novela científica”, “novela satírica”).

The top ranks of literary-historical labels in the corpus are of a similar kind but in part differently ordered. The “novela romántica” is also most important in the corpus. Labels with more weight than in the overall bibliography are “novela social”, “novela de costumbres”, “novela naturalista”, and “novela abolicionista”, for instance, and labels that are less relevant in the corpus are, for example, “novela histórica”, “novela realista”, “novela criminal”, and “novela gauchesca”. The different top ranks illustrate where there have been shifts in the composition by subgenre due to the selection of works for the corpus. Some were made on purpose, for example, the inclusion of more novels of customs as a counterbalance to the great majority of historical novels or a preferred inclusion of Cuban novels to strengthen the smallest country group. Others depend on the availability of the novels. Many crime and gaucho novels can be classified as low-prestige novels and are not yet readily available in digital format.

The differences between explicit subgenre signals, implicit signals, and literary-historical subgenre assignments underline that the views on what a subgenre of the novel is differ considerably, depending on the practice and the purpose of the labeling. There are only some intersections. The labels assigned to the historical editions probably served a number of functions:

- to clearly mark a novel as an instance of a known and popular subgenre of the time (e.g., “novela histórica”, “novela de costumbres”),
- to associate a novel with a certain literary current or school by using terms that were loaded with such semantics (e.g., “estudio”, “cuadros”),
- to signal that the language and content of a novel are autochthonous (e.g., “novela original”, “novela americana”),
- or to give the novel other attributes inscribing it in the general history of genres, playing with terms, or attracting attention towards its particular way of (re)presentation (e.g., “leyenda”, “episodios”, “memorias”, “drama”), among other possibilities.

Literary-historical assignments of subgenre labels, on the other hand, usually have the primary function of classifying the works in question according to established critical schemes. They are much more uniform and concentrate on aspects of style and content that allow for connecting the novels with subgenres

- that are rich in tradition (e.g., “novela histórica”, “novela sentimental”, “novela de aventuras”),
- that are associated with certain literary periods and currents (“novela romántica”, “novela realista”, etc.),
- or that are of special critical interest regarding the novelistic production of a certain cultural-historical space (“novela gauchesca”, “novela indigenista”, “novela abolicionista”).

Only in part do these two practices of labeling individual novels with (sub)generic terms overlap.

The top lists of subgenre labels show that the labeling is more systematic in the literary-historical approach: there are not just a few thematically oriented labels but a whole range of them, and not just one or two labels relating to literary currents but labels for all the currents that have been recognized as relevant in the nineteenth century. The historical practice, on the other hand, is much more selective. "Novelas históricas" and "novelas de costumbres" were commonly explicitly named, but "novelas sentimentales", for example, only exceptionally. In addition, there is historical evidence in the bibliography for the labels "novela realista" and "novela naturalista", but not for the "novela romántica". This does not prove that there was no awareness of the latter, but apparently, some types of novels were more part of a general, given generic pool than others.

4.1.5.2 Discursive Levels of Subgenre Labels

Even though literary-historical approaches to subgenres aim to systematize the field, in sum, the resulting set of labels is still a conglomerate of different perspectives on the novels, even if to a lesser degree than the historical labels. This becomes very clear when the same novel is labeled with several terms at the same time. The works of the Mexican writer Victoriano Salado Álvarez, for example, have been classified both as historical and realist novels (Fernández-Arias Campoamor 1952, 84–85; Read 1939, 293–294) and some works by the Cuban Cirilo Villaverde both as novels of customs and romantic novels (Remos y Rubio 1935, 166–180; Suárez-Murias 1963, 23–24). This makes it difficult for a stylistic analysis of subgenres that aims to select and compare subsets of novels from a corpus. Therefore, the explicit, implicit, and literary-historical labels have been sorted according to a system of discursive categories, as explained above in chapter 3.2.3, resulting in several sets of labels that belong to different discursive levels but whose comparison is more meaningful on each level. It has to be reminded, though, that this system is artificial. In what follows, overviews of the different sets of subgenre labels based on the discursive model of subgenre terms are given for the bibliography and the corpus. First, summaries of how many labels there are on which levels are presented.[426]

How many different labels on the various discursive levels are there in the whole bibliography? In total, there are 124 different terms. Most of them belong to the thematic group (39 %), followed by the mode the novel is represented in linguistically or narratively (36 %), the cultural-geographical and linguistic identity of the novel (20 %), the medium that the novel uses (11 %), the intention of the author or narrator (10 %), the relationship between the novel and reality (8 %), the literary current of the novel (5 %), and the attitude the author or narrator has towards what is represented in the novel (4 %). The diversity of thematic labels is not surprising, and that there is only a small set of labels related to literary currents is expectable as the number of different currents is limited. The broad range of labels referring to the mode of representation

[426] See figures 144 and 145 in the appendix of figures. In figure 144, the number of different labels related to the realization of the discursive act is smaller than the sum of labels related to its three subgroups ("semantic", "syntactic", and "medium") because the same label can be associated with several levels. For example, the labels "novela filosófica" and "novela psicológica" are categorized both as realization/semantic/theme and as realization/syntactic/mode.representation, because the terms point both to certain themes (e.g., general considerations about the meaning of life in a philosophical novel or the focus on personal, emotional states of characters in a psychological novel) and also to a certain way of representation (e.g., an argumentative style in philosophical novels and an introspective narrative style in psychological novels).

and also to the identity was not expected, though, as these aspects are usually not focused on in studies of subgenres of the Spanish-American novel.

How many different subgenre labels are there in Conha19? In the corpus, there are 90 different subgenre labels, which are distributed similarly over the different levels when compared to BibACMé.[427] 44 % of the different labels are thematic, 36 % are related to the mode of representation, 14 % to the identity, 13 % to the medium, 10 % to the intention, 9 % to the mode of reality, 7 % to the literary current, and 3 % to the attitude.

The significance of the different discursive levels changes to a certain degree if not the number of *different* labels in each category is considered, but the *overall number* of labels belonging to them. How often have such labels been assigned?[428] In total, on the different levels, 3,193 labels were assigned to the novels in the bibliography and 1,317 to the novels in the corpus.[429] Most labels are of the thematic type (38 %), followed by the mode of representation (23 %), literary currents (15 %), the mode of reality (10 %), identity (10 %), the attitude (2 %), medium (2 %), and intention (1 %). So in terms of quantity (instead of diversity), labels related to literary currents and the reality mode are more important, while labels associated with the mode of representation, with identity, the medium, intention, and attitude are less relevant. The picture is similar for the corpus, only that literary currents have a bit more weight than modes of representation.[430] In Conha19, 40 % are thematic labels, 20 % are related to literary currents, 18 % to the mode of representation, 8 % to the identity, 8 % to the mode of reality, 2 % to the attitude, 2 % to the medium, and 1 % to the intention. Clearly, thematic labels are most important both when the number of kinds and the overall number of assignments is considered.

The impression that the overall top lists of labels give is also confirmed when one differentiates between different sources of labels. All in all, there are 108 different explicit subgenre labels in the bibliography and 34 different literary-historical labels. The corpus includes 70 different explicit labels and 32 different literary-historical labels. This means that the diversity of explicit labels doubles the diversity of literary-historical labels in the corpus and is three times higher in the bibliography. Regarding the overall number of labels, 1,669 explicit labels and 1,120 literary-historical labels were assigned to the works in the bibliography, and 564 explicit and 686 literary-historical labels were assigned to the works in the corpus. So on average, a literary-historical label is assigned more often than an explicit label.[431] The following table 23 summarizes the importance of the different discursive levels for explicit versus literary-historical labels in the bibliography.[432]

---

[427] The visualization for the corpus is available at https://github.com/cligs/data-nh/blob/master/corpus/overview/subgenres-labels-number-corpus.html. Accessed 15 September 2020.

[428] See figure 145 in the appendix of figures. When the overall number of labels was determined for the various categories, identical labels stemming from different kinds of sources were only counted once for each novel (e.g., if a novel was explicitly labeled as "novela histórica" and also classified as such by literary historians).

[429] Again, on each level, if the same label is assigned to a work by different sources, it is only counted once.

[430] The visualization for the overall number of labels in the different categories in the corpus is available at https://github.com/cligs/data-nh/blob/master/corpus/overview/subgenres-labels-amount-corpus.html. Accessed 15 September 2020.

[431] To determine the overall number of labels, they were counted on each discursive level, so a homonymic label on different levels is counted several times. On the other hand, if several literary-historical sources mentioned the same label for a work, it was only counted once per level.

[432] See the corresponding visualizations "subgenres-labels-number-explicit-bib.html", "subgenres-labels-amount-explicit-bib.html", "subgenres-labels-number-litHist-bib.html", "subgenres-labels-amount-litHist-bib.html" at https://github.com/cligs/data-nh/tree/master/corpus/overview. Accessed 15 September 2020.

Rank	Different explicit labels		Amount explicit labels		Different literary historical labels		Amount literary historical labels	
1	mode.representation	43	mode.representation	707	theme	25	theme	558
2	theme	37	theme	363	current	6	current	385
3	identity	25	identity	302	mode.representation	5	mode.reality	121
4	mode.medium	14	mode.reality	203	mode.reality	4	mode.attitude	35
5	mode.intention	13	mode.medium	48	mode.intention	2	mode.representation	10
6	mode.reality	9	mode.intention	28	mode.attitude	2	mode.intention	7
7	mode.attitude	4	current	11	mode.medium, identity	1	mode.medium, identity	2
8	current	2	mode.attitude	7	-	-	-	-

Table 23: Ranks of discursive levels of subgenre labels, explicit vs. literary historical (Bib-ACMé).

When differentiating by the type of source (explicit versus literary-historical labels), thematic labels are also generally important. Regarding the overall number of labels, the mode of reality also has some importance for both explicit and literary-historical labels. The relevance of the other levels depends more on the provenance of the labels. For explicit labels, the mode of representation and identity are prominent, and for literary-historical labels, in particular, the literary currents. The ranks do not change for the corpus, apart from minor shifts in the last positions.[433] Regarding the importance of the linguistic and cultural-geographical identity of the novels, it has to be said that this aspect is, of course, also a topic in the critical discourse about the novels, but in a different way than in the historical practice of labeling the works. For literary-historical studies, usually, the linguistic and cultural-geographical frame is set from the beginning on. Either it concentrates on one national space, e.g., "the Mexican novel", or on Spanish America or Latin America as a whole and then differentiating the novels by country. However, usually, no difference is made between one work or the other and the attribution to the cultural-geographical space is made based on general extra-textual parameters. On the other hand, in literary histories, works are also reviewed in terms of how their content reflects local or foreign realities, but usually in more general terms and without explicitly categorizing novels in that way. Furthermore, the question of the own and the foreign is discussed more in aesthetic categories.[434]

4.1.5.2.1 Theme

So far, the discursive model has served to give a general overview of what kind of subgenre labels are relevant in Bib-ACMé and Conha19. Proceeding to the individual levels and the actual labels, it becomes clearer what these subgenres are and which ones dominate from a quantitative point of view. The different levels are presented here in the order of their relevance for analyses of the novels in the corpus from a quantitative point of view. Starting with the thematic labels, of all the works in the bibliography, 695 (84 %) have such a label. In the corpus, all the works have been assigned at least one thematic label. There are 48 different thematic labels in the

---

[433] For the corpus, see the charts "subgenres-labels-number-explicit-corp.html", "subgenres-labels-amount-explicit-corp.html", "subgenres-labels-number-litHist-corp.html", "subgenres-labels-amount-litHist-corp.html" at https://github.com/cligs/data-nh/tree/master/corpus/overview. Accessed 15 September 2020.

[434] See, for instance, examples of statements on the relationship to the European literatures and questions of emancipation in Rössner's literary history of Latin America. About nineteenth-century Caribbean literature, he writes: "Das Streben nach Unabhängigkeit ist nicht nur eine politische Angelegenheit, es bestimmt auch das literarische Leben. Einerseits orientieren sich die karibischen Literaten des 19. Jhs. an europäischen, besonders spanischen und französischen Vorbildern […], andererseits bemühen sie sich darum, diese Modelle nicht einfach zu kopieren, sondern sich deren theoretisches Gedankengut mit Berücksichtigung all der Spezifika ihres amerikanischen Lebensraums kreativ anzuverwandeln" (Rössner 2007, 153). In this context, there are also discussions of individual authors and works: "José López Portillo y Rojas hingegen distanziert sich von französischen Vorbildern. In dem Vorwort zu seinem Roman La parcela (1898) weist er die ästhetische Wortkunst Flauberts oder der Brüder Goncourt ebenso zurück, wie er auch die Obszönität Zolas meiden möchte. Stattdessen bezieht er sich auf die Spanier Galdós und vor allem Pereda, deren Vorliebe für das naturverbundene Leben in der Provinz auch in Portillos Bauernroman ihren thematischen Niederschlag findet. Die Reserve dem französischen Kulturerbe gegenüber hat sich auch in dem Roman Fuertes y débiles (1919) erhalten. Das hier porträtiert porfiristische Gesellschaftssystem krankt daran, dass sich der französische Positivismus nicht einfach auf die mexikanischen Verhältnisse übertragen lässt und an den espíritus débiles der científicos scheitert" (Rössner 2007, 148).

bibliography, of which 18 are assigned to at least 10 works.[435] In the bibliography and in the corpus, as well, four thematic subgenres are predominant: the sentimental novel, the historical novel, the social novel, and the novel of customs. Especially in the corpus, also the political novel is significant. Comparing bibliography and corpus, the order and ratio of the top thematic subgenres are different. In Conha19, the social novel and the novel of customs are relatively more important, and the sentimental and historical novels are less relevant. However, the absolute amount of these four subgenres is more balanced in the corpus than in the bibliography.

An analysis of the sources for the top 18 thematic labels shows the different statuses they have as explicit historical subgenres, implicitly signaled subgenres, and literary-historically discussed subgenres.[436] As can be seen, the historical novel and the novel of customs are the most important historically explicit subgenres, followed by the "leyenda" and the "novela policial". Especially the sentimental novel is implicitly signaled. All four top subgenres play an important role as literary-historical subgenres. Although this list of thematic labels is already the result of a process of systematization, two kinds of one-to-many relationships persist: the same label can also play a role on other discursive levels (the "novela histórica", "leyenda", and "novela contemporánea", for example, are also among the labels referring to the relationship between the novel and reality) and each novel cannot just carry one, but several thematic labels. The latter is very often the case.[437] In the bibliography, 16 % of the novels do not have any thematic label, 45 % have just one, 23 % have two, 11 % have three, and 5 % have more than three thematic labels. In the corpus, all the novels have thematic labels: 38 % have just one, 33 % have two, 18 % have three, and 11 % have more than three different thematic labels. The difference between the bibliography and the corpus shows that the more is known about the works, the more differentiated (and less clear) their categorization is in terms of subgenres. Besides the one-to-many relationships, i.e., one work to many thematic subgenre labels, there are also relationships of overlap or inclusion among the different thematic labels. For example, the "novela abolicionista" can be considered a special type of social novel, and the "leyenda" is associated with historical content just as the "novela histórica".[438] Relationships of this kind also become visible in the most frequent combinations of thematic subgenre labels, which are listed in table 24 below.[439]

The most frequent combination of thematic subgenre labels both in the bibliography and the corpus is "novela de costumbres" and "novela social". Most top combinations are very frequent in both Bib-ACMé and Conha19, even if the ranks are not exactly the same. Combinations that

---

[435] See figure 146 in the appendix of figures, which shows how many works have these 18 labels in Bib-ACMé and how many works there are in Conha19 with the same label.

[436] See figure 147 in the appendix of figures. The different source types are stacked for each subgenre label to highlight their proportions, but the sums can be bigger than the ones of the number of works carrying the label because the same label can have various types of sources. This applies to all the following charts of subgenre label sources.

[437] See figure 148 in the appendix of figures.

[438] Such connections between different labels on the same level are listed in table 10 above (not exhaustively, but for some obvious cases).

[439] Combinations of labels that are at the same time part of combinations of more labels are counted each time (the combination of "novela de costumbres" and "novela social", for example, is also counted for works that have a combination of "novela de costumbres", "novela gauchesca", and "novela social"). Combinations with the same number of assignments are ordered alphabetically. The whole list of combinations is available at https://github.com/cligs/data-nh/blob/master/corpus/overview/subgenres-label-combinations-theme.csv. Accessed September 29, 2020.

Rank	Bib-ACMé		Conha19	
	Labels	**Number of assignments**	**Labels**	**Number of assignments**
1	*novela de costumbres*, *novela social*	58	*novela de costumbres*, *novela social*	48
2	*novela histórica*, *novela sentimental*	51	*novela sentimental*, *novela social*	31
3	*novela sentimental*, *novela social*	41	*novela de costumbres*, *novela sentimental*	30
4	*novela de costumbres*, *novela sentimental*	37	*novela histórica*, *novela sentimental*	26
5	*leyenda*, *novela histórica*	29	*novela de costumbres*, *novela sentimental*, *novela social*	16
6	*novela histórica*, *novela política*	18	*novela de costumbres*, *novela histórica*	14
7	*novela de costumbres*, *novela histórica*	16	*novela de costumbres*, *novela política*	12
8	*novela de costumbres*, *novela sentimental*, *novela social*	16	*novela histórica*, *novela política*	10
9	*novela histórica*, *novela indigenista*	15	*novela política*, *novela social*	9
10	*leyenda*, *novela sentimental*	13	*novela abolicionista*, *novela social*	8

Table 24. Top combinations of thematic subgenre labels.

are only in the top ten of the bibliography are highlighted in blue, and the ones that are only in the top list of the corpus are highlighted in orange. As can be seen, combinations including the "leyenda" and "novela indigenista" have more weight in the bibliography and combinations including the "novela política" and the "novela abolicionista" have more weight in the corpus. Clearly, combinations of the "novela sentimental" with other types are the most frequent ones that are not characterized by semantic overlap.

Quantity distributions as the ones described for thematic subgenre labels here (a few frequent groups and many infrequent ones) and complexities (multiple assignments and interrelationships) have consequences for a digital quantitative analysis that aims to analyze novels in terms of subgenre categories. First, for very small groups, it is hardly possible to achieve general results as they will be influenced very much by the particular characteristics of the few individual works that form the group. Furthermore, groups of very different sizes cause problems when evaluating the performance of a categorization task. In the case of a large majority, good results could simply be achieved by always choosing that group. To avoid that, the groups to compare would have to be balanced down to the smallest size by undersampling. Another strategy is oversampling, where instances of the smallest group are duplicated, which again can cause problems of overfitting

and a lack of generalizability.[440] This study will concentrate on the largest groups of novels that are represented in the corpus to avoid these problems as far as possible. In the case of thematic subgenres, these are sentimental, historical, and social novels, novels of customs, and political novels.

One way to solve the problem of multiple assignments is to decide on primary labels. For thematic labels and those relating to literary currents, this was done by choosing the label

- that was most prominent in the signals,
- that was most often assigned by literary historians or mentioned by them as the primary label,
- that corresponded to the same level of generality as the major groups,
- or that was most informative.

However, it was not possible to define only one criterion because the sources of the labels, the kind of different labels, and the amount of available information about the subgenres differ from case to case. By choosing a primary label, it becomes possible to separate groups of novels on a specific discursive level. It has to be reminded, though, that this is another step of interpretation and simplification. In the following, the results of choosing a primary subgenre for the thematic labels are discussed.

The process of choosing a primary label changes the order of the top thematic subgenres both in the bibliography and in the corpus.[441] In Bib-ACMé, the historical novel becomes more important than the sentimental novel, and in Conha19, also the historical novel rises from the fourth to the top position and changes place with the social novel. This shows that both the sentimental and the social are often secondary thematic aspects in the novels, but at the same time, they can also be primary concerns. The primary labels can be used as target values in a standard categorization task, keeping in mind that part of the works has secondary labels when evaluating the results. Another possibility to handle multiple assignments of subgenre labels on the same discursive level is to include them all in the analysis. For a categorization task, this would mean allowing multiple target values, as, for example, in a multi-label classification.[442] In both cases, one has to be aware of using subgenre labels that stem from multiple sources and that have been assigned to the works by people with different perspectives on them and different interests in them. The goal of such a categorization can only be to examine to what extent the collective labelings actually reflect characteristics of the texts, and to find out to which traits of the texts they correspond. If literary-historical sources are included, these already introduce an element of systematization, so the digital text analysis then aims to analyze in what way this system matches the one found based on a computational treatment of the texts. The collectivity of labelings also applies to explicit historical labels when the whole set of novels is concerned, and to a lesser degree also per work, for example, when there is a label mentioned on a title page and the same or another one in a foreword or introduction written by someone who is not the author or editor of the work. Usually, though, historical labels are not multiple because they stand for

[440] See, for instance, Alpaydin (2016, 37–41) on the issue of generalization in supervised learning approaches. See also Branco, Torgo, and Ribeiro (2015) on the challenges of imbalanced data distributions.
[441] See figure 149 in the appendix of figures.
[442] For overviews of multi-label learning methods, see Elkafrawy and Mausad (2015) and Madjarov et al. (2012).

a will to mark a novel as a representative of a certain subgenre and do not aim to describe the text systematically in all its aspects. Still, cases of multiple assignments also exist for historical labels, especially on different discursive levels, but also on the same ones. An extreme case is the novel "Los bandidos de Río Frío" (1892, MX) by Manuel Payno which has the subtitle "Novela naturalista, humorística, de costumbres, de crímenes y de horrores", including two thematic labels ("novela de costumbres", "novela de crímenes"), two labels of intention ("novela humorística", "novela de horrores") and one label indicating the literary current ("novela naturalista"). Another combination that occurs three times in the bibliography is that of "novela histórica" and "novela de costumbres".[443]

4.1.5.2.2 Literary Currents

Going on to another level of subgenre labels, the ones related to literary currents, 405 works (49 %) in the bibliography and 201 works (79 %) in the corpus have such labels, so compared to thematic labels, the coverage of works is lower in both the bibliography and the corpus. This is mainly because the assignment of a work to a literary current is, above all, a critical task, so there are fewer sources than for thematic labels, as the latter draw from explicit and implicit signals and literary-historical sources alike.

To how many works were the labels related to literary currents assigned in Bib-ACMé and Conha19?[444] In the whole bibliography, there are only six different labels of literary currents. The ranks of the subgenre labels related to literary currents are the same for the bibliography and the corpus. Of the works that are associated with a literary current, most were labeled as "novela romántica", followed by "novela realista" and "novela naturalista". Three other subgenre labels related to literary currents occur but do only play a minor role from a quantitative point of view: "novela modernista", "novela verista", and "novela clasicista". Comparing Bib-ACMé and Conha19, realist and naturalistic novels are relatively more important in the corpus than in the bibliography. Regarding the sources for the labels related to literary currents, literary-historical sources and implicit signals play the most important role.[445]

As with thematic subgenres, multiple assignments also occur with literary currents. All the novels that were labeled as naturalistic novels were also marked as realist novels here because Naturalism is considered a current that evolved from and is closely related to Realism, and there are literary-historical sources for Spanish-American novels that do not differentiate clearly between the two. On the other hand, realist novels are not automatically also considered naturalistic novels, so the label "novela naturalista" is a more specific marker. Furthermore, there are eight novels in the bibliography, seven of which are also in the corpus, with the labels "novela romántica" and "novela realista" in combination.[446] Three novels in the bibliography and two in

---

[443] This combination is found in the subtitles of the novels "Calvario y Tabor. Novela histórica y de costumbres" (1868, MX) by Vicente Riva Palacio, "Julia. Novela histórica y de costumbres" (1868, MX) by Manuel Martínez de Castro, and "Astucia. Novela histórica de costumbres mexicanas" (1865–1866, MX) by Luis Gonzaga Inclán.

[444] See figure 150 in the appendix of figures.

[445] See figure 151 in the appendix of figures.

[446] These are "Cecilia Valdés" (1882, CU) by Cirilo Villaverde, "Abismos" (1890, AR) by Manuel Bahamonde, "Amalia" (1891, MX) by José Rafael Guadalajara, "Los bandidos de Río Frío" (1892, MX) by Manuel Payno, "Angelina" (1893,

the corpus are labeled both as "novela romántica" and "novela naturalista".[447] Concerning the multiple labels for literary currents, it must be remarked that it was not always easy to evaluate the literary-historical sources because the influences of several different literary currents in the works are often mentioned. Very differentiated descriptions are not fully represented in the subgenre labels. Instead, the main labels that were mentioned in the literary-historical sources were collected, for instance, the heading of a chapter in which a novel is included or the dominant literary current to which it is attributed in the source.[448] Multiple labels were only assigned where different literary historians classified a novel differently or where works are clearly and repeatedly described as representatives of several currents.[449]

As the literary currents are related to literary periods, looking at the publication dates of the novels that were labeled as belonging to a certain current can contribute to clarifying the question of when the different currents dominated and replaced each other. To this end, the distribution of works in the bibliography over the years per current is analyzed here, considering the publication date of the first known edition.[450] The analysis confirms several aspects that have been highlighted in literary-historical literature regarding the periodization of Spanish-American novels belonging to specific literary currents. First, it can be confirmed that different currents that followed each other temporarily in Europe were in vogue simultaneously in Spanish America. For instance, the majority of the naturalistic novels were published at the same time as the realist novels, approximately between 1880 and 1900.[451] Furthermore, the *novela modernista* came up approximately at the same time as the other currents of the late nineteenth century, around 1890. That the Romantic novel was a phenomenon that persisted during the whole century is also confirmed by evaluating the publication dates of the novels that have been labeled as romantic.[452] The earliest works classified as *novela romántica* were published in 1836, and the latest one in

---

MX), "Historia vulgar" (1904, MX), "La Calandria" (1890, MX), and "Los parientes ricos" (1901, MX) by Rafael Delgado. The novel "Abismos" is not part of the corpus.

[447] These are again "Abismos" (1890, AR) by Manuel Bahamonde, "La Calandria" (1890, MX) by Rafael Delgado and "Los bandidos de Río Frío" (1892, MX) by Manuel Payno.

[448] Writing about novels of customs, for example, Fernández-Arias Campoamor (1952, 56), states: "Los novelistas románticos que fueron costumbristas constituyen el puente tendido entre el romanticismo y el realismo [...] el costumbrismo como inclinación extensa y generalizada se inicia en el romanticismo". However, as he subsumes the section "Costumbristas" under "El Romanticismo", the novels mentioned in that chapter are interpreted as having been labeled as romantic novels by him here.

[449] This was the case, for example, with the works of Rafael Delgado, which have been described as both romantic and realist (Gálvez 1990, 105–106; Varela Jácome [1982] 2000, sec. 2.1.3).

[450] See figure 152 in the appendix of figures.

[451] There are even earlier cases of naturalistic novels: "La familia Unzúazu" (1868, CU) is the earliest novel that has been classified as "novela naturalista".

[452] Varela Jácome, for instance, comments on the delayed absorption of the romantic current in Spanish America and the chronological overlap of different literary currents: "La novela romántica no se aclimata en Hispanoamérica hasta el año 1846. Esto significa un claro asincronismo, con respecto a la narrativa de Europa y Estados Unidos, debido a la conflictividad ideológica y la carencia de modelos culturales idóneos. [...] Al margen de los obstáculos históricos, se produce, con una secuenciación discontinua, la introducción de modelos narrativos foráneos. [...] La novela indianista, iniciada muy temprano, en 1832, con Netzula, de Lafragua, se desarrolla en estratificación con los otros metagéneros románticos; en su época culminante, con Cumanda (1871), de Mera, se superpone sobre la narrativa de tendencia realista, a las últimas manifestaciones del ciclo coinciden, incluso, con la incorporación de las técnicas naturalistas" (Varela Jácome [1982] 2000, sec. 1.1.3).

1905.[453] The majority of the romantic novels was published between 1860 and 1886, though. The simultaneous publication of romantic and realist novels can also be attested.[454] Moreover, the 1880s, which have been described as a turning point regarding the prevalence of literary currents (Rössner 2007, 200; Varela Jácome [1982] 2000, sec. 3), can be confirmed as the decade in which the proportions of romantic novels versus novels of other literary currents shifted. However, based on Bib-ACMé, this shift can be dated more precisely here. Three-thirds of the romantic novels were published up to 1886, and three-thirds of the realist and naturalistic novels from 1888 and 1887 onwards, respectively. So instead of the year 1880 itself, 1887 is the year with a quantitative move from romantic to post-romantic currents.

Concerning categorization tasks, the literary-historical labels are suitable for standard classification because the number of different classes is manageable, and apart from the realist-naturalistic relationship, multiple labels are the exception. At least the groups romantic versus realist-naturalistic are both represented with many members. One useful application would be to train a model with the novels in the corpus that have labels related to literary currents and use the model to also label the 55 works for which no such label was found.

4.1.5.2.3 Mode of Representation

The next level of subgenre labels analyzed is related to the mode of representation. In Bib-ACMé, 556 works (67 %), and in Conha 19, 188 works (73 %) have a label related to the representational mode. Compared to the previously examined levels, in Bib-ACMé, these are fewer works than those with thematic labels but more than with labels related to literary currents. For Conha19, the proportion of works with a label related to the mode of representation is comparable to those with a label related to literary currents but less than the proportion of works with thematic labels. In the bibliography, 45 different labels belong to this level. Of these, 17 are assigned to at least five works.

The distribution of the top labels on this discursive level is quite different from the one of literary currents and themes, though, because the general label "novela" is by far dominating.[455] It is assigned to 49 % of the works in the bibliography and to 52 % of the works in the corpus. In the bibliography, also the labels "memorias" (assigned to 7 % of the works), "episodios" (7 %), "drama" (3 %), and "historia" (3 %) have a certain, although minor importance. In the corpus, labels that are assigned to at least 3 % of the works are "episodios", "drama", "memorias", "cuento", "historia", and "cuadros", but compared to the bibliography, "memorias" and "episodios" are underrepresented.

---

[453] Namely "Camila o la tiranía de Juan Manuel de Rosas" (AR, 1836) by Agustín Fontanella, "Sepulcros blanqueados" (MX, 1836) by Juan Antonio Mateos, and "Lina Montalván, o el terremoto que destruyó el Callao y la ciudad de Lima en 1746" (AR, 1905) by José Victoriano Cabral.

[454] The earliest novel classified as *novela realista* is "El negro Francisco" (1839, CU) by Antonio Zambrana y Vázquez. Even if this is an outlier, also the second and third quartiles of realist novels lie within the scope of the romantic novels, as the plot in the appendix of figures shows.

[455] See figure 153 in the appendix of figures.

Rank	Bib-ACMé		Conha19	
	Labels	Number of assignments	Labels	Number of assignments
1	*novela*, *memorias*	39	*novela*, *memorias*	6
2	*novela*, *episodios*	33	*novela*, *episodios*	5
3	*novela*, *episodios*, *memorias*	19	*novela*, *cuadros*	4
4	*novela*, *historia*	14	*novela*, *crónica*	3
5	*novela*, *historia*, *episodios*	7	*novela*, *historia*	3

Table 25. Top combinations of subgenre labels related to the mode of representation.

The sources for the labels referring to the representational mode are almost exclusively explicit historical signals.[456] Only the terms "memorias" and "crónica" were also part of the literary-historical discussion of the novels.

On this level of subgenre terms, multiple assignments occur as well. In the bibliography, 123 works – i.e., 15 % of all the works in the bibliography and 22 % of the works with labels on this level – have combinations of several labels related to the mode of representation. In the corpus, there are 42 works with more than one label of this kind, which corresponds to 16 % of all the works in the corpus and 22 % of the corpus works with labels on this level. The top 5 combinations in the bibliography and the corpus are summarized in table 25.[457]

In the bibliography, even the triple combinations "novela-episodios-memorias" and "novela-historia-episodios" are found repeatedly because they occur in series of historical novels, in particular the "Episodios nacionales mexicanos" by Enrique Olavarría y Ferrari. In the corpus, on the other hand, the combinations "novela-cuadros" and "novela-crónica" are more important.

All in all, only the presence or absence of the label "novela" is meaningfully analyzable with a quantitative approach. Beyond that, the wide range of different labels related to the mode of representation lends itself more to a qualitative analysis of the respective novels.

4.1.5.2.4 Mode of Reality

In total, there are ten different subgenre labels in the bibliography that involve the relationship of the text to reality. In Bib-ACMé, 279 works (34 %) have such a label, and in Conha19, 95 works

[456] See figure 154 in the appendix of figures.

[457] Combinations of labels that are at the same time part of combinations of more labels are counted each time (the combination "novela-memorias", for example, is also counted for works that have "novela-episodios-memorias"). The whole list of combinations is available at https://github.com/cligs/data-nh/blob/master/corpus/overview/subgenres-label-combinations-mode-representation.csv. Accessed September 29, 2020.

(37 %), so this discursive level is quantitatively less relevant than thematic labels, labels related to literary currents, and labels related to the mode of representation.

As with subgenre labels related to the mode of representation, also the labels related to the mode of reality are quantitatively dominated by a single label, the "novela histórica".[458] In Bib-ACMé, 29 % of all the works have that label, and in Conha19, 32 %. In the bibliography, also the "leyenda" plays a role, as there are 44 works with that label, which corresponds to 5 % of all the works. On the other hand, in the corpus, the label "leyenda" is underrepresented. There are only a few instances of the other labels associated with the mode of reality. In addition, the variety of labels in this group is much smaller than for the mode of representation.

The sources for the different labels related to the mode of reality are mainly explicit signals, but for the historical novel also implicit signals and literary-historical sources are important. Other labels that are a subject of discussion in literary histories are the "novela científica" and the "novela de misterio".[459]

Clearly, in the novels, the historical perspective is the most important aspect of the relationship between the text and reality, at least when subgenre labels are evaluated. On this discursive level, only the opposition of historical versus non-historical novels is suited for a comparative, quantitative analysis of subgenres.

4.1.5.2.5 Identity

In the bibliography, there are 25 different subgenre labels related to linguistic, geographical, and socio-cultural identity. These were assigned to 273 works (33 %) in Bib-ACMé and 101 works (39 %) in Conha19 so that the share of works having such a label is comparable to the ones with a label related to the mode of reality. Nevertheless, the variety of identity labels is larger, and this level is less dominated by a single label.[460]

The most important identity label is the general term "novela original", carried by 113 (14 %) of the works in the bibliography and 36 (also 14 %) of the works in the corpus. It is followed by the labels related to the three selected countries ("novela mexicana", "novela cubana", "novela argentina"), by other general labels ("novela nacional", "novela regional"), and by labels referring to the American continent ("novela americana", "novela criolla", "novela india"). Among the various identity labels of minor importance, there are several related to the countries' capitals ("novela bonaerense", "novela porteña", "novela habanera")[461], to specific regions or cities in Mexico or Cuba ("novela yucateca", "novela suriana",[462] "novela tapatía",[463] "novela de Tabasco", "novela camagüeyana"), and to Mexican indigenous people ("novela mixteca", "novela azteca"). Furthermore, there are references to European regions and culture ("novela romana", "novela franco-argentina", "novela siciliana", "novela kantabro-americana", "novela andaluza"). Comparing the bibliography and the corpus, the label "nacional" is underrepresented in the corpus, while

[458] See figure 155 in the appendix of figures.
[459] See figure 156 in the appendix of figures.
[460] See figure 157 in the appendix of figures.
[461] In the case of the "novela mexicana", it cannot be decided if it refers to the country or the capital.
[462] According to the "Diccionario de la lengua española" of the Spanish Royal Academy, "suriana" means "coming from the south of Mexico" (Real Academia Española (RAE) 2023b).
[463] "Coming from Guadalajara" (Real Academia Española (RAE) 2023c).

the labels referring to the three countries of Argentina, Mexico, and Cuba are overrepresented. Apart from the "novela regional", the sources of all the identity labels are explicit signals.[464]

Combinations of identity labels are not very frequent. Only the label "episodios nacionales mexicanos", containing both the identity label "nacional" and "mexicano", occurs 24 times in the bibliography and two times in the corpus. The label "novela original" is also combined with other identity markers, but only a few times. There are two "novelas cubanas originales", one "novela americana original", and one "novela mexicana original", both in the bibliography and the corpus.[465]

In a quantitative approach, the subgenres related to the linguistic, geographic, and socio-cultural identity could be analyzed in the following setups:

- *novela original* versus novels not carrying that label
- labels referring to the American context versus novels not carrying such a label[466]
- labels referring to the three countries of interest versus novels not carrying such a label[467]

It is suggested here that one could concentrate on the most frequent identity label, "novela original", and analyze the other labels as groups combining several individual labels that refer to a similar spatial context. In the corpus, a group of 36 "novelas originales", 67 novels with a label referring to the American context, and 30, 22, and 9 novels with a label related to the Mexican, Cuban, and Argentine context, respectively, can be compared to the other novels in the corpus not carrying such labels.[468]

4.1.5.2.6 Medium

Fourteen different subgenre labels in Bib-ACMé refer to a medial aspect of the novel. In the bibliography, only 47 novels (6 %) carry a label of this group, and in the corpus, only 23 novels (6 %), which is marginal from a quantitative point of view.

In the bibliography, the three most frequent medial labels are "escenas" (12 novels), "cuadros" (11 novels), and "páginas" (8 novels).[469] In the corpus, they are "cuadros" (8 novels), "páginas" (3 novels), and "esbozos" (also 3 novels). Some of the subgenre labels related to medial aspects are connected to novels of customs but also to social, realist, and naturalistic novels, as in the following titles:

- "El guajiro. Cuadro de costumbres cubanas" (1842, CU) by Cirilo Villaverde
- The series of short novels "Del natural. Esbozos contemporáneos" (1889, MX) by Federico Gamboa

---

[464] See figure 158 in the appendix of figures.

[465] See the file https://github.com/cligs/data-nh/blob/master/corpus/overview/subgenres-label-combinations-identity.csv, which lists the combinations of identity labels in Bib-ACMé and Conha19. Accessed September 29 2020. Combinations of labels that are at the same time part of combinations of more labels are counted each time (the combination "novela cubana-novela original", for example, is also counted for the work that has the combination "novela cubana-novela original-novela regional").

[466] The labels "novela kantabro-americana" and "novela franco-argentina" are assigned to the group of the American context.

[467] The label "novela franco-argentina" is assigned to the group of Argentine novels.

[468] See figure 159 in the appendix of figures.

[469] See figure 160 in the appendix of figures.

- "Escenas populares. Cuadros vivos de la clase ínfima del pueblo mexicano. Novela original de costumbres" (1901, MX) by Pablo Zayas Guarneros
- "La sociedad y sus víctimas. Escenas bonaerenses" (1902, AR) by Matías Calandrelli

Others are associated with sentimental novels, for example, "Amalia. Páginas del primer amor" (1891, MX) by José Rafael Guadalajara or "Páginas íntimas" (1895, MX) by Manuel Blanco. However, there are also historical or political novels with labels referring to a medial aspect, for instance, "Campaña y Guarnición. Escenas de la vida militar" (1892, AR) by E. Mayer, "El Señor Gobernador. Breves apuntamientos sobre cosas nacionales del siglo pasado" (1901, MX) by Manuel H. San Juan, or "Vía Crucis. Páginas de ayer" (1910, CU) by Emilio Bacardí Moreau. The sources of the labels associated with medial aspects are almost exclusively explicit signals.[470]

Because the frequency of these labels is low, especially in the corpus, they are not examined further in a quantitative setup but only considered for interpreting the results of other analyses.

4.1.5.2.7 Attitude

Subgenre labels related to the attitude the author or narrator has towards what is presented in the novel are not frequent. In the bibliography, five different labels of this level were found. In Bib-ACMé, they are assigned to 57 (7 %), and in Conha19, to 32 (13 %) of the works.[471]

The main subgenre in this group is the political novel, which is at the same time a thematic subgenre label. In the bibliography, there are 51 political novels, and in the corpus, 28. Besides the political novel, also the "novela satírica" is a label that has been assigned to novels by literary historians, but only four times in the bibliography and three times in the corpus. The other three labels, "reseña", "novela festiva", and "elegía", all go back to explicit signals and occur only once each. This group of labels is not analyzed further because of its minor importance.

4.1.5.2.8 Intention

The last group of subgenre labels considered here are the ones related to the intention the author or narrator pursues with the novel. In the bibliography, there are 13 different labels of this kind. In Bib-ACMé, 34 works (4 %), and in Conha19, 15 novels (6 %) carry an intention label, making this discursive level the least important in quantitative terms.[472]

In the bibliography, there are four labels related to the intention that are assigned to at least five works, each: the "novela moralista" (13 works), "estudio" (7 works), "novela humorística" (6 works), and "novela didáctica" (5 works). In the corpus, "estudio" is the most frequent with five works, and the other labels only occur three times or less. The "novela moralista" and the "novela didáctica" are subgenres that are a topic in literary histories, but the other labels all have explicit or implicit signals as their source. Roughly, the different labels of this group can be classified into labels of entertainment ("novela humorística", "novela cómica", "entretenimientos", "novela curiosa", "juguete", "comedia de carácter", "novela de horrores") and instruction ("novela moralista",

[470] See figure 161 in the appendix of figures.

[471] The number of works associated with them as well as the kind of sources for the labels that refer to the attitude of the author or narrator, are visualized in figures 162 and 163 in the appendix of figures.

[472] The number of works per label, as well as the sources of the labels, are given in figures 164 and 165 in the appendix of figures.

"estudio", "novela didáctica", "novela de propaganda", "lecturas", "novela enciclopédica"). Because this last group of subgenre labels is small, it is not examined further in the following analyses.

Reflecting upon the empirically driven discursive model of subgenre labels that was used here to organize the subgenres contained in the bibliography and the corpus and which served to get an overview of their quantitative distributions, it can be concluded that the model is helpful in enhancing the comparability of subgenre labels. Generic labels usually refer to certain semiotic or discursive levels of a literary work, which can be quite different, for example, the spatial context versus the syntactical realization of a work, and which cannot be compared directly in a useful way.

The specific model was created based on the metadata collected for Bib-ACMé and Conha19 and helped determine which discursive levels are relevant from a quantitative point of view. Applying it to the metadata and evaluating the resulting quantities clarifies which levels have a certain weight, either regarding the variety of labels or the number of labels assigned to the novels. The relevance of some levels was expected because they are traditionally a focus of critical concern: thematic labels and labels related to literary currents. Other levels – the mode of representation, the mode of reality, and identity – resulted in having a certain quantitative significance, as well. This highlights perspectives on subgeneric terms that have not been discussed widely so far and that are mainly derived from explicit historical subgenre signals. It has to be analyzed if and to what extent the historical practices of assigning such labels to the novels actually correspond to textual patterns. In addition, some levels are dominated by individual subgenre labels, which trigger the overall quantitative importance of the respective level, for example, the general term "novela" on the level of the representational mode or the "novela histórica" on the level of the reality mode. All in all, it became clear that compared to the overall variety of subgenre labels, relatively few of them are very frequent.[473]

In Bib-ACMé, 89 labels, which corresponds to 72 % of all the different labels in the bibliography, are only assigned to up to 9 works. In Conha19, this is true for 73 labels, which is 81 % of all the different labels in the corpus. In the bibliography, only 13 labels are assigned to 50 works or more.[474] Regarding the sources of labels, there are 108 different explicitly signaled labels and 34 different literary-historical labels. All these numbers are based on normalized terms, so variances in spelling, formulation, and syntactic constructions are not causing the broad range of subgenre labels. Following these numbers, it can be assumed that in the historical practice of explicitly labeling novels, in particular, creativity and the emphasis on individuality are important factors besides the wish to mark a work as belonging to some established or widely practiced subgenre. In literary-historical approaches, on the other hand, subgenres are studied independently of their quantitative weight in the whole production of novels. In comparison, a digital quantitative

[473] This observation is supported by figure 166 in the appendix of figures, which summarizes how many labels are assigned to how many works.

[474] These are: "novela", "novela romántica", "novela sentimental", "novela histórica", "novela social", "novela de costumbres", "novela realista", "novela original", "novela naturalista", "novela mexicana", "memorias", "episodios", and "novela política". Complete lists of the different subgenre terms in Bib-ACMé and Conha19 and the number of works to which they are assigned are available at https://github.com/cligs/data-nh/blob/master/corpus/overview/subgenres-works-per-label-bib.csv and https://github.com/cligs/data-nh/blob/master/corpus/overview/subgenres-works-per-label-corp.csv. Accessed October 1, 2020.

approach to the subgenres sets different focuses of analysis, or rather, it is only usefully applicable to a subset of the whole range of generic signals and classifications.

Some aspects of the nature of generic terms that complicate classificatory and comparative studies of subgenres are not solved by applying the discursive model. Actually, the semiotic and discursive models of generic terms are above all descriptive models and do not claim to tackle these issues:

- generic terms can be semantically complex and loaded with meaning on several discursive levels (the term "novela histórica", for instance, refers to certain themes but also to the relationship between the text and reality)
- even if generic terms refer to the same discursive level, they can have a different, narrower or broader, semantic scope, which can lead to relationships of inclusion or overlap (the "novela abolicionista", for instance, can be understood as a special kind of "novela social", while both are terms on the thematic level); in addition, no clear line is drawn between generic and subgeneric terms (the term "novela", for example, is more general than "novela epistolar", but both refer to the mode of representation)
- the degree of relationships between a work and the generic terms is not controlled (none, just one, or several generic terms of different levels or of the same level can be associated with one work) so that the generic information is not necessarily complete on all the levels or unique on individual levels (for instance, the literary current of a part of the novels was not determined; on the other hand many novels had multiple labels, for example, on the thematic level)
- nothing is said about the sources of the generic terms and how they are defined. As for general language terms, also the semantics of generic terms is determined by their use, synchronously and diachronically, by numerous different speakers, writers, and critics, at least if the terms are not restricted to selected scholarly defined ones; on the other side, assigning terms to the different discursive levels involves presupposing that they cover certain semantic aspects so that another level of uncertainty is introduced

When classificatory tasks are designed based on the subgenre labels in the bibliography and the corpus, which have a collective background, are potentially semantically multifaceted, historically bound, and neither exhaustive nor unique, and when these tasks are based on how the labels are ordered in the discursive model of subgenre terms, this constitutes a simplifying choice of perspective on a more complex system of generic relationships. To conclude this section of metadata analysis of the subgenres, only the constellations of labels selected for further analysis are presented in more detail, i.e., differentiating the overviews also by country, time period, and related to corpus-specific metadata and characteristics of the texts. This is done in the following subchapters.

4.1.5.3 Subgenre Labels Selected for Text Analysis

In the following, the two discursive levels of subgenre labels that were chosen for text analysis are analyzed further on the metadata level: primary thematic subgenres and primary literary currents. In the text analysis part, the setups for thematic labels are further reduced by concentrating on

the three most frequent subgenres (*novela histórica*, *novela sentimental*, and *novela de costumbres*), but in the following section, also the less frequent ones are examined in terms of metadata.

4.1.5.3.1 Primary Thematic Labels

This chapter analyzes the proportions of novels with a particular primary thematic subgenre label in Bib-ACMé and Conha19. Here, only the top primary thematic subgenres are analyzed in detail, summarizing the remaining ones as "other". Primary thematic subgenres are considered "top" if they cover at least 5 % of the works in the corpus. Although the proportion may be higher in the bibliography, the coverage in the corpus is decisive because, in the end, only the full text of the novels in the corpus is analyzed. First, the general proportions of novels having a certain primary thematic subgenre label in the bibliography and the corpus are assessed.[475] This general overview serves as a reference point for the overviews differentiating by further parameters such as country, time period, etc. In Bib-ACMé, 16 % of the works do not have any thematic subgenre label, whereas in Conha19, all the works were labeled thematically. In both resources, historical novels are most frequent, followed by sentimental novels. In the corpus, novels of customs, social, and political novels have more weight than in the bibliography.

Next, the primary thematic subgenres are analyzed by the three countries Mexico, Argentina, and Cuba.[476] As for historical novels, in the bibliography, they are overrepresented in the Mexican works (38 %). In the Argentine novels, and even more in the Cuban ones, historical novels are underrepresented (19 % and 12 %, respectively). In the corpus, the distribution of historical novels is more balanced by country. However, even there, the Mexican works have a greater proportion of historical novels (32 %) than the Argentine (23 %) and Cuban ones (20 %). Sentimental novels are slightly underrepresented in the Mexican and Argentine works contained in Bib-ACMé (19 %, respectively), and clearly overrepresented in the Cuban works (33 %). Also for this subgenre, the distribution is a bit more balanced in the corpus. The novels of customs are proportionally overrepresented in the Cuban works in the bibliography and also in the corpus. Social novels are above average in Argentine novels and below average in novels of other countries, both in Bib-ACMé and Conha19. Finally, in the bibliography, political novels are a bit overrepresented in the Argentine works and underrepresented in the Mexican and Cuban works, slightly in the first and more in the latter case. In the corpus, in contrast, there are relatively more political novels from Mexico, while the Argentine novels correspond to the average, and for Cuba, there are no political novels at all. So regarding the distribution of primary thematic subgenres by country, the differences between Argentine, Cuban, and Mexican works that are visible in the bibliography are, for the most part, also reflected in the corpus, where they are balanced out a bit. In summary, there are more historical novels from Mexico, more sentimental novels and *novelas de costumbres* from Cuba, and more social novels from Argentina. In the following, the distribution of primary thematic subgenres is given per decade, for Bib-ACMé and Conha19, respectively.[477]

---

[475] See figure 167 in the appendix of figures for a visualization of these proportions.
[476] See figure 168 in the appendix of figures.
[477] See figures 169 and 170 in the appendix of figures.

The most important point to conclude for the bibliography is that all of the top primary thematic subgenres occur in almost all of the decades. The political novel is missing in the 1830s and 1840s, and the social novel in the 1840s. In the corpus, there are some more decades without works of individual subgenres, especially the first and last decades. There is only one novel of customs from the 1830s and one novel of a different subgenre. There are only three novels from the 1910: one political novel, one social novel, and one historical novel. In the more central decades, no political novel is included in the 1840s, 1860s, and 1870s. No social novel is present in the 1840s and 1870s. In contrast, historical novels, sentimental novels, and novels of customs are represented in all the central decades of the corpus.

Regarding the relative amount of different subgenres over the decades, in the bibliography, the proportion of historical and sentimental novels is higher up to the 1860s. After that, especially the social novel gets more significant. When compared to the bibliography, in the corpus, especially the 1860s stand out as a decade with an over-proportional number of historical novels and the 1890s with an above-average number of social novels. The change of proportions of primary thematic subgenres over time becomes clearer if the two periods before 1880 and in or after the year 1880 are compared.[478]

In Bib-ACMé, in particular, the number of sentimental novels is lower after 1880 (17 % instead of 28 % before), and the number of social novels higher (14 % instead of 6 %). The political novel, which is in general not very frequent, also raises from 1 % of the works before 1880 to 4 % after that year. Furthermore, the share of works without any thematic subgenre label is considerably higher in the later period (19 % instead of 10 %). On the other hand, the differences between the proportion of historical novels and novels of customs are not very big between the two periods (4 % in the case of the historical novels and 3 % for the novels of customs). In general, and regardless of the shifts of proportions, the absolute number of novels of all the primary thematic subgenres is higher in the second period than in the first one.

In Conha19, in contrast, the absolute number of sentimental novels and novels with other subgenres drops for the period after 1880. On the other hand, the relative increase of social and political novels is even higher than in the bibliography. There is almost no change in the proportion of novels of customs between the two periods, but the relative importance of the historical novel drops more in the corpus than in the bibliography. It can be assumed that the change of proportions of the various thematic subgenres over time reflects the preferences of the different literary currents, which overlap and succeed each other during the nineteenth century. The romantic current is apparently more closely related to semantic and historical themes, and the realist and naturalistic novels to social and political topics. However, the historical novel persists as an important subgenre throughout the whole century, and also the novel of customs remains relevant. It is striking that the only two thematic subgenres that are frequently explicitly named in the novels' subtitles, the *novela histórica* and the *novela de costumbres*, are the ones that are less subject to change in terms of relative quantities over time, which means that they are less bound to the dominant literary currents than the other thematic subgenres.

---

[478] See figures 171 for the bibliography and 172 for the corpus in the appendix of figures. The percentages in the two figures indicate the proportion of novels of a certain subgenre in comparison to all the novels in the respective period.

In what follows, the distribution of primary thematic subgenres is analyzed in relationship to corpus-specific metadata and text characteristics, namely the prestige of the texts, the narrative perspective, the continent and time period of the setting, and text length. The first of the corpus-specific aspects to be analyzed is the prestige of the texts.[479] Historical novels are equally present both in high- and low-prestige novels. The sentimental novels, in contrast, are underrepresented among the high-restige novels and overrepresented in the low-prestige ones. Interestingly, of the novels of customs, relatively more are classified as high- than as low-prestige. The social novels are slightly overrepresented in the high-prestige group and the political novels in the low-prestige one. Especially in the latter case, the difference is not considered significant because of the comparatively low number of political novels in the corpus. It can be concluded that one of the primary thematic subgenres, the sentimental novel, is clearly marked as having a low-prestige branch, but none of the other subgenres sticks out as a particular high-prestige subgenre.

Considering the narrative perspective, the great majority of historical novels are written in the third person. Social novels also primarily have a third-person narrator. The first-person narrator is most frequent in sentimental novels but also overrepresented in the group of political novels. In the novels of customs, both narrative perspectives are almost equally in use. It is noticeable that 20 % of the novels with a first person narrator have a primary thematic subgenre that is part of the "other" group, so apparently, this perspective is favored in some of the subgenres that are less frequent.[480]

For the continent of the setting, there are also correlations with some of the primary thematic subgenres.[481] Almost half of the novels set in Europe (46 %) are sentimental novels. *Novelas de costumbres* rarely have a European setting – they make only up 4 %. The political novels are exclusively set in America. There are social novels set on both continents, but also for that subgenre, the American setting is more prominent. Only the proportion of historical novels is almost the same for the American and the European setting. A hypothesis that follows from this is that European models of sentimental novels were more often just copied and that models of other subgenres were more often adapted and appropriated to reflect the local circumstances, or even that they developed more into independent varieties of the subgenres.

Correlations between the different primary thematic subgenres and the time period of the setting become visible, as well.[482] Obviously, most of the novels set in the past are historical novels (85 %), and there are no primarily sentimental or political novels in this group at all. Most of the novels set in the recent past are also historical novels (68 %), but here there are also some representatives of the other subgenres, in particular sentimental novels and novels of customs. Novels with a contemporary setting are led by the sentimental subgenre, *novelas de costumbres*, and social novels. However, also 7 % of the novels with a contemporary setting are primarily

---

[479] See figure 173 in the appendix of figures, in which the proportions of primary thematic subgenres are displayed for high- and low-prestige novels.

[480] See figure 174 in the appendix of figures.

[481] See figure 175 in the appendix of figures.

[482] See figure 176 in the appendix of figures. For the figure, the time period was evaluated in relationship to the year of the first known publication of the novels. For details on how the time period was determined, see chapter 3.3.3.1.6 ("Text Classification with Keywords") above.

classified as historical novels, which shows how widely the concept of historicity was interpreted in the novels.

Analyzing the lengths of the works of different primary thematic subgenres reveals that all the groups overlap.[483] There are short novels among all the subgenres, and differences become only visible regarding the median and the variance of length. The longest novels are historical novels, followed by novels of customs and sentimental novels. Apart from the outliers, the subgenres with the highest median length are historical novels (89,000 tokens) and social novels (57,000 tokens), and the ones with the lowest sentimental novels (40,000 tokens) and political novels (43,000 tokens). The novels of customs lie in between with a median of 48,000 tokens, so the median sentimental and political novels are less than have as long as the median historical novel.[484] A test for significance shows that the difference in length is significant for the historical versus all other types of thematic subgenres, but not for the other pairs of subgenres.[485] Analyzing the two central quartiles, the variety of length is greatest for historical novels (50,000 to 139,000 tokens) and novels of customs (34,000 to 94,000 tokens) and lowest for social novels (37,000 to 76,000 tokens). The ratio of variances and, thereby, the difference in statistical variance between selected pairs of subgenres is greatest for the historical versus the political novel, the historical versus the social, and the historical versus the sentimental novel and lowest for the sentimental novels when compared to the novels of customs.[486] As with the development of the number of works associated with the primary thematic subgenres over time, also in terms of text length, the two *explicit* subgenres, *novela histórica* and *novela de costumbres*, stick out as the subgenres with the longest novels on the one hand (considering the outliers and upper fences) and the greatest variety of length on the other hand (in terms of the ranges of the two central quartiles). It can be hypothesized that these are signs of long-lived subgenres for which there is historical variability and for which experimentation to the extremes has taken place. When statistical differences in the distributions of lengths are considered, the historical novel is the one that is significantly longer than the other types of thematic subgenres and, at the same time, the one with a comparatively bigger variance in text length.

4.1.5.3.2 Primary Literary Currents

In this chapter, the primary literary currents to which the novels in the bibliography and the corpus were assigned are analyzed.[487] In the bibliography, the literary current is only known

---

[483] See figure 177 in the appendix of figures.

[484] The numbers are rounded to the next thousand.

[485] The script that was used for the significance tests and the calculation of variances is available at https://github.com/cligs/scripts-nh/blob/master/analysis/sign.py. Accessed January 3, 2021. As the data is not normally distributed, the Mann-Whitney U test was used to check for the statistical significance of the different text lengths. The p-value for *novela histórica* versus *novela sentimental* is 2.8e-06, for *novela histórica* versus *novela de costumbres* 0.002, for *novela histórica* versus *novela social* 0.0005, and for *novela histórica* versus *novela política* 0.02. For the *novela sentimental* versus *novela social*, the p-value is at the limit of significance with 0.047, but for all the other combinations of subgenres, the p-value is higher than 0.05.

[486] The following ratios of variance were calculated: 5.9 for *novela histórica* versus *novela política*, 5.2 for *novela histórica* versus *novela social*, 4.1 for *novela histórica* versus *novela sentimental*, and 0.4 for *novela sentimental* versus *novela de costumbres*.

[487] The general proportions of primary literary currents in Conha19 and Bib-ACMé are visualized in figure 178 in the appendix of figures.

for half of the works (49 % or 405 novels). Normally, this information is available from literary-historical accounts of the novels and not from explicit historical subgenre labels. As many of the novels in the bibliography have not been the focus of critical literary-historical work, this information is missing. In the corpus, the number of novels for which the literary current is known is much higher (79 % or 201 novels), which again shows that the corpus represents works that are better known and more canonized than the works in the whole bibliography.[488] Here, only the primary literary currents assigned to the novels are analyzed. Secondary literary currents are not taken into account here. Examples are novels that are borderline cases or mixtures of romantic and realist novels but also all naturalistic novels that can as well be understood as realist novels in a general sense.

In the bibliography, the largest group of novels by literary current are the romantic novels, with 32 % or 263 novels. This is also the case in the corpus, where the romantic novels have a share of 45 % or 116 novels. In the bibliography, the amount of realist and naturalistic novels is equal (8 % or 66 novels each), while in the corpus, there are a bit more naturalistic works (18 % or 45 novels) than realist ones (14 % or 35 novels). Other literary currents only have a minor quantitative importance in both the bibliography and the corpus. Compared to the bibliography, the relationship between romantic novels on the one side and realist and naturalistic novels on the other is more balanced in the corpus. However, it is difficult to say if this turns the corpus further away from the population of novels or not because the literary current is unknown for so many novels in the bibliography.

When the proportions of literary currents are viewed by country, some differences become visible.[489] In Bib-ACMé, the share of novels for which the literary current is unknown is similar in the three countries, but in the corpus, it is highest for the Cuban novels and lowest for the Mexican ones, with the Argentine novels in-between. The higher number of unknown cases for Cuba reflects that apart from a few very well-known works, novels from that country appear to be less studied. Regarding the distribution of the different literary currents by country, in the bibliography, approximately one-third of the novels are romantic ones, also by country. In Conha19, however, the proportion of romantic novels is bigger for Mexico (55 %) than for the other two countries. Most realist novels come from Mexico, in the bibliography, and also in the corpus. The naturalistic novels have the most weight in Argentina, both in Bib-ACMé and Conha19.

Next, the distribution of the novels' primary literary currents over time is analyzed, considering the number of works per decade in the bibliography and the corpus, respectively.[490] What becomes clearly visible is that the romantic novel is a phenomenon that persisted throughout the whole nineteenth century. It was the dominant current up to the 1870s and only gradually gave way to the other currents in the following decades. The first realist and naturalistic works are found in the 1860s and 1870s, but it was from the 1880s onwards that these two currents gained more weight. If one compares the bibliography and the corpus, the proportions are similar in both, only that there are relatively more realist and naturalistic novels in the corpus. If only the relationship

[488] The known instances of romantic, realist, naturalistic, and modernist novels could be used to determine the literary current of the 55 novels for which this information is still lacking by means of text classification.

[489] See figure 179 in the appendix of figures.

[490] See figures 180 and 181 in the appendix of figures.

between realist and naturalistic novels is examined, the latter are a bit overrepresented in the corpus in the 1890s and 1900s. Considering the period before and after 1880,[491] the dominance of the romantic novel in the early period again stands out, both in Bib-ACMé and Conha19. Moreover, even after 1880, the number of romantic novels is higher (in the bibliography) or comparable (in the corpus) to the amount of either realist or naturalistic works. Concerning the number of works for which the literary current is unknown, it is very high in the bibliography for the period after 1880, amounting to 59 % of all the novels. In the corpus, the proportion of works with unknown currents is more balanced before and after 1880. As was seen before in chapter 4.1.3.1, where the general number of works was assessed independently of the subgenres, the number of works doubled from the 1870s to the 1880s. For the 1880s and 1890s, the corpus contains approximately 30 % of the novels that are registered in the whole bibliography, but for the 1900s, only 20 %. So the high proportion of novels after 1880 in Bib-ACMé for which the literary current is unknown shows that there is a mass of works that is still mostly unexplored.

We now turn to the analysis of the corpus-specific metadata, starting with the proportions of high- and low-prestige novels in Conha19 for the different literary currents.[492] The most striking difference is the higher proportion of novels without a literary current label in the low-prestige group, which is at 34 %, compared to 16 % in the high-prestige group. This result is mainly due to the way that the prestige of the novels was determined. Only works of which at least one new edition was published between 1960 and 2020 were considered high-prestige novels. The high proportion of novels without known literary current in the low-prestige group is just another perspective on them as a group of texts which has been largely forgotten or has not been investigated in the last 60 years.[493]

The next aspect that is analyzed is the kind of narrator that the novels related to the different literary currents have.[494] In general, the third-person narrator is much more frequent in the corpus than the first-person narrator: there are 44 novels (17 %) with a first-person narrator and 212 novels (83 %) with third-person narrator. Interestingly, the proportion of realist novels in the first-person narrator group is considerably higher than in the third-person group. On the other hand, naturalistic novels are overrepresented in the third-person group, and also the romantic novels have a higher proportion in the latter one. In his overview of the history of Latin-American literature, Dill introduces the realist novel as follows: "Die Realität wurde nicht costumbristisch-dokumentarisch kopiert, vielmehr mit literarischer Inszenierung ein Ähnlichkeits- oder Realitätseffekt (*effet du réel*) durch einen heterodiegetischen Erzähler erzeugt, der dem impliziten Leser die neue Gesellschaft und adäquate Verhaltensweisen modellierte" (Dill 1999, 159). So the third-person narrator is usually seen as typical for the realist novel, but the overview of the novels in Conha19 suggests that there is a subgroup of realist novels narrated in first person and that the choice of this narrative perspective is a factor that differentiates the realist from the naturalistic novel, where the third person is proportionally more important. Checking which ones are the realist novels with a first-person narrator reveals that they belong to several different thematic subgenres: the sentimental novel, the historical novel, the novel

[491] See figures 182 and 183 in the appendix of figures.
[492] See figure 184 in the appendix of figures for a visualization of these proportions.
[493] See chapter 3.3.3.1.6, where the collection of prestige metadata for the novels in the corpus is outlined.
[494] See figure 185 in the appendix of figures.

of customs, and the political novel, so one specific type of thematic subgenre is not responsible for the number of realist novels with first-person narrator. In addition, they were written by different authors from the three countries.[495]

The next property to be analyzed is the continent of the novels' setting.[496] Comparing the proportions of novels with an American and European setting reveals that the romantic and also the realist novels are preponderant for the American setting. Naturalistic novels have a similar proportion in both groups, which means that the European setting is relatively more important in them when compared to romantic and realist novels. Apparently, the French origin of the naturalistic current had an influence on the choice of the setting in some cases. Nevertheless, the numbers of the continent of the setting must be interpreted with caution because the great majority of the novels are set in America (90 %, 231 novels) and only a few in Europe (9 %, 24 novels).[497] Furthermore, the numbers for Europe are biased because the proportion of novels for which the literary current is not known is considerably higher than in the case of an American setting. Here there is a correlation with the countries of origin of the novels because knowledge about the literary current is missing for many Cuban novels, and novels from Cuba are also the ones that have a European setting more often than novels from Argentina or Mexico.

How does the time period of the setting relate to the literary currents? Clear tendencies are visible in this respect:[498] the romantic novel has high proportions in all three categories but is overrepresented in the groups of novels with a setting in the past or recent past. The realist novel primarily has a contemporary setting or one in the recent past, and the naturalistic novel is, above all, set in the present. That the romantic novel is inclined towards the past is in line with the fact that the historical novel was a very popular thematic subgenre in that current. However, the results also show that the romantic novel is a multi-faceted phenomenon because all the different time periods of the setting are covered by it to a significant extent. That the naturalistic novel primarily has a contemporary setting confirms its role as the type of novel that served to depict the process of social, economic, and technological modernization going on in the Spanish-American countries in the last decades of the nineteenth century. In terms of text style, the preferences for certain time periods of the setting that the novels associated with the different literary currents have might influence, for example, the usage of temporal expressions or verb forms, which is an aspect that needs to be investigated further.

Finally, the differences in text length between the novels of the three main literary currents are analyzed.[499] The romantic novels have a median length of 66,000 tokens, the realist novels of 56,000, the naturalistic novels of 51,000, novels with other literary currents of 62,000, and

---

[495] In the corpus, the realist novels which have a first-person narrator are the four novels of the Mexican writer Emilio Rabasa: "La bola" (1887), "La gran ciencia" (1887), "El cuarto poder" (1888)", and "Moneda falsa" (1888), furthermore the sentimental realist novel "Angelina" (1893, MX) by Rafael Delgado, the historical realist novel "Las ranas pidiendo rey. Confesiones de una afrancesada (1861–1862)" (1903, MX) by Victoriano Salado Álvarez, and the novels "La gran aldea" (1884, AR) by Lucio Vicente López, "Mi tío el empleado" (1887, CU) by Ramón Meza, "Don Perfecto" (1902, AR) by Carlos María Ocantos, and "Divertidas aventuras del nieto de Juan Moreira" (1910, AR) by Roberto Payró.

[496] See figure 186 in the appendix of figures.

[497] The only novel that is neither set in America nor Europe is a science fiction novel set on the planet Mars.

[498] See figure 187 in the appendix of figures.

[499] See figure 188 in the appendix of figures.

novels with an unknown literary current of 42,0000 tokens.[500] Of most interest are the differences between the romantic, realist, and naturalistic novels because the group of novels with another literary current only consists of 5 novels, and the "none" group is probably mixed. A test for significance shows that the differences in length between the romantic, realist, and naturalistic novels are not statistically significant, though.[501] Even if the differences in median text length are not significant, the group of romantic novels has a bigger variance in length than the other two currents, which are more similar regarding the variation of the novels' length.[502] That the length of romantic novels varies more is plausible because the group of romantic novels in the corpus is bigger (116) than that of the realist (35) or naturalistic (45) novels, but it can also be a sign of the variability of romantic novels in themselves. Romantic novels can be part of several distinct thematic subgenres, of which, for example, historical novels can be very long and sentimental novels are usually shorter. In addition, romantic novels spread over the whole nineteenth century, whereas the realist and naturalistic novels were concentrated in the decades 1880 to 1910. The general comparison of text length by decade that was made in chapter 4.1.3.2 above showed that the variability of text length was lowest from the 1870s to the 1890s, and very long novels were the exception in these decades. This suggests that the form of the novel in terms of length was more stabilized in the last decades of the nineteenth century. In Spanish America, romantic novels participated in all the phases that the novel as a genre underwent in the nineteenth century, so they were variable, whereas the realist and naturalistic novels were anchored in a more precise literary-historical moment. That the variance of the length of the three major currents is smallest for the naturalistic novels can be interpreted as a sign for the relatively uniform and defined form of the novels participating in that subgenre.

Finally, the results of the metadata analysis on the novels in the bibliography and the corpus will be briefly summarized and evaluated here in overview. All in all, the proportions of works in the corpus and the bibliography are very similar in most aspects. In the corpus, some balancing can be noted, for example, concerning the distribution of subgenres in general or by country. This means that the corpus deviates from the distributions in the bibliography to mitigate certain tendencies of imbalance and to achieve a dataset that is more suited for an analysis of subgroups, even if it doesn't represent the proportions of the sampling frame closely. In addition, there are effects that are due to the fact that the sources for the corpus are limited because specific works are difficult to access in digital format. As a whole, the corpus is not entirely balanced in all its aspects. Its closeness to the distributions of subgroups in the bibliography is bigger than their balance in terms of similar numbers in the corpus itself, for example, by subgenre. The overviews of the subgenres selected for analysis highlighted which metadata and also textual factors are connected to the subgenres. They pointed out some interdependencies (e.g., regarding narrative perspective or text length). Some influencing factors only became visible from a single

[500] The numbers are rounded to the next thousand.

[501] The python script that was used for the significance tests is available at https://github.com/cligs/scripts-nh/blob/master/analysis/sign.py. Accessed January 3, 2021. The data is not normally distributed, so the Mann-Whitney U test was used. The p-value for *novela romántica* versus *novela realista* is 0.22, for *novela romántica* versus *novela naturalista* 0.16, and for *novela realista* versus *novela naturalista* 0.40.

[502] The ratio of the variance between romantic and realist novels is 3.7, between romantic and naturalistic novels 2.3, and between realist and naturalistic novels 0.6.

perspective, whereas others manifested themselves repeatedly, such as the very long Mexican historical novels, for instance.

Such an analysis of metadata in connection with the subgenres is of interest in itself, but it also provides useful background knowledge for further textual analysis to be able to evaluate possible biases in the results. Especially, differences between the works of the three countries and also some chronological trends became visible. When conducting analysis on the texts, it must be kept in mind that the sampling frame and also the corpus are not entirely balanced datasets – instead, there are many different factors influencing each other and some asymmetries that are already characteristic of the underlying set of novels. In the next chapter, the textual features used for text analysis and the methods used to categorize the novels by subgenre are presented.

4.2 Text Analysis

In this chapter, text analyses are carried out with the novels of the corpus based on two different sets of text features. The first set consists of general features derived from the most frequent words in the corpus, and the second set consists of topic features created with topic modeling. In chapter 4.2.1, it is explained what the two main types of features are and how they were created. In the second part of this chapter (4.2.2), two different methods are used to categorize the novels. First, statistical classification is employed to classify the novels by subgenre. Three thematic subgenres (*novela histórica*, *novela sentimental*, and *novela de costumbres*) as well as three literary currents (*novela romántica*, *novela realista*, *novela naturalista*) were chosen for the analysis as the quantitatively most relevant groups in the corpus. The classification methods that were used and the results of the analysis are discussed in chapter 4.2.2.1. A family resemblance analysis is realized in chapter 4.2.2.2, and a proposal is made for how the calculation of text similarities, network analyses, and community detection can be used to that end. In the discussion of the results, selections must be made as the amount of data that results from all the different analyses is huge. The overall results are presented in every case, but the study of individual subgenres and novels is only deepened in some cases.

4.2.1 Features

Two main types of features were selected as a basis for the genre categorization tasks: general and semantic features. The general features include different sets of tokens that are most frequent in the corpus (words, word n-grams, and character n-grams). They are called "general" here because, in their basic forms, these features are not filtered based on their linguistic characteristics, which means that all types of tokens are included independently of specific grammatical or semantic properties that they might have. The second feature group is called "semantic" because the feature sets are built on lexical units and on tokens that convey meaning by the way they occur together in the texts. In particular, the method of topic modeling is used. The general features are presented in chapter 4.2.1.1, and the semantic features are outlined in chapter 4.2.1.2.

In digital literary and stylometric studies, the features that are termed "general" here are commonly understood as "stylistic" features in a narrow sense, meaning that they are suitable to differentiate between different writing styles of authors, periods, or genreswithout being too

closely connected to the contents of the texts. This is especially the case when the number of most frequent items chosen is limited, for example, to a few hundred, so that mainly function words are included (Burrows 2002, 268; Juola 2006, 33–34; Stamatatos 2009, 539–542).[503] For the purpose of text classification, the use of such features is often considered a good baseline against which other approaches can be compared because they have often been tested and proved successful for a variety of tasks, also for the classification of texts by genre (Hettinger et al. 2016). In this dissertation, the general features are used for both reasons, to cover elements of style in the narrower sense and to use the results based on this feature type as a starting point for the evaluation of other feature types. However, precisely because general features in the high-frequency spectrum are usually only weakly semantically loaded, their interpretation can be more difficult than for inherently semantic features.

The semantic features were chosen as an alternative that is easier to inspect and that can be linked to expectations about known subgenres of the novel, especially the thematic ones. A love theme, for example, is assumed to occur in sentimental novels and also often in the ones belonging to the romantic current. Historical novels are expected to include topics related to political events or military confrontations, and novels of customs descriptive topics that represent different spheres of life.[504] These examples show that hypotheses about connections between specific subgenres of the novel and different kinds of semantic characteristics of the texts are easily formulated. The usage of semantic text features aims to make it possible to check whether or not such assumptions hold when a large number of texts is analyzed quantitatively, but it also enables the discovery of unexpected correlations. Although semantic features are more closely related to the contents of the texts than general token-based features, they can still be considered stylistic features in a wide sense because how a theme is developed in terms of topics in a text is a stylistic choice. This becomes clear when the relationship between the general literary concept of theme and the specific textual topic features is analyzed. In text-centered approaches, "theme" is usually defined as a non-surface element characterizing the underlying semantics of a narrative literary text and as an element that can be connected to linguistic manifestations in the text.[505] Topic features, in contrast, serve to directly measure characteristics of the textual surface. These cues can subsequently be used to make inferences about higher-order semantic structures of the texts, which can be interpreted in terms of generic facets. So the literary theoretical concepts and the Natural Language Processing terms are different in the way the texts are approached. The NLP concepts have a more direct relationship to text style, as it is understood here.[506]

This raises the question of how a common conceptual ground can be found for the literary-theoretical and the NLP terms. Here, no direct and specific digital modeling and formalization of literary theoretical concepts is intended. Such a mapping of concepts to the requirements of formal text analysis would need an extensive discussion of the formalization procedures and

---

[503] If higher amounts of the most frequent items are chosen, also the general features include tokens that are content words, which have more specific semantic properties than function words. This is an effect of the size of the feature space rather than due to preliminary considerations regarding the semantics of the features.

[504] See the presentation of the different subgenres in chapter 2.3.

[505] An overview of different disciplinary approaches to the analysis of themes and topics is given in Anz (2007).

[506] See chapter 2.2 for the working definitions of literary style used in this study.

also the development of entirely new tools or at least advanced adaptations of existing ones.[507] Instead, for the purpose of this dissertation, the problem is approached by assuming a loose connection between literary theoretical terms and existing terms employed in text mining, which is established by clarifying the differences and similarities. This is discussed in chapter 4.2.1.2 for the relationship between literary themes and topics. In the same way, only loose hypotheses are formulated regarding the relationships between definitions of subgenres of the novel and characteristics of the textual surface. The subgenres are not defined formally and are not directly linked to linguistic properties in a top-down approach. The main reason for this reserve to explicit modeling is that the assignment of the novels in the corpus to the subgenres has not been done on the basis of specific, unambiguous definitions but instead by collecting and interpreting generic signals and attributions of different kinds and provenience. A connection between the textual patterns found and the literary conceptions of the subgenres is explored ex-post, based on the analyses results. That way, mappings between textual evidence and generic categorizations emerge from mainly exploratory procedures and can be described as stylistic cues, stylistic traits, and generic facets.

In the context of the analysis of subgenres of the novel conducted here, the cues are the feature values that turn out to be distinctive for a particular textual category, but the generic facets are not predetermined. Instead, in the case of classification, the cues are interpreted in terms of the subgenre labels that are associated with the text category to look for generic facets. In the analyses of family resemblance structures, it is examined to what extent the cues characterizing the text types found with different feature sets are related to the subgenres at all – on the different discursive levels that these are defined on. In parallel, connections to other determining textual and contextual factors are checked, e.g., the narrative perspective, the country of origin, or the period of publication. So instead of building a direct chain from theories about the subgenres and their properties to expected stylistic characteristics of the texts, the empirical findings are interpreted regarding their relationship to the subgenre labels and other influencing factors (which could turn out to be relevant facets). The hypothesis that is tested is of a very general kind: “the cues are generic”, meaning that it is tested whether there is a statistically relevant correlation between subgenre labels and the distinctive features of text categories.

4.2.1.1 General Features: MFW

Several sets of general features were prepared. They can be distinguished by the unit that is counted, by the number of most frequent items taken into account, and by the kind of normalization applied to the resulting counts. The values that were selected on these three levels are summarized in table 26.

Regarding the token units, words are the classical option for a “bag” representation of the texts, which is needed as a basis for most machine learning algorithms (Müller and Guido 2016, 330; Scikit-learn developers 2007–2023d). For a bag-of-words model, the text is tokenized into individual units, which are then counted. As a result, a corpus is represented by the token counts for each document. In such a model, the order of the units does not play a role anymore. By

[507] See, for example, first approaches to model literary space conducted by Barth and Viehhauser (2017).

Level			Selected values
token unit	basic units		words, word 2-grams, word 3-grams, word 4-grams, character 3-grams, character 4-grams, character 5-grams
	character subtypes	n-gram	“all”, “word” (mid-word, multi-word), “affix-punct” (prefix, end-punct)
frequency range (MFW)			100, 200, 300, 400, 500, 1000, 2000, 3000, 4000, 5000
normalization			tf, tf-idf, z-scores

Table 26. Parameters for general feature sets.

using word n-grams, i.e., sequences of n words, as the basic unit, the relevance of word order can be reintroduced into a bag-of-words model to a certain extent. Word 2- to 4-grams were chosen to test different ranges of word sequences. While word 2-grams are most frequent in the whole corpus, 3- and 4-grams are closer to whole phrases and are more special features that might be relevant for the distinction of subgenres. Diminishing the unit to character sequences, on the other hand, has several other advantages. First, it makes the model less dependent on the exact spelling of the words. In a word-based approach, a difference in spelling results in different features. In a character-based approach, the words are split into several sequences anyhow so that the effect of orthographic differences in specific parts of the words is smaller. For the corpus at hand, this is of interest because some spelling errors persist as a result of the text digitization process.[508] Second, character n-grams that include blank space and punctuation marks cover aspects of the text string that are not captured by the classical bag-of-words approach, where words are usually split at blank spaces and punctuation marks are ignored. Third, character n-grams can be modeled so that they cover specific linguistic substructures of a morphological, thematic, or stylistic nature. Here, both general character n-grams (called “all” in the above table) and specifically modeled ones (called “word” and “affix-punct”) are used. How the different types of character n-grams are generated is illustrated with an example sentence given in example 49.[509]

```
«Espero que algún día, vos amaréis también mis rosas blancas».
```

Example 49. Example sentence for character n-gram creation.

The general approach, which is implemented, for example, in the Python library scikit-learn (Scikit-learn developers 2007–2023f), uses all types of characters in the string and yields the following 3-grams for the example sentence: “«Es”, “Esp”, “spe”, “per”, “ero”, “ro_”, “o_q”, “_qu”, “que”, “ue_”, “e_a”, “_al”, “alg”, “lgú”, “gún”, “ún_”, “n_d”, “_dí”, “día”, “ía,”, “a,_”, “,_v”, “_vo”, “vos”, etc.[510] As Sapkota et al. state, such general character n-grams are very effective for text classification, for example, by author. However, because of the mixed nature of the n-gram types, it is not clear why they are so useful: “Character n-grams are the single most successful feature in authorship attribution […], but the reason for their success is not well understood. One

[508] See chapter 3.3.2 on text treatment.
[509] The sentence is taken from the novel “Lucía Miranda” (1860, AR) by Eduarda Mansilla de García.
[510] Underscores are used to represent blank spaces to enhance the readability of the n-grams.

hypothesis is that character n-grams carry a little bit of everything: lexical content, syntactic content, and even style by means of punctuation and white spaces […]. While this argument seems plausible, it falls short of a rigorous explanation" (Sapkota et al. 2015, 93). The authors of the cited paper test this hypothesis by designing different types of character n-grams and by using each type separately in authorship attribution tasks to see which kinds of n-grams are successful. They use three types of super categories: n-grams covering morpho-syntactic features (affix-like n-grams, called "affix"), thematic content (word-like n-grams, called "word"), and style (punctuation-based n-grams, called "punct"). The n-gram types associated with the three super categories are called "prefix", "suffix", "space-prefix", and "space-suffix" for the affix group, "whole-word", "mid-word", and "multi-word" for the word group, and "beg-punct", "mid-punct", and "end-punct" for the punctuation-based group.[511] Sapkota et al. (2015, 96–98) found out that the affix-like n-grams are most successful in their single-domain authorship attribution task. For cross-domain authorship attribution, also the punctuation-based features are very strong.[512]

The subtypes of n-grams that are modeled here are based on the propositions of Sapkota et al., but because the issue is subgenre classification and not authorship attribution, the selection of n-gram subtypes is led by hypotheses about the features' relevance for this categorization task. Here, the subtypes "mid-word" and "multi-word" are selected and combined into a word-based character n-gram feature set (called "word"). It aims to cover aspects of the thematic content that is supposed to be different from subgenre to subgenre. Furthermore, the subtypes "prefix" and "end-punct" are used in a set combining morpho-syntactic and stylistic characteristics (called "affix-punct"). Its goal is to test whether the differences between the subgenres may also be captured with features that are not primarily content-based.[513] The main reason for not only using general character n-grams is that they are dominated by blank spaces and the short most frequent words in different variants if only a certain number of top most frequent items is used (see table 28 below illustrating the resulting top most frequent tokens for different types of the general features). The intuition is that these features can hardly be interpreted in terms of the different subgenres. Regarding the chosen range of character numbers, it was decided to use 3-, 4-, and 5-grams because 3-grams are the minimum that is needed to be able to construct the different character n-gram subtypes, and 4-grams have been successfully used in subgenre classification before (Hettinger et al. 2016). The 5-grams are included as an alternative candidate. Examples of the n-gram subtypes are listed in table 27, using the above-mentioned example sentence for their creation.

---

[511] For details on how these are implemented, see Sapkota et al. (2015, 94–95).

[512] With "single-domain", they mean that the corpus consists of texts about a single topic written by different authors. "Cross-domain" means a corpus with multiple different topics.

[513] "Whole-word" was not selected because the general bag-of-words approach already covers this feature type. Of the affix-like features, "space-prefix" and "space-suffix" were not selected because they are congruent with "prefix" and "suffix", the only difference being that they are one character shorter because they start or end with a blank space. "Suffix" was not selected because, in the Spanish language, the suffixes are highly influenced by how pronouns are used (if they are used freely before or attached to verb forms after them). As has been shown in the chapter on text treatment above (see 3.3.2), the difference in pronoun use is connected to the type of edition and its publication year, so it cannot be reliably associated with the subgenres. Of the three punctuation-based n-gram types, only "end-punct" was used because it is the only one capturing phrase-level structures of the sentences.

Group	Subtype	Subtype definition	Examples (for 3-grams)
word	mid-word	"a character n-gram that covers n characters of a word that is at least n + 2 characters long, and that covers neither the first nor the last character of the word"[514]	spe per lgú mar aré réi amb mbi bié osa lan anc nca
word	multi-word	"n-grams that span multiple words, identified by the presence of a space in the middle of the n-gram"	o_q e_a n_d s_a s_t n_m s_r s_b
affix-punct	prefix	"a character n-gram that covers the first n characters of a word that is at least n + 1 characters long"	esp alg ama tam ros bla
affix-punct	end-punct	"a character n-gram whose last character is punctuation, but middle characters are not"	ía, as»

Table 27. Definition and examples of character n-gram subtypes.

The examples highlight the differences between the character n-gram subtypes. The most evident one is that the number of resulting n-grams varies. The general approach using all types of character n-grams for the example sentence results in 60 3-gram tokens. In contrast, there are only 13 mid-word, eight multi-word, six prefix, and two end-punct tokens, so depending on the rules set up for the character n-gram subtypes, the feature space is reduced considerably because only some types of words or cross-word sequences meet the conditions.

Besides the different token units, the frequency ranges were chosen so as to cover stylistic features in the narrower sense with only a small number of most frequent words but also a mix of function and content words with several thousand most frequent words. Three types of normalization were chosen for the token counts: term frequency (tf), term frequency – inverse document frequency (tf-idf), and z-scores. The raw, absolute counts are computed as a basis for the normalization steps but are not directly used in the categorization because they depend heavily on the length of the texts, which varies considerably.[515] The term frequency is a relative frequency balancing out the effects that different text lengths have. It is calculated by dividing the absolute number of occurrences of a term *t* in document *d* by the frequency of the term *t'*, which is the term with the maximum frequency in *d* (see formula 1).

$$\mathrm{tf}(t,d) = \frac{f_d(t)}{max_{\mathrm{t}' \in d} f_d(\mathrm{t}')} \qquad f_d(t) := \text{frequency of term t in d}$$

Formula 1. Term frequency.

For tf-idf, the tf-scores are weighted by the inverse document frequency, which is defined as the logarithm of the number of all documents divided by the number of documents that contain the term *t*. This normalization takes into account the structure of the whole corpus: terms that

[514] The source of this and the other subtype definitions in the table is Sapkota et al. 2015, 94f.
[515] See the overviews on text length in chapter 4.1.3.2.

occur in many documents of the corpus are considered less important than terns that only occur in a few documents. This gives the function words, which are likely to occur in all documents, a weaker weight compared to the more specific content words. Here, the tf-idf-scores were calculated with the Python library scikit-learn, which uses a specific implementation of this weighting scheme (see formulas 2 and 3).[516]

$$\text{tf} - \text{idf}(t, d) = \text{tf}(t, d) * \text{idf}(t)$$

Formula 2. Tf-idf.

$$\text{idf}(t) = log(\frac{|D|}{|\{d \in D : t \in d\}|}) + 1 \qquad D := \text{corpus of documents}$$

Formula 3. Idf.

The third kind of normalization used, the z-scores, also involve the relationship between the different documents in the corpus, but in a different way than the tf-idf-scores. A z-score measures the number of population standard deviations by which the value of a score is above or below the population mean. This results in positive values for scores that are above the mean, zero values for scores that are equal to the mean, and negative values for scores that are below the mean. The population values are calculated by determining the standard deviation and mean for each feature in the corpus (see formula 4).

$$Z = \frac{f_d(t) - \mu_t}{\sigma_t} \qquad \mu_t := \text{mean of term t in D}$$
$$\sigma_t := \text{standard devidation of term t in D}$$

Formula 4. Z-score.

Like tf-idf-scores, z-scores upgrade the importance of terms that are not so frequent relative to the ones that are very frequent, but in the case of z-scores, it is not decisive in how many of the documents in the corpus they occur. In addition, the range of the resulting values is different than for tf and tf-idf because z-score zero values indicate that a term frequency is equal to the population mean but not that the term in question has a frequency of zero in the document. Instead, terms that do not occur at all in documents get negative z-score values. This different treatment of zero values can have consequences when the texts are categorized.

Furthermore, z-scores themselves do not take into account the effects of different text lengths by computing relative values per document as in tf and tf-idf. By applying the z-score transformation to absolute values, text length effects would only be controlled indirectly and only to a certain extent because the spread of the absolute values in the corpus is factored in through the standard deviation. A term with a higher standard deviation is penalized by the denominator in the formula if it is above the mean and revalued if it is below the mean. However, very long texts would tend to get higher z-score values. In addition, it would not be clear if higher z-score

[516] The 1 that is added to the idf secures that terms that occur in all documents will not be ignored (Scikit-learn developers 2007–2023g).

values are due to text length or to above-average uses of a term. It does therefore make sense to apply the z-score-normalization on top of tf-values, which is done here. By considering several normalization strategies in the feature sets, these can be compared to see which effect they have on the kind of features that turn out to be relevant for the categorization. While tf and tf-idf-scores are generally used in document classification, z-scores have especially been used in the context of authorship attribution (Burrows 2002; Evert et al. 2017; Müller and Guido 2016, 338–340). In sum, the combinations of parameters for the general features listed above result in 390 different MFW-based feature sets.[517]

The bag-of-words representations of the novels were created on the basis of full-text files that were extracted from the linguistically annotated TEI files.[518] This was done to be able to remove tokens that were annotated as proper nouns, such as person and place names, from the files. This preprocessing step is considered important because especially the names of the protagonists can have distorting effects on the lists of MFW, as they occur very often. By taking out these named entities, of course, conclusions of the form "the female name 'Blanca' is overrepresented in sentimental novels" or "the place name 'Madrid' is mentioned above-average in Cuban novels" cannot be drawn anymore. Instead, the features are intended to lead to insights into more general linguistic-stylistic characteristics of the texts. In addition to the replacement of proper nouns, a list of stop words was created, consisting of the lists of proper names and place names that were compiled for the spell check of the text files.[519] The stop word list is used by scikit-learn's CountVectorizer (Scikit-learn developers 2007–2023f) in the process of creating the bag-of-words models that are based on word units.[520] The stop word list was applied to the full texts in a customized manner before the character n-gram features were created.[521] The combination of entity removal and stop word list has two advantages. A stop word list alone would not work to detect proper names that also have a general meaning in the language, and there are a lot of such names in Spanish, for example, "Blanca" (white), "Clemencia" (mercy), "Rosa" (rose), "Gloria" (fame), "Hidalgo" (nobleman), "Salvador" (saviour), "Gil" (silly), "Cortés" (polite). On the other hand, the stop word list can compensate for some errors of the named entity recognition.

Several visualizations were created to characterize the resulting feature sets.[522] It is, for example, useful to know if a feature set is sparse, i.e., if many of the features have mostly zero

---

[517] The common abbreviation "MFW" is used in the general sense of "most frequent items" here and also includes the most frequent n-grams. The features were generated with a Python script available at https://github.com/cligs/scripts-nh/blob/master/features/general_features.py. The resulting feature sets are available at https://github.com/cligs/data-nh/tree/master/analysis/features/mfw. Both links were accessed on November 4, 2020.

[518] See chapter 3.3.5 for details about the linguistic annotation. The annotated TEI files were used in their corrected form and are available at https://github.com/cligs/conha19/tree/master/annotated_corr and the corresponding full-text files at https://github.com/cligs/conha19/tree/master/txt_annotated_corr. Accessed December 14, 2020.

[519] See chapter 3.3.2 about the text treatment. The resulting stop word list is available at https://github.com/cligs/data-nh/blob/master/analysis/features/stopwords/mfw_stopwords.txt. Accessed November 4, 2020.

[520] Per default, the CountVectorizer uses the token pattern `(?u)bww+b`, which looks for sequences of word characters separated by word boundaries. For the Spanish texts in the corpus, the pattern was slightly modified to also cover words consisting of just one character, such as "y" or "a": `(?u)bw+b`.

[521] The special subtypes of character n-gram units were created with a Python approach designed specifically for this purpose (see the link to the script in footnote 517) because the CountVectorizer only supports the general character n-grams that include all types of characters.

[522] For this purpose, also the script https://github.com/cligs/scripts-nh/blob/master/features/general_features.py was

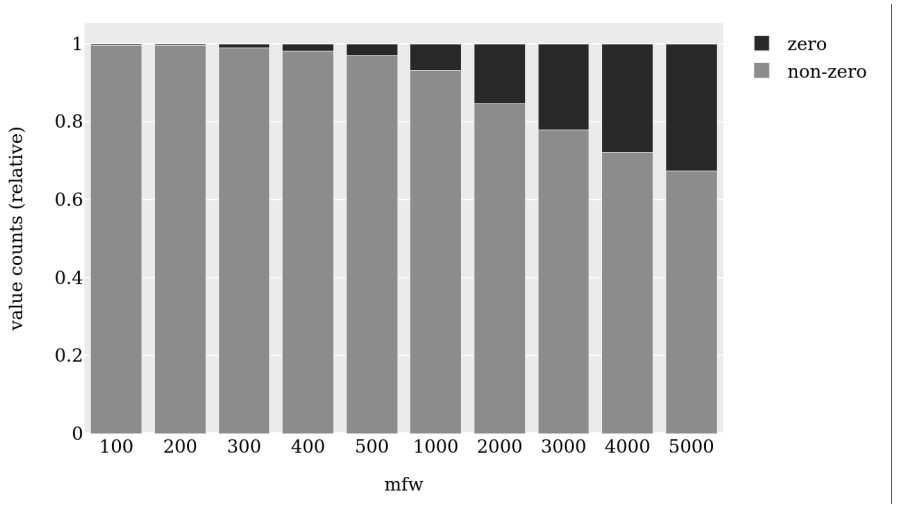


Figure 36. Proportions of zero values in MFW feature sets.

values or if almost all the features are present in all instances of the corpus. It is also of interest to know how much the feature values vary. Both these characteristics have an influence on how well specific categorization methods work and if further steps of preprocessing are needed before the feature sets can be used for categorization.[523] Figure 36 shows how the overall proportion of zero values increases from 100 to 5,000 MFW from less than 1 % to one-third of all the values.[524]

Looking in more detail into how many features have how many zero values in figures 37 and 38, it can be seen that the zero values in MFW100 stem from a few features only, of which one has over 40 zero values. 94 % of the features are never zero. In contrast, in MFW5000, only 7 % of the features are never zero. The form of the histogram shows that there are many features that are zero in about half of the texts. For the frequencies between 100 and 5,000 MFW, the overall proportion of zero values lies between these extremes, and the bulk of features has zero values in an increasing part of the corpus.

Character n-grams are much more stable regarding the proportion of zero values than words. For example, there are no zero values for the 100 most frequent character 5-grams. For the 5000 most frequent ones, the amount increases only to 1 %. For word n-grams, on the other hand, the proportions of zero values increase abruptly from the 3-grams on. The 100 most frequent word 2-grams, for instance, have less than 1 % of zero values. In contrast, the 100 most frequent word

used. The resulting plots are available at https://github.com/cligs/data-nh/tree/master/analysis/features/mfw/overviews. Accessed November 4, 2020.

[523] Random forest classifiers, for example, tend not to work very well with high-dimensional sparse data. Support Vector Machines (SVM), on the other hand, usually handle these well but work better when features have similar scales (Müller and Guido 2016, 90, 103–106).

[524] This chart only shows the feature sets based on single words. Charts displaying the proportions of zero values in the feature sets based on other token units (word ngrams and character ngrams) are available on GitHub. See footnote 522 above.

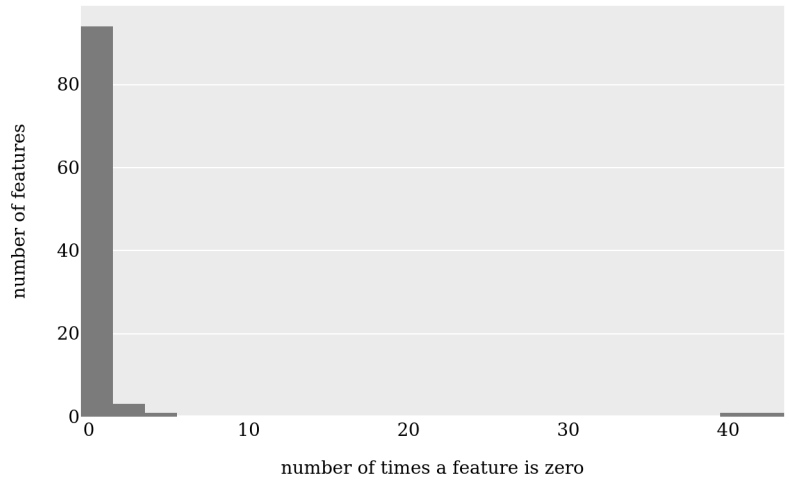


Figure 37. Distribution of zero values in MFW100.

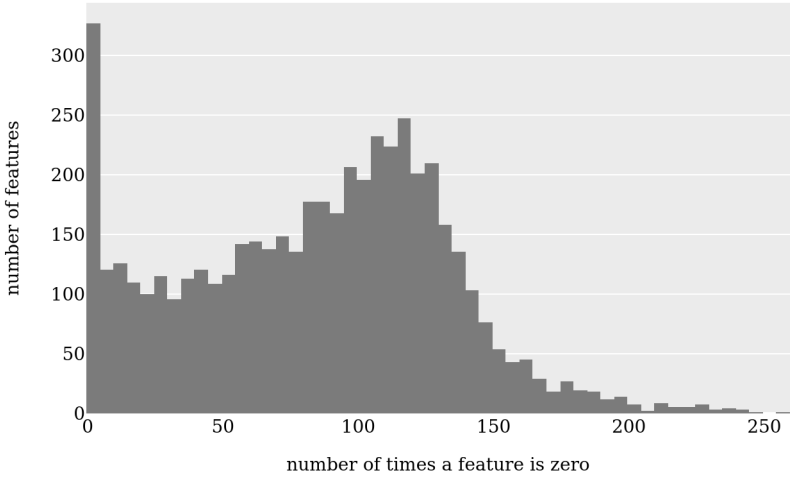


Figure 38. Distribution of zero values in MFW5000.

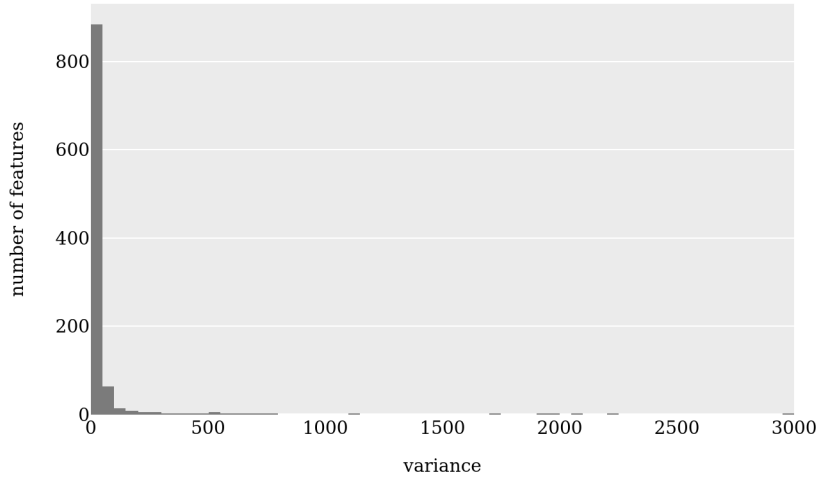


Figure 39. Variances of the 1000 MFW (absolute values).

3-grams have 11 %, and the word 4-grams 43 %. The 5000 most frequent word 2-grams have 32 % of zero values, the word 3-grams 57 %, and the word 4-grams 82%.[525]

In the following figures 39 to 42, it is illustrated for MFW1000 how the different normalization strategies reduce the variance of the feature values. For absolute values, the variances range between 6 and 2,976, for tf-scores between 0.001 and 0.150, for tf-idf-scores between 0.0006 and 0.0581, and for z-scores between 0.999999999999998 and 1.000000000000003, so the absolute values of variance are smallest for the tf- and tf-idf-scores, but the range of variance is smallest for the z-scores.[526] In addition, the distribution of the variances changes from a right-skewed distribution for the tf- and tf-idf-scores to a left-skewed distribution for the z-score values, for which the variance is expressed in relative terms and, therefore, more balanced for the different frequency ranges of the features. With z-scores, most features vary to a similar degree, and only a few features are more consistent throughout the texts in the corpus. A hypothesis of the effect that these differently normalized features have on classification tasks is that less frequent features will have a higher probability of being selected as important features if the variances of the features are more balanced.[527]

The top 10 features for different frequency sectors and token units are listed in table 28 to get a sense of the kind of features contained in the matrices of the most frequent items.

For the word unit, the frequency range 1–10 contains exclusively function words (propositions, conjunctions, articles, and pronouns) and the particle “no”. In the range 101–110, the nouns “amor” and “vez” enter the list, and there are several verbs and adverbs with general meanings. Range 501–510 contains more specific nouns (“voluntad”, “alegría”, “pasos”), semantically more

[525] For details, see the overviews at https://github.com/cligs/data-nh/tree/master/analysis/features/mfw/overviews. Accessed November 4, 2020.

[526] The last number or decimal of the values was rounded.

[527] See, for instance, the remarks made on the scikit-learn website on how different feature variances influence the selection of coefficients in linear models (Scikit-learn developers 2007–2023c).

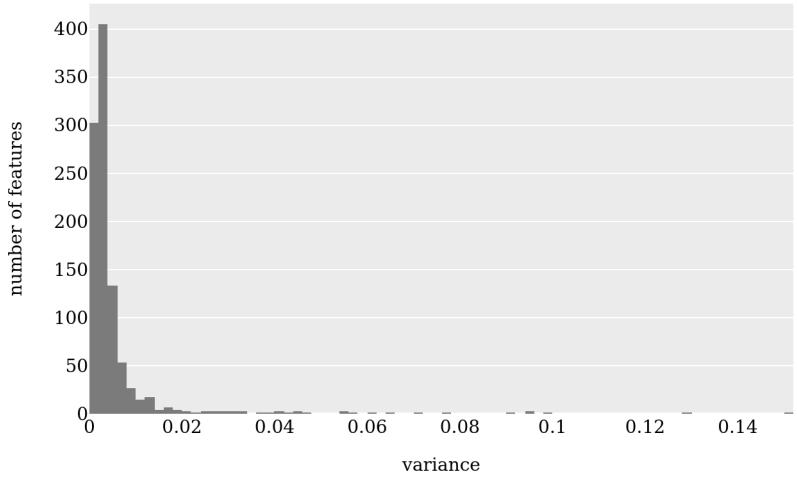


Figure 40. Variances of the 1000 MFW (tf-scores).

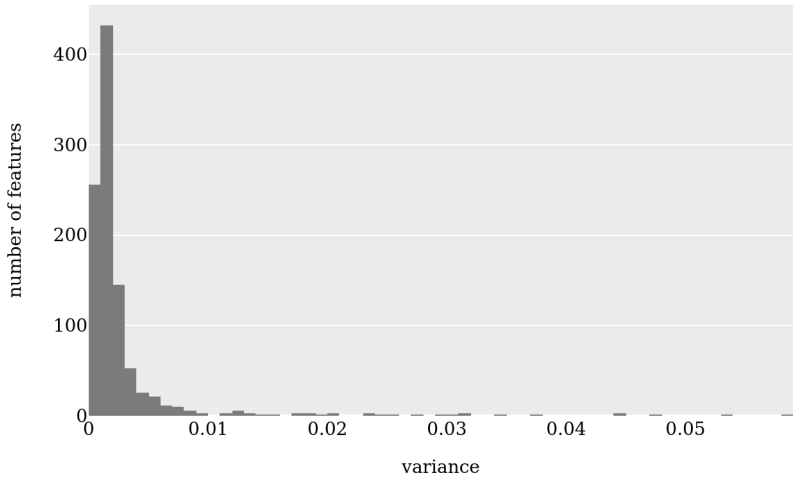


Figure 41. Variances of the 1000 MFW (tf-idf-scores).

Frequency sector (MFW)	Token unit	Tokens
1–10	word	de el la que y a en se los no
101–110	word	esto habían allí hay desde todas amor vez toda mucho
501–510	word	alto última voluntad pensar dirigió delante_de alegría siquiera además pasos
1001–1010	word	daban no_obstante basta tranquilo hablando vivo mozo batalla cólera poco_a_poco
4091–5000	word	injusticia apoyada sensaciones bendito cuadra insignificante personalmente distinción significa fijo
1–10	word 2-gram	de_el a_el de_la a_la en_el en_la de_los de_su que_se lo_que
101–110	word 2-gram	como_si si_no todas_las por_su pero_no sobre_el y_le por_los a_una la_joven
501–510	word 2-gram	calle_de dijo_que como_los una_palabra era_de sobre_las mi_madre sobre_los el_primero que_tiene
1001–1010	word 2-gram	de_otro tenía_que en_toda el_instante sus_hijos con_lo ver_lo hombres_de preguntó_el dueño_de
4091–5000	word 2-gram	esos_hombres eso_y mañana_de sin_querer en_ti qué_había dando_se sus_soldados paz_y amigo_el
1–10	char 4-gram	_de_ _el_ que_ _que _la_ _en_ _con _se_ o_de a_de
101–110	char 4-gram	ada_ _cua más_ as_d a_ca a_y_ o_co ue_l _más e_qu
501–510	char 4-gram	es._ as._ ar_e _cam adre ía_d ue_c _ni_ no_p nte,
1001–1010	char 4-gram	o,_q s,_a _tem an_c s,_l olvi s,_q uego aber ber_
4091–5000	char 4-gram	neci pitá siti rmas ría, lia_ rici no_f _obr vend
1–10	char 4-gram (*word*)	o_de a_de s_de e_la de_l e_el de_e a_el ente ando
101–110	char 4-gram (*word*)	el_a ones enta n_qu os_e ue_n ntra o_nha ar_s mbre
501–510	char 4-gram (*word*)	nera ar_d otra ra_p o_re s_ve es_l s_el ende echa
1001–1010	char 4-gram (*word*)	nqui dich ía_q casi sibl feli l_mi alid uant dici
4091–5000	char 4-gram (*word*)	rcan masi idam clas ifes mbia pica e_be mó_e uili
1–10	char 4-gram (*affix-punct*)	ente ando para ment esta ento ante como habí cont
101–110	char 4-gram (*affix-punct*)	ismo hast pera ensa algu acio ales hora prim vida
501–510	char 4-gram (*affix-punct*)	grad inad cura cult rinc ingu vers engo camp oras
1001–1010	char 4-gram (*affix-punct*)	sueñ dorm fere ritu alud denc creí pend saca lara
4091–5000	char 4-gram (*affix-punct*)	dudo line aza. nel, mutu sinu genu oso; iro, decr

Table 28. Most frequent tokens.

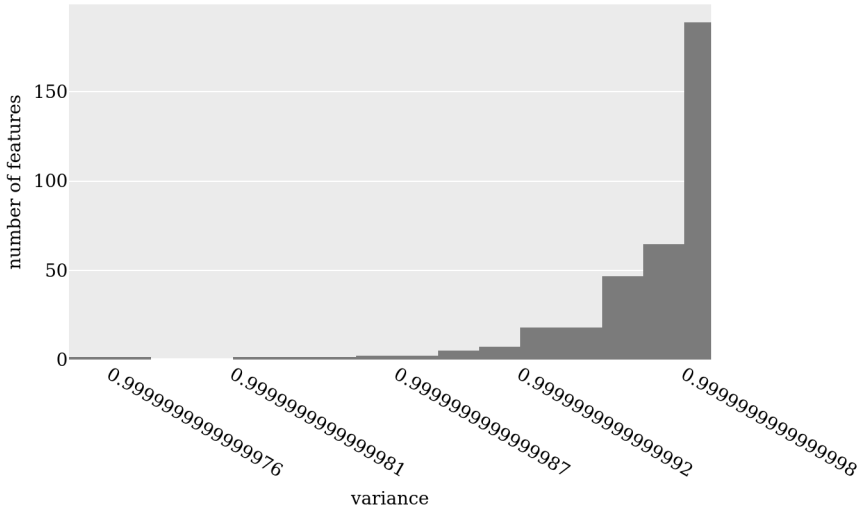


Figure 42. Variances of the 1000 MFW (z-scores).

specific verb forms (“pensar”, “dirigió”), and adjectives (“alto”, “última”). On the positions 1001–1010 and 4091–5000, there are nouns with quite specific meanings (“mozo”, “batalla”, “cólera”, “injusticia”, “sensaciones”, “cuadra”, “distinción”), verbs with more different tenses (gerund: “hablando”, imperfect: “daban”, participle: “apoyada”), and also adjectives and adverbs with increasingly specific meanings (“tranquilo”, “bendito”, “insignificante” and “no_obstante”, “poco_a_poco”, “personalmente”). The word lists from the different frequency sectors show how the feature space gets more complex in terms of the grammatical forms of words that are included and also regarding the semantic specificity of the words as more most frequent items are included.

In the word 2-grams, function words dominate the most frequent tokens in all frequency ranges. On positions 1–10, the 2-grams consist entirely of function words. In the range 101–110, the particle “no” (“si_no”, “pero_no”), a noun (“la_joven”), and an adjective (“todas_las”) appear as one part of the word combinations. The first verbs are visible in the range 501–510 (“dijo_que”, “era_de”, “que_tiene”). A 2-gram without function words appears in the 4091–5000 most frequent items (“sin_querer”). So due to the combination of two words, the semantic specificity of individual words increases later than for the one-word unit. As for the character n-grams, it is directly visible that the most frequent items of the classic n-grams contain many combinations of white spaces and short function words (e.g., “_de_”, “o_de”, “a_de”). Nouns and verbs can be recognized from the 501–510 MFW group onwards (“adre”, “olvi”, “aber”, “rmas”). The special types of n-grams – the ones only including mid- and multi-word n-grams (“word”) and the ones consisting only of prefixes and word endings with punctuation marks (“affix-punct”) – are more homogeneous throughout the different frequency ranges.

The general features presented here are open with regards to the grammatical and semantic characteristics and have the advantage that they cover a broad spectrum of the linguistic material in the full-text basis, the extent of which depends on the chosen frequency range and token unit. Therefore, they lend themselves well to exploratory analyses. A disadvantage coming along is that

these features can be difficult to interpret. In the high frequency ranges, this is because semantic elements are scarce, and in the broader frequency spans, it is because of the heterogeneity of the features. The preparation of the semantic features that are used as a counterweight and an alternative to these general features is outlined in the next chapter.

4.2.1.2 Semantic Features: Topics

The method used to determine thematic elements in the texts is topic modeling, a text-mining method that is unsupervised and not deterministic. Unsupervised means that the set of topics is not predefined but emerges from the text collection that is analyzed. A method that is not deterministic does not produce the same output every time it is run, even if the start conditions are fixed, which means that an element of randomness is involved in the process. The goal of topic modeling is to uncover hidden semantic relationships between the words used in the documents of a large text collection and thereby determine what the texts in the collection are about.[528] The semantic relationships between the words are described as *hidden* because a topic model involves the basic units *corpus*, *document*, *word*, and *topic*, of which the first three – in their input version – are predefined structures and processable parts of the text collection's initial state. The topics, on the other hand, constitute a medium layer between the words and the document, or the words and the corpus. The topic layer is inferred from the distributional characteristics of the words in the documents with a probabilistic approach but is not inherently manifest in the text strings and their structural organization. The basic idea comes from the field of distributional semantics and was pointedly put by J. R. Firth already in 1957: "You shall know a word by the company it keeps!" (Firth 1968, 179).[529] The essence of the quote is that words that repeatedly precede and follow a particular word of interest in a defined textual context contribute to its meaning so that words that occur in the same contexts tend to have similar meanings. This distributional hypothesis is also the theoretical basis for topic modeling.[530] Words that occur in the same context in many documents of a text collection are grouped together into topics because it is assumed that their meanings are related.[531]

---

[528] The method was initially developed in the context of Information Retrieval as a general approach to model text corpora: "The goal is to find short descriptions of the members of a collection that enable efficient processing of large collections while preserving the essential statistical relationships that are useful for basic tasks such as classification, novelty detection, summarization, and similarity and relevance judgements" (Blei, Ng, and Jordan 2003, 993).

[529] The cited version of the paper is a reprint of Firth 1957. The wider context of this quote is: "The *placing* of a *text* as a constituent in a context of situation contributes to the statement of meaning since situations are set up to recognize *use*. As Wittgenstein says, 'the meaning of words lies in their use.' […] The day-to-day practice of playing language games recognizes customs and rules. It follows that a text in such established usage may contain sentences such as 'Don't be such an ass!', 'You silly ass!', 'What an ass he is!' In these examples, the word *ass* is in familiar and habitual company, commonly collocated with *you silly—*, *he is a silly—*, *don't be such an—*. You shall know a word by the company it keeps! One of the meanings of *ass* is its habitual collocation with such other words as those above quoted" (Firth 1968, 179).

[530] For a general discussion of the foundations of the distributional hypothesis, see Sahlgren 2008.

[531] What exactly the "context" is and which scope it has depends on the implementation of the distributional analysis. The two main approaches are to either use surrounding words or to use text regions in which the words occur together as a basis. Current implementations of topic modeling, for instance, Latent Dirichlet Allocation (LDA), usually follow the latter strategy (Sahlgren 2008, 33). Depending on the kind and size of the context, Sahlgren

In the topic modeling approach, the terms "word", "document", "corpus", and "topic" have specific meanings. *Words* correspond to tokens that the text is split into and the tokens that are selected for the topic modeling analysis. They can but need not concur with words in a linguistic sense, depending on how the tokens are defined. Often a list of stop words is applied before the topic model is built so that not all the words that are part of the initial text string stay in the resulting model. The goal is to remove words that carry little semantic value, such as function words, in order to get topics that consist mainly of content words whose distributional relationships are easy to interpret as being of a semantic kind. The term "document" has different meanings, depending on the level on which it is used in a topic modeling workflow. In the narrow sense, "document" is defined as a structure resulting from the topic modeling process. In a wider sense, though, "document" also has a special meaning as an input structure for the topic modeling algorithm. To be able to differentiate between these different uses of the term "document", they are distinguished here by indexing the term. For the purpose of this dissertation, $document_S$ means the original source document in the form of a continuous text string of characters, $document_{In\text{-}1}$ is the input document that the topic modeling algorithm gets, $document_{In\text{-}2}$ is the input document that the topic modeling algorithm creates internally and uses as a basis for the modeling process, and $document_{Out}$ is the output document resulting in the final topic model.

$Documents_{In\text{-}1}$ are the text units that constitute the contextual frame in which the co-occurrence of the words is analyzed and counted. They can but need not correspond to entire $documents_S$ of a text collection such as books or articles. Alternatively, they can be combinations of $documents_S$ or subparts of them. In practice, the size of the $documents_{In\text{-}1}$ for topic modeling is often chosen with the aim of balancing the length of the text units and optimizing them for the algorithm to produce good results. Usually, very long texts and texts with a significant variance in length, such as novels, are segmented into smaller units.[532] For a topic model, the $documents_{In\text{-}1}$ are not kept as sequences of words, punctuation marks, and blank spaces. Instead, internally, they are converted to a collection of word counts following the bag-of-words approach, the $documents_{In\text{-}2}$, as demonstrated with the sentence in example 50 and the resulting word count matrix in table 29.

```
El 4 de mayo de 1840, a las diez y media de la noche, seis hombres atravesaban el patio
    de una pequeña casa de la calle de Belgrano, en la ciudad de Buenos Aires.
```

Example 50. Example sentence from the novel "Amalia" (1855, AR) by José Mármol.

In the example, the proper names "Belgrano" and "Buenos Aires" are treated as stop words and are not included in the word count matrix. Furthermore, all the tokens are converted to

---

differentiates between two types of semantic similarity. According to him, a wider, document-oriented context leads to models that capture "semantic relatedness (e.g. 'boat' - 'water')", while a narrower, word-oriented context models "semantic similarity (e.g. 'boat' - 'ship')" (Sahlgren 2015). As topic modeling is document-oriented, topics are characterized by semantic relatedness.

[532] So far, the decisions for a certain length of the $documents_{In\text{-}1}$ follow empirical experience or rules of thumb, because there are no theoretical foundations for this parameter of text preprocessing for topic modeling yet. In a survey on LDA-based topic modeling in Digital Humanities, Du notes: "The common preprocessing procedures include lemmatization, part-of-speech (POS) tagging and document chunking. [...] Chunking allows us to capture topics which only appear at certain points. My survey pays particular attention to the reasons of applying (or not) a preprocessing procedure in practice. [...] Document chunking is very diverse: the chunk-size could be several hundred or several thousand words, or a page of a book, or to split a book into ten equal segments. But almost no approach explained the reason of their chunking choices" (Du 2019).

4	1840	a	atravesaban	calle	casa	ciudad	de
1	1	1	1	1	1	1	7
diez	**el**	**en**	**hombres**	**la**	**las**	**mayo**	**media**
1	2	1	1	3	1	1	1
noche	**patio**	**pequeña**	**seis**	**una**	**y**		
1	1	1	1	1	1		

Table 29. Word count matrix for the first sentence of the novel "Amalia" (1855, AR) by José Mármol.

calle	casa	ciudad	hombre	mayo	noche	patio
1	1	1	1	1	1	1

Table 30. Word count matrix with word lemmas.

lowercase before counting them. Punctuation marks and blank spaces are removed, and the order of the words is suspended. Already in this single sentence, the counts give an impression of the kind of words that dominate from a quantitative point of view: "de" (7), "la" (3), and "el" (2), which are all function words. Table 30 shows how the matrix is further reduced if the sentence is preprocessed by only selecting the lemmas of nouns.

It becomes clear that the decisions on how to preprocess the texts significantly influence what a word and a document_S become in the topic modeling. In the case of lemmatization, also morphological information is lost, and by selecting only one type of word category, the initial sentence is reduced to content buzzwords. Furthermore, it is not only the structure inside of sentences that is reduced. The sentences themselves and higher-order structures inside of the documents_S are also not preserved.

In the context of topic modeling, also the term "corpus" has several meanings if both the input and output states of the model are considered. As for the documents, I differentiate between $corpus_S$ as the collection of documents_S, i.e., the original full-text files, $corpus_{In\text{-}1}$ as the collection of $\text{documents}_{\text{In-1}}$, that is, the set of full-text snippets that the topic algorithm receives as an input and that already can have been preprocessed by stop word removal, linguistic annotation and selection, or chunking. The $corpus_{In\text{-}2}$ is the collection of $\text{documents}_{\text{In-2}}$, which is a matrix of terms and their counts per document.[533] Finally, there is also a $corpus_{Out}$, which is part of the final topic model.

The term "topic" is only defined on the output level, as this structure does not exist on the input levels and is found by the algorithm. Technically, a topic is a probability distribution over words. For each word that is part of the $\text{corpus}_{\text{In-2}}$'s vocabulary[534], a probability value indicates how important it is for the topic at hand (Steyvers and Griffiths 2007, 2). Table 31 illustrates the

[533] This is the so-called "term-document matrix", "a data structure, a computationally tractable (to use McCarty's term) representation of the texts able to be modeled by a computational process" (Burton 2013).

[534] Here, "vocabulary" means the set of different words contained in the corpus.

Topic 77			Topic 10		
Word	**Probability**	**Count**	**Word**	**Probability**	**Count**
amor	0.09260	8839	bandido	0.05592	763
corazón	0.06372	6082	jefe	0.02844	388
alma	0.03663	3496	ladrón	0.02375	324
pasión	0.02052	1959	robo	0.01752	239
felicidad	0.01882	1796	policía	0.01708	233
palabra	0.01761	1681	crimen	0.01700	232
ser	0.01235	1179	compañero	0.01634	223
sentimiento	0.01187	1133	bandolero	0.01158	158
flor	0.01173	1120	lugar	0.01099	150
mirada	0.01113	1062	justicia	0.00880	120
pensamiento	0.01009	963	sociedad	0.00777	106
placer	0.00950	907	camarada	0.00770	105
esperanza	0.00819	782	silencio	0.00770	105
cielo	0.00777	742	comandante	0.00696	95
ángel	0.00776	741	camino	0.00660	90

Table 31. Top 15 words of two example topics.

structure of topics by showing the 15 top words of two different topics that are part of a topic model created for the corpus Conha19.[535]

An important insight is to realize that each topic consists of all the words in the corpus vocabulary and that the differences between the topics are the result of the different weights that the words have in each topic. When humans interpret topics, usually, they only examine a certain number of words – those with the highest probability. The semantics of the topics emerges from the combination of high-probability words. A human inspector can evaluate the semantic relationships between the terms and label the topics with titles. A topic title can consist of individual words but also descriptions. In the above example, topic 77 could be entitled “love” or “love and feelings” and topic 10 “crime” or “crime and society”. Computers can also evaluate the resulting topics by calculating the similarity and semantic relationships of the most important terms in each topic by using external semantic resources, for example, word embeddings,[536] word nets,[537] or dictionaries. Computers could also label topics, for example, by finding superordinate

[535] The example topic model is available at https://github.com/hennyu/papers/tree/master/family_resemblance_dsrom19/features/topicmodel. Accessed November 12, 2020. It was created with the following parameters: 100 topics, 5,000 iterations, and a hyperparameter optimization interval of 100. The texts were preprocessed by lemmatizing them and only using nouns. The original input full-text files were chunked into segments with a length of 1,000 tokens. A list of stop words was applied, which contained the 50 most frequent nouns plus some nouns that were added manually. The structure of the table corresponds to the examples shown in Steyvers and Griffiths (2007, 2).

[536] Word embeddings are another method from the area of distributional semantics. In word embeddings, words from a vocabulary are converted to vectors of real numbers, and semantic relationships can be inferred from the proximity or distance of the word vectors to each other (Sahlgren 2015).

[537] Word nets are lexical databases of semantic relations between words, for example, synonymy or hyponymy (Miller 1995).

Document number	Document name	Topic 0	Topic 1	Topic 2	Topic 10	Topic 77
0	nh0217§0030.txt	0.04178	0.00009	0.00003	0.00362	0.00029
1	nh0107§0052.txt	0.00002	0.00384	0.00004	0.00004	0.00406
2	nh0103§0012.txt	0.00002	0.00010	0.00004	0.00005	0.14777

Table 32. Example $documents_{Out}$ from a topic model.

terms. Nevertheless, for creating the topics, no explicit semantic knowledge is necessary because the algorithm bases the assumptions on the relationships between the terms in a topic only on statistical, distributional patterns. Another advantage of topic modeling is that polysemy is not problematic. As every word occurs in every topic, the same word can be important in several topics where its meaning is determined by the surrounding words. In the table, the "probability" column indicates how probable a word is in a specific topic, and the "count" column shows how often a word has been assigned to a topic across all documents. When the counts are compared, the love topic is more important in the corpus than the crime topic because the number of tokens assigned to it is higher. In total, the love topic has 95,452 token assignments, and the crime topic has only 13,644.[538] Inside each topic, the probabilities indicate the relative importance of each word. In the love topic, the first words have higher probabilities than the first words of the crime topic. However, the differences between the probabilities decrease for the lower word ranks, which means that the love topic is more dominated by a few very important words than the crime topic.

In the topic model output, the $document_{Out}$ is a probability distribution over topics. For each topic that was defined for the corpus, a probability value is given for every document, and topics with higher probabilities are considered especially relevant for the document in question. It follows from this that the $corpus_{Out}$ is a matrix of probability distributions over topics. The structure of three example $documents_{Out}$ is given in table 32, showing the probabilities of five selected topics of the model.[539]

Here, the names of the documents are a combination of the novels' identifiers in Conha19 ("nh0217") and the numbers of the text segments that served as $documents_{In\text{-}1}$ ("§0030"). To be able to evaluate the probabilities of the topics in the entire novels, the values for the individual segments have to be aggregated again, for example, by using average probabilities. The different probability values show how important the topics are in the $documents_{Out}$. Regarding the love topic (topic 10) and the crime topic (topic 77), the table shows that the segment of the first document in the list ("nh0217§0030") has a higher probability for the love topic than the other two segments from other documents and the third segment ("nh0103§0012") has a higher probability for the crime topic than the preceding ones. The identifier "nh0217" belongs to the novel "El espejo de Amarilis" (1902, MX) by Laura Méndez de Cuenca and "nh0103" to the novel

[538] This information about token counts per topic can be found in the diagnostics file of the tool MALLET, which was used to create the model (McCallum 2018a).

[539] See footnote 535 above for information about the underlying topic model. In the table, the topic probability values are rounded.

"El mendigo de San Ángel" (1865, MX) by Niceto de Zamacois. The first one is a novel of customs involving romantic plot elements, and the latter is a historical novel. The results of the topic model seem reasonable, since a prominent love topic can be expected in a novel with a romantic plot, and it is also plausible that the crime topic would carry greater weight in a historical novel.

The topic model was created with the tool MALLET (McCallum 2002), in which the topic modeling algorithm Latent Dirichlet Allocation (LDA) is implemented (Blei, Ng, and Jordan 2003; Blei 2012). Other mathematical and technical approaches to topic models exist, but LDA is the most prominent current technique and is widely used.[540] LDA is based on a generative model, assuming the documents are generated based on the topics. The method works by choosing a distribution over topics for each document and subsequently choosing a topic from this distribution for each word in each document. Finally, a word is chosen from the topic's distribution over words. The process starts with an initial distribution[541] and approximates it to the data by iterating over the words in the corpus.[542] The most important parameter that needs to be set for the algorithm is the number of topics to model.

For humanists who want to evaluate results obtained from topic modeling or who are interested in using the method to produce their own results, it is fundamental to recognize the different meanings of the basic terms *word*, *document*, *corpus*, and *topic* in the context of topic modeling compared to the context of thematic analyses based on linguistic and literary theories. In the "Reallexikon of German Literary Studies", for instance, the term "Thema" is defined as "Die einem Text zugrundeliegende Problem- oder Gedankenkonstellation" (Schulz 2007, 634), which means the central topic or subject of a text and corresponds roughly to the English term "theme". This sense of *topic* is different from the $\text{topics}_{\text{TM}}$ resulting from topic modeling[543] in that it is defined on a more abstract level. The underlying theme of a text can be interpreted from its linguistic material but does not necessarily have to be directly present in terms of formulations using the word or words that describe the theme. For example, one could think of a short story describing a romantic dinner and containing dialogues of a couple, but in which the word *love* is never used. Nevertheless, a reader could conclude that the central topic of the story is just love.

The $\text{topics}_{\text{TM,}}$ on the other hand, are closer to the textual surface. A topic model captures how topics are realized in text segments. However, to find one or several central themes of a text, the $\text{topics}_{\text{TM}}$ need to be interpreted on a more general level.[544] A term that is better suited to be related to the $\text{topics}_{\text{TM}}$ is the text-linguistic term *thematische Entfaltung*. It focuses on how the

---

[540] A forerunner of LDA is, for example, Latent Semantic Analysis (LSA), which is not probabilistic (Landauer and Dumais 1997; Landauer, Foltz, and Laham 1998). Newer approaches are Nonnegative Matrix Factorization (NMF) and a network-based approach using a stochastic block model (SBM) (Arora, Ge, and Moitra 2012; Gerlach, Peixoto, and Altmann 2018). In the context of deep learning, the method lda2vec has been developed, which combines the learning of word, document, and topic vectors (Moody 2016).

[541] This is the Dirichlet distribution which gives the algorithm its name (Blei, Ng, and Jordan 2003, 996–997).

[542] This process is called Gibbs sampling (Steyvers and Griffiths 2007, 7–9).

[543] The index *TM* is used here to differentiate the general term *topic* from the term as it is defined in the context of the topic modeling method.

[544] For a useful discussion of the similarities and differences of $\text{topics}_{\text{TM}}$ to literary theoretical terms on thematic aspects, see Horstmann (2018). Besides the term *Thema*, Horstmann also compares the German terms *Motiv*, *Topos*, and *Sujet* to $\text{topics}_{\text{TM}}$ and comes to the conclusion that they all cover different thematic and content-related aspects of literary texts.

theme or central topic of a text is unfolded in the overall content of the text. Brinker, Cölfen, and Pappert (2014, 57–80) define the thematic unfolding as a combination and linkage of relational and logical-semantically defined categories that express the relationship of partial topics present in individual parts and substructures of the text to its central topic. They mention justification and specification as examples of such relational categories. Obviously, the same central topic can be unfolded in different ways. Brinker, Cölfen, and Pappert also define a set of basic forms of thematic unfolding: descriptive, narrative, explicative, and argumentative thematic unfolding.[545] Brinker, Cölfen, and Pappert stress that the thematic unfolding is influenced significantly by communicative and situational factors, such as the intention or purpose of the communication, which leads to several different possibilities of unfolding the same central topic. However, little is known about the factors which have an effect on the exact unfolding. From the point of view of digital literary studies and especially stylistics, this means that also stylistic intentions can have an influence on the thematic unfolding, or, looked at from another perspective, that the exact thematic unfolding can be interpreted as an intended or unintended stylistic effect. The distribution of $\text{topics}_{\text{TM}}$ in the documentsOut of a topic model could be interpreted as the result of thematic unfolding. On the level of the $\text{corpus}_{\text{Out}}$, the specific set of $\text{topics}_{\text{TM}}$ can be understood as resulting from the thematic unfolding across all documents in the text collection. The thematic unfolding itself, as defined by the combination of relational and logical-semantical categories, though, is located on a more abstract level, which is intermediate between the $\text{topics}_{\text{TM}}$ and the theme of the text. Brinker, Cölfen, and Pappert point out that the theme or central topic of the text is the shortest possible summary of the textual content and that only a reader's interpretation can achieve this reduction:

> Man muss sich überhaupt darüber im Klaren sein, dass die textanalytische Bestimmung des Themas primär auf interpretativen Verfahren beruht; es kann hier keine 'mechanische' Prozedur geben, die nach endlich vielen Schritten automatisch zur 'richtigen' Themenformulierung führt. […] Die Bestimmung des Themas ist vielmehr abhängig von dem Gesamtverständnis, das der jeweilige Leser von dem Text gewinnt. Dieses **Gesamtverständnis** ist entscheidend durch die beim Emittenten vermutete Intention bestimmt, d. h. durch die kommunikative Absicht, die der Sprecher / Schreiber mit seinem Text *nach der Meinung des Rezipienten verfolgt.* (Brinker, Cölfen, and Pappert 2014, 53–54)

Regarding the possibilities of a mechanical (or computational) method to determine the theme of a text, I take a more moderate position. Even if a reader is indispensable as the last instance of interpretation, formal methods can be used to evaluate how a theme is thematically unfolded in a text. Furthermore, the basic types of thematic unfolding have been taken up by Schöch and Rißler-Pipka in a topic modeling analysis of literary texts. In a contribution to a conference

---

[545] In German, the term *Themenentfaltung* is defined as "die gedankliche Ausführung des Themas" and, more specifically, "Die Entfaltung des Themas zum Gesamtinhalt des Textes kann als Verknüpfung bzw. Kombination relationaler, logisch-semantische definierter Kategorien beschrieben werden, welche die internen Beziehungen der in den einzelnen Textteilen (Überschrift, Abschnitten, Sätzen usw.) ausgedrückten Teilinhalte bzw. Teilthemen zum thematischen Kern des Textes (dem Textthema) angeben (z. B. Spezifizierung, Begründung usw.)" (Brinker, Cölfen, and Pappert 2014, 57). These concepts go back to Brinker (1992).

panel on drama analysis, Schöch and Rißler-Pipka analyzed the distribution of argumentative, narrative, descriptive, and discursive topics in 1,100 novels, 800 dramatic texts, and 1.8 million Wikipedia articles in French to see if the proportions of topic types vary by genre.[546] Moreover, different forms of thematic unfolding are also studied in computational studies in general.[547] These approaches perform much of the text-analytical determination of topics and theme and push the role of a human reader or a human interpreter of analytical results to a higher level.

Whatever the relationship between literary-theoretical and linguistic terms that cover content-related aspects and $\text{topics}_{\text{TM}}$, so far, the systematic analysis of content has not been a key concern in literary scholarship:

> 'Inhaltsanalyse' (engl. content analysis) ist eigentlich kein literaturwissenschaftlicher Begriff. Diese [in den Sozialwissenschaften verbreitete] hier stark vereinfacht beschriebene, mit diversen Kontrollverfahren zur Einhaltung von Intersubjektivitätsmaßstäben begleitete, insgesamt sehr aufwändige und inzwischen bei der Texterfassung, -bearbeitung und -auswertung weitgehend computergestützte Vorgehensweise ist lieraturwissenschaftlichen Umgangsformen mit Texten fast völlig fremd. Allenfalls in der Computerphilologie [...] zeigt man sich mit ihnen vertraut. (Anz 2007, 55)

As Anz points out, systematic content analysis has a longer tradition in the social sciences than in literary studies. According to him, the core interest of literary studies is much more the specific literary function of the texts, which manifests itself in particular linguistic and structural forms (Anz 2007, 57). In this regard, the method of topic modeling, with its dependence on the textual surface, is closer to the concern of literary scholarship than content analysis because it allows us to examine the relationships between content and form more directly. In computational literary studies, topic modeling has indeed been taken up with interest and applied to a variety of literary texts with success (see, among others, Jockers 2013; Rhody 2012; Schöch 2017c). Literary scholars that used topic modeling soon discovered that the relationships between the terms in the topics are not necessarily content-related and do not even need to be of a semantic nature. When applied to literary texts, topic modeling can also discover rhetoric structures and elements of discourse (Jannidis 2016, 27).[548] That such structures were especially noticed when topic modeling was applied to literary texts shows that the method, which originally aimed at enabling content analysis, has been developed and optimized on the basis of non-literary text types. It also indicates that the distributional hypothesis could be understood in a wider sense as not only applying to semantic relationships but also to discourse relationships. However, to what extent non-thematic topics can be produced also depends on the kind of preprocessing applied to

---

[546] For the abstract of the conference panel, see Willand et al. (2017). The results of the contribution by Schöch and Rißler-Pipka about topic types can be found in the presentation, which can be downloaded at https://github.com/christofs/dramenanalyse-dhd/. Accessed November 14, 2020.

[547] See, for example, the study on Argument Mining using topic modeling presented by Lawrence and Reed (2017).

[548] The method has even been applied intentionally to discover non-thematic structures, for example, by Rhody: "By locating 'figurative language' as an aspect of address for topic modeling, I choose to constrain my consideration of poetic texts and agree to a caricature of poetry that hyper-focuses on its figurative aspects so that we can better understand how topic modeling, a methodology that deals with language at the level of word and document, can be leveraged to identify latent patterns in poetic discourse" (Rhody 2012).

Parameter	Selected values
Number of topics	50, 60, 70, 80, 90, 100
Optimization interval	50, 100, 250, 500, 1000, 2500, 5000, None

Table 33. Parameters for topic feature sets.

the texts. If only nouns are processed, it is more probable that the words in the topics are held together by semantic, thematic, and content-related connections.

For the analysis of subgenres of the novel pursued here, this traditional way of using topic models is intended. Therefore, the linguistically annotated version of the novels is used, and only noun lemmas are selected, aiming at semantic and content-related topics. This way of preprocessing the texts has also been applied in other digital literary studies that used topic modeling for genre distinction and categorization.[549] Primarily stylistic aspects, on the other hand, are intended to be covered by the general, most frequent words-based features, so that the noun-based topics function as a counterpart and as semantic features. As for the general features, also for the topics, several feature sets were prepared using different parameters, as summarized in the following table 33.

The first parameter that was varied is the number of topics. Here, a range of 50 to 100 topics is considered reasonable for the corpus with 256 novels. In his study of French classical drama, Schöch (2017c) used 60 topics for a collection of 391 novels. Schöch, Henny et al. (2016) used 70 topics for a corpus of 150 Spanish and Spanish-American nineteenth-century novels, and Hettinger et al. (2016) found the best performance for classifying 628 German novels with 100 topics. In general, a lower number of topics produces more general results, and a higher number covers more specific thematic aspects of a text collection.

The second parameter that was selected for variation here is the optimization interval, that is, the interval at which the Dirichlet hyperparameters for the LDA model are optimized during the iterations of the topic modeling process. The two hyperparameters, alpha (α) and beta (β), influence the form of the topic probability distribution over words and the document probability distribution over topics, respectively. The lower the parameters are, the more the resulting distribution is concentrated on single values. Otherwise, it is more even. It follows that a lower alpha value leads to topics with a few dominant and many unimportant words and a lower beta parameter to documents that are dominated by a few topics rather than by a larger number of roughly equivalent topics. In the MALLET implementation of LDA, which was used here,[550] the parameter *optimize-interval* indirectly influences the form of the two types of distribution. A lower rate of optimization results in more even distributions and a higher value in more skewed ones. By not setting this parameter at all, no hyperparameter optimization is performed, which leads to even probability distributions (McCallum 2018b; Wallach, Mimno, and McCallum 2009). Schöch (2016, 2017c) has analyzed how different optimization intervals influence the results

[549] For instance, only noun lemmas are used in Hettinger et al. (2016) and Schöch, Henny et al. (2016). Noun, verb, adjective, and adverb lemmas are used by Schöch (2017c).

[550] MALLET was used in version 2.0.8RC3.

for subgenre classification in French novels and drama.[551] The range of intervals that has been selected here varies from 50 to 5,000 and also includes None. With a fixed number of 5,000 iterations, this results in 50, 20, 10, 5, 2, 1, and no optimizations.

As to the preprocessing of the texts, the following procedure was followed: first, the linguistically annotated versions of the TEI corpus files were used to extract the noun lemmas for each novel and to create full-text files only containing the nouns. As for the general features, also here proper names were excluded.[552] Then, the novels are segmented into chunks of 1,000 tokens, which are fed into the topic modeling workflow. A stopword list consisting of the 50 most frequent nouns and of proper names and place names was created and used by MALLET.[553] For text segmentation, modeling, and post-processing of the results, the tool tmw (Topic Modeling Workflow) was used. Tmw was developed by Christof Schöch and Daniel Schlör in the context of the CLiGS project. It is a set of python scripts that especially supports the pre- and post-processing of the texts that are used for the topic modeling. It also includes a set of functions to visualize the topic modeling results. Its development is organized in a GitHub repository (see Schöch and Schlör 2017).[554] The tmw scripts have been slightly adapted to create the topic models used here. Mainly the routines for calling the core functions were changed.[555] An important feature of tmw is the text segmenting procedure. Before the novels are processed with MALLET, they are split into segments of a defined length,[556] so that the topic models are created for the segments, not the entire texts. In the post-process step, tmw recombines the segments of the novels and calculates average probability values. The segmenting step is considered important to counter the effects of different text lengths, and the recombination of the segments is necessary to be able to interpret the results per novel. Finally, because topic modeling is not deterministic, five different models were produced for each parameter constellation.[557] Through the different

---

[551] He recommends setting the optimization interval depending on the goal of the analysis: "If your goal is to identify small numbers of texts about specific themes in a large collection, then a lot of optimization may be good. However, if your goal is to identify topics typical of certain authors, periods, genres or some other reasonably large subset of your collection, then it may be better to optimize a bit less" (Schöch 2016).

[552] The script for creating the noun lemma files is available at https://github.com/cligs/scripts-nh/blob/master/corpus/derivative_formats/get-plaintext-annotated-nouns.xsl. The resulting plain-text files are available at https://github.com/cligs/conha19/tree/master/txt_annotated_nouns. Accessed November 15, 2020.

[553] The proper names and place names in the stopword list can cover cases that were missed by the named entity recognition. The stopword list for the topic models is available at https://github.com/cligs/data-nh/blob/master/analysis/features/stopwords/topics_stopwords.txt. Accessed November 14, 2020. The list was refined by inspecting selected topic models and adding words with very general meanings that dominated some topics and reduced their interpretability.

[554] The tool was presented at a workshop at the DH2017 conference in Montréal by the CLiGS team (Betz et al. 2017). A more recent presentation about topic modeling that includes a description of tmw is Schöch (2019).

[555] The version of tmw which was used here is available at https://github.com/cligs/scripts-nh/tree/master/features/tmw. Accessed November 14, 2020.

[556] Tmw is even able to apply a smooth segmenting that respects paragraph boundaries, but this feature was not used here. Instead, a fixed segment length was defined.

[557] The script that applies the different parameter settings, controls the topic modeling workflow, and calls the functions of tmw is available at https://github.com/cligs/scripts-nh/blob/master/features/topics.py. The resulting topic models are available at https://github.com/cligs/data-nh/tree/master/analysis/features/topics/. Accessed November 14, 2020.

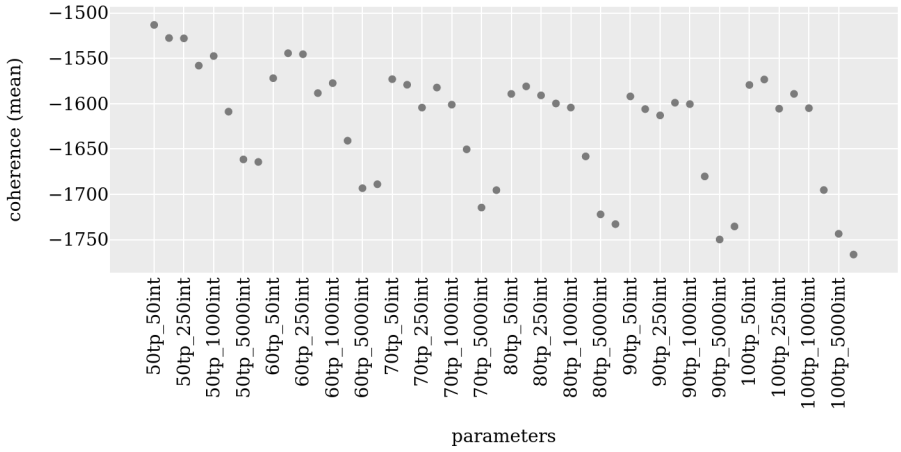


Figure 43. Mean coherence of the topic models with different parameter settings.

combinations of topic modeling parameters and the repetitions of the modeling process, 240 different topic models were created for the whole corpus.

As for the MFW-based features, also for the topics, some general characteristics of the feature sets are evaluated here. Unlike for MFW, the number of zero values is not relevant for the topics, though, as no zero values were found in the whole set of topic models created for the corpus of Spanish-American novels. This means that each topic has at least some probability in each document, even if it is very small.[558] Another aspect of the models that was evaluated is the coherence of the topics. The tool MALLET includes a set of diagnostic measures which can be saved as an additional output of the topic modeling process and can be used to assess the formal quality of the topic models. One of the measures is the coherence metric, which evaluates whether the words in a topic actually occur together in the texts of the corpus. This is examined by taking each pair of top words in a topic and calculating "the log of the probability that a document containing at least one instance of the higher-ranked word also contains at least one instance of the lower-ranked word" (McCallum 2018a). The resulting scores are negative, and values closer to zero mean that words co-occur more often. For the models at hand, the number of top words was set to 50. As the coherence measure depends on the top words, it is expected that lower numbers of top words lead to coherence scores that a closer to zero than the scores for models with a higher number of top words. However, as the number of top words is the same for all the models here, the resulting coherence values can be compared. A summary is given in figure 43.

For the figure, the mean coherence of all the topics in a topic model was calculated. This was done for each combination of the number of topics and optimization intervals. As there are five different models for each setting, their mean coherences were averaged. The results show two trends. First, the coherence of the topics tends to decrease with an increasing number of topics,

[558] The charts summarizing this finding can be viewed at https://github.com/cligs/data-nh/tree/main/analysis/features/topics/overviews. Accessed December 22, 2020.

Figure 44. Example topics.

and second, the coherence also decreases when the optimization intervals are bigger (so that the hyperparameters are optimized less often). What can also be seen is that the values drop more for bigger optimization intervals than for higher numbers of topics. It makes sense that less optimization leads to topics that are "further away" from the actual distribution of words in the corpus so that the topics tend to be more abstract. It is more difficult to explain that higher topic numbers produce less coherent topics. A hypothesis for this is that with more topics, many of them are more special, so it is less probable that texts containing some top words of a topic also contain all the others. Interestingly, though, the coherence values go up again from 90 to 100 topics for the lower optimization intervals. Regarding the usage of the topic models for subgenre categorization, the coherence values are of interest in two ways. One is the question of whether models with more coherent topics are better feature sets for classification, and the other is whether a better formal coherence of the topics also facilitates their interpretation as features related to the subgenres. Four example topics are shown in figure xxx to get an impression of the kind of topics that are part of the different models.

Each word cloud depicts one topic and includes its 40 most important words. The top left topic stems from a model with 50 topics and an optimization interval of 50, and the top right from a model also with 50 topics but without hyperparameter optimization. The bottom left topic is part of a model that has 100 topics and an optimization interval of 50, and the bottom right topic

was produced in a model also with 100 topics but without hyperparameter optimization. With optimization, the topics can be ranked because they can have different overall probabilities in the corpus. The tenth most important topic was selected from the two models with optimization (as indicated by the first number in parentheses after the topic number). Although the overall number of topics is different in the two models, the tenth topic is similar in both. Without optimization, all the topics in the model have the same weight in the whole corpus, so of the two models without optimization, two topics with similar words were selected. The two topics to the left can be interpreted as representing a ball. In the top left topic, the three most important words are "baile", "salón", and "amor", and in the bottom left one, "baile", "salón", and "brazo". The other words in the topics are also related to a ball situation and cover music, dance, social life and conversation, and dresses. The main difference between the topic from the 50-topics model and the one from the 100-topics model is that in the former, words about love and family relationships are more prominent ("amor", "mamá", "novio", "corazón", and "papá"), suggesting that here the ball situation is connected with central plot elements, whereas the latter has a stronger focus on the ball as a social event and seems to be more descriptive, and hence more specialized ("brazo", "traje", "movimiento", "efecto", "concurrencia", "reunión"). The two topics to the right can be interpreted as covering travel to or in a city. In the topic from the 50-topics model, the three most important words are "coche", "carruaje", "ciudad", and the three top words in the topic from the 100-topics model are "coche", "carruaje", and "cochero". As in the case of the ball topics, also here the first topic includes some top words which add elements from another type of situation. The word "hotel" is more important in that topic than in the second one, and there are the words "mesa", "juego", "jugador", "reloj", and "oro", which seem to describe a situation at a gaming table. Such a subtopic is not noticeable in the second topic. Overall, from the point of view of a human interpreter, all four topics stemming from models that were created with different topic modeling parameters have good quality and are semantically coherent, with differences only on the level of detail. Interestingly, the two topics from the 100-topics model give a more coherent impression than the two topics from the 50-topics model, which is contrary to the outcome of the MALLET diagnostics and shows that the degree of formal coherence does not necessarily coincide with the semantic coherence that can be observed in the topics by a human interpreter. In any case, topics as semantic features are mostly easy to interpret, so it can be expected that the feature sets resulting from topic modeling are useful for gaining new insights into the subgenres of the novel.

4.2.2 Categorization

Two types of categorization methods are used to analyze the novels of Conha19. These are based on the sets of general features (most frequent words), and of semantic features (topics) that were presented in chapter 4.2.1 above. As a first step, a classification is applied. It aims primarily to select the best feature set of each type (general and semantic). It also has the purpose of examining how well the novels can be classified by subgenre at all. This is done in chapter 4.2.2.1. As a second step, a family resemblance analysis is conducted in chapter 4.2.2.2, which is based on text similarity calculations and rankings, network analysis, and community detection. This specific combination of techniques is proposed here as one possible implementation of the family resemblance concept. In contrast to the classification, the family resemblance analysis does not

start from predefined subgenre labels. It is an open and exploratory technique that uses the similarities between the texts of the novels – as represented in the feature sets – to group them. Only afterwards are the resulting *families* of texts compared to the subgenre labels and other metadata categories. By using the kind of feature sets that were successful also in the classification tasks, the family resemblance analysis starts from a reliable basis regarding the relevance of the features for categorization by genre. On the other hand, the influence of other contextual and textual factors, such as authorship, time period, country, narrative perspective, or setting, on the resulting categories, can be better explored in a bottom-up approach, i.e., in a feature-based categorization approach without prior labeling of the texts. The family resemblance analysis also has the advantage that the categories are constituted based on a network of relationships between individual texts. It does not presuppose that every feature is present in every text in the same way. Partial overlaps in the feature distributions are enough to connect the texts. Like that, texts can be distinct in detail but similar in general and still be grouped together. This is a second kind of openness of the family resemblance method compared to classification. The details about the two methods are discussed in the respective subchapters, including the algorithms and implementations used to apply them.

In the bibliography and the corpus, all kinds of subgenre labels were collected for the novels and sorted according to a discursive model for subgenre terms. A quantitatively relevant selection of them was analyzed on a metadata level in chapter 4.1.5.3 ("Subgenre Labels Selected for Text Analysis"). In this chapter on categorization, further selections are made for constellations of subgenres that are analyzed on the textual level. There are several reasons for the selections. First, critical literature is only available for some of the discursive levels of subgenre terms and only for some of the subgenres on those levels. The existence of critical approaches to the subgenres is important in order to be able to formulate hypotheses based on previous knowledge and research results. In addition it shows which subgenres have been at the center of interest of literary scholars. Referring back to the existing discourse on subgenres of the novel in literary scholarship increases the chance that the quantitative text analytical approaches also find a response there. Furthermore, there is the chance that previous results are confirmed or critically examined from a different methodological standpoint. On the other hand, discursive levels of subgenre terms and specific types of subgenres that have not been the focus of literary scholars yet can be a new ground that is worth exploring with the help of digital text analysis. As formulated in the chapter on the features (see chapter 4.2.1 above), in these cases, the main hypothesis to be checked is whether there is any detectable and significant relationship between the subgenre labels and the texts. Here, the focus is on the quantitatively and qualitatively most relevant and critically established subgenres from the levels of theme and literary current. As thematic subgenres, the three subgenres with the most frequent primary labels in the corpus have been selected: historical novels, sentimental novels, and novels of customs. For the literary currents, romantic, realist, and naturalistic novels are compared.

With this selection, two levels of the discursive model of subgenre terms are covered, and different types of label sources are included. The labels related to theme and literary current are critically established, preferably based on interpretations made by other scholars who have classified the novels in question. If no such classifications were available, the labels were assigned by the author of this study based on explicit and implicit subgenre signals that were collected

and encoded in detail for the novels in the corpus.[559] Another aspect that has been pointed out here is that one novel can have several different subgenre labels, even on a single discursive level. As a simple approach to model multiple subgenre terms of the same kind, one primary label was selected for the levels of theme and literary current, marking the remaining ones as secondary. In the classification analysis, these primary labels are employed. Another option would have been to conduct a multilabel classification, allowing one text to pertain to several different subgenres at once. A text analysis considering this more complex and, at the same time, diffusing modeling of subgenre assignments is left as a future task. Furthermore, the textual analysis of the difference between critically established thematic labels, on the one hand, and purely explicit and historical thematic labels, on the other hand (i.e., above all the *novelas históricas* and the *novelas de costumbres*), is not conducted here. It is assumed that such an analysis will bring to light different nuances of the historical and the current contemporary conceptions of the subgenres, and to capture these on the level of the texts is considered an advanced task for the future. However, as the metadata about the subgenre terms and their assignment to the novels has been captured on all these levels, the information can still be used to analyze their impact as influencing factors on the results of the text analysis, together with the other metadata categories. It can be expected that a reduced setup of subgenre comparisons leads to clearer results and facilitates their interpretation, especially when an open approach such as the family resemblance analysis is used. As there are no simple measures to evaluate the categories emerging from a network-based approach yet, a manual inspection of the results and underlying feature distributions is indispensable in that case.[560]

4.2.2.1 Classification

The method which is used here to group the novels into discrete classes of texts is statistical classification, as it is defined and implemented in the context of machine learning. According to Alpaydin, machine learning is

> programming computers to optimize a performance criterion using example data or past experience. We have a model defined up to some parameters, and learning is the execution of a computer program to optimize the parameters of the model using the training data or past experience. The model may be predictive to make predictions in the future, or descriptive to gain knowledge from data, or both. (Alpaydin 2016, 3)

Machine learning is applied in cases in which it is very difficult or not possible to know the rules that connect a certain type of data to a characteristic that is attributed to this data. In the case of genres, this means that there are texts about which we have some knowledge: we know, for example, which words are used in the texts, how they are organized syntagmatically, to which grammatical categories they belong, and so on. On the other hand, we know that certain texts

---

[559] See chapter 3.2.3 on the assignment of subgenre labels to the novels in the bibliography and chapter 3.3.4 for the novels in the corpus.

[560] As explained below in chapter 4.2.2.2, the categories in the family resemblance network are built through community detection. A discussion of the challenges to evaluating network communities, including a proposition to use ground truth, can be found in Yang and Leskovec (2015).

belong to specific genres because it is indicated in the paratexts of their publication, because the books they are published in are offered in a genre-specific section of a bookstore, because some literary scholar has said that they belong to a genre, etc. We may have some ideas about how the input characteristics of the texts are related to the output labels, but we do not know enough to be able to design corresponding algorithms for the computer that could convert the input (the features) to the output (the category labels) in a direct way.

In machine learning, the computer uses the data that has been labeled with some output value to learn the rules from it. This is done by using a model of a predefined type (for example, a linear model aiming to describe connections in the data with linear relationships), and adapting and optimizing this model by setting its parameters in a way that fits the data and their labels best. The goal of training such a model is to be able to automatically label new cases of the data by applying the trained model to them, i.e., to make predictions. In the case of genre, this would mean being able to assign a genre label to a new, unknown text by handing it to the computer. As Alpaydin states, another goal can be to inspect which part of the data the computer used to make the predictions to learn something about the rules that connect the input data to the output labels. For the genres, this would mean reproducing which features of the texts, for example, the frequency of a certain word or the distribution of word categories, really are relevant for the decision to assign the texts to certain genres.[561]

Approaches in machine learning are broadly classified into unsupervised and supervised learning methods. In an unsupervised case, knowledge is extracted from the data without information about some output value or category. The data may, for example, be grouped based only on how the features are distributed. In a supervised approach, on the other hand, a specific outcome is known for a given input. Classification belongs to the latter group because the goal is to find a model for the relationships between the input data and the output class labels (Müller and Guido 2016, 27, 133).

Usually, the aim of a trained machine learning model, or more specifically, a classifier, is not just to be able to treat the data from which it has learned. Instead, it should be a general model, in the sense that it can differentiate between instances of the classes independently of the specific data set. For the corpus of nineteenth-century Spanish-American novels at hand, for instance, this implies that a classifier trained to recognize the difference between historical and non-historical novels should be able to make correct predictions for another corpus. A different corpus may be built using the same population of nineteenth-century Spanish-American novels as represented in the bibliography but selecting different authors and works. This would mean that the classifier can tell us something about the general difference between historical and non-historical novels of that period and cultural-geographical context, and not just about how the genre *novela histórica* is realized in the specific works in the selected corpus. By training a classifier not just on nineteenth-century but also on twentieth-century novels, one could create a model that is independent of the time period and learn a more abstract data-based concept of the historical novel. Similarly, also the cultural-geographical and linguistic context could be broadened or narrowed down, for example, to only learn about the historical novel in Mexico. As in literary studies in general, these examples show that also in digital literary studies applying

[561] This description of machine learning is based on Alpaydin (2016, 1–3).

machine learning and classification, the design of the corpus is decisive for the kind of conclusions that can be drawn from the text analysis results.[562]

To be able to build a model that is general for the selected context, a supervised machine learning task is usually performed in several successive steps. First, part of the data is used as a *training set* to build the model and find the best parameters for it. Another part of the data is kept as a *validation set* to be able to check how well the model performs if it has to classify data on which it has not been initially trained. Subsequently, the model is refined by adjusting its parameters so that it performs best on the validation set. That way, the validation set becomes part of the training process of the model. By repeating the process of splitting the data into different training and validation sets (*cross-validation*), the effect of random chance in selecting a specific training set can be reduced. The parameters of the model should be fit in a way that makes the model neither too specific for the data (*overfitting*) nor too general (*underfitting*) (Alpaydin 2016, 39–41). If, for example, a model is built with the goal of differentiating between historical and sentimental nineteenth-century Spanish-American novels, but the training set is dominated by a certain type of historical novel, then it would be disadvantageous if the model learns too many details about the special types of historical novels when compared to sentimental novels. Such special types of historical novels are, for example, novels that were part of large series of historical novels published by a few Mexican authors, such as the "Episodios nacionales mexicanos" (1902–1903, MX) by Victoriano Salado Álvarez, or the "Leyendas históricas de la independencia" (1886–1913, MX) by Ireneo Paz. If other types of historical novels are presented to the model as a test case, it could happen that the specialized model is not able to classify them correctly. This could happen, for instance, if an individual Argentine or Cuban novel dealing with the conquest or colonial times is presented to a model trained with Mexican historical novels. Such a model would be overfitting. Conversely, if a model does not predict the classes of the data it was trained on well, it has probably learned too little about the data structures representing the classes and is underfitting. Finally, a fully trained and validated model can be used with a third part of the data, the *test set*, which has not been part of the training cycle at all.

Obviously, not just the type of model and the selection of the best parameters for it are decisive for the quality of a classification task, but also how – based on what data – the instances of the classes are represented in the machine learning process. The selection of the text features for genre classification should be based on good hypotheses about their relevance for the problem. Here, it was decided to use the most frequent words and topics as two comparatively generic types of features to classify the nineteenth-century Spanish-American novels by subgenre. Both types of features have already been used successfully in the classification of literary genres and other text types. Hettinger et al. (2016), for instance, used the 3,000 most frequent words, the 1,000 most frequent character 4-grams, and topic models to classify a corpus of German nineteenth-century novels by subgenre. They achieved accuracy scores between 70 and more than 90 %, depending on the feature set used and on the constellation of subgenres. Schöch (2017c) classified French dramas of the Classical Age and Enlightenment by subgenre, using a set of topic models with varying parameters. In Schöch's study, the accuracy reached between 70 and 87 % for different numbers of topics, optimization intervals, and classifiers. Schöch also

[562] For a formal definition of the generalization problem, see Alpaydin (2016, 23–27, 37–41).

tested to classify the dramatic subgenres based on the most frequent words. He found the best results with 3,500 MFW and a z-score transformation of the word frequencies.

Although the most frequent words and topics are quite different types of features, both cover much of the underlying textual material. The hypotheses that the distributions of most frequent words and topics in the texts can be related to subgenre distinctions are fairly general.[563] By testing different sets of the two types of features, the results of the classification can help to refine the hypotheses about the features' relevance to capture the differences between the subgenres.[564]

Regarding the type of model used for the classification, it was decided to compare three different kinds of classifiers: k-Nearest Neighbours (KNN), linear Support Vector Machine (linear SVM), and Random Forest (RF). By comparing the results of different classifiers, it can be assessed if the different feature sets work well or not with all of them or if this depends on the kind of classifier. The importance that single features have in the classification process can also be checked by comparing their relevance in different models. In general, many different algorithms for supervised machine learning with many variants exist.[565] A type of classifier that has repeatedly shown very good results for the classification of literary texts is the linear SVM (Bei 2008; Hettinger et al. 2015, 2016; Schöch 2017c). This algorithm has been described as giving good results for high-dimensional data sets and also for data sets that are sparse (Müller and Guido 2016, 69). These are both typical characteristics of features that are extracted from texts, which might explain why SVMs work so well for literary text classification. The KNN algorithm is comparatively simple and can be considered a baseline option against which the results of the other classifiers can be checked. It is not expected to work well with high-dimensional and sparse datasets, though. RF classifiers do not tend to work very well with such data either (Müller and Guido 2016, 37, 46, 90). Nevertheless, they are widely used and approach the data in an entirely different way than SVMs, so that it is worth testing them as an alternative algorithm. Furthermore, the dimensionality and sparseness of the different feature sets prepared for the corpus varies, as was shown in the chapters 4.2.1.1 and 4.2.1.2 above. A set with a lower number of topics or less MFW might also work well with KNN or RF.

The three chosen algorithms depend on different parameters that need to be set before the models are trained with data. Because much variance is already introduced here by selecting different feature sets and different types and constellations of subgenres that are to be analyzed, it was decided not to vary the model parameters systematically for all settings but to conduct preliminary tests with selected feature sets and subgenres. The parameters that turn out to be good choices in this preparatory step are chosen and fixed for the subsequent experiments. In

---

[563] See chapter 4.2.1 on the chosen feature sets for details.

[564] It would also be possible to use more specific types of features, for example, temporal expressions (including dates or words related to duration or repetitive events). By using such special features, the textual material is reduced much more, making it more difficult for classifiers to optimize a model for the differences between subgenres. As a consequence, the hypotheses about the relevance of these features for genre classification would have to be much stronger, for example, based on assumptions about how temporal expressions are used in historical novels versus other subgenres, and the classification could serve to confirm or reject the hypotheses.

[565] For an overview of seven important and popular families of classification algorithms, see Müller and Guido (2016, 31–121). In the Python library scikit-learn, there are 17 groups of classifiers (Scikit-learn developers 2007–2023l).

Classifier	Parameter	Parameter values
KNN	n_neighbors	3, 5, 7
KNN	weights	uniform, distance
KNN	metric	Manhattan, Euclidean
SVM	C	1, 10, 100, 1000
RF	max_features	sqrt(n_features), log2(n_features)

Table 34. Parameters for classifiers.

the following table 34, the parameter values that were tested are given for the three classifiers in question.[566]

KNN classifies new data by looking for the nearest data points for which the class is known (the *neighbors*). The class to which most of the nearest neighbors belong is assigned to the new data. For this algorithm, three parameters are varied: the number of neighbors taken into consideration for the classification decision, the method applied to weight neighbors, and the distance metric used to calculate how far the neighbors are away from the data point in question. A uniform weight means that all neighbors have the same influence on the decision, whereas a distance-based weight means that neighbors that are nearer are weighted higher than neighbors that are further away (Müller and Guido 2016, 37–38; Scikit-learn developers 2007–2023j). The difference between the Manhattan and the Euclidean distance is that the former sums up the differences between every feature for the two texts that are compared, and the latter uses the direct distance of the two feature vectors (Evert et al. 2017, ii7).[567]

As a linear SVM belongs to the class of linear models, it uses a linear function of the input data to make predictions about new data. In such a function, a weight (coefficient) is determined for each feature, and the prediction can be understood as "a weighted sum of the input features" (Müller and Guido 2016, 47).[568] If it is smaller than zero, the negative class is predicted; otherwise, the positive class is. That way, the linear function that is learned defines a decision boundary for the classification. The most important parameter to vary in an SVM is the C parameter, which regularizes the learning process. The higher the value of C is, the more the model tries to learn a function that fits the training data in the best possible way, and a lower C means that the model tries to find low coefficients (Müller and Guido 2016, 58–60). The C parameter is thus directly connected to the question of over- and underfitting. As SVMs are sensitive to different feature scales, for this classifier, all the feature sets except the ones based on z-scores were further processed by rescaling them to a range of 0 to 1.[569]

---

[566] For the three algorithms, the Python implementations in scikit-learn were used (Scikit-learn developers 2007–2023k, 2007–2023j, 2007–2023e).

[567] In the cited paper, it is discussed in-depth which effect the choice of different distance measures has on the results of authorship attribution tasks. Such a systematic investigation of the role that different distance measures have on genre classification has not been conducted yet.

[568] Besides the feature weights, for the linear function, also an intercept or y-axis offset is learned.

[569] This means that the minimum value for every feature is 0, and the maximum value is 1. The same kind of preprocessing of the data was used by Hettinger et al. (2016). For the feature scaling, the function "scale_feature_sets()" in the script "classification.py" was used. See https://github.com/cligs/scripts-nh/blob/master/analysis/classification.py.

General features			Topic features	
MFW	Token units	Normalization	Topics	Optimization interval
100	word	tf-idf	50	100
1000	word 3-grams	z-scores	100	1000
5000	character 3-grams			
Subgenre constellations				
novela histórica versus others				
novela romántica versus others				
novela sentimental versus *novela de costumbres*				
novela realista versus *novela naturalista*				

Table 35. Experiments for parameter evaluation.

Random Forests use an ensemble of decision trees to make predictions. Decision trees learn a set of rules involving if-else questions. The rules are processed hierarchically until they lead to final decisions (Müller and Guido 2016, 72). The advantage of random forests over simple decision trees is that they are less likely to overfit because each tree in the collection differs from the others, and the effects of overfitting can be reduced by averaging the results. To achieve that the trees are different, randomness is introduced into the learning process by selecting varying data and features for each decision tree.[570] Each node in the decision process uses a subset of the features (Müller and Guido 2016, 85–86). The `max_features` parameter controls how many of all the features are made available to each decision node in the trees.

Table 35 lists the selected feature sets and subgenre constellations that were tested in the preliminary study with different classifier parameters, to be able to decide on which ones to use for the subsequent classification tasks. The various combinations result in 88 different settings.

An important point when selecting the training and test data for the classification is the size of the different classes, that is, the number of novels in each class. Because the number of novels for each subgenre differs in the corpus, undersampling is used here as a strategy to balance the classes. This means that the number of instances of each class is set to the size of the smallest class. There are, for example, 116 romantic novels and 85 non-romantic novels in the corpus. With undersampling, the subgenre constellation "*novela romántica* versus other" is performed with a set of 85 romantic novels, which are selected randomly from all the romantic novels, and 85 non-romantic novels. The undersampling process is repeated ten times for each setting to make sure that the random selection does not have too much influence on the results. For the topic models, all five models that were produced for each combination of topic modeling parameters are used, and the resulting scores are averaged.

The scaled feature sets are stored as new files in the same location as the original ones, and they have the additional term "MinMax" at the end of their filename. See https://github.com/cligs/data-nh/tree/master/analysis/features/mfw for the general features and https://github.com/cligs/data-nh/tree/master/analysis/features/topics for the topic features.

[570] The results of Random Forests can vary depending on the different random states that are used. To make the classification results reproducible, here, the `random_state` parameter of the RF classifier is always set to 0.

In the parameter study, grid search classifications are performed for each type of classifier, with the different classifier parameters and for each feature and data combination. A grid search consists in creating a grid of parameter values, for example, the different values for the number of neighbors and metrics for KNN, and running a set of classifications to find out which combinations of the parameters work best. A method that facilitates this procedure in the context of Machine Learning is scikit-learn's GridSearchCV, which was used here (Scikit-learn developers 2007–2023i). The method includes the possibility of performing cross-validation. Here, 10-fold cross-validation was used, meaning that each combination of parameters is tested in ten different splits of training and test data.[571] Part of the results that Scikit-learn's GridSearchCV function returns are the ranks of the different parameter values resulting from average test scores over all folds. Here the results of the parameter study were evaluated by counting how often each parameter value had the first rank, that is, how often it led to the best mean accuracy test score.[572] That way, the test scores themselves are not considered, so rather than finding the parameter values which lead to the highest scores for specific data and feature combinations, it was analyzed which values yield the best score most often for all the different settings. The goal was to find parameter values that are generally a good choice so that they could be used for all the following analyses. However, as the classifications were performed separately with the two main feature types (general, MFW-based features versus topics),[573] different parameter values were chosen for them whenever the results of the parameter study suggested that. The main reason for keeping the main types of feature sets separate is to allow for their interpretation in terms of different levels of text style.

The results of the parameter study for the KNN classifier are summarized in figures 45 to 47 for the three parameters `n_neighbors`, `weight`, and `metric`. Concerning the number of neighbors that are decisive for the classification, all three candidates (3, 5, and 7) reached the first rank almost equally often. A number of 7 neighbors is a bit better than fewer ones (for MFW and for topic features), so this number was used in all the following experiments. Why could a higher number of neighbors be advantageous? One hypothesis is that the novels' individual feature distributions must be checked against several neighboring ones to be able to decide on the subgenre, which means that resemblances to several works are more relevant than the similarity to a few works that represent the subgenre in a homogeneous way.

Regarding the method used for weighting the influence of the neighbors on the classification decision, distance-based weights work better than uniform weights for both feature types, only

---

[571] The main methods used to perform the parameter study are the functions `parameter_study()` and `evaluate_parameter_study()`, which are defined in the Python script classification.py available at https://github.com/cligs/scripts-nh/blob/master/analysis/classification.py. Accessed December 16, 2020. The results of all the grid searches for the three classifiers were stored together in the files grid-searches-KNN.csv, grid-searches-SVM.csv, and grid-searches-RF.csv, which can be viewed at https://github.com/cligs/data-nh/tree/main/analysis/classification/parameter_study. Accessed December 16, 2020.

[572] The mean is calculated from the test scores of the ten splits from cross-validation.

[573] In principle, it would also be possible to combine the two main types of feature sets. This was done, for example, by Hettinger et al. (2015). In their approach to classifying German novels by subgenre, they experimented with combining feature sets to see if it improved their results. They used Principal Component Analysis (PCA) to make the sizes of the feature sets comparable and found out that some combinations improved the results, whereas others led to lower accuracy scores.

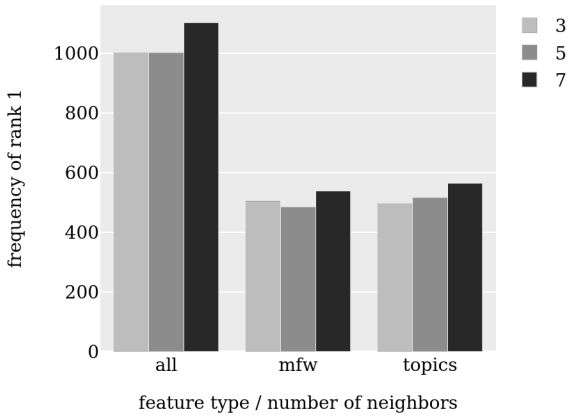


Figure 45. Frequency of rank 1 for different values of `n_neighbors` (KNN).

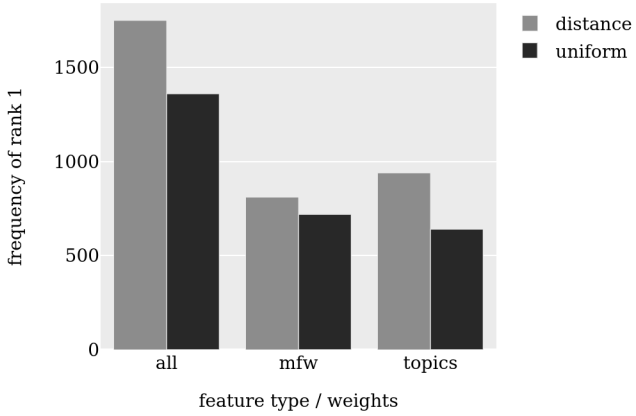


Figure 46. Frequency of rank 1 for different values of `weights` (KNN).

slightly in the case of MFW and a bit more for topics, as figure 46 shows. Even in this case, the difference between both types of weighting is not striking. A hypothesis to explain the tendency towards distance-based weighting is that the novels rather form classes on relationships of similarity between individual works than on uniform similarities to several other works.

The difference in performance between the parameter values is clearest for the metric, as shown in figure 47. The Manhattan distance works best in more than half of the cases. Both for MFW and even more for topics, the Manhattan distance reaches the top mean accuracy test score more often than other metrics. This shows that measuring the distance between the feature vectors as they are in the different dimensions (with tf-idf or z-scores) works better than measuring the distance between the vectors in one step. So distances between individual features play an important role in the classification by subgenre. To summarize, for KNN, the parameters are set to 7 neighbors, distance-based weights, and Manhattan metric.

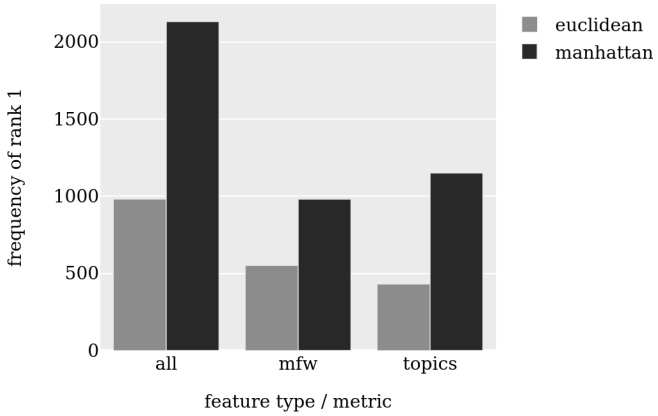


Figure 47. Frequency of rank 1 for different values of `metric` (KNN).

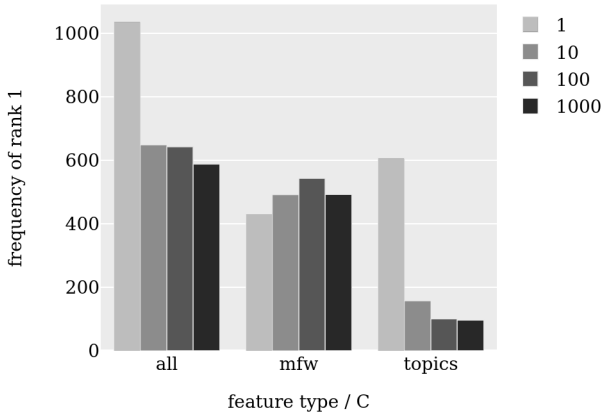


Figure 48. Frequency of rank 1 for different values of `C` (SVM).

For the SVM classifier, only the C parameter was tested. The results in figure 48 show that the value 1 works best for topics. For MFW, a value of 100 leads to the best mean score more often than other values, but the differences are small. As a lower value of C means that the coefficients for the linear model are preferably lower so that the model is less specialized on the data it is trained on, the models that are built with it are more general, which, in principle, is good. A hypothesis to explain that a lower value of C is much better with topic features and not so decisive with MFW is that topics are primarily content-related. When the topics are too specific, they might not be characteristic of the subgenres of the novel but rather of individual texts. Primarily stylistic features, on the other hand, are more flexible when a classifier uses them to build models for subgenres. Because of these results of the parameter study, in the following analyses, the C parameter is set to the value 1 for the topic features and to the value 100 for MFW-based features.

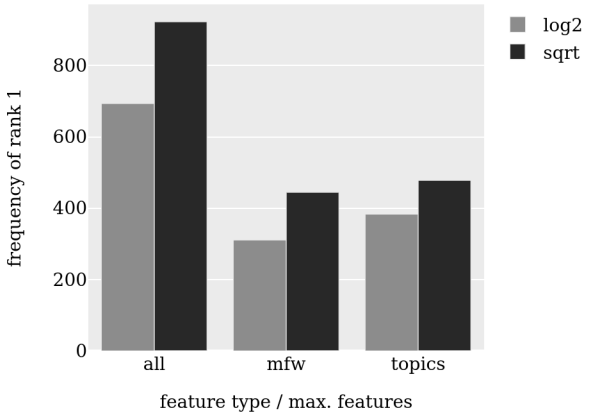


Figure 49. Frequency of rank 1 for different values of `max_features` (RF).

For the third classifier, RF, it was tested whether the maximum number of features used in each tree is better determined by taking the logarithm to the basis 2 or the square root of the overall feature number. The results in figure 49 indicate that the square root is more successful for both MFW and topic features, so this parameter value is used for RF in the following. With the square root, the number of features chosen is higher than with the logarithm. In addition, it increases faster with an increasing number of features, so using more of the features that are available is especially useful when decision trees are built based on MFW, but also when they are built based on topics.

A conclusion to be drawn from the different results of the parameter study for the three classifiers and their respective parameters is that the choice of the parameter values does not make all the difference for the classification results. All of the parameter values reached rank 1 for several data and feature sets, but some are more successful in the majority of cases. By fixing the values for subsequent analyses, a common basis is found for comparing classification results for different subgenres and feature sets, even if this means that for some constellations, the results could still be improved by adjusting the parameters individually.

Before the classification results are presented in the following subchapters on thematic subgenres (4.2.2.1.1) and literary currents (4.2.2.1.2), the general classification workflow is outlined here. In a first step, the discursive level on which subgenres are analyzed is chosen (i.e., *theme* or *current*). Then the individual subgenres which are contrasted on each level are selected, for instance, *novela histórica* versus *novela sentimental* or *novela histórica* versus all other novels. To keep the evaluation of the classification results simple and comparable, only two classes are used each time. For these, the data is selected so that the classes have the same size. As said before, the data selection process is done randomly and repeated ten times to make sure that the results do not depend on the specific selection. If the two classes always have the same size, the classification baseline can be set at 50 % for all the constellations. Next, the classification is performed with the three types of classifiers and all the different feature sets that were prepared. The classifier parameters are fixed with the values that were chosen in the preliminary parameter

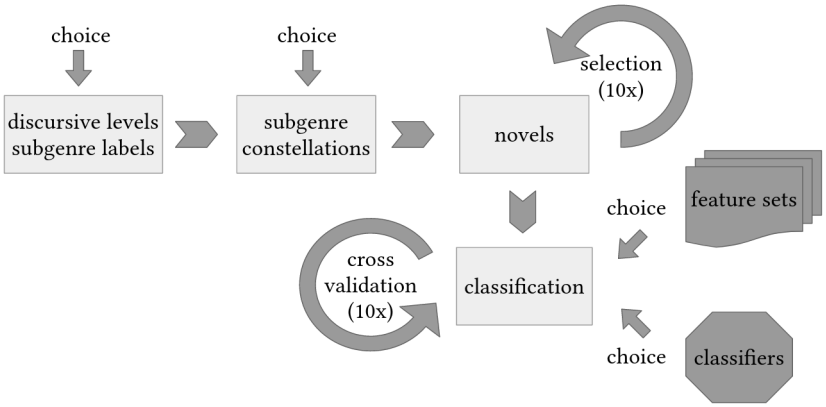


Figure 50. Classification workflow.

study, and the classification is performed with 10-fold cross-validation. A graphic overview of the classification workflow is given in figure 50.[574]

The results are evaluated to find out which sets of MFW-based and topic features work best with which classifier. The best constellation of classifiers and features is chosen for each discursive level. Following the assumption that different kinds of features and hence different levels of style can be decisive for the various discursive levels of subgenres, independent decisions are made for thematic subgenres and literary currents. However, for the different subgenre constellations on the same discursive level, no individual choice is made. Here the best classifier, MFW-based, and topic feature sets found for the discursive level are chosen to further evaluate the results of the subgenre classification. For example, the results for the *novela romántica* can be directly compared to the results for the *novela realista*. Beyond inspecting accuracy values and F1 scores, further interpretation of the results is concerned with two aspects:

1. the importance of individual features in the feature set for the classification, for instance, which word n-grams or which topics are decisive to differentiate between two types of novels, and
2. the number of times that novels are classified correctly or misclassified.

The first aspect leads to insights about the text types of the subgenres: Which features are characteristic of the subgenres when they are contrasted with others? The second aspect aims to find out which works are typical for specific textual subgenres. Which works are classified

[574] The implementation of the classification tasks is included in the script classification.py, which is available at https://github.com/cligs/scripts-nh/blob/master/analysis/classification.py. Accessed January 9, 2021. The repetition of the data selection process with undersampling for the bigger class is also implemented in that script. For the cross-validation, scikit-learn's function `cross_validate()` was used. With that function, all the estimators that are trained for the different cross-validation runs can be returned. This was done, and the results of each cross-validation run were stored in CSV tables. The mean accuracy values and standard deviations, which are discussed in the result section, are calculated with an own python script based on the collected cv-runs in the CSV tables. It was necessary to get the results for every estimator to be able to analyze the feature importances and predictions of each one (Scikit-learn developers 2007–2023h).

correctly most often (the true positives)? Which ones are frequently misclassified and are rather untypical or mixed instances of a subgenre? If a novel has the label of a specific class but is always or almost in every case misclassified (a false negative), it is considered a member of the conventional genre (because it has the corresponding genre label) but not of the literary text type (because it is not recognized as being a member of it). If, on the other hand, a novel is always or almost in every case seen as an instance of the class, but it does not have the label (a false positive), then it is considered a member of the text type (because it is textually similar to the other novels that are part of this type) but not of the conventional literary genre (because it does not have the corresponding label). Finally, novels that do not have the label of the class in question and are never or almost never classified as instances of it (the true negatives) are neither part of the conventional genre nor the text type.[575] That way, the classification results are interpreted in terms of the distinction between conventional literary genre, literary text type, and textual genre to find out the degree to which the conventional genre and a text type that can be associated with it overlap. This gives insight into the textual coherence of the genre as a whole, but it also allows us to find prototypical and untypical members of a genre and to locate individual novels on the levels of convention and text in relationship to the genre they participate in (or not).

The feature importances and classification results vary for each run with the ten different data selections in the undersampling process and the ten repetitions in the cross-validation procedure. Therefore, all the values are collected, and the average results are analyzed. For the MFW, the feature importances and classification results for the individual novels can be averaged on the level of the different feature sets (for example, for all the data selections and cross-validations performed with 100 MFW and tf-idf normalization) because the kinds of features stay the same: which ones are the 100 MFW does not change. For the topics, in contrast, it is not possible to summarize the feature importances and novel classifications on such a general level because the topic features are different in each topic model that is produced. Even if the number of topics, the iterations, and optimization intervals are fixed, the topics themselves are not consistent throughout the five topic modeling repetitions because of the probabilistic procedure. So in the case of the topics, one specific topic model must be chosen. Only then the features and the individual classifications of the novels can be evaluated and their importance can be averaged. Here, the first of the five topic models that are produced for each topic model parameter constellation is chosen as a representative.[576]

---

[575] Another, more sophisticated way to analyze this would be to check the probabilities that the classifiers calculate for an instance to belong to the different classes. The implementations of KNN, SVM, and RF in scikit-learn include the possibility of inspecting the probability calculations, but for the SVM it is noted that these values may be inconsistent with the general prediction (Scikit-learn developers 2007–2023m).

[576] Analyzing feature importances for the three types of classifiers involves different concepts and techniques. For SVMs, the coefficients of the linear model can be interpreted (Scikit-learn developers 2007–2023k). For RFs, the trained estimators in scikit-learn include a list of feature importances which is based on the average contribution of each feature to reduce misclassifications in all the different trees of the forest (Scikit-learn developers 2007–2023e). Finally, for KNN, no feature importances are returned because the classification is made based on similarities between whole data points, which means that all the features of the neighboring points are decisive. So to interpret how important individual features were in this case, other external methods have to be used. One possibility is to use the feature set which worked best with KNN and to calculate how distinctive different features are for the

In this setup, the models that the classifiers build based on specific feature sets are interpreted as sets of literary text types. They are considered sets of text types and not individual text types because the classifiers learn to differentiate between several classes. Because here only two classes are compared each time, two text types are learned, or rather, it is learned how two text types can be delimited and distinguished, one for the positive and the other for the negative class. That they depend on a specific set of textual features means that the text types are constituted on a certain stylistic level (e.g., the 1,000 MFW and all the linguistic material that is covered by them or 50 topics and all the thematic distinctions that can be made based on them). Furthermore, specific stylistic cues are determined for the text types, for instance, specific topics that have great importance as features for the classification. When the results are evaluated, several stylistic cues can be interpreted as forming stylistic traits of the literary text types in question. For instance, if several adjectives and nouns referring to opposites (e.g., good-bad, city-countryside) turn out to be significant features in an MFW-based model, these stylistic cues can be subsumed as a trait of opposites or antagonisms. When a range of different feature sets is used in several classifications to analyze the same subgenre, such traits can be collected and interpreted as different facets of a text type.

The text types are not only represented by the classification models but also by the set of texts that participate in them. However, these texts are not equivalents of the conventional genre. It is not that the texts carrying specific labels are contrasted directly to view their features. Instead, the feature importances that result from trying to classify the novels by their subgenre label are contrasted. In the latter case, the text type consists of the texts that are repeatedly found as true positives plus the ones that are frequently determined as false positives. On the other side, the texts that carry the conventional label but do not fit textually are not included when the importance of the different features is evaluated. In the next chapter, the classification results for the thematic subgenres are presented. As the general classification workflow was already presented in this chapter, the details about the steps taken to perform the classifications are not repeated in the result sections.

4.2.2.1.1 Thematic Subgenres

This chapter presents the results of the classification of novels by thematic subgenres. First, the distribution of the primary thematic subgenres in the corpus is presented in figure 51.

The four most frequent types of primary thematic subgenres are the *novela histórica*, with 67 novels; the *novela sentimental*, with 55 novels; the *novela de costumbres*, with 50 novels; and the *novela social*, with 45 novels. Smaller groups are the political novels (13), the science fiction and crime novels (5 each), and the anti-slavery novels (4). The latter can also be subsumed under social novels but are marked separately because historically and by literary critics, they were categorized with this more specific label. Twelve novels participate in other primary thematic

---

classes in question. For example, this can be done by evaluating which features differ most from the mean values of the features in both the positive and negative class and ranking these features by their distinctiveness. If the full-texts are used as a basis (instead of the specific feature set), zeta-scores can be calculated for the features to find the ones that are distinctive for the classes. For an overview of the zeta-measure of distinctiveness and its variants, see Schöch, Schlör et al. (2018).

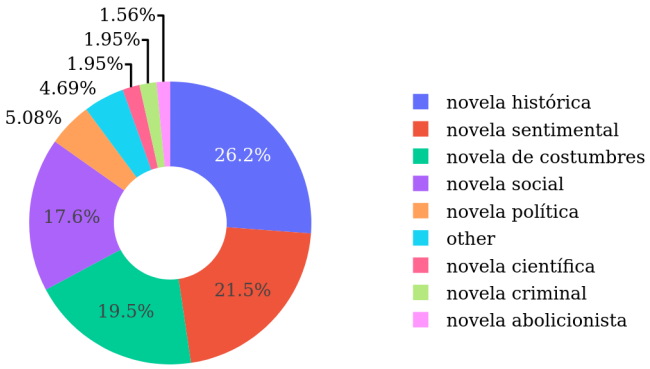


Figure 51. Primary thematic subgenres in the corpus.

subgenres, which all occur only one to three times. The classification of primary thematic subgenres is conducted for the following selected subgenre constellations:

- *novela histórica* versus other novels
- *novela sentimental* versus other novels
- *novela de costumbres* versus other novels
- *novela histórica* versus *novela sentimental*
- *novela histórica* versus *novela de costumbres*
- *novela sentimental* versus *novela de costumbres*

With these six constellations, the three most frequent primary thematic subgenres are analyzed. They correspond to the subgenres that developed in the romantic period and to the ones that literary historians have often distinguished as relevant thematic subgenres in nineteenth-century Spanish America.[577] The presentation of the results starts with the topic features and then goes on to the MFW-based ones. The results for the topic features are discussed in more detail, inspecting feature importances, numbers of correct and wrong classifications, and topic profiles of selected novels. For the MFW, the overall classification results are presented to see how well these features work, but a deeper analysis of the features and novels is left as a future task.

First, it is analyzed which classifiers worked best to classify the thematic subgenres in all six constellations.[578] For this, the accuracy and F1 score values for all classification runs were evaluated, and top values, mean values, and the standard deviation (SD) were calculated. In total, 144,000 classification runs were considered for the topic features, including all topic model parameter constellations, topic modeling repetitions, repeated data selections, and 10-fold cross-validation.[579] The results for the topic features are shown in table 36.

---

[577] See the presentation of thematic subgenres from a literary-historical point of view in chapter 2.3.1.

[578] The classification was performed with the script https://github.com/cligs/scripts-nh/blob/master/analysis/classification.py. All the results are available at https://github.com/cligs/data-nh/tree/main/analysis/classification/themes. Accessed January 3, 2021. Tables with the results of all the classification runs can be found in the subfolder "results_data", summaries in tabular form in "results_summaries", and visualizations of results in "visuals".

[579] For some classification runs, no F1 scores were available. This happens if the test set only contains instances of one class (either the positive or the negative one) because then it is not possible to calculate precision and recall

Classifier	Feature type	Top accuracy	Mean accuracy	SD accuracy	Top F1	Mean F1	SD F1
KNN	topics	1.0	0.77	0.14	1.0	0.78	0.14
SVM	**topics**	**1.0**	**0.80**	**0.13**	**1.0**	**0.80**	**0.14**
RF	**topics**	**1.0**	**0.80**	**0.14**	**1.0**	**0.80**	**0.15**

Table 36. Classification results for primary thematic subgenres (topics).

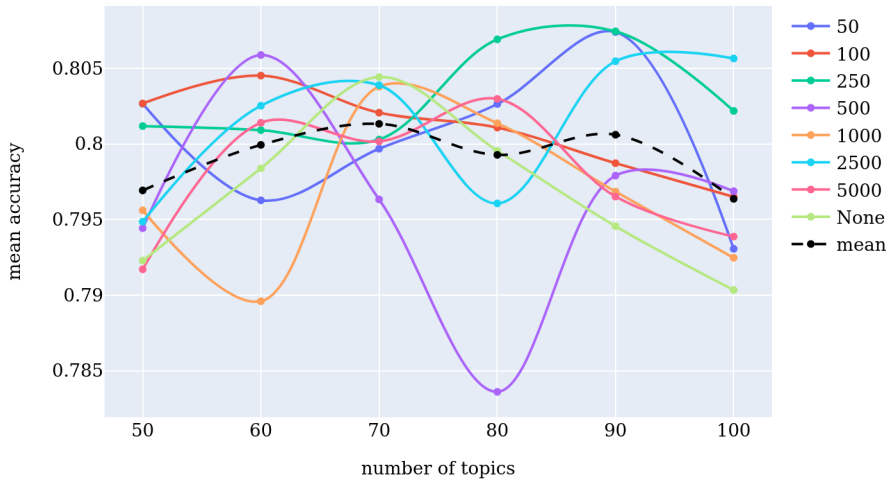


Figure 52. Classification results for topic feature sets (SVM, varying number of topics, and optimization intervals).

In the case of the topics, SVM and RF work best with a mean accuracy of 0.80, but the results for KNN are also quite good, with a mean accuracy of 0.77. The relatively high standard deviation indicates that either the results vary for the different topic model parameters or for the different subgenre constellations, or both. For the following, more detailed analyses of the results, SVM is chosen because the standard deviation is lower than for RF, even if there is only a difference of one percentage point. Once the best classifier is chosen, the next step consists in evaluating the feature sets in more detail. The goal is to find out which influence the different parameter values have on the results and to find the ones that are best suited for the classification of primary thematic subgenres. The classification results for the different topic modeling parameters are presented in figure 52.

The top mean accuracy of 0.81 is reached with several different combinations of topic numbers and optimization intervals: with 60 topics and an interval of 500, 80 topics and an interval of 250, and 90 topics and an interval of 50, 250, or 2,500.[580] If one evaluates the general development of

values. The missing F1 scores were just left out for the calculations of the means and standard deviations. For the calculation of accuracy values, in turn, this does not constitute a problem. The values in the table are rounded to two decimal places.

[580] Accuracies that are mentioned in the text are rounded to two decimal places. Because the range between the lowest and highest mean accuracies is so small, different third or fourth decimal places lead to differences in the plots.

the results for different topic numbers and optimization intervals, for some intervals, there are trends, and for others, oscillations. With an interval of 100, for example, the score is highest at 60 topics and decreases for higher topic numbers. The score for no optimization at all rises up to 70 topics and then falls again. For intervals of 50 and 250, the scores get higher with more topics. Except for the interval of 2,500, which means very little optimization, all the scores drop with more than 90 topics. This suggests that the models get too specific to detect subgenres if they have more topics. Otherwise, it appears that the number of topics interacts with the optimization interval so that specific combinations which produce topics and topic distributions that are not too general and not too specific work well to model thematic subgenres.[581]

Overall, the differences between the mean accuracies using different kinds of topic features are very low, ranging from 0.78 for 80 topics and an interval of 500 to the highest value of 0.81. This means that the higher standard deviation observed in the general accuracy mean for topic features is not due to the different topic modeling parameters but must be connected to the different kinds of thematic subgenres or to the influence of individual novels. With such similar results, the decision to choose a certain combination of topic numbers and optimization intervals does not only need to be based on the top mean accuracies but can also take into account the kind of topics that are likely to result from a model with a specific combination of parameters. Here, 90 topics are chosen because they lead to the highest mean score several times and also because the topics are more specific than they would be in a topic model with a lower number of topics so that the resulting text types can be described in more detail. As for the optimization interval, a number of 250 is chosen so that the topics are still relatively balanced regarding the different weights they have in the corpus and individual texts.

Having decided on the kind of classifier to use and on the specific kinds of features, the classification results for the different constellations of thematic subgenres can now be inspected. The results for the topic models with 90 topics and an optimization interval of 250, using SVM as a classifier, are shown in table 37. These average results are based on 500 classification runs for different data samples and the 10-fold cross-validation.

The constellation *novela histórica* versus *novela sentimental* has the highest mean accuracy with 0.89. The second best result is reached for *novela histórica* versus *novela de costumbres*, and the third for *novela histórica* versus other. This shows clearly that the historical novel is the subgenre that can best be distinguished from the other two thematic subgenres that were chosen here but also from the big group of all other kinds of novels, which include not only sentimental novels and novels of customs, but also crime novels, science fiction novels, social and political novels, and more. This result confirms the expectation that the *novela histórica* as a subgenre, which is by convention firmly established in Argentina, Cuba, and Mexico in the nineteenth century, also is united textually. Although there are different subtypes of historical novels dealing with different past (or not-so-past) epochs and different specific topics, the subgenre can be classified well by using topic features. Furthermore, it is astonishing that it can be separated best from the sentimental novel because there are many historical novels with a sentimental plot and also

[581] See also Schöch, who analyzes the effects that different optimization intervals have on the resulting topic models of collections of literary texts and who notes: "if your goal is to identify topics typical of certain authors, periods, genres or some other reasonably large subset of your collection, then it may be better to optimize a bit less. In any case, it seems to me that it is quite possible to do too much or too little optimization for a given task" (Schöch 2016).

Subgenre 1	Subgenre 2	Top accuracy	Mean accuracy	SD accuracy	Top F1	Mean F1	SD F1
novela histórica	other	1	0.82	0.10	1	0.82	0.09
novela sentimental	other	1	0.81	0.12	1	0.80	0.13
novela de costumbres	other	1	0.71	0.14	1	0.72	0.16
novela histórica	*novela sentimental*	1	0.89	0.11	1	0.89	0.10
novela histórica	*novela de costumbres*	1	0.84	0.10	1	0.84	0.10
novela sentimental	*novela de costumbres*	1	0.77	0.13	1	0.75	0.16

Table 37. Classification results for primary thematic subgenres (SVM, 90 topics, optimization interval of 250).

because most historical and sentimental novels in the corpus are romantic novels that have much in common. This result confirms what most literary histories of nineteenth-century Spanish-American novels describe (see chapter 3.1.3 on the *novela sentimental*): that the sentimental novel is an important and recognizable thematic subgenre as well. After the historical novel, the sentimental novel has the second-best results. The mean accuracy for *novela sentimental* versus other novels is at 0.81, which is only slightly worse than for the historical versus the other novels. Apparently, the sentimental novel can be separated well from other subgenres, even if there are many novels that also have a sentimental plot. However, the results also show that this is not so easy if sentimental novels are compared to novels of customs. Here, the mean accuracy is at 0.77, which is the worst result for the direct comparisons of subgenres. This confirms what has has been remarked by Janik (2008, 67–68): that the description of local customs in novels and the expansion of the *cuadros de costumbres* to longer fictional narrative works in the form of novels of customs needed the sentimental plot as a basic structure, at least in the romantic variant of the subgenre. The results are even worse when the novels of customs are contrasted with all other novels (with an accuracy of 0.71), which can be interpreted as a sign that *costumbrista* elements are not only found in sentimental novels but also in other subgenres. As the group of other novels also contains all the social and political novels that are associated with the realist (or naturalistic) current, it might well be the case that this causes lower accuracy. The two results for the novels of customs can be interpreted as a sign of mixtures of sentimental novels and novels of customs, but also similarities of the novels of customs with other social and political novels in terms of the topics that they treat. Thus, for example, Kohut has noted that there are not only romantic, but also realistic and naturalistic novelas de costumbres (Kohut 2016, 196). However, mean accuracies of over 0.70 do mean that even for the novels of customs, a model can be learned that classifies the texts better than by chance, only that this thematic subgenre is textually less coherent than the others. Overall, the standard deviations for the accuracies (and

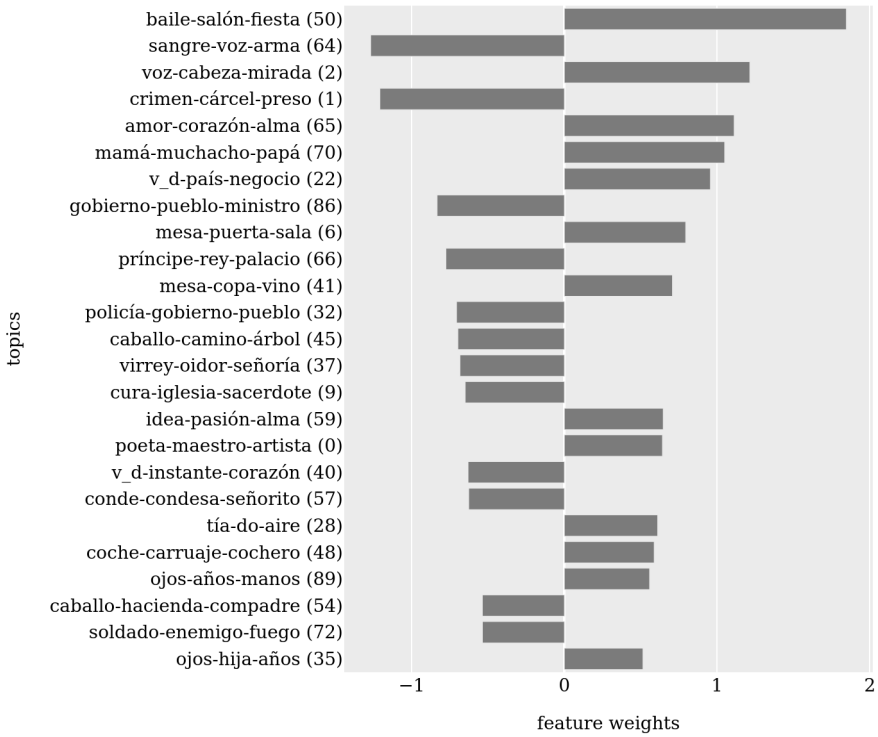


Figure 53. Feature weights (topics) for historical versus sentimental novels.

F1-scores) are still quite high, which means that the results also depend on the individual novels that are selected for the training and test sets.

The next step consists in having a look at the coefficients that have the highest values in each subgenre constellation to inspect the characteristics of the text types that were learned by the classifiers. Starting with the historical novels, figure 53 shows the top 25 coefficients for the classification of *novela histórica* versus *novela sentimental*. As these weights are based on a specific topic model, they represent the average weights of 100 classification runs (ten different selections of novels and ten cross-validation runs, as explained in chapter 4.2.2.1 above on the classification workflow). In the plot, the topics are identified by the number they got in the topic modeling process (0 to 89 for the 90 topics) and the three words that are most important in each topic.[582]

[582] Word clouds, which visualize the 40 top words of each topic in the model, are available at https://github.com/cligs/data-nh/tree/main/analysis/features/topics/5_visuals/90tp-5000it-250in-0/wordles. The file containing all the top words for the topics can be found at https://github.com/cligs/data-nh/blob/main/analysis/features/topics/3_models/topics-with-words_90tp-5000it-250in-0.csv and the topic probabilities per novel at https://github.com/cligs/data-nh/blob/main/analysis/features/topics/4_aggregates/90tp-5000it-250in-0/avgtopicscores_by-idno.csv. Accessed January 6, 2021.

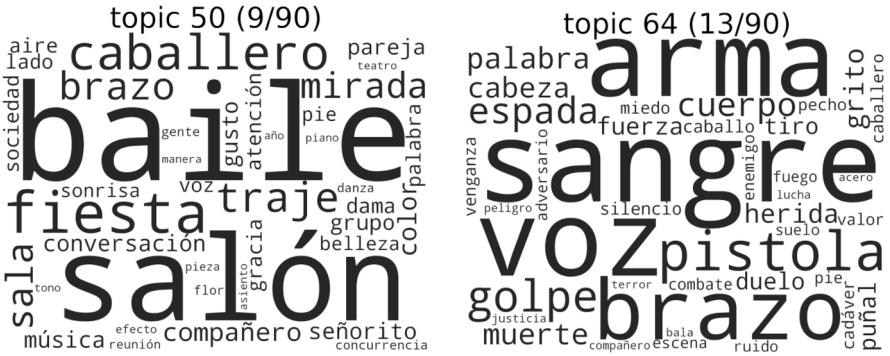


Figure 54. Most distinctive topics for historical versus sentimental novels.

The bars with negative values are the topics characteristic of historical novels and those with positive values of sentimental novels.[583] The most important topic for distinguishing historical from sentimental novels is “baile-salón-fiesta”, which is a topic concerned with a ball situation that is typical for sentimental novels. The second most important distinguishing topic is “sangre-voz-arma”, this time for historical novels. The topic covers a situation in a fight or battle. The word clouds of the two topics are given in figure 54. It is interesting that both top topics describe situations that are typical for the two subgenres in question and that can be considered characteristic scenes and elements of the plot. As the topics consist only of nouns and verbs are not included, the dynamic aspect of the topics does not stand out directly, but even the nouns describe actions if they are analyzed together. In the case of the ball situation, for example, the words “mirada”, “brazo”, “sonrisa”, “conversación”, and “palabra” point to acts of the characters. In the fight situation, also the word “brazo” occurs, as well as “cabeza”, “golpe”, “grito”, “espada”, “pistola”, “puñal”, “combate”, “lucha”, “herida”, and “muerte”, so in both cases parts of the body and physical actions are involved. The ball topic contains also words that describe the setting of the situation, for example, “salón”, “sala”, “pieza”, “teatro”, “flor”, “música”, “piano”, “gente”, “reunión”, “concurrencia”, “sociedad”. That the word “teatro” is part of the ball topic could be a sign that similar thematic elements are relevant if the situation is a theater performance as a social event. The fight topic is less descriptive and contains more words related to emotions and sounds: “miedo”, “peligro”, “terror”, “venganza”, “ruido”, and “silencio”.

All in all, the top feature weights for historical versus sentimental novels are balanced in that there are as many features that are distinctive for the historical as for the sentimental novel. They also alternate by importance in the range of top 0 to 25. Besides the fight topic, other highly weighted topic features for the historical novels are concerned with crime and prison (“crimen-cárcel-preso”), politics and administration (“gobierno-pueblo-ministro”, “policía-gobierno-pueblo”, “virrey-oidor-señoría”), monarchy and nobility (“príncipe-rey-palacio”, “conde-condesa-señorito”), the countryside (“caballo-camino-árbol”, “caballo-hacienda-compadre”), church (“cura-iglesia-sacerdote”), and another topic about battles (“soldado-enemigo-fuego”). These topics point to

[583] The sign of the feature weights is determined by the SVM classifier and is not directly related to the order of the classes as first and second.

Figure 55. Topics “v_d-instante-corazón” and “tía-do-aire”.

different subtypes of historical novels dealing with colonial and contemporary history and different urban and rural surroundings.

Among the top topics for historical novels, there is also one which is a bit more difficult to interpret: “v_d-instante-corazón”. It is dominated by the historical form of address “vuestra merced”, which is actually not a noun but was misinterpreted by the linguistic tagger because of the abbreviation “vd”. The topic, which is shown in figure 55, is a bit more abstract and mixed than the other top topics for historical novels. It seems to be about physical (“pecho”, “pena”, “doctor”) and mental states (“alma”), feelings (“temor”, “duda”, “amor”, “placer”, “satisfacción”, “esperanza”, “corazón”, “desgracia”, “felicidad”), and thoughts (“pensamiento”, “idea”), but it also contains other words that are not related to these aspects. The words “instante”, “puerta”, and “vista” together with “fisonomía”, “cabeza”, “rostro”, “gracia”, and “virtud” seem to describe a moment in which someone is observed and the appearance of a person is described. The impression that this topic gives is that it joins together several thematic aspects that are not prototypical for historical novels but are accompanying material related to the plots and the representation of the characters’ feelings.

Besides the ball topics, the other top distinctive topics for sentimental novels cover private conversation (“voz-cabeza-mirada”), love (“amor-corazón-alma”, “idea-pasión-alma”), family relationships (“mamá-muchacho-papá”), material aspects (“v_d-país-negocio”), interiors and meals (“mesa-puerta-sala”, “mesa-copa-vino), art (“poeta-maestro-artista”), and travel or movement (“coche-carruaje-cochero”). The three topics “tía-do-aire”, “ojos-años-manos”, and “ojos-hija-años” need a closer look. In the first one, which is visualized in figure 55, the word “do” sticks out because it is not a regular Spanish word.[584] The novels in which this topic has the highest probability were checked, and the “do” is a spelling error (instead of “de”) that remained unnoticed in some of these novels.[585] Moreover, the other top words of this topic suggest that it can be interpreted as

[584] The word “do” does not appear in the word cloud of the topic.

[585] In the novel “Alma de niña” (nh0082), for instance, it occurs 58 times, and in the novel “Auras de Abril” (nh0233), 20 times. The spell-checking result files for these novels show that the spell-checker did not recognize this word as an error. See https://github.com/cligs/conha19/blob/main/spellcheck/results/spellcheck_nh0082.csv and https://github.com/cligs/conha19/blob/main/spellcheck/results/spellcheck_nh0233.csv. A plot showing the top 20

being about outward appearances and physical encounters because it contains words referring to people (“muchacho”, “hijas”, “mujeres”, “viejita”, “virgen”), body parts (“cabeza”, “ojos”, “labios”), clothes (“camisa”, “pana”, “abrigo”, “alfiler”), meeting places (“sitio”, “calle”, “playa”, “pieza”, “sillas”, “cama”, “flores”, “vela”), and also the word “besos” and “belleza” occur in it. The word “aire” can then be read in its sense of “look” and “airs and graces”. The other two topics that were not directly clear from the three top words (“ojos-años-manos” and “ojos-hija-años”) can be interpreted better when more of the top words are considered. The first one has mixed meanings and appears more abstract than the other top topics for sentimental novels. It can be described as treating views, nature, body, and time. The second one is about looking, also talking, mainly female persons, and emotions. In the first case, the word “años” probably refers to time, and in the second case, it may refer to age.[586] Summing up the findings about the topics that are most important for the classifier to distinguish historical from sentimental novels, it can be stated that most topics can be interpreted easily from the top words and some by examining them more closely. The topics of the historical novels are more easily recognizable as themes, whereas some topics of the sentimental novels are more subtle or mixed but also make sense for the concept of the subgenre. For the aspects covered by the topics that are at first sight less clear, it may have been favorable to also have verbs and adjectives and not only nouns in the topics. In what follows, not all the topics can be analyzed in the same depth as the ones that are most distinctive for historical versus sentimental novels, and the presentation of the results concentrates on the three top topic words unless there are aspects that need to be clarified or are of special interest. Nevertheless, a look into more top words of the topics and, if needed, also into the texts of the novels shows that the topics can be interpreted in more depth to characterize the subgenres that are classified. The next subgenre constellation for which the feature weights are analyzed is *novela histórica* versus *novela de costumbres* (see figure 56).

The positive bars to the right are the topics that are distinctive for the historical novels and the negative bars to the left are the ones that are distinctive for the novels of customs. Interestingly, the top topic for novels of customs is the same as for sentimental novels: “baile-salón-fiesta”, which shows that the text types of these two kinds of conventional subgenres have common stylistic traits. The top topic for the historical novels is, in this case, the topic of battles: “soldado-enemigo-fuego”. The number of top topics for each subgenre is less balanced in this case than for the constellation *novela histórica* versus *novela sentimental* because, among the top 25, there are 10 topics for the historical novels and 15 for the novels of customs. This is a sign that the novels of customs are more diverse in terms of characteristic topics. Besides the first one, the other topics with high weights that characterize the novels of customs are concerned with lower class work (“muchacho-dinero-año”), with tobacco, sugar, and coffee plantations and work on them (“gallo-finca-negro”, “amo-estancia-muchacho”), which clearly points to the Cuban novels, with work on a ranch (“cutter-indio-tierra”), which is a topic that seems to be related to cattle and sheep breeding in Argentina, with the countryside (“pueblo-molino-sol”, “caballo-camino-

novels for this topic is available at https://github.com/cligs/data-nh/blob/main/analysis/features/topics/5_visuals/90tp-5000it-250in-0/topItems/idno/tI_by-idno-028.png. A January 7, 2021.

[586] See the word clouds at https://github.com/cligs/data-nh/blob/main/analysis/features/topics/5_visuals/90tp-5000it-250in-0/wordles/wordle_tp089.png and https://github.com/cligs/data-nh/blob/main/analysis/features/topics/5_visuals/90tp-5000it-250in-0/wordles/wordle_tp035.png, respectively. Accessed January 7, 2020.

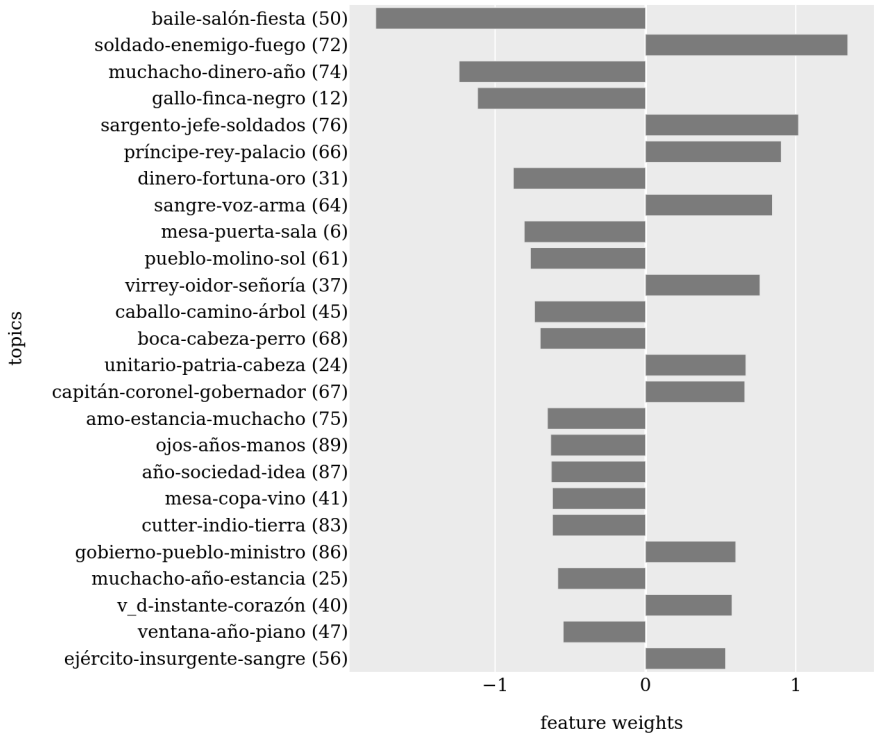


Figure 56. Feature weights (topics) for novels of customs versus historical novels.

árbol", "muchacho-año-estancia"), money, business, and gambling ("dinero-fortuna-oro"), meals ("mesa-copa-vino"), the description of rooms ("mesa-puerta-sala"), the description of characters ("boca-cabeza-perro"), including pets (dogs and cats), youth and student lifestyle ("ventana-año-piano"), and reflections about society ("año-sociedad-idea"). Of these topics, the one about meals was also among the top topics for sentimental novels. Furthermore, the mixed topic "ojos-años-manos" about views, nature, body, and time appears again. Among the top topics for the novels of customs, there are several that cover rural life and surroundings, which confirms that the *novelas de costumbres* were oriented towards that sphere. In addition, the aspect that the life and working conditions of people are realistically described in the novels of customs becomes visible in the top topics. Finally, there are several descriptive topics as well as a reflective one, which is in line with the aim of the novel of customs to represent different areas of society closely and also to provide a social critique or vision. For inspection, the two descriptive topics "mesa-puerta-sala" and "boca-cabeza-perro" are given in figure 57.

The other top topics for the historical novels are very similar to the ones that also appeared in contrasting the historical novels with the sentimental ones, and they are semantically much more homogeneous than the various top topics of the novels of customs. Again, elements of colonial history and also contemporary politics are present. A topic that is new here is, for example, "unitario-patria-cabeza", which is related to the struggle between Unitarians and Federalists that took place in Argentina in the first half of the nineteenth century. The last constellation of

Figure 57. Topics “mesa-puerta-sala” and “boca-cabeza-perro”.

thematic subgenres for which the coefficients of the linear classification model are analyzed is *novela de costumbres* versus *novela sentimental* (see figure 58).

Among the top 25 topics, there are 14 that are typical for the novels of customs and 11 that are typical for the sentimental novels. Many topics that were already relevant in the other subgenre constellations appear again, for example, the lower class work topic “muchacho-dinero-año”, the youth and student life topic “ventana-año-piano”, or the countryside topic “muchacho-año-estancia” as topics that are distinctive for the novels of customs, and the love topics “amor-corazón-alma” and “idea-pasión-alma”, as well as the private conversation topic “voz-cabeza-mirada” as topics that are typical for the sentimental novels. In addition, some new topics appear, for instance, “carta-papel-duda” for the sentimental novel, which addresses the writing and reading of letters, or “tío-sobrino-primo” for the novel of customs, which is a topic about relatives and cliques. Furthermore, some topics change the side, that is, they become distinctive for another subgenre than in the other constellations that were already examined. This is the case for the airs and looks topic “tía-do-aire”, which is now typical for the novel of customs but was important for the recognition of sentimental novels when they were contrasted with historical ones, and also for the abstract views, nature, body, and time topic “ojos-años-manos”, which was distinctive for novels of customs in contrast with historical novels and is now typical for sentimental novels when compared to novels of customs. This fluctuation shows that even though some characteristic traits of the subgenres remain constant independently of the subgenre constellation, others are relative and depend more on the kind of subgenres that are compared. The fact that topics are, in one case, typical for the sentimental novel and, in the other, for the novels of customs is again a sign of how these two subgenres are intertwined thematically. That topics are in one case typical for the sentimental novel and in the other for the novels of customs is again a sign of how these two subgenres are intertwined thematically. The plots for the constellations of the individual thematic subgenres against all other novels are not shown in detail here. What is interesting about them is that in all three cases, there are more top topics for the positive class than for the other group, which shows that the classifier focuses on the

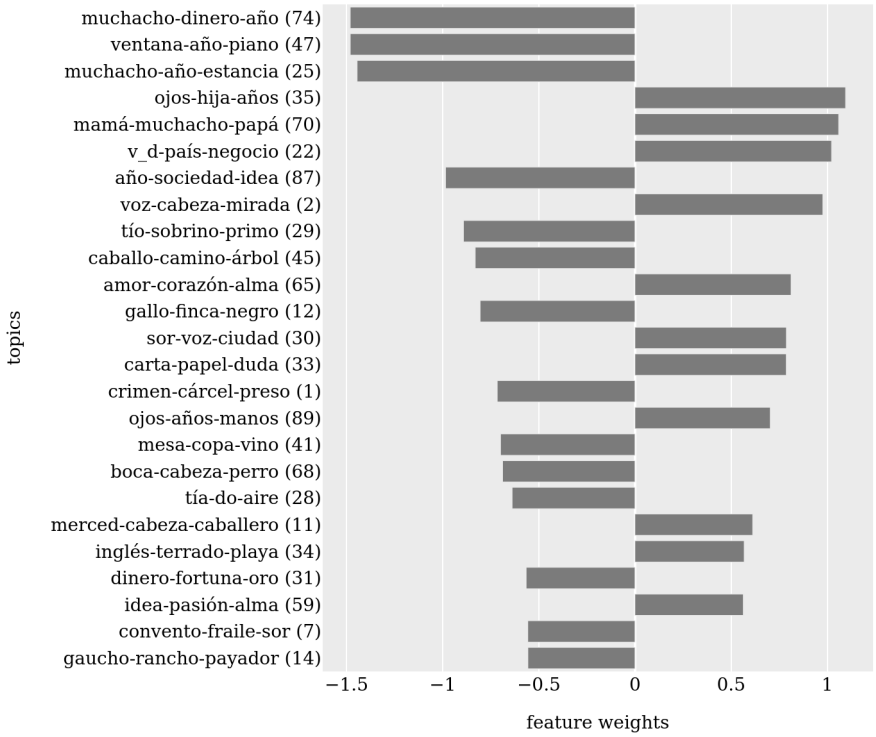


Figure 58. Feature weights (topics) for novels of customs versus sentimental novels.

aspects that are specific for the individual subgenre that is compared to the big group of novels with many different subgenres, which makes complete sense.[587]

Besides the feature importances, also the cases of correct and false classifications of individual novels were analyzed for each subgenre constellation, with the aim to find out how many and which of the texts that are part of the conventional subgenres are typical or untypical for the text types. The results of this analysis are summarized in the form of histograms, which show the distributions of true positives, false positives, and false negatives. The true positives are interpreted as instances of the conventional genre, as well as of the text type and hence of the textual genre. The false positives are instances of the text type but not of the conventional and textual genre, and the false negatives are part of the conventional genre but not of the text type and, therefore, also not of the textual genre. For each novel in the corpus, it was counted how often it fell into one of the three groups in absolute and relative terms. As the data selection process was random and repeated several times (ten times for the undersampling and ten times for the cross-validation), not every novel of the corpus is present in every classification, and some

[587] The three plots with the feature importances for *novela histórica* versus other novels, for *novela sentimental* versus other novels, and for *novela de costumbres* versus other novels can be seen at https://github.com/cligs/data-nh/tree/main/analysis/classification/themes/visuals and are called “feat_imp_SVM_90t_250oi_topic-rep_0_novela_histórica_other.html”, “feat_imp_SVM_90t_250oi_topic-rep_0_novela_sentimental_other.html”, and “feat_imp_SVM_90t_250oi_topic-rep_0_novela_de_costumbres_other.html”, respectively. Accessed January 7, 2020.

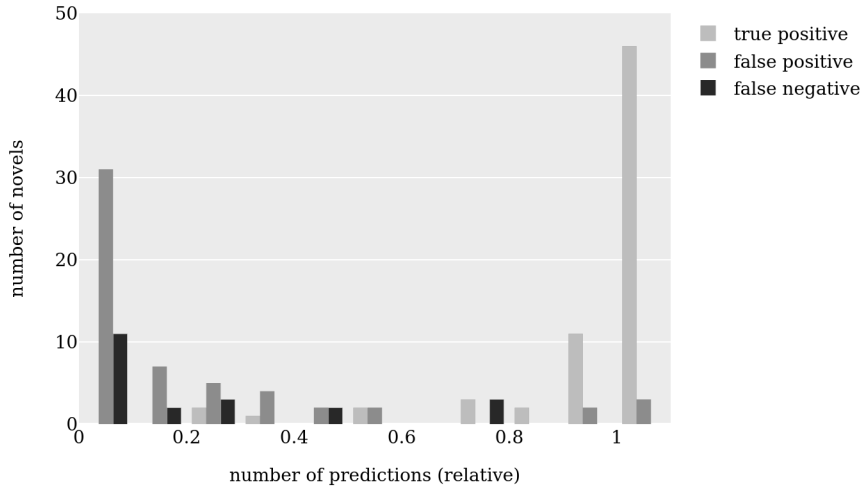


Figure 59. Predictions for *novela histórica* versus other novels (topics).

novels were classified more often than others. The relative numbers of correct classifications and misclassifications can still be used to examine how prototypical the novels are for the subgenres. The chart for *novela histórica* versus other novels is given in figure 59.[588]

Of special interest are the bars on the right side because they mean that the novels covered by them were classified correctly or wrong in many cases. In the histogram, the bars are grouped in steps of 10 %, starting with 0–9 % and ending with 90–99 %, and then 100 %. So the rightmost group of bars stands for the novels that were classified correctly or wrong in 100 % of the cases. If the novels were classified correctly, they can be interpreted as prototypical instances of the subgenre. If not, there is a discrepancy between the conventional and the textual genre for these novels. Of course, if a novel was only classified once and then correctly, it would appear as prototypical. That characterization would be less sure than for a novel that was classified 100 times and each time correctly, so these details have to be taken into account when this kind of chart is interpreted. For the constellation *novela histórica* versus other novels, 252 of the 256 novels in the corpus were classified at least once, among them all the 67 historical novels. Of these, 46 (69 %) were classified correctly in all 100 classification runs and can thus be considered the prototypical core of the text type *novela histórica*. Among them are, for example, the romantic historical novel "La cruz y la espada" (1866, MX) by Eligio Ancona and the realist historical novel "Puebla" (1903, MX) by Victoriano Salado Álvarez, which both carry the explicit historical subgenre label "novela histórica" and for which there is also agreement among literary historians to classify them as historical novels. As an example of such prototypical historical novels, the

[588] The charts about the classification results for individual novels are available in the "visuals" folder in the data-nh GitHub repository (see the previous footnote). The data was also collected in CSV files (called "misclassifications…"), which can be viewed in the folder https://github.com/cligs/data-nh/tree/main/analysis/classification/themes/results_summaries. Accessed January 7, 2020.

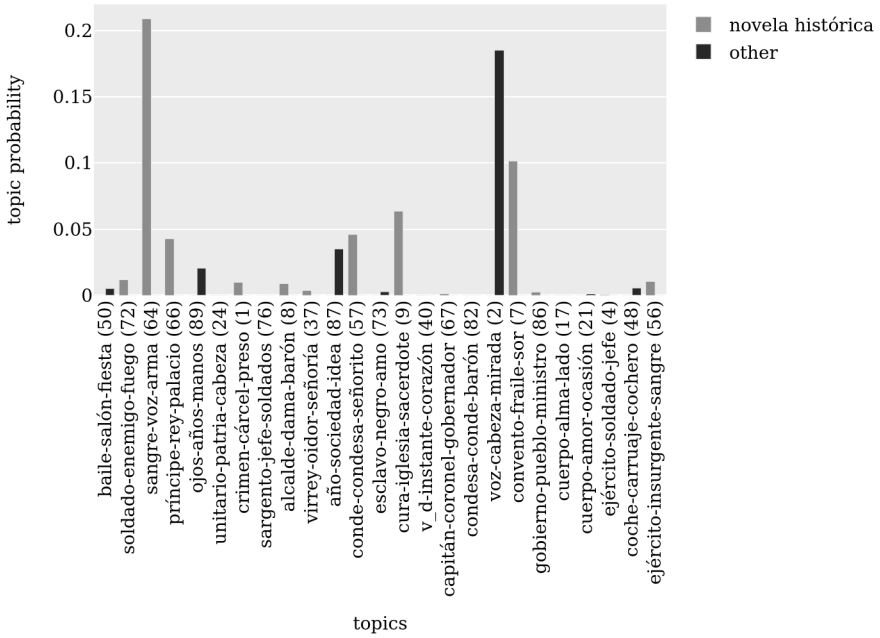


Figure 60. Top topics for *novela histórica* versus other novels in the novel "La cruz y la espada".

probabilities of the topics that are most distinctive for historical versus other novels are visualized for the novel "La cruz y la espada" in figure 60.[589]

In the plot, the topics are ordered by the weight that they have for the distinction of historical and sentimental novels in general, so that "baile-salón-fiesta" is the topic with the most weight and "ejército-insurgente-sangre" with the least. The colors indicate for which subgenre the topics are typical. The plot for "La cruz y la espada" shows that several topics that are distinctive for historical novels also have higher probabilities in this novel. For example, this is the case for "sangre-voz-arma", "príncipe-rey-palacio", "conde-condesa-señorito", "cura-iglesia-sacerdote", and "convento-fraile-sor". Especially the last four topics characterize this novel as one that is set in a more remote past (it is about the Spanish conquest of Mexico) because they refer to monarchy, nobility, and church. So the novel is a specific subtype of the historical novel, but still always classified correctly. On the one side, it has a high probability for the topic "sangre-voz-arma", which is a general topic for historical novels, but there are also top topics of historical novels which are not so important in "La cruz y la espada", as, for example, "soldado-enemigo-fuego", "virrey-oidor-señorita", "capitán-coronel-gobernador", or "gobierno-pueblo-ministro". The first one of these is a general topic about battles. The second would be typical for a novel that is set in the colonial era because it mentions the viceroy and "oidor", which was a judge in the colonial judicial system. The last two would be typical for a novel that treats contemporary history. This shows that the top topics of historical novels, in general, combine elements of different subtypes

[589] The probabilities are used in the form they have after the MinMax scaling, which sets the range of the individual features to [0,1]. This form is used here because it was the form of the features that were also used by the classifier.

of the subgenre. As the case of "La cruz y la espada" shows, it is not necessary for each individual novel to have high values for all of these topics in order to be always classified correctly as a historical novel. This means that not only the idea of prototypically organized categories is present in the statistical classes that are determined in the machine learning process, but also the idea of family resemblance: The novels can have high probabilities of specific subsets of the most distinctive topics, but they do not all need to have the same distributional profile. However, the aspect of family resemblance that is covered by the statistical classification is also limited because the boundaries between classes are still strictly drawn, and there are no loose networks of novels with overlapping similarities. What can be done is to interpret the novels inside of a class as members of a family that share different characteristics, or to analyze novels at the edges between two classes, that is, novels that are often misclassified for each class, to see to what extent they share properties.

In the case of "La cruz y la espada", not only the topics that are distinctive for historical novels are of interest, but also the ones that are typical for other types of novels because this novel also has high probabilities for some of these topics. For example, a top topic that has much weight in the novel is "voz-cabeza-mirada", and also "año-sociedad-idea" and "ojos-años-manos" have higher weights. However, in sum, these topics are less dominant than the ones that are typical for the historical novel. What this shows is that elements that are typical for other subgenres can be present and have a certain weight, but as long as they are not dominating, the novel is still classified correctly. So the class of a novel is really determined in quantitative terms: which topics "win" in terms of numbers? This also means that a prototypical historical novel can have elements of other subgenres, as well. This aspect is particularly useful in the case of the nineteenth-century Spanish-American historical novels, of which many were romantic and included a sentimental plot or were more realistic and contained passages with descriptions of customs. In this view, instances of prototypes are not necessarily pure but have certain feature values and distributions that are quantitatively dominating and can be interpreted as salient.

Returning to the group of prototypical historical novels that were always classified correctly, there are also novels that have been discussed not only as historical novels but also in other terms by literary historians. For example, the novel "Amalia" (1855, AR) by José Mármol has been described as a political novel, a historical novel, and also a sentimental novel, but is here always classified as a historical novel. Regarding the historical genre conventions, an interesting aspect of this novel is that the main title, "Amalia", is typical for sentimental novels because it refers to a female first name, but the novel also has the subtitle "novela histórica" in most editions that are published from 1874 onwards. It was therefore labeled with the primary thematic label "novela histórica" here. The results of the SVM classifier make clear that the sentimental plot elements do not prevent this novel from being classified as a historical one. The four novels of the series "Dramas militares", which were published between 1884 and 1886 by Eduardo Gutiérrez and in which the protagonist is an Argentine gaucho, are also always classified as historical novels.

There are three novels that resulted as false positives in every case and two novels for which this happened in more than 90 % of their classifications so that they can be interpreted as belonging to the text type of the historical novel, although they did not have the primary thematic label "novela histórica". One of them is "Las gentes que son así (Perfiles de hoy)" (1872, MX) by José Tomás de Cuéllar, which is a novel of customs that is part of the series "La linterna mágica". It

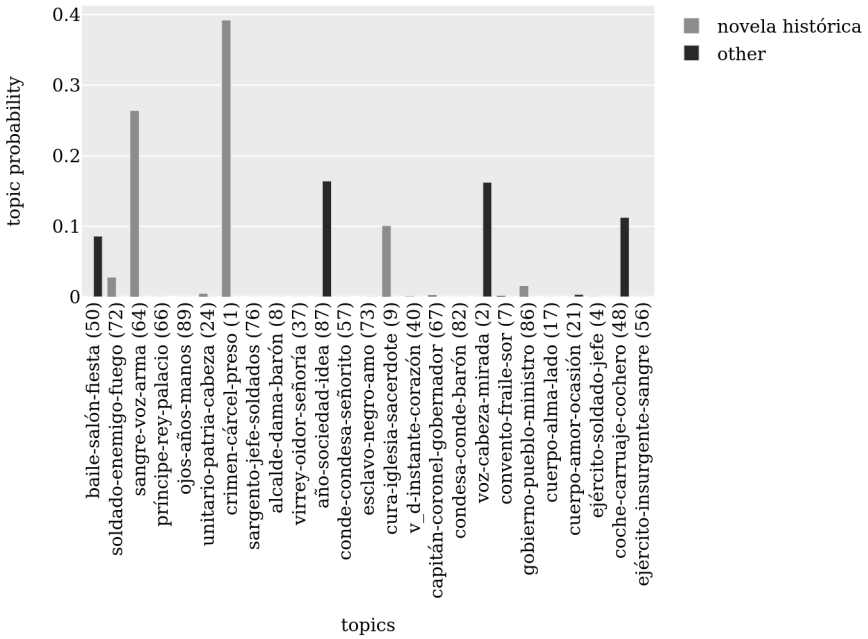


Figure 61. Top topics for *novela histórica* versus other novels in the novel "Las gentes que son así".

was classified as a historical novel 60 times out of 60. Another one, which was classified 39 times as a historical novel and one time as "other", is "Los bandidos de Río Frío. Novela naturalista, humorística, de costumbres, de crímenes y de horrores" (1892, MX) by Manuel Payno, which has been assigned the primary thematic label "novela de costumbres" in the corpus. Already the subtitle shows that it refers to several genre conventions at once. It is about banditry and has also been classified as a novel of customs by Brushwood:

> Manuel Payno, for example, published another serial novel *Los bandidos de Río Frío*, from 1889 to 1891. The subtitle even states that it is 'naturalistic', but it has much Romantic overstatement that identifies it with an earlier time. Taking the theme of banditry that had by that time become very popular, Payno wrote another essentially *costumbrista* novel. (Brushwood 1966, 116)[590]

In these two example cases, it is not entirely clear why they are repeatedly classified as historical novels, so the top topic profiles of them are checked to see if they explain these results (see figures 61 and 62).

As to the novel "Las gentes que son así", a look into the probabilities that it has for the topics that are most distinctive for historical novels reveals that especially the topics "crimen-cárcel-preso" and "sangre-voz-arma" have high values. The first one is about crime, justice, and imprisonment, and the second one is a general topic about fights. These topics are not exclusively historical but

[590] The publication date in the corpus differs from the one mentioned by Brushwood because it refers to the year of the first book publication.

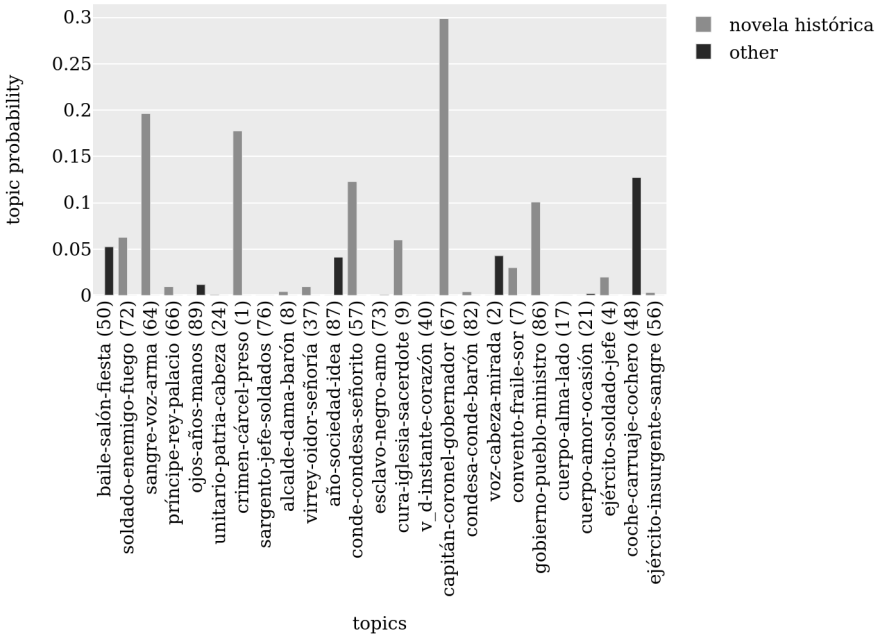


Figure 62. Top topics for *novela histórica* versus other novels in the novel “Los bandidos de Río Frío”.

can as well occur in novels of other subgenres. However, because they are *typical* for historical novels, the novel of customs “Las gentes que son así” in which they are so important, is mistaken as a historical novel, even if it does not have high probabilities for topics that are more specific for historical novels, such as “soldado-enemigo-fuego” or “sargente-jefe-soldados”. What this demonstrates is that the classifier has no idea about the conventional genres and does not make a difference between necessary, typical, or sufficient features. It just compares the topic values and distributions, and if it finds that they are similar to the ones typically found in a certain subgenre, the novel is classified as such. The example of “Las gentes que son así” is one where the novel has textual similarities to historical novels and is recognized as part of the corresponding text type, but it should not be treated as part of the textual genre, because by convention and also in terms of necessary features it is no historical novel.

The second case, “Los bandidos de Río Frío”, has another quality. The probabilities for the top topics that are typical for historical novels show that some of them also have high values in this novel, above all “capitán-coronel-gobernador”, “sangre-voz-arma”, “crimen-cárcel-preso”, “conde-condesa-señorito”, and “gobierno-pueblo-ministro”. They are more than in the novel “Las gentes que son así”, and even though the general fight and crime topics appear again, there are also topics concerned with the military, politics, and nobility. The novel “Los bandidos de Río Frío” was published in 1892 but is set in the 1830s. The male protagonist has a military position, and also social and political problems of banditry, insecurity, and corruption are treated in the novel. That the occupation of the protagonist has an influence on the topics of the novel and that this, in turn, affects how the novel is classified in terms of subgenre again makes clear that a

machine learning classifier cannot distinguish between different reasons for which topics are present in the texts. However, the combination of themes which are treated in "Los bandidos de Río Frío", together with the fact that it treats a period that is several decades away from its publication time, does not make it far-fetched to compare it to other novels of contemporary history, so here, the result is not considered so misleading as in for "Las gentes que son así".

The three other novels that frequently resulted as false positives are "Misterios del corazón" (1875/1897, AR) by Rafael Barreda, "La hija de Tutul Xiu. Novela yucateca" (1884, MX) by Eulogio Palma y Palma, and "La Chapanay" (1884, AR) by Pedro Echagüe. The first one was classified as a historical novel 50 out of 50 times, the second one 30 out of 30 times, and the third one 28 times as a historical novel and 2 times as "other". Barreda's novel has the primary label "novela sentimental" and the secondary thematic label "novela histórica" because it is set in the Rosas' era. Therefore, it is not surprising that it is mistaken as a historical novel, but what is surprising is that this happens every time. It seems that, in this case, the setting has a bigger influence on the text type than the intended primary theme of the novel. Or, from another perspective, aspects of the sentimental novel are not untypical for nineteenth-century Spanish-American historical novels and are part of the historical novel's text type. The case of "La hija de Tutul Xiu. Novela yucateca" can also be easily explained because it is an indianist novel set in the pre-Spanish period, and it also has the secondary thematic label "novela histórica" in the corpus. So here, the problem was that only the more specific thematic label was considered in the classification and not the more general secondary one. As to "La Chapanay", it has the primary thematic label "novela de costumbres". This novel is an account of the life of Martina Chapanay, a historical Argentine personality of the nineteenth century which is only of regional importance. Lichtblau summarizes the novel's contents as follows: "Regional types, customs, and the daily life of the San Juaneses are vividly portrayed in the novel as we see Martina first as a rebellious young girl, then as a member of a band of highwaymen, and finally as a contrite woman aiding the forces of law and order" (Lichtblau 1997, 299). This description identifies it as a novel of customs, but "La Chapanay" also contains passages in which the narrator expresses his political opinion and criticizes the Rosas regime (Lichtblau 1997, 299). So similarly to "Misterios del corazón", where it was the historical background that approximated the novel stylistically to other conventional historical novels, here there are argumentative passages that influence the text style in terms of topics. In addition, this is another case of a hybrid between the novel of customs and the historical novel.

As to the novels that were labeled as historical novels but often classified as other types (the false negatives), there is none that was misclassified in every run. There are three significant cases, though. The first one is "Los esposos" (1893, AR) by Lola Larrosa de Ansaldo, which was classified as "other" 77 times and as a historical novel 23 times. In the corpus, the novel has the primary thematic label "novela histórica" because it had this subtitle in the first edition. The second edition, in contrast, does not have that subtitle anymore. Moreover, Lichtblau describes the novel as follows: "A cloyingly romantic novel, Los esposos interweaves several plots that demonstrate the author's self-righteous morality and her traditional beliefs about women's place in the home" (Lichtblau 1997, 530). Finally, it also has the main title "Los esposos", which seems rather typical for a sentimental novel. In the corpus, it has therefore been assigned a secondary subgenre label: "novela sentimental". As it seems, the explicit historical label that this novel

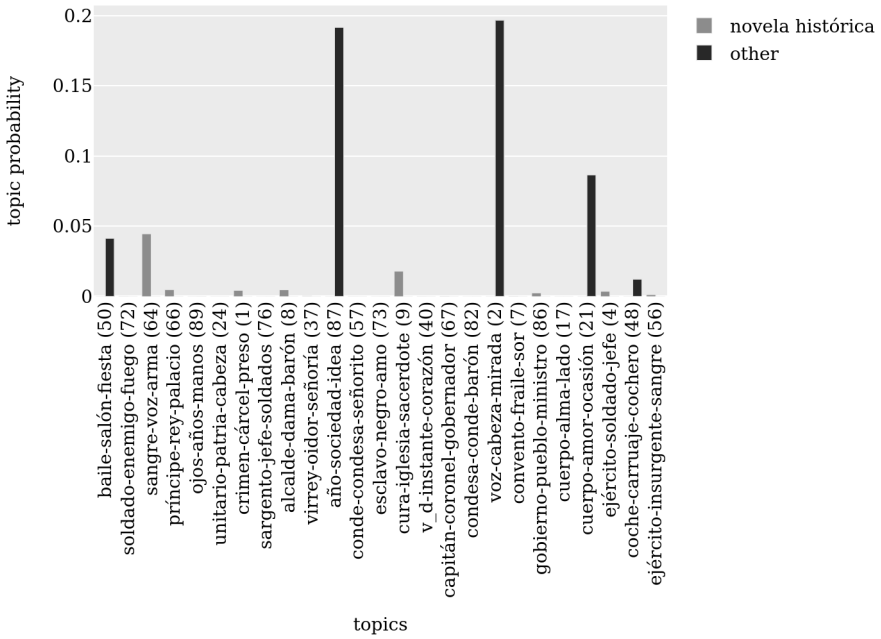


Figure 63. Top topics for *novela histórica* versus other novels in the novel "Los esposos".

carries is not entirely detached from the textual material and fits in 23 % of the cases, but the novel is textually closer to another subgenre, most probably the sentimental novel. A look into the weight that the top distinctive topics for *novela histórica* versus "other" have in this novel confirms that it does not correspond to the historical novels textually (see figure 63). The topic with the highest value is "voz-cabeza-mirada", which is indeed one of the top topics for sentimental novels when compared to others, but the other two topics, "año-sociedad-idea" and "cuerpo-amor-ocasión", are neither distinctive for historical nor for sentimental novels, and also not for the novels of customs,[591] but for other subgenres. As the novel was published in 1893, it is possible that it includes elements that are typical for the realist or naturalist novel. The topic "año-sociedad-idea" points in the first direction and "cuerpo-amor-ocasión" in the second one.

The second false negative is "Vía Crucis" (CU) by Emilio Bacardí Moreau, of which the first part, "Páginas de ayer", published in 1910, is included in the corpus. It was classified as "other" 73 times and as a historical novel 27 times. This novel is an account of the Cuban struggle for independence in the second half of the nineteenth century, which means that it treats contemporary historical events. It has nonetheless been characterized as a historical novel by several literary historians, for example, by Remos y Rubio, who compares Bacardí's style to the one of Pérez Galdós, a Spanish writer who published the famous series of historical novels entitled "Episodios Nacionales". He also relates Bacardí Moreau's style to the ones of Alexandre Dumas and Walter Scott (Remos

[591] At least when compared to the "other" group.

y Rubio 1935, 42–43).[592] However, Remos y Rubio also remarks that the novel "Vía Crucis" contains accomplished *costumbrista* passages: "Por el alto valor costumbrista que hay en ella, por lo admirable de las descripciones locales, como cafetales, campiñas, amenas, hogares criollos, etc., ha sido estimada *Vía Crucis* por algunos críticos, superior a *Cecilia Valdés*" (Remos y Rubio 1935, 44). The similarity with other novels of customs, especially the Cuban ones that also contain descriptions of coffee plantations, might be an aspect that leads to the frequent misclassifications in the case of "Vía Crucis". Its top topic profile is visualized in figure 64. It becomes visible that this novel has high weights for topics that are typical for the sentimental novel ("voz-cabeza-mirada", "baile-salón-fiesta", and "ojos-años-manos") and indeed a topic about slavery ("esclavo-negro-amo"), but none of the topics that are distinctive for novels of customs versus other novels are strong in this novel. Instead, the two topics "año-sociedad-idea" and "cuerpo-amor-ocasión" are important, which are the same ones as in "Los esposos". As "Vía Crucis" was published in 1910, it is a novel that is clearly not a romantic historical novel and also Remos y Rubio remarks on the new kind of historical novel that develops in the early twentieth century and of which "Vía Crucis" is an example: "La novela histórica se oscurece con el advenimiento de las nuevas tendencias, hasta que a partir de 1910 la reaniman con nuevos bríos y moderna fisonomía, el selecto Rodríguez Embil y el ameno narrador santiaguero, EMILIO BACARDI MOREAU" (Remos y Rubio 1935, 42). So the misclassifications of "Los esposos" and "Vía Crucis" can also be interpreted as signs of a new type of historical novel related to Realism, Naturalism, and Modernism, which is less frequent in general and in the corpus than the romantic type of historical novel and is therefore not recognized well, because it is not learned as part of the model of the quantitatively dominant historical novel. These results underline how important the construction of the corpus is for the digital text analysis, and the attention to balance subparts of it. However, if a certain type of novel simply was not frequent enough to be balanced against another type (as the modernist historical novel, for instance, against the romantic historical novel), then qualitative analysis or an approach that is different from statistical classification is more meaningful to describe its characteristics.

The third novel that carries the label "novela histórica" but is often misclassified is the novel "Las ranas pidiendo rey. Confesiones de una afrancesada (1861–1862)" (1903, MX) y Victoriano Salado Álvarez. It is part of the series of historical novels "Episodios Nacionales Mexicanos". The novel was classified as historical 30 times and as "other" 70 times. Like "Vía Crucis", this novel treats events of the recent past. Furthermore, it is written in the form of a diary and thus in the first person, which is quite unusual for a historical novel. Here it is assumed that especially the latter aspect turns it into a novel that is stylistically closer to other subgenres than the historical novel, but the narrative perspective does not need to have much influence on the topics. The plot for the novels' top topic probabilities is given in figure 65 to check that. What it shows is that the most important topic which is not distinctive for historical novels is "año-sociedad-idea", followed by "baile-salón-fiesta". The second one is typical for sentimental novels, and it might have more weight here because of the personal account that is given of the events, but not necessarily. It can also be an element that is inherited from romantic historical novels with a sentimental plot. The

[592] Remos y Rubio mentions 1914 as the publication date, but he refers to the second part of the novel (or the novel with both parts).

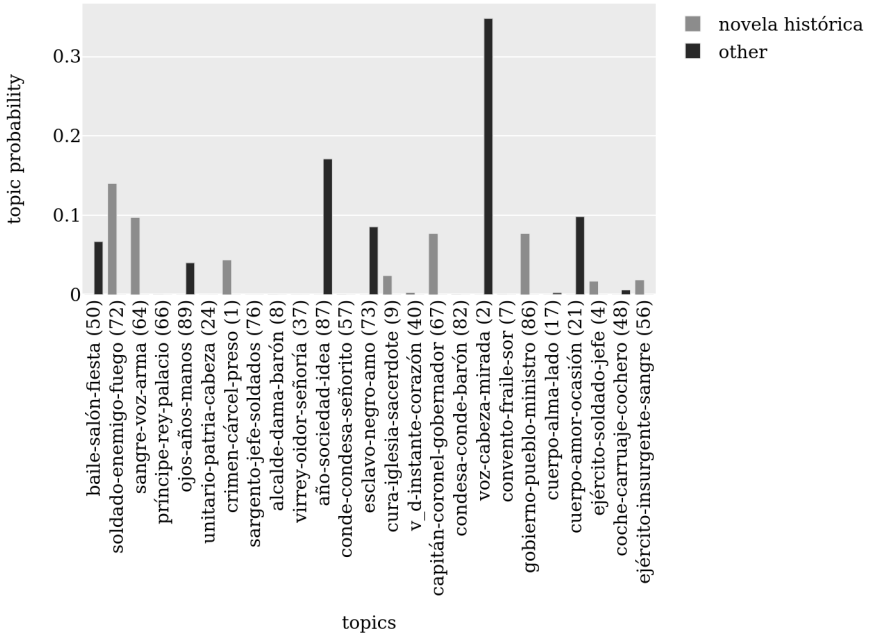


Figure 64. Top topics for *novela histórica* versus other novels in the novel "Vía Crucis".

other topic about intellectual work and society is the same that had much weight in "Los esposos" and "Vía Crucis". Now, "Las ranas pidiendo rey" was published in 1903, and the historical novels of Victoriano Salado Álvarez are commonly attributed to the realist current. So this is yet another example of a later type of historical novel that challenges the classifier.

The findings for the textual coherence of the historical novel can be summarized as follows: The majority of novels are prototypical historical novels and are both part of the conventional and the textual genre. There are only a few novels that are consistently misclassified or classified wrongly in more than 70 % of the cases. Five novels are regular false positives and interpreted as members of the text type but were, for different reasons, not labeled as members of the conventional genre. Three novels are persistently false negatives or members of the conventional genre but not of the text type, for individual reasons but also because they were published in the late nineteenth or early twentieth century. In all of these cases, the results can be explained by comparing and relating the textual and the conventional generic levels to each other, so their misclassifications are not considered real errors, except in the case of "Las gentes que son así". Howver, also in that case, the decision of the classifier can be explained. Besides, there are groups of novels that are only sometimes misclassified. For instance, 13 of the "novelas históricas" are classified as other types only in up to 20 % of the cases. They are still considered members of the text type but less prototypical ones. On the other hand, there are 38 novels of other subgenres that are classified as historical novels in up to 20 % of the cases. These are members of other text types that have similarities with the historical text type and that are stylistically located at the edge of the historical prototype category. For the sentimental novels and the novels of customs, no

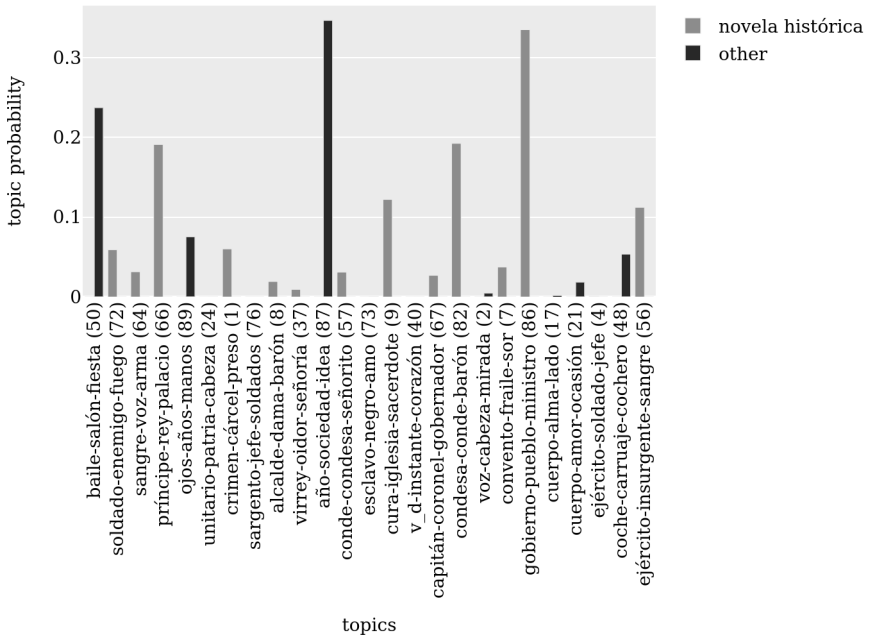


Figure 65. Top topics for *novela histórica* versus other novels in the novel “Las ranas pidiendo rey”.

individual cases are discussed, but their *generic profiles* are evaluated on a general level. The histograms for the constellations *novela sentimental* versus “other” and *novela de costumbres* versus “other” are given in figures 66 and 67.

All of the 55 sentimental novels in the corpus were included in the classifications for the constellation *novela sentimental* versus “other”. In total, 243 of the 256 novels were part of this contrast at least once. Like the historical novel, also the sentimental novel has a strong prototypical core, but it is smaller than in the case of the historical novel. Of the sentimental novels, 33 (60 %) were classified correctly in 100 of 100 cases, compared to 67 % of the historical novels. For the sentimental novel, there are twelve instances that are regular false positives (seven with 100 %, four with 90 or more, and one with 80 %) and three instances that are very frequent false negatives (one of 100 % and the other two with 97 and 89 %). So the number of novels that were not labeled as sentimental novels but are recognized as such on the textual level is higher than for the historical novels, while the number of novels that carry the label but are not textually congruent with the subgenre is the same. Otherwise, the distribution is similar to the one of the historical novel. Less prototypical are 15 novels that have the label but are not recognized as sentimental novels in up to 10 % of the classifications. At the edge of the category are 34 novels that are classified as sentimental novels in up to 20 % of the cases but were not labeled as such.

The novel of customs has 50 instances in the corpus, all of which were included in the classifications of *novela de costumbres* versus “other”. 241 novels out of the 256 in the whole corpus participated in these classifications. The histogram for the novel of customs looks a bit different

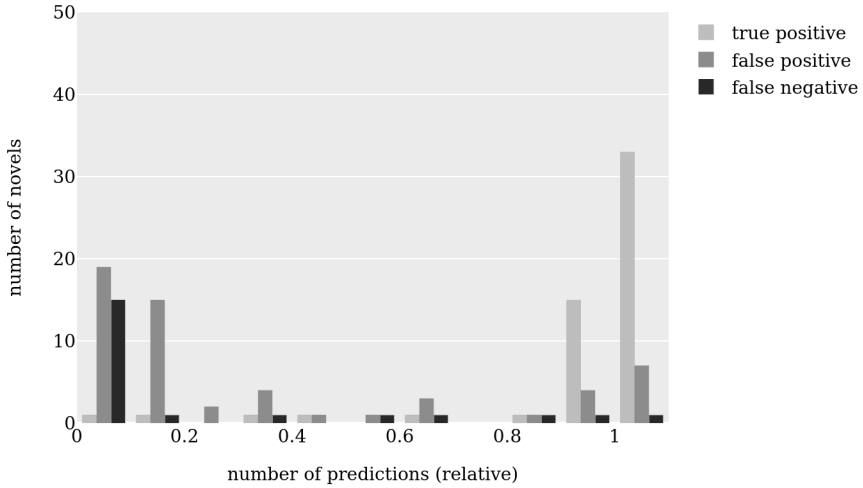


Figure 66. Predictions for *novela sentimental* versus other novels (topics).

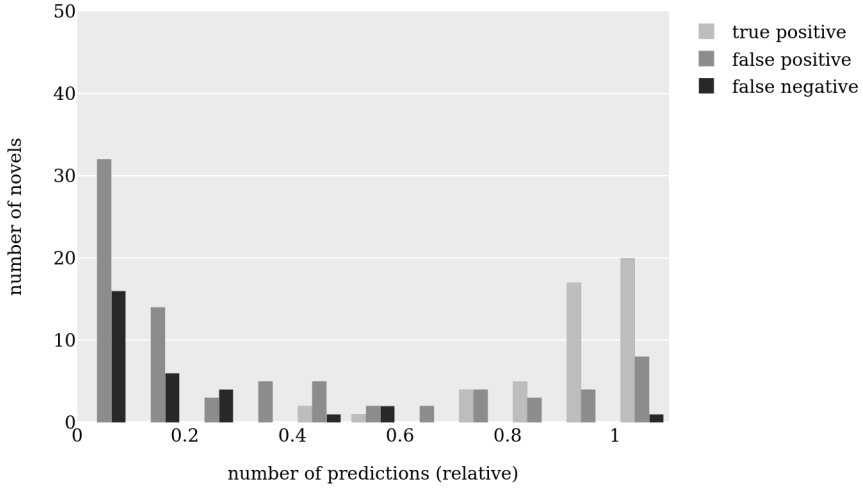


Figure 67. Predictions for *novela de costumbres* versus other novels (topics).

than the ones for the historical and the sentimental novel. The group of novels that are always or almost always classified correctly is smaller, and there are more false positives which are relatively frequent. The amount of frequent false negatives, in contrast, is not especially high. This means that the prototypical core of this textual genre is smaller: 20 novels (40 %) are classified correctly in 100 of 100 cases. In addition, more novels that are not labeled as novels of customs still belong to the text type (according to these results): 19 novels are classified as *novelas de costumbres* in 70 % of the cases or more. On the other hand, there is only one novel which is labeled as novel of customs and always classified as "other": the Cuban anti-slavery novel "El negro Francisco" (1873, CU) by Antonio Zambrana y Vázquez. An interpretation of these results is that the description of customs is a textual element in many of the nineteenth-century Spanish-American novels. However, it is not always marked on the level of historical genre convention and also not always considered as the primary thematic element by literary historians even if it dominates a text quantitatively. As was seen before, there are many different topics that are typical for the novels of customs, which might contribute to this lesser degree of overlap between the larger text type and the smaller conventional genre. The results for all three thematic subgenres demonstrate that the primary thematic labels that the novels have are not totally disconnected from the stylistic characteristics of the texts; quite the contrary. Nevertheless, there are also novels in which the conventional subgenre identity does not correspond to the textual one, either completely, as in the extreme cases, or in part. All constellations are possible on the way from a prototypical core to the edge of a textual genre, but some are more frequent than others, depending on the kind of subgenre that is analyzed.[593]

In what follows, the classification results for the MFW-based features are presented. These features consist of three main groups: words, word n-grams, and character n-grams. The results for each group are reported separately. In total, 18,000 classification runs were considered for basic MFW, 54,000 for word n-grams, and 162,000 for character n-grams.[594] The results for the MFW-based features are summarized in table 38.[595]

With the KNN classifier, a mean accuracy of 0.68 is reached with MFW features and also with character n-gram features. The mean accuracy for word n-grams is only 0.64. With SVM, MFW yield the best mean accuracy value of 0.77, character n-grams 0.76, and the word n-grams 0.75, so here, the results for the three subtypes of MFW-based features are very close. The RF works best with MFW features, which lead to a mean accuracy of 0.78, followed by character n-grams with 0.76 and word n-grams with 0.74. So MFW are the best token unit across all classifiers, and character n-grams work equally well for KNN, but word n-grams work less well for all three classifiers. SVM and RF are almost equally successful and have the same average accuracies across the different types of feature sets (MFW, word n-grams, and character n-grams).

---

[593] Here only the plots for the contrast of individual subgenres with the "other" group were shown. Another possibility is to analyze the direct subgenre comparisons to investigate how two selected subgenres relate to each other in terms of prototypicality.

[594] The different numbers of runs are due to the different amounts of parameter constellations for each feature type. For basic MFW, there is just one token unit (word); for word n-grams, there are three; and for character n-grams, nine. In addition, all of the ten repetitions with the different data samples and ten cross-validation runs per constellation are included.

[595] The scores and standard deviations are rounded to two decimal places.

Classifier	Feature type	Top accuracy	Mean accuracy	SD accuracy	Top F1	Mean F1	SD F1
KNN	**MFW**	**1.0**	**0.68**	**0.15**	**1.0**	**0.68**	**0.18**
	MFW word n-grams	1.0	0.64	0.14	1.0	0.64	0.17
	MFW character n-grams	**1.0**	**0.68**	**0.15**	**1.0**	**0.69**	**0.17**
	all	1.0	0.67	0.15	1.0	0.67	0.17
SVM	**MFW**	**1.0**	**0.77**	**0.15**	**1.0**	**0.77**	**0.16**
	MFW word n-grams	1.0	0.75	0.16	1.0	0.74	0.17
	MFW character n-grams	1.0	0.76	0.15	1.0	0.76	0.16
	all	1.0	0.76	0.15	1.0	0.76	0.17
RF	**MFW**	**1.0**	**0.78**	**0.14**	**1.0**	**0.77**	**0.15**
	MFW word n-grams	1.0	0.74	0.15	1.0	0.73	0.16
	MFW character n-grams	1.0	0.76	0.15	1.0	0.76	0.15
	all	1.0	0.76	0.15	1.0	0.75	0.16

Table 38. Classification results for primary thematic subgenres (MFW).

Because RF is slightly better for the basic MFW, it is chosen here for evaluating the classification results for primary thematic subgenres with MFW-based features further. Figure 68 displays the classification results for different numbers of MFW and tf-idf versus z-scores.[596]

It is directly visible that the lower numbers of MFW are not suited for classification by thematic subgenre. On the other hand, the results stay at a high level, approximately from the 1,000 MFW onwards. As was shown in the chapter on the MFW-based features, in the range of 1,000 MFW, there are already verbs, adjectives, adverbs, and nouns, so it can be concluded that these content words are essential for the classification by thematic subgenre.[597] As to the difference between tf-idf and z-scores, the former are better up to 3,000 MFW and the latter from 4,000 MFW onwards. When one rounds to two decimal places, the top mean accuracies of 0.81 are reached with tf-idf and 3,000 or 2,000 MFW.[598] Because of the above results, tf-idf and 3,000 MFW

[596] The results for tf-scores are not visible here because they are almost identical to those for z-scores because the latter are based on the tf-scores, and the scaling of the features is not decisive for the RF classifier.

[597] See chapter 4.2.1.1 above.

[598] Hettinger et al. (2015) also found out that 3,000 MFW worked best for the classification of subgenres of the novel with SVM. This is astonishing because they classified nineteenth-century novels in German language and not Spanish. Here, also the plot for SVM was checked, and there the highest mean accuracies are at 3,000 and 4,000 MFW. As Hettinger et al. also analyzed mainly thematic subgenres (adventure novels, social novels, and education novels), it can be hypothesized that a range of 2,000 to 4,000 MFW is a good feature choice for the classification of thematic subgenres, independently of the language and classifier.

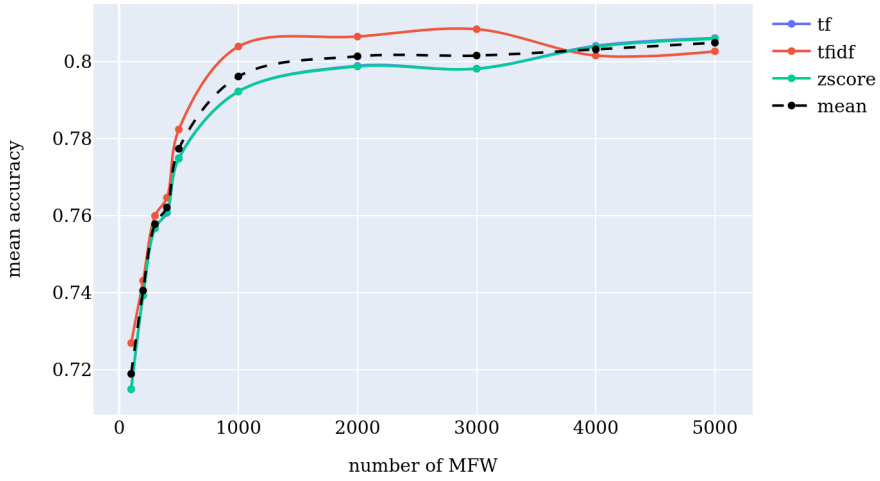


Figure 68. Classification results for MFW feature sets (RF, varying number of MFW and normalization technique).

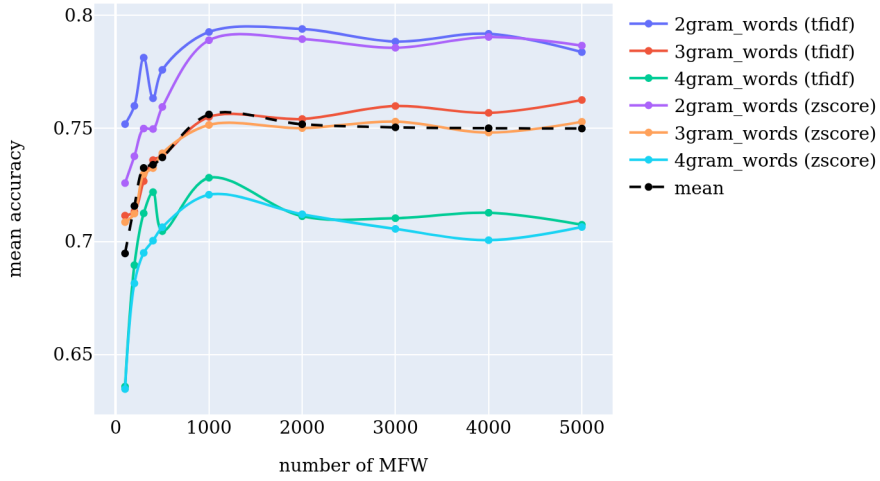


Figure 69. Classification results for word n-gram feature sets (RF, varying number of MFW, grams, and normalization technique).

are chosen to analyze the classification results further. Next, word n-gram features are examined. The classification results relying on these feature sets are summarized in figure 69.

In general, word 2-grams give higher mean accuracies than 3-grams or 4-grams, which was expected because the more words are involved, the less frequent the combinations are in the corpus and potentially less helpful to classify texts by genre if they only occur in a few texts. Here, too, the results are generally better from 1,000 MFW onwards than with fewer MFW, but they also drop again for the 2-grams and 4-grams, even if slowly. The differences between tf-idf values and z-scores are very small. The best mean accuracy is achieved with tf-idf, word 2-grams, and 2,000 MFW, which is at 0.79. If the values are rounded to two decimal points, the same result

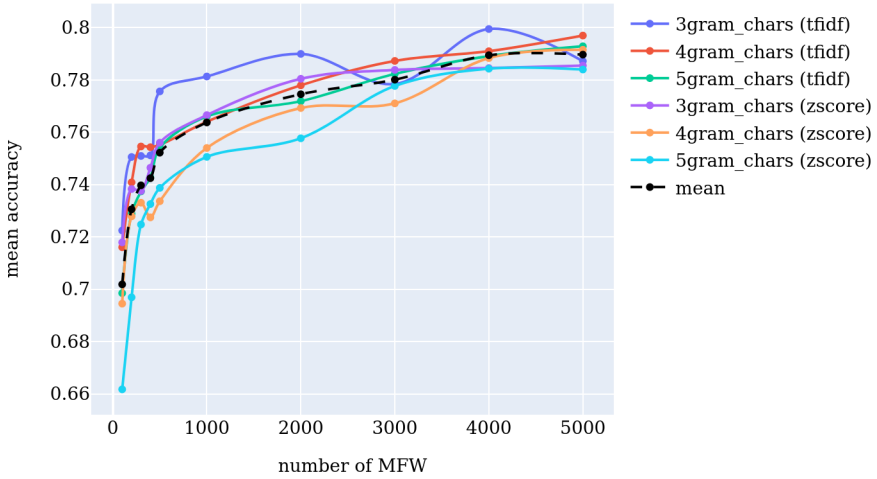


Figure 70. Classification results for classic character n-gram feature sets (RF, varying number of MFW, grams, and normalization technique).

is achieved with 2-grams from 1,000 to 4,000 MFW with both tf-idf and z-scores, and also with 5,000 MFW and z-scores, so the differences between these different constellations are minimal.

The third set of MFW-based feature sets that is examined are the character n-grams. Here, different parameter constellations for three n-gram subtypes are evaluated. The first subtype are the "classic" character n-grams containing all characters, punctuation marks, and blank spaces. The classification results for these are summarized in figure 70.[599]

With character n-grams, the best mean accuracy is achieved with the combination of tf-idf, 3-gram characters, and 4,000 MFW, reaching 0.80. With values rounded to two decimal points, the same result is achieved with tf-idf, 4-gram characters, and 5,000 MFW.[600] Besides the classic n-gram features, also two special types of n-grams were created. The first of them only uses mid-word and multi-word n-grams and is called "word n-grams" and the second type only uses prefix n-grams as well as n-grams ending with punctuation marks and is called "affix-punct".[601] In the following, it is checked if they have advantages over the classic n-grams. Figure 71 visualizes the results for the "word" character n-grams.

With the "word" character n-grams, the best result is an average accuracy of 0.80, which is reached with tf-idf, 3-grams, and 1,000 MFW and also with tf-idf, 4,000 MFW, and 4-grams or 5-grams, as well as with tf-idf, 5-grams, and 3,000 MFW (rounded values). The top mean accuracy is the same as for the classic n-gram features, so it was no advantage to only use n-grams that are derived from words or word boundaries. The classification results for the second special type of n-gram features, the "affix-punct" character n-grams, are given in figure 72.

---

[599] The default n-grams are provided by the CountVectorizer of scikit-learn. See chapter 4.2.1.1 on the MFW-based features for details.

[600] Hettinger et al. (2015) used the top 1,000 character 4-grams for the classification of German novels by subgenre.

[601] See chapter 4.2.1.1 above, in which these special types of n-gram features are explained.

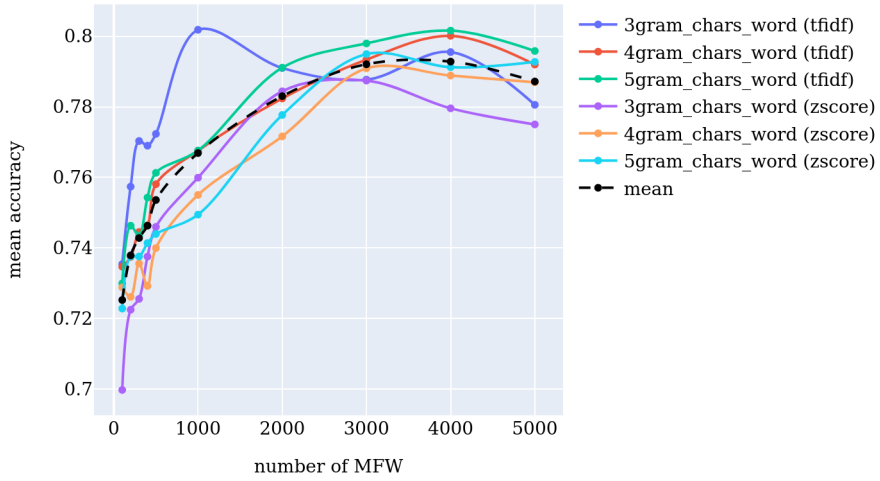


Figure 71. Classification results for “word” character n-gram feature sets (RF, varying number of MFW, grams, and normalization technique).

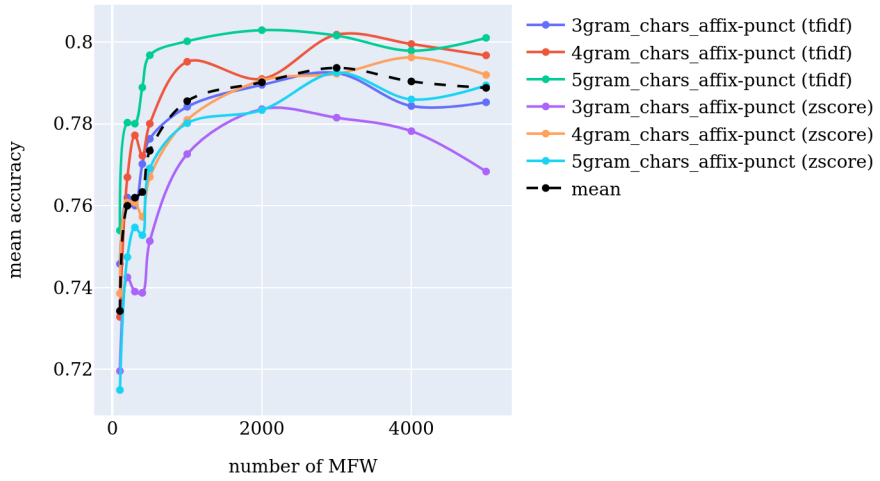


Figure 72. Classification results for ”affix-punct” character n-gram features sets (RF, varying number of MFW, grams, and normalization technique).

In this group of feature sets, the highest mean accuracies of 0.80 are achieved with character 5-grams and tf-idf values for several numbers of MFW (1,000 to 5,000, with accuracies rounded to two decimal places). The 4-grams with tf-idf reach this value several times, too. Again, there is no difference in the mean accuracy compared to the other types of n-grams, so for the classification of the nineteenth-century Spanish America novels by thematic subgenres, it is not decisive which kind of n-gram features are used.

When the findings for different token units are summarized, it can be observed that the top mean accuracies are all very similar, as they range from 0.79 in the case of word 2-grams over 0.80

Subgenre 1	Subgenre 2	Top accuracy	Mean accuracy	SD accuracy	Top F1	Mean F1	SD F1
novela histórica	other	1	0.86	0.09	1	0.86	0.09
novela sentimental	other	1	0.78	0.14	1	0.78	0.16
novela de costumbres	other	1	0.69	0.15	1	0.71	0.15
novela histórica	*novela sentimental*	1	0.90	0.08	1	0.90	0.09
novela histórica	*novela de costumbres*	1	0.88	0.10	1	0.87	0.12
novela sentimental	*novela de costumbres*	1	0.74	0.12	1	0.72	0.15

Table 39. Classification results for primary thematic subgenres (RF, 3,000 MFW, tf-idf).

for the character n-grams to 0.81 for normal MFW. All of these token units seem equally suited to classify the novels by thematic subgenres. As to the numbers of MFW, higher numbers of 1,000 or more tokens work better in all cases. Regarding the normalization techniques, tf-idf was, in general, the one that reached the top values, but the differences between the different maximum mean accuracies were small in many cases. Only the standard MFW features are chosen here to inspect the classification results for the different subgenre constellations, with 3,000 MFW and tf-idf. The results for this combination, using the RF classifier, are given in table 39 below.[602]

What is striking about the results for MFW is that the ranking of subgenre constellations is the same as for the topic features. The constellation that is easiest to distinguish is *novela histórica* versus *novela sentimental*, which reaches a mean accuracy of 0.90, the second best *novela histórica* versus *novela de costumbres* with 0.88, and the third best *novela histórica* versus "other" with 0.86. Again, the historical novel is the subgenre that can be separated best from the other thematic subgenres, followed by the sentimental novel. For the latter, a score of 0.78 is achieved if the subgenre is compared to all other novels and a score of 0.74 if it is contrasted with the novel of customs. As with the topics, also with MFW, the *novelas de costumbres* have the lowest mean accuracy of all constellations when they are classified against all other novels. In part, it is understandable that 3,000 MFW leads to similar results as topics because content words and, thereby, semantic features which can capture thematic elements or other semantic surface structures are also part of this feature set. Nevertheless, in the MFW set, not only nouns are included, but all kinds of word categories, and there is no layer of hidden semantic distributions that is analyzed, but the (normalized) word frequencies are directly used as features. So obtaining such similar classification results with different feature sets means that the results are meaningful regarding the characteristics of the different subgenres.

[602] The scores and standard deviations are rounded to two decimal places.

Another noticeable aspect of all the results for the classification of thematic subgenres is that the standard deviations were not reduced considerably from the first average results based on all feature sets and parameter constellations to the individual subgenre constellations. They were about 0.15 in the first overviews and are still at about 0.10 in the direct comparison of subgenres with a certain feature set. This means that approximately two-thirds of the variation are to due to different selections of novels for the various classification runs, which shows how important it is to have a corpus and subcorpora for the different subgenres as large as possible and how important it is to prevent the classifiers from focusing too much on specific novels. Furthermore, it means that in the case of the nineteenth-century Spanish America novels analyzed here, the thematic subgenres are not categories of very homogeneous texts but of texts that share textual features so that they are recognizable as instances of the text types, but they do so to different degrees. Furthermore, that the results for different classifiers, feature sets, and feature parameters are so similar indicates that the approximately 20 % of accuracy that is missing to reach the perfect classification is probably not due to a wrong choice of features or settings in the classification procedure but due to the discrepancies between the conventional genres and their text types. This could also be shown in the discussion of examples of historical novels that were frequently misclassified.

4.2.2.1.2 Literary Currents

The classification of the novels by their primary literary currents has several purposes. As for the thematic subgenres, the results of the classification are analyzed to see which classifiers and feature sets work best to capture the differences between the various literary currents. The best constellations are then chosen to see how well the different constellations of literary currents can be classified. A short analysis of the features that are decisive in the classification as well as overviews of how often individual novels were classified correctly or wrong is given for the MFW-based features in this case but not for the topics. Word-based features are considered more interesting for the literary currents because they are not primarily thematically defined, even if ranges of different topics are also characteristic for them. For the literary currents, style in a narrower sense is assumed to play a more important role.[603] Figure 73 shows the distribution of primary literary currents in the corpus.[604]

Of the 256 novels, there are 55 for which the primary literary current is unknown.[605] Romantic novels are most frequent with 116 instances, followed by 45 naturalistic, 35 realist, and 5 modernist novels. The following subgenre constellations are selected for the classification of primary literary currents:

---

[603] For an overview of the characteristics of the romantic, realist, and naturalistic novels, see chapter 2.3.2.

[604] This overview was generated with the function `plot_overview_literary_currents_primary()` in the script https://github.com/cligs/scripts-nh/blob/master/analysis/classification.py. The resulting chart is available at https://github.com/cligs/data-nh/blob/main/analysis/classification/literary-currents/visuals/overview-primary-currents-corp.png. Accessed October 29, 2020. For more extensive overviews of the novels by primary literary current, see chapter 4.1.5.3.2.

[605] A future task for the literary currents is to use the classification models to determine the labels of the novels for which no indication of literary current could be found, neither through explicit or implicit paratextual signals nor via literary-historical publications.

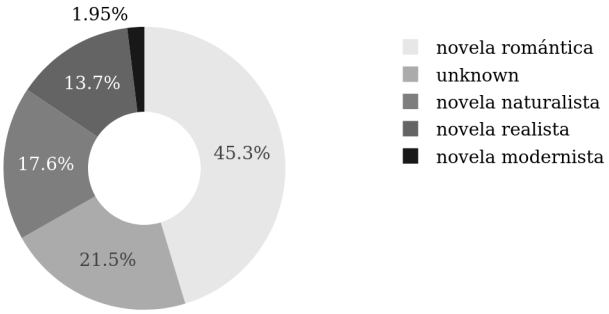


Figure 73. Primary literary currents in the corpus.

Classifier	Feature type	Top accuracy	Mean accuracy	SD accuracy	Top F1	Mean F1	SD F1
KNN	topics	1.0	0.83	0.15	1.0	0.83	0.15
SVM	**topics**	**1.0**	**0.85**	**0.14**	**1.0**	**0.85**	**0.15**
RF	topics	1.0	0.83	0.15	1.0	0.84	0.15

Table 40. Classification results for primary literary currents (topics).

- *novela romántica* versus other novels
- *novela realista* versus other novels
- *novela naturalista* versus other novels
- *novela romántica* versus *novela realista*
- *novela romántica* versus *novela naturalista*
- *novela realista* versus *novela naturalista*

The novels for which the literary current is unknown were not included in classifications because they could be instances of the positive class. As for the thematic subgenres, also here the results for topic features are presented first, and then the ones for MFW-based features. First, it is determined which of the three classifiers, KNN, SVM, and RF, worked best. The overall results for the classification by literary current with topic features are shown in table 40. The results are based on 144,000 classification runs for each classifier, including all parameter variations, repetitions of data selection, and cross-validation steps.[606]

SVM works best with a mean accuracy of 0.85, but KNN and also RF follow closely with 0.83 each. These mean accuracies are higher than the ones for the thematic subgenres, which were 0.77 and 0.80 for topic features. This was not expected for the topic features because especially the novels of the romantic current are subdivided into specific thematic subgenres. Apparently, despite the different conventional and textual thematic subgenres, the romantic novels all have topics in common when contrasted with novels of the two later realist and naturalistic currents. The possibility that the topics are not thematic in a narrow sense also has to be kept in mind because they can also represent other structures (descriptions, argumentation, motifs, etc.). That

[606] The scores and standard deviations in the table are rounded to two decimal places.

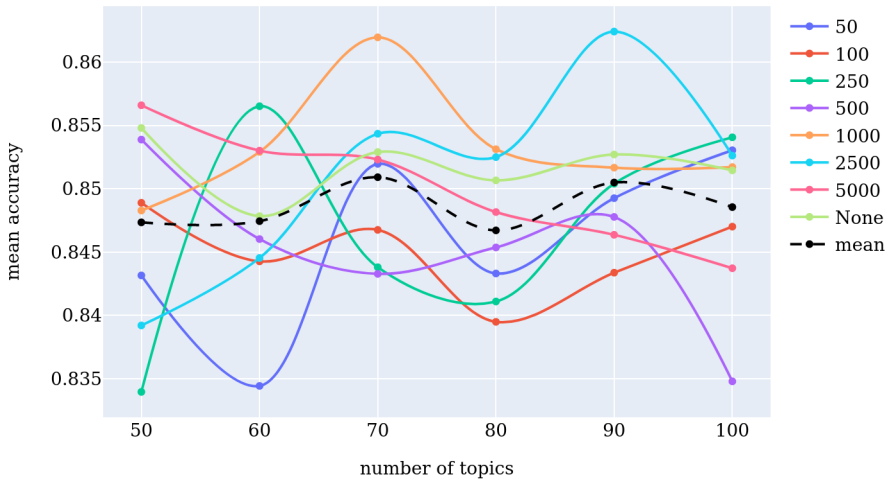


Figure 74. Classification results for topic feature sets (SVM, varying number of topics and optimization intervals).

so high results are already reached on average for all kinds of features and subgenre constellations demonstrates that the literary currents were not merely conventional movements but that the novels associated with them had textual and stylistic traits in common. The next step consists in evaluating which feature parameters worked best with the SVM classifier. An overview of the results for different numbers of topics and optimization intervals is given in figure 74.

The range of the mean accuracies goes from 0.83 to 0.86, so the differences are small between the different parameter constellations. The top mean accuracies are reached with 70 topics and an optimization interval of 1,000 and with 90 topics and an interval of 2,500.[607] Regarding the developments of different optimization intervals with an increasing number of topics, the curves for intervals of 50, 100, and 250 first fluctuate and then rise from 80 topics onwards. Intervals of 500 and 5,000 have downward trends, the curves of 1,000 and 2,500 mainly fluctuate but also fall with more topics, and the curve for no optimization at all is the most stable one. Here, the best combination of 90 topics and an optimization interval of 2,500 is chosen to analyze the classification results for literary currents further. That a constellation with less hyperparameter optimization works better for the literary currents than for the thematic subgenres is a sign that it is helpful if the topics are more evenly distributed across the corpus and in the individual novels, maybe because the literary currents are more general phenomena than the thematic subgenres.

With the decision for a certain combination of parameters for the topic feature set, the classification results for the different subgenre constellations can be inspected. The results for the literary currents with SVM, 70 topics, and an optimization interval of 1,000 are listed in table 41. These results are based on 500 classification runs for each constellation because of five topic modeling repetitions, ten random data selections, and 10-fold cross-validation.

The mean accuracies range from 0.77 to 0.92, so there are clear differences as to how well the different literary currents can be distinguished from each other. The best mean accuracy

[607] Slight differences in the plot are due to the small range of the values. Here the scores are rounded to two decimal points.

Sub-genre 1	Sub-genre 2	Top accuracy	Mean accuracy	SD accuracy	Top F1	Mean F1	SD F1
novela romántica	other	1	0.92	0.07	1	0.92	0.07
novela realista	other	1	0.80	0.16	1	0.80	0.17
novela naturalista	other	1	0.85	0.12	1	0.85	0.12
novela romántica	*novela realista*	1	0.90	0.10	1	0.90	0.12
novela romántica	*novela naturalista*	1	0.92	0.09	1	0.92	0.09
novela realista	*novela naturalista*	1	0.77	0.16	1	0.78	0.16

Table 41. Classification results for primary literary currents (SVM, 90 topics, optimization interval of 2,500).

of 0.92 is achieved for the classification of the *novela romántica* versus other novels and the *novela romántica* versus the *novela naturalista*. That the latter constellation yields the best results could be expected because the naturalistic novel can be considered a further development and specialization of the realist novel, carrying the realist aesthetic to the extreme so that it is also poetically more distant from the romantic novel. The best result for the *novela romántica* versus other novels with 0.92 and also the third best one for the *novela romántica* versus the *novela realista* with a mean accuracy of 0.91 was not expected for several reasons: because of the smooth transition of Romanticism into Realism in Spanish America, the existence of several novels that combine elements of both currents and the novels that included realistic elements before Realism (in particular the novels of customs). Furthermore, the romantic novel included several thematic subgenres, each of which is recognizable by its own topics, as was seen in the previous chapter. As it seems, the difference between romantic and realist novels is still big enough to lead to such good classification results. The classification of naturalistic novels against all others achieves a mean accuracy of 0.85, the realist novel versus the others 0.80, and the realist versus the naturalistic novels 0.77. That the score is lowest for the *novela realista* versus the *novela naturalista* could be expected because of the temporal overlap of both literary currents in nineteenth-century Spanish America but also because of the similar aesthetic concept of these two currents. The realist novel can be interpreted as the current that has a middle position between the romantic and the naturalistic novel. Therefore, it is not surprising that it has the lowest mean accuracy when it is contrasted with all the other romantic and naturalistic works. Next, the results achieved with MFW-based features and the different classifiers are analyzed for the literary currents. An overview of the results for the three main types of MFW features (words, word n-grams, and character n-grams) is given in table 42.

As with topics, so with MFW-based features, the SVM classifier achieved the best results with a mean accuracy of 0.86 for the basic MFW. The second-best mean accuracy is 0.84 for RF, and the worst is 0.78 for KNN. For all three classifiers, word n-gram features produced worse results than

Classifier	Feature type	Top accuracy	Mean accuracy	SD accuracy	Top F1	Mean F1	SD F1
KNN	MFW	1.0	0.77	0.14	1.0	0.76	0.15
	MFW word n-grams	1.0	0.69	0.16	1.0	0.71	0.17
	MFW character n-grams	**1.0**	**0.78**	**0.15**	**1.0**	**0.79**	**0.15**
	all	1.0	0.75	0.15	1.0	0.75	0.16
SVM	**MFW**	**1.0**	**0.86**	**0.12**	**1.0**	**0.86**	**0.13**
	MFW word n-grams	1.0	0.78	0.15	1.0	0.79	0.16
	MFW character n-grams	1.0	0.84	0.13	1.0	0.84	0.14
	all	1.0	0.83	0.13	1.0	0.83	0.14
RF	**MFW**	**1.0**	**0.84**	**0.13**	**1.0**	**0.84**	**0.14**
	MFW word n-grams	1.0	0.77	0.15	1.0	0.77	0.16
	MFW character n-grams	1.0	0.82	0.14	1.0	0.81	0.14
	all	1.0	0.81	0.14	1.0	0.81	0.15

Table 42. Classification results for primary literary currents (MFW).

single words. For KNN, the best results were achieved with the character n-grams, and for SVM and RF, with words. As the SVM classifier worked best, it was chosen to further evaluate the results. The mean accuracies for different values of MFW and the three normalization techniques are shown in figure 75.

For the basic MFW features, the highest mean accuracy of 0.88 is reached with 3,000 MFW and tf-idf values. The differences between tf, tf-idf, and z-scores are minimal. The mean accuracies for the other numbers of MFW from 1,000 up to 5,000 are also at 0.87 or 0.88 if the values are rounded to two decimal points. So the most important aspect is to use 1,000 MFW as a minimum, which was also the case for the thematic subgenres. As 3,000 MFW and tf-idf resulted in the highest mean accuracy, that combination of parameters was kept for the analysis of further results. Figure 76 visualizes the classification results for the word n-gram features.

The results for word n-grams are similar to the ones for thematic subgenres in that word 2-grams produce better results than word 3-grams or 4-grams. Here, too, the best mean accuracy of 0.87 is reached with 3,000 MFW for tf-idf and also for z-scores. As for word features, also for word n-grams, the results get much better with 1,000 MFW or more. It is of interest that the word 2-grams are almost as good as the word features which reached 0.88 because interpreting word 2-gram features might reveal further relevant aspects about the subgenres. Next, the classification results for normal character n-gram features are given in figure 77.

For the character n-grams, the highest mean accuracy of 0.86 is achieved with character 4-grams, 4,000 MFW, and tf-idf, so it is slightly worse than for word 2-grams or basic MFW. Other

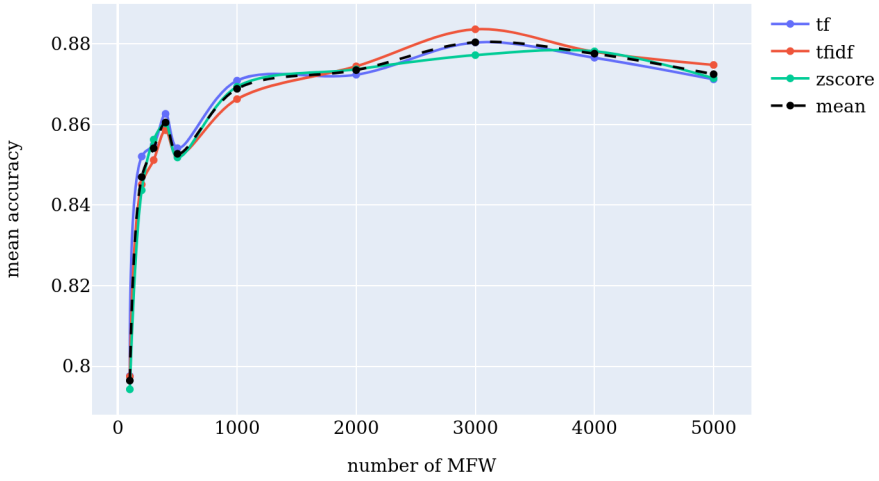


Figure 75. Classification results for MFW feature sets (SVM, varying number of MFW and normalization technique).

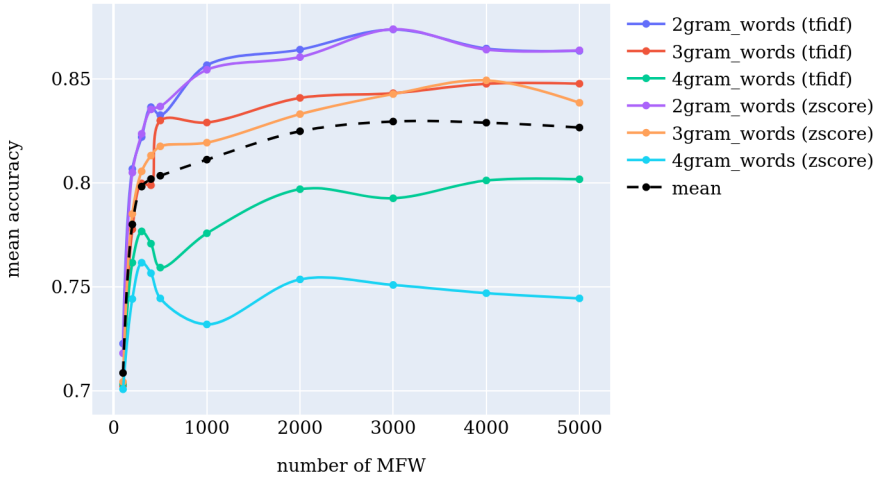


Figure 76. Classification results for word n-gram feature sets (SVM, varying number of MFW, grams, and normalization technique).

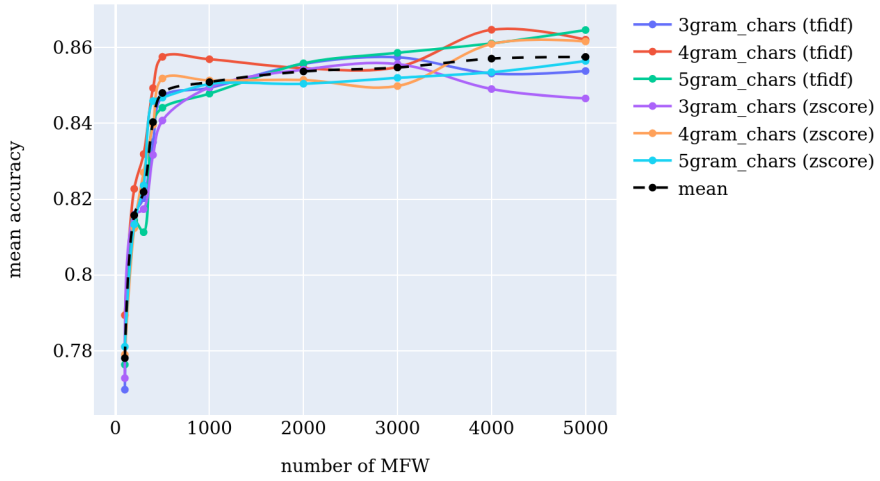


Figure 77. Classification results for classic character n-gram feature sets (SVM, varying number of MFW, grams, and normalization technique).

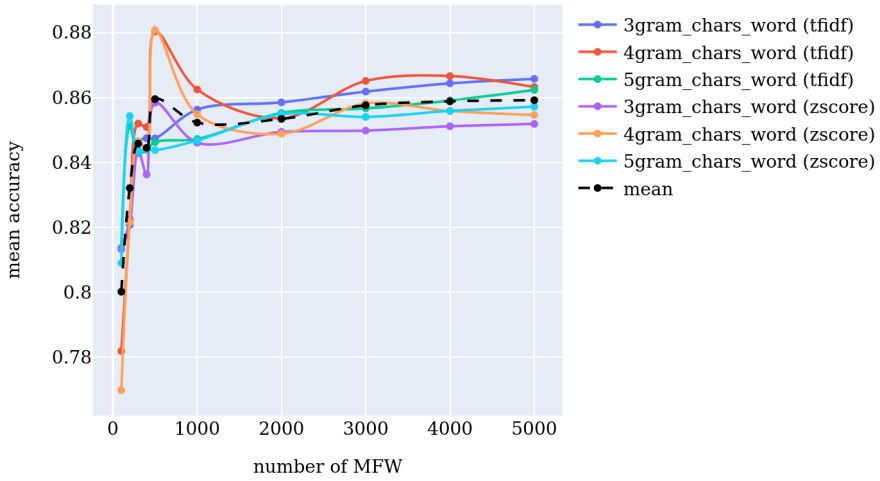


Figure 78. Classification results for "word" character n-gram feature sets (SVM, varying number of MFW, grams, and normalization technique).

n-gram units lead to very similar results, but interestingly, with more than 3,000 MFW, the scores drop a bit for character 3-grams but rise for the 4-grams and also the 5-grams. As more content words are included with higher numbers of MFW, maybe it is advantageous for the classification of subgenres when a bigger part of these words is captured with the larger character n-grams. In contrast to word units or word 2-grams, the results for character n-grams already stabilize largely with 500 MFW. The results for the special n-gram type "word", which only contains character n-grams from the inside of words and n-grams that include the end of one word and the beginning of another, are shown in figure 78.

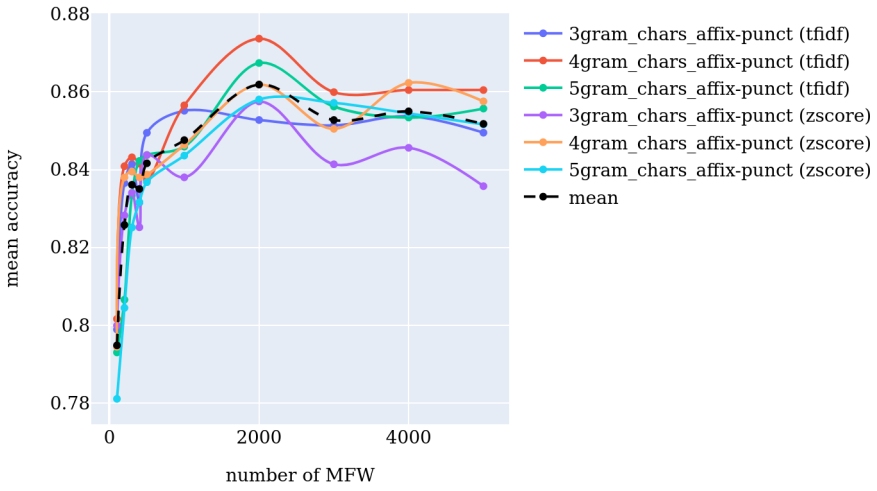


Figure 79. Classification results for "affix-punct" character n-gram feature sets (SVM, varying number of MFW, grams, and normalization technique).

Here the highest mean accuracy of 0.88 is reached with 500 MFW, character 4-grams, and tf-idf values or z-scores. This is higher than for the normal character n-grams, but this special character n-gram type does not outperform other feature types completely because it is equal to the highest score that was achieved with basic MFW. The other special n-gram type, "affix-punct", which only includes prefixes, (i.e., word beginnings), and punctuation marks, yields 0.87 as the best score. The same is achieved with character 4-grams, this time with 2,000 MFW and tf-idf values, as shown in figure 79. That 4-grams are the best token unit for character n-grams was also found out by Hettinger et al. (2015), so they seem to be generally useful for subgenre classification.

In table 43, the classification results for the different constellations of literary currents are given, based on the SVM classifier, 3,000 MFW, and tf-idf values.[608]

With MFW features, the subgenre constellation for which the highest mean accuracy was reached in the classification is the *novela romántica* versus the other novels with 0.93. The same constellation also had the highest value with topic features, but here it is even slightly higher. The second best result is 0.92 for the *novela romántica* versus the *novela realista*, followed by 0.90 for the *novela romántica* versus the *novela naturalista*. The contrast of the realist novel with all others yields a mean accuracy of 0.87, the realist novel against the naturalistic novel 0.86, and the worst result is 0.83 for the naturalistic novel versus the other novels. Compared to the topic features, here, all values are higher except the ones involving the naturalistic novel. This result is intriguing because it seems that the naturalistic novel is better captured with features related to themes than with word features. The change goes in the other direction for the realist novel. Here, the contrast between the *novela realista* versus the other novels is 7 % higher than with topic features, and the opposition of the *novela realista* and the *novela naturalista* is 9 % higher here than with topic features. The *novela romántica* versus the *novela realista* is only 1 % better with MFW than with topic features, though. So the realist novel wins if word features are used

[608] The scores are rounded to two decimal points.

Subgenre 1	Subgenre 2	Top accuracy	Mean accuracy	SD accuracy	Top F1	Mean F1	SD F1
novela romántica	other	1	0.93	0.06	1	0.93	0.06
novela realista	other	1	0.87	0.12	1	0.87	0.12
novela naturalista	other	1	0.83	0.12	1	0.82	0.13
novela romántica	*novela realista*	1	0.91	0.09	1	0.90	0.10
novela romántica	*novela naturalista*	1	0.90	0.10	1	0.90	0.11
novela realista	*novela naturalista*	1	0.86	0.14	1	0.87	0.13

Table 43. Classification results for primary literary currents (SVM, 3,000 MFW, tf-idf).

instead of topics. The result is so interesting because the label "novela naturalista" is used in subtitles of several novels in the corpus. In addition, its status as a (thematic) subgenre versus a literary current or movement has been discussed, for example, by Schlickers, who confirms that the naturalistic novel was a concept that was consciously applied and communicated by contemporary writers in the nineteenth century.[609]

To get an insight into the kind of MFW features that are relevant for the classification by literary current, the two contrasts of the *novela romántica* versus the *novela realista* and the *novela realista* versus the *novela naturalista* are analyzed here. Figure 80 shows the top feature weights for the romantic versus the realist novels.

The features that are distinctive for the realist novels are on the left side and the ones for the romantic novels on the right side. Among the top 25 features, there are 17 which are important for the realist novel and only 8 for the romantic novel, which is a sign that the realist novel is stylistically more homogeneous than the romantic novel. Among the top words for realist novels, there are several adjectives, of which the most important one is "rojo". The others are "duro", "gran", "duros", "excelente", and "listo". In the list of the romantic novels, there is no adjective. A more frequent use of adjectives could be due to more descriptive passages in the realist novels. Nouns, too, are only present in the list for realist novels: "corriente", "grupo", "política", "envidia", "hoja", "importancia". So politics is an important topic in realist novels. Groups of people are also relevant, and "envy" points to their materialist orientation. The other three nouns are not so easily interpreted at first sight. Furthermore, there are some verbs in the past tense: "mostraba", "habló", and "escribió". The last one, together with "hoja", seems to indicate that people frequently write things down in realist novels. The words that are distinctive for the romantic novel include the first person plural verb form "hemos", the participle "quedado", the adverbs "verdaderamente"

[609] In the corpus, the novels "¿Inocentes o culpables?" (1884, AR) by Juan Antonio Argerich and "Los bandidos de Río Frío" (1892, MX) by Manuel Payno carry the label "novela naturalista" in their subtitles. About the discussion to treat the naturalistic novel as a subgenre or a movement, see Schlickers (2003, 16).

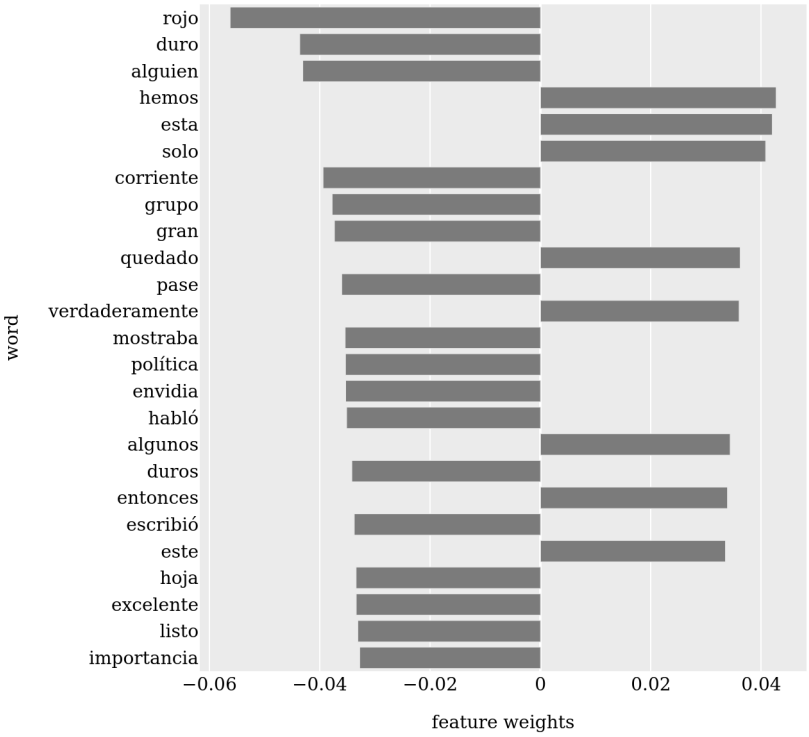


Figure 80. Feature weights (MFW) for realist versus romantic novels.

and “entonces”, the demonstrative pronouns “esta” and “este”, and the adjective “solo”. The fact that the word “lonely” is typical for the romantic novel fits well with the concept of the individual romantic hero or heroine. Furthermore, it is noticeable that the words with the top weights for the two different literary currents belong to different grammatical categories. The feature importances of the naturalistic versus realist novel are shown in figure 81.

Here, the words that are distinctive for the naturalistic novels are on the left side, and the ones that are typical for the realist novels are on the right. Again, the proportion of top features is 8:17, with more features for the realist novels. The nouns that are typical for the naturalistic novels are “mitad” and “pasajeros”. The first one is an indication of quantity. The passengers could either be connected to transport in a city or to travel. The word “costado” is here probably the participle “cost” and not the noun “side” and alludes to the importance of money in naturalistic novels. The adverbs “hasta” and “recién” indicate that time also matters in novels of this literary current. The verb “convencer” is the only infinitive in the whole list. To convince someone means the need to discuss and the will to achieve a certain goal. The words “cual” and “fuesen” are not easily interpreted. As to the words that are distinctive for the realist novels, there are some which did already appear in contrast with the romantic novels. Again, there are several verbs in the past tense: “traía”, “dieron”, and “oyeron”. The nouns “figura”, “gracia”, and “voces” are new. The verbs in past tense could mean that narrative passages in which the actions of characters are described are more frequent in realist novels than in romantic and naturalistic ones. Altogether,

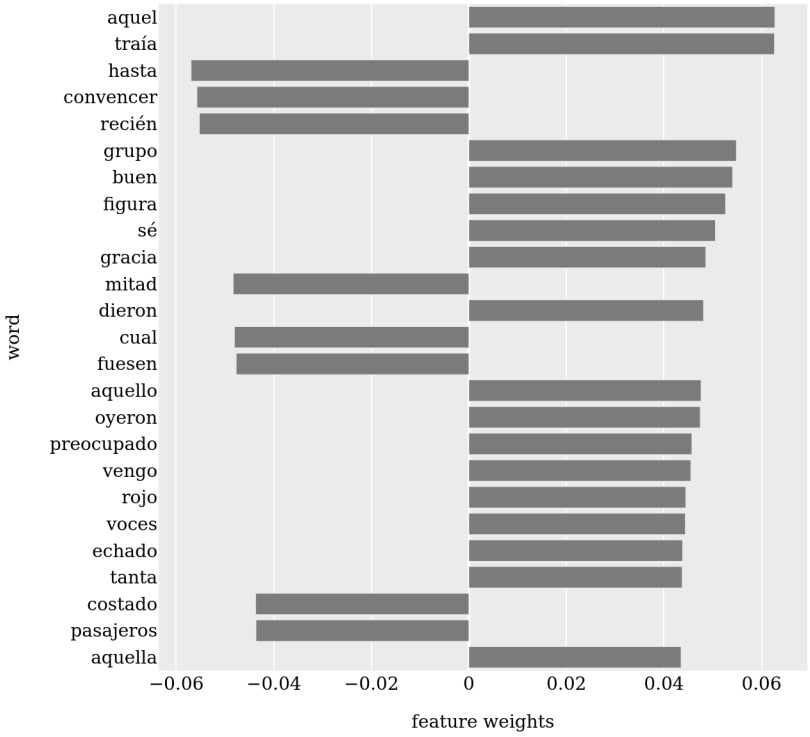


Figure 81. Feature weights (MFW) for naturalistic versus realist novels.

the MFW features that are weighted highly by the SVM classifier can be interpreted well as stylistic cues of the literary currents in the same way as the topic features could be explained for the thematic subgenres. To conclude the discussion of the classification results for literary currents, the *classification profiles* with the proportions of true positives, false positives, and false negatives are visualized in figures 82 to 84. On the one hand, they serve to see how many novels were always classified correctly and are part of the conventional as well as the textual genres. On the other hand, it is checked whether there are any cases of novels that are misclassified very frequently and are only part of either the text type or the conventional genre.

What is striking is that there are no cases of novels that are false positives or false negatives in more than 20 % of the cases for none of the literary currents. This was very different for the thematic subgenres. It means that there are more considerable discrepancies between the conventional genres and the text types for the thematic subgenres than for the literary currents. A hypothesis is that the literary currents that the novels belong to are mostly determined by literary historians. Therefore, they could be closer to the textual characteristics of the texts than, for example, explicit historical labels of thematic subgenres that might be misleading. Another hypothesis is that the style of novels that are part of a certain literary current is easier to recognize than the primary theme of a novel. At least in some cases of the novels, there are diverging opinions also among literary historians about which primary topic they have. Furthermore, the primary themes of the novels are not necessarily congruent with the surface topics, which can

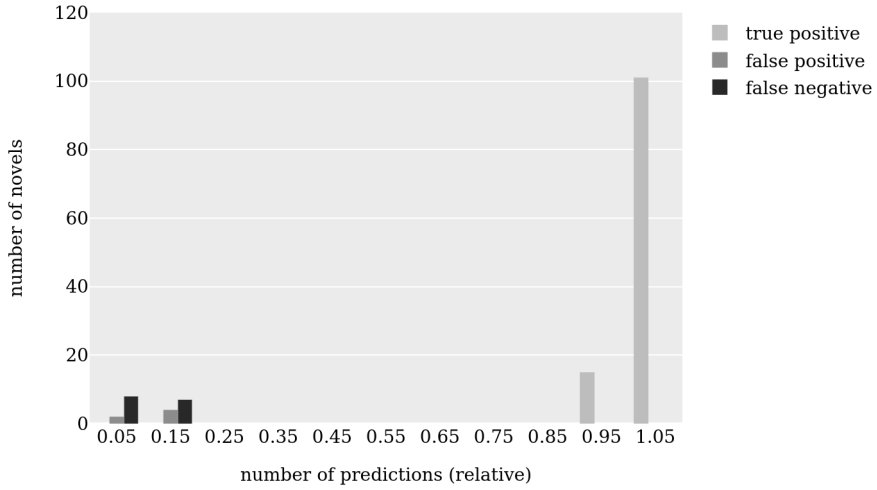


Figure 82. Predictions for *novela romántica* versus other novels (MFW).

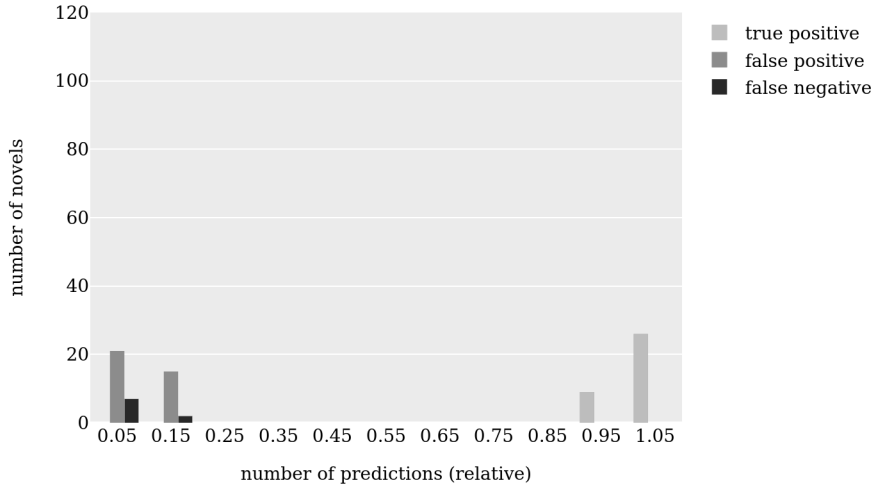


Figure 83. Predictions for *novela realista* versus other novels (MFW).

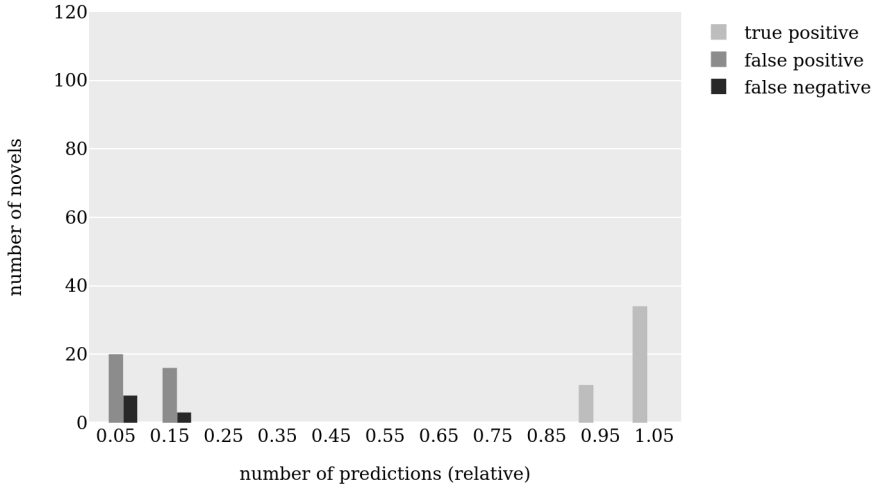


Figure 84. Predictions for *novela naturalista* versus other novels (MFW).

lead to misclassifications. All in all, the classification results for the literary currents showed that these are indeed categories with a stylistic unity.

4.2.2.2 Family Resemblance: Network Analysis

In the previous chapter, statistical classification was used to categorize the novels by subgenre. Overall, high accuracy values of 70 % and more could be achieved when the classifiers tried to recognize which texts match which subgenre labels, so the congruence of the conventional genres and text types is relatively high. Furthermore, the analysis of misclassified examples showed that there are individual cases of less prototypical works or special cases in which the conventional label of a novel does not correspond so well to its textual and stylistic characteristics. As classification *aims* to match labels and texts, it might cover internal differentiations of the text types. The limits between the different classes are strictly drawn, and the connections that can exist between individual novels or groups of novels that are part of different conventional subgenres are cut off. Therefore, a method relying on network analysis is proposed here as a "family resemblance analysis" with the aim of providing a more open means of categorization. It is very likely that the groups of novels found by this approach are influenced by a number of different factors that determine the style of the texts, such as authorship, period, or narrative perspective. This is not viewed as a principal drawback because such factors are the ones that contribute to the organic whole of the subgenres' style if the subgenres are understood as groups of historical texts set in a specific geographical, cultural, and temporal context and as texts written by individual authors who shape the subgenres through their works. They are only disturbing if they dominate the categorization completely so that the groupings are no longer about subgenres as the target of the stylistic analysis but about some other factor.

In the previous chapter, it was shown that even the results of classification could be interpreted in terms of prototypical structures and family resemblance. However, in classification, the

similarity between different texts is still determined based on whole feature distributions and not of overlapping parts of it. Even if it is enough for a historical novel to have high weights for some topics that are distinctive for historical novels in order to be classified as such, it is not possible in standard classification that a novel is grouped with novels of one subgenre based on some topics and categorized with novels of another subgenre based on other topics that it has – a decision for one class is always made. Even with multilabel classification, an approach where each sample can be recognized as an instance of several different classes, the assignments are made based on internal comparisons of two classes, and it is decided in each case if the sample is a member of the positive or the negative class. In the network-based family resemblance analysis, in contrast, links between individual novels are decisive. If a group of several novels turns out to have particularly strong links, they are considered members of one category. However, each member still also has links with other novels outside of the "nuclear family". Furthermore, the similarities are determined for each pair of novels and not for one novel against all others in the group so that partial similarities, understood as individual resemblances, can hold the group together.

Regarding the analysis of the subgenres, the family resemblance analysis has several functions. It can serve to find out how a subgenre is organized internally: by looking at the network of similarities between the individual novels, do subgroups, i.e. individual *families*, emerge, and which traits hold them together? Can the prospectively historical novels, for example, be distinguished from the other novels belonging to that subgenre, or the romantic historical novels from the realist and modernist ones? Or are there differences by country or over time? As can be seen, no preliminary assumption is made here as to the kind of connections that the family resemblance analysis might reveal for the internal structure of subgenres. There can be diachronic shifts but also synchronic variations in the subgenres. Another possibility is to analyze novels of several subgenres or the whole corpus of novels together to see how they are connected stylistically when no strict boundaries are applied. That way, it can be tested whether pure or mixed types of subgenres become visible as *families*. It is useful to test the feature sets with classification beforehand to make it probable that the textual features used are relevant for the distinction of subgenres at all. Here, topics are used as features, for which it was shown in the previous chapter that they work well to classify nineteenth-century Spanish-American novels by subgenre. In particular, the historical novel and the sentimental novel are examined, with a focus on the historical novel, to analyze the internal substructure of these subgenres. In a third network, historical and sentimental novels are compared. In the following, first, the method for the creation of the family resemblance network is outlined. Then the subcorpus of novels that was used for the family resemblance analyses is presented, and the resulting networks are discussed.

4.2.2.2.1 Method

To create the network for the family resemblance analysis, first, the similarities between all the individual novels were calculated for the chosen feature set using cosine similarity.[610] After that,

[610] Cosine similarity measures the cosine of the angle between two text vectors. See https://en.wikipedia.org/wiki/Cosine_similarity. Accessed April 28, 2020.

the resulting textual similarities were mapped onto a network structure. The novels themselves constitute the nodes in the network. The network relationships (or edges) were determined using the three nearest neighbors of each text, which were selected from a ranking of the text similarities. The strength (or weight) of the edges was calculated by summing up the similarity values of the neighbors.[611]

In the overall similarity matrix, there are relationships between all the texts. Reducing the number of connections for the network to the three nearest neighbors makes the network less complex and the closest relationships in it more salient. This, in turn, enhances the interpretability of the network. The choice of three is arbitrary and could be varied. However, using more than one nearest neighbor makes the results of the network more stable, as Eder has shown (Eder 2017, 56–60). Eder introduced the idea of visualizing nearest neighborships based on textual similarities in a network structure, intending to make the results of stylometric cluster analysis more reliable. This technique is adapted here with a different aim: to formalize the family resemblance concept for genre analysis.

In addition to the creation of the basic network structure, community detection was used to explore the *families* of novels in the network. Communities are sets of nodes in a network that are more densely connected to each other than to nodes outside.[612] Different algorithms for the detection of network communities exist.[613] Here, the Louvain modularity algorithm was used. It is based on modularity optimization, which means that possible divisions of the network are checked and optimized to reach high modularity so that the nodes which share a community are more likely to be connected with each other than with other nodes that are not community members. The Louvain algorithm optimizes local communities iteratively until the best global modularity of the network is reached. Here, non-overlapping communities are built, meaning that each node is only part of one community in the result. As it looks for densely connected data points, the Louvain algorithm is suitable to detect *families* of novels. To that end, the novels are represented as vectors in a space of textual features, in which they can be closer to each other or further away, and these similarities (or distances) are interpreted in terms of network links. Furthermore, the algorithm is also comparatively efficient and has been implemented in Python, which is used to create and visualize the network here.[614]

Reflecting on how the concept of family resemblance is formulated by Wittgenstein and in literary genre theory and how it is implemented here, the following observations can be made. Using similarity relationships between novels based on feature distributions means that not the presence or absence of a trait or a set of traits determines the connection between members of a family and the difference to other families, but the numerical strengths of the features in combination and this only between pairs of individual novels.[615] The similarities of the various

---

[611] Because the closest neighborship depends on the perspective, it was calculated for each node. If two nodes are mutually closest, the strength of the edge increases.

[612] For general information about community structures in networks, see https://en.wikipedia.org/wiki/Community_structure. Accessed April 28, 2020.

[613] For a review of different community detection algorithms for networks, see Javed et al. (2018).

[614] See https://github.com/taynaud/python-louvain for the Python module implementing the Louvain community detection. Accessed April 28, 2020. The algorithm itself is presented in Blondel et al. (2008).

[615] Of course, zero values are also possible in the feature matrices and could be interpreted as *absent*, but it would not be proportionate to consider all values that are greater than zero as *present*. A possibility to model the features in a

pairs are summed up and interpreted in terms of neighbor rankings to build the families. This transfers the idea of partial and overlapping similarities to a quantitative approach because nearest neighborship between novel A and another novel B, on the one hand, and novel B and C, on the other, does not mean that A and C must also be closest in a direct comparison.

Second, when communities are calculated and interpreted as families, the boundaries of the categories are retroactively sharpened because communities are clusters, which have members and non-members. This is an advantage that balances out the looseness of the original family resemblance concept. However, there is a significant difference between these communities and classes in a logical sense because the former emerge from a network of similarities and not from the condition of shared common features. The communities mark a boundary between one group of dense relationships and another, they cut off the family at a certain point but they do not lever out the basic idea of family resemblance, because the links to other families and family members in the network can still be explored.[616]

4.2.2.2.2 Data

The subcorpus of novels used for the family resemblance analysis includes 83 novels first published between 1840 and 1910. Of these, 40 are historical novels, and 43 are sentimental novels. 32 of the novels are Argentinian, 35 are Mexican, and 16 are Cuban. The novels were written by 74 different authors, 11 of them female and 72 male. To prevent the authorial signal from interfering too much with the genre signal, for each subgenre, only one novel per author was chosen. Nevertheless, if authors wrote novels in both subgenres, these are included. There are nine authors with both a historical and a sentimental novel in the subcorpus. Figure 85 shows the distribution of the novels by decade and subgenre.[617]

As features, a topic model with 100 topics and an optimization interval of 100 was used, which is considered a medium degree of specification, given that the overall corpus contains 256 novels. The feature set was generated for the whole corpus Conha19 with the goal of having more stable topics that represent the novel of the time in a better way than if they had been based on the smaller subcorpus. For the network analysis, only the features for the novels in the subcorpus are used.[618]

---

different way would be to define a threshold value and convert all values below it to zero and all values above it to 1 to get a binary distinction. Good reasons would have to be given for the value at which to set the threshold.

[616] From the many decisions taken so far to formalize the family resemblance concept, it becomes clear that variants of this approach are possible. For example, the similarity measure used, the number of nearest neighbors considered, the way to determine the strength of the edges, and the kind of community detection algorithm could be varied. As is generally the case with feature-based categorization, also the selection of the features and their modeling and parametrization are subject to choice. Further empirical studies and serial analyses are needed to test the effects that such variation has on the results.

[617] The metadata of the subcorpus used for this analysis is available as "metadata.csv" in the folder "corpus_metadata" at https://github.com/cligs/data-nh/tree/main/analysis/family-resemblance/. All other data related to this family resemblance analysis, including results and figures, can be found in the same GitHub folder. The Python scripts used are available at https://github.com/cligs/scripts-nh/tree/master/analysis/family_resemblance. Accessed January 8, 2021.

[618] This topic model was created separately from the ones used for the classification in the previous chapter because the work on the family resemblance network was done earlier. The topic model used here was built with the tool MALLET (McCallum 2002) and pre- and post-processed with tmw (Schöch and Schlör 2017). The texts were

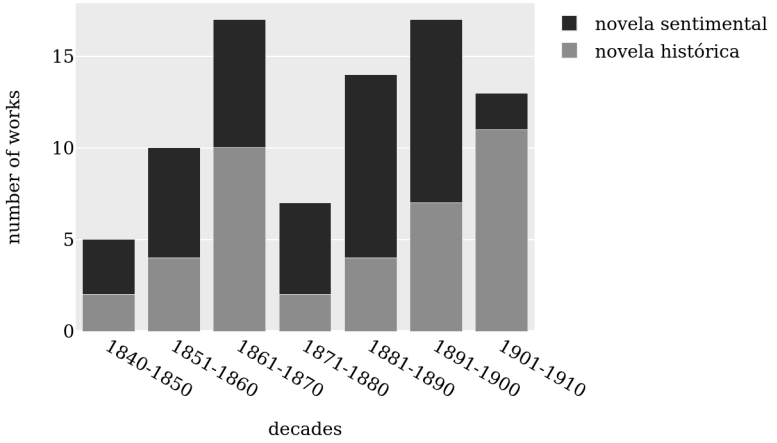


Figure 85. Subcorpus for the family resemblance analysis.

Figure 86. Examples of topics for the family resemblance analysis.

Shortcut	Subgenre(s)	Number of novels	Number of clusters (*families*)
HIST	historical novels	40	6
SENT	sentimental novels	43	6
HIST-SENT	historical + sentimental novels	83	8

Table 44. Overview of the family resemblance networks produced.

In figure 86, the top 40 words of four of the resulting 100 topics are visualized. They exemplify the range of themes covered in the novels. The first topic is about love and feelings ("amor", "corazón", "alma", "amor", "pasión"); the second topic is dominated by politics ("gobierno", "ministro", "guerra", "poder"), the third one is about crime and banditry ("bandido", "jefe", "ladrón", "robo"), and the fourth one about religion and colonization ("sacerdote", "dios", "español", "guerrero"). The first number in parentheses indicates the rank of the topic by its probability in the whole corpus, so the topics are of different importance for the whole collection of texts. The lower the topic rank, the more important the topic is, so the love topic is a very general one, the politics and crime topics are still rather common, and the colonization topic is more special.

4.2.2.2.3 Results

With the approach outlined in the previous section, three kinds of networks were produced, two for the individual subgenres and one for the two subgenres combined, as shown in table 44.[619]

The last column indicates how many *families*, that is, clusters based on the communities in the network, were produced. The number of clusters is identical for both historical and sentimental novels when they are analyzed separately. Given that the number of novels doubles when the two subgenres are combined, the number of resulting clusters does not grow proportionally, indicating that there is an overlap between the subgenres. The discussion of the results focuses on the historical novels.[620] Figure 87 shows the first network for historical novels and topics (HIST-topics). The communities detected are indicated by the different colors of the nodes.

An important question for the interpretation of the network is which kinds of novels constitute the different *families*. Before looking at different clusters in detail, an overview of the cluster sizes was generated, and the possible influence of some text-external and -internal factors on the clusters was calculated, as displayed in figures 88 and 89.

---

lemmatized with the TreeTagger (Schmid 1994), using the Spanish parameter file, and only nouns were kept. In addition, a list of stop words was prepared based on the 50 most frequent nouns and adapted manually. To this, some more stop words were added after inspecting the results of the topic model (e.g., proper names or very general nouns). Before running the topic modeling, the texts were first lemmatized and then segmented into chunks with a length of 1,000 tokens. Besides the number of topics, the topic model was created with 5,000 iterations. The feature matrices, both for the full and the reduced corpus, can be viewed on GitHub (see the previous footnote).

[619] The script calling the various functions of the network analysis for the different setups is available at https://github.com/hennyu/papers/blob/master/family_resemblance_dsrom19/analysis/run_scripts.py. Accessed May 6, 2020.

[620] The overall results can be inspected on GitHub, though.

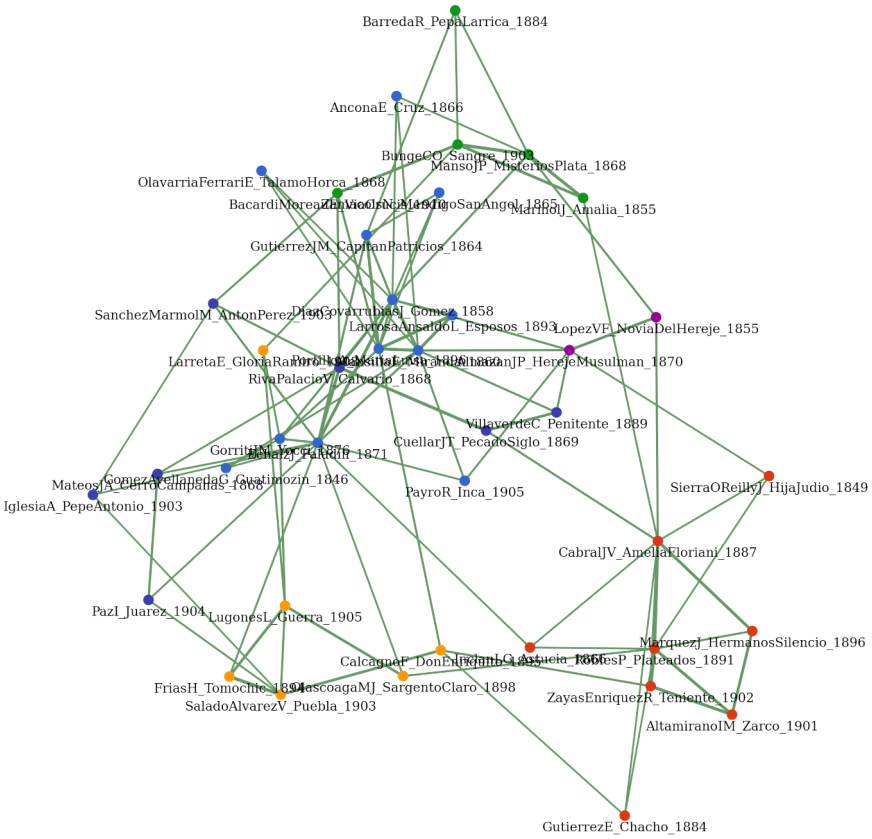


Figure 87. Network of historical novels based on topics (HIST).

Four of the resulting six clusters are evenly sized, with 8 novels each, and the other two are smaller. Cuban novels are only contained in clusters 1, 2, 3, and 5. Clusters 1, 4, and 5 are dominated by Mexican novels and clusters 2 and 3 by Argentine novels. Cluster 3 is an Argentine-Cuban cluster, and cluster 5 is a Mexican-Cuban cluster. Even if there are some tendencies regarding the distribution of novels by country in the different clusters, there is no cluster consisting only of novels from one country, and it should also be kept in mind that the overall number of novels in the individual clusters is quite small. The narrative perspective is not significant for the historical novels because there is only one novel with a homodiegetic narrator, the others all have a heterodiegetic narrator. The five historical novels written by female authors are distributed over the three clusters 2, 3, and 4, so there is no clear female cluster. Regarding the distribution of the novels over the years, there is also much overlap, as all the clusters have earlier and later novels. Apart from one outlier, cluster 2 is rather late, and cluster 4 is mostly filled with earlier works.

When one looks at one cluster in detail, it is possible to retrace the family resemblance relationships. In table 45, the novels contained in cluster 3 are listed together with their nearest neighbors (N-1, N-2, N-3), including the weight of the edge to the respective neighbor. The

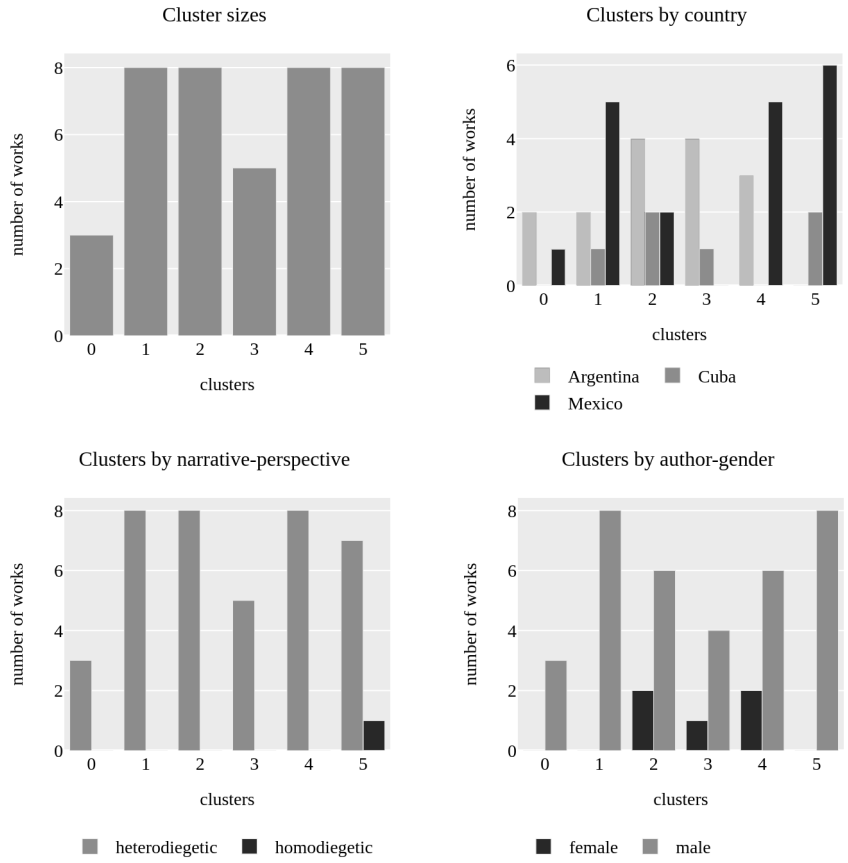


Figure 88. Overview of cluster metadata in the network HIST.

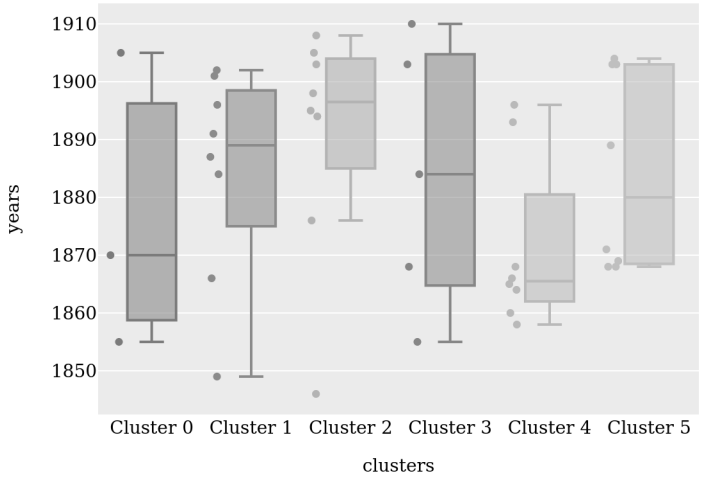


Figure 89. Clusters by year in the network HIST.

Idno	Author	Title	N-1		N-2		N-3	
nh0017	Mármol	Amalia	Misterios	1.4	Sangre	1.2	Cl 1	0.3
nh0081	Bunge	La novela de la sangre	Misterios	1.5	Crucis	1.2	Amalia	1.2
nh0094	Manso	Los misterios del plata	Sangre	1.5	Amalia	1.4	Cl 4	0.6
nh0160	Barreda	Pepa Larrica	Cl 4	0.4	Misterios	0.4	Sangre	0.4
nh0166	Bacardí-Moreau	Vía Crucis	Sangre	1.2	Cl 4	0.6	Cl 5	0.6

Table 45. Nearest neighbors in cluster 3 of the network HIST.

strongest relationship exists between "La novela de la sangre" (1903, AR) by Carlos Octavio Bunge and "Los misterios del Plata" (1868, AR) by Juana Manso de Noronha because they are mutually closest to each other. Other bilateral nearest neighborships between novels in the cluster are highlighted in lighter orange. The novel "Pepa Larrica" (1884, AR) by Rafael Barreda has two nearest neighbors in the cluster, but the relationships are only unilateral. Boxes that are not highlighted show which nearest neighbors are outside of the current cluster. It becomes clear that some novels are central members of the *family* while others are rather distant relatives.

The topic distributions for the five novels are visualized in figure 90 to see what topics are decisive for the relationships in this cluster of historical novels. The axis on the top shows the absolute value that the topic achieved in each novel, and the axis to the left shows the individual 100 topics.[621] In addition to the lines for the five novels in the cluster, a black dashed line indicating the mean topic values for all the historical novels in the network is added. The topics are ordered by importance in the whole corpus of 256 novels from top to bottom so that more general topics are at the top and more special topics are further down. Some topics of interest are labeled, the black ones being particularly important for this cluster and the red ones less important when compared to all the historical novels in the corpus. What makes the family approach visible is that not all the decisive topics are equally relevant for the individual novels in the cluster. For example, the topics "sacerdote-dios-español" and "fortaleza-batería-plaza" are underrepresented in the whole cluster, but "amor-corazón-alma" and "soldado-fuego-columna" are only partly less relevant. The first corresponds to the mean for the novel "Amalia" and the second reaches almost the mean for "Vía Crucis". Topics that are overrepresented in several novels in the cluster are "voz-palabra-brazo", "idea-espíritu-instante", "pueblo-ley-país", "calle-puerta-voz", "agua-cuerpo-sangre", "gobierno-ministro-guerra", "puerta-espíritu-cabeza", and "cabeza-rosa-asesino". They stand for the general characteristics of the family: historical novels that are not so much mixed with love stories, not focused on military actions and not about the Conquest or colonial history, but about political ideas and conditions, and about (inter)personal contacts and states, about voices, words, and bodies. However, as specific topic values are not necessary conditions, some of the novels have their own special topics. The topic "mar-buque-puerto" is specific for "Los

[621] In a strict sense, the topics are categories and not numerical values and should be visualized as bars rather than lines. The line plot was chosen here because it facilitates seeing the differences between the data series.

misterios del Plata", "negro-esclavo-amo" for "Vía Crucis", the only Cuban novel in this cluster, and "capitán-voz-revolución" for "Pepa Larrica".

A more general overview of the topics that are distinctive for the different clusters in the network of historical novels is given in figure 91. In the heatmap, the yellower the boxes, the more important the topics are for the cluster, and the bluer, the lesser important they are. The distinctiveness was calculated by normalizing the topic values to z-scores. Here, only the top 30 most distinctive topics are shown. The values in parentheses at the end of the topic labels indicate the ranks of the topics in the whole corpus, so the topic "voz-palabra-brazo", for example, is much more general than "fortaleza-batería-plaza".

The distinctive topics of cluster 3 that were already discussed can be recognized in the heatmap. The smallest cluster 0 seems to be about the Conquest and colonial history, as the most distinctive topics are concerned with Indians and Spaniards ("indio-español-tierra"), rural church ("cura-fraile-pueblo"), and seafaring ("mar-buque-puerto"). In Cluster 1, topics about military campaigns and rural life prevail, making one think about internal struggle, bandits, and gauchos. The topics that are distinctive for this cluster are about conversation and social roles of bandits ("palabra-asunto-razón", "bandido-jefe-razón", "bandido-jefe-ladrón"), provincial police forces ("caballero-comisario-provincia"), horses ("caballo-amo-instante"), soldiers and military actions ("manera-soldado-muerte", "ejército-prisionero-jefe"), and ranches ("hacienda-compadre-pueblo"). Cluster 2 is not so easy to interpret. It is also about military action ("soldado-fuego-columna", "sargento-cerro-gruta", and "ejército-guerra-ciudad"), but there are other, individual topics. Cluster 4 is clearly romantic, as it contains topics about love ("corazón-alma-lágrima", "amor-corazón-alma"), about the missionary work, colonial administration and aristocracy ("sacerdote-dios-español", "alcalde-dama-barón"), and illness ("instante-doctor-sitio"). This fits well with the observation that it contains mostly earlier novels. The last cluster is politico-historical with top topics about the military ("soldado-jefe-coronel", "fortaleza-batería-plaza"), government ("gobierno-ministro-guerra"), and the time of the French intervention in Mexico ("francés-emperador-estudiante").

The results for the network of sentimental novels and the one containing both types of subgenres are only summarized briefly here. Regarding the metadata, the cluster sizes vary more for the sentimental novels.The biggest cluster has 14 novels, and the smallest one has only two. All the clusters are mixed by country. Among the sentimental novels, there are more with an autodiegetic and a homodiegetic narrator, and the narrative perspective has an influence on the results. The smallest cluster, for instance, consists solely of autodiegetic texts featuring topics related to inner life and landscape. For the sentimental novels, there are also clearer tendencies of topic changes over time, as figure 92 shows. The early cluster 0 is romantic with letters, dance, aristocracy, and much emotionality. The three later clusters are the ones dominated by interiorization, and the mid-century cluster 5 is worldly about food, marriage, business, and money.

When both subgenres are analyzed together, they are not neatly sorted into different *families*. As can be seen in figure 93, there are clusters dominated by one subgenre – clusters 1, 3, and 6 by sentimental novels and clusters 2, 4, and 7 by historical novels – but there is no cluster containing only novels of one subgenre and the clusters 0 and 5 are entirely mixed. In the combined network, the cluster sizes vary moderately from 7 to 13 novels. Here, too, there is no clear tendency for countries. Different narrative perspectives are not concentrated in single clusters, so this aspect

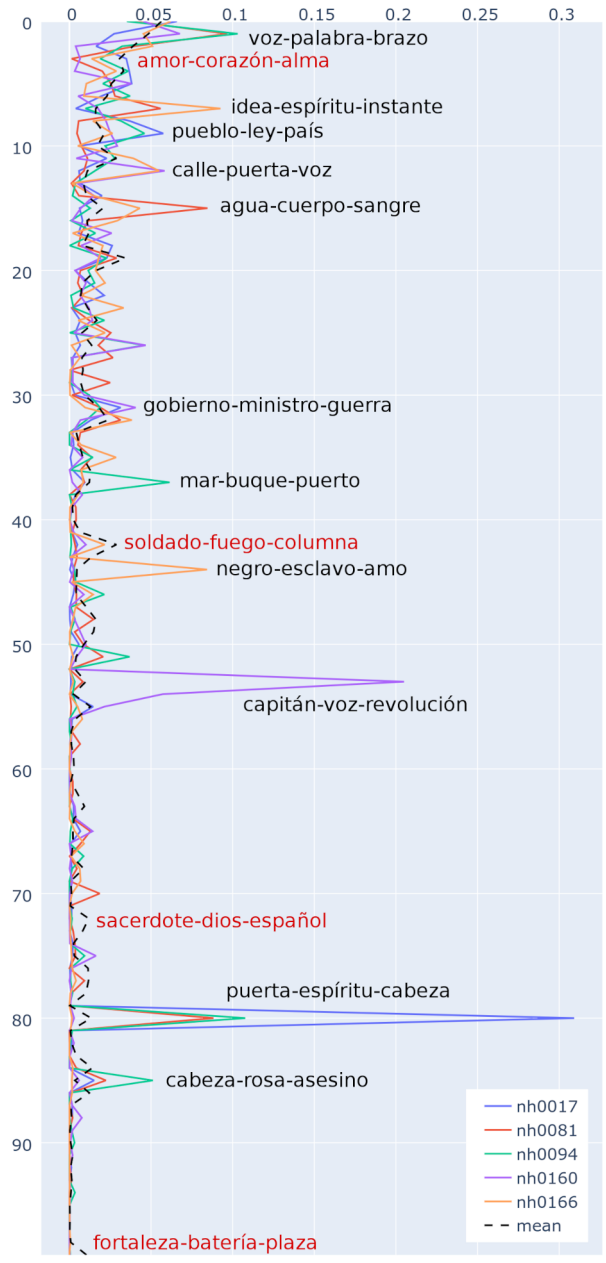


Figure 90. Topic scores for cluster 3 in the network HIST.

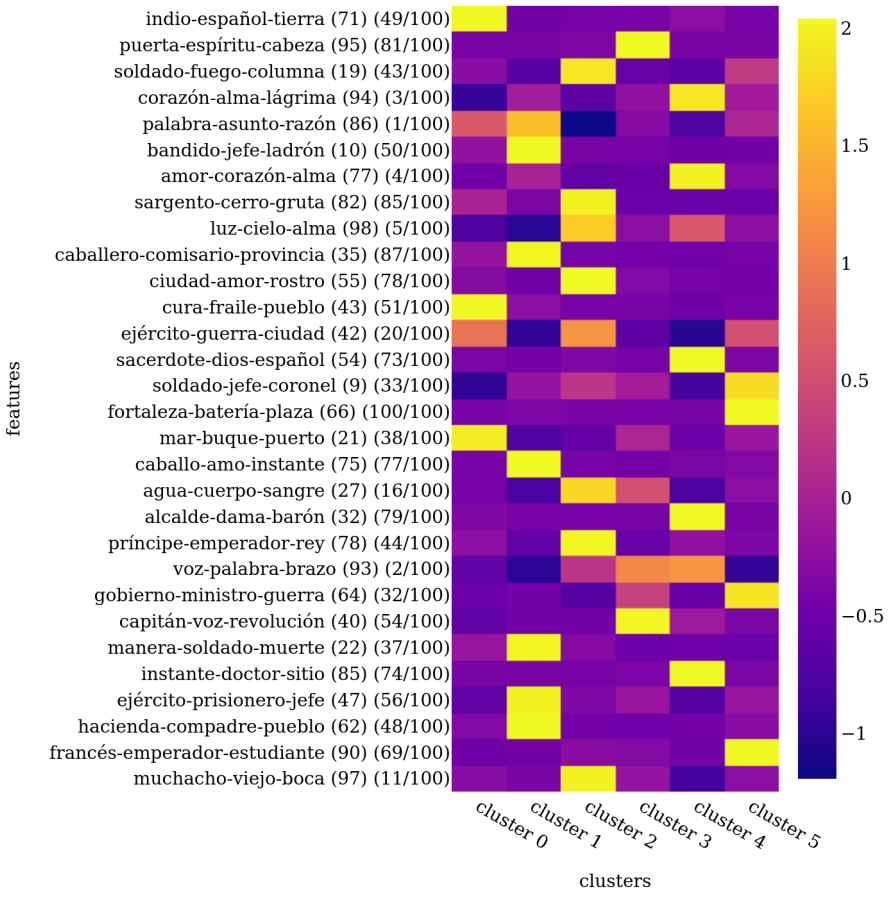


Figure 91. Top distinctive topics in the clusters of the network HIST.

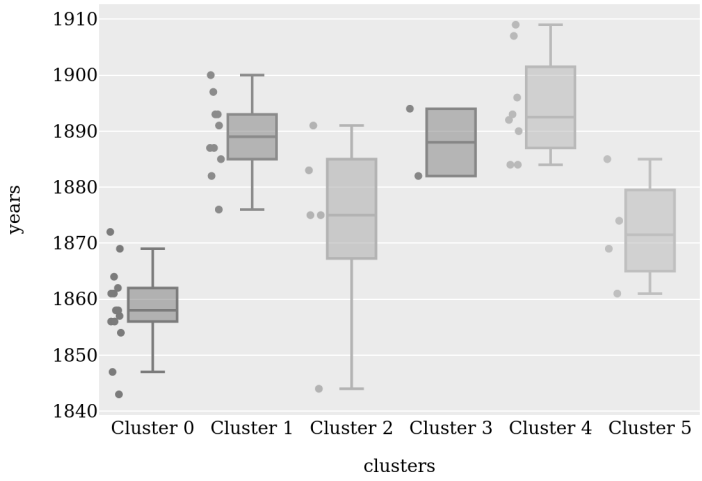


Figure 92. Clusters by year in the network SENT.

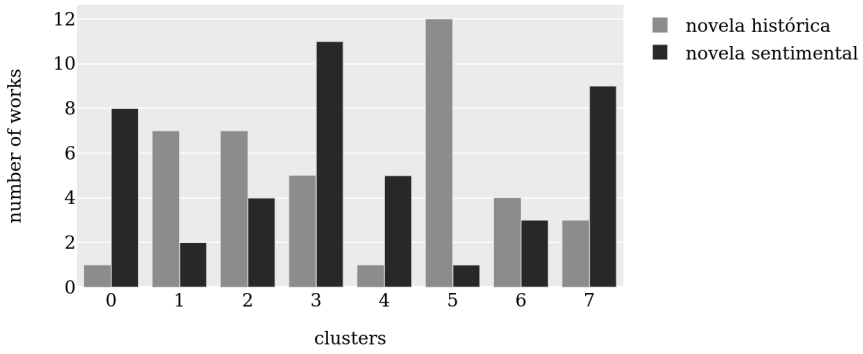


Figure 93. Clusters by subgenre in the combined network.

observed for the sentimental novels alone disappears when they are analyzed in the more general setup. Regarding the distribution by years, cluster 1 is early, clusters 3, 4, and 6 are late, and the others are mixed. The topics that are distinctive for the different *families* reflect the relative purity or mixture of subgenres as well as the preferences of the early versus late nineteenth century.

To conclude, with the analysis of topics in nineteenth-century Spanish-American historical and sentimental novels in a network-based approach, a proposal was made here how the concept of family resemblance that has been introduced into genre theory in the 1960s and argued for by several genre theorists also recently, can be applied in a digital genre stylistics approach. When one looks at the current strategies to categorize genres in this field, the majority focuses on classificatory groupings based on the assumption of features that are common to all members of a class. Yet, there are also alternative ways to analyze genres in digital stylistics. Especially stylometric network analyses implicitly contain the idea of overlapping similarities and unsharp boundaries characteristic of the family resemblance approach. Here these two scenarios were brought together. With the chosen approach to compare feature distributions of the novels in terms of nearest neighborships and to organize the resulting network of similarities into communities interpreted as *families*, the original idea of family resemblance is adjusted for the digital analysis. First, because rather than the presence or absence of individual textual features, the degree of their joint presence in individual pairs of novels is decisive, and second, because communities or clusters found in the similarity network constitute a way to delimit the "families" retroactively, without changing the underlying concept of intertwining shared characteristics of individual members of the groups.

For the Argentine, Mexican, and Cuban historical and sentimental novels, the analysis confirmed that there are subtypes of the subgenres that have been described in literary-historical approaches, such as a novel with a historical setting and a sentimental plot or a historical novel focusing on contemporary political conditions in contrast to novels about historical events about colonial times. In addition, influences of the narrative perspective on subtypes of the sentimental novel became visible. Analyzing both types of subgenres together resulted in mixed groups as well as some that are dominated by one subgenre. While the country the novels were published in does not have a clear impact on the resulting *families* of novels, the year of publication has an

influence in some cases when the preferred and avoided topics reflect the literary development in the nineteenth century.

The family resemblance analysis is a categorization method that is more open than classification, and if it is applied to more than one type of novel, it allows the novels to be grouped based on criteria other than only their conventional subgenre. All kinds of factors that have an influence on the surface features – as the topics in the analysis conducted here – potentially have an impact on the structure of the family resemblance network if these factors are not controlled beforehand through the composition of the corpus, as was done here with authorship. So depending on how many and which determining factors of style are permitted to enter the network, the family resemblance analysis reveals different stylistic connections between the works, conventional subgenre being one of them. All in all, the results show that features common to all novels of a conventional subgenre cannot be expected and that the textual factors that influence the subgroups or *families* of subgenres are diverse. There is not one decisive factor, each *family* has its own traits that hold it together, and inside of each one, there are additional individual traits as well as connections to other *families*.

The algorithm producing the family resemblance network and the resulting data offer an empirical ground on which literary historians can look for sense in genre historical terms. The idea of family resemblance that is implemented here is different from Wittgenstein's metaphor because it is not about resemblances caused by biological kinship or because of the use of the same word for concepts that do not have necessary common semantic features but partial similarities in meaning. It is also not about conventional relationships between the literary works and does not say anything about the historico-cultural and communicative relevance of the connections inside the network. Here the concept is interpreted in terms of textual families. As the family resemblance network is based on stylistic features, it might reveal previously unrecognized textual similarities in addition to confirming known ones on a broader textual basis. By not presupposing strict uniformity inside and strict boundaries between the text types, it might come closer to the multi-faceted genre that the novel is.

5 Conclusion

A central question raised in the introduction was how literary genres can be conceived in theoretical terms as categories that can capture the common aspects that humans see in literary works of the same genres. Moreover, the theoretical conception of literary genres should be able to grasp the common features of literary works belonging to a genre on a textual level. In the context of computational literary studies, this applies not only to the analysis of textual properties of genres by literary scholars but also by computers. To that end, concepts of genre stemming from literary theory were discussed in the theoretical part of this dissertation and were related to the aims and procedures of digital genre stylistics.

Regarding the ontological status of genres, it was argued that they can best be understood as communicative norms or conventions which have an influence on the stylistic form of the literary texts that participate in them. Surface cues of the texts that are related to genre labels can be interpreted as traces or "normative facts", in the terms of Hempfer (1973), which are left by the genre conventions. In this way, even a digital stylistic text analysis can capture elements of communicative phenomena that are, at least to a certain degree, determined by factors that lie outside of the syntactic and semantic level of literary texts.

As a second aspect of the theoretical part of this dissertation, the usefulness of semiotic models of genres for the modeling of generic terms was highlighted. They allow to organize conventional signals that are transported by genre labels on different linguistic and contextual levels so that stylistic analyses can focus on selected semiotic levels. Semiotic levels include the thematic one, which is, for instance, primarily addressed by sentimental or social novels, or the level concerned with the relationship of the texts to reality, for which historical, science fiction, or fantastic novels are examples. These are only two of the possible semiotic levels to which generic names can refer. Most of the genre labels do not refer to one level exclusively, but it is helpful to decide on a specific dimension of genre discourse that is examined, to compare genres or subgenres on a similar level. To make decisions for selected levels of analysis is especially relevant for digital genre stylistics because corpus studies are usually performed in a contrastive setting, either by comparing one subgenre to a larger corpus covering a superordinate major genre or by opposing individual genres or subgenres directly. Analyzing the discursive levels that genre labels refer to can also provide useful hints about the textual traits that could be relevant for them and can lead to hypotheses about the stylistic characteristics of the texts participating in the genre in question. For thematic subgenres, for instance, topic features can be a good choice. However, it must be assumed that most genres are defined on several textual levels at once. In any case, in digital stylistic analysis, the textual traits must all be captured on the textual surface, and they must therefore be formalized and leveled down to text style, as is done by using topics as indicators for thematic developments. As a result, other aspects of text surface style may enter the feature space. In the case of topics, for example, not only thematic elements become visible but also aspects of the setting, plot, or specific literary motifs. This shows that a connection between genre conventions, textual traits of genres, and surface style can be assumed and deduced, or induced, but that this connection is a complex one, which still needs to be investigated further in theoretical and empirical terms.

A third finding of the theoretical part was that for digital genre stylistics, it is advantageous to have separate concepts of literary text types, conventional genres, and textual genres. In this way, not only theoretically driven genre analyses, starting from certain concepts of textual genres are enabled, but also exploratory ones concerned with groupings of texts that emerge from stylistic features. This is important because it is not necessarily directly possible to clarify the relationship between literary definitions of genres and surface style captured in computational analyses. Theoretically, such a relationship is most often hypothesis-driven, and practically, many tools that formalize literary concepts are still missing. In literary genre theory, there are proposals of how to distinguish text types from genres, but they usually adhere to a relationship of dependence between both. Here it was proposed to completely separate the levels of literary text types and literary conventional genres to be able to define and find intersections of both as a result.

In the part on the different literary theoretical concepts of genres as categories (classes, prototypes, and families), it was laid out how they can be applied fruitfully by using statistical categorization methods. Classification, clustering, or network analysis constitute different options to analyze literary text types and to relate them to conventional genres. It was argued that there is no exclusive relationship between each of these text categorization methods and the different literary theoretical concepts of genre categories. Instead, each method covers different aspects of several types of categories. Statistical classification, for instance, can also be used to determine the prototypicality of literary texts as instances of genres to a certain degree. On the other hand, networks can not only be used for family resemblance analyses, but also allow the formation of clusters in the networks, which then represent delimited groups.

In the chapter on style, a definition of literary genre style was formulated that applies the concept of style formulated by Herrmann, Schöch, and van Dalen-Oskam (2015) to the field of literary genre analysis. In this definition, the distinction between communicatively determined, conventional genres and formally determined literary text types is included. Furthermore, a distinction is made between higher-level stylistic traits and low-level surface cues, and it is formulated which types of linguistic features are considered stylistic cues. In this way, the relationships between literary genres and their stylistic characteristics on the one hand and linguistic features of the surface of texts on the other hand is differentiated and made clear.

In the last part of the Concepts chapter, selected thematic subgenres and literary currents of the nineteenth-century Spanish-American novel were presented with a view to genre-stylistic characteristics. In this way, the textual analysis of these subgenres conducted in the Analysis chapter of the dissertation was prepared and related to established literary-historical knowledge.

The various considerations of the theory section contribute to relating established theoretical concepts of literary and linguistic studies to computer-based analyses of literary genres. On the one hand, this should promote a theoretical foundation of textual analyses of digital genre stylistics. On the other hand, it should also be possible to relate the results obtained with digital genre stylistic analyses to previous results obtained with classical methods. They should not be detached from previous research in literary history, but should enter, be, and remain in dialogue with it. Recent developments in the digital humanities show that the need for a stronger theoretical foundation is seen. For example, a working group on digital humanities theory has been established in the DHd association (AG Digital Humanities Theorie 2023). In relation to

the results of this dissertation, the next step will be to see to what extent the suggestions made here for concepts of digital genre stylistics can be useful not only for the analyses in this specific thesis, but also in other application studies. Furthermore, the conceptual considerations of genre categories and computational categorization can still be extended and need to be supported by further empirical studies.

In the empirical part of this thesis, the methodology for the creation of a digital bibliography serving as a sampling frame and a text corpus aimed at the analysis of nineteenth-century Spanish-American novels with regard to subgenres was discussed and presented in detail. The approaches taken were outlined in practical terms and by example, addressing the specific questions, challenges, and solutions found for the bibliography and corpus at hand. Still, the proceedings followed were also the result of general considerations and discussions in the CLiGS project. They can be taken as a proposal for how to prepare a digital corpus of literary texts for genre analysis in a good way. The findings of these reflections in the project led to the creation of the "textbox" as an exemplary set of text collections for digital literary analysis (Schöch et al. 2019). For this dissertation in particular, the considerations and implementation of data modeling in the digital bibliography on novels, their subgenres, and the editions in which they were published is also of a general character and it goes beyond what was worked out on text corpora in the project group. Own decisions have also been made for the digital text corpus regarding its composition and the modeling of the metadata and text structure. These have mainly resulted from the subject matter of the nineteenth-century Spanish-American novels.

One question was, for example, how to define the limits of the corpus in generic terms – how to determine if a literary work can be considered a novel in several hundreds of cases without being able to read every single work? Here, the decision was to start from a very general formal definition of the novel and to check the fulfillment of its requirements in part through quantitative analysis and in part by evaluating available metadata and by checking the information in paratexts manually. The selection of subgenres, on the other hand, was not predetermined, so that the bibliography Bib-ACMé and the corpus Conha19 constitute general resources covering the Spanish-American nineteenth-century novel as it is represented by works from the three countries Argentina, Cuba, and Mexico. It is hoped that this corpus will be used in the future by other researchers for digital text analysis, which may also have a different focus than that pursued here. For example, the corpus could also be used for the study of authorship, for an analysis that examines chronological developments within the nineteenth century, or for the analysis of specific motifs or themes in the texts. Individual texts from Conha19 encoded in TEI could also serve as a starting point for the development of digital critical editions of selected works. Finally, a desirable future development is to add novels from other Spanish-American countries to the corpus, or texts from other major genres (for example, drama), in order to make cross-genre analyses possible.

As the analysis of subgenres was the goal of this thesis, a particular focus in the corpus-building process was on the information about the subgenres of the novels, which was collected by evaluating different literary-historical sources and explicit and implicit historical genre labels and signals. All of this material represents the conventional level of the novels' subgenres. It was organized by encoding it in TEI according to an empirically induced model of discursive levels of subgenres, including the following levels:

- thematic subgenre labels
- labels referring to literary currents
- labels related to the cultural-geographical and linguistic identity
- labels pointing to the relationship between the novel and reality
- labels concerned with the mode that a novel is narrated in
- labels reflecting the author's or narrator's intention or attitude
- labels that refer to the medium that a novel uses and the mode that it is represented in linguistically or narratively

An interesting question is whether other extensive empirical studies of genre signals and genre assignments will arrive at similar categorizations, i.e., whether a general trend of the types of genre signals and, in particular, their frequencies can be identified, or whether the results obtained here are specific to Spanish-American novels. Only further corpus-based studies of historical and literary critical genre assignments will be able to show this.

The information on subgenres was evaluated in the metadata analysis part to see which subgenres were frequent enough to be analyzed on a textual level with quantitative methods. It was found that several of the discursive levels initially proposed by Raible (1980) and Schaeffer (1983) are present in literary critical and also explicit historical subgenre labels but that only a few of them are quantitatively relevant. The frequent ones were found in particular on the thematic level ("novela histórica", "novela de costumbres", "novela sentimental", etc.) and the contextual levels of literary current ("novela romántica", "novela realista", "novela naturalista", and so on) as well as of the cultural, geographical, or linguistic identity (as expressed through the labels "novela argentina", "novela mexicana", "novela cubana", or "novela original"). For the text analysis part, it was decided to focus on the thematic subgenres and the literary currents. Labels related to the cultural, geographical, or linguistic identity have been evaluated in a research article outside of the scope of this dissertation (Henny-Krahmer 2022). Thus, not all categories of genre labels were effectively evaluated in this thesis. The great variety of subgenre labels found for the novels in Bib-ACMé and Conha19 shows how multifaceted the novel as a genre is and further analyses that can proceed from the material encoded here are possible. As a result, quantitative digital text analysis is only one of several possible approaches for analyzing the subgenres of the nineteenth-century Spanish-American novel. The big pool of different subgenres and generic terms that are not so frequent can only meaningfully be analyzed qualitatively.

Besides assessing different levels and perspectives on subgenres of the novels, the metadata analysis also served to check how well the corpus represents the material collected in the more extensive bibliography in quantitative terms. The relationship between both resources was found to be approximately proportional in most but not all aspects. For example, the corpus contains more works written by well-known, high-prestige authors than the bibliography. This is not surprising because these are the works that have primarily been digitized. Still, this aspect of the corpus can still be improved. It is desirable to enlarge the corpus in the future and include even more works of lesser-known authors than the ones it already contains. That would make the corpus more representative of the whole production of Argentine, Cuban, and Mexican nineteenth-century novels. As the corpus is published with this dissertation, it can be hoped

that it provides a starting point for building a more extensive collection of nineteenth-century Spanish-American novels.

In the text analysis part, statistical classification was used as the first method for categorizing the novels by their subgenre. To this end, one primary subgenre label was determined for each novel, and the classification was done for thematic subgenres and literary currents. Both subgenres and currents can be classified well on the basis of most frequent words and topic features. Experiments with three types of classifiers (K-nearest neighbor, Support Vector Machine, and Random Forest) and several different feature sets were run. In the case of MFW, different token units, numbers of words, and normalization techniques were used. For topics, different numbers of topics and optimization parameters for the modeling were chosen. The results showed that both subgenres and currents are textually coherent to degrees of 70 to 90 %, depending on the type of features and subgenre. Textual coherence here refers to the degree to which the communicatively determined subgenre assignments of the novels coincide with their class assignments as determined by text classification, measured in classification accuracy.

In general, the differences between the two feature types were minor. In the case of thematic subgenres, there was almost no difference in the accuracy values between most frequent words and topics, and for literary currents, most frequent words worked a bit better than topics.

The classification results for the literary currents were better than for the thematic subgenres. This was not expected from the beginning because the currents are broader phenomena which, at least in the case of the realist and romantic novels, include several thematic subgenres. While the realist and naturalistic novels were mainly published in the last two decades of the nineteenth century and the first decade of the twentieth century, the romantic novel was present from the early decades of the nineteenth century up to the last ones, with a dominant role up to the 1870s. So also diachronically, especially the romantic novels constitute a general phenomenon. The classification results suggest that the literary currents nonetheless represent a level of the novels that can be captured better stylistically than the thematic subgenres.

That literary currents can be better recognized stylistically than thematic subgenres was also visible in the visualizations of the amount of correct and wrong classifications for individual novels of the corpus. For the thematic subgenres, there were cases of regular false positives and false negatives, that is, novels that carry the respective genre label but were classified as another subgenre in more than 70 % of the cases or novels that do not carry the label but were recognized as members of the subgenre almost in every case. Such instances of novels were not present in the results for the literary currents. The same could be observed for the "middle range" of novels that are misclassified in around 30 to 70 % of the cases, which happened for the thematic subgenres but not for the literary currents. This concludes that the levels of genre convention and text type are more congruent for the literary currents than for the thematic subgenres. When evaluating these results, it has to be kept in mind that the labels for literary currents were mainly collected from literary-historical sources. In contrast, the thematic labels are also present as explicit historical labels on the texts, so it is more probable that there are discrepancies between convention and text type than for critically established labels. Nevertheless, it may as well be the case that there is less consensus on the major theme of novels than on their dominant style in terms of literary currents.

The text classification in the context of this dissertation has not yet taken into account the multiple subgenre assignments recorded in Bib-ACMé and Conha19. For the purposes of this thesis, the most plausible primary subgenre label was selected as the classification target in order to establish a classification baseline in a classical setting in the first place. This has not been done before for Spanish-American novels. The next obvious step is applying multi-label classification procedures to include secondary and further subgenre assignments. Especially in the case of thematic subgenres, there are very often multiple classifications on a communicative level. These occur in the historical subtitles of the novels (e.g., "novela histórica de costumbres") and especially when literary scholars describe the novels in terms of thematic subgenres. It can therefore be assumed that such multiple attributions are also textually and stylistically relevant. Another way to further map the complexity of the novels in terms of their subgenres would be to include chapter structures or to divide the novels into text sections to classify them individually. The corpus Conha19 is well prepared for both approaches.

As an alternative approach to text categorization, a family resemblance network analysis was conducted, with a focus on the internal structure of individual subgenres and a focus on the historical novel. It showed that specific subtypes of historical novels are obscured in the classification approach because they do not represent the quantitatively dominant type of romantic historical novels. However, these subtypes of historical novels become visible as a family when the corpus of historical novels is partitioned into network communities based on individual similarity relationships between the texts. When two subgenres were combined in the network, it became clear that several factors other than genre influence the style of the novels, for instance, the period of publication or the narrative perspective. These results show that statistical classification is a powerful method that yields very good results but also occludes interior subdivisions and stylistic differences of the novels that are part of the subgenres, so for exploratory and refining analyses, a family resemblance network analysis is a good alternative.

Returning to the observation made at the outset that categorization is a basic human need, the analyses here have shown that human assignments of texts to genres, in terms of communicative phenomena, function differently from computer-assisted stylistic genre classification. The classifiers are powerful tools that can recognize which subgenres the novels belong to in most cases. However, they do not (yet) understand how the text style is related to the perceived conventional genre of the texts. This led to errors and unexpected results in some individual cases. The classifiers are more strict in interpreting the textual surface structures. Because they are different in that way from a human author or reader, through the analysis of mistakes that the statistical models make, they bring the opportunity to understand better how genres function as communicative phenomena. This understanding is supported by alternative categorization methods, such as the family resemblance approach presented in this thesis. Research in digital genre stylistics and computational literary studies needs to continue to engage with literary theory, literary history, and computational approaches to develop their own suitable methods for their objects of study.

In the future, the text analysis can be developed further by using other features, especially those that can be related more directly to literary concepts, such as character constellations or literary space. In many cases, however, the computational caption of such features on the textual surface has yet to be developed. Another enhancement for the analysis would be not to

analyze the novels as a whole, as was done here, but to evaluate the results for text segments, chapters, or different parts of the narration, such as direct speech versus narrator text. The textual characteristics of genres may also be defined in connection to these substructures of the texts and not only for the entire texts. Furthermore, classification and network analysis are just two of the further available options of computational text categorization methods that could be employed for genre analysis. Especially the concept of genres as prototypical structures could be approached with clustering methods, for example, which has not been done in the scope of this dissertation. Moreover, of course, there should be more empirical studies based on other literatures and corpora in order to be able to identify broader trends in the history of genres and to be able to come to more general conclusions than was possible in this work on the nineteenth-century Spanish-American novel. Thus, future research directions in digital genre stylistics could be envisioned here but have yet to be carried out. Hopefully, the spirit of Open Science will be followed in future genre-stylistic research, as it has been done in this work, by making the texts, data, and scripts all openly available, as far as legally possible.

References

Adamzik, Kirsten, ed. 2001. *Kontrastive Textologie: Untersuchungen zur deutschen und französischen Sprach- und Literaturwissenschaft.* Tübingen: Stauffenburg-Verlag.

AG Digital Humanities Theorie. 2023. Digital Humanities Theorie. Theorie und Theoriebildung in den digitalen Geisteswissenschaften. https://web.archive.org/web/20230208165032/https://dhtheorien.hypotheses.org/.

Álamo Felices, Francisco. 2011. *Los subgéneros novelescos. Teoría y modalidades narrativas.* Almería: Universidad Almería.

Alburquerque-García, Luis. 2015. "Poética de la literatura de viaje." In *Cartografía de la literatura de viaje en Hispanoamérica*, edited by Daniar Chávez and Marco Urdapilleta, 19–34. México: Universidad Autónoma del Estado de México.

Alegría, Fernando. 1959. *Breve historia de la novela hispanoamericana.* México: Ed. de Andrea.

Alpaydin, Ethem. 2016. *Machine Learning: The New AI.* Cambridge, Mass.: The MIT Press.

Altamirano, Ignacio Manuel. 1868. *Revistas literarias de México.* México: T. F. Neve.

Anderson Imbert, Enrique. 1954. *Historia de la literatura hispanoamericana.* México: Fondo de Cultura Económica.

Anderson Imbert, Enrique. 1995. *Historia de la literatura hispanoamericana.* Vol. 1: La colonia. Cien años de república. 2nd ed. México: Fondo de Cultura Económica.

Andresen, Melanie, and Heike Zinsmeister. 2019. *Korpuslinguistik.* Tübingen: Narr Francke Attempto.

Anz, Thomas. 2007. "Inhaltsanalyse." In *Handbuch Literaturwissenschaft. Gegenstände – Konzepte – Institutionen*, edited by Thomas Anz, vol. 2, *Methoden und Theorien*, 55–69. Stuttgart, Weimar: J.B. Metzler.

Ardao, Arturo. 1980. *Genesis de la idea y el nombre de América Latina.* Caracas: Centro de Estudios Latinoamericanos Rómulo Gallegos.

Armas, Emilio de. 1997. "Cuba. 19th- and 20th-Century Prose and Poetry." In *Encyclopedia of Latin American Literature*, edited by Verity Smith, 235–242. London/Chicago: Fitzroy Dearborn Publishers.

Arora, Sanjeev, Rong Ge, and Ankur Moitra. 2012. "Learning Topic Models – Going beyond SVD." In *Proceedings of the 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, 1–10. Washington D.C.: IEEE Computer Society. https://doi.ieeecomputersociety.org/10.1109/FOCS.2012.49.

Avellaneda, Gertrudis Gómez de. (1871) 2008. *La Baronesa de Joux: leyenda fundada en una tradición francesa (en formato HTML).* Alicante: Biblioteca Virtual Miguel de Cervantes. https://www.cervantesvirtual.com/nd/ark:/59851/bmcng4q2.

Barth, Florian, and Gabriel Viehhauser. 2017. "Digitale Modellierung literarischen Raums." In *DHd2017. Konferenzabstracts.* https://doi.org/10.5281/zenodo.4622732.

Barthes, Roland. (1953) 2002. *Le Degré zero de l'écriture.* Reprint, Paris: Seuil.

Bawarshi, Anis S., and Mary Jo Reiff. 2010. *Genre: An Introduction to History, Theory, Research, and Pedagogy.* West Lafayette: Parlor Press and the WAC Clearinghouse. https://web.archive.org/web/20230210055352/https://wac.colostate.edu/docs/books/bawarshi_reiff/genre.pdf.

Behnel, Stefan. 2022. "Validation with lxml." lxml – XML and HTML with Python. https://web.archive.org/web/20230611112928/https://lxml.de/validation.html.

Behrens, Irene. 1940. *Die Lehre von der Einteilung der Dichtkunst. Vornehmlich vom 16. bis 19. Jahrhundert.* Halle/Saale: Max Niemeyer Verlag.

Bei, Yu. 2008. "An Evaluation of Text Classification Methods for Literary Study." *Literary and Linguistic Computing* 23: 327–343. https://doi.org/10.1093/llc/fqn015.

Berlin-Brandenburgische Akademie der Wissenschaften, ed. 2022. "Deutsches Textarchiv. Grundlage

für ein Referenzkorpus der neuhochdeutschen Sprache." DTA. Accessed November 6, 2022. https://web.archive.org/web/20221106163539/https://www.deutschestextarchiv.de/.

Bernecker, Walter L., ed. 1992. *Handbuch der Geschichte Lateinamerikas.* Vol. 2: Lateinamerika von 1760 bis 1900. Stuttgart: Klett-Cotta.

Betz, Katrin, Ulrike Henny-Krahmer, Christian Pölitz, Christof Schöch, and Albin Zehe. 2017. "The 'topic modeling workflow (tmw)'." Talk presented at the workshop 'Let's Develop an Infrastructure for Historical Research Tools' at the conference DH2017, Montreal, Canada. Accessed November 14, 2020. https://christofs.github.io/tmw-dh/#/.

Biber, Douglas. 1992. "The multi-dimensional approach to linguistic analyses of genre variation: An overview of methodology and findings." *Computers and the Humanities* 26 (5–6): 331–345. https://doi.org/10.1007/BF00136979.

Biber, Douglas. 1993a. "Representativeness in Corpus Design." *Literary and Linguistic Computing* 8 (4): 243–257. https://web.archive.org/web/20230128095417/http://otipl.philol.msu.ru/media/biber930.pdf.

Biber, Douglas. 1993b. "The Multi-Dimensional Approach to Linguistic Analyses of Genre Variation: An Overview of Methodology and Findings." *Computers in the Humanities* 26 (5–6): 331–345. https://doi.org/10.1007/BF00136979.

Biblioteca Nacional de Cuba José Martí. 2011. "Bibliografía Nacional Cubana." https://web.archive.org/web/20190702105833/http://bdigital.bnjm.cu/?secc=bibliografias.

Biblioteca Nacional de España. 2023. "Biblioteca Digital Hispánica." https://web.archive.org/web/20230603173847/http://bdh.bne.es/bnesearch/Inicio.do.

Binongo, José Nilo G., and M. W. A. Smith. 1999. "A Bridge Between Statistics and Literature: The Graphs of Oscar Wilde's Literary Genres." *Journal of Applied Statistics* 26 (7): 781–787. https://doi.org/10.1080/02664769922025.

Blei, David M. 2012. "Probabilistic Topic Models." *Communications of the ACM* 55 (4): 77–84. https://doi.org/10.1145/2133806.2133826.

Blei, David M., and John D. Lafferty. 2006. "Dynamic topic models." In *Proceedings of the 23rd International conference on Machine learning (ICML)*, 113–120. https://doi.org/10.1145/1143844.1143859.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent dirichlet allocation." *Journal of Machine Learning Research* 3: 993–1022. https://web.archive.org/web/20230310095853/https://dl.acm.org/doi/pdf/10.5555/944919.944937.

Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. "Fast unfolding of communities in large networks." *Journal of Statistical Mechanics: Theory and Experiment* 10: 155–168. https://10.1088/1742-5468/2008/10/P10008.

Bonheim, Helmut. 1992. "The Cladistic Method of Classifying Genres." *Yearbook of Research in English and American Literature (REAL)* 8: 1–32.

Botrel, Jean-François. 2001. "La novela, género editorial (España, 1830–1930)." In *La novela en España en los siglos XIX y XX. Historia, sociedad, búsqueda identitaria*, edited by Paul Aubert, 35–51. Madrid: Casa de Velázquez. https://web.archive.org/web/20230606180953/https://books.openedition.org/cvz/2631.

Branco, Paula, Luís Torgo, and Rita P. Ribeiro. 2015. "A Survey of Predictive Modelling under Imbalanced Distributions." *ACM Computing Surveys* 49 (2): 1–50. https://doi.org/10.1145/2907070.

Brinker, Klaus. 1992. *Textlinguistik.* Heidelberg: Groos.

Brinker, Klaus, Hermann Cölfen, and Steffen Pappert. 2014. *Linguististische Textanalyse. Eine Einführung in Grundbegriffe und Methoden.* 8th ed. Berlin: Erich Schmidt Verlag.

Brunner, Annelen. 2013. "Automatic recognition of speech, thought, and writing representation in German narrative texts." *Literary and Linguistic Computing.* https://doi.org/10.1093/llc/fqt024.

Brunner, Annelen. 2015. *Automatische Erkennung von Redewiedergabe: ein Beitrag zur quantitativen Narratologie.* Narratologia: contibutions to narrative theory. Vol. 47. Berlin, Boston: De Gruyter.

Brushwood, John S. 1966. *Mexico in its Novel. A Nation's Search for Identity.* Austin: University of Texas Press.

Bundesamt für Justiz. n.d.a "Gesetz über Urheberrecht und verwandte Schutzrechte (Urheberrechtsgesetz). § 64 Allgemeines." Gesetze im Internet. https://web.archive.org/web/20230423112139/https://www.gesetze-im-internet.de/urhg/__64.html.

Bundesamt für Justiz. n.d.b "Gesetz über Urheberrecht und verwandte Schutzrechte (Urheberrechtsgesetz). § 66 Anonyme und pseudonyme Werke." Gesetze im Internet.https://web.archive.org/web/20230423112720/https://www.gesetze-im-internet.de/urhg/__66.html.

Bundesamt für Justiz. n.d.c "Gesetz über Urheberrecht und verwandte Schutzrechte (Urheberrechtsgesetz). § 70 Wissenschaftliche Ausgaben." Gesetze im Internet.https://web.archive.org/web/20230423113034/https://www.gesetze-im-internet.de/urhg/__70.html.

Burnard, Lou. 2014. *What is the text encoding initative? How to add intelligent markup to digital resources.* Encyclopédie numérique 3. Marseille: OpenEdition Press. https://doi.org/10.4000/books.oep.426.

Burrows, John. 2002. "'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship." *Literary and Linguistic Computing* 17 (3): 267–287. https://doi.org/10.1093/llc/17.3.267.

Burrows, John. 2007. "All the Way Through: Testing for Authorship in Different Frequency Strata." *Literary and Linguistic Computing* 22 (1): 27–47. https://doi.org/10.1093/llc/fqi067.

Burton, Matt. 2013. "The Joy of Topic Modeling." May 21, 2013. https://web.archive.org/web/20211012091043/http://mcburton.net/blog/joy-of-tm/.

Byszuk, Joanna, Michal Woźniak, Mike Kestemont, Albert Leśniak, Wojciech Łukasik, Artjoms Šeļa, and Maciej Eder. 2020. "Detecting Direct Speech in Multilingual Collection of 19th-century Novels." In *Proceedings of LT4HALA 2020 – 1st Workshop on Language Technologies for Historical and Ancient Languages*, 100–104. Marseille: European Language Resources Association (ELRA). https://web.archive.org/web/20230611135104/https://aclanthology.org/2020.lt4hala-1.15.pdf

Calderón, Mario. 2005. "La novela costumbrista mexicana." In *La república de las letras. Asomos a la cultura escrita del México decimonónico*, edited by Belem Clark de Lara and Elisa Speckman Guerra, 315–324. Vol. 1: Ambientes, asociaciones y grupos. Movimientos, temas y géneros literarios. México: Universidad Nacional Autónoma de México.

Calvo Tello, José. 2018. "Genre Classification in Novels: A Hard Task for Humans and Machines?" In *EADH 2018: Data in Digital Humanities. Conference Abstracts.* Galway: National University of Ireland. https://web.archive.org/web/20230304103733/https://eadh2018.exordo.com/files/papers/46/final_draft/20181205_genre_classification_human_vs_machines.pdf.

Calvo Tello, José, ed. 2021a. "Corpus of Novels of the Spanish Silver Age (CoNSSA)." Version 1.0.0. GitHub.com. Accessed December 9, 2022. https://github.com/cligs/conssa.

Calvo Tello, José. 2021b. *The Novel in the Spanish Silver Age. A Digital Analysis of Genre Using Machine Learning.* Digital Humanities Research, vol. 4. Bielefeld: Bielefeld University Press. https://doi.org/10.14361/9783839459256.

Calvo Tello, José, Ulrike Henny-Krahmer, and Christof Schöch. 2018. "Textbox. Análisis del léxico mediante corpus literarios." In *Historia del léxico español y Humanidades Digitales*, edited by Dolores Corbella, Alejandro Fajardo, and Jutta Langenbacher, 225–253. Berlin: Peter Lang.

Calvo Tello, José, Daniel Schlör, Ulrike Henny, and Christof Schöch. 2017. "Neutralizing the Authorial Signal in Delta by Penalization: Stylometric Clustering of Genre in Spanish Novels." In *Digital Humanities 2017. Conference Abstracts, Montréal, Canada, August 8–11, 2017*, 181–184. Montreal: McGill University & Université de Montréal. https://web.archive.org/web/20230212053238/https://dh2017.adho.org/abstracts/037/037.pdf.

Cambaceres, Eugenio. (1885) 2000. *Sin rumbo (en formato HTML).* Alicante: Biblioteca Virtual Miguel de Cervantes. https://www.cervantesvirtual.com/nd/ark:/59851/bmc1v5d1.

Cambaceres, Eugenio. 2008. “Sin rumbo.” Wikisource. https://web.archive.org/web/20230422143111/https://es.wikisource.org/wiki/Sin_rumbo.

Campo y Valle, Ángel de. 2009. *La Rumba.* Colección Autores del Siglo XIX. México: Instituto Latinoamericano de la Comunicación Educativa. http://web.archive.org/web/20160615221017/http://bibliotecadigital.ilce.edu.mx/Colecciones/ObrasClasicas/_docs/Rumba.pdf.

Cárabes, Celia Miranda, and Jorge A. Ruedas de la Serna. 1998. *La novela corta en el primer romanticismo mexicano.* 2nd ed. Mexico: Universidad Nacional Autónoma de México.

Cárrega, Hemilce. 1986. *Las novelas argentinas de Carlos María Ocantos.* Buenos Aires: Febra Editores.

Centre National de Ressources Textuelles et Lexicales (CNRTL). 2012. “RASTAQUOUÈRE.” Portail lexical. https://web.archive.org/web/20230611105723/https://www.cnrtl.fr/etymologie/rastaquou%C3%A8re.

Centro Biblioteca Virtual Miguel de Cervantes. 2023. “Biblioteca Virtual Miguel de Cervantes.” Accessed March 28, 2023. http://www.cervantesvirtual.com/.

Chaves, José Ricardo. 2011. “Huellas y enigmas de la novela corta en el siglo XIX.” In *Una selva infinita. La novela corta en México (1872–2011)*, edited by Gustavo Jiménez Aguirre, Gabriel M. Enríquez Hernández, Esther Martínez Luna, Salvador Tovar Mendoza, and Raquel Velasco, 109–127. México: Fundación para las Letras Mexicanas.

Chávez, Daniar, and Marco Urdapilleta. 2015. “Prólogo.” In *Cartografía de la literatura de viaje en Hispanoamérica*, edited by Daniar Chávez and Marco Urdapilleta, 9–17. México: Universidad Autónoma del Estado de México.

CLiGS. n.d. “Call for Papers: Digital Stylistics in Romance Studies and beyond.” CLiGS – Computergestützte literarische Gattungsstilistik. Accessed October 23, 2022. http://web.archive.org/web/20221023113851/https://cligs.hypotheses.org/digital-stylistics-in-romance-studies-and-beyond/call-for-papers.

Computational Stylistics Group. 2023. “Resources.” https://web.archive.org/web/20230423092924/https://computationalstylistics.github.io/resources/.

Cranenburgh, Andreas van, and Corina Koolen. 2015. “Identifying Literary Texts with Bigrams.” In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, 58–67. Denver, Colorado: Association for Computational Linguistics. http://dx.doi.org/10.3115/v1/W15-0707.

Creative Commons. n.d. “Public Domain Mark 1.0.” https://web.archive.org/web/20230610120916/https://creativecommons.org/publicdomain/mark/1.0/deed.en.

Croce, Benedetto. 1905. *Aesthetik als Wissenschaft des Ausdrucks und allgemeine Linguistik. Theorie und Geschichte.* Leipzig: E.A. Seemann.

Cuéllar, José Tomás de. 1890. “Prólogo.” In *Ensalada de pollos. Novela de estos tiempos que corren (1871) tomada del carnet de Facundo (José T. de Cuéllar).* Vol. 1 of *La linterna mágica. Segunda época.* Barcelona: Tipo-Litografía de Hermenegildo Miralles. http://web.archive.org/web/20230128094558/http://cdigital.dgb.uanl.mx/la/1080046422_C/1080046436_T2/1080046436_01.pdf.

Daireaux, Godofredo. (1945) 2001. *Los dioses de la Pampa (en formato HTML).* Alicante: Biblioteca Virtual Miguel de Cervantes. https://www.cervantesvirtual.com/nd/ark:/59851/bmcp55j4.

Daireaux, Godofredo. 2006. “Los dioses de la Pampa.” Wikisource. https://web.archive.org/web/20230328202403/https://es.wikisource.org/wiki/Los_dioses_de_la_Pampa_%28Versi%C3%B3n_para_imprimir%29.

Danneberg, Lutz, and Jürg Niederhauser, eds. 1998. *Darstellungsformen der Wissenschaften im Kontrast: Aspekte der Methodik, Theorie und Empirie.* Tübingen: Narr.

Derrida, Jacques. 1980. “The Law of Genre.” Translated by Avital Ronell. *Critical Inquiry* 7 (1): 55–81.

Díaz Covarrubias, Juan. n.d. *El diablo en México.* Obras clásicas de siempre. Biblioteca Digital del ILCE. https://web.archive.org/web/20230423115244/http://bibliotecadigital.ilce.edu.mx/Colecciones/ObrasClasicas/_docs/El_diablo_en_Mexico-Juan_Diaz_Covarrubias.pdf.

Dill, Hans-Otto. 1999. *Geschichte der lateinamerikanischen Literatur im Überblick.* Stuttgart: Reclam.

Du, Keli. 2019. “A Survey on LDA Topic Modeling in Digital Humanities.” In *Proceedings of DH2019:*

'Complexities'. Utrecht: Utrecht University. https://web.archive.org/web/20220121042220/https://dev.clariah.nl/files/dh2019/boa/0326.html.

Dubrow, Heather. (1982) 2014. *Genre.* Reprint, London: Routledge.

Duff, David, ed. 2010. *Modern Genre Theory.* Harlow: Longman.

Eder, Maciej. 2017. "Visualization in stylometry: Cluster analysis using networks." *Digital Scholarship in the Humanities* 32 (1): 50–64. https://doi.org/10.1093/llc/fqv061.

Elkafrawy, Passent, Amr Mausad, and Heba Esmail. 2015. "Experimental Comparison of Methods for Multi-Label Classification in Different Application Domains." *International Journal of Computer Applications* 114 (19): 1–9. http://web.archive.org/web/20230513205320/https://research.ijcaonline.org/volume114/number19/pxc3901666.pdf.

Eroms, Hans-Werner. 2008. *Stil und Stilistik: eine Einführung*. Berlin: Schmidt.

Ertler, Klaus-Dieter. 2002. *Kleine Geschichte des lateinamerikanischen Romans: Strömungen – Autoren – Werke.* Tübingen: Gunter Narr.

Escalona Rios, Lina. 2006. "El trabajo bibliográfico en México." In *Recursos bibliográficos y de información*, edited by Hugo Alberto Figueroa Alcántara and César Augusto Ramírez Velázquez, 185–215. México: Universidad Nacional Autónoma de México. http://hdl.handle.net/10391/4727.

Estrada, Santiago. 1866. *La flor de las tumbas.* Buenos Aires: Imprenta del Siglo.

Evert, Stefan, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, and Thorsten Vitt. 2017. "Understanding and explaining Delta measures for authorship attribution." *Digital Scholarship in the Humanities* 32 (Supplement 2): ii4–ii16. https://doi.org/10.1093/llc/fqx023.

Expert Advisory Group on Language Engineering Standards (EAGLES). 1996. "EAGLES. Recommendations for the Morphosyntactic Annotation of Corpora." https://web.archive.org/web/20230610174614/https://home.uni-leipzig.de/burr/Verb/htm/LinkedDocuments/annotate.pdf.

Fellbaum, Christiane, ed. 1998. *WordNet: An Electronic Lexical Database.* Cambridge, MA: MIT Press.

Fernández-Arias Campoamor, José. 1952. *Novelistas de Mejico: esquema de la historia de la novela mejicana (de Lizardi al 1950).* Madrid: Ediciones Cultura Hispánica.

Fernández Prieto, Celia. 1996. "Poética de la novela histórica como género literario." *Signa. Revista de la Asociación Española de Semiótica* 5: 185–202. https://www.cervantesvirtual.com/nd/ark:/59851/bmc7p9c7.

Ferrer, José Luis. 2018. *La invención de Cuba: Novela y nación (1837–1846).* Madrid: Editorial Verbum.

Fièvre, Paul, ed. 2007–2022. "Théâtre Classique." Accessed December 10, 2022. https://www.theatre-classique.fr.

Firth, John Rupert. 1968. "A synopsis of linguistic theory 1930–1955." In *Selected Papers of J. R. Firth 1952–1959*, edited by F. R. Palmer, 168–205. Bloomington, London: Indiana University Press.

Fischer, Frank, Ingo Börner, Mathias Göbel, Angelika Hechtl, Christopher Kittel, Carsten Milling, and Peer Trilcke. 2019. "Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama." In *Proceedings of DH2019: 'Complexities'*. Utrecht: Utrecht University. http://web.archive.org/web/20220303001044/https://dev.clariah.nl/files/dh2019/boa/0268.html.

Fischer, Frank, and Jannik Strötgen. 2017. "Corpus of German-Language Fiction (txt)." figshare. https://doi.org/10.6084/m9.figshare.4524680.v1.

Fischer, Frank, Peer Trilcke, Julia Jennifer Beine, and Boris Orekhov, eds. n.d. "Drama Corpora Project." Accessed December 6, 2022. https://dracor.org/.

Fishelov, David. 1993. *Metaphors of genre: the role of analogies in genre theory.* University Park, PA: Pennsylvania State Univ. Press.

Fix, Ulla, Andreas Gardt, and Joachim Knape, eds. 2008. *Rhetoric and Stylistics. An International Handbook of Historical and Systematic Research.* 2 vols. Berlin, New York: De Gruyter.

Fludernik, Monika. 2009. "Roman." In *Handbuch der literarischen Gattungen*, edited by Dieter Lamping and Sandra Poppe, 627–645. Stuttgart: Kröner.

Foote, Jonathan. 2000. "Automatic Audio Segmentation Using A Measure of Audio Novelty." In *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia*, 452–455. Vol. 1. New York: IEEE. https://doi.org/10.1109/ICME.2000.869637.

Forster, E. M. 1927. *Aspects of the novel.* New York: Harcourt, Brace & Company.

Fowler, Alastair. 1982. *Kinds of Literature. An Introduction to the Theory of Genres and Modes.* Oxford: Clarendon Press.

Frech, Susana. 1998–2017. "Países, sus capitales y gentilicios* correspondientes Es < En." Susana Frech. Traducciones Profesionales. https://web.archive.org/web/20210615063220/http://www.susana-translations.de/paises.htm.

Fricke, Harald. 1981. *Norm und Abweichung. Eine Philosophie der Literatur.* München: Beck.

Fricke, Harald. 2010. "Definitionen und Begriffsformen." In *Handbuch Gattungstheorie*, edited by Rüdiger Zymner, 7–10. Stuttgart, Weimar: J.B. Metzler.

Frow, John. 2015. *Genre. The New Critical Idiom.* 2nd ed. London: Routledge.

Fubini, Mario. 1971. *Entstehung und Geschichte der literarischen Gattungen.* Tübingen: Max Niemeyer Verlag.

Fundación para las Letras Mexicanas A.C. 2018. "Enciclopedia de la literatura en México." https://web.archive.org/web/20230603174401/http://www.elem.mx/.

Gálvez, Marina. 1990. *La novela hispanoamericana (hasta 1940).* Madrid: Taurus.

Gansel, Christina. 2011. *Textsortenlinguistik.* Göttingen: Vandenhoeck & Ruprecht.

García, Germán. 1952. *La novela argentina: Un itinerario.* Buenos Aires: Editorial Sudamericana.

García Berrio, Antonio, and Javier Huerta Calvo. 2009. *Los géneros literarios: sistema e historia. Una introducción.* 5th ed. Madrid: Cátedra.

Gemeinböck, Iris. 2016. "Representativeness in corpora of literary texts: introducing the C18P project." *MATLIT: Materialidades da Literatura* 4 (2): 29–48. https://doi.org/10.14195/2182-8830_4-2_2.

Genette, Gérard. 1987. *Seuils.* Paris: Seuil.

Genette, Gérard. 2014. "The Architext." In *Modern Genre Theory*, edited by David Duff, 210–218. New York: Routledge.

Gerlach, Martin, Tiago P. Peixoto, and Eduardo G. Altmann. 2018. "A network approach to topic models." *Science Advances* 4 (7): 1–11. https://dx.doi.org/10.1126/sciadv.aaq1360.

Ghiano, Juan Carlos. 1957. "La novela gauchesca." *La Biblioteca*, Época 2, 9 (1): 17–38.

Gianitsos, Efthimios Tim, Thomas J. Bolt, Pramit Chaudhuri, and Joseph P. Dexter. 2019. "Stylometric Classification of Ancient Greek Literary Texts by Genre." In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, Minneapolis, MN, USA, June 7, 2019*, 52–60. Minneapolis: Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/W19-2507.

Gius, Evelyn, Katharina Krüger, and Carla Sökefeld. 2019. "Korpuserstellung als literaturwissenschaftliche Aufgabe." In *DHd2019. Digital Humanities: multimedial & multimodal. Konferenzabstracts, Universitäten zu Mainz und Frankfurt, 25. bis 29. März 2019*, edited by Patrick Sahle, 164–166. Frankfurt & Mainz: Verband Digital Humanities im deutschsprachigen Raum e.V. https://doi.org/10.5281/zenodo.2596095.

Gius, Evelyn, Christof Schöch, and Peer Trilcke, eds. 2022–2023. Journal of Computational Literary Studies (JCLS). Darmstadt: Universitäts- und Landesbibliothek Darmstadt. https://web.archive.org/web/20230210112118/https://jcls.io/.

Gnutzmann, Claus. 1990. *Kontrastive Linguistik.* Frankfurt am Main: Lang.

Gnutzmann, Rita. 1998. *La novela naturalista en Argentina (1880–1900).* Amsterdam, Atlanta: Rodopi.

Goić, Cedomil. 1980. *Historia de la novela hispanoamericana.* 2nd ed. Valparaiso: Ed. Univ. de Valparaiso.

Goić, Cedomil. 2009. *Brevísima relación de la historia de la novela hispanoamericana.* Madrid: Biblioteca Nueva.

González, Joaquín Víctor. (1905) 2001. *Mis montañas (en formato HTML).* Alicante: Biblioteca Virtual Miguel de Cervantes. https://www.cervantesvirtual.com/nd/ark:/59851/bmcw37r4.

Gopnik, Alison, Andrew Meltzoff, and Patricia Kuhl. 2009. *The Scientist in the Crib. What Early Learning Tells Us About the Mind.* New York: HarperCollins. First published 1999.

Gorriti, Juana Manuela. (1889) 2001. *La tierra natal (en formato HTML).* Alicante: Biblioteca Virtual Miguel de Cervantes. https://www.cervantesvirtual.com/nd/ark:/59851/bmc222t4.

Gülich, Elisabeth, and Wolfgang Raible. 1872. *Textsorten. Differenzierungskriterien aus linguistischer Sicht.* Frankfurt am Main: Athenäum-Verlag.

Gutiérrez, Eduardo. (1880) 2016a. *El Jorobado.* Wikimedia Commons. https://web.archive.org/web/20230325193326/https://upload.wikimedia.org/wikipedia/commons/c/c3/El_Jorobado_-_Eduardo_Gutierrez.pdf.

Gutiérrez, Eduardo. (1880) 2016b. *Juan Moreira.* Wikimedia Commons. https://web.archive.org/web/20230325193551/https://upload.wikimedia.org/wikipedia/commons/2/21/Juan_Moreira_-_Eduardo_Gutierrez.pdf.

Gutiérrez, Eduardo. (1893) 2016. *Carlo Lanza. Episodios curiosos.* Wikimedia Commons. https://web.archive.org/web/20230325192637/https://upload.wikimedia.org/wikipedia/commons/e/e0/Carlo_Lanza_-_Eduardo_Gutierrez.pdf.

Gutiérrez, Eduardo, and Bartolomé R. Aprile. (1944) 2015. *Juan Cuello. Novela histórica de Eduardo Gutiérrez, versificada por Bartolomé R. Aprile.* Berlin: Ibero-Amerikanisches Institut – Preußischer Kulturbesitz. http://resolver.iai.spk-berlin.de/IAI00005CB300000000.

Gutiérrez, Eduardo, and Silverio Manco. (1948) 2015. *El rastreador. Novela histórica de Eduardo Gutiérrez, versificada por Silverio Manco.* Berlin: Ibero-Amerikanisches Institut – Preußischer Kulturbesitz. http://resolver.iai.spk-berlin.de/IAI00005C9700000000.

Gutiérrez, Eduardo, and Apolinario Sierra. (1944) 2015. *Aparicio Saravia. Novela histórica por Eduardo Gutiérrez, versificada por Apolinario Sierra.* Berlin: Ibero-Amerikanisches Institut – Preußischer Kulturbesitz. http://resolver.iai.spk-berlin.de/IAI00005CB200000000.

Gymnich, Marion, and Birgit Neumann. 2007. "Vorschläge für eine Relationierung verschiedener Aspekte und Dimensionen des Gattungskonzepts: Der Kompaktbegriff Gattung." In *Gattungstheorie und Gattungsgeschichte*, edited by Marion Gymnich, Birgit Neumann, and Ansgar Nünning, 31–52. Trier: WVT.

Gymnich, Marion, Birgit Neumann, and Ansgar Nünning, eds. 2007. *Gattungstheorie und Gattungsgeschichte.* Trier: WVT.

HathiTrust. 2008–2023. "HathiTrust Digital Library." https://www.hathitrust.org/. Accessed March 28, 2023.

Hempfer, Klaus W. 1973. *Gattungstheorie. Information und Synthese.* München: Fink.

Hempfer, Klaus W. 2014. "Some Aspects of a Theory of Genre." In *Linguistics and Literary Studies/Linguistik und Literaturwissenschaft. Interfaces, Encounters, Transfers/Begegnungen, Interferenzen und Kooperationen*, edited by Monika Fludernik and Daniel Jacob, 405–422. Berlin: De Gruyter.

Henny, Ulrike, and Frederike Neuber. 2017. "Criteria for Reviewing Digital Text Collections, version 1.0." Institut für Dokumentologie und Editorik. https://web.archive.org/web/20230418162046/https://www.i-d-e.de/publikationen/weitereschriften/criteria-text-collections-version-1-0/.

Henny-Krahmer, Ulrike. 2017. "Bib-ACMé: Bibliografía digital de novelas argentinas, cubanas y mexicanas (1810–1930)." In *III Congreso de la Sociedad Internacional Humanidades Digitales Hispánicas. Sociedades, políticas, saberes (Libro de resúmenes)*, edited by Nuria Rodríguez Ortega, 99–104. Málaga: Universidad de Málaga. https://web.archive.org/web/20200514082600/https://humanidadesdigitaleshispanicas.es/wp-content/uploads/2020/03/Actas-HDH2017.pdf.

Henny-Krahmer, Ulrike, ed. 2017–2021. "Bib-ACMé. Bibliografía digital de novelas argentinas, cubanas y mexicanas (1830–1910)." Version 1.2. Zenodo. https://doi.org/10.5281/zenodo.4453491.

Henny-Krahmer, Ulrike. 2018. "Exploration of Sentiments and Genre in Spanish American Novels." In *Digital Humanities 2018. Puentes–Bridges. Book of Abstracts. Mexico City, 26–29 June 2018*, 399–403. Mexico City: Red de Humanidades Digitales. https://web.archive.org/web/20200702225303/https://dh2018.adho.org/exploration-of-sentiments-and-genre-in-spanish-american-novels/.

Henny-Krahmer, Ulrike, ed. 2021a. "Corpus de novelas hispanoamericanas del siglo XIX (conha19)." Version 1.0.1. Zenodo. https://doi.org/10.5281/zenodo.4766987.

Henny-Krahmer, Ulrike. 2021b. "Data accompanying the dissertation 'Genre analysis and corpus design: 19th century Spanish American novels (1830–1910)'." Version 1.0.0. Zenodo. http://doi.org/10.5281/zenodo.4451928.

Henny-Krahmer, Ulrike. 2021c. "Features for the classification of Spanish American 19th century novels by subgenre." Version 1.0.0. Zenodo. http://doi.org/10.5281/zenodo.4449494.

Henny-Krahmer, Ulrike. 2021d. "Scripts accompanying the dissertation 'Genre analysis and corpus design: 19th century Spanish American novels (1830–1910)'." Version 1.0.0. Zenodo. https://doi.org/10.5281/zenodo.4445877.

Henny-Krahmer, Ulrike. 2022. "Novelas originales y americanas. A Digital Analysis of References to Identity in Subtitles of Spanish American 19th Century Novels." *apropos [Perspektiven auf die Romania]* 9: 14–36. https://doi.org/10.15460/apropos.9.1893.

Henny-Krahmer, Ulrike, Katrin Betz, Daniel Schlör, and Andreas Hotho. 2018. "Alternative Gattungstheorien. Das Prototypenmodell am Beispiel hispanoamerikanischer Romane." In *DHd 2018. Kritik der digitalen Vernunft. Konferenzabstracts, Köln, 26.2.-2.3.2018*, edited by Georg Vogeler, 105–112. Köln: Universität zu Köln. https://doi.org/10.5281/zenodo.4622412.

Henny-Krahmer, Ulrike, and Christof Schöch. 2016. "How good are our texts, really? Quality assurance for literary texts from various sources." CLiGS – Computergestützte literarische Gattungsstilistik. https://web.archive.org/web/20230422152455/https://cligs.hypotheses.org/371.

Herrmann, J. Berenike, Christof Schöch, and Karina van Dalen-Oskam. 2015. "Revisiting Style, a Key Concept in Literary Studies." *Journal of Literary Theory* 9 (1): 25–52. https://doi.org/10.1515/jlt-2015-0003.

Hesselbach, Robert, José Calvo Tello, Ulrike Henny-Krahmer, Christof Schöch, and Daniel Schlör, eds. Forthcoming. *Digital Stylistics in Romance Studies and Beyond.* Heidelberg: Heidelberg University Publishing.

Hettinger, Lena, Martin Becker, Isabella Reger, Fotis Jannidis, and Andreas Hotho. 2015. "Genre Classification on German Novels." In *Proceedings of the 26th International Workshop on Database and Expert Systems Applications (DEXA)*, 249–253. Valencia. https://doi.org/10.1109/DEXA.2015.62.

Hettinger, Lena, Isabella Reger, Fotis Jannidis, and Andreas Hotho. 2016. "Classification of Literary Subgenres." In *DHd2016. Modellierung – Vernetzung – Visualisierung. Die Digital Humanities als fächerübergreifendes Forschungsparadigma. Konferenzabstracts. Universität Leipzig 7. bis 12. März 2016*, 160–164. Duisburg: nisaba verlag. https://doi.org/10.5281/zenodo.4645368.

Holmes, David I. 1998. "The Evolution of Stylometry in Humanities Scholarship." *Literary and Linguistic Computing* 13 (3): 111–117. https://doi.org/10.1093/llc/13.3.111.

Hoover, David L., Jonathan Culpeper, and Kieran O'Halloran. 2014. *Digital Literary Studies: Corpus Approaches to Poetry, Prose and Drama.* New York, London: Routledge.

Horstmann, Jan. 2018. "Topic Modeling." forTEXT. Literatur digital erforschen. https://web.archive.org/web/20230316111225/https://fortext.net/routinen/methoden/topic-modeling.

Ianes, Raúl. 2018. "La ficción de la lengua en las Novelas argentinas de Carlos María Ocantos: una lectura histórica." *Decimonónica* 15 (2): 14–28. https://web.archive.org/web/20230328191041/https://www.decimononica.org/wp-content/uploads/2018/07/Ianes_15.2.pdf.

Iguiniz, Juan Bautista. 1926. *Bibliografía de novelistas mexicanos: ensayo biográfico, bibliográfico y crítico.* México: Imprenta de la Secretaria de relaciones exteriores.

Íñigo Madrigal, Luis, Manuel Alvar, and Fernando Aínsa, eds. 1982. *Historia de la literatura hispanoamericana.* 3 vols. Madrid: Cátedra.

Instituto de Investigaciones Bibliográficas. n.d. "Módulo de búsqueda." Bibliografía Mexicana Siglo XIX. https://web.archive.org/web/20230603165352/http://bd.iib.unam.mx/iib/proyectos/sigloxix/modulo.html.

Instituto de Literatura y Lingüística de la Academia de Ciencias de Cuba. 1999. *Diccionario de la literatura cubana (en formato HTML).* Alicante: Biblioteca Virtual Miguel de Cervantes. https://www.cervantesvirtual.com/nd/ark:/59851/bmckh0j1.

International Federation of Library Associations and Institutions (IFLA). 2009. *Functional Requirements for Bibliographic Records. Final report.* https://repository.ifla.org/handle/123456789/811.

Internet Archive. n.d. "Internet Archive." https://web.archive.org/web/20230603161417/https://archive.org/.

Jacobs, Jürgen. 1986. "Bildungsroman und Pikaroroman. Versuch einer Abgrenzung." In *Der moderne deutsche Schelmenroman. Interpretationen*, edited by Gerhart Hoffmeister, 9–18. Amsterdamer Beiträge zur neueren Germanistik, vol. 20. Amsterdam: Rodopi.

Janik, Dieter. 2008. *Hispanoamerikanische Literaturen. Von der Unabhängigkeit bis zu den Avantgarden (1810–1930).* Tübingen: Narr Francke Attempto.

Jannidis, Fotis. 2007. "Computerphilologie." In *Handbuch Literaturwissenschaft. Gegenstände – Konzepte – Institutionen*, edited by Thomas Anz, vol. 2, *Methoden und Theorien*, 27–40. Stuttgart, Weimar: J.B. Metzler.

Jannidis, Fotis. 2010. "Methoden der computergestützten Textanalyse." In *Methoden der literatur- und kulturwissenschaftlichen Textanalyse*, edited by Ansgar Nünning and Vera Nünning, 109–132. Stuttgart, Weimar: J.B. Metzler.

Jannidis, Fotis. 2016. "Quantitative Analyse literarischer Texte am Beispiel des Topic Modeling." *Der Deutschunterricht* 68 (5): 24–35.

Jannidis, Fotis, Leonard Konle, and Peter Leinen. 2019. "Makroanalytische Untersuchung von Heftromanen." In *DHd2019. Digital Humanities: multimedial & multimodal. Konferenzabstracts, Universitäten zu Mainz und Frankfurt, 25. bis 29. März 2019*, edited by Patrick Sahle, 167–173. Frankfurt & Mainz: Verband Digital Humanities im deutschsprachigen Raum e.V. https://doi.org/10.5281/zenodo.4622093.

Jannidis, Fotis, Leonard Konle, Albin Zehe, Andreas Hotho, and Markus Krug. 2018. "Analysing Direct Speech in German Novels." In DHd 2018. *Kritik der digitalen Vernunft. Konferenzabstracts, Köln, 26.2.-2.3.2018*, edited by Georg Vogeler, 114–118. Köln: Universität zu Köln. https://doi.org/10.5281/zenodo.4622454.

Jannidis, Fotis, Markus Krug, Martin Toepfer, Frank Puppe, Isabella Reger, and Lukas Weimer. 2015. "Automatische Erkennung von Figuren in deutschsprachigen Romanen." In *DHd2015. Konferenzabstracts.* https://doi.org/10.5281/zenodo.4623273.

Jauß, Hans Robert. 2014. "Theory of Genres and Medieval Literature." In *Modern Genre Theory*, edited by David Duff, 127–147. New York: Routledge.

Javed, Muhammad Aqib, Muhammad Shahzad Younis, Siddique Latif, Junaid Qadir, and Adeel Baig. 2018. "Community detection in networks: A multidisciplinary review." *Journal of Network and Computer Applications* 108: 87–111. https://doi.org/10.1016/j.jnca.2018.02.011.

Jeffries, Lesley, and Daniel McIntyre. 2010. *Stylistics.* Cambridge: Cambridge University Press.

Jockers, Matthew L. 2013. *Macroanalysis. Digital Methods & Literary History.* Topics in the Digital Humanities. Urbana, Chicago, and Springfield: University of Illinois Press.

Juola, Patrick. 2006. *Authorship Attribution.* Boston, Mass.: Now Publishers.

Kahle, Günther. 1993. *Lateinamerika Ploetz. Die Geschichte der lateinamerikanischen Länder zum Nachschlagen.* 2nd ed. Freiburg/Würzburg: Ploetz.

Kaiser, Dorothee. 2002. *Wege zum wissenschaftlichen Schreiben. Eine kontrastive Untersuchung zu studentischen Texten aus Venezuela und Deutschland.* Tübingen: Stauffenburg-Verlag.

Kaiser, Dorothee. 2008. "Ensayo o artículo científico? Una comparación de tradiciones discursivas en Alemania y Latinoamérica." In *Le style, c'est l'homme: unité et pluralité du discours scientifique dans les langues romanes*, edited by Ursula Reutner, 285–304. Frankfurt am Main: Lang.

Keckeis, Paul, and Werner Michler. 2020. "Einleitung: Gattungen und Gattungstheorie." In *Gattungstheorie*, edited by Paul Keckeis and Werner Michler, 7–48. Berlin: Suhrkamp.

Kessler, Brett, Geoffrey Numberg, and Hinrich Schütze. 1997. "Automatic detection of text genre." In *ACL '98/EACL '98: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, 32–38. Stroudsburg, PA, USA: Association for Computational Linguistics. http://dx.doi.org/10.3115/976909.979622.

Kim, Evgeny, Sebastian Padó, and Roman Klinger. 2017. "Prototypical Emotion Developments in Literary Genres." *DH2017. Conference Abstracts.* Montréal: McGill University & Université de Montréal. https://web.archive.org/web/20230211105146/https://dh2017.adho.org/abstracts/203/203.pdf.

Klauk, Tobias, and Tilmann Köppe. 2014. "Bausteine einer Theorie der Fiktionalität." In *Fiktionalität. Ein interdisziplinäres Handbuch*, edited by Tobias Klauk and Tilmann Köppe, 3–31. Berlin, Boston: De Gruyter.

Klausnitzer, Ralf, and Guido Naschert. 2007. "Gattungstheoretische Kontroversen? Konstellationen der Diskussion von Textordnungen im 20. Jahrhundert." In *Kontroversen in der Literaturtheorie – Literaturtheorie in der Kontroverse*, edited by Ralf Klausnitzer and Carlos Spoerhase, 369–412. Bern: Peter Lang.

Kleinschmidt, Erich. 2003. "Prosa." In *Reallexikon der deutschen Literaturwissenschaft*, edited by Klaus Weimar, Harald Fricke, and Jan-Dirk Müller, 168–172. Berlin, New York: De Gruyter.

Kohut, Karl. 2016. *Kurze Einführung in Theorie und Geschichte der lateinamerikanischen Literatur (1492–1920).* Berlin: Lit Verlag.

Kompetenzzentrum – Trier Center for Digital Humanities. 2023. "Computational Literary Studies. A Bird's Eye View of Literature." https://web.archive.org/web/20230210111714/https://tcdh.uni-trier.de/en/thema/computational-literary-studies.

Konle, Leonard. 2019. "Word Embeddings für literarische Texte." Master's thesis, Würzburg: Julius-Maximilians-Universität Würzburg. https://web.archive.org/web/20230305090725/https://lekonard.github.io/blog/Konle_Thesis.pdf.

Köppe, Tilmann. 2014. "Die Institution Fiktionalität." In *Fiktionalität. Ein interdisziplinäres Handbuch*, edited by Tobias Klauk and Tilmann Köppe, 35–49. Berlin, Boston: De Gruyter.

Krieg-Holz, Ulrike, and Lars Bülow. 2016. *Linguististische Stil- und Textanalyse: eine Einführung.* Tübingen: Narr Francke Attempto.

Lacey, Nick. 2000. *Narrative and Genre: Key Concepts in Media Studies.* Basingstoke: Macmillan.

Lamping, Dieter, ed. 1990. *Gattungstheorie und Gattungsgeschichte: ein Symposium.* Wuppertal: Bergische Universität, Gesamthochschule Wuppertal.

Lamping, Dieter. 2010. "Komparatistische Gattungsforschung." In *Handbuch Gattungstheorie*, edited by Rüdiger Zymner, 270–273. Stuttgart, Weimar: J.B. Metzler.

Landauer, Thomas K., and Susan T. Dumais. 1997. "A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge." *Psychological Review* 104: 211–240. https://psycnet.apa.org/doi/10.1037/0033-295X.104.2.211.

Landauer, Thomas K., Peter W. Foltz, and Darrell Laham. 1998. "Introduction to latent semantic analysis." *Discourse Processes* 25: 259–284. https://doi.org/10.1080/01638539809545028.

Lawrence, John, and Chris Reed. 2017. "Mining Argumentative Structure from Natural Language text using Automatically Generated Premise-Conclusion Topic Models." In *Proceedings of the 4th Workshop on Argument Mining*, 39–48. Copenhagen: Association for Computational Linguistics (ACL). http://dx.doi.org/10.18653/v1/W17-5105.

Leech, Geoffrey, and Mick Short. 2007. *Style in Fiction. A Linguistic Introduction to English Fictional Prose.* 2nd ed. Harlow, England: Pearson Education Limited.

Lefere, Robin. 2013. *La novela histórica: (re)definición, caracterización, tipología.* Madrid: Visor Libros.

Lichtblau, Myron I. 1959. *The Argentine Novel in the Nineteenth Century.* New York: Hispanic Institute in the United States.

Lichtblau, Myron. 1997. *The Argentine novel: an annotated bibliography.* Lanham, Maryland: Scarecrow.

Lindstrom, Naomi. 2004. *Early Spanish American Narrative.* Austin: University of Texas Press.

Löfquist, Eva. 1995. *La novela histórica chilena dentro del marco de la novelística chilena. 1843–1879.* Göteborg: Acta Universitatis Gothoburgensis.

Lüdeling, Anke, and Merja Kytö, eds. 2008. *Corpus Linguistics: An International Handbook.* 2 vols. Handbooks of Linguistics and Communication Science (HSK). Berlin, Boston: De Gruyter.

Lukács, Georg. 1955. *Der Historische Roman.* Berlin: Aufbau-Verlag.

Madjarov, Gjorji, Dragi Kocev, Dejan Gjorgjevikj, and Sašo Džeroski. 2012. "An extensive experimental comparison of methods for multi-label learning." *Pattern Recognition* 45 (9): 3084–3104. https://doi.org/10.1016/j.patcog.2012.03.004.

Manning, Christopher D., and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing.* Cambridge, Mass: The MIT Press.

Mansilla, Lucio Victorio. (1870) 2001. *Una excursión a los indios ranqueles. Tomo Primero (en formato HTML).* Alicante: Biblioteca Virtual Miguel de Cervantes. https://www.cervantesvirtual.com/nd/ark:/59851/bmcn8760.

Martínez Cantón, Clara Isabel. 2008. "El indigenismo en la obra de Vargas Llosa." *Espéculo. Revista de estudios literarios* 38. https://web.archive.org/web/20210226164843/https://webs.ucm.es/info/especulo/numero38/vllindig.html.

Mata, Óscar. 1999. *La novela corta mexicana en el siglo XIX.* México: Universidad Nacional Autónoma de México.

Maxwell, Richard. 2009. *The Historical Novel in Europe, 1650–1950.* Cambridge: Cambridge University Press.

McCallum, Andrew. 2002. "MALLET: A Machine Learning for Language Toolkit." Accessed November 13, 2020. http://mallet.cs.umass.edu.

McCallum, Andrew. 2018a. "Topic model diagnostics." *MALLET: A Machine Learning for Language Toolkit.* https://web.archive.org/web/20200221035417/https://mallet.cs.umass.edu/diagnostics.php.

McCallum, Andrew. 2018b. "Topic modeling." *MALLET: A Machine Learning for Language Toolkit.* https://web.archive.org/web/20201112052435/http://mallet.cs.umass.edu/topics.php.

Meléndez, Concha. 1961. *La novela indianista en Hispanoamerica (1832–1889).* Rio Piedras: Universidad de Puerto Rico.

Meyer-Minnemann, Klaus. 1979. *Der spanischamerikanische Roman des Fin de siècle.* Tübingen: Niemeyer.

Miller, George A. 1995. "WordNet: A Lexical Database for English." *Communications of the ACM* 38 (11): 39–41. https://doi.org/10.1145/219717.219748.

Milling, Carsten, Frank Fischer, and Mathias Göbel, eds. 2021. "French Drama Corpus (FreDraCor): A TEI P5 Version of Paul Fièvre's 'Théâtre Classique' Corpus." GitHub.com. Accessed December 9, 2022. https://github.com/dracor-org/fredracor.

Mitjans, Aurelio. (1918) 2010. *Historia de la Literatura cubana (en formato HTML).* Biblioteca Virtual Miguel de Cervantes. https://www.cervantesvirtual.com/nd/ark:/59851/bmc1g0w8.

Molina, Hebe Beatriz. 2011. *Como crecen los hongos. La novela argentina entre 1838 y 1872.* Buenos Aires: Teseo.

Molina, Sintia. 2001. *El Naturalismo en la novela cubana.* Lanham, Maryland: University Press of America.

Moody, Christopher E. 2016. "Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec." arXiv.org. https://doi.org/10.48550/arXiv.1605.02019.

Moretti, Franco. 2008. "The Novel: History and Theory." *New Left Review* 52: 111–124. Accessed January 28, 2023. https://newleftreview.org/issues/ii52/articles/franco-moretti-the-novel-history-and-theory.

Müller, Ralph. 2010. "Kategorisieren." In *Handbuch Gattungstheorie*, edited by Rüdiger Zymner, 21–23. Stuttgart, Weimar: J.B. Metzler.

Müller, Andreas C., and Sarah Guido. 2016. *Introduction to Machine Learning with Python: a Guide for Data Scientists.* Sebastopol, CA: O'Reilly.

Murata, Makoto. 2014. "RELAX NG home page." https://web.archive.org/web/20230604105524/https://relaxng.org/.

National Science Board, ed. 2005. "Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century." National Science Foundation. https://web.archive.org/web/20230207100814/https://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf.

Navarro, Joaquina. 1955. *La novela realista mexicana.* México: Compañía General de Ediciones.

Navarro-Colorado, Borja, ed. 2020. "Corpus of Spanish Golden-Age Sonnets." Version 1.0.0. GitHub.com. Accessed December 6, 2022. https://github.com/bncolorado/CorpusSonetosSigloDeOro.

Navarro-Colorado, Borja, María Ribes Lafoz, and Noelia Sánchez. 2016. "Metrical annotation of a large corpus of Spanish sonnets: representation, scansion and evaluation." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 4360–4364. Portorož, Slovenia: European Language Resources Association (ELRA). http://web.archive.org/web/20220315081224/https://aclanthology.org/L16-1691.pdf.

Neumann, Birgit, and Ansgar Nünning. 2007. "Einleitung: Probleme, Aufgaben und Perspektiven der Gattungstheorie und Gattungsgeschichte." In *Gattungstheorie und Gattungsgeschichte*, edited by Marion Gymnich, Birgit Neumann, and Ansgar Nünning, 1–28. Trier: WVT.

Neuschäfer, Hans-Jörg, ed. 2001. *Spanische Literaturgeschichte.* 2nd ed. Stuttgart/Weimar: J.B. Metzler.

Nielsen, Lars Holm. 2013. "ZENODO - An Innovative Service for Sharing All Research Outputs." Talk presented at the Joint OpenAIRE/LIBER Workshop, Ghent. http://dx.doi.org/10.5281/zenodo.6815.

Oakes, Michael P. 2003. *Statistics for corpus linguistics.* Edinburgh: Edinburgh Univ. Press.

Oakes, Michael P. 2009. "Corpus Linguistics and Stylometry." In *Corpus Linguistics*, edited by Anke Lüdeling and Merja Kytö, 1070–1090. Vol. 2. Berlin: De Gruyter. https://doi.org/10.1515/9783110213881.2.1070.

Ocantos, Carlos María. (1913) 2007. "Quilito." Project Gutenberg. https://web.archive.org/web/20230422145502/https://www.gutenberg.org/files/23035/23035-h/23035-h.htm.

Odebrecht, Carolin, Lou Burnard, and Christof Schöch, eds. 2021. "European Literary Text Collection (ELTeC)." Version 1.1.0. COST Action Distant Reading for European Literary History (CA16204). https://doi.org/10.5281/zenodo.4662444.

Olea, Ismael. 2021. "Lemarios y listas de palabras del español." GitHub.com. https://web.archive.org/web/20230609200732/https://github.com/olea/lemarios.

Olea Franco, Rafael. 2011. "Narrativa e identidad hispanoamericanas. De Fernández de Lizardi a Borges." In *La literatura hispanoamericana*, edited by Julio Ortega, 23–134. La búsqueda perpetua: lo propio y lo universal de la cultura latinoamericana 3. México: Secretaría de Relaciones Exteriores, Dirección General del Acervo Histórico Diplomático.

OCLC. 2001–2023. "WorldCat." https://www.worldcat.org/de. Accessed March 28, 2023.

OCLC. 2010–2021a. "Iglesia, Álvaro de la, 1859–1940." VIAF. Virtual International Authority File. https://viaf.org/viaf/120788045/.

OCLC. 2010–2021b. "VIAF. Virtual International Authority File." https://web.archive.org/web/20230423111630/https://viaf.org/.

OCLC. 2023. "Inside WorldCat." https://web.archive.org/web/20230325164614/https://www.oclc.org/en/worldcat/inside-worldcat.html.

Padró, Lluís. n.d.a. "FreeLing Home Page." https://web.archive.org/web/20230610172727/https://nlp.lsi.upc.edu/freeling/.

Padró, Lluís. n.d.b. "Linguistic Data." FreeLing Home Page. https://web.archive.org/web/20230610173053/https://nlp.lsi.upc.edu/freeling/index.php/node/12.

Padró, Lluís. n.d.c. "Multiword Recognition Module." FreeLing 4.0 User Manual. https://web.archive.org/web/20230610173340/https://freeling-user-manual.readthedocs.io/en/v4.0/modules/locutions/.

Padró, Lluís. n.d.d. "Tagset for Spanish (es)." FreeLing 4.0 User Manual. https://web.archive.org/web/20230610173624/https://freeling-user-manual.readthedocs.io/en/v4.0/tagsets/tagset-es/.

Padró, Lluís. n.d.e. "Using analyzer Program to Process Corpora." FreeLing 4.0 User Manual. https://web.archive.org/web/20230610173731/https://freeling-user-manual.readthedocs.io/en/v4.0/analyzer/.

Padró, Lluís, and Evgeny Stanislovsky. 2012. "FreeLing 3.0: Towards Wider Multilinguality." In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA*, 2473–2479. Istanbul, Turkey: ELRA. https://web.archive.org/web/20230610172457/http://www.lrec-conf.org/proceedings/lrec2012/pdf/430_Paper.pdf.

Paz, Ireneo. (1883) 2017. *Doña Marina.* Berlin: Ibero-Amerikanisches Institut – Preußischer Kulturbesitz. http://resolver.iai.spk-berlin.de/IAI00006A0B00000000.

Peñaranda Medina, Rosario. 1994. *La novela modernista hispanoamericana.* Valencia: Universitat de Valencia.

Phillips-López, Dolores. 1996. *La novela hispanoamericana del modernismo.* Genève: Ed. Slatkine.

Pi-Suñer Llorens, Antonia. 2005. "Entre la historia y la novela. Ireneo Paz." In *La república de las letras. Asomos a la cultura escrita del México decimonónico*, edited by Belem Clark de Lara and Elisa Speckman Guerra, 379–392. Vol. 3: Galería de escritores. México: UNAM.

Portillo, Andrés. (1896) 2020. *María Luisa. Leyenda histórica.* Würzburg: CLiGS. Accessed January 28, 2023. https://github.com/cligs/conha19/blob/master/tei/nh0100.xml.

Prendes, Manuel. 2003. *La novela naturalista hispanoamericana. Evolución y direcciones de un proceso narrativo.* Madrid: Ediciones Cátedra.

Princeton University. 2023. "lexnames(5WN)." WordNet. A Lexical Database for English. https://web.archive.org/web/20230610175939/https://wordnet.princeton.edu/documentation/lexnames5wn.

Raible, Wolfgang. 1980. "Was sind Gattungen? Eine Antwort aus semiotischer und textlinguistischer Sicht." *Poetica* 12: 320–349.

Ramírez, José María. (1868) 2008. "Full text of 'Una rosa y un harapo: Novela original'." Internet Archive. https://web.archive.org/web/20230328190153/https://archive.org/stream/unarosayunharap00ramgoog/unarosayunharap00ramgoog_djvu.txt.

Read, John Lloyd. 1939. *The Mexican Historical Novel. 1826–1910.* New York: Instituto de las Españas en los Estados Unidos.

Real Academia Española (RAE). 2023a. "Diccionario de la lengua española. (DLe)." Accessed June 11, 2023. https://dle.rae.es/.

Real Academia Española (RAE). 2023b. "suriano, na." Diccionario de la lengua española (DLe). http://web.archive.org/web/20230601151138/https://dle.rae.es/suriano.

Real Academia Española (RAE). 2023c. "tapatío, a." Diccionario de la lengua española (DLe). http://web.archive.org/web/20230601151450/https://dle.rae.es/tapat%C3%ADo.

Remos y Rubio, Juan J. 1935. *Tendencias de la narración imaginativa en Cuba.* La Habana: Casa Montalvo-Cárdenas. https://dloc.com/UF00078289/00001/images.

Remos y Rubio, Juan J. 1945. *Historia de la literatura cubana.* Vol. 2: Romanticismo. La Habana: Cárdenas y Compañía.

Rhody, Lisa M. 2012. "Topic Modeling and Figurative Language." *Journal of Digital Humanities* 2 (1). https://web.archive.org/web/20230316135657/https://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/.

Rinas, Karsten. 2015. "Zum linguistischen Status des Absatzes." *Aussiger Beiträge: germanistische Schriftenreihe aus Forschung und Lehre* 9: 139–157.

Rivas, Mercedes. 1990. *Literatura y esclavitud en la novela cubana del siglo XIX.* Sevilla: Escuela de Estudios Hispano-Americanos.

Rojas Mix, Miguel. 1987. "La cultura hispanoamericana del siglo XIX." In *Historia de la Literatura Hispanoamericana*, edited by Luis Íñigo Madrigal, 55–74. Vol. 2: Del neoclasicismo al modernismo. Madrid: Ediciones Cátedra.

Romanos de Tiratel, Susana. 2004. "La bibliografía nacional Argentina: una deuda pendiente." In *Proceedings of the 70th IFLA General Conference and Council*, 1–11. Buenos Aires. https://web.archive.org/web/20230603163155/https://archive.ifla.org/IV/ifla70/papers/046s_Tiratel.pdf.

Rosch, Eleanor, and Carolyn B. Mervis. 1975. "Family Resemblances: Studies in the Internal Structure of Categories." *Cognitive Psychology* 7: 573–605.

Rosell, Sara V. 1997. *La novela antiesclavista en Cuba y Brasil, siglo XIX.* Madrid: Ed. Pliegos.

Rössner, Michael. 2007. *Lateinamerikanische Literaturgeschichte.* 3rd ed. Stuttgart, Weimar: J.B. Metzler.

Ruhl, Klaus-Jörg, and Laura Ibarra García. 2000. *Kleine Geschichte Mexikos. Von der Frühzeit bis zur Gegenwart.* München: C. H. Beck.

Sahle, Patrick. 2013. *Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels.* Vol. 2: Befunde, Theorie und Methodik. Schriften des Instituts für Dokumentologie und Editorik 8. Norderstedt: BoD.

Sahlgren, Magnus. 2008. "The Distributional Hypothesis." *Rivista di Linguistica* 20 (1): 33–53. https://web.archive.org/web/20230310101109/https://www.italian-journal-linguistics.com/app/uploads/2021/05/Sahlgren-1.pdf.

Sahlgren, Magnus. 2015. "A brief history of word embeddings (and some clarifications)." *Linked in.* https://web.archive.org/web/20230310102225/https://www.linkedin.com/pulse/brief-history-word-embeddings-some-clarifications-magnus-sahlgren.

Sáinz de Medrano, Luis. 1987. "Cirilo Villaverde." In *Historia de la Literatura Hispanoamericana*, edited by Luis Íñigo Madrigal, 145–153. Vol. 2: Del neoclasicismo al modernismo. Madrid: Ediciones Cátedra.

Sánchez, Luis Alberto. 1953. *Proceso y contenido de la novela hispano-americana.* Madrid: Editorial Gredos.

Sánchez Mármol, Manuel. 2011. *Obras completas I: novelas.* Edited by Manuel Sol. Colección Manuel Sánchez Mármol. Villahermosa: Universidad Juárez Autónoma de Tabasco.

Sapkota, Upendra, Steven Bethard, Manuel Montes-y-Gómez, and Thamar Solorio. 2015. "Not All Character N-grams Are Created Equal: A Study in Authorship Attribution." In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 93–102. Denver, Colorado: Association for Computational Linguistics. http://dx.doi.org/10.3115/v1/N15-1010.

Sarmiento, Domingo Faustino. (1845) 2000. *Vida de Juan Facundo Quiroga (en formato HTML).* Edited by Benito Varela Jácome. Alicante: Biblioteca Virtual Miguel de Cervantes. https://www.cervantesvirtual.com/nd/ark:/59851/bmc18359.

Saxonica. n.d. "Running XSLT from the Command Line." Saxonica. XSLT and XQuery Processing. https://web.archive.org/web/20230610171712/https://www.saxonica.com/html/documentation12/using-xsl/commandline/.

Schaeffer, Jean-Marie. 1983. *Qu'est-ce qu'un genre littéraire?* Paris: Seuil.

Schlickers, Sabine. 2003. *El lado oscuro de la modernización: estudios sobre la novela naturalista hispanoamericana.* Madrid, Frankfurt: Iberoamericana/Vervuert.

Schmid, Helmut. 1994. "Probabilistic Part-of-Speech Tagging Using Decision Trees." In *Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK*, 44–49. https://web.archive.org/web/20230603115230/https://www.cis.uni-muenchen.de/~{}schmid/tools/TreeTagger/data/tree-tagger1.pdf.

Schmitz-Emans, Monika. 2010. "Écriture und Gattung." In *Handbuch Gattungstheorie*, edited by Rüdiger Zymner, 107–109. Stuttgart: J.B. Metzler.

Schnur-Wellpott, Margrit. 1983. *Aporien der Gattungstheorie aus semiotischer Sicht.* Tübingen: Narr.

Schöch, Christof. 2013. "Fine-tuning Stylometric Tools: Investigating Authorship and Genre in French Classical Theater." In *Digital Humanities 2013. Conference Abstracts, Lincoln, NE, USA, July 16–19*, 383–386. Lincoln, NE, USA: University of Nebraska-Lincoln. https://web.archive.org/web/20230304104934/http://dh2013.unl.edu/abstracts/ab-270.html.

Schöch, Christof. 2016. "Topic Modeling with MALLET: Hyperparameter Optimization." The Dragonfly's Gaze. https://web.archive.org/web/20230316145457/https://dragonfly.hypotheses.org/1051.

Schöch, Christof. 2017a. "Aufbau von Datensammlungen." In *Digital Humanities. Eine Einführung*, edited by Fotis Jannidis, Hubertus Kohle, and Malte Rehbein, 223–233. Stuttgart: J.B. Metzler.

Schöch, Christof, ed. 2017b. "theatreclassique." Accessed December 9, 2022. https://github.com/cligs/theatreclassique.

Schöch, Christof. 2017c. "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama." *Digital Humanities Quarterly* 11 (2). https://web.archive.org/web/20230211105751/http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html.

Schöch, Christof. 2018. "Zeta für die kontrastive Analyse literarischer Texte. Theorie, Implementierung, Fallstudie." In *Quantitative Ansätze in den Literatur- und Geisteswissenschaften*, edited by Toni Bernhart, Marcus Willand, Sandra Richter, and Andrea Albrecht, 77–94. Berlin, München, Boston: De Gruyter. https://doi.org/10.1515/9783110523300-004.

Schöch, Christof. 2019. "Distributional Semantics and Topic Modeling: Theory and Application." Workshop given at the *Baltic Summer School of Digital Humanties: Essentials of Coding and Encoding*, Riga, July 2019. Accessed November 14, 2020. https://christofs.github.io/riga/#/.

Schöch, Christof, José Calvo Tello, Ulrike Henny-Krahmer, and Stefanie Popp, eds. 2018. "The CLiGS textbox." Version 4.0.0. Zenodo. https://doi.org/10.5281/zenodo.597430.

Schöch, Christof, José Calvo Tello, Ulrike Henny-Krahmer, and Stefanie Popp. 2019. "The CLiGS Textbox: Building and Using Collections of Literary Texts in Romance Languages Encoded in TEI XML." *Journal of the Text Encoding Initiative.* Rolling Issue. https://doi.org/10.4000/jtei.2085.

Schöch, Christof, Ulrike Henny, José Calvo Tello, Katrin Betz, and Daniel Schlör. 2017. "Epochenschwellen als Phasen beschleunigter literarischer Entwicklung?" Talk presented at the *Forum Junge Romanistik*, Göttingen, March, 2017. Accessed March 3, 2023. https://christofs.github.io/fjr17/#.

Schöch, Christof, Ulrike Henny, José Calvo Tello, Daniel Schlör, and Stefanie Popp. 2016. "Topic, Genre, Text. Topics im Textverlauf von Untergattungen des spanischen und hispanoamerikanischen Romans (1880–1930)." In *DHd 2016. Modellierung, Vernetzung, Visualisierung. Die Digital Humanities als fächerübergreifendes Forschungsparadigma. Konferenzabstracts*, 235–239. Leipzig: Universität Leipzig. https://doi.org/10.5281/zenodo.4645380.

Schöch, Christof, and Steffen Pielström. 2014a. "Für eine computergestützte literarische Gattungsstilistik." *DHd2014. Konferenzabstracts.* Passau: Universität Passau. https://doi.org/10.5281/zenodo.4623620.

Schöch, Christof, and Steffen Pielström. 2014b. "Die Principal Component Analysis für die Differenzierung von Autorschaft, Form und Gattung literarischer Texte." Talk presented at the *<philtag n=12/>*, University of Würzburg, September 19, 2014.

Schöch, Christof, and Daniel Schlör. 2017. "tmw – Topic Modeling Workflow." GitHub. Accessed November 14, 2020. https://github.com/cligs/tmw.

Schöch, Christof, Daniel Schlör, Stefanie Popp, Annelen Brunner, Ulrike Henny, and José Calvo Tello. 2016. "Straight Talk! Automatic Recognition of Direct Speech in Nineteenth-Century French Novels." In *Digital Humanities 2016: Conference Abstracts*, 346–353. Kraków: Jagiellonian University and Paedagocial University. https://web.archive.org/web/20230325081511/https://dh2016.adho.org/abstracts/31.

Schöch, Christof, Daniel Schlör, Albin Zehe, Henning Gebhard, Martin Becker, and Andreas Hotho. 2018. "Burrows' Zeta: Exploring and Evaluating Variants and Parameters." In *Digital Humanities 2018. Puentes–Bridges. Book of Abstracts. Mexico City, 26–29 June 2018*, 274–278. Mexico City: Red de Humanidades Digitales. https://web.archive.org/web/20230212045250/https://dh2018.adho.org/burrows-zeta-exploring-and-evaluating-variants-and-parameters/.

Schröter, Julian. 2019. "Gattungsgeschichte und ihr Gattungsbegriff am Beispiel der Novellen." *Journal of Literary Theory* 13 (2): 227–257.

Schröter, Julian. Forthcoming. "Machine-Learning as a Measure of the Conceptual Looseness of Disordered Genres: Studies on German Novellen." In *Digital Stylistics in Romance Studies and Beyond*, edited by Robert, Hesselbach, José Calvo Tello, Ulrike Henny-Krahmer, Christof Schöch, and Daniel Schlör. Heidelberg: Heidelberg University Publishing.

Schulz, Armin. 2007. "Thema." In *Reallexikon der deutschen Literaturwissenschaft*, edited by Harald Fricke, 634–635. Vol. 2. Berlin, New York: De Gruyter.

Scikit-learn developers. 2007–2023a. "Clustering." *Scikit-learn.* https://web.archive.org/web/20230304125710/https://scikit-learn.org/stable/modules/clustering.html.

Scikit-learn developers. 2007–2023b. "Clustering, sec. K-means." *Scikit-learn.* https://web.archive.org/web/20230304125710/https://scikit-learn.org/stable/modules/clustering.html.

Scikit-learn developers. 2007–2023c. "Common pitfalls in interpretation of coefficients of linear models." *Scikit-learn.* https://web.archive.org/web/20230304130148/https://scikit-learn.org/stable/auto_examples/inspection/plot_linear_model_coefficient_interpretation.html.

Scikit-learn developers. 2007–2023d. "Feature extraction, sec. The Bag of Words representation." *Scikit-learn.* https://web.archive.org/web/20230304131525/https://scikit-learn.org/stable/modules/feature_extraction.html.

Scikit-learn developers. 2007–2023e. "sklearn.ensemble.RandomForestClassifier." *Scikit-learn.* https://web.archive.org/web/20230304130404/https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

Scikit-learn developers. 2007–2023f. "sklearn.feature_extraction.text.CountVectorizer." *Scikit-learn.* https://web.archive.org/web/20230304130529/https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html.

Scikit-learn developers. 2007–2023g. "sklearn.feature_extraction.text.TfidfTransformer." *Scikit-learn.* https://web.archive.org/web/20230304130653/https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html.

Scikit-learn developers. 2007–2023h. "sklearn.model_selection.cross_validate." *Scikit-learn.* https://web.archive.org/web/20230304130816/https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_validate.html.

Scikit-learn developers. 2007–2023i. "sklearn.model_selection.GridSearchCV." *Scikit-learn.* https://web.archive.org/web/20230304131032/https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html.

Scikit-learn developers. 2007–2023j. "sklearn.neighbors.KNeighborsClassifier." *Scikit-learn.* https://web.archive.org/web/20230304131239/https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html.

Scikit-learn developers. 2007–2023k. "sklearn.svm.SVC." *Scikit-learn.* https://web.archive.org/web/20230304131430/https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html.

Scikit-learn developers. 2007–2023l. "Supervised learning." *Scikit-learn.* https://web.archive.org/web/20230304125409/https://scikit-learn.org/stable/supervised_learning.html.

Scikit-learn developers. 2007–2023m. "Support Vector Machines, sec. Scores and probabilites." *Scikit-learn.* https://web.archive.org/web/20230304123130/https://scikit-learn.org/stable/modules/svm.html.

Scott, Mike. 1997. "PC analysis of key words — and key key words." *System* 25 (2): 233–245.

Siegel, Erik. 2022. *Schematron. A language for Validating XML.* Denver: XML Press.

Siemens, Ray, and Susan Schreibman, eds. 2008. *A Companion to Digital Literary Studies.* Oxford: Blackwell.

Sierra, Justo. 2018. *Confesiones de un pianista.* Edited by Karla Ximena Salinas Gallegos. La novela corta. Una biblioteca virtual. Novelas en tránsito. Segunda serie. México: Universidad Nacional Autónoma de México. https://web.archive.org/web/20200322100041/http://www.lanovelacorta.com/novelas-en-transito-2/confesiones-de-un-pianista.pdf.

SIG-DLS. n.d. "Goals." Digital Literary Stylistics (SIG-DLS). http://web.archive.org/web/20221023111813/https://dls.hypotheses.org/activities/about/about.

Smith, Verity, ed. 1997. *Encyclopedia of Latin American Literature.* Chicago, Illinois: Fitzroy Dearborn Publishers.

Sommer, Doris. 1993. *Foundational Fictions. The National Romances of Latin America.* Berkeley: University of California Press.

Sowinski, Bernhard. 1999. *Stilistik: Stiltheorien und Stilanalysen.* Stuttgart: Metzler.

Spang, Kurt. 1998. "Apuntes para una definición de la novela histórica." In *La novela histórica. Teoría y comentarios.*, edited by Kurt Spang, Ignacio Arellano, and Carlos Mata, 63–125. 2nd ed. Pamplona: EUNSA. https://web.archive.org/web/20160504022949/http://www.culturahistorica.es/spang/novela_historica.pdf.

Sparrow de García Barrío, Constance. 1977. *The abolitionist novel in nineteenth century Cuba.* Baltimore: Morgan State College.

Spillner, Bernd. 2001. "Stilistik." In *Grundzüge der Literaturwissenschaft*, edited by Heinz Ludwig Arnold and Heinrich Detering, 234–256. 4th ed. München: dtv.

Stamatatos, Efstathios. 2009. "A Survey of Modern Authorship Attribution Methods." *Journal of the American Society of Information Science and Technology* 60 (3): 538–556. https://doi.org/10.1002/asi.21001.

Steinecke, Hartmut. 2007. "Roman." In *Reallexikon der deutschen Literaturwissenschaft*, edited by Klaus Weimar, Harald Fricke, and Jan-Dirk Müller, 317–323. Berlin, New York: De Gruyter.

Steyvers, Mark, and Tom Griffiths. 2007. "Probabilistic Topic Models." In *Handbook of latent semantic analysis*, edited by Thomas K. Landauer, Danielle S. McNamara, Simon Dennis, and Walter Kintsch, 427–448. Mahwah, NJ: Lawrence Erlbaum Associates. https://web.archive.org/web/20220927113904/https://cocosci.princeton.edu/tom/papers/SteyversGriffiths.pdf.

Strube, Werner. 1993. *Analytische Philosophie der Literaturwissenschaft. Untersuchungen zur literaturwissenschaftlichen Definition, Klassifikation, Interpretation und Textbewertung.* Paderborn: Schöningh.

Suárez-Murias, Marguerite C. 1963. *La novela romántica en Hispanoamérica.* New York: Hispanic Institute in the United States.

Taylor, John R. 2003. *Linguistic categorization.* 3rd ed. New York: Oxford University Press.

Text Encoding Initiative Consortium. n.d.a "Text Encoding Initiative." https://web.archive.org/web/20230423104650/https://tei-c.org/.

Text Encoding Initiative Consortium. n.d.b "Getting Started with P5 ODDs." https://web.archive.org/web/20230423104437/https://tei-c.org/favicon.ico.

Text Encoding Initiative Consortium. 2023a. "att.global.responsibility." In: *TEI P5: Guidelines for Electronic*

Text Encoding and Interchange, 839–840. Version 4.6.0. Revision f18deffba. https://web.archive.org/web/20230423102337/https://tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf.

Text Encoding Initiative Consortium. 2023b. "Components of Bibliographic References." In: *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, 146–164. Version 4.6.0. Revision f18deffba. https://web.archive.org/web/20230423102337/https://tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf.

Text Encoding Initiative Consortium. 2023c. "The TEI Header and Its Components." In: *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, 22–23. Version 4.6.0. Revision f18deffba. https://web.archive.org/web/20230423102337/https://tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf.

Text Encoding Initiative Consortium. 2023d. "<ab>." In: *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, 868–870. Version 4.6.0. Revision f18deffba. https://web.archive.org/web/20230423102337/https://tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf.

Text Encoding Initiative Consortium. 2023e. "<back>." In: *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, 933–936. Version 4.6.0. Revision f18deffba. https://web.archive.org/web/20230423102337/https://tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf.

Text Encoding Initiative Consortium. 2023f. "<floatingText>." In: *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, 1192–1194. Version 4.6.0. Revision f18deffba. https://web.archive.org/web/20230423102337/https://tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf.

Text Encoding Initiative Consortium. 2023g. "<front>." In: *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, 1206–1209. Version 4.6.0. Revision f18deffba. https://web.archive.org/web/20230423102337/https://tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf.

Text Encoding Initiative Consortium. 2023h. "<nationality>." In: *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, 1455–1457. Version 4.6.0. Revision f18deffba. https://web.archive.org/web/20230423102337/https://tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf.

Text Encoding Initiative Consortium. 2023i. "<quote>." In: *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, 1581–1583. Version 4.6.0. Revision f18deffba. https://web.archive.org/web/20230423102337/https://tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf.

Text Encoding Initiative Consortium. 2023j. "<textClass>." In: *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, 1782–1783. Version 4.6.0. Revision f18deffba. https://web.archive.org/web/20230423102337/https://tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf.

Text Encoding Initiative Consortium. 2023k. "<titlePage>." In: *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, 201–203. Version 4.6.0. Revision f18deffba. https://web.archive.org/web/20230423102337/https://tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf.

TextGrid, ed. n.d. "The Digital Library in the TextGrid Repository." TextGrid. Virtuelle Forschungsumgebung für die Geisteswissenschaften. https://web.archive.org/web/20221106162919/https://textgrid.de/en/digitale-bibliothek.

Theisen, Joachim. 2016. *Kontrastive Linguistik.* Tübingen: Narr.

Thomé, Horst. 2007. "Werk." In *Reallexikon der Deutschen Literaturwissenschaft. Neubearbeitung des Reallexikons der deutschen Literaturgeschichte*, edited by Klaus Weimar, Harald Fricke, and Jan-Dirk Müller, 832–834. Vol. 2. Berlin, New York: De Gruyter.

Thompson, W., and B. Thompson. 1991. "Overturning the Category Bucket." *Byte* 16 (1): 249–256.

Todorov, Tzvetan. 1970. *Introduction à la littérature fantastique.* Paris: Seuil.

Todorov, Tzvetan. 2014. "The Origin of Genres." In *Modern Genre Theory*, edited by David Duff, 193–209. New York: Routledge.

Tophinke, Doris. 1997. "Zum Problem der Gattungsgrenze – Möglichkeiten einer prototypentheoretischen Lösung." In *Gattungen mittelalterlicher Schriftlichkeit*, edited by Barbara Frank, Thomas Haye, and Doris Tophinke, 161–182. Tübingen: Narr.

Torres-Rioseco, Arturo. 1933. *Bibliografía de la novela mejicana.* Cambridge, Massachusetts: Harvard University Press.

Trelles, Carlos Manuel de. 1911. *Bibliografía Cubana del Siglo XIX.* 8 vols. Matanzas: Imprenta de Quirós y Estrada.

Underwood, Ted. 2015a. "Can we date revolutions in the history of literature and music?" *The Stone and the Shell. Using large digital libraries to advance literary history.* October 3, 2015. https://web.archive.org/web/20230303135942/https://tedunderwood.com/2015/10/03/can-we-date-revolutions-in-the-history-of-literature-and-music/.

Underwood, Ted. 2015b. *Understanding Genre in a Collection of a Million Volumes.* White Paper Report. Urbana-Champaign: University of Illinois. http://dx.doi.org/10.17613/M6W07V.

Underwood, Ted. 2016. "The Life Cycles of Genres." *Journal of Cultural Analytics* 2 (2). https://doi.org/10.22148/16.005.

Underwood, Ted. 2019. *Distant Horizons: Digital Evidence and Literary Change.* Chicago: The University of Chicago Press.

Universidad Nacional Autónoma de México. 2008–2023. "La Novela Corta. Una biblioteca virtual." https://web.archive.org/web/20230328173719/https://www.lanovelacorta.com/.

VanderPlas, Jake. 2017. *Python Data Science Handbook. Essential Tools for Working with Data.* 2nd ed. Sebastopol, CA: O'Reilly.

Varela Jácome, Benito. (1982) 2000. *Evolución de la novela hispanoamericana en el siglo XIX (en formato HTML).* Alicante: Biblioteca Virtual Miguel de Cervantes. https://www.cervantesvirtual.com/nd/ark:/59851/bmct14z8.

Villegas Cedillo, Alberto. 1984. *La novela popular mexicana en el siglo XIX.* San Nicolás de los Garza: Universidad Autónoma de Nuevo León.

Vivas, Eliseo. 1968. "Literary Classes: Some Problems." *Genre* 1: 97–105.

Voßkamp, Wilhelm. 1977. "Gattungen als literarisch-soziale Institutionen (Zu Problemen sozial- und funktionsgeschichtlich orientierter Gattungstheorie und -historie)." In *Textsortenlehre – Gattungsgeschichte*, edited by Walter Hinck, 27–44. Heidelberg: Quelle & Meyer.

Wallach, Hanna M., David Mimno, and Andrew McCallum. 2009. "Rethinking LDA: Why Priors Matter." *Advances in Neural Information Processing Systems* 22: 1973–1981. https://web.archive.org/web/20230316142452/https://mimno.infosci.cornell.edu/papers/NIPS2009_0929.pdf.

Weber, Dietrich. 1998. *Erzählliteratur. Schriftwerk, Kunstwerk, Erzählwerk.* Göttingen: Vandenhoeck & Ruprecht.

Weidacher, Georg. 2017. "Fiktionalität und Fiktionalitätssignale." In *Handbuch Sprache in der Literatur*, edited by Anne Betten, Ulla Fix, and Berbeli Wanning, 373–390. Berlin, New York: De Gruyter.

Werlich, Egon. 1975. *Typologie der Texte. Entwurf eines textlinguistischen Modells zur Grundlegung einer Textgrammatik.* Heidelberg: Quelle & Meyer.

Wikimedia Commons. 2019. "Category:Files from Academia Argentina de Letras." Accessed March 16, 2020. https://commons.wikimedia.org/wiki/Category:Files_from_Academia_Argentina_de_Letras.

Wikimedia Foundation. 2023. "Wikimedia Commons." https://web.archive.org/web/20230603175401/https://commons.wikimedia.org/wiki/Main_Page.

Wikipedia. 2022. "Anexo:Nombres de países en español." Wikipedia. https://web.archive.org/web/20230422153154/https://es.wikipedia.org/wiki/Anexo:Nombres_de_pa%C3%ADses_en_espa%C3%B1ol.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3. https://doi.org/10.1038/sdata.2016.18.

Willand, Marcus, Peer Trilcke, Christof Schöch, Nanette Rißler-Pipka, Nils Reiter, and Frank Fischer. 2017. "Aktuelle Herausforderungen der Digitalen Dramenanalyse." In *DHd 2017. Digitale Nachhaltigkeit. Konferenzabstracts*, 46–49. Bern: Universität Bern. https://doi.org/10.5281/zenodo.4622643.

Wittgenstein, Ludwig. 2009. *Philosophical Investigations.* Edited by P.M.S. Hacker and Joachim Schulte. New York: Wiley.

Wolfzettel, Friedrich. 1999. *Der spanische Roman von der Aufklärung bis zur frühen Moderne.* Tübingen/Basel: Francke.

Yang, Jaewon, and Jure Leskovec. 2015. "Defining and Evaluatign Network Communities based on ground-truth." *Knowledge and Information Systems* 42: 181–213. https://doi.org/10.1007/s10115-013-0693-z.

Yin, Filippa B. 1992. "Díaz Covarrubias, Juan." In *Dictionary of Mexican Literature*, edited by Eladio Cortés, 195–196. Westport, Connecticut; London: Greenwood Press.

Zehe, Albin, Daniel Schlör, Ulrike Henny-Krahmer, Martin Becker, and Andreas Hotho. 2018. "A White-Box Model for Detecting Author Nationality by Linguistic Differences in Spanish Novels." In *Digital Humanities 2018. Puentes–Bridges. Book of Abstracts. Mexico City, 26–29 June 2018*, 519–522. Mexico City: Red de Humanidades Digitales. https://web.archive.org/web/20230212050806/https://dh2018.adho.org/en/a-white-box-model-for-detecting-author-nationality-by-linguistic-differences-in-spanish-novels/.

Zeuske, Michael. 2002. *Kleine Geschichte Kubas.* 2nd ed. München: C. H. Beck.

Zipfel, Frank. 2010. "Gattungstheorie im 20. Jahrhundert." In *Handbuch Gattungstheorie*, edited by Rüdiger Zymner, 213–216. Stuttgart: J.B. Metzler.

Zipfel, Frank. 2014. "Fiktionalitätssignale." In *Fiktionalität. Ein interdisziplinäres Handbuch*, edited by Tobias Klauk and Tilmann Köppe, 97–124. Berlin, Boston: De Gruyter.

Zó, Ramiro Esteban. 2015. *Emociones escriturales. La novela sentimental latinoamericana.* Saarbrücken: Editorial Académica Española.

Zum Felde, Alberto. 1954. *Índice crítico de la literatura hispanoamericana.* 2 vols. México: Editorial Guaranía.

Zymner, Rüdiger. 2003. *Gattungstheorie. Probleme und Positionen der Literaturwissenschaft.* Paderborn: mentis.

Zymner, Rüdiger, ed. 2010. *Handbuch Gattungstheorie.* Stuttgart: J.B. Metzler.

Zymner, Rüdiger. 2017. "Narrative Gattungen." In *Grundthemen der Literaturwissenschaft: Erzählen*, edited by Martin Huber and Wolf Schmid, 365–383. Berlin: De Gruyter.

Appendix

Sources of the Novels in the Corpus

The table is also available at https://github.com/cligs/data-nh/blob/master/corpus/corpus-sources/overview-sources.csv. Accessed March 28, 2020. All the links in the table were accessed on December 8, 2019.

Table 46. Sources of the novels in the corpus.

Source	Link	Novels
Autores de Concordia	www.autoresdeconcordia.com.ar	1
Bayerische Staatsbibliothek digital	https://www.digitale-sammlungen.de/	1
Biblioteca Digital Argentina	http://www.biblioteca.clarin.com/pbda/novela[622]	11
Biblioteca Digital de Castilla y León	https://bibliotecadigital.jcyl.es	1
Biblioteca Digital de la Biblioteca Nacional de Cuba José Martí	http://bdigital.bnjm.cu/	2
Biblioteca Digital del ILCE (Instituto Latinoamericano de la Comunicación Educativa)	http://bibliotecadigital.ilce.edu.mx/	2
Biblioteca Digital Hispánica	http://bibliotecadigitalhispanica.bne.es/	6
Biblioteca Digital Mexicana del Bicentenario	http://www.bicentenario.gob.mx/[623]	1
Biblioteca Digital Trapalanda	http://trapalanda.bn.gov.ar[624]	1
Biblioteca Virtual Antorcha	http://www.antorcha.net/index/biblioteca/literatura.html	4
Biblioteca Virtual Miguel de Cervantes	http://www.cervantesvirtual.com/	39
Colección Digital UANL (Universidad Autónoma de Nuevo León)	https://cd.dgb.uanl.mx/	10
Conaculta (Secretaría de Cultura del Gobierno de México)	https://mexicana.cultura.gob.mx/es/repositorio[625]	2
Digital Library of the Caribbean (dLOC)	https://dloc.com/	1
El Libro Total	https://www.ellibrototal.com/ltotal/	4
EnCaribe: Enciclopedia de Historia y Cultura del Caribe	http://www.encaribe.org/	3
Fondo antiguo y colecciones especiales de la Dirección General de Bibliotecas de la UNAM (Universidad Nacional Autónoma de México)	http://dgb.unam.mx/index.php/catalogos/fondo-antiguo	4
Fondo de Cultura Económica	https://fondodeculturaeconomica.com/	1
Google Books	https://books.google.de/	6

[622] This website is not available anymore.

[623] This website has broken links and does not seem to be supported anymore.

[624] This website is not available anymore.

[625] The digital repository of the Mexican government's Secretary of Culture was relaunched during the preparation of this dissertation. The two novels in the corpus were taken from the old version of the repository and could be downloaded at http://dgb.conaculta.gob.mx/coleccion_sep and http://impresosmexicanos.conaculta.gob.mx/libros. These links are not accessible anymore. The new version of this digital repository whose link is given in the table is called "Mexicana. Repositorio del Patrimonio Cultural de México".

Source	Link	Novels
HathiTrust Digital Library	https://www.hathitrust.org/	24
Ibero-Amerikanisches Institut	https://digital.iai.spk-berlin.de/viewer/	22
iBooks Store	https://books.apple.com/do/book/	3
Individual website	-	1
Internet Archive	https://archive.org/	26
La novela corta. Una biblioteca virtual	https://www.lanovelacorta.com/	6
Libros México	https://www.librosmexico.mx	1
Project Gutenberg	https://www.gutenberg.org/	6
University library	-	37
Wikimedia Commons	https://commons.wikimedia.org/wiki/Main_Page	19
Wikisource	https://es.wikisource.org/wiki/Portada	11
Sum:	-	226

Appendix of Figures

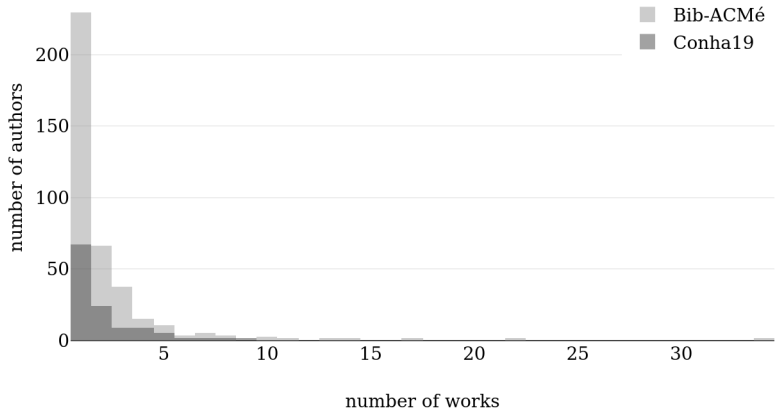


Figure 94. Number of works per author.

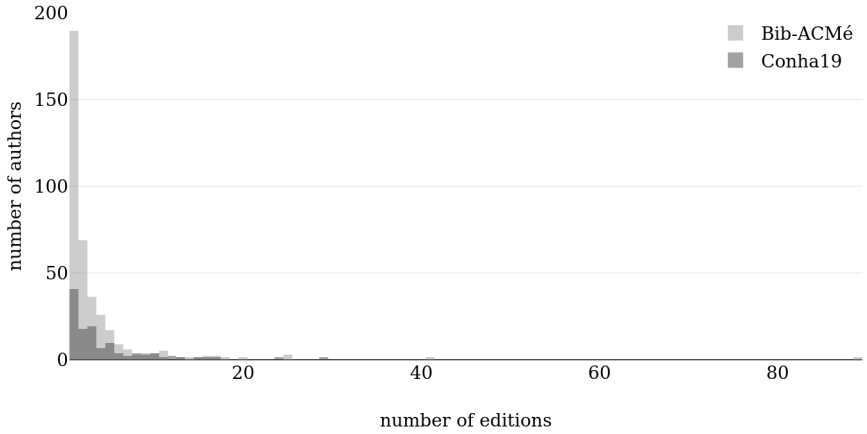


Figure 95. Number of editions per author.

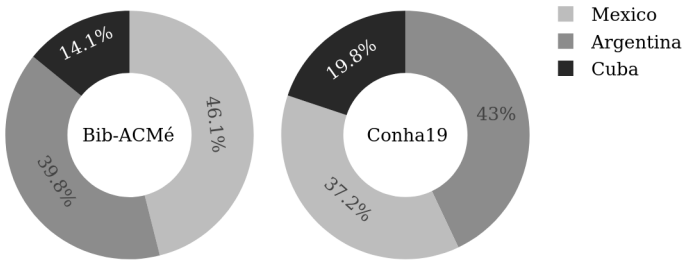


Figure 96. Authors by country.

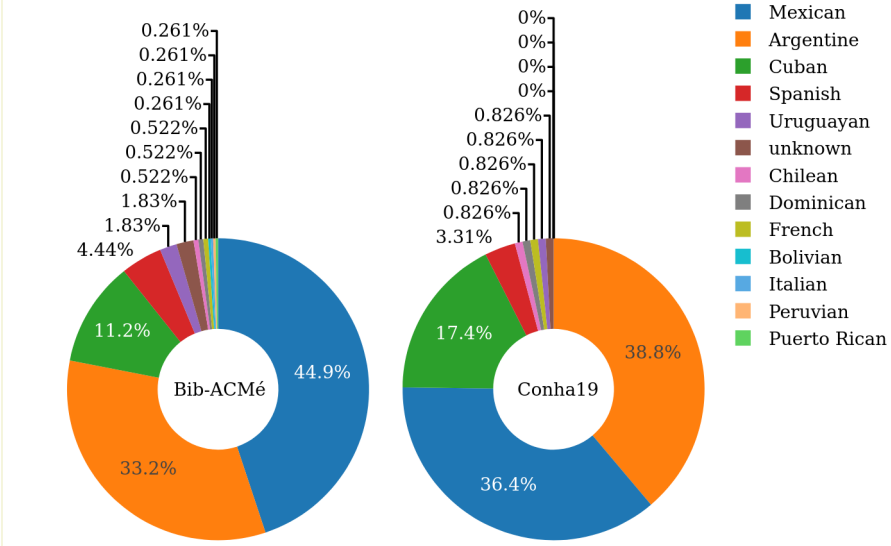


Figure 97. Authors by nationality.

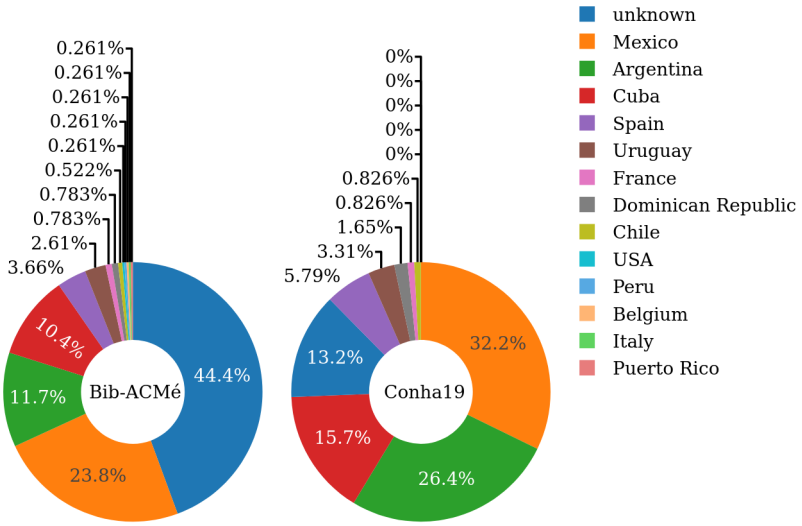


Figure 98. Authors by country of birth.

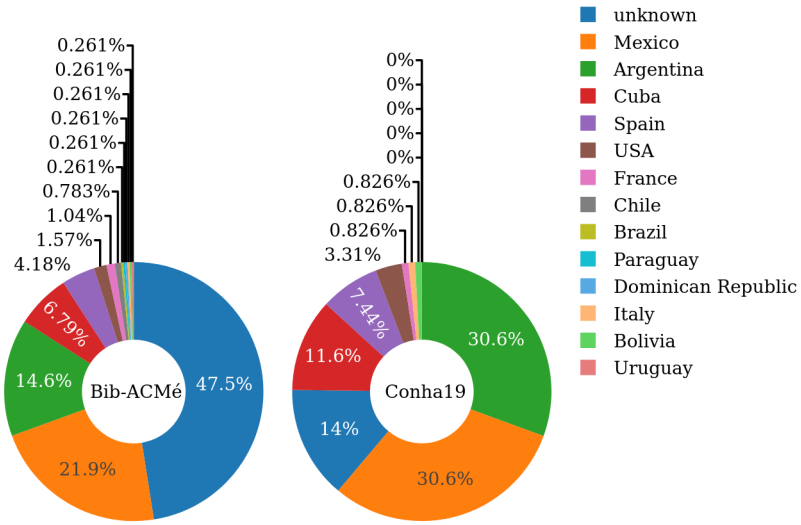


Figure 99. Authors by country of death.

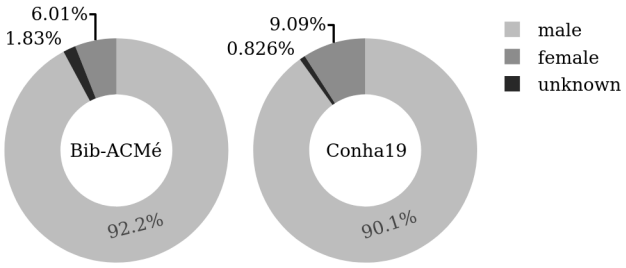


Figure 100. Author gender.

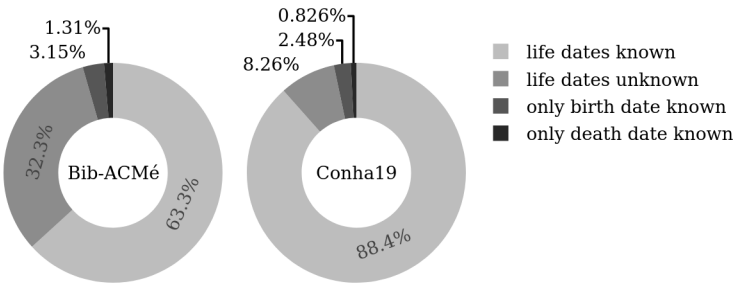


Figure 101. Knowledge of the authors' life dates.

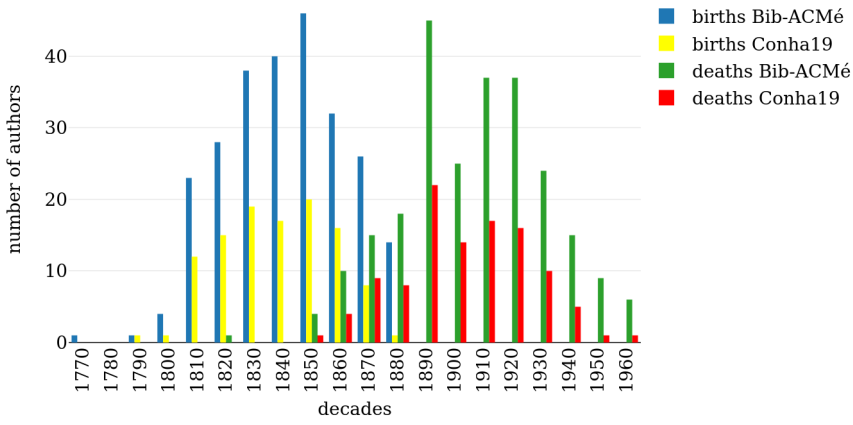


Figure 102. Births and deaths of authors by decade.

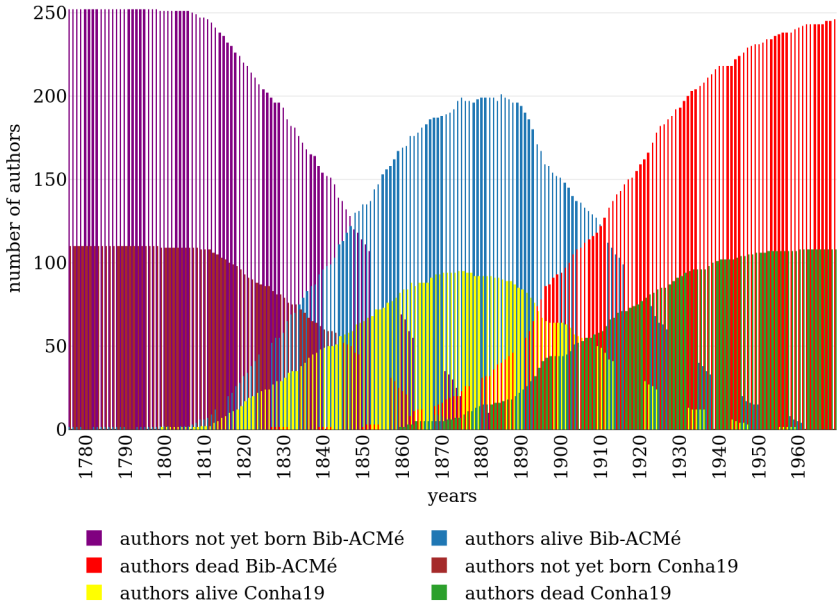


Figure 103. Authors alive per year.

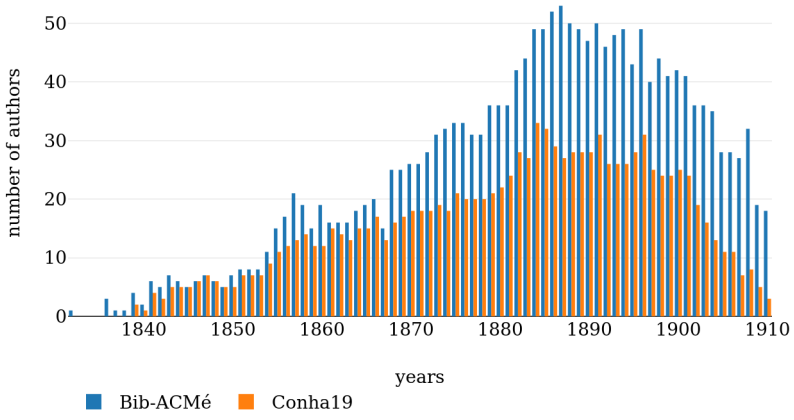


Figure 104. Number of active authors per year.

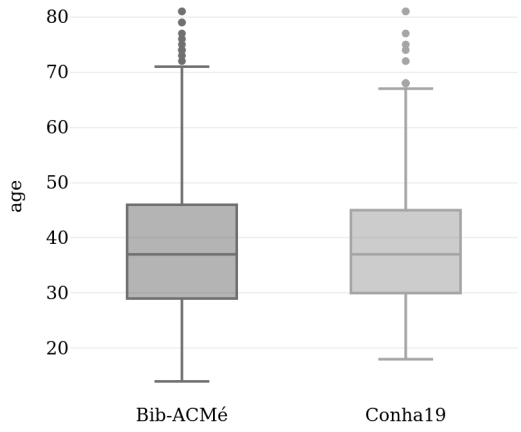


Figure 105. Author ages when publishing novels.

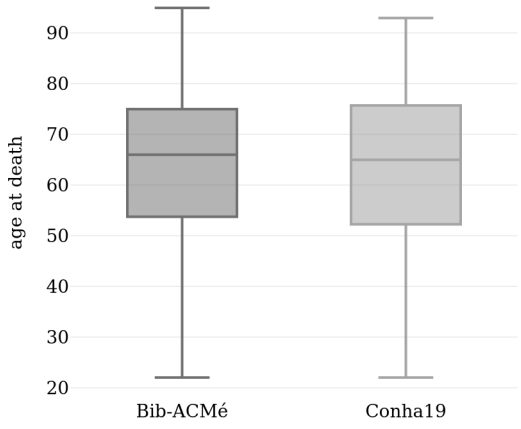


Figure 106. Authors' age at death.

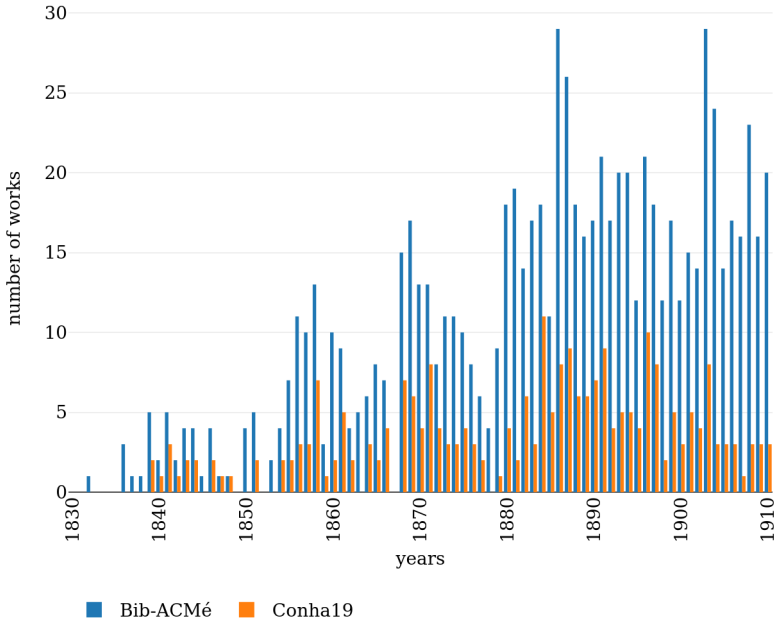


Figure 107. Number of works per year in Bib-ACMé and Conha19.

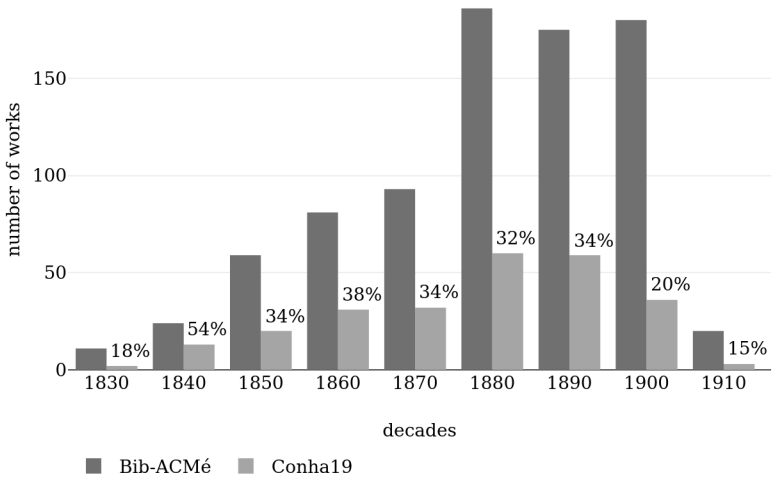


Figure 108. Works by decade in Bib-ACMé and Conha19.

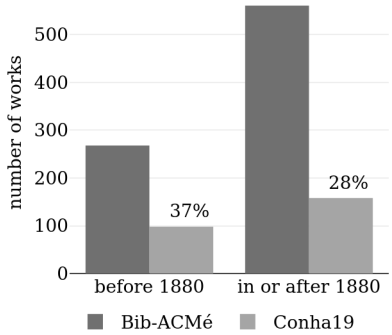


Figure 109. Works before and after 1880.

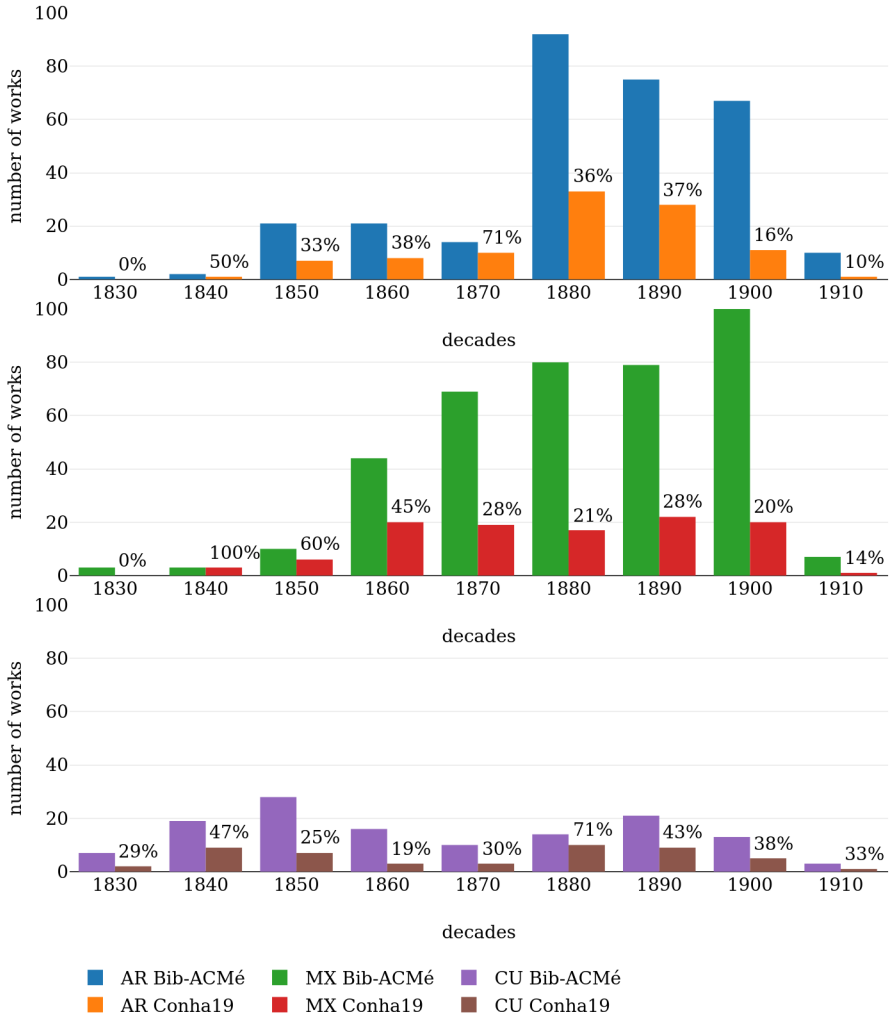


Figure 110. Works by decade and country.

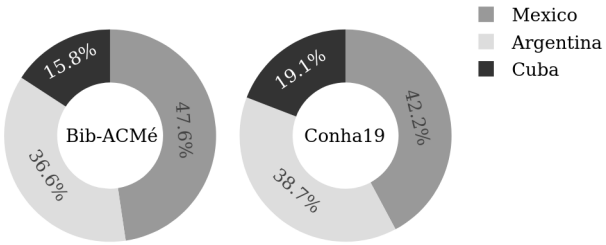


Figure 111. Works by country in Bib-ACMé and Conha19.

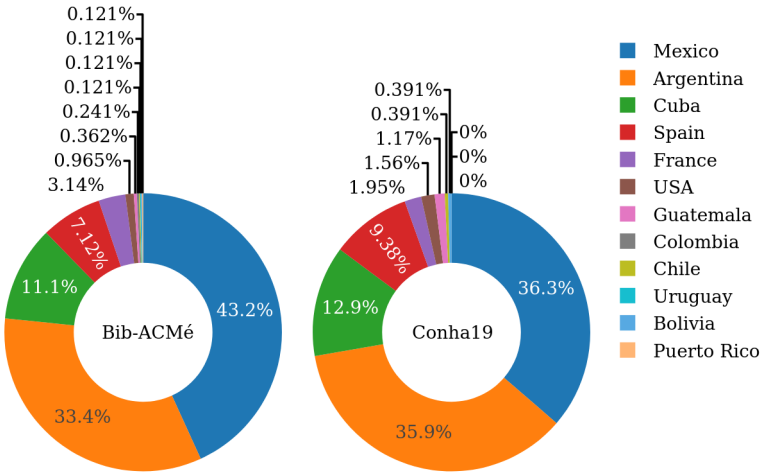


Figure 112. Publication countries of first editions.

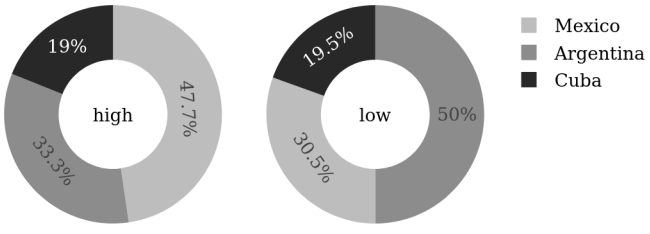


Figure 113. High and low prestige novels by country.

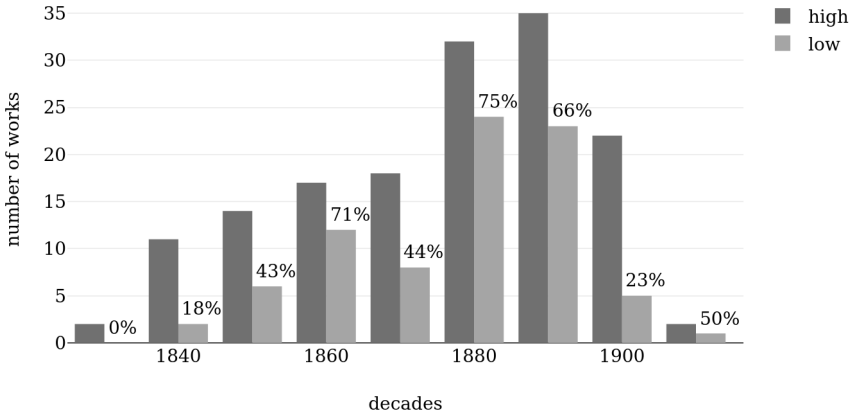


Figure 114. High and low prestige novels by decade.

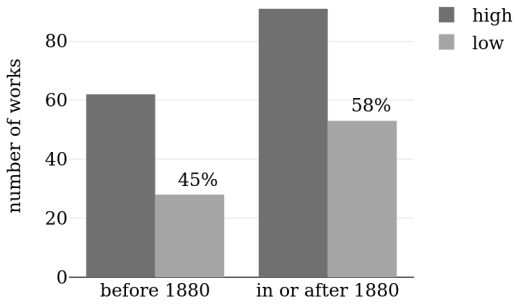


Figure 115. High and low prestige novels before and in or after 1880.

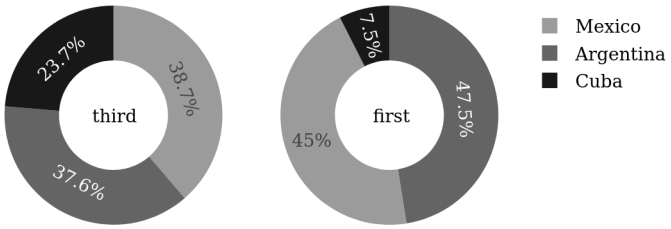


Figure 116. Narrative perspective by country.

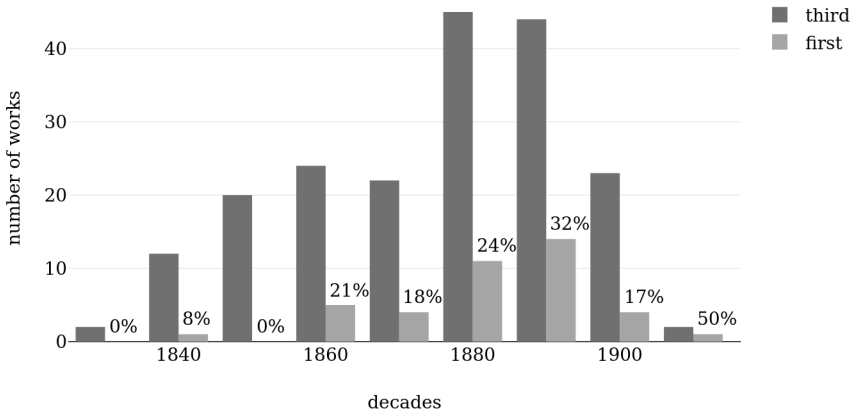


Figure 117. Narrative perspective by decade.

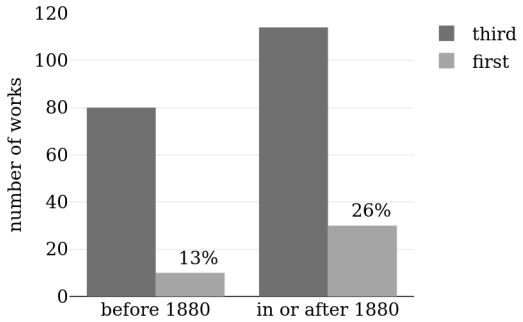


Figure 118. Narrative perspective before and in or after 1880.

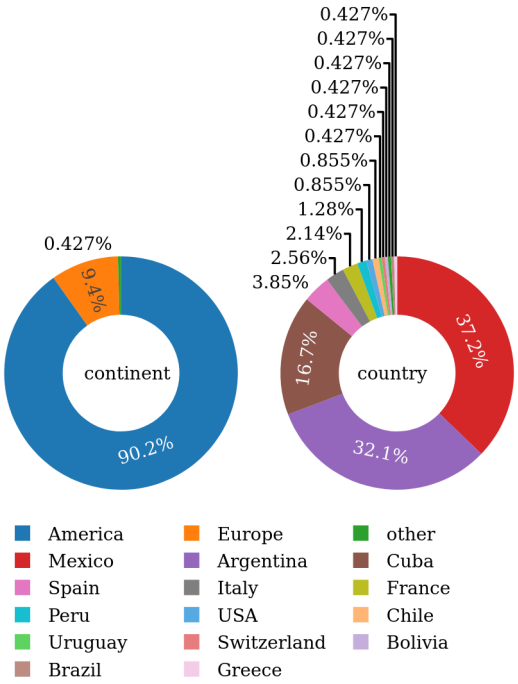


Figure 119. Continent and country of the setting.

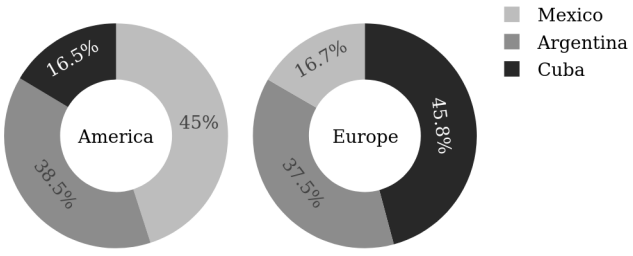


Figure 120. Continent of the setting by country.

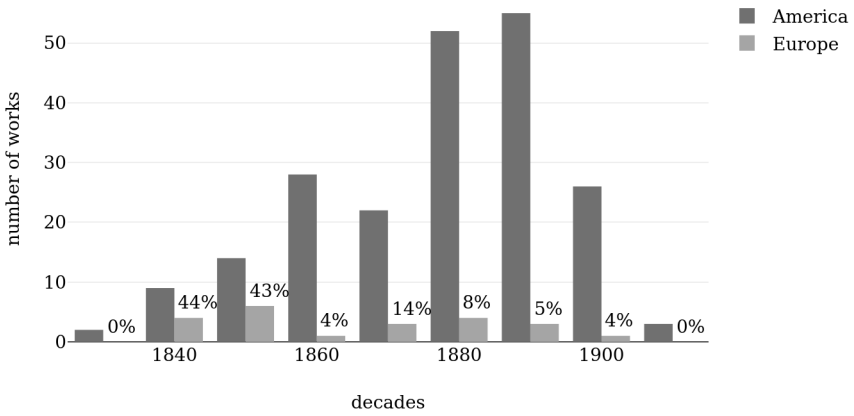


Figure 121. Continent of the setting per decade.

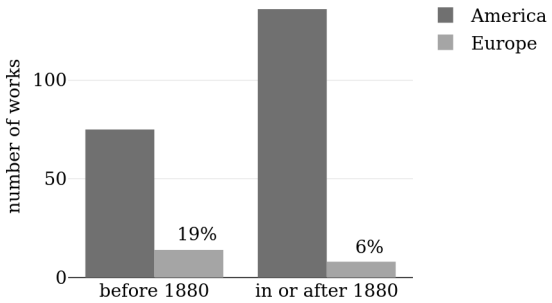


Figure 122. Continent of the setting before and in or after 1880.

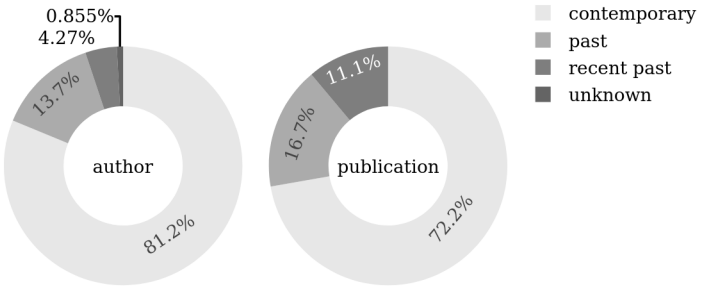


Figure 123. Time periods of the setting relative to the authors' birth year and publication year.

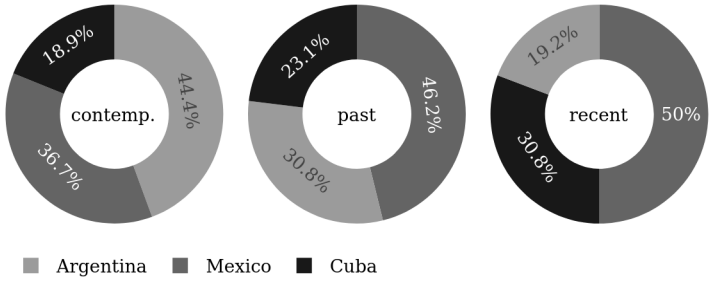


Figure 124. Time periods of the setting by country.

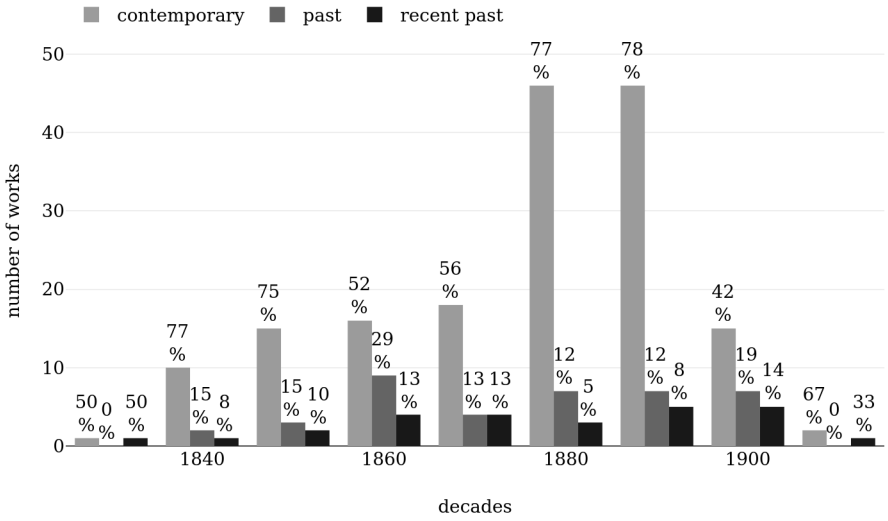


Figure 125. Time period of the setting per decade.

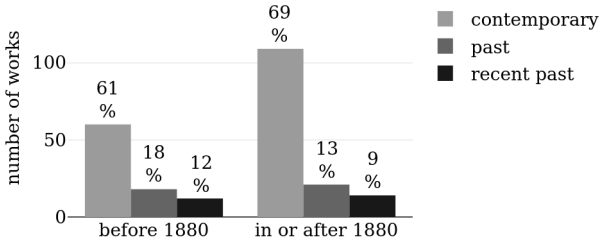


Figure 126. Time period of the setting before and in or after 1880.

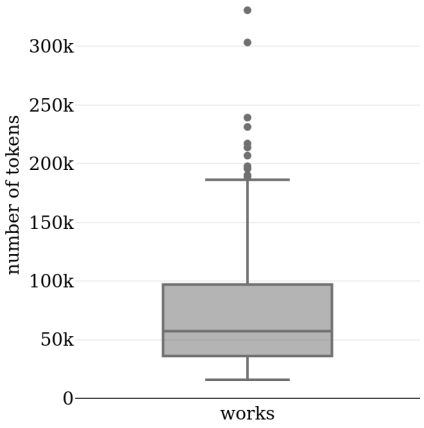


Figure 127. Length of the novels in the corpus.

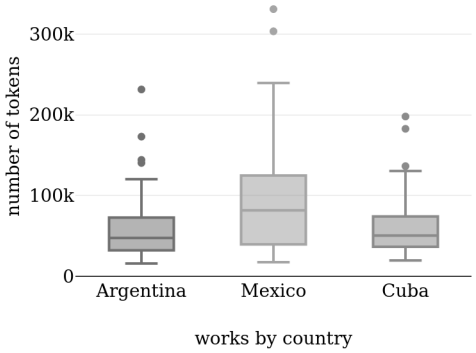


Figure 128. Length of the novels by country.

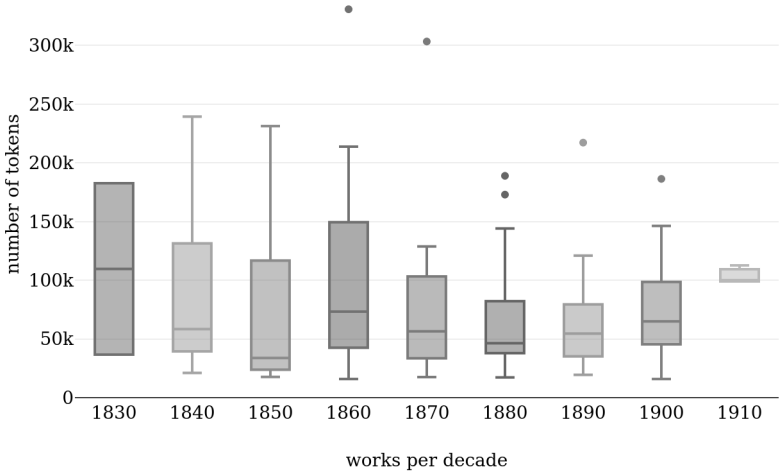


Figure 129. Length of the novels per decade.

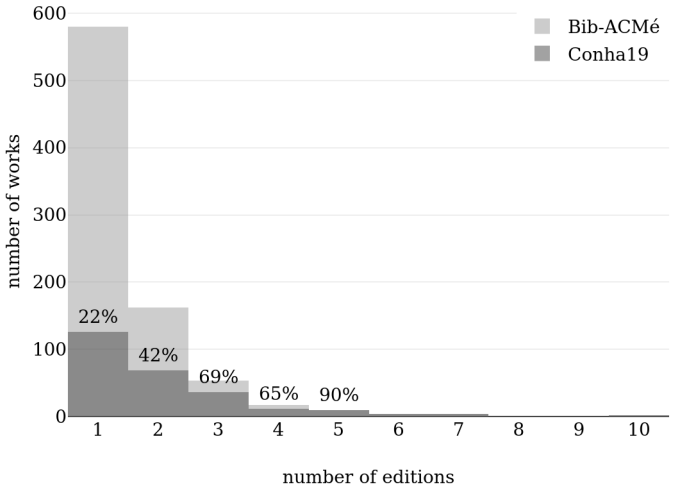


Figure 130. Number of editions per work in Bib-ACMé and Conha19.

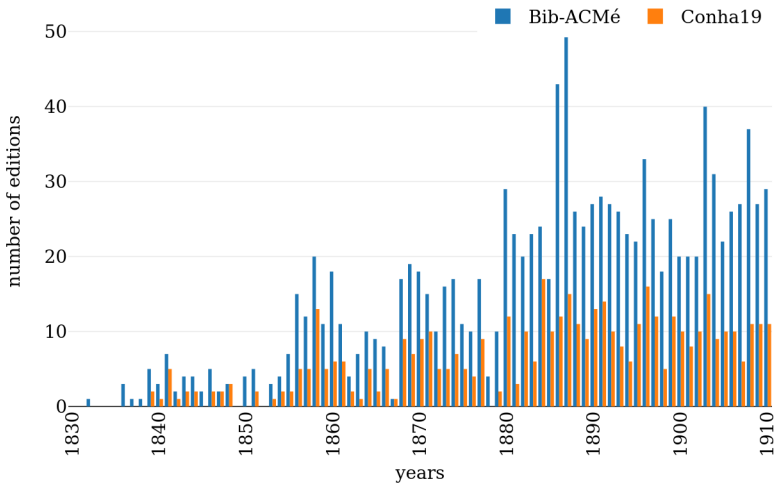


Figure 131. Editions per year in Bib-ACMé and Conha19.

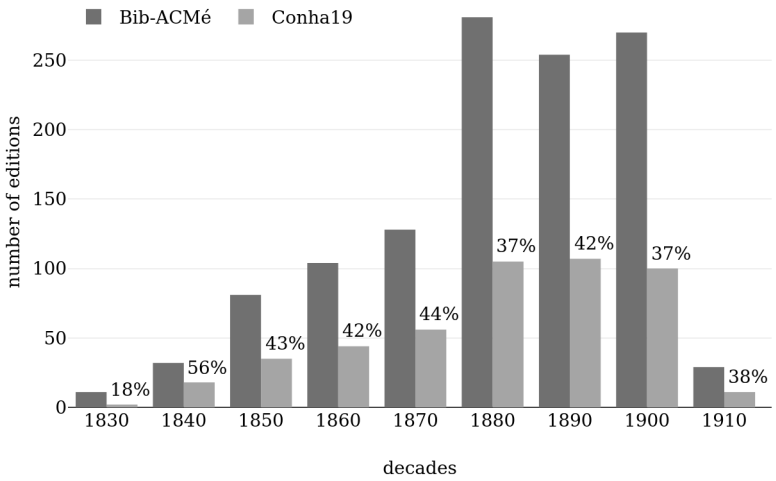


Figure 132. Editions per decade in Bib-ACMé and Conha19.

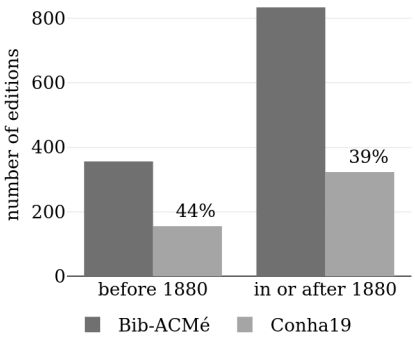


Figure 133. Editions before and in or after 1880.

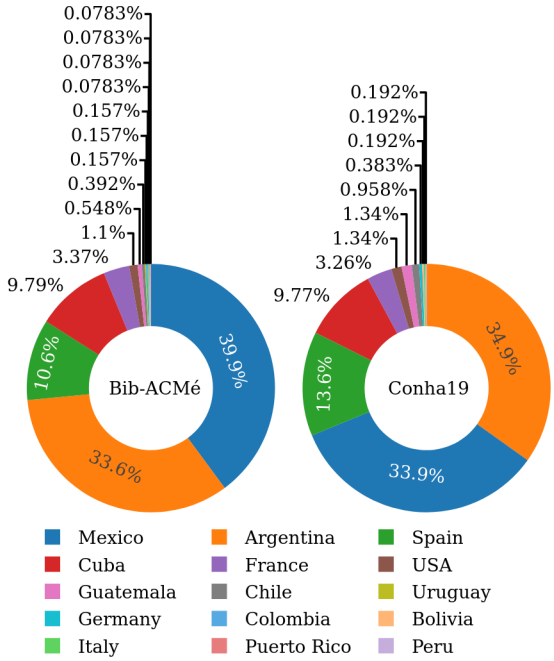


Figure 134. Editions by country in Bib-ACMé and Conha19.

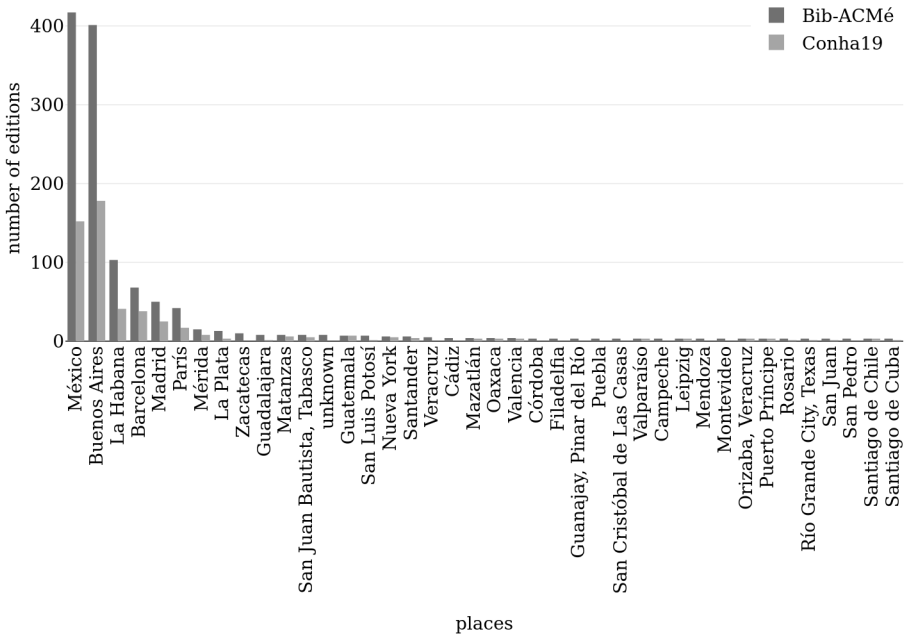


Figure 135. Editions by place of publication in Bib-ACMé and Conha19.

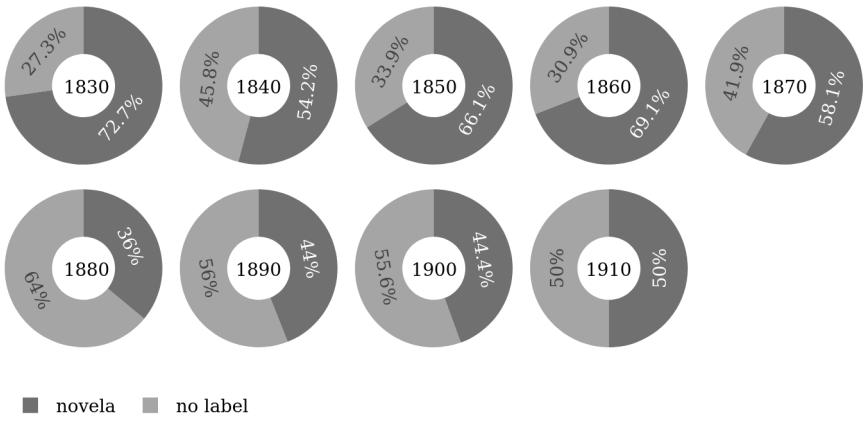


Figure 136. Works with the label “novela” by decade.

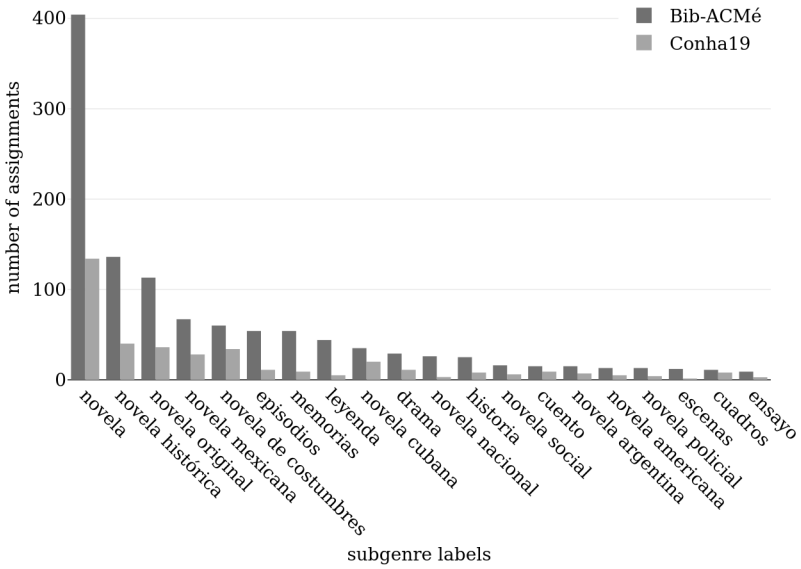


Figure 137. Top 20 most frequent explicit subgenre labels in the bibliography.

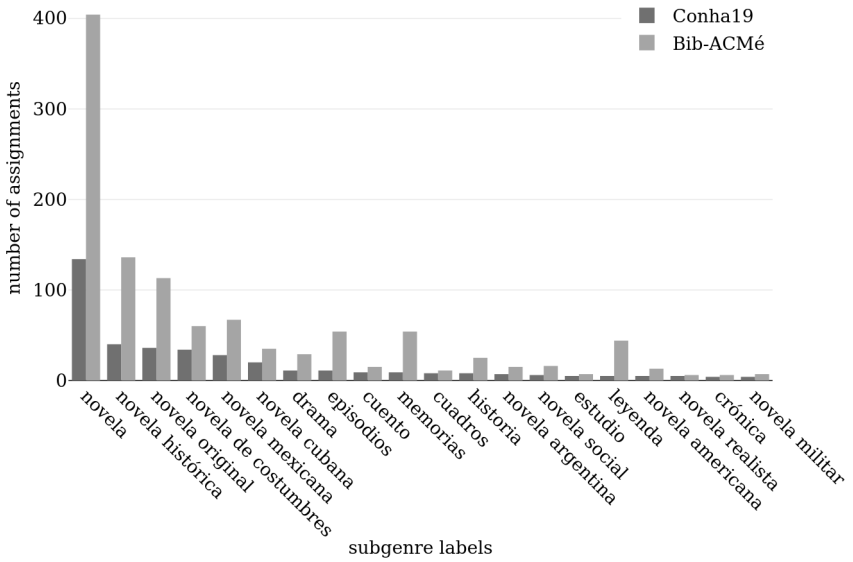


Figure 138. Top 20 most frequent explicit subgenre labels in the corpus.

Figure 139. Works with an “identity label” by decade.

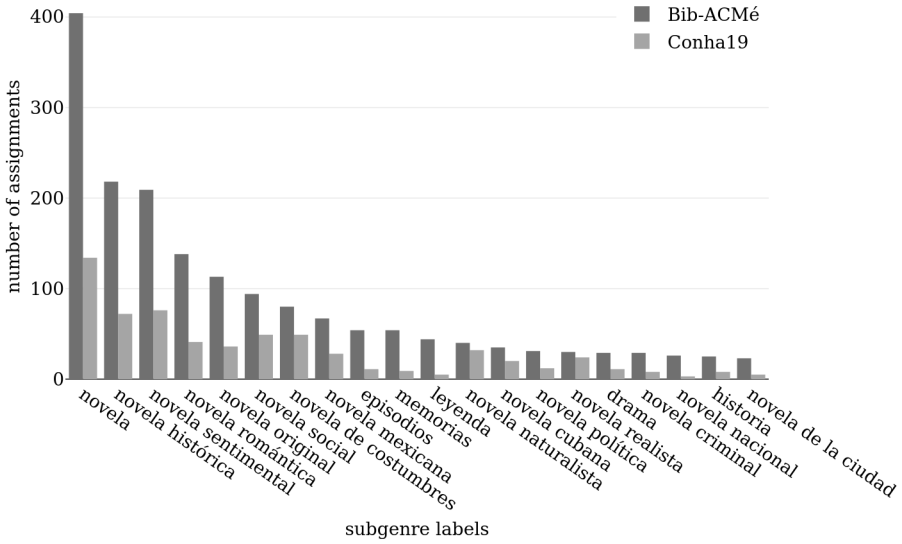


Figure 140. Top 20 most frequent subgenre signals in the bibliography.

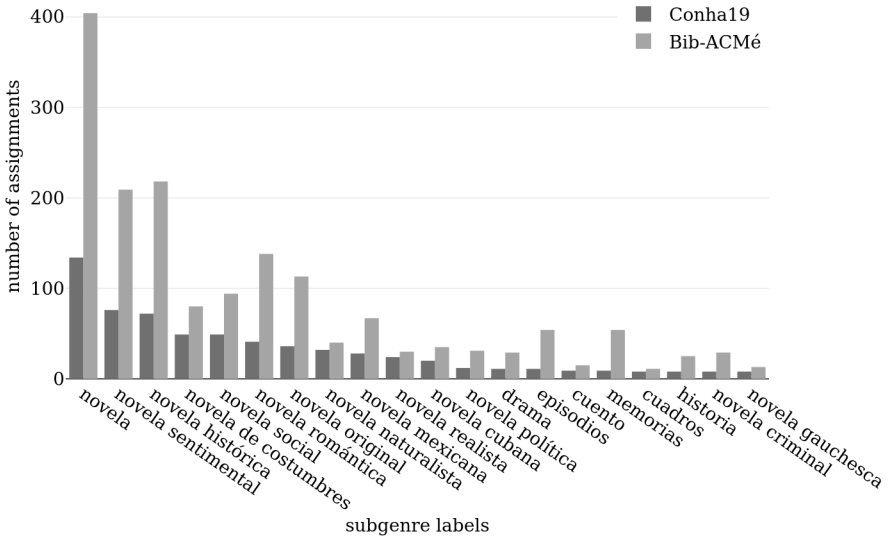


Figure 141. Top 20 most frequent subgenre signals in the corpus.

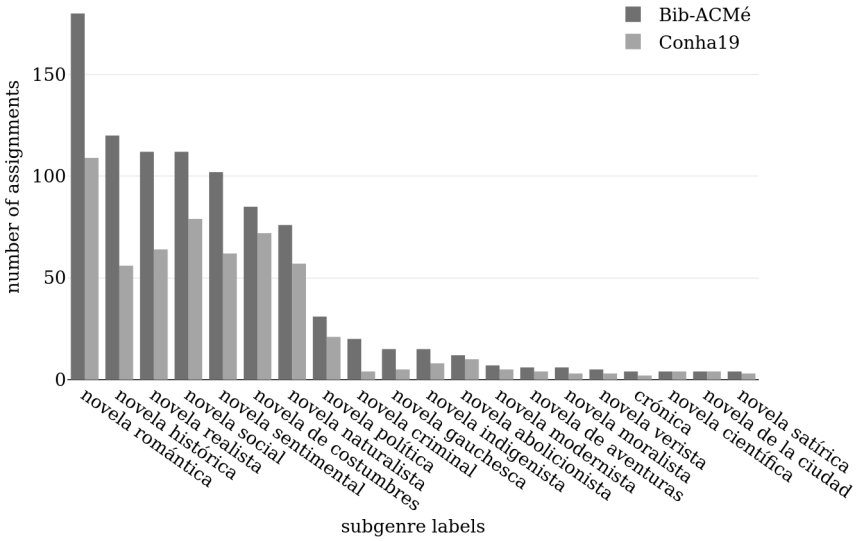


Figure 142. Top 20 most frequent literary historical subgenre labels in the bibliography.

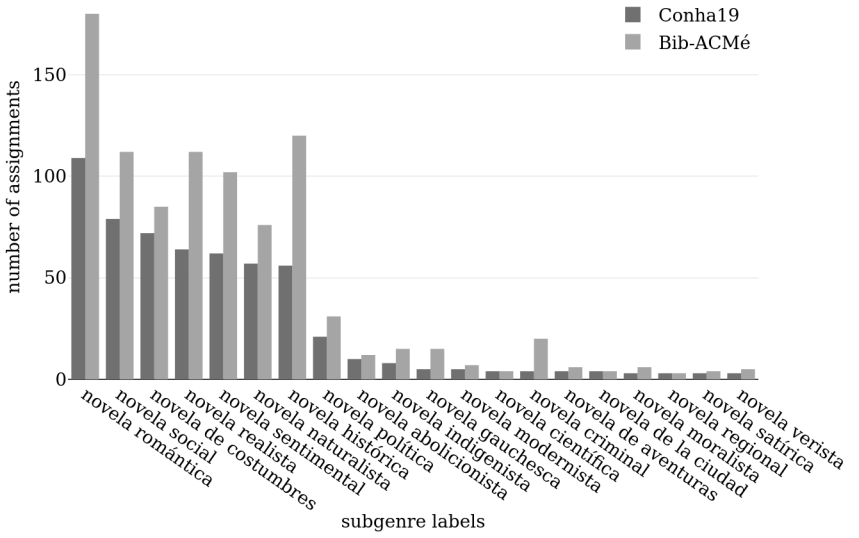


Figure 143. Top 20 most frequent literary historical subgenre labels in the corpus.

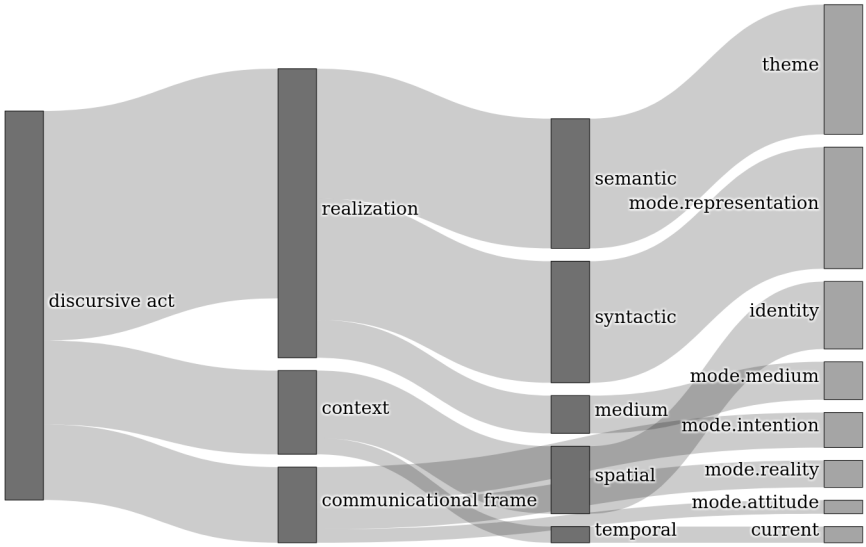


Figure 144. Number of different subgenre labels on discursive levels (in Bib-ACMé).

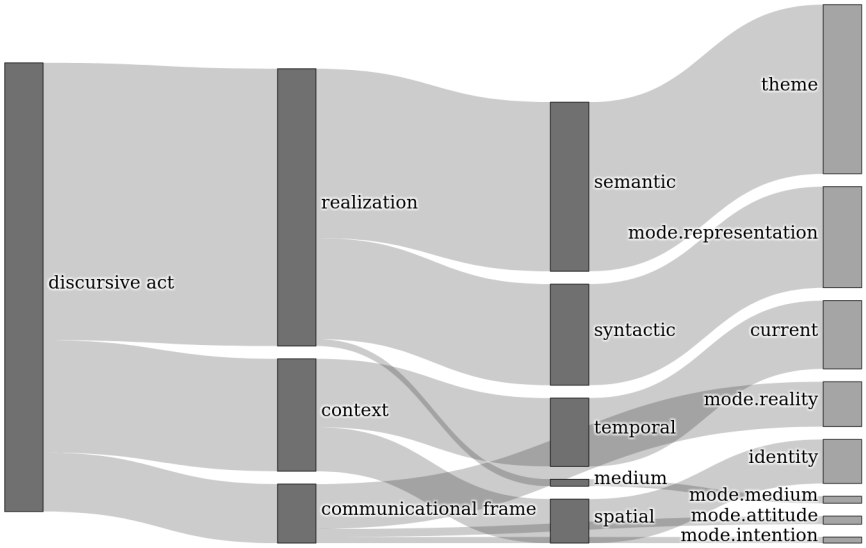


Figure 145. Overall number of subgenre labels on discursive levels (in Bib-ACMé).

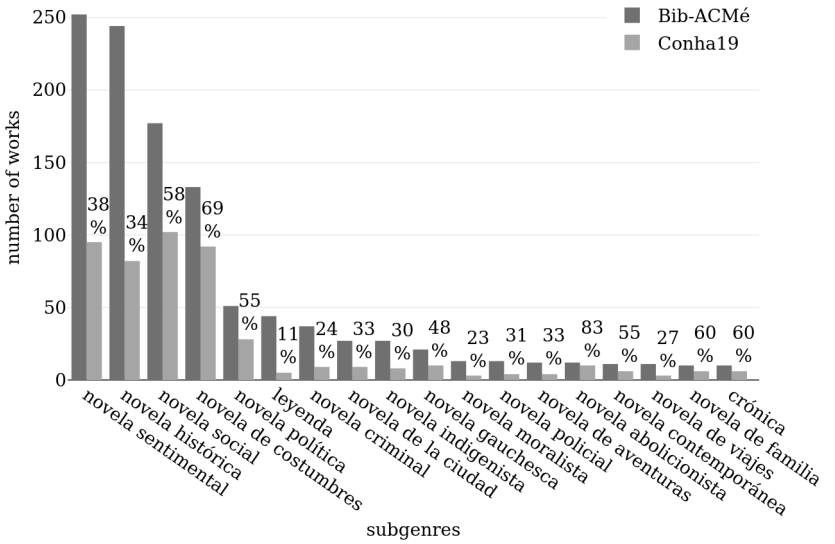


Figure 146. Thematic subgenre labels in Bib-ACMé and Conha19.

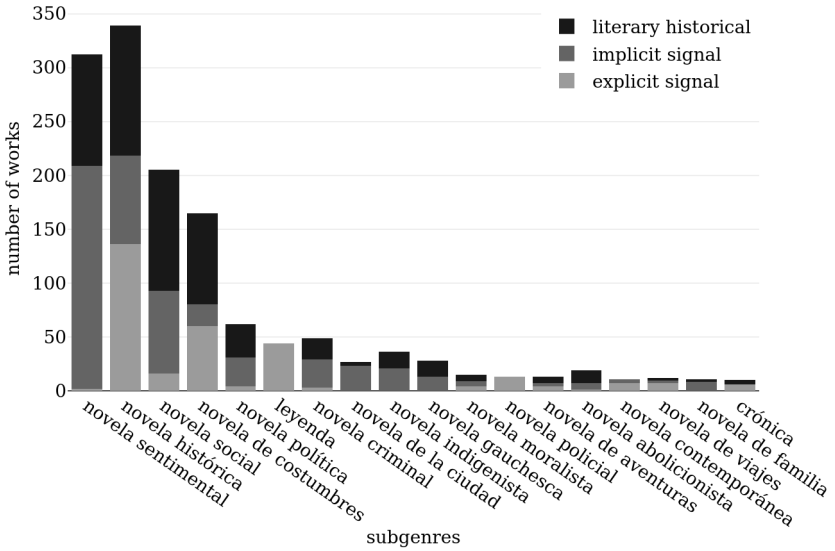


Figure 147. Sources of thematic subgenres in Bib-ACMé.

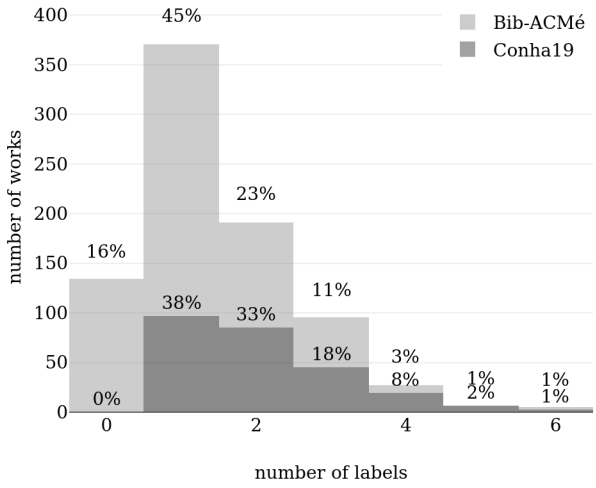


Figure 148. Number of thematic labels per work.

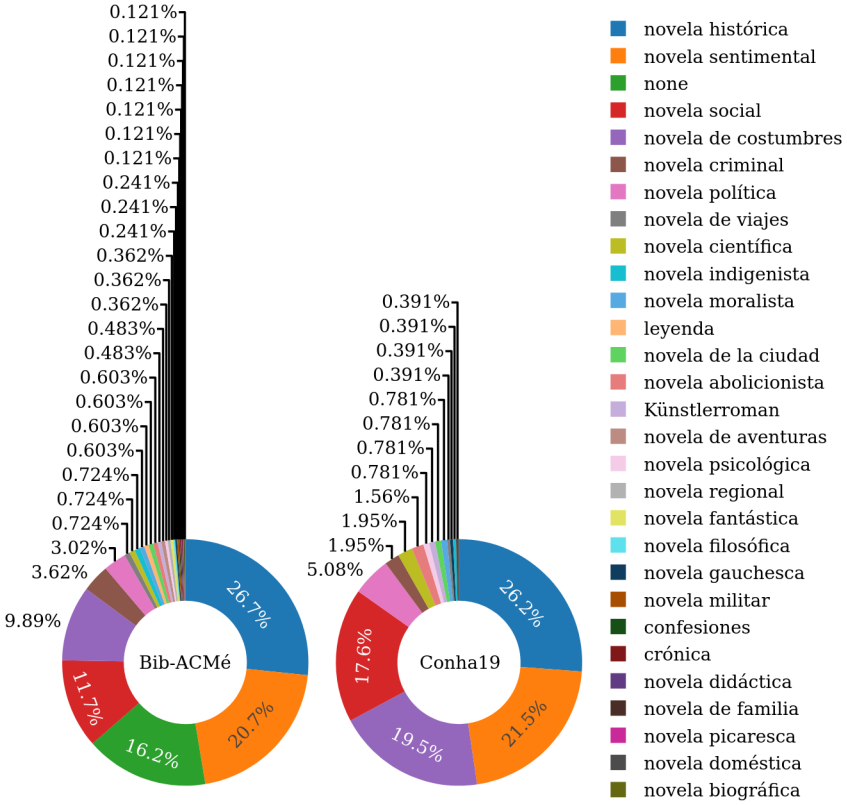


Figure 149. Primary thematic subgenres of the works.

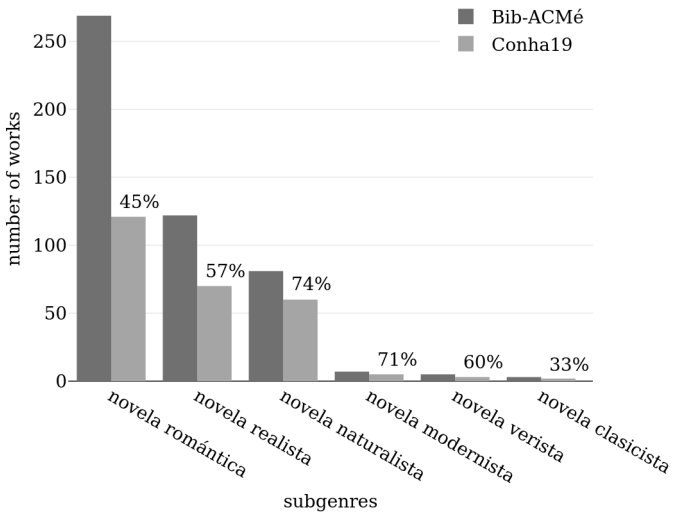


Figure 150. Subgenre labels related to literary currents in Bib-ACMé and Conha19.

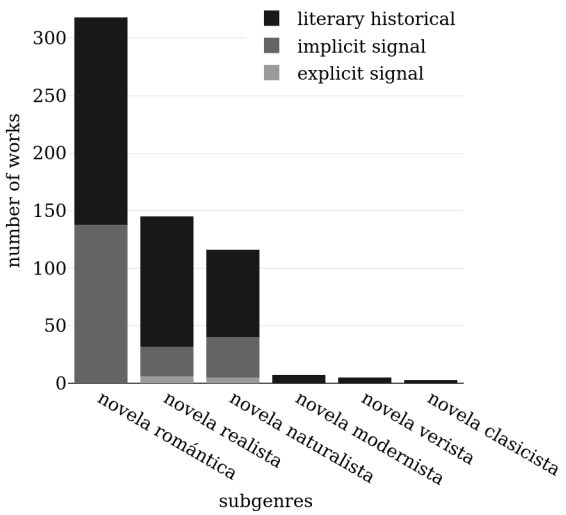


Figure 151. Sources of subgenre labels related to literary currents in Bib-ACMé.

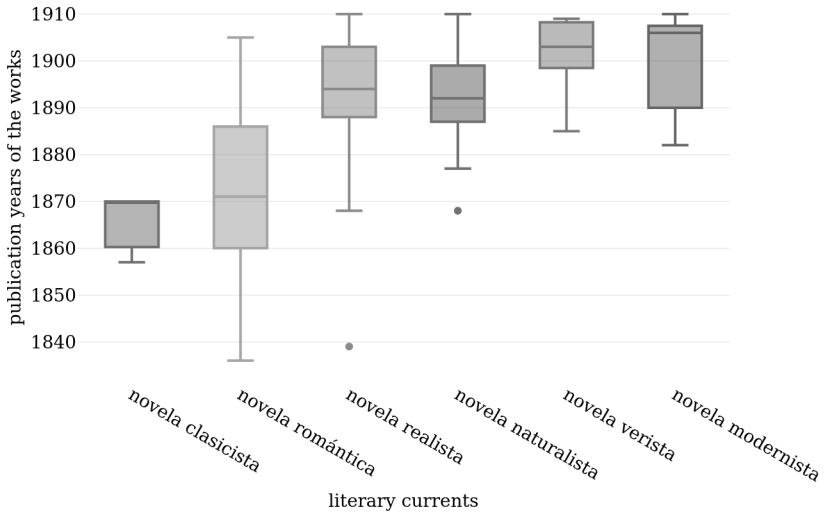


Figure 152. Publication years of works by literary current in Bib-ACMé.

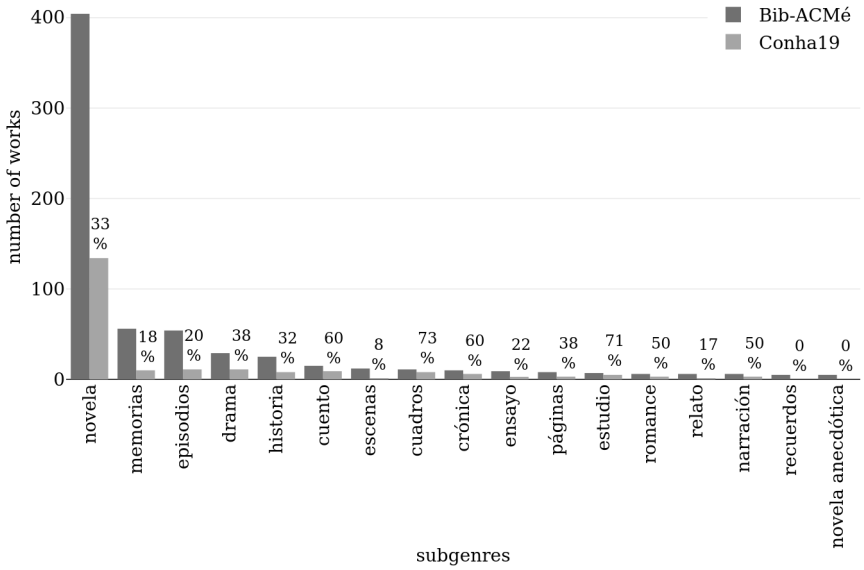


Figure 153. Subgenre labels related to the mode of representation in Bib-ACMé and Conha19.

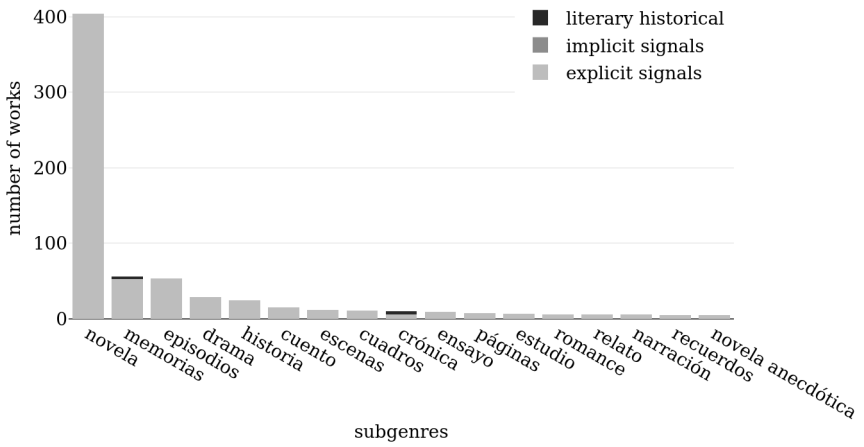


Figure 154. Sources of labels related to the mode of representation in Bib-ACMé.

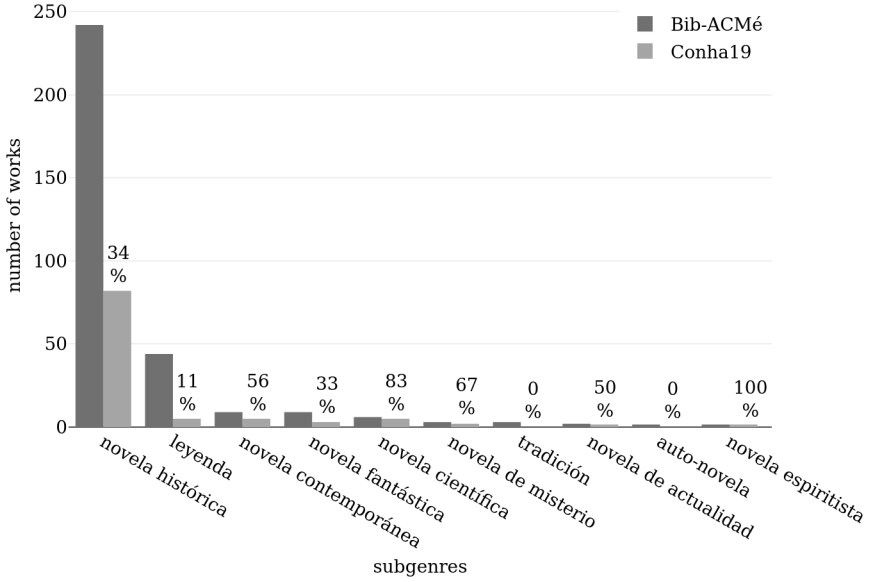


Figure 155. Subgenre labels related to the mode of reality in Bib-ACMé and Conha19.

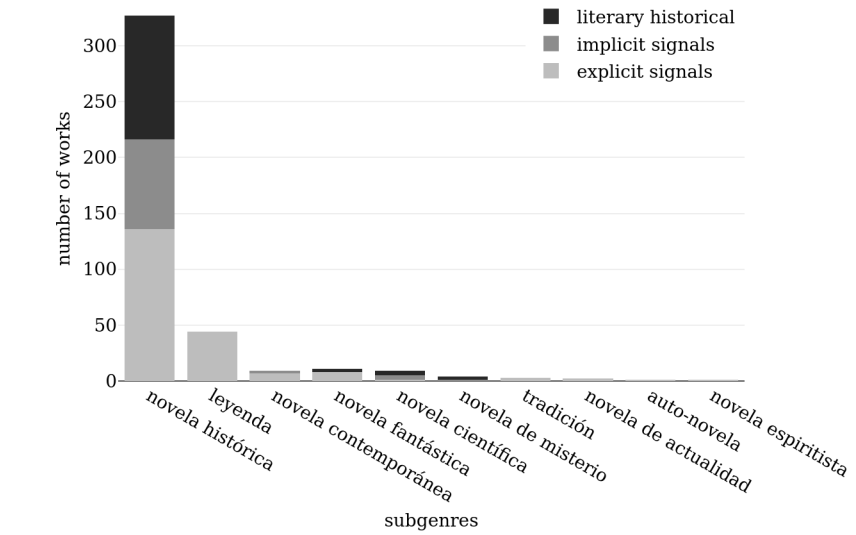


Figure 156. Sources of subgenre labels related to the mode of reality in Bib-ACMé.

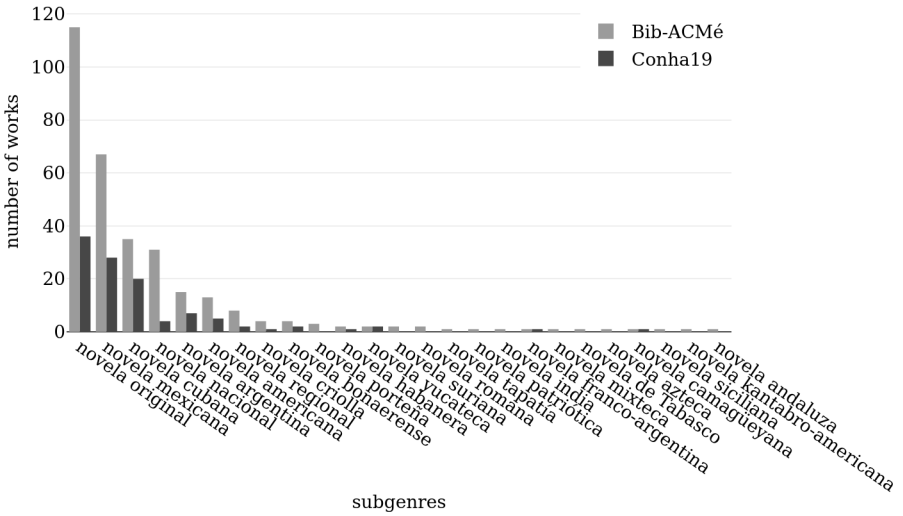


Figure 157. Subgenres related to the linguistic, geographical, and socio-cultural identity.

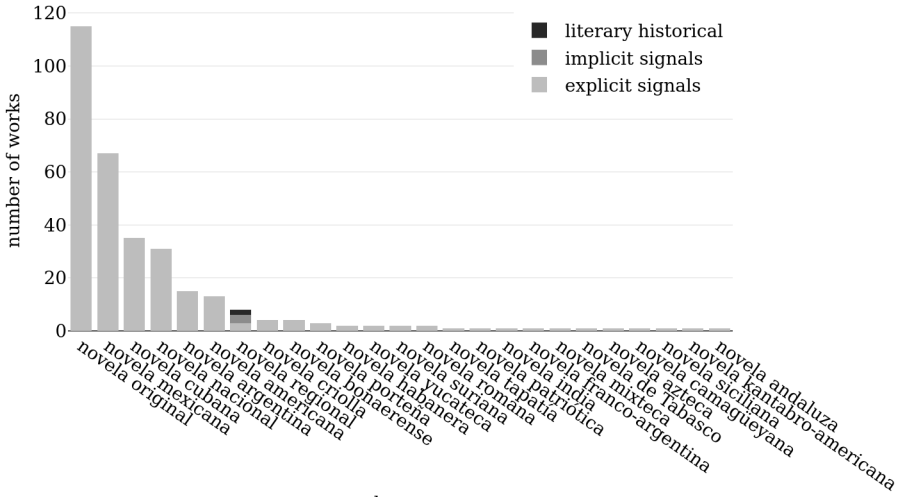


Figure 158. Sources of identity subgenre labels in Bib-ACMé.

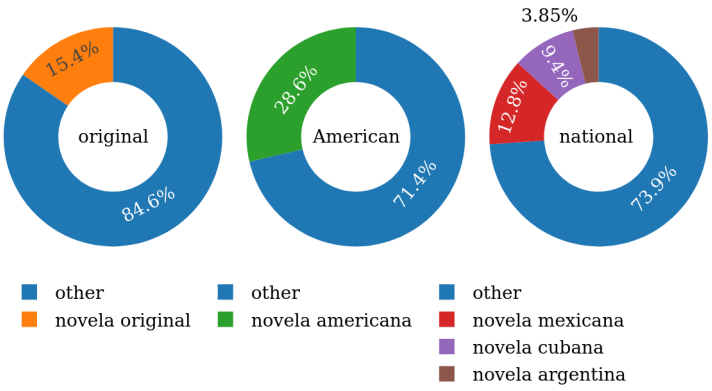


Figure 159. Constellations of identity groups in Conha19.

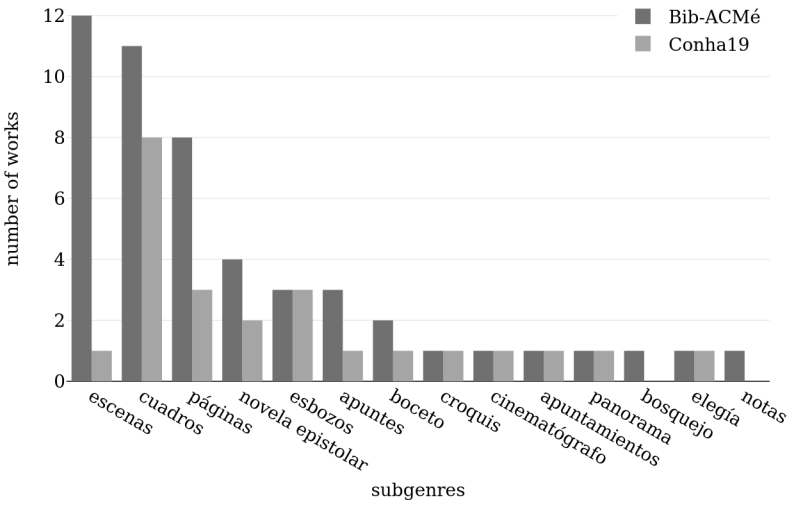


Figure 160. Subgenre labels related to medial aspects in Bib-ACMé and Conha19.

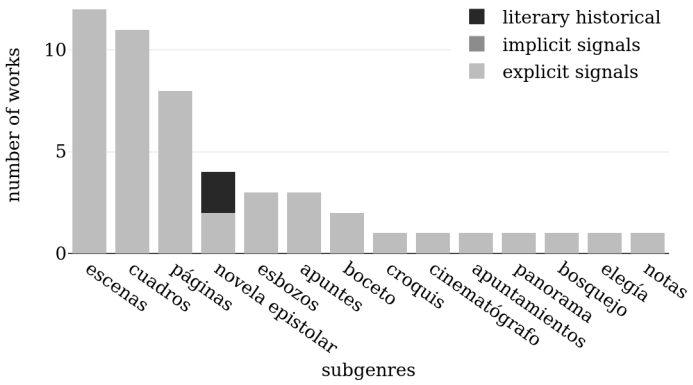


Figure 161. Sources of the subgenre labels related to medial aspects in Bib-ACMé.

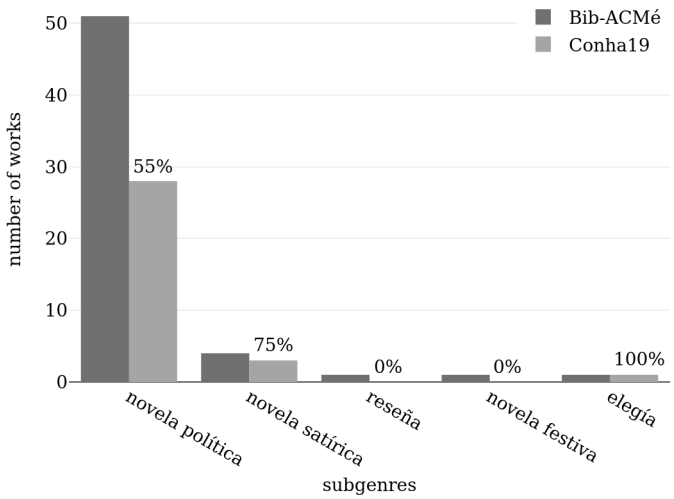


Figure 162. Subgenre labels related to the attitude in Bib-ACMé and Conha19.

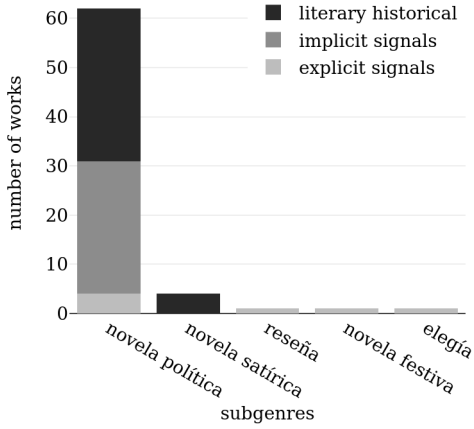


Figure 163. Sources of subgenre labels related to the attitude in Bib-ACMé.

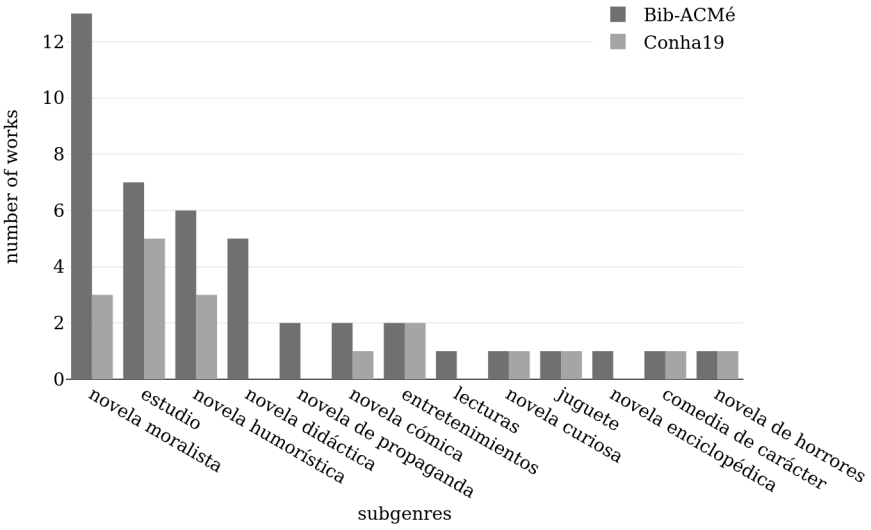


Figure 164. Subgenre labels related to the intention in Bib-ACMé and Conha19.

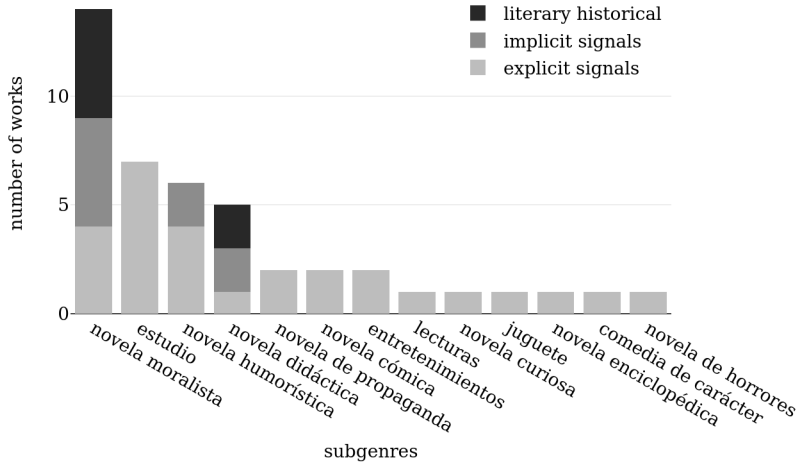


Figure 165. Sources of subgenre labels related to the intention in Bib-ACMé.

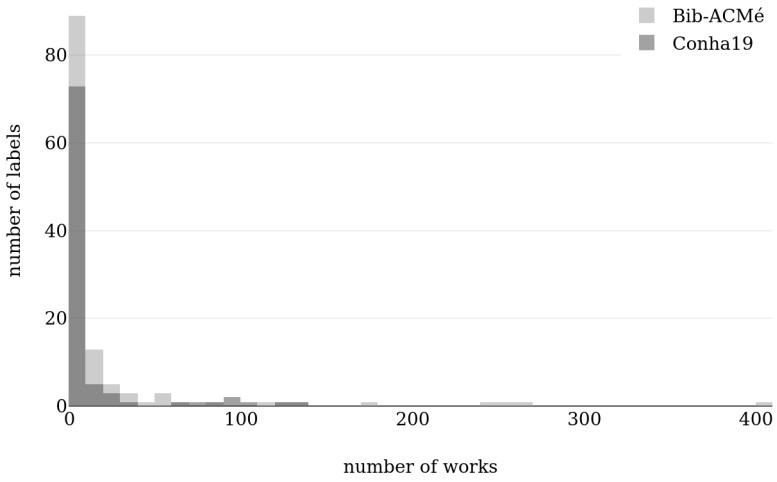


Figure 166. Number of works per subgenre label.

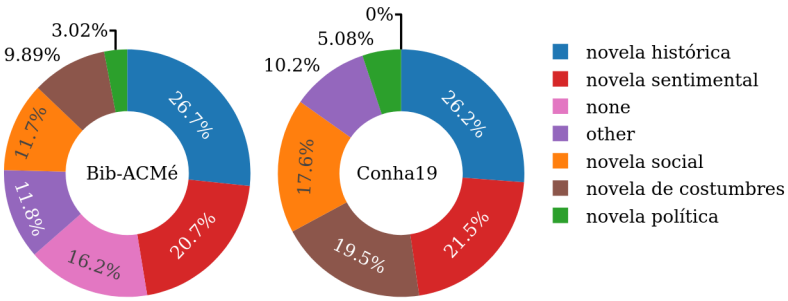


Figure 167. Primary thematic subgenres in Bib-ACMé and Conha19.

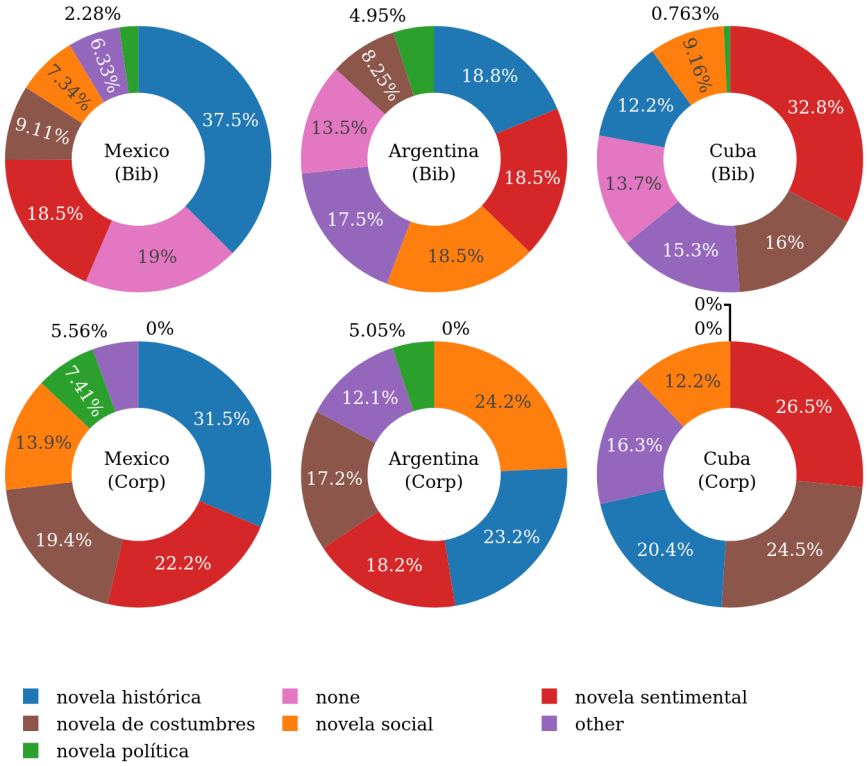


Figure 168. Primary thematic subgenre labels in Bib-ACMé and Conha19 by country.

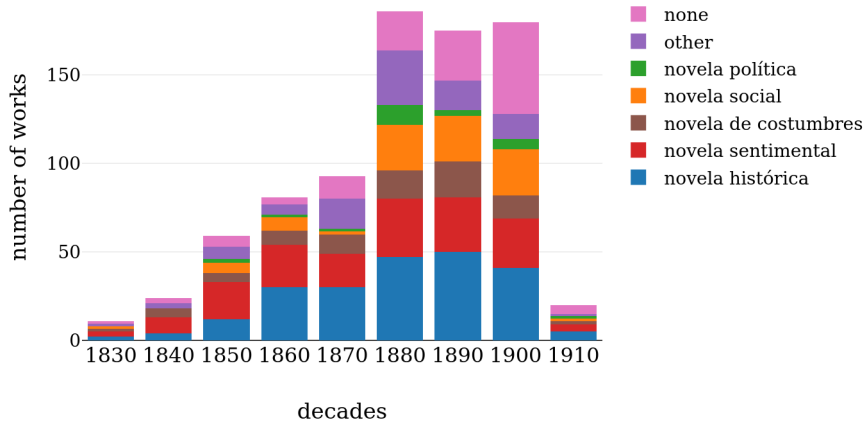


Figure 169. Primary thematic subgenre labels in Bib-ACMé per decade.

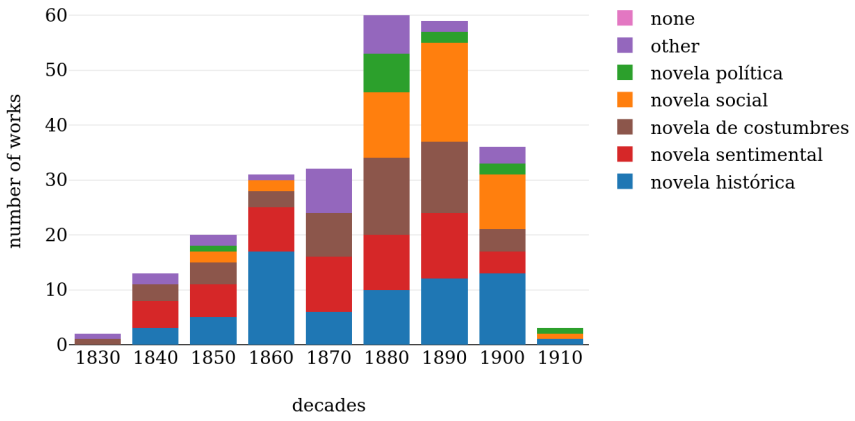


Figure 170. Primary thematic subgenre labels in Conha19 per decade.

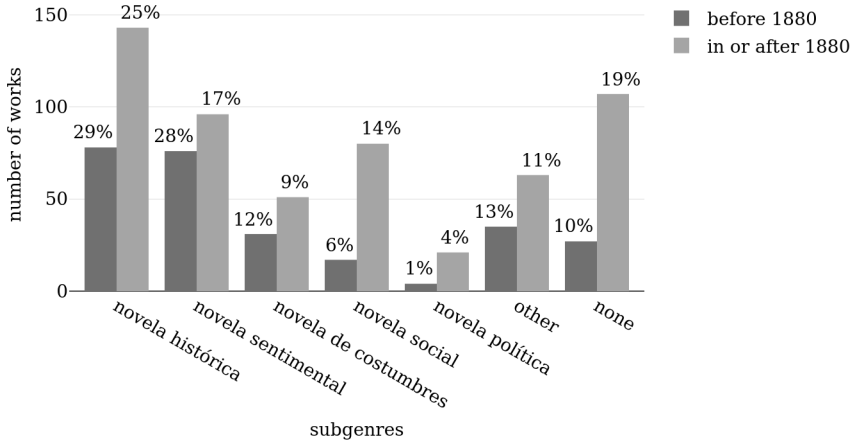


Figure 171. Primary thematic subgenres in Bib-ACMé before and in or after 1880.

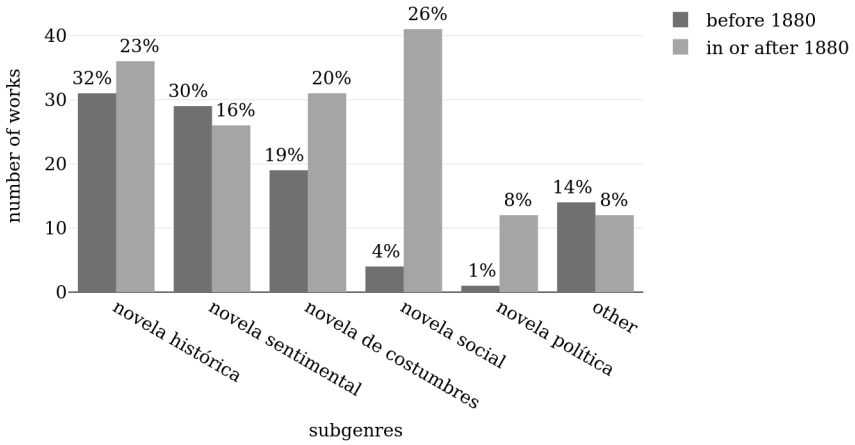


Figure 172. Primary thematic subgenres in Conha19 before and in or after 1880.

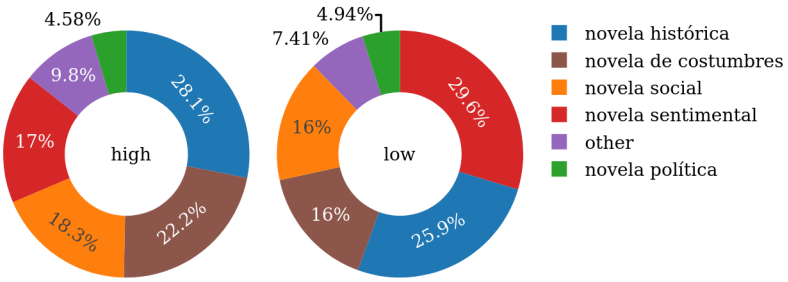


Figure 173. Primary thematic subgenre labels in Conha19 by prestige.

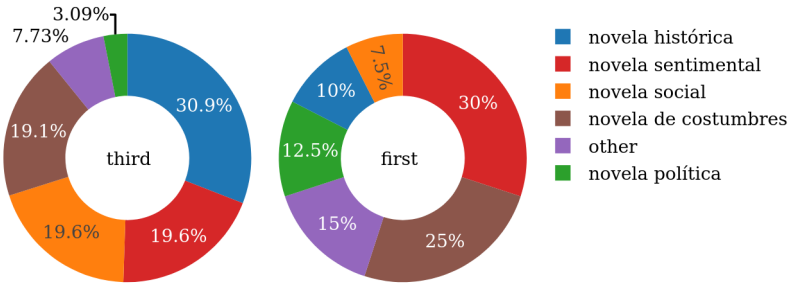


Figure 174. Primary thematic subgenre in Conha19 by narrative perspective.

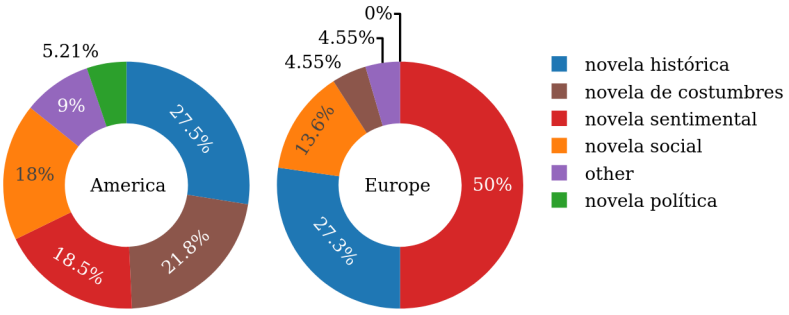


Figure 175. Primary thematic subgenres in Conha19 by continent of the setting.

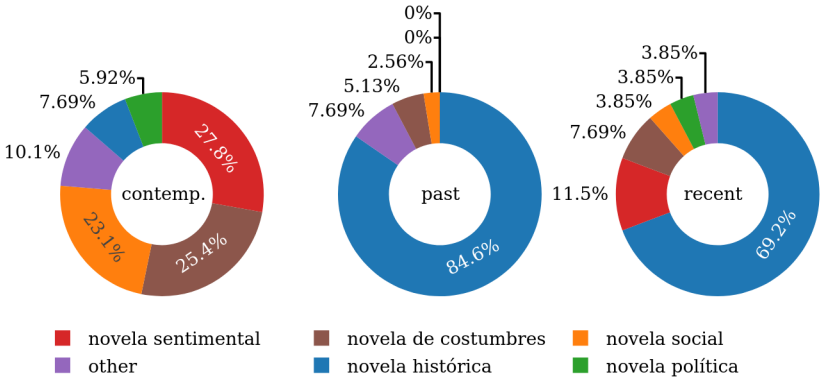


Figure 176. Primary thematic subgenres in Conha19 by time period of the setting.

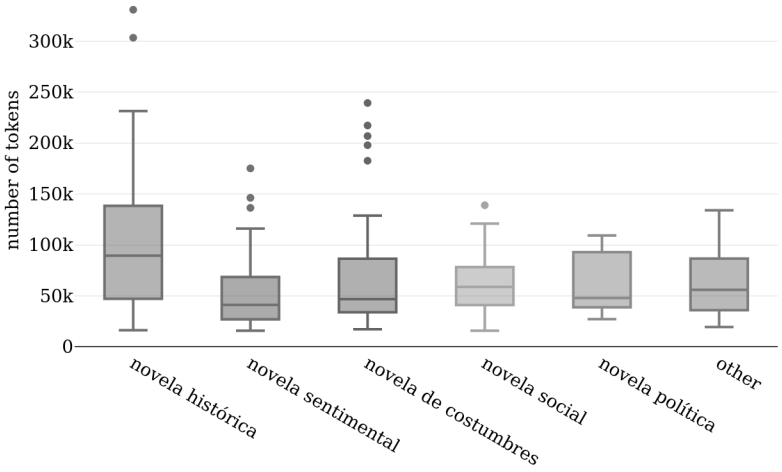


Figure 177. Work lengths in tokens by primary thematic subgenre in Conha19.

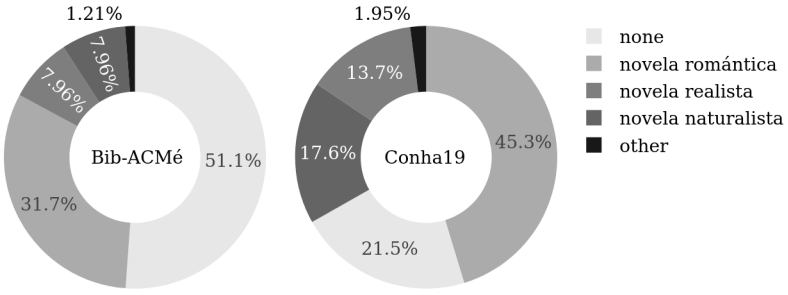


Figure 178. Primary subgenres related to literary currents in Bib-ACMé and Conha19.

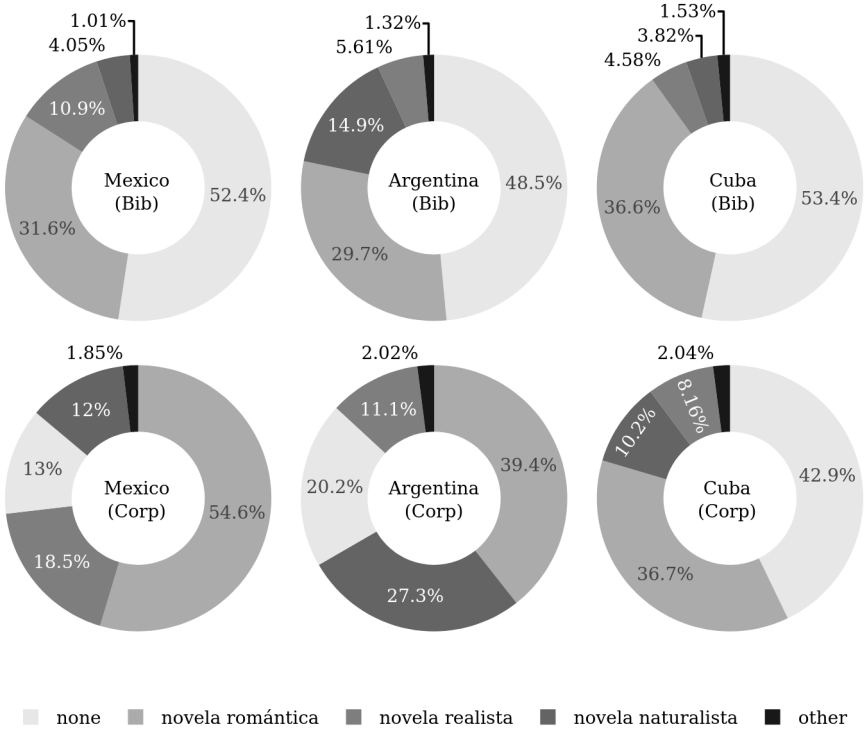


Figure 179. Primary subgenre labels related to literary currents in Bib-ACMé and Conha19 by country.

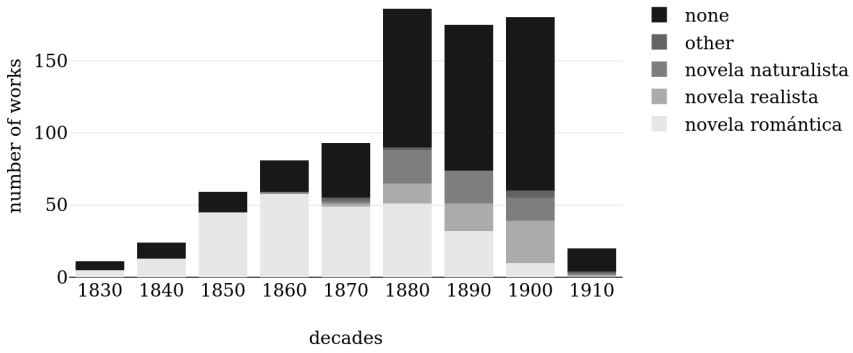


Figure 180. Primary subgenre labels related to literary currents in Bib-ACMé by decade.

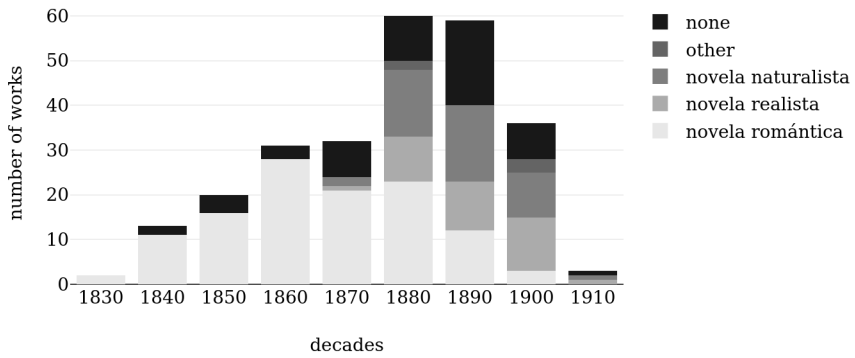


Figure 181. Primary subgenre labels related to literary currents in Conha19 by decade.

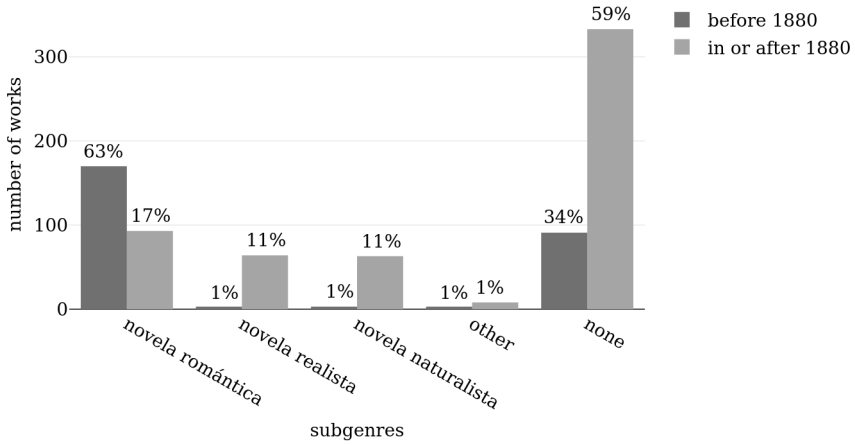


Figure 182. Primary subgenres related to literary currents in Bib-ACMé before and in or after 1880.

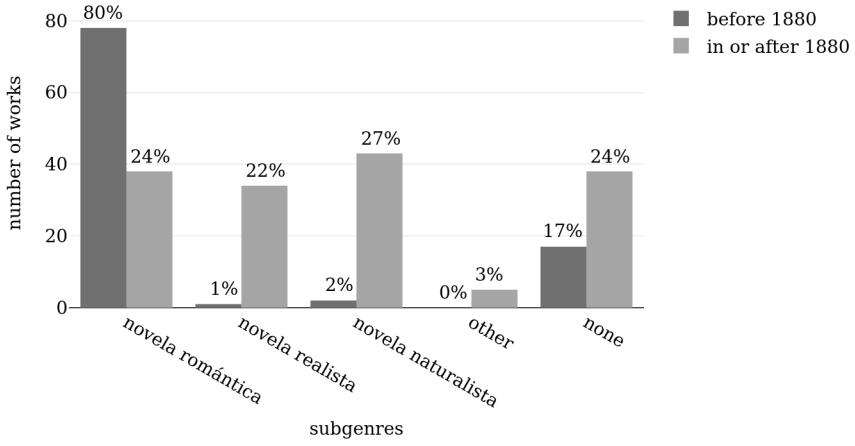


Figure 183. Primary subgenres related to literary currents in Conha19 before and in or after 1880.

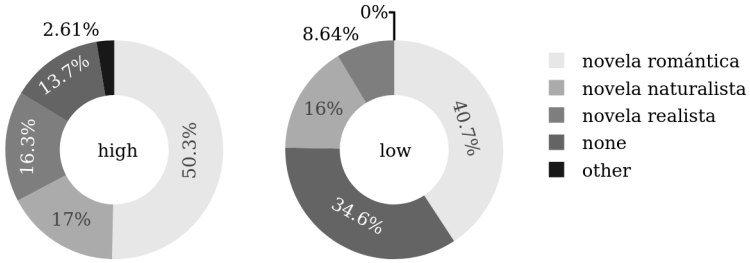


Figure 184. Primary subgenre labels related to literary currents in Conha19 by prestige.

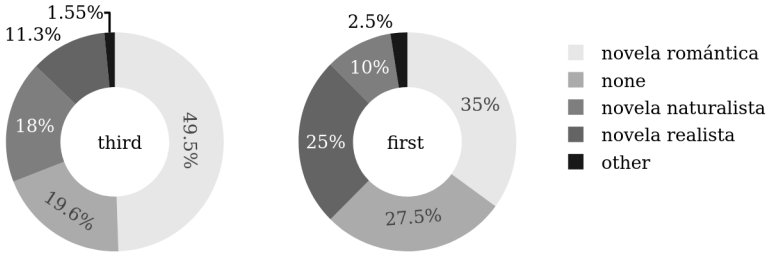


Figure 185. Primary subgenre labels related to literary currents in Conha19 by narrative perspective.

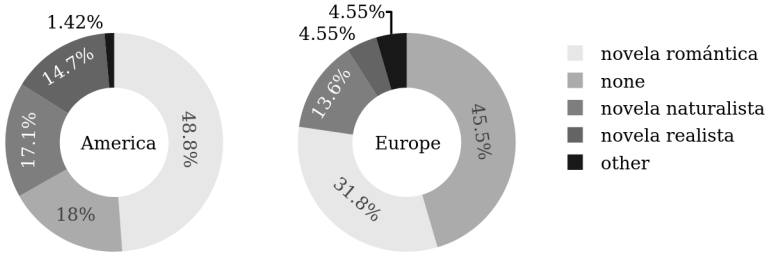


Figure 186. Primary subgenre labels related to literary currents in Conha19 by continent of the setting.

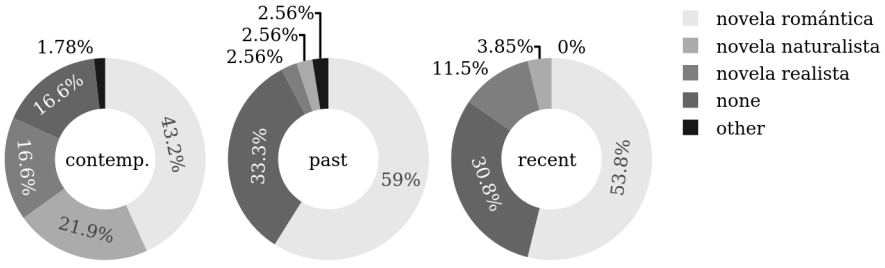


Figure 187. Primary subgenres related to literary currents in Conha19 by time period of the setting.

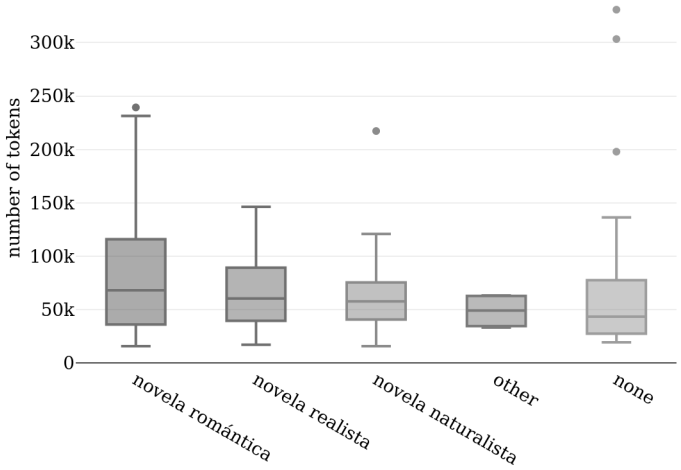


Figure 188. Work length in tokens by primary subgenres related to literary currents in Conha19.

Index of Figures

Index of Tables

Index of Examples