PHILOSOPHISCHE FAKULTÄT I

DER

JULIUS-MAXIMILIANS-UNIVERSITÄT WÜRZBURG

WISSENSCHAFTLICHE ARBEIT ZUR ERLANGUNG DES
AKADEMISCHEN GRADES EINES

MAGISTER ARTIUM (M.A.)

THEMA:

Semantic Relations in WordNet and the BNC

EINGEREICHT VON:

Oliver Ferschke

WÜRZBURG, 2009

FACH:

Englische Sprachwissenschaft

Table of contents

Abstract

Die vorliegende Arbeit beschäftigt sich mit der Untersuchung von semantischen Relationen in der lexikalischen Datenbank WordNet, sowie im British National Corpus (BNC)

Im ersten Kapitel werden die grundlegenden Begriffe erklärt, die zum Verständnis der weiteren Ausführungen bekannt sein müssen. Ausgehend von einer Begriffsbestimmung von *Semantik* und der Betrachtung verschiedener Arten von Wortbedeutungen, wird schließlich der Begriff der konzeptuellen Bedeutung eingeführt. Nach einem kurzen Exkurs in die Komponentenanalyse wird schließlich zum zentralen Thema dieser Arbeit übergeleitet, den Bedeutungsrelationen. Grundlegend wird zwischen logischen und semantischen Relationen unterschieden, von denen letztere weiter in lexikalische und konzeptuelle Relationen unterteilbar sind.

Im zweiten Kapitel sollen nun die eben erwähnten lexikalischen und konzeptuellen Relationen näher besprochen werden. Im Zuge dessen wird die lexikalische Datenbank WordNet vorgestellt, die das Lexikon der englischen Sprache mit Hilfe eben dieser Relationen in der Form eines semantischen Netzes organisiert. Nachdem zunächst kurz die Überlegungen angesprochen werden, die der Entwicklung von WordNet zugrunde lagen, werden anschließend die grundlegenden Eigenschaften von semantischen Netzen erklärt. Der erste Abschnitt des Kapitels schließt mit der Erörterung einiger Vorteile, die sich aus der netzartigen Organisation der Daten in WordNet ergeben. Die genaue Art und Beschaffenheit dieser Daten ist das Thema des zweiten Abschnitts, der die Vierteilung des semantischen Netzes anspricht und Gründe dafür nennt warum WordNet nach Wortklassen unterteilt ist. Der Hauptteil des Kapitels folgt im dritten Abschnitt, welcher sich um die eigentlichen semantischen Relationen dreht. Diese sind in dieser Arbeit nicht wie im Großteil der Literatur zu diesem Thema nach den entsprechenden Wortarten geordnet, für welche die Relationen anwendbar sind, sondern nach den grundlegenden strukturellen und konzeptuellen Eigenschaften dieser Relationen selbst. Zunächst wird die zentrale lexikalische Relation Synonymie besprochen, welche die Grundbausteine von WordNet, die Synsets, im Innersten zusammenhält.

Im Anschluss werden unter dem Überbegriff „hierarchische Relationen“ eine Reihe von Relationen besprochen, die für die hierarchische Organisation besonders der Nomen verantwortlich sind. Hyponymie verbindet Konzepte durch eine Unterbegriff-Oberbegriff Beziehung, wobei der Unterbegriff, das Hyponym, eine Spezialisierung des allgemeineren Hyperonyms ist. Meronymie, wiederum, verbindet Teilkonzepte mit ihrem Ganzen, also z.B. die Finger mit der Hand. Hierdurch lassen sich komplexe Objekte durch ihre Einzelteile definieren. Durch die Kombination dieser beiden Relationen sind aufwändige hierarchische Strukturen herstellbar. Auch wenn für die Verben die hierarchische Ordnung nicht von größter Wichtigkeit ist, beschreibt der letzte Teil des Abschnitts über hierarchische Relationen dennoch einige Varianten, die ähnliches auch für die Verben leisten.

An die Diskussion über die hierarchischen Relationen schließt sich die Betrachtung der letzten größeren Klassen von Relationen an, nämlich jene, die Ähnlichkeit und Gegensatz ausdrücken. Bei diesen Relationen, welche vor allem für die Organisation der Adjektive wichtig sind, ist es besonders interessant, dass Gegensatz häufig nicht ohne Ähnlichkeit zu definieren ist.

Nachdem die wichtigsten Relationen besprochen wurden, wird schließlich noch auf ein Phänomen hingewiesen, dass Benutzern von WordNet bekannt sein sollte – dem Einfluss von Polysemie auf semantische Relationen, welcher in allen Wortklassen beobachtbar ist. Der Abschnitt schließt mit einer Auflistung weniger wichtiger Relationen, denen bis dahin noch keine Beachtung geschenkt wurde.

Auch wenn WordNet eine gut durchdachte Auswahl an Inhalten besitzt, gibt es dennoch immer wieder Anwendungen die bestimmte Daten, Relationen oder Funktionen in WordNet vermissen. Der letzte Abschnitt im Kapitel zu WordNet will daher einen Überblick über einige Unzulänglichkeiten des semantischen Netzes geben.

Das dritte Kapitel beschäftigt sich mit einer Sichtweise von Sprache, die besonders durch ihren empirischen Charakter geprägt ist. Die Korpuslinguistik ist eine relativ neue Disziplin in der Sprachwissenschaft, die eine große, repräsentative Auswahl an Texten hernimmt und durch deren Analyse Rückschlüsse auf die Beschaffenheit der Sprache zieht. Die Grundprinzipien dieser Disziplin werden zu Beginn des dritten Kapitels beschrieben und mit anderen Sichtweisen kontrastiert.

Im darauffolgenden Abschnitt wird dann der British National Corpus im Besonderen betrachtet, da dieser den folgenden Untersuchungen zu Grunde liegt. Das anschließende Kapitel beschäftigt sich mit den Unterschieden zwischen dem wissensbasierten Ansatz von WordNet und dem empirischen Ansatz der Korpuslinguistik, welche besonders im Hinblick auf das spätere Vorhaben der Extraktion semantischer Relationen aus Textkorpora untersucht werden. Dieses Vorhaben wird schließlich im letzten Abschnitt beschrieben. Es werden mehrere Möglichkeiten aufgezeigt, wie man mit überraschend geringem Aufwand eine große Menge semantischer Relationen aus dem BNC extrahieren kann. Ein Problem stellt meist jedoch die Beurteilung der Güte und die Klassifikation dieser Daten dar. Neben detailliert beschriebenen Ansätzen zur Extraktion von semantischen Relationen mit Hilfe lexikalischer Muster werden auch andere Möglichkeiten kurz angesprochen.

Im letzen Kapitel wird eine weitere linguistische Disziplin in das bisherige Gesamtbild mit einbezogen, die kognitive Linguistik. Entstammend aus der kognitiven Psychologie will die Prototypentheorie die herkömmlichen linguistischen Modelle um den Aspekt erweitern, der die menschliche Wahrnehmung und Sprachverarbeitung betrifft. Der erste Abschnitt des Kapitels erklärt die Prototypentheorie, indem sie kontrastiv zu zwei weiteren Theorien dargestellt wird. Im Anschluss wird ein Projekt vorgestellt welches es sich zum Ziel gesetzt hat zwei Konzepte der Prototypentheorie, die Basisebenen und Prototypen, im semantischen Netz von WordNet zur Anwendung zu bringen. Mittels eines halbautomatischen Verfahrens sollen diejenigen Elemente in WordNet-Hierarchien identifiziert werden, die besondere psychologische Eigenschaften besitzen.

Dieses Verfahren und die hierzu entwickelte Software werden im Hauptteil des Kapitels beschrieben. Es wird abgeschlossen mit der Analyse von Daten, die in einer kleinen Studie erhoben wurden, welche dieses Verfahren eingesetzt hat. Auch wenn die konkreten erhobenen Daten nur teilweise brauchbar waren, lassen die Ergebnisse dieser Studie dennoch hoffen, dass Verfahren dieser Art bestehende semantische Netze um wertvolle Daten, wie zum Beispiel die Basislevelinformationen, erweitern können.

Introduction

It is not always easy to define what a word means. We can choose between a variety of possibilities, from simply pointing at the correct object as we say its name to lengthy definitions in encyclopaedias, which can sometimes fill multiple pages. Although the former approach is pretty straightforward and is also very important for first language acquisition, it is obviously not a practical solution for defining the semantics of the whole lexicon. The latter approach is more widely accepted in this context, but it turns out that defining dictionary and encyclopaedia entries is not an easy task. In order to simplify the challenge of defining the meaning of words, it is of great advantage to organize the lexicon in a way that the structure in which the words are integrated gives us information about the meaning of the words by showing their relation to other words.

These *semantic relations* are the focal point of this paper. In the first chapter, different ways to describe meaning will be discussed. It will become obvious why semantic relations are a very good instrument to organizing the lexicon.

The second chapter deals with WordNet, an electronic lexical database which follows precisely this approach. We will examine the semantic relations which are used in WordNet and we will study the distinct characteristics of each of them. Furthermore, we will see which contribution is made by which relation to the organization of the lexicon. Finally, we will look at the downside of the fact that WordNet is a manually engineered network by examining the shortcomings of WordNet.

In the third chapter, an alternative approach to linguistics is introduced. We will discuss the principles of corpus linguistics and, using the example of the British National Corpus, we will consider possibilities to extract semantic relations from language corpora which could help to overcome the deficiencies of the knowledge based approach.

In the fourth chapter, I will describe a project the goal of which is to extend WordNet by findings from cognitive linguistics. Therefore, I will discuss the development process of a piece of software that has been programmed in the course of this thesis. Furthermore, the results from a small-scale study using this software will be analysed and evaluated in order to check for the success of the project.

1 A wide view on semantic relations

Semantics is the branch of linguistics which studies meaning in language. While this definition is as short as it is simple and even understandable for the non-linguist, it is actually not a satisfying definition for a professional in linguistics. (Löbner 2003, 3) The main reason for this is that the word *meaning*, which is here used to define *semantics*, has many facets and can be examined from various points of view. In order to comprehend the concerns of semantics, one must take a closer look at the meanings of *meaning* in linguistics.

Leech (1974, 10–27) distinguishes seven types of meaning, which can further be subcategorized into 3 categories; thematic meaning, associative meaning and conceptual meaning. "Thematic meaning is mainly a matter of choice between alternative grammatical instructions […]" (Leech 1974, 22 f.), lexemes or intonations in order to create emphasis. Associative meaning is a summary term for several meaning types with open-ended and indeterminate character. *Meaning* in the associative sense is very subjective and sensitive to the communication situation and the particular sender and recipients in the communication process. (Leech 1974, 14–22) For this thesis, however, the third category of meaning is important. The conceptual meaning "has a complex and sophisticated organization [...]" with contrastive features and a constituent structure. (Leech 1974, 11) It represents the core sense of a word and includes only the basic semantic properties that are needed to distinguish it from other words. It does not contain any connotations and is thus also called *denotative meaning*. (Herbst et al. 1991, 160) As Leech points out, it is not always easy to differentiate between the conceptual meaning and the other peripheral categories. Therefore, the meanings of two words can sometimes not be compared by only relying on one level, but have to be analysed on other planes of meaning as well. (Leech 1974, 25)

What we are now interested in is how to describe the conceptual meaning of a word. There are, however, two facts that we have to note first. One is that there are various approaches to semantics with different perspectives on meaning, which all have an extrinsic frame of reference. So far, there is no comprehensive semantic system with solely an intrinsic frame of reference. This means that semantic descriptions of meaning are always sensitive to the environment that they are used in and can only be understood in the context of this environment. (Leech 1974, 4) The other fact is that any semantic system has to use language in order to describe the meaning of language. This is even the case when the description is not carried out by words of a natural language, but by mathematical or chemical symbols or any other kind of scientific formula. (Leech 1974, 5)

Encyclopaedic descriptions of word meanings are very hard to compile. Even for alleged simple words, it is virtually impossible to create a dictionary entry that offers a sufficient description for any imaginable context. One only needs to imagine a description of the abstract concept of *love* which covers all its facets and is suitable as well for a philosopher as for a psychiatrist and a journalist. Because of the dependency on the different extrinsic frames of reference which the various users of this word have, such a universal encyclopaedic description cannot exist. Consequently, other means of meaning representation are necessary. We basically have the choice between looking further inside the meaning of a single word by examining its constituents or we can study the relations of one word to other words.

A method that follows the first approach is called *componential analysis*. It breaks down the complex meaning of a word into its constituents which are more easily to describe. We can regard componential analysis as a recursive [1]definition of meaning. This becomes clear by looking at the following diagram

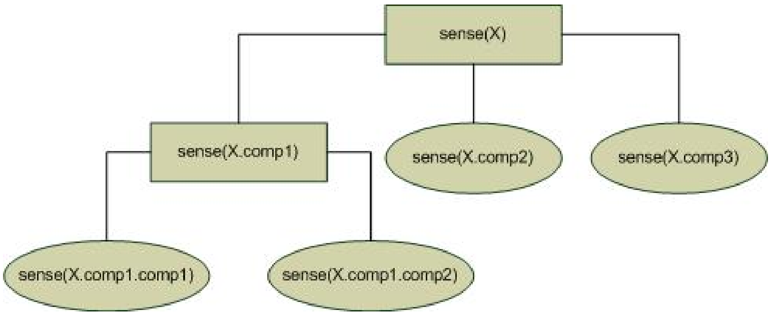


figure 1: Componential analysis

[1] The solution of a recursive problem depends on the solutions to smaller instances of the same problem. (Graham et al. 1990, 1)

The meaning of a word is described by its components which are again described by their components until we have reached an atomic level. Most of those atomic components are semantic oppositions. Their meaning can be defined much more easily than the meaning of the original word. (Leech 1974, 95 ff.)

Methods following the other approach, as already mentioned above, try to define the meaning of a word by its relationships to other words. The principle here is that we do not have to specify the meaning of a word down to the greatest detail, but by looking at the position of the word in the network of relations and by examining its links to other lexemes, we can deduce a surprisingly comprehensive definition of its meaning. The way this method works can be compared to divide&conquer-algorithms. The complexity of defining the meaning of one word is shared with all the other words. Every word uses others to define itself. Instead of having to know much about a single lexeme, we only have to know little about many other words. The relations between these pieces of information then give us comprehensive descriptions of meaning. For example, the definition of *sports car* is not that hard if we know what a *car* is. We only have to know about the differences between the two. The term *car*, in turn, may be defined by a definition of all its components. This again makes it easier to gain a complete picture of the meaning of this word. Relations of this kind will be the focal point of the investigations in the following chapters.

When browsing the literature that deals with relations between words and sentences, one comes across many types of relations and even more technical terms which describe them. However, we can basically distinguish two types of relations, logical and semantic relations.

The logical view of language is a separate complex topic that will not be discussed here in detail. In a nutshell, from the logics' point of view, sentences are compositions of single propositions. The logical combination of the truth values of these propositions can be resolved to maintain the truth value of the whole sentence, for example

(1) Marc is ten years old = true *(single proposition)*

(2) Marc has red hair = true *(single proposition)*

(3) Marc is ten years old and has red hair = true *(composition of propositions)*

In sentence (3), the single propositions (1) and (2) have been combined with the logical operator *AND*, which results in the truth value *true.* The problem is that we cannot logically combine every proposition with any other proposition.

Therefore, we need logical relations between the propositions that dictate which combinations are possible and which are not. We distinguish four different relations

(1) Implication (A implies B: if A is true, B must be true)
(2) Equivalence (Two propositions always have the same truth value)
(3) Contradiction (Two propositions always have opposite truth values)
(4) Contrariety (A is contrary to B: if A is true, B must be false)

For two related propositions A and B in the same context of utterance, we can deduce the truth value of one proposition from the other. The following propositions, for example, are contrary.

(a) It is hot today
(b) It is cold today

Consequently, if we know that (a) is true, we also know that (b) must be false in the same context of utterance. The relation also dictates that both propositions may not be combined with the operator AND. (Löbner 2003, 80–102)

In order to identify logical relations between single words, we have to form a sentence with each word that expresses a single proposition. Depending on the relation between the resulting test sentences, we can then identify one of the following logical word relations

(5) Equivalence (test sentences are equivalent)
(6) Subordination (test sentences form an implication)
(7) Incompatibility (denotations of both words are disjoint)
(8) Complementarity (incompatibility between binary oppositions[2])
(Löbner 2003, 102–106)

[2] Mathematical definition: A set X that is complementary to the set Y includes everything that is not included in set Y. Absolute complementarity therefore does not exist in natural language, but it is used for binary oppositions. (Löbner 2003, 105 f.)

Logical relations between words and sentences only make statements about truth values and denotations. They are not relations between meanings. This can be seen best when looking at the following two contradictory, but non-contingent sentences

(a) Ice is hot

(b) Bees can fly.

These sentences are contradictory, because (a) is always false and (b) is always true. However, there are no relations between their meanings. (Löbner 2003, 97) Also for contingent sentences, logical relations give only limited information about their meaning. Therefore, semantic relations are required to make statements about the relations between word meanings.

These relations will be the subject of investigation in the following chapters. Before stepping further into this topic, one last terminological distinction has to be made. Semantic relations are generally divided into two categories, the lexical relations and the conceptual semantic relations. The difference between the two is that lexical relations link words, whereas conceptual semantic relations link concepts. (Fellbaum 1999b, 9)

2 WordNet

In the following chapter, the electronic lexical database *WordNet* will be introduced. We will first discuss the ideas behind its design and implementation and we will also touch on its development since the project has first started in 1985. Afterwards, a brief overview of the content of WordNet will be given. The largest part of the chapter, however, will be concerned with a detailed examination of the semantic relations that have been used in WordNet. Although the database has grown considerably during the previous two decades, there are still limitations that should be considered in a complete discussion of WordNet. This will be the topic of the last section in this chapter.

2.1 Introducing WordNet

In the mid-fifteenth century, Johannes Gutenberg changed our view on language. His letterpress made it possible to put manifold texts on paper and thus make them available to the public. From then on, it was by far easier than in the days before Gutenberg to preserve information and knowledge by the printed word. At the end of the twentieth century, technology once again began to have a stake in our view on language. The ability to store any kind of text on paper, or later electronically, had bestowed great amounts of data on us, but it was not before the advent of the computer that this cornucopia of information could be processed to the extent that a new insight into language became possible.

WordNet was one product of this computational revolution. At first, componential lexical semantics had been regarded as the appropriate approach to computational natural language processing. However, it turned out not to be the best theory for this purpose, because even after ten years of research, the basic atoms of language could not be found. Therefore, an alternative was needed. In the nineteen-eighties cognitive psychologists and computational linguists followed an idea that had been around since the German linguist Jost Trier (1931), who first introduced the study of semantic fields (*Wortfeldforschung*) in the nineteen-thirties. Based on his findings, Quillian (1966) discovered that humans organize their conceptional knowledge in network structures. G.A. Miller adapted this idea and built a small semantic network of 45 nouns. The question that arose from this proof of concept was whether this approach could scale up to become a model of the whole lexicon. Thus, the WordNet project was created to answer this question. As the database of WordNet grew, it became not only apparent that relational semantics was a satisfying theory for computational linguistics, but it also seemed that WordNet had "intrinsic merit of its own." (Miller 1999a, xvi–xvii)

Today, WordNet has developed into a very large semantic network, containing over 150.000 unique strings. (Cognitive Science Lab, Princeton University 22.08.2007) Semantic networks are "a class of knowledge representation formalisms", consisting of *nodes*, which represent *objects*, *concepts* or *situations*, and *arcs*, which represent the *relations* between the nodes. (Barr et al. 1981, 180)

A node in WordNet represents a single semantic concept. It is encoded as a so-called *synset*, which is a set of lexical items that have the same meaning in a given context, i.e. they can be used interchangeably in one sentence without changing its meaning. (Miller 1999b, 24) A lexical item usually has more than one possible meaning. Consequently, it can be part of more than one synset at the same time.

The arcs in WordNet represent semantic relations. As we have already seen in chapter 0, there are two categories of semantic relations. The majority of the arcs establishes links between the synsets and thus represents the conceptual semantic relations. A quarter of the arcs, however, relates the individual lexical items to each other and thus represents the lexical relations.[3] The various kinds of relations that exist in WordNet will be subject of discussion in chapter 3.3.

The implementation of the lexicon as an electronic semantic network has several advantages over the static print format. In printed representations, the underlying structure of the lexicon is covert. WordNet explicitly shows this structure and puts less emphasis on the single word entries. The advantage for the user is that they are not restricted in their searches to a predefined order, but they can access the information along various paths in the network. (Fellbaum 1999b, 7–9)

Although WordNet shares several characteristics with traditional print representations of the lexicon, like thesauri and dictionaries, its structure offers additional

[3] The proportion between lexical and conceptual relations are taken from a statistic of pointers in WordNet 3.0 in (Finlayson 16.04.2008, 6)

information that cannot be learned from plain word lists. Thesauri, for example, only provide entries for concepts that are lexicalized. (Fellbaum 1999b, 7–9) Not every concept, however, that may occur in a person's mind, is lexicalized by the English language, or any other language for that matter.

> For instance, it is arguable that most of us have a non-lexicalised concept of *uncle-or-aunt*. We have many beliefs and expectations about uncles-or-aunts (i.e. siblings of parents, and, by extension, their spouses). It makes sense to assume that these beliefs and expectations are mentally stored together in a non-lexicalised mental concept, which has the lexicalised concepts of uncle and aunt as sub-categories. (Sperber, Wilson 1997, 6)

Semantic networks underlie the rules of the *closed-world assumption*. They only work correctly if the information provided is complete. (Russell, Norvig 2007, 355) That is why the non-lexicalized concepts that do not even appear in thesauri disturb the structure of the network. Without them, the connectedness of the whole network would not be guaranteed. This means that either semantic networks are inadequate models of the lexicon or, what is more likely, our lexicon is incomplete. By creating a semantic net of a whole language, the lexical gaps in the lexicon can be identified, which would not be easy without the information of such a relational network. In WordNet, the gaps are filled with pseudo words that describe these non-lexicalized concepts, for example *bad person* or *wheeled vehicle*. (Fellbaum 1996, 223 f.)

2.2 The contents of WordNet

Although, WordNet is one large semantic network[4], it consists of four separate independent sub-networks, one for each word class. In each subnet, concepts can only be lexicalized by a lexeme from the corresponding word class, according to the assumption that only verbs or deverbal nouns can lexicalize concepts that express actions, only nouns can name things, persons etc. and only modifiers can refer to the attribute values of the other two.

[4] To be more precise, the network has one large connected component that includes more than 99% of all words. (Steyvers, Tenenbaum 2005, 53)

The main reason for this separation is that each word class supports a different range of semantic relations. Therefore, the nodes in each subnet can only be connected by relations that are available in the corresponding word class. (Grabowski et al. 1996, 212 f.)

Furthermore, each word class has its own means of organization. Nouns are arranged in hierarchical taxonomies, whereas verbs are organized in semantic domains. Among the modifiers, only the adjectives have a particular order, which we call clusters. Adverbs do not have a dedicated topology in WordNet. Both aspects, the organization and, above all, the relations in the individual subnets will be the main focus of chapter 3.3.

The overall connectedness of WordNet is accomplished by pointers that interlink the four subnets. The following table illustrates how they are linked together to form a single semantic network. (Fellbaum 1999c, part 1)

Subnet	**➔**	**Connected with subnet**	**Type of connection**
verb	→	noun	*verb domain*
adjective (descriptive)	→ →	verb noun	*principal part of* *adjective domain*
adjective (relational)	→	noun	*related nouns*
adverb	→	adjective	*derived from*

Table 1: Interconnections between the four sub-networks in WordNet

“Although WordNet contains compounds, phrasal verbs, collocations, and idiomatic phrases, the word is the basic unit.”(Fellbaum 1999b, 4) Smaller phrases sometimes have to be used to paraphrase the already mentioned lexical gaps. Besides that, most of the synsets are accompanied by explanatory glosses that facilitate the distinction of the different word senses. (Fellbaum 1999b, 5–7)

The data of WordNet were first based on the Brown Corpus, but several other sources have been assimilated over the time. New words, however, have always been integrated manually. (Miller 1999a, xviii ff.)

2.3 Semantic relations in WordNet

A question that developers of semantic networks like WordNet have to face is how many semantic relations actually exist and, if there is a very large or even infinite number, which of them are the most important ones for the structuring of the lexicon. The answer to this question depends on the environment and the application in which the relations are finally examined.

In domain specific applications, for instance, one can identify many semantic relations that are only valid in this particular field. Rosario and Hearst (2001), for example, have discovered almost forty different semantic relations only within the medical domain by empirically analysing biomedical texts. With the same method, one can identify new relations for other domains as well.

WordNet, however, is not domain specific. Therefore, it must only contain semantic relations that are not restricted to a definite field. Analogous to the fact that the English language has closed and open word classes, we can also identify closed and open classes of semantic relations. "The closed class corresponds to those relationships expressed linguistically through paradigmatic relationships [....]. The open class corresponds to those [...] expressed linguistically through syntagmatic relationships." (Green 2001, 6) Since WordNet treats the four open word classes separately, it is automatically restricted to the closed class of paradigmatic relations.[5] (Fellbaum 1999b, 9) From this class, the developers of WordNet have selected the relations that "were believed to be the semantic relations of broadest applicability and greatest familiarity."(Miller 1999b, 36)

The following section will discuss the synonymy relation, which is the most important lexical relation in the whole semantic network. The sections thereafter will then account for the different semantic relations within the four major word classes. Because these semantic relations are the central topic of this thesis, they will not be ordered by their respective word classes, like many other papers on this topic do, but they will be arranged according to their characteristic features and their ability to organize the lexicon. After all major relations have been described, we will discuss the role of polysemy in context of these relations. Finally, in section

[5] See also chapter 3.4

3.3.5, a few minor relations are listed which are not of vital importance for the organization of the words, but which should still be accounted for.

2.3.1 Synonymy

The lexical relation of *synonymy* can be regarded as the central semantic relation in WordNet. As we have seen in chapter 3.1, the nodes in WordNet's semantic network consist of all the lexical items that refer to a single concept. These lexical items are linked by the synonymy relation, hence the name *synset*. Figure 2 illustrates the concept of synonymy between two symbols using Ogden and Richard's semiotic triangle. According to this definition, two words are synonymous if they always induce the same thought and in doing so refer to the same referent. If this is the case for any imaginable context, the two words are absolute synonyms i.e. they are identical in respect of both all central and peripheral semantic traits. (cf. Cruse 1986, 265 ff.)

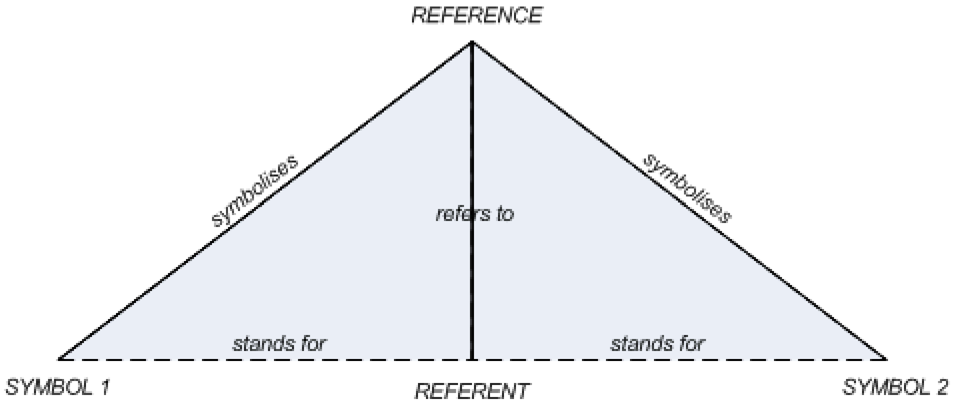


figure 2: Semiotic triangle for (absolute) synonyms

Absolute synonymy, however, is very rare if not nonexistent. It would be uneconomical, if a language had several words that do not differ in the slightest aspect of meaning. That is why most synonyms actually diverge in at least one minor semantic trait. Consequently, the level of synonymy has to be graded on a scale between absolute synonymy and non-synonymy and it depends both on the degree of semantic overlap and on the degree of implicit contrastiveness. (cf. Cruse 1986)

In WordNet, synonymy resembles to a large extent what Cruse (1986, 88) calls *cognitive synonymy*. "X is a cognitive synonym of Y if (i) X and Y are syntactically identical, and (ii) any grammatical declarative sentence S containing X has equivalent truth conditions to another sentence S^1, which is identical to S except that X is replaced by Y." (Cruse 1986, 88) Cognitive synonymous words can still differ in either their presupposed or their evoked meaning, which means that they do not induce exactly the same thought into the recipient, as figure 2 implies, but that they either co-activate other concepts or that they impose usage restrictions. Cruse analyses these possible differences between cognitive synonyms in greater detail. (Cruse 1986, 270–285) WordNet, on the other hand, "does not account for such [...] differences by means of relations among synsets [...]. However, the information given in parentheses (a gloss and one or more sample sentences) often spells out the specific usage restrictions associated with individual [words][6]." (Fellbaum 1999a, 73)

Cognitive synonymy is a bidirectional relation. This means that for two given cognitive synonyms X and Y, one can replace both X by Y and also Y by X. This characteristic is necessary, because every item in one synset should be able to lexicalize the same concept.

There are, however, unidirectional synonymy relations, which disqualify for the use in synset definition. Nevertheless, these relations are still necessary to account for in a lexical database. These kinds of synonyms, which Cruse (1986) calls *plesionyms*, often have an additional hyponymous relation (cf. chapter 3.3.2.1). They allow us to replace a lexical item by its more general hypernym, but we may not do so the other way round. In the WordNet program, these synonyms are accessible via a *synonymy query*, which basically returns the direct hypernyms of the current search word. A synonymy query for the word *lie*, for example, returns, among others, the more general terms *untruth* and *falsehood*, because a lie necessarily refers to a statement that is "false" or "not true". The reverse query, a search for the synonyms of *untruth* and *falsehood*, does not return the word *lie*. This is also logical, because a false statement does not necessarily have to be a lie. Consequently, because of the asymmetry in this synonymy relation, there is no synset that contains as well *lie* as *untruth* or *falsehood*.

[6] Fellbaum originally talks about '*verbs*', but this is also true for cognitive synonyms from other word classes.

2.3.2 Hierarchical relations

As we have seen so far, synonymy is such an important relation in WordNet, because it connects lexical items to form synsets which again are the building blocks of the semantic network. The most important conceptual semantic relations between these synsets, however, are the hierarchical relations. There are various possibilities how a network can be structured. The fact that WordNet has mostly a hierarchical structure is due to the predominant influence of these hierarchical relations.

Hierarchies have always been used to organize knowledge. Aristotle already tried in his text *Katēgoriai* to place “every object of human apprehension under one of ten categories”. (“Categories (Aristotle).” 13.11.2008). These categories have later been picked up by Porphyr, who created the *abor porphyriana*, a hierarchical classification system which is ordered from the most general concept at the top to the most specific concept at the bottom. Each level in the hierarchy differs in the *differentiae specificae*, i.e. the properties that differentiate the more specific level from the generic level. (Mihatsch 2006, 1)

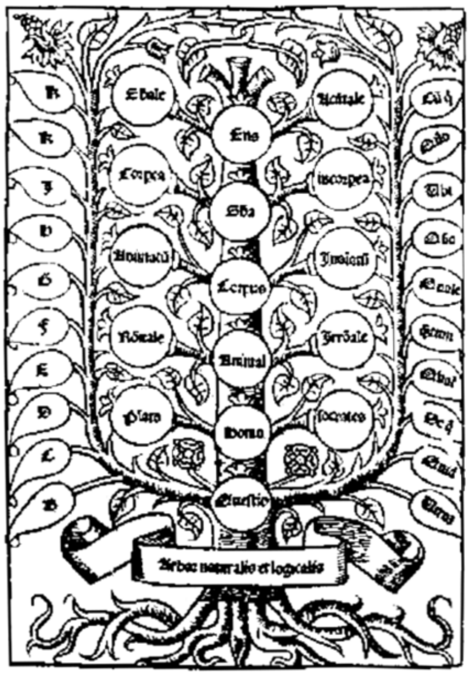


figure 3: Arbor porphyriana {Mihatsch 2006 #69: 2}

The benefit of a hierarchical model of categories is what Rosch (1978, 28 f.) calls *cognitive economy*. “A category system [provides] maximum information with the least cognitive effort. […] [It] reduce[s] the infinite differences among stimuli to behaviorally and cognitively usable proportions."(Rosch 1978, 28 f.)[7] Each concept within the hierarchy has not to be defined by the totality of its properties, but only by the description of its superordinate and the differentiae specificae. This allows us not only to describe and store concepts more economically, we can furthermore select the level of abstraction on which we want to refer to a referent. In choosing the concept that is as near as possible to the top of the hierarchy and still contains all the properties that are needed to express the specific thought or concept, we can avoid an overspecification of the statement. For example, we do not have to say *"The straight chair collapsed when he sat down"*, when the specific type of chair is irrelevant for the statement. However, we do have to make this specification in the sentence *"Please do not use the armchair, it is broken."*, because here we need the additional information to know which chair is meant by the speaker.

Rosch's (1978, 29) second principle of categorization "asserts that […] the perceived world is not an unstructured total set of equiprobable concurring attributes. Rather, the material objecs of the world are perceived to possess […] high correlational structure." Accordingly, this hierarchical categorization must also be an adequate instrument to structure the lexicon hierarchically, because words describe the perceived world. Both these facts and the discovery of cognitive scientists that the human mental lexicon also makes use of these structures (Anderson et al. 2001, 153–156), have made hierarchical relations that important for WordNet.

In the following, generic hierarchies will be referred to as *taxonomies*[8]. In order to be able to organize concepts in taxonomies, the semantic relations between superordinate and subordinate levels have to be defined. A specific taxonomic structure then gets its name from this relation. For example, a taxonomy that uses a hyponymy relation can be referred to as a hyponymous structure. On the horizontal dimension, all concepts on one level are generically called *co-taxonyms,* or *co-hyponyms* in the above example. The following sections will discuss the various relations that occur in WordNet's taxonomies in detail. Before we come to that, we want to take a brief look at the structural properties of taxonomies.

[7] Rosch's definition of *category* is similar to the definition of *synsets*. "By category is meant a number of objects that are considered equivalent."(Rosch 1978, 30)

[8] In contrast to Cruse's definition (1986, 137), who defines *taxonomy* as a sub-species of *hyponomy*.

Although taxonomic structures can be represented in many ways, the most common way to display them is as tree structures. A *tree* is a concept from graph theory and is defined as an "acyclic connected graph, where each node has a set of zero or more children nodes, and at most one parent node". ("Tree (data structure)." 12.11.2008) Every tree has exactly one node that has no parent. This node is called the *root* of the tree[9]. It contains the most generic concept which is also the greatest common denominator or alternatively the organizational unit for all the concepts that follow further down the hierarchy. Therefore, great care has to be taken in choosing the appropriate root concept. In WordNet, these roots are called *unique beginners*. They have been chosen according to psycholinguistic evidence about how people organize their lexical knowledge. (Fellbaum 1999a, 72) For the noun class, 25 unique beginners had been discovered at first, which have then further been grouped so that the number of beginners was reduced to 11 (cf. appendix 1). The hierarchies headed by these 11 synsets "correspond to relatively distinct semantic fields, each with its own vocabulary." (Miller 1999b, 28) "As of WordNet 2.1, the[se] hierarchies have been merged into a single hierarchy headed by the unique beginner{entity}."(Scriver 11.05.2006, 12, cf. appendix 2) Nevertheless, it might still be useful to keep the original division in mind.

As it has already been mentioned, the verb class, unlike nouns, is organized in semantic domains. Accordingly, unique beginners had to be selected for each of the 14 domains. However, "[w]ithin a single semantic field it is frequently the case that not all verbs can be grouped under a single unique beginner; some semantic domains can be represented only by several independent trees." (Fellbaum 1999a, 72)

All taxonomies in WordNet fulfil the requirement of connectedness, which means that within one hierarchy, every node can be related to every other node by the particular semantic relations that are used in the hierarchy. There are however instances of taxonomies in which the precondition of acyclicity is violated, which could lead to problems with the reasoning process. This will be further explained in the following section.

[9] To be precise, every node that has at least one child is the root of a subtree within the taxonomy. That is why the tree structures are usually defined recursively. (Knuth 2007, 308) In this case, however, we refer to the single root of the whole taxonomy.

Figure 4 shows the hierarchical relations that will be discussed in the following sections and their distribution among the word classes.

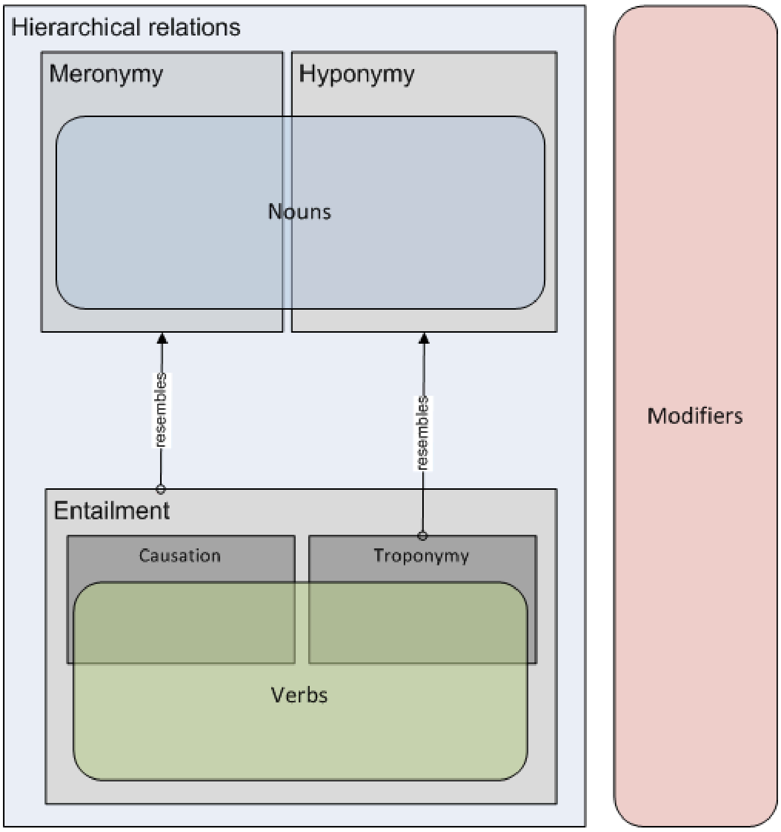


figure 4: Hierarchical relations in WordNet

2.3.2.1 Hyponymy

Under the label of hyponymy, we actually want to discuss two separate reflexive relations, *hypernymy* and *hyponymy*, which are most important for the nouns in WordNet.

Hypernymy is the hierarchical relation of *extensional inclusion.* (Mihatsch 2006, 36) It is a unidirectional relation that connects a *hypernym* or superordinate concept to one or more *hyponyms* or subordinate concepts. In reverse, every hyponym is connected to its hypernym by the hyponymy relation, which is the hierarchical relation of *inheritance* and *intensional inclusion* (Mihatsch 2006, 36) (cf. figure 5). The effect of these relations is that every subordinate concept inherits the properties that have been defined in the superordinate concept and adds its own individual characteristics. For example, the concept *animal* is the hypernym of the concepts *bird* and *cat*, because both bird and cat share the features that are common for all animals. In turn, *bird* and *cat* are hyponyms of *animal*. For these concepts, only the features have to be defined that have not already been inherited by the superordinate concept *animal*.

Extensional inclusion means that the extensions of all hyponyms are included in the extension of their hypernym, which means that every instance of a hyponym concept can also be referred to by any instance of the hypernym concept. In our example, we can choose to refer to the bird or to the cat as an animal.

Intensional inclusion means that every hyponym includes all the properties of its hypernym. In this case we could also speak of feature inheritance.

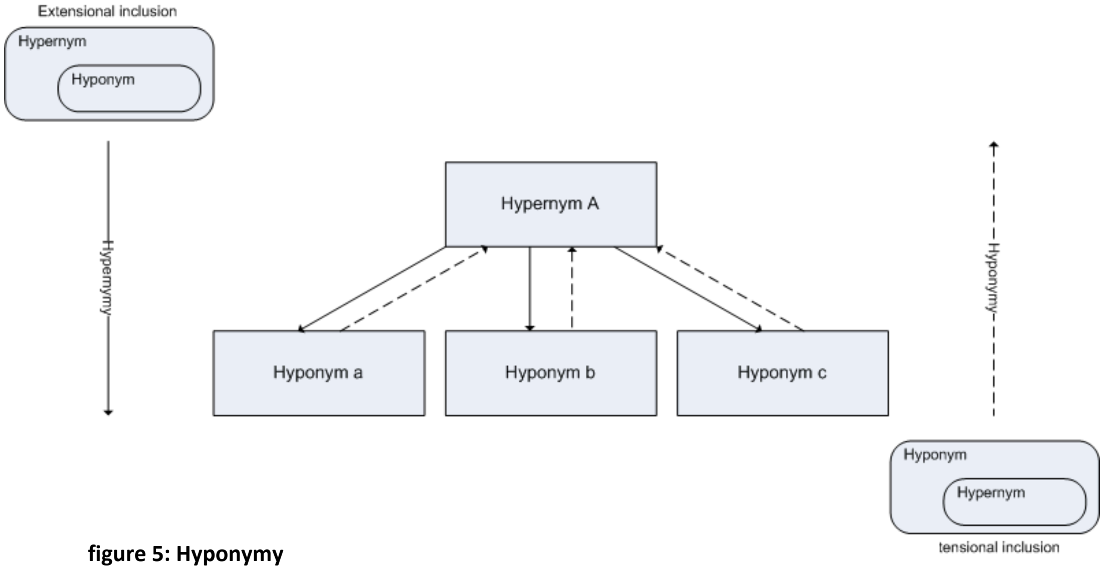


figure 5: Hyponymy

In WordNet, only hyponymy relations are directly coded in the database. The hypernymous connections have to be inferred by reversing the direction of the hyponymous ones. (Miller 1999b, 26)

As it already has been mentioned before, the hierarchical relations are assigned to the vertical dimension of the taxonomy. On the horizontal axis, the individual nodes of a hyponymous structure are related by *co-hyponymy*. All nodes on one level which share the same hypernym are called *co-hyponyms*. As is the case with hypernymy relations, the co-hyponymy relation has also to be inferred in WordNet, which can be done by “combining a generic step and a specific step” (Miller 1999b, 26). Co-hyponyms are similar to non-gradable contrary antonyms. They have disjoint extensions, but their intensions overlap in the features that have been inherited from the hypernym. (Mihatsch 2006, 36) According to Cruse (1986, 137), co-hyponyms are often, but do not need to be, incompatible. He argues that both *novel* and *paperback* are direct hyponyms of *book*, although they are not incompatibles. This can also be seen in WordNet. However, I regard this to be an inconsistency in the hierarchy. Both *novel* and *paperback* refer to *book* as a physical entity, but they each refer to a different set of features. The intension of the first basically includes the information that the book may contain a story; the intension of the latter contains properties regarding the appearance of the book. The extensions of the two overlap, which is inconsistent with the previous definition of co-hyponymy. Strictly speaking, this is no case of co-hyponymy, but of hyponymy, because a novel could be a paperback and the other way round.

In his example, Cruse most likely referred to *novel* in the abstract sense as "an invented story in prose, long enough to fill a complete book" (Hornby 2000, 757). In WordNet, we get an adequate example when replacing *novel* by *cookbook.* On the lexical level, both are hyponyms of *book*. However, because WordNet uses hyponymy as a conceptual relation, *cookbook* and *paperback* are hyponyms of two different senses of *book, {book}* and *{book, volume}*. Therefore, they are no real co-hyponyms, but they are so-called *cousins*. (Miller 1999b, 42) This will further be elaborated when discussing the influence of polysemy on semantic relations in WordNet.

A major problem of hyponymies in WordNet is that there is only one type of hyponymy relation (consequently there is also only one type of hypernymy relation). There are, though, different ways to interpret the relation between a subordinate A and a superordinate B. We could either say "A **IS A** B" or "A **IS USED AS A KIND OF** B". WordNet does not differentiate between these two interpretations of hyponymy, which are also called *formal* and *telic*. This is not a problem as long as a concept has only formal or only telic connections. If it has both, however, we have to establish two hyponymy relations for the same concept. This, however, is a violation of the already mentioned precondition of acyclicity. In creating a link to two superordinates, we automatically "create" the inferred reflexive relations, which ultimately causes a circle in the tree graph. The result is that there can be more than one path from a node to the root of the hierarchy. With linear reasoning, we are only able to discover one path. For the synset {written_agreement}, for example, we would either be able to retrieve the hypernym {legal_document} or {agreement}, but linear reasoning does not account for two possibilities (Miller

1999b, 35). This dilemma of multiple inheritance is also called the *diamond problem,* because of the shape of the inheritance structure. (Truyen et al.)

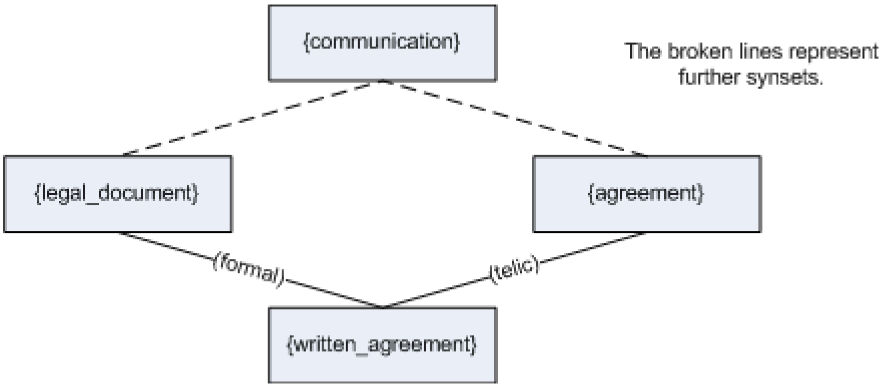


figure 6: The diamond problem

There are basically two solutions to the problem, which would allow the usage of linear reasoning in WordNet without exceptions. Either, the hyponymy relation has to be split into two different relations, which would result in two separate acyclic hierarchies (figure 8). Alternatively, the critical nodes with two parent concepts could be split into two separate concepts, one for the formal sense and one for the telic sense (figure 7). (cf. Truyen et al. and Miller 1999b, 35)

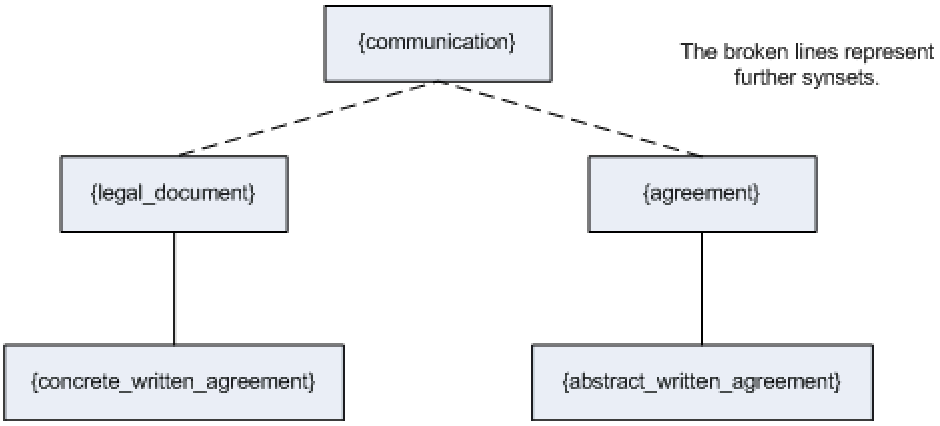

{communication}
The broken lines represent further synsets.
{legal_document}
{agreement}
{concrete_written_agreement}
{abstract_written_agreement}

figure 7: Solution to the diamond problem (1)

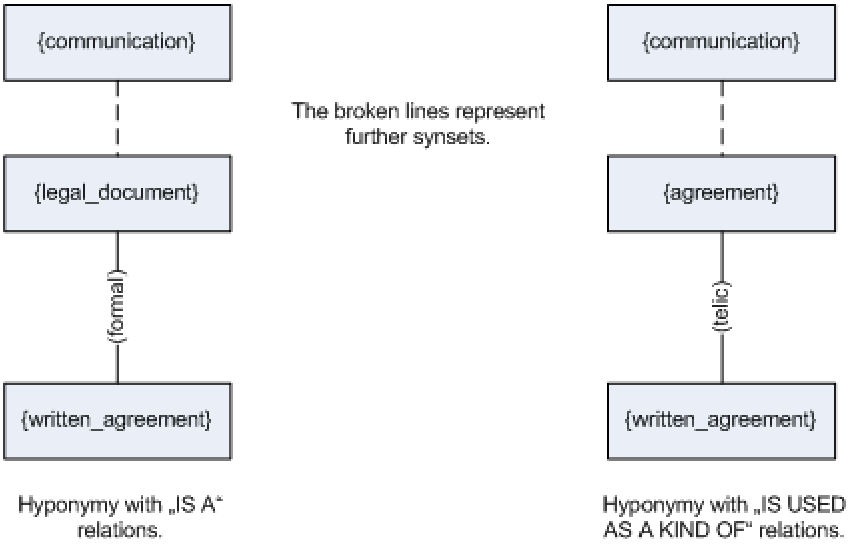

{communication}
{legal_document}
(formal)
{written_agreement}
Hyponymy with „IS A" relations.
The broken lines represent further synsets.
{communication}
{agreement}
(telic)
{written_agreement}
Hyponymy with „IS USED AS A KIND OF" relations.

figure 8: Solution to the diamond problem (2)

Hyponymy is an inheritance system. The features that are inherited, however, are not explicitly defined in WordNet. They are only provided "by the explanatory glosses that accompany (nearly all) synsets." On first sight, this is disadvantageous, but this practice avoids another problem that arises from linear reasoning in non monotonous systems. If every synset would also have a defined set of features, the concept *{penguin}* would probably inherit the feature *"can_fly"* from the concept *{bird}*. In WordNet, however, the explanatory gloss describes *penguins* as "short-legged flightless birds of cold southern especially Antarctic regions having webbed feet and wings modified as flippers", which prevents erroneous inheritance. (Miller 1999b, 34) Of course, the glosses must not contain any information that is incompatible with any subordinate concept; otherwise they would have to use hedges to make clear that the gloss may not be true for all hyponyms of the respective concept. This is a step towards *default reasoning*, which acts on the assumption that every feature fits any subordinate concept unless it is explicitly known that the feature does not apply. (Russell, Norvig 2007, 354–360)

2.3.2.2 Meronymy

Another hierarchical relation, which can only be applied to the word class of nouns, is *meronymy*. In contrast to hyponymy, meronymy does not dissect concepts on the feature level, but it "deals with the relation between an object and its constituents or proper parts" (Miller 1999b, 37). Like hyponymies, meronymous structures utilise two separate reflexive relations, *meronymy* and *holonymy*. The former is directly coded in WordNet, whereas the holonymy relation has to be inferred.

We can test whether an object W_m is a meronym of an object W_h in checking if the phrase W_m *is a part of* W_h is acceptable. The other way round, we can check for holonymy in validating the phrase W_h *has a* W_m. (Miller 1999b, 37) The pair IS A PART OF/HAS A PART is, however, not the only type of meronymous relation, even though it is the most common. Cruse (1986, 157–180) has identified many more types of meronymous relations and corresponding sentence frames to account for any shade of variation a whole-part relation could have. WordNet uses three different types of meronyms, which are sufficient for the majority of the cases. Besides the already mentioned CONSTITUENT PART OF relation, there is the SUBSTANCE OF and the MEMBER OF relation.

Although the semantic network connects concepts with all of the available relations, reasoning in meronymies can only be done along arcs of the same meronymy type at a time. If this constraint did not exist, rather odd inferences could be made, like "*the bark is part of a forest"* as the transitive statement of the two propositions "*the bark is PART OF a tree*" and "*a tree can be MEMBER OF a forest*". (Miller 1999b, 38–39) Furthermore, not only the relations have to be of the same type in order to make reasonable inferences, but also the synsets have to be "of the same general type" (Cruse 1986, 168), which means that they have to descend from the same unique beginner (cf. chapter 3.3.2.1).

Multiple inheritance was a rare exception in hyponymous structures. However, multiple holonymy pointers occur very often in meronymous taxonomies. (Miller 1999b, 38) The reason for this can be found in cognitive psychology. People make abstractions of the perceived world. Although the immediate memory still stores the full perception in all its detail, the richness of detail decreases during further cognitive processing in the brain. People therefore tend to forget unnecessary details; they strip all the information from their mental image that is not needed to understand its basic significance. (Anderson et al. 2001, 142–147) They remember complex objects as combinations of generic concepts which basically always have the same function but still differ from case to case. Miller (1999b, 38) gives the meronyms *point* and *handle* as examples. They are parts of a wide variety of holonyms, but there is no doubt that the handle of a coffee mug is very different from the one of a brush. However, they serve the same purpose.

Because of the fact that many nodes have multiple holonyms, we cannot speak of meronymies as tree structures any more. A meronymous hierarchy is a *general directed graph*, in which related nodes are connected both by a meronymy relation in one direction and a holonymy relation in the other direction. It does not have a single explicit root node, because there is nothing like the unique beginners of hyponymies. One could, however, define synsets that either have meronyms or holonyms (and not both) as *end points* in the hierarchy.

“Since meronyms are distinguishing features that hyponyms can inherit, meronymy and hyponymy become intertwined in complex ways. For example, if {beak} and {wing} are meronyms of {bird}, and if {robin} is a hyponym of {bird}, then, by inheritance, {beak} and {wing} must also be meronyms of {robin}.“ (Miller 1999b, 38) One must, however, keep in mind that whole-part relations can pose restrictions on hyponymy structures. If the concept {fur} would be a meronym of the concept {cat} this would automatically deny the existence of a hairless cat as a hyponym of {cat}. Therefore, it is important to establish meronymy relations at the appropriate hyponymy level. (cf. Miller 1999b, 38)

2.3.2.3 Entailment, troponymy and causation

In this section I will discuss three semantic relations that are closely related and which exclusively apply to the class of verbs.

For some, it might be a bit unusual to speak of *entailment* as a hierarchical relation, because this relation of implication is very much connected with logics. However, it is used differently in semantics, logic and pragmatics. (Bublitz 2001, 136) In WordNet, the semantic interpretation applies. Entailment is defined as “the relation between two verbs V_1 and V_2 that holds when the sentence *Someone* V_1 logically entails the sentence *Someone* V_2.”(Fellbaum 1999a, 77) This means, if the first sentence is true, the second sentence must be true as well. For example, “*John is snoring*” entails “*John is sleeping*”. The reverse, however, does not necessarily have to be true. Therefore, we call entailment a *unilateral relation*. (Fellbaum 1999a, 77)

Entailing verb pairs can be regarded as part-whole statements; s*noring* can be seen as a part of *sleeping*. This lets us draw the analogy to meronymy. In fact, entailment can be said to be meronymy among verbs. Besides the part-whole relation, entailing verbs have something else in common. Because verbs refer to actions, states or events, they all have a temporal feature. Two verbs in an entailing relation can either describe distinct activities that occur simultaneously, like *drive* and *ride*, or they can refer to activities where one of which is included in the other, like *snore* and *sleep*. This means that entailing verbs are temporally inclusive. (Fellbaum 1999a, 78)

The following diagram contrasts verbal entailment with nominal meronymy.

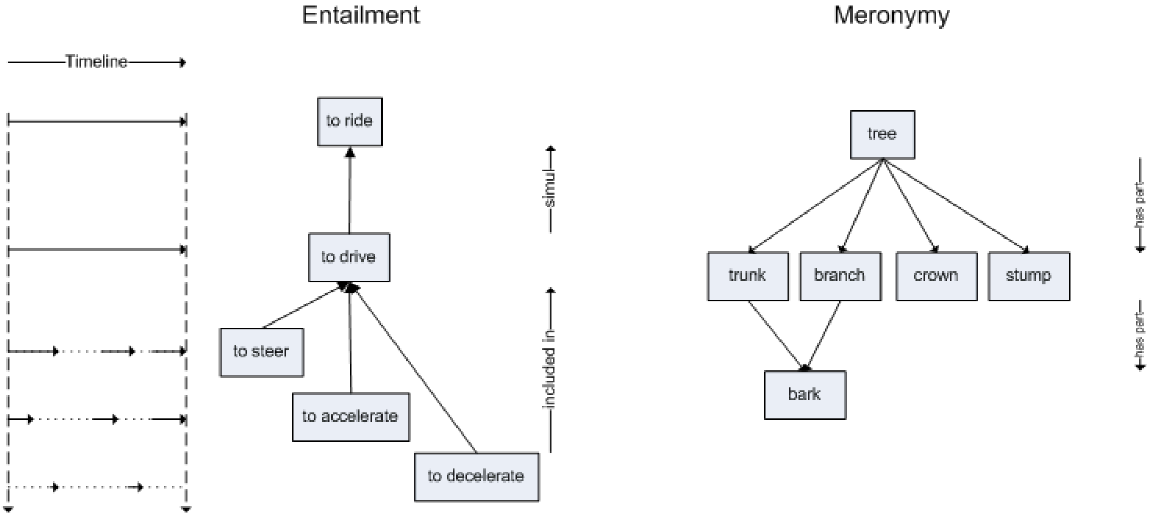


figure 9: Entailment vs. Meronymy

In addition to this meronymy-like hierarchy, the verbs can also be structured in a way similar to hyponymy. To achieve this, Fellbaum and Miller have coined the expression troponymy. It is a special kind of entailment, which has the additional requirement of temporal identity. Whereas *to steer* requires only part of the time of *to drive*, the action of *marching* is always temporally coextensive with the action of *walking*. Marching is therefore not really a part of the action of *to walk*, but it is a particular way of walking and thus called a *troponym* of *to walk*. Accordingly, the arcs of troponymy relations cannot be labelled with *IS A* or *IS USED AS A KIND OF,* like in hyponymous structures, but they have rather to be paraphrased with *IS A PARTICULAR WAY TO* or *IS A MANNER OF*. (Fellbaum 1999a, 79–80) Due to the precondition of temporal identity, troponymy is not a unilateral relation like general entailment, but it has a reverse relation also called *hypernymy*, which relates the individual troponyms to their superordinate hypernyms.

A third semantic relation, which is also derived from entailment, is the *cause-relation*. It has the least hierarchical character of the three, but is still worth mentioning in this section, because it is often used in verbal taxonomies together with entailment relations. The cause-relation connects a *causative verb* with another verb that describes the action or state resulting from the first. An example for a pair of causative and resultative verbs are the words *show* and *see*. The action of showing an object to someone results in the action of someone seeing this object.(Fellbaum 1999a, 83) This alone is not yet a hierarchy, but the cause-result pairs can be very well combined with entailment relations.

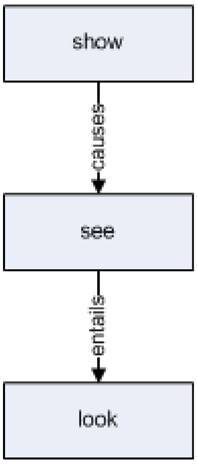


figure 10: Causation and entailment

“In addition, WordNet contains CAUSE pointers from causative, transitive verbs to the corresponding anticausative (inchoative), intransitive sense of the same word[.]”(Fellbaum 1999a, 83) For example, the lexical pair *break-break* stands for a cause relation between the synsets *{break, bust}* and *{break, wear, wear out, bust, fall apart}*.

2.3.3 Oppositeness and similarity of meaning

This section will discuss *antonymy*, a lexical relation which is present in all word classes, but which is most important for the organization of adjectives in WordNet. In this context, the *similarity* relation will also be mentioned, which is another integral part of adjective clusters and is needed to deduce indirect antonym pairs.

Linguists usually distinguish several types of opposites, *antonymy* being only one among them. WordNet, however, uses the term in a wider sense and links all word pairs which express directly opposing concepts with the same relational pointer *ANTONYM*.

The largest group within the class of modifiers in WordNet are the *descriptive adjectives*. In contrast to the hierarchically structured nouns, they are organized in adjective clusters for which the antonymy relation is of vital importance. These clusters are usually defined by a central pair of direct antonyms, like *fast/slow* or *light/heavy.* Not for all adjectives, however, can direct antonyms be found. Therefore, they are connected to adjectives which are similar in meaning and which do have a direct antonym by the *SIMILAR_TO* pointer. This creates an adjective cluster consisting of two half clusters which are linked by the antonymy relation of the defining, central antonymy pair. Thus, every adjective from one half-cluster can be related to the central adjective of the other half-cluster by indirect antonymy in inheriting antonymy through similarity. In *figure 11*, for example, *quick* is an indirect antonym of *slow*, because it is similar to its direct antonym *fast*. (Miller 1999c, 48–52)

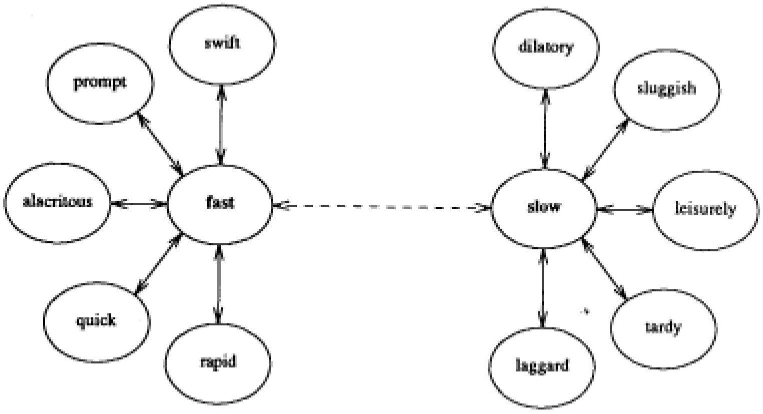


figure 11: Adjective cluster *fast/slow (Miller 1999c, 51)*

Although an adjective has at most one direct antonym, it can have several indirect ones, because it can be member of more than one cluster.

Besides descriptive adjectives, WordNet distinguished between *participial adjectives*, a subcategory of the descriptive ones, *relational adjectives* and *adverbs*. These three classes of modifiers are by far less likely to have antonyms, both direct and indirect ones. Therefore, to make the organization simpler, WordNet treats participial and relational adjectives that have antonyms as if they were descriptive adjectives. Adverbs which are derived from verbs inherit their properties from their respective stem adjective to which they are connected by the *DERIVED_FROM* pointer. This way, antonyms of adverbs can be deduced by combining a 'derived from'-search with an 'antonymy'-search.

Special positions are taken by *colour terms* and *quantifiers*. The only pair of direct antonyms among colour terms is *black* and *white*. All the other terms form a single cluster with the central antonymy pair *chromatic/achromatic* (cf. figure 12). Quantifiers, on the other hand, are frequently included in the class of determiners. However, because they share many characteristics with adjectives, inter alia, antonymy, they are treated as descriptive adjectives in WordNet. (Miller 1999c, 56–61)

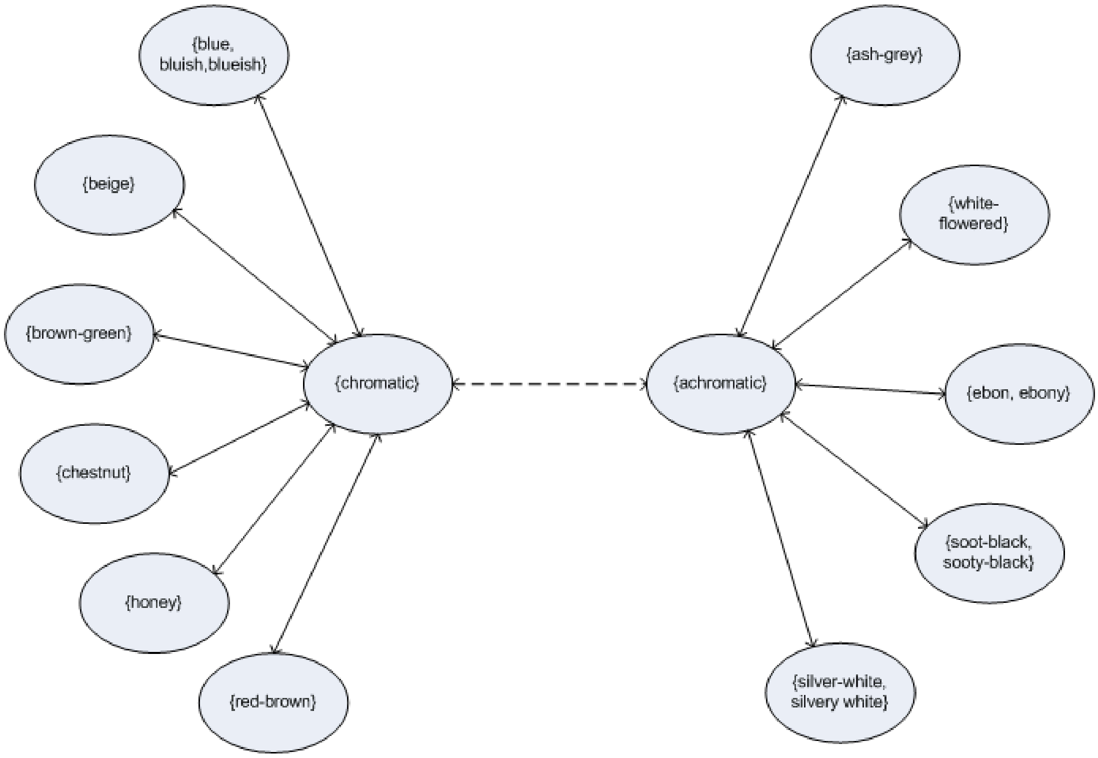


figure 12: A selection of colour terms

In contrast to the modifiers, the organisation of nouns is not much influenced by antonymy. As long as concrete nouns are concerned, there is very little semantic opposition, because physical referents can hardly be compared in terms of opposition. Abstract nouns, however, can frequently be linked by the antonymy relation. It is not surprising that most occurrences of semantic opposition can be found among the deadjectival nouns. Antonymy does not automatically interact with the hierarchical organization of the nouns, because it is a lexical relation between two specific word forms. It is not automatically inherited by subordinate concepts in the noun hierarchy and has to be entered for any two nouns for which the relation should exist. Interestingly, antonymous nouns nearly always have the same hypernym and often even are co-hyponyms. (Miller 1999b, 39–40)

In the class of verbs, semantic opposition mostly concentrates on stative and change-of-state verbs. Like adjectival antonymy, verbal opposition is usually further subcategorized by linguists, because not every semantic opposition has the same properties (cf. Cruse 1986). In WordNet, however, only the antonymy pointer is used to link antonymous verbs. As we have seen in 3.3.2.3, troponymy is similar to hyponymy. Therefore, we can analogously find that many antonymous verbs are co-troponyms.

Furthermore, verbal antonymy interacts with entailment in a way that temporal inclusion is substituted by a kind of backward presupposition. However, backward presupposition is not properly coded in WordNet (cf. section 3.4), which is why the occasional interdependence between entailment and antonymy is not further discussed here. (Fellbaum 1999a, 81–83)

2.3.4 Influence of polysemy on semantic relations

Although the English language consists of a vast number of words, many of them have more than one meaning. Consequently, a WordNet query will usually return more than one synset in which the search term occurs. This phenomenon of *polysemy* has some notable effects on the semantic relations we have discussed so far.

We can identify three different effects of polysemy on the class of nouns, one of which we have already mentioned in section 3.3.2.1 – *cousin nodes* are hyponyms of two different hypernyms which are similar in meaning. “For example, if two senses of the noun *fish* are related as an animal and as a food, then all the matching hyponyms for these two senses of *fish* (which includes *perch*, *sole*, *bass*, etc.) will also bear that relation.” (Miller 1999b, 42) *Sister nodes*, on the other hand, are co-hyponyms, i.e. direct hyponyms of the same hypernym, which have identical strings.

They can therefore be assumed to be similar in meaning. Finally, *twins* are pairs of synsets which have three or more words in common. For example, two of the four senses of duo consist of the same set of words *{duet, duette, duo}.* Although both synsets refer two different concepts, they qualify to be twin nodes and are therefore similar in meaning. (Miller 1999b, 43)

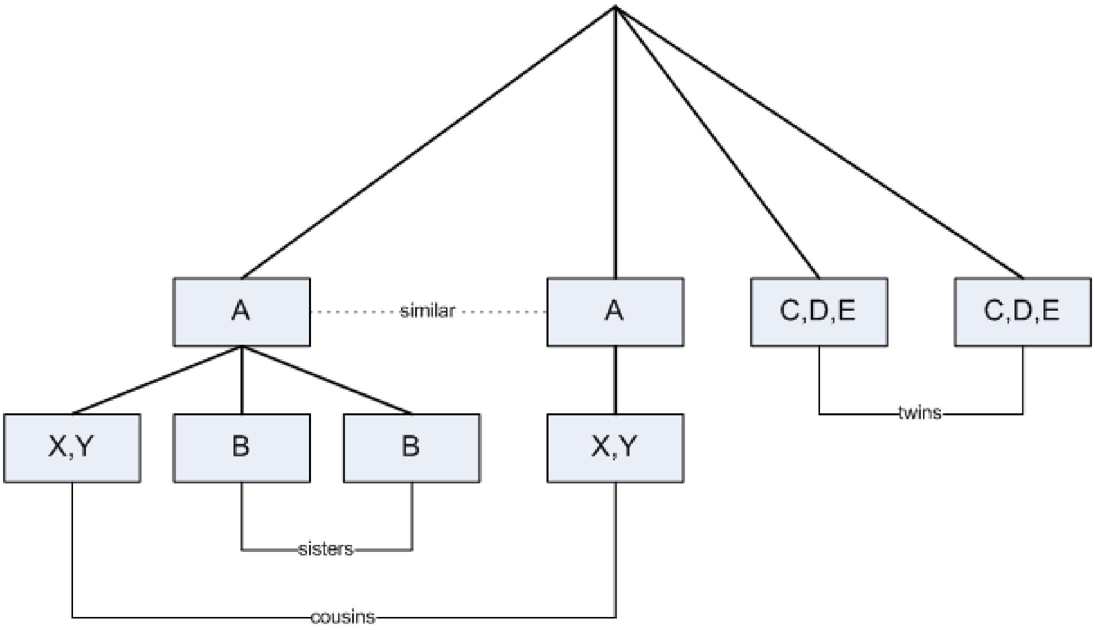


figure 13: Influence of polysemy on hyponymy among nouns

The effect of polysemy on adjectives is not as concrete as its effect on the nouns. Here, polysemy seems to create a kind of selectional preference of the nouns these adjectives occur with. "For example, the sense of *old* meaning 'not young' frequently modifies hyponyms of *person* [...] whereas *old* meaning 'not new' frequently modifies hyponyms of *artefact*."(Miller 1999c, 54)

The biggest influence of polysemy on semantic relations can be observed among the class of verbs. According to Fellbaum (1999a, 84), this is because the English language has comparatively few verbs – "[they] are approximately twice as polysemous as nouns."(Fellbaum 1999a, 84) As a consequence of this strong polysemy, new derivates of semantic relations emerge. Usually, semantic relations exist only between distinct word forms. For highly polysemous verbs, however, we can also see that individual senses are linked by similar relations. Fellbaum (1999a, 85ff.) names the resulting phenomena *autorelations*. So far, three autorelations have been identified, *autohyponymy*, *autoantonymy* and *entailment*.

Autohyponymy exists when two senses of a verb are related by troponymy, “that is to say, a semantically more elaborate verb is sometimes expressed by the same surface form as its more general superordinate.” (Fellbaum 1999a, 85) For example, depending on the context, *behave* can either mean ‘*conduct oneself*’ or ‘*conduct oneself well*’. (Fellbaum 1999a, 85–86)

Autoantonymy, on the other hand, means that a word can have two senses that are antonymous. For example, *to string* can either mean ‘*to provide with strings*’ (“To string an instrument”) or ‘*to remove the string from something*’ (“To string the beans”). (Fellbaum 1999a, 88–89)

Finally, *entailment* can also hold between two senses of a polysemous verb. This is mostly the case when the context is not totally clear. In the sentence

“*The president drove on to the Capitol.*” (Fellbaum 1999a, 88),

one cannot clearly decide if *to drive* is meant in the sense of ‘*to operate and steer a vehicle*’ or in the sense of ‘*to have oneself carried in a vehicle*’. “In the former case, his driving (operating the car) necessarily entails his driving (being a passenger)” (Fellbaum 1999a, 88)

2.3.5 Other relations

So far, the major semantic relations in WordNet have been described. There are, however, a few other relations that are not vital to the organization of the words in WordNet, but which are still worth mentioning.

Similar to the adverbs, which are usually derived from adjectives and therefore connected with them by the DERIVED_FROM relation, participal adjectives are most of the time derived from verbs. Because these adjectives carry most of the meaning of the verb, they are linked to the respective verbs with the PRINCIPAL_PART_OF pointer. For example, *breaking* is the PRINCIPAL_PART_OF *to break*. (Miller 1999c, 58)

Descriptive adjectives, on the other hand, have a similar relation with nouns. They describe the value for the attribute that is defined by the noun. Therefore, descriptive adjectives are connected to their related noun by the reflexive relation

ATTRIBUTE. The adjective *heavy*, for example, is “a value of” the noun *weight*. (Miller 1999c, 48–49)

The HAS_INSTANCE relation can only apply to subordinates of the synset {event}. It is a special kind of hyponym relation that links concrete instances of an event to the appropriate concept node. For example, the synset {civil war} has the instance “American Civil War”.

The last relation that will be mentioned in this section is not really a relation, let alone a semantic relation, but it shall nevertheless be listed here for the sake of completeness. *Morphy* is a sub-module of WordNet that offers a set of morphology functions. “Although only base forms of words are usually stored in WordNet, searches may be done on inflected forms. [Each search string is therefore analyzed by morphy] to generate a form that is present in WordNet.” (Cognitive Science Lab, Princeton University 14.12.2006)

2.4 What does WordNet not contain?

WordNet has become very important for many researchers of the English language and it has been adopted for many practical applications. Nevertheless, there are some aspects that are – deliberately or not – not contained in WordNet.

2.4.1 Syntax and context

A limitation that has already been mentioned earlier is the focus on paradigmatic relations. “WordNet contains no information about the syntagmatic properties of the words [,]” because of the separate treatment of the four word classes (Fellbaum 1999b, 5). Consequently, beyond the explanatory glosses, WordNet also bears no contextual information about the words. Researchers have created smaller semantic concordances (cf. Fellbaum 1999b, 13) and semantically tagged corpora, like the *SemCor* corpus, to provide some contextual information for the words in WordNet. However, no mapping of a large corpus, like the *British National Corpus*, to WordNet-senses has been implemented so far. Besides contextual and syntactical information, such a mapping would also provide frequency information, which has been omitted in WordNet due to copyright reasons in the beginning of the project. (Miller 1999a, xviii)

2.4.2 Semantic relations

For reasons that have been explained in section 3.3, WordNet only contains a small selection of semantic relations. While this decision is certainly understandable, it would nevertheless be advantegous for many applications to have other relations at hand. Although "WordNet is not likely to incorporate new relations in the near future [...] [it still can be] customized for a particular purpose." (Fellbaum 1999b, 12) For example, semantic relations acquired by corpus linguistic methods, as it will be demonstrated in chapter 3, can be added to the existing relations in WordNet's semantic net.

Generally, WordNet's sense distinctions can be said to be very fine grained, which turns sense disambiguation into a hard task, especially due to the lack of contextual information. Some of the semantic relations, however, deserve further differentiation. The hyponymy pointer, as we have seen in section 3.3.2.1, actually represents more than one relation, which sometimes causes inconsistencies in the reasoning process. The antonymy pointer is even more universally used, because it stands for any kind of semantic opposition without any further distinction, even though antonymy is usually highly differentiated by linguists (cf. Cruse 1986).

2.4.3 Topical information

"Because [WordNet] focuses on the semantics of words and concepts rather than on semantics at the text or discourse level, [it] contains no relations that indicate the words' shared membership in a topic of discourse."(Fellbaum 1999b, 10) The purely hierarchical organization of nouns, instead of the use of co-occurence relations, entails that there is not necessarily a connection between topically related concepts, like, for example, *racquet, ball, net* and *court game*. The words are rather spread over the various lexicographer files[10] like *noun.person*, *noun.artifact* or *noun.location*. (Miller 1999b, 34)

2.4.4 Other

In section 3.3.2.3, we have discussed the entailment relation and two other relations derived from it. Fellbaum (1999a, 83–84) also mentions a fourth entailment relation between verbs, namely *backward presupposition*.

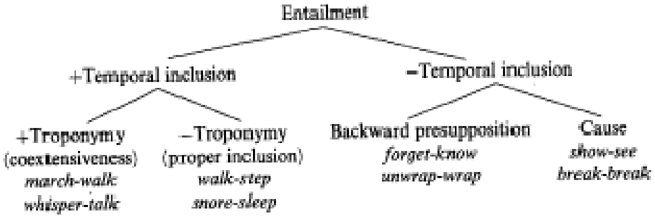


figure 14: Different entailment relations in WordNet (Fellbaum 1999a, 84)

Backward presupposition can be regarded as a kind of constraint in verbal taxonomies. It specifies which action already has to be carried out before another action can begin. For example, *to forget* presupposes *to know*, because "one can only forget something that one has known."(Fellbaum 1999a, 82) Presupposition is important for many applications in natural language processing (NLP), because reliable knowledge extraction from natural texts can only work with certain pragmatic information, one of which being the *presupposition constraint*. This, however, is not implemented in WordNet.

Another feature which is quite interesting for NLP but which is not properly included in WordNet is *semantic distance*. Especially word sense disambiguation would benefit from the information about *semantic relatedness* and *semantic similarity*. In the first instance, one might think that the hierarchical structure of WordNet already bears this information, but as Scriver (11.05.2006) has found out in his thesis, semantic distance is not equatable with the distance of two concepts in the semantic network. WordNet, at best, can tell us that two concepts are similar or related, but it gives no hint to what degree they are related. There are, however, several approaches that promise to deliver this information and which might be

[10] The *lexicographer files* are the sources that have to be manually created by linguists. They are then further processed by the so-called *grinder*, which creates the machine readable database used by the WordNet application.

worth including in future versions of WordNet. This would then provide *relatedness* and *similarity values* for any two (noun-) concepts in the network.

The last remark about the shortcomings of WordNet relates to the adjectives. When talking about the semantics of adjectives, *markedness* and *gradation* usually play important roles. According to Katherine Miller (1999c), gradation even has to be regarded as a separate semantic relation. However, markedness and gradation are not easy to represent in semantic networks. Therefore, and because they are not vital to the organization of the adjectives, they have not been coded in WordNet. (Miller 1999c, 52–54) To obtain these kinds of information, other sources besides WordNet have to be incorporated.

The features and relations incorporated in WordNet are carefully selected, because WordNet's character of a general, domain-independent lexical database ought to be sustained. However, in the recent past it turned out that applications in natural language processing and information retrieval can profit a lot from additional semantic information. (Hearst 1999, 131) These needs can be satisfied by adding this information to WordNet as it is needed. Through various different programming interfaces, WordNet can be accessed by third party tools which can extend the semantic net or create new ways of interaction and visualization. In chapter 5, an approach will be presented to add additional psycholinguistic features to WordNet that are missing in the original version.

3 The British National Corpus

> The *British National Corpus* is a collection of over 4000 samples of modern British English, both spoken and written, stored in electronic form and selected so as to reflect the widest possible variety of users and uses of the language. Totalling over 100 million words, the corpus is currently being used by lexicographers to create dictionaries, by computer scientists to make machines 'understand' and produce natural language, by linguists to describe the English language, and by language teachers and students to teach and lean it – to name but a few of its applications. (Aston, Burnard 1998, n.pag.)

In the following chapter, I will give a short overview of the school of corpus linguistics. I will briefly glance over different types of corpora and their potential for linguists and other researchers of language. The main focus will then be on the British National Corpus (BNC), how the corpus approach differs from the one we have become acquainted with in the previous chapter, and how corpora can be used to acquire and examine semantic relations.

3.1 What is corpus linguistics?

3.1.1 Traditional linguistics, Chomsky and the corpus

In order to really understand the principles of corpus linguistics and its separation from other linguistic schools, we have to face a rather philosophical question first – the nature of language. Lakoff (2005, 157) put it accurately in saying that philosophy

> [...] matters more than most people realize, because philosophical ideas that have developed over the centuries enter our culture in the form of a world view and affect us in thousands of ways. Philosophy matters in the academic world, because the conceptual frameworks upon which entire academic disciplines rest usually have roots in philosophy – roots so deep and invisible that they are usually not even notices.

Although, linguistics can be subdivided in many ways and is influenced by various other sciences, we want to distinguish between three main schools – the traditional view of structural linguistics, the Chomskyan approach of generative, universal grammar and finally the empirical, corpus linguistic view on language.

The traditional view is the one that most foreign learners of the English language are confronted with during the learning process, because it is the predominant way of looking at language in schools and other facilities that teach language. It is the approach of grammar books and bilingual vocabulary lists. Structural linguists are often called “armchair linguists”, because they try to organize and explain language by thinking about it. They compile books with grammar rules and vocabulary without considering the mental representation of language or the actual usage of language in discourse. (Teubert, Cermáková 2008) Within the traditional school of language, we can identify a certain circularity. Grammarians derive grammar rules from natural language by thinking about how language can be structured. At the same time, these rules are the basis for learners of the language which consequently define how language is structured. The awareness of this “chicken or the egg” dilemma indicates that the traditional view on language cannot be the last word on this subject, albeit its grammar rules are undisputedly helpful for foreign learners of a language.

For Noam Chomsky and followers of his theories, rules are something different than for the traditionalists. They are not descriptions of existing language, but rather the innate algorithms of native speakers that allow them to generate an endless amount of grammatically correct language. “Rather than being tools for language analysis, they [are] the metaphysically real essence of language.” (Teubert, Cermáková 2008, 9) Linguists following this school of thought, usually deny the famous Sapir-Whorf hypothesis that our thinking is determined and restricted by our language. Instead, every thought of any human being is not carried out in their mother tongue, but in a special language of thought, which is often called *mentalese*. The universal grammar is therefore not a grammar of any spoken language, but a grammar of our mental language. The innate generative rules, which are specific to our native language, in the end, transform our thought from *mentalese* to the natural language which we are then able to speak and write. (Pinker 2007 , Teubert, Cermáková 2008, 32) “Knowing a language, then, is knowing how to translate mentalese into strings of words and vice versa.” (Pinker 2007, 73)

Corpus linguists neither see language simply as the “miraculous [...] faculty we all are born with”, nor merely as “an acquired skill enabling us to take an active part in verbal communication" (Teubert, Cermáková 2008, 35). *Corpus linguistics* is an empirical discipline that "sees language as a social phenomenon" (Teubert, Cermáková 2008, 37) for the comprehension of which we must consider both the *meaning* of the language and our *understanding* of this meaning. Meaning is nothing that is inherent to the word itself, but it is the sum of all its *usages* and *paraphrases* within the discourse community. (Teubert, Cermáková 2008, 78) No single person can, of course, be acquainted with every usage and paraphrase of a word. The idea a person has of the meaning of a word is therefore only part of the total picture and consists of every instance the person has ever heard or seen someone use this word. This also means that no two persons have the exact same understanding of the meaning a word, because their experiences with this word are most likely not identical. (Teubert, Cermáková 2008, 80) U*nderstanding* is only a subset of *meaning*, the latter of which containing all possible ways one can understand a word.

That is the point where the corpus comes into play. Although, a corpus can also not be a perfect reflection of word meaning, it certainly offers a more comprehensive picture to the individual. "Corpus linguistics puts us into a position where we can inform ourselves what use others have made of language."(Teubert, Cermáková 2008, 134)

3.1.2 What is a corpus?

"A corpus is a collection of naturally-occurring language text, chosen to characterize a state or variety of a language. In modern computational linguistics, a corpus typically contains many millions of words [...]". (Sinclair 1995, 171) Ideally, a comprehensive corpus should be a representative collection of samples of the whole discourse. However, this is easier said than done, because we do not simply have the whole discourse at hand and we are not in the position just to select representative samples from it. Therefore, a corpus can never be representative in the narrow, statistical sense. It should, nevertheless, be as diverse as possible and contain examples from all parts of the discourse, spoken and written, formal and informal. Only this way, it can help to understand most of the facets of meaning words can have and expand the very restricted personal understanding of the lexicon. One must, however, always keep in mind, that the corpus can never be more than a scattered mirror, reflecting the whole discourse only in parts.

3.1.3 Corpus types

Corpora have many different possible applications. Therefore, it is not surprising that there exist various corpus types ranging from large, generic *opportunistic corpora* to rather small scale, domain specific *special corpora*. (Teubert, Cermáková 2008, 65–77) The classification of corpus types is not uniform among textbooks. Sinclair (1995) basically distinguishes between two main types, the *sample corpus* and the *monitor corpus*, whereas the short overview of Aston and Burnard (1998) goes more into the different varieties of special corpora, like the *genre specific corpus* and *spoken language corpora*. For a good first impression of the different faces corpora can have, I want to elaborate further on the classification by Teubert and Čermáková (2008), who categorized corpora, according to their content, into five different types, *reference corpora, opportunistic corpora, monitor corpora, parallel corpora and special corpora*.

"*Reference corpora* contain the standard vocabulary of a language. They are the corpus linguist's main resource to learn about meaning. If they are large enough, they reveal the contexts into which words are usually embedded[.]"(Teubert, Cermáková 2008, 67) Its texts are selected to be as representative for the whole of the language as possible, within the already mentioned constraints. A reference corpus represents "what the discourse community agrees to be what a fairly educated member of the middle class would read outside of work, mostly in printed form, but also handwritten or typed: and in principle at least, it should also contain a sample of what they would hear [...]"(Teubert, Cermáková 2008, 67).

The *opportunistic corpus*, on the other hand, "is based on the assumption that each and every corpus is imbalanced" (Teubert, Cermáková 2008, 70). Consequently, there is no need to carefully select the texts to create such a balance. It therefore contains as many texts as possible and usually consists of a larger number of smaller special corpora. (Teubert, Cermáková 2008, 70–71)

Monitor corpora are used to study language change. In contrast to the other corpus types, they have to be updated frequently. A monitor corpus keeps track of the date of origin of the texts that are added and thus allows comparing the usage of particular words in the course of time. (Teubert, Cermáková 2008, 71–73)

The penultimate corpus type in the list, the *parallel corpus*, consists of “original texts in one language and their translation into another (or several other languages)” (Teubert, Cermáková 2008, 73). Parallel corpora are often used by translators as a tool for sense disambiguation, because common dictionaries seldom bear enough information that is needed for adequate translations. (Teubert, Cermáková 2008, 73–76)

Finally, *special corpora* are mostly small scale corpora which have been created for special purposes. They are used to monitor specific phenomena and are often deduced from larger corpora, like reference corpora or opportunistic corpora. Due to their concentration on certain aspects and their small size, the application spectrum of special corpora is usually very limited. However, because of the evolution of specialist software and the increasing availability and decreasing prices of powerful computer systems, special corpora can today be easily created on demand, which makes them very useful for projects where statistical representativeness is not crucial.

An additional sixth category, the *internet corpus*, sticks out from this categorization, because it is basically the source that differentiates this type from the others. Today, a lot of the everyday communication and journalistic publication is made online. In addition to that, many of the traditional print media are also accessible via the internet. That is why the World Wide Web is such an important source for language researchers. Internet corpora and, to a certain extent, internet search engines[11] make these resources accessible for linguists.

3.2 The British National Corpus

As announced, the remainder of chapter 0 will deal with the British National Corpus (BNC). According to the reference guide of the latest (XML-)version of the BNC, the British National Corpus is a sample corpus, or to stay within our categorization of corpus types, a reference corpus, which is “composed of [over 2400] text

[11] Here, an *internet corpus* is understood as a static corpus of texts from the internet that has been compiled (and maybe further processed) for offline or online usage. *Search engines*, on the other hand, perform live searches directly on the data in the web.

samples generally no longer than 45,000 words"(Burnard 2007, ch. 1.2). It is a synchronic, general corpus of British English containing both written and spoken language. Due to the great effort that is needed to transcribe spoken language, the latter has only a share of 10 percent of the total 100 million words in the BNC. Further information about the composition of the BNC, like a detailed breakdown of the used texts into text types, domains, publication dates and many more properties, can be found in the latest reference guide. (Burnard 2007, ch. 1)

In principle, a corpus is just a collection of plain texts. However, linguists, more often than not, need information beyond the mere words. Language is more than just a long sequence of letters and punctuation marks; it bears a vast amount of information that is not directly contained in the words. "There is a lot of information that can't really be disseminated in that linear way [...]". (Erin McKean). Thus, a corpus that only contains plain texts would lack those important pieces of information like, for example, paralinguistic phenomena for spoken texts or layout information for the written texts. In addition to that, it is also useful to have grammatical information directly at hand, because they can be used for more sophisticated queries in the corpus. The BNC therefore contains so-called metadata which further comments and describes the information contained in the plain texts. It is attached to the text using the *eXtensible Markup Language* (XML)[12], which basically allows corpus designers to add further attributes and classification data to any lexical unit of any text in the corpus. The corpus user can later use the XML-tags in their queries. This allows them to perform very broad searches, for examples queries for specific parts of speech, or they can use the tags for more detailed queries, for example in restricting a search for "will" to all its occurrences as a full verb. Lexicographer Erin McKean has made some interesting remarks about the amount and the importance of those tags[13]. She said that the visible text of the corpus is like the tip of an iceberg making up only a small percentage of the whole. This becomes obvious when looking at the source files of the BNC (cf. appendix 3), where the actual original text makes up only a small part of the complete data. The BNC reference guide (2007) contains a full specification of the BNC XML schema, which is both

[12] XML is used for the current version of the BNC. Older versions used the XML predecessor SGML, which is also a markup language.

[13] Although she was talking about corpus based dictionaries, this is also true for corpora themselves.

interesting for developers of software that is supposed to access the BNC and for users of existing BNC query software, like the XAIRA-Client[14].

3.3 The differences between the WordNet-approach and the corpus-approach to semantic relations

WordNet and the BNC follow very different approaches to the English language. Therefore we cannot make a general, comprehensive comparison, but we want to look at the differences that are especially important for the examination of semantic relations. This will show us the advantages we could gain from enriching our knowledge-based semantic relations from WordNet with relations obtained by an approach based on text evidence.

The first, most important and also most obvious difference between WordNet and the BNC is that semantic relations are directly available in WordNet, whereas they are not explicitly coded in the BNC. In section 4.4 we will see how semantic relations can be acquired from the data available in the corpus, but first, we want to consider the implications of the facts that WordNet is a network of predefined semantic relations and that the BNC has no semantic information at all. On the one hand, WordNet makes it very easy for the researcher to access, evaluate and use the semantic information, because it can be easily obtained by a simple WordNet query. The BNC, on the other hand, does not divulge this information without further linguistic work. At first sight, this might seem like a pure disadvantage on the part of corpus linguistics, but when we reconsider this fact, it can also be seen as a great chance. As it has already been mentioned in chapter 3, WordNet contains a rather small number of selected semantic relations – the ones which have been decided by the creators of WordNet to be the most useful and the most universal relations. At this point, we can draw the analogy to the theory of education (*Bildungstheorie*) of the German philosopher Theodor W. Adorno. He was moaning the decay of classical education for the benefit of modern *half knowledge* (*Halbbildung*). In his eyes, half knowledge was even worse than no knowledge at all

[14] XAIRA stands for *XML Aware Indexing and Retrieval Architecture.* It is a piece of software designed by the *Oxford University Computing Services* that generally allows accessing any corpus that is coded in XML. It is the standard software used for querying the XML-BNC. Oxford University Computing Services et al. 16.05.2007

(*Unbildung*), because the uneducated theoretically still had the chance to get a proper classical education, whereas the half-educated could never return to that state of mind that makes real education possible. (Dörpinghaus et al. 2006, 104–115) Semantic relations in WordNet can be seen analogous to Adorno's half knowledge. It is implied that certain relations do not exist, because they are not contained in the semantic network, whereas the necessity of constant discovery of the semantic relations in the BNC could lead to new insights on language.

A second, important difference between the two approaches can be seen in the difference between the basic building blocks of WordNet and of corpora. As we have already seen, synsets form the basic building blocks in WordNet, which again mostly consist of single words. Consequently, semantic relations can only exist between single words. Corpus linguistics, on the other hand, is not fixated on the word as the basic unit of meaning. The limitation on single word units of meaning implies that the so-called *'slot-and-filler'-model* is substantial enough to describe our usage of language. The open-choice principle assumes that we use grammar as a pattern in which we just have to fill in the single words from our lexicon in order to express whatever comes to our mind. (Sinclair 1995, 109–110) Sinclair (1995, 109 ff.), however, argues that we are seldom able to make those single choices, but that we rather have to choose between "a large number of semi-preconstructed phrases that constitute single choices". That is what Sinclair calls the *idiom principle*. In language corpora, like the BNC, the idiom principle is reflected by *collocations*, which are pairs or groups of words that are chosen together as a single unit of meaning and which do not necessarily have to be adjacent. (Sinclair 1995, 115) In accordance with the empirical nature of corpus linguistics, collocations can be obtained by statistical calculations, like e.g. the *log-likelihood* (Krenn 2000). Furthermore, the measure can be extended to include grammatical implications. "The collocation of a word with a particular grammatical class of words [is] termed *colligation*." (Aston, Burnard 1998, 14) Both collocation and colligation are useful tools for identifying semantic relations in the BNC. With their help, semantic relations can be identified which involve lexical units larger than single words.

As we have already seen in the discussion about the shortcomings of WordNet, the semantic network provides little or no topical, contextual and syntagmatic information. For some applications this might be unproblematic, but often the lack of context entails difficulties with sense disambiguation. This is especially the case for very frequent words. “There is a broad general tendency for frequent words [...] to have less of a clear and independent meaning than less frequent words or senses.”(Sinclair 1995, 113) Without contextual information, we face such problems that were described in section 3.4 of the previous chapter. “[W]ith the very frequent words, we are reduced to talking about uses rather than meanings. The tendency can be seen as a progressive delexicalization [...]”(Sinclair 1995, 113), which inevitably also affects the semantic relations between those words. Semantic relation acquired with the help of corpora could potentially be embedded in context to avoid those kinds of difficulties. We would not have to face the problem that we are not able to choose the appropriate sense of a word, because we could rely on an extensive context provided by the corpus.

Furthermore, the empirical nature of corpus linguistics would allow for semantic relations obtained in such a way to be enriched by frequency information, which could result in semantic networks with weighted edges. This could improve the reasoning in these networks, because the frequencies might be used as transition probabilities.

The last difference between the two approaches that will be mentioned here is regarding the lexical content of WordNet and corpora. As we have seen earlier, corpora like the BNC contain a balanced sample of the English language. This means, words are expected to occur as frequently in the corpus as they occur in the real discourse community. This poses a strong contrast to WordNet. In order to sustain the complex structure of the semantic network, WordNet has to contain a disproportionally large amount of special and infrequent vocabulary. Intermediary terms like “*chordate*” or “*proboscidean*” are needed to build correct hierarchies, but they are usually only part of the special vocabulary of professionals. Also the words at the bottom level of hyponymy structures are mostly rather infrequent and uncommon. This very fact also caused a problem in the project that is presented in chapter 5. Corpus-based information could help to deal with these difficulties.

3.4 Acquisition of semantic relations in the BNC

The previous section has shown that there are several aspects of semantic relations which WordNet does not account for. This alone is reason enough to seek for additional sources of semantic information. Moreover, the development of WordNet-like resources by hand very expensive, which causes a problem in meeting the constantly rising demands for domain specific semantic information (Schulte im Walde, Koller 06.08.2007, 6, cf. chapter 3.3). Corpora can fill these information gaps. However, because semantic relations are not exactly on show in large language corpora, we have to find ways to discover them within the pack of data. In order to acquire semantic relations from corpora, we have to perform a kind of reverse engineering of language.

Semantic relations form the base of our knowledge organization. When we want to use this knowledge in our written or spoken communication, we have to paraphrase the semantic relations by distinct linguistic structures. By re-discovering these structures from corpus-texts, we can draw conclusions about the original relations.

3.4.1 Discovery of semantic relations using patterns

A well established approach to the acquisition of semantic relations from unrestricted text is the utilization of surface patterns. The idea is to look for patterns which are (a)“highly specific”, (b)“easy to recognize” and which (c)“occur frequently [...] in many different text genres”. (Schulte im Walde, Koller 06.08.2007, 8) Methods following this approach basically take place in three steps:

(A) Discovery of relevant patterns
(B) Preparation of the patterns, pattern matching on the texts
(C) Filtering, evaluation and classification of the results

In step (A), the patterns have to be acquired which, supposedly, encode the semantic relations. To do so, a *top-down* or a *bottom-up* approach can be pursued. In the top down approach, the desired relation types are specified a priori. What is left to be done is therefore to discover the patterns that encode these relations. In the bottom-up approach, the relations are not defined from the outset. According to the principle of data-mining, all “interesting” patterns are therefore considered, i.e. all patterns that meet the basic requirements (a)-(c). In this case, however, the third step has to determine whether a pattern really encodes any semantic relation and which one it is. (Siemen 2006, 3–4)

An example for a top-down approach is the *Lexicosyntactic Pattern Extraction (LSPE) method* developed by Hearst (1992). According to this method, patterns for the relations used in WordNet can be acquired by the following algorithm:

> 1. Decide on a lexical relation that is of interest (e.g., meronymy).
> 2. Decide a list of word pairs from WordNet in which this relation is known to hold (e.g., house-porch).
> 3. Extract sentences from a large text corpus in which these terms both occur, and record the lexical and syntactic context.
> 4. Find the commonalities among these contexts and hypothesize that the common ones yield patterns that indicate the relation of interest.
>
> (Hearst 1999, 135)

Alternatively, the algorithm can be adjusted so that it does not rely on WordNet as a word source and therefore also permits the acquisition of patterns for other relations. (Hearst 1992, 541) Step 1, however, is the reason why Hearst’s approach, and many others that are based on his idea, is a top-down approach, because the semantic relations have to be decided on from the beginning.

By a manual test run of the procedure, it became obvious that hyponymy relations and relations that are related to hyponymy, like e.g. the group/member-relation, are the best candidates for this approach. For other relations, like meronymy, ambiguous results have been obtained. (Hearst 1992, 542)

For the hyponymy relation, this algorithm has lead to the following lexico-syntactic patterns

`such NP as {NP ,}* {(or|and)} NP`
(e.g. "... works by such authors as Herrick, Goldsmith, and Shakespeare.")

`NP {, NP}*{,} or other NP`
(e.g. "Bruises, ... , broken bones, or other injuries ...")

`NP {, NP}*{,} and other NP`
(e.g. "... temples, treasuries, and other important civic buildings.")

`NP {,} including {NP ,}* {or|and} NP`
(e.g. "All common-law countries, including Canada and England ...")

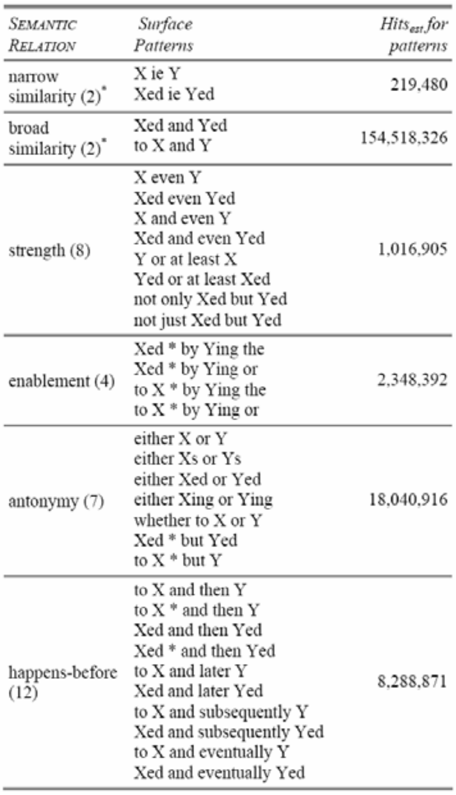
`NP {,} especially {NP ,}* {or|and} NP`
(e.g. "... most European countries, especially France, England, and Spain.")

(cf. Hearst 1992, 541)

Having discovered enough patterns, we can transcribe them into a form that is suitable for a corpus query. This will take place in step (B). Before we come to that, we will first have a look at an example for a bottom-up approach to semantic relation acquisition.

An example for an approach of that kind can be found in the *VerbOcean* project. In their paper, Chklovski and Pantel (2004) described a way to extract semantic verb relations from unrestricted text[15]. Their approach begins with the identification of highly associated verb pairs in the corpus. Afterwards, these pairs are analyzed in respect to their strength of association, which gives a hint whether the pairs really denote a semantic relation. The result is a list of verb pairs which are very likely to be related by a semantic relation. In order to get good results, this list is further filtered to remove less frequent items. From that point on, the procedure continues similar to step 3 of Hearst's pattern detection algorithm and results in a set of patterns that encode semantic relations. The actual type of the respective relation, however, has to be determined manually. A test run of the procedure has produced a total of 35 patterns encoding 6 distinct semantic relations, as can be seen in figure 15 (Chklovski, Pantel 2004, 4)

[15] The original approach was concerned with the extraction of semantic verb relations from the web. The procedure should, however, work with large text corpora as well. Therefore, I will continue speaking of *corpus* rather than *the web*.



Semantic Relation	Surface Patterns	$Hits_{est}$ for patterns
narrow similarity (2)*	X ie Y Xed ie Yed	219,480
broad similarity (2)*	Xed and Yed to X and Y	154,518,326
strength (8)	X even Y Xed even Yed X and even Y Xed and even Yed Y or at least X Yed or at least Xed not only Xed but Yed not just Xed but Yed	1,016,905
enablement (4)	Xed * by Ying the Xed * by Ying or to X * by Ying the to X * by Ying or	2,348,392
antonymy (7)	either X or Y either Xs or Ys either Xed or Yed either Xing or Ying whether to X or Y Xed * but Yed to X * but Y	18,040,916
happens-before (12)	to X and then Y to X * and then Y Xed and then Yed Xed * and then Yed to X and later Y Xed and later Yed to X and subsequently Y Xed and subsequently Yed to X and eventually Y Xed and eventually Yed	8,288,871

figure 15: Semantic relations and patterns identified by VerbOcean

In step (B), the acquired patterns have to be translated into the *Corpus Query Language (CQL)*, so that a pattern matching can be performed on the BNC documents. The patterns from VerbOcean do not cause any trouble here. For example, the pattern

```
either X or Y
```

translates into the CQL-query

```
<seq>
   <lemmas ls="BNC">
      <word>either</word><word>ADV</word>
   </lemmas>
   <pos><word>_</word><poscode key="pos">VERB</poscode></pos>
   <lemmas ls="BNC">
      <word>or</word><word>CONJ</word>
   </lemmas>
   <pos><word>_</word><poscode key="pos">VERB</poscode></pos>
</seq>
```

The same works very well for the other patterns, too. The Corpus Query Language is described shortly in the BNC Handbook (Aston, Burnard 1998, 210), but it is easier to create a query with the XAIRA Query Builder[16] and then export the CQL query phrase.

Hearst's patterns, unfortunately, cannot be translated that simple, because the BNC does not identify noun phrases (NP) or any other phrases for that matter. Therefore, this information has to be obtained separately. For a NP-detection in the BNC, we would need a context free grammar consisting (CFG) of phrase structure rules that exclusively contain non-terminal[17] elements on the right side of each rule that are contained as tags in the BNC sources. For example, the rule

NP → proper_noun

would be directly applicable, because the BNC does contain the POS[18]-tag *NP0,* which identifies proper nouns.(Burnard 2007) Therefore we could formulate a CQL-representation for that rule.

A non-exhaustive CFG for noun phrases, which could be translated into CQL, could look like this:

NP → pronoun
NP → proper_noun
NP → determiner NOMINAL
NP → determiner adjective NOMINAL
NP → determiner NOMINAL PP
NOMINAL → noun
NOMINAL → noun NOMINAL
PP → preposition
PP → preposition NP

It has to be checked whether this small set of rules already suffices for the semantic relation identification process. Furthermore, it has to be ensured that the CQL does not have any problems with recursive definitions (cf. e.g. the line in bold). A CQL-representation for this grammar can be found in appendix 4 - appendix 6.

[16] A function within the XAIRA client.

[17] Non-terminals are any elements that are not further substituted by other (phrase structure) rules.

[18] POS = Part of speech

Hearst uses a “regular-expression-based noun phrase recognizer” (Hearst 1999, 135–136). Unfortunately, the paper referenced by Hearst does not give a lot of information about this piece of software. Such an external tool would, however, be a possible alternative to the above mentioned solution.

Once all patterns are translated into CQL, the queries can be executed in the BNC in order to obtain the final data. The last step (C) is then concerned with the post-processing of this data. Depending on how specific and restrictive the utilized patterns were, the resulting data has to undergo a statistical filtering that removes irrelevant and infrequent result sets. What we are then left with are the potential semantic relations in a raw form. To be of any use, they have yet to be transformed into a standardized format, e.g. logical propositions, a semantic network or any other well defined form.

For relations resulting from patterns like the ones used by Hearst, we have further to decide how much of the individual noun phrases we want to use in the definitions of the semantic relations. Context specific information and unnecessary modifiers, for example, should be disregarded. This, however, depends on the prospective application of the semantic data. (Schulte im Walde, Koller 06.08.2007, 10–11)

3.4.2 Other approaches to semantic relation discovery from corpora

The pattern approach that has just been discussed is not the only way to discover semantic relations. A considerable number of other approaches have been developed to date and the field is far from being covered exhaustively. Hearst gives a whole list of attempts to find better and more effective methods to extract semantic information from corpora. (Hearst 1999, 146–148)

A promising approach, for example, is the usage of co-occurence information. Heyer et al. (22.04.2001) have found out that some semantic relations are more likely to link a specific word with its collocates than others. The relations which are found by a collocation analysis can be classified into three classes, *symmetric*, *anti-symmetric* and *transitive*, but the final naming of each relation has to be done manually or with the help of other tools. By iterating the procedure, which means that the “[...] collocation sets themselves are subjected to the collocation analysis again and again” (Heyer et al. 22.04.2001, 5), it is possible to obtain “[...] collocation sets carrying a homogeneous semantic relation.”(Heyer et al. 22.04.2001, 5)

It is, furthermore, possible to use this approach to compute relations not only between words and their collocates but also to determine which relation, if there is any, holds between any two words A and B. This is done by compiling both collocate sets of A and B. If these sets overlap, i.e. A and B have at least one joint collocate C, the relations A-C and B-C can be used to deduce the relation between A and B.

Besides using collocation as a measure of association, it is also possible to rely on colligation. In a corpus-based approach, Nastase, Sokolava and Szpakowicz (2006) have used what they called *grammatical collocations* in order to identify noun-modifier relations. By adding grammar to the equation, this approach, in a way, meets Sinclair’s claim to take both lexis and grammar into account when studying the meaning of language. (Sinclair 2000)

4 WordNet and the Prototype Theory

"In order to function in the world, all creatures, including humans, need to be able to group different entities together as instances of the same kind. Our cognitive apparatus does this for us automatically, most of the time. We 'automatically classify things around us as 'books', 'pencils', 'trees', 'coffee-cup', and so on. [...]" (Taylor 2007, xi) From the linguistic standpoint, categorization is interesting, because "[c]ategorizing something very often involves naming it" (Taylor 2007, xi) and because "language itself is an object of categorization."(Taylor 2007, xii) The hierarchical relations in WordNet by the same token constitute a hierarchy of categories. WordNet's categories are to be assigned to the *classical theory of categorization*, which "capture[s] our intuitions about 'essences' [and] make[s] possible an elegant account of the semantic relations that hold between words and between sentences."(Taylor 2007, 35) However, this theory, which bases class membership solely on a matchup of defining features, has been proven to be too strict and not flexible enough to be suited as the only model of categorization. More often than not, we need categories with fuzzy boundaries rather than classes with clear-cut shapes in order to account for the endless variety of entities in the real world that need to be categorized.

In the following chapter, an alternative theory of categorization, namely the prototype theory, will be discussed. Afterwards, I will introduce a project idea that tries to apply this theory to the hierarchies in WordNet in order to extend them by so-called *basic-level* and *prototype information*. Finally, the results of a small scale test run of this project will be evaluated.

4.1 Prototype theory

Many scholars have engaged themselves in studying categorization in language. The philosopher Wittgenstein was one of the first to notice that the classical approach might not be the perfect solution. When he studied the category *game*, he found that it was by no means easy to define the features something must necessarily possess to be called a *game*. He came to the conclusion that it might not be possible to characterize categories by defining properties alone, but rather by *family resem-*

blance. This was the first step to a more fuzzy definition of criteria for category membership. (Lakoff 2005, 16)

Others like J.L. Austin, Lotfi Zadeh, Floyd Lounsbury, Brent Berlin and Paul Kay acknowledged the thoughts of Wittgenstein and engaged in their own research of human categorization. Roger Brown was then the first who realized that categories are not all equal. In observing children's first language acquisition, he found out that there is a certain level of categorization which children learn first and that this level has distinctive properties that distinguish it from the others. (Lakoff 2005, 14) He found it to be "the level of distinctive actions [,] [...] the level which is learned earliest and at which things are first named [, and] the level at which names are shortest and used most frequently."(Lakoff 2005, 32) Brent Berlin and his associated confirmed and extended Browns theory of what was later named the *basic level* (Lakoff 2005, 14), but it was not before the work of Eleanor Rosch that all these rather special case-studies in that field were put into a "general perspective" which is now known as "'the theory of prototypes and basic-level categories,' or 'prototype theory.'"(Lakoff 2005, 39)

With the *prototype theory*, Rosch (1978, 28–30) introduces two basic principles for the formation of categories, *cognitive economy* and *perceived world structure*. The former principle "asserts that the task of category systems is to provide maximum information with the least cognitive effort [whereas the latter] asserts that the perceived world comes as structured information rather than as arbitrary or unpredictable attributes." (Rosch 1978, 29)

Although the view on categories within the prototype theory differs from the way the classical theory sees them, the overall hierarchical structure is valid for both approaches. The prototype theory distinguishes between the vertical and the horizontal dimension of a hierarchy of categories, whereupon "[t]he vertical dimension concerns the level of inclusiveness of the category [...] [and t]he horizontal dimension concerns the segmentation of categories at the same level of inclusiveness." (Rosch 1978, 30) Both dimensions deserve closer attention.

The vertical dimension of categorization is governed by the hierarchical relation-pairs hyponymy/hypernymy and meronymy/holonymy which we have discussed earlier. However, the taxonomies that we know from WordNet are based on the assumption that every level in the hierarchy is of equal importance. “Yet, as Berlin [...] and Hunn [...] have shown for Tzeltal plant and animal taxonomies, the level of the biological genus is psychologically basic” (Lakoff 2005, 46) – a fact that is not reflected in WordNet’s hierarchies. “Rosch and her associates have extended the study of basic-level effects from cognitive anthropology to the experimental paradigm of cognitive psychology.”(Lakoff 2005, 46) Their research showed that the basic level is:

> - The highest level at which category members have similarly perceived overall shapes.
> - The highest level at which a single mental image can reflect the entire category.
> - The highest level at which a person uses similar motor actions for interacting with category members.
> - The level at which subjects are fastest at identifying category members.
> - The level with the most commonly used labels for category members.
> - The first level named and understood by children.
> - The first level to enter the lexicon of a language.
> - The level with the shortest primary lexemes.
> - The level at which terms are used in neutral contexts. [...]
> - The level at which most of our knowledge is organized.
>
> (Lakoff 2005, 46)

The fact that a single level in any hierarchy of categories sticks out by so many distinctive qualities shows that not every level in the hierarchy can be of equal importance, as WordNet suggests. According to “Rosch and her co-workers [...], basic-level distinctions are ‘the generally most useful distinctions to make in the world,’ since they are [...] at the most general level at which one can form a mental image.” (Lakoff 2005, 49) Consequently, it would be very useful to identify these basic levels within the WordNet hierarchies to gain a better view on the cognitive organization of language.

“If the classical theory were both correct and complete, no member of a category would have any special status.” (Lakoff 2005, 40) This, however, seems to be as erroneous an assumption as the notion that every level in a hierarchy is of the same significance. The prototype theory therefore also accounts for the internal structure of categories which is represented by the horizontal axis of categorization. According to the classical theory, categories have clear cut shapes and membership in a category is defined by a set of defining features. But, this definition has been proven to be oversimplified. Categories do not always have clear cut shapes. In fact, class boundaries are more often than not fuzzy. Labov , for one, has made classification experiments in which he asked the participants to name drawings of different shaped receptacles. He found out that there were some containers that were named cup, bowl or vase by all of the subjects, whereas for some containers a variation in the description could be noticed. This variation is a hint on the fuzzy boundaries of the categories vase, bowl and cup. The artefacts without any variation in their denomination were the most typical instances of their respective category. Rosch later called them *prototypes* or *cognitive reference points* of the category. (Lakoff 2005, 41) With the help of these prototypes, class membership can be defined without relying alone on defining features. Prototypes serve as fixed points for measuring family resemblance. They might vary among different cultures, languages (Taylor 2007, 45) and also change over time. (Taylor 2007, 59)

In opposition to what was first assumed, "[p]rototypes [alone] do not constitute any particular processing model for categories."(Lakoff 2005, 44) Nevertheless, adding prototype information to the classical definition of categories helps to overcome the deficiencies in respect to fuzzy category boundaries and might even account for the often neglected internal structures of categories[19] that are, too, not represented in WordNet. (Lakoff 2005, 45) The availability of typicality ratings and prototype information for lexical items in WordNet could also be used to filter WordNet items according to their typicality. This way, the already mentioned problem of the proportionally large number of infrequent and uncommon words could be mastered.

[19] Not all categories have fuzzy boundaries. The category *bird*, for example, has a clear-cut shape. It can be clearly decided which entities belong to this category and which do not. Nevertheless, prototype effects are noticeable in this category – some birds are more typical than others. These effects, however, must result from some internal structure rather than from fuzzy category boundaries. Lakoff 2005, 45

In the following section, an idea for a project will be introduced which aims at the semi-automatic extension of WordNet with basic level and prototype information.

4.2 Project idea

The main goal of the project is the identification of the basic levels for any hyponymous hierarchy in WordNet, because, as of the current version 3.0, this data is not included in the semantic net. Furthermore, some prototype information is acquired so that the horizontal dimension of categorization is also accounted for. In this section, the general ideas of the project will be introduced. The subsequent section then describes the methodology in greater detail by discussing the actual software implementation. Afterwards, the data obtained by a small-scale study using this software will be analyzed and the results evaluated.

Although it is known that the basic level tends to be found somewhere in the middle of the hierarchy, "[...] basic level terms do not always occur on the same level [...]. Rather, the level is liable to vary depending on the nature of the category in question." (Taylor 2007, 54) Furthermore, basic levels are "human-sized", which means that they cannot be considered independent from people. (Lakoff 2005, 51) They are not innate to the physical world, but they are part of the perceived world and subjective to a specific knower. (Rosch 1978) "They depend not on objects themselves, [...] but on the way people interact with object: the way they perceive them, image them, organize information about them and behave toward them with their bodies." (Lakoff 2005, 51)

This obviously makes an attempt to automatically identify basic levels in WordNet hierarchies a very complex task, because at the current state of affairs, a computer does not exactly equal a *human knower* with an *embodied mind*. Therefore, the approach of this project is to divide the overall task into two parts – automatic pre-processing and manual selection – which are finally followed by a statistical analysis.

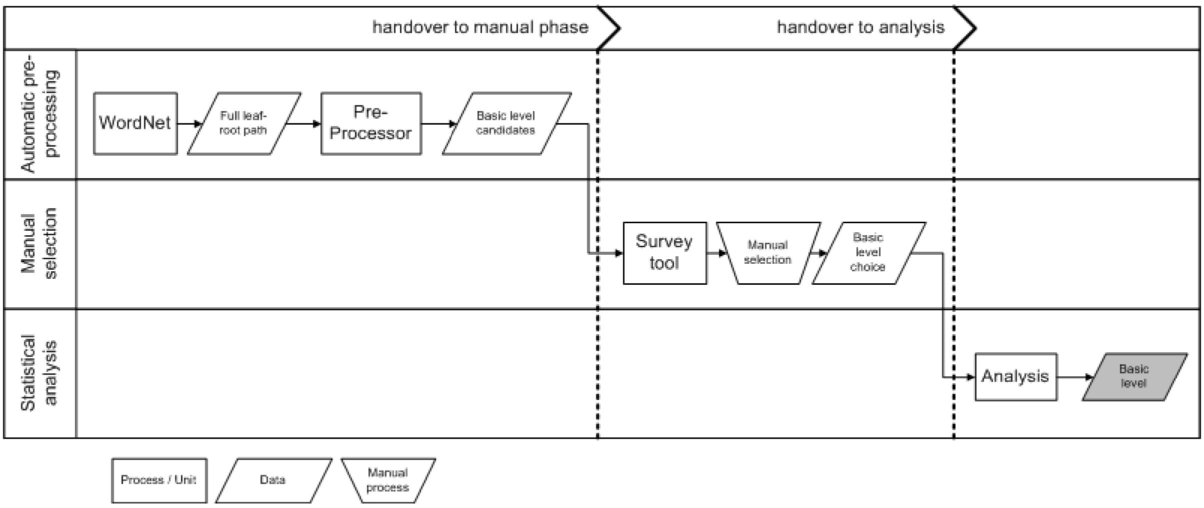


figure 16 : Project workflow

In the first phase, full *leaf-root paths*[20] are extracted from hyponymous WordNet hierarchies. Because hyponymous taxonomies are usually trees, each hierarchy has exactly as many leaf-root paths as it has leaf nodes. Each of these paths is then pre-processed, i.e. every level of each hierarchy is analyzed in terms of basic level properties, which will be discussed in more detail in the next section. The pre-processor finally terminates with a set of basic level candidates that provides the data base for the second phase.

In the second phase, a web-based survey application automatically compiles one-choice questions using the basic level candidates. From each candidate-set, the participants in the study are asked to choose one item that is, for them, especially psychologically salient and which, in their eyes, best represents the leaf synset of the hierarchy from which the basic level candidates were taken. The exact procedure will be elaborated in the following section. Subsequent to these basic level questions, the participants are shown their chosen candidates again. They are then asked to rate the typicality of every lexical item from the leaf synset that appeared in the respective basic level question earlier. It is expected that the items which are rated most typical in respect to their corresponding basic level are also prototypical for the whole semantic field they belong to.

[20] In a tree, a *leaf-root path* is the node sequence that connects a leaf-node with the root node. As trees are by definition acyclic, the leaf-root paths are unique for every leaf-node. Occurrences of cycles in WordNet-hierarchies can be managed as described in chapter 2.

Ultimately, as soon as the survey has been completed by all participants, the data resulting from the questionnaire is statistically analyzed so that a final and representative selection of basic levels can be made.

4.3 Software design and implementation

4.3.1 General technical decisions

The software has been completely programmed in the JAVA programming language, which has been proven to be a perfect choice for these kinds of application.

The preprocessor is a standalone desktop application. It accesses WordNet via the *MIT Java WordNet Interface (JWI)* available from the *MIT Computer Science And Artificial intelligence Laboratory* (http://projects.csail.mit.edu/jwi/). Furthermore, it has a link to the British National Corpus, which provides the necessary frequency information. This link is established by a XML-RPC[21] connection to the XAIRA_DAEMON[22], which accesses the XML data of the corpus (http://xaira.sourceforge.net/). The most recent version of the java interface to the daemon at the time of publication is 1.23. This version contains a bug that prohibits a successful connection to the corpus. The responsible software developer from *Oxford University Computing Services*, however, provided a fix for that problem. It will be included in future versions of the interface.

The preprocessor communicates indirectly with the survey application and the data analysis tool via a jointly used MYSQL database.

The survey application and the analysis tool are both JSP web applications running on an Apache Tomcat application server[23]. They get all the necessary data from the database and do not require any further connections to WordNet or the BNC.

[21] Extensible Markup Language Remote Procedure Call

[22] A *daemon* is a computer program that runs in the background and offers a service, which, in this case, is a high level access to the BNC XML data.

[23] More precisely, Apache Tomcat should be described as a *servlet container*. This, however, would call for further explanation, which is not necessary in this context.

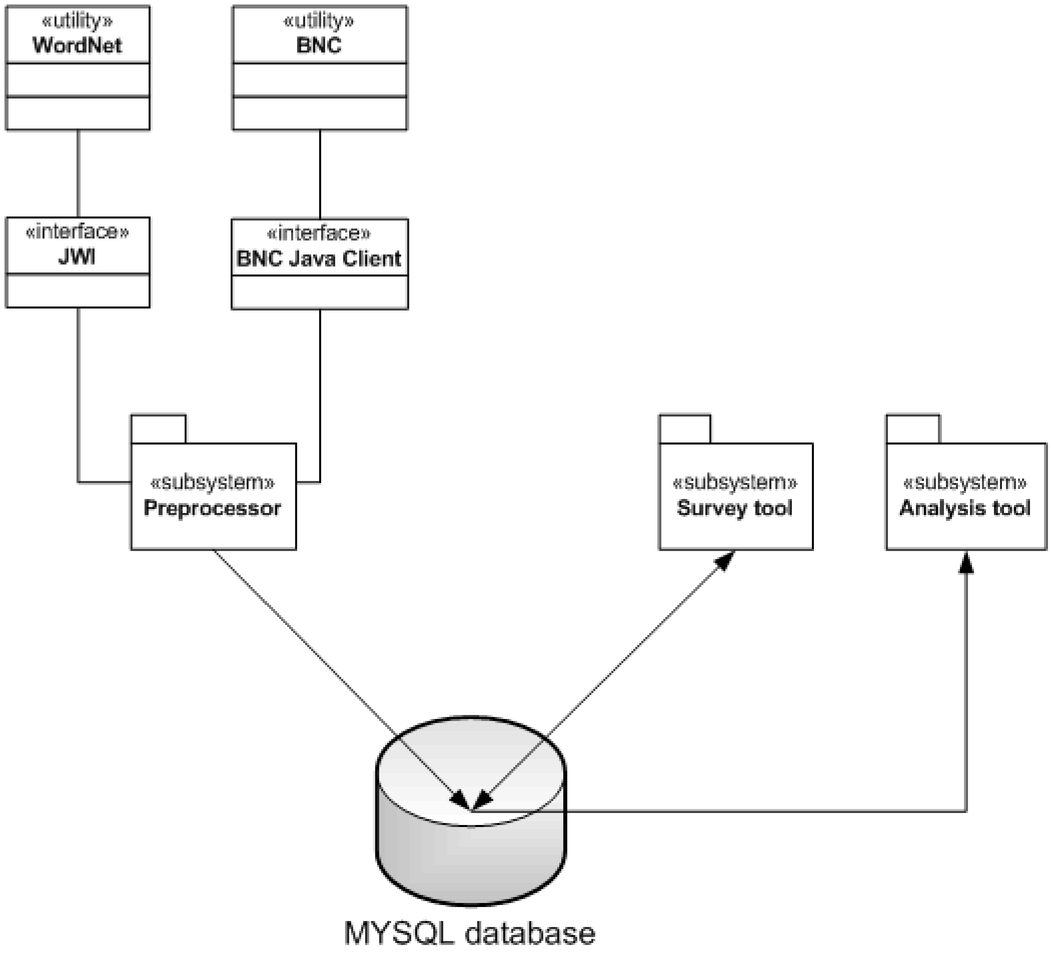


figure 17: Overview over the software architecture

4.3.2 The preprocessor

As it has already been mentioned in section 5.2, the task of the preprocessor module is to extract the data from WordNet and make it available as POJOs[24]. This way, the data can be further processed in order to identify the basic level candidates, which are then written to the database.

To start the preprocessing procedure, the user has to execute the *PreprocessorStart*-class which expects a single word as input. The preprocessor then performs a WordNet query to obtain all hyponym hierarchies in which the specified word is contained. Depending on the program settings, either all senses of the word are considered or only the first sense is taken into account. In order to cover all nouns in WordNet, the preprocessing would have to be carried out for every unique

[24] POJOs = Plain Old Java Objects. A technical term which stresses that the data has the form of ordinary JAVA objects that can easily be processed by any JAVA-based program. Richardson 2006, xix–xx

beginner. For the small-scale study in the course of this thesis, only the first senses of a chosen set of query words have been used.

The next step in the procedure is to extract all leaf-root paths from the hierarchies and store them in a dedicated data structure that simplifies further processing (cf. figure 18). To achieve this, all leaf-nodes in each hierarchy are identified. Afterwards, the path from each leaf node to the respective hierarchy root is followed and every node in that path is stored in a *Hierarchy* object[25]. Within that *Hierarchy* object, which depicts a single leaf-root path, each node is represented by a *LevelObject*. This *LevelObject* contains a link to the synset data and also the information at which level the synset appeared in the original hierarchy, which will be important for the later basic level rating. In the case of tangled hierarchies, the "diamond problem" is solved as described in figure 8, by considering every possible path separately[26]. For the small-scale test within this thesis, only those leaf-root paths are used whose leaf-synsets have a *combined BNC-frequency*[27] larger than a specified limit. This limit varied between 500 and 3000.

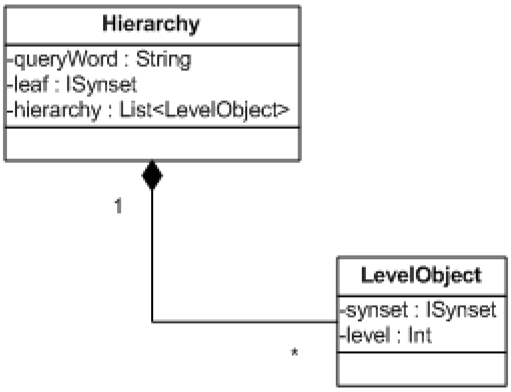


figure 18: Data structure for leaf-root paths

After all the leaf-root paths have been extracted, they can further be processed to indentify the basic level candidates. This takes place in several steps. The criteria for the candidates are based on the criteria for basic levels discovered by Rebecca Green (2005), who was working on a method to fully automatically identify

[25] Originally, the Hierarchy-object was supposed to represent a whole WordNet hierarchy. The usage, however, has changed during the software development process. Therefore it is important to keep in mind that the Hierarchy object represents a single leaf-root path and not a complete hierarchy.
[26] In this case, no such distinction between *formal* and *telic* is made.
[27] The sum of the frequencies of all lexical units contained in the synset.

basic levels in WordNet. These criteria have been modified to fit the needs for a semi-automatic approach, which lead to the following processing steps.

(1) Crop hierarchy (SS)
(2) Lexical analysis (LI)
(3) Check relative properties[28] (SS)
(4) Check qualifying absolute properties (SS)
(5) Check disqualifying absolute properties (SS)

(LI) = performed on lexical items, (SS) = performed on synsets

In step (1), the top 50% of the hierarchy is being disqualified right from the beginning, because experiments have shown that the upper half of a hierarchy only contains concepts that are too generic to be at the basic level. The cropping point, however, can be set to a different percentage by changing the value in the preprocessor setup.

In step (2), a lexical analysis is performed on every lexical unit in every remaining synset. At this point, lexical items consisting of more than one word are disqualified. Furthermore, a length check is performed, keeping only the words with a length between 2 and 15 characters. Finally, the BNC frequency for each remaining item is checked. Every item with a frequency below 10 is automatically disqualified. In order to perform this step, the original WordNet synsets have to be altered.

Step (3) originally consisted of two different tests checking the properties of each synset relative to the whole hierarchy. Here, the synset with the smallest number of children in the hierarchy and the synset with the least relations to other concepts are disqualified. This check has, however, been disabled, because it has produced too many false negatives. It is possible that it could be used with a finer grained voting algorithm.

Steps (4) and (5) are concerned with the absolute properties of the synsets. In step (4), two characteristics are checked that qualify a synset to be a basic level candidate. "If the name of a concept is included within the name of a more specific

[28] The utilization of the *relative property check* has lead to too many false negatives. Because too many good candidates have been disqualified this check has been disabled.

concept, it probably names the basic level category."(Green 2005, 8) Furthermore, all synsets that have more than one part are also candidates for the basic level.

Step (5) finally checks two characteristics that disqualify synsets to be basic level candidates. Any synset without children is automatically disqualified, because the basic level usually occurs in the middle of a hierarchy and not at the leaf-level. Ultimately, all the synsets are disqualified that have more than X succeeding levels. X can be set to any arbitrary value. The default limit was set to 5. Concepts with more than five succeeding levels are more likely to be at a superordinate level and not the basic level.

After these checks have all been performed, only the basic level candidates are left. They are then written to the database so that the data can be used by the survey application and the analysis tool. This is done by performing an object-relational mapping of the data structure in figure 18 to the entity-relationship model of the database. (cf. appendix 8) At this point, even the synset data is mapped to the database, as the synsets had to be disassembled in step (2) anyway and because the web applications are not allowed to access WordNet directly.

4.3.3 The survey application

The survey application has been implemented as a JSP web application running on an Apache Tomcat Server. The necessary data for the survey is completely contained in the database. However, for each item that originated from WordNet, the respective *lexical item id* or *synset id* is available to make later reference to the semantic net possible.

Each survey is associated with a *user id* that has either to be entered at the login page (cf. appendix 9) or that is passed on automatically by a personal URL. This was necessary to make statistical calculations easier and less error-prone. Furthermore, it is convenient for the user, because they can adjourn the survey at any time and come back to it at a later time without having to answer all the questions again.

After the first login, each participant is shown the instruction page that can be seen in appendix 10. It describes the overall procedure and gives an illustrated example for the questions that await the user.

The following survey is separated into the basic level part and the prototype part. The basic level section is the main element of the survey. In this part, each data set that has been produced by the preprocessor is retrieved from the database, mapped to the data structure displayed in figure 19 and then transformed into a question form (cf. appendix 11) by the survey application. This form is then presented to the participant, whose task it is to make a single basic level choice for each available data set.

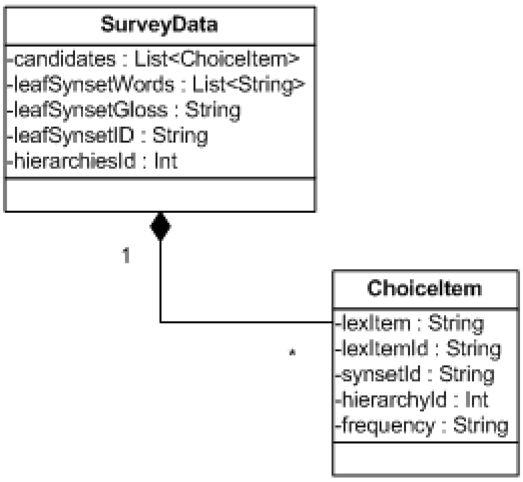


figure 19: Data structure for survey questions

On the top of each form, the lexical items are shown for which the participant must pick out the basic level in the list of basic level candidates. The entries in this candidate list are shuffled every time the question is displayed, so that no artificial order or organization can influence the decision of the participant. The selection criteria had been explained to the user on the introductory instruction page as follows:

> A good analogy is, to think of it as a game, in which you get a word (or a couple of words) and you have to choose a term that describes this word without giving away too much specific information, but which is also a term that includes the basic characteristics of the original term, so that another person will be able to guess the original word.
>
> **For example:**
>
> For the term "aircraft carrier" you have, among others, the choices "warship", "ship" and "vessel"

- The term warship is too specific and thus gives away too much information about the original term. It would not be hard to guess the original word "airship carrier"
- The term "vessel" is far too generic. It is virtually impossible to guess "aircraft carrier" in a short amount of time.
- The term "ship" would be a good choice, because it describes the basic properties of "aircraft carrier", but it is not too specific. (see appendix 10)

A similar, shorter instruction is repeated in the hint-section on each question form. Because many of the question terms are very specific and uncommon for the non-native speaker, some of the participants might not be familiar with them. Therefore, a definition can be shown by clicking the “*Show definition*” link (cf. appendix 11, appendix 12). This definition, which is basically the explanatory gloss of the leaf-synset, is not displayed by default, because it might influence the choices of the participants by utilizing one of the basic level candidates in the definition text.

Subsequent to the basic level part, the prototype survey follows, in which the users are asked to make typicality ratings for the leaf concepts associated with the selected basic levels from the first part (cf. appendix 13). Although prototypes exist on every level in a hierarchy (cf. section 5.1), the ratings were restricted to these leaf concepts. The scale for this rating has been set to 5 levels of typicality, because a finer grained rating would probably produce less meaningful results because of the inability of the participants to make such fine grained decisions. However, the scale can easily be changed in the setup of the survey application. In addition to the ratings of the leaf concepts, the participants also have the possibility to enter an item of their own choosing[29], which, in their opinion, is especially typical for the respective basic level term. This way, even items that do not originally occur in WordNet are accounted for.

The user choices are written to the database after the submission of every single question instead of being stored in bulk at the end of the survey, because the participants should have the option to interrupt and resume the survey at any time.

[29] Later referred to as *free text prototype*

For this reason, however, it is not possible to return to a question once it has been submitted.

4.3.4 The data analysis tool

The data analysis tool is, like the survey tool, also a web application that can be used to display all the relevant data in the database and to statistically analyze the choices made by the participants of the survey.

The entry point of the tool is an overview of all data sets that have been created by the preprocessor (cf. appendix 14). For each data set, the leaf-synset, the explanatory gloss and the candidate-choice ratio[30] is listed. The list entries are automatically clustered according to the original query used for the preprocessing (e.g. ‘furniture’, ‘vehicle’, etc.).

A click on any item in the list leads to the respective *basic level analysis page* (cf. appendix 15, appendix 16). This page lists the basic level candidates for the particular hierarchy and the choices that have been made so far in the survey. For each choice, the number of users that have chosen the term and the resulting percentage are shown. Furthermore, it is also possible to list these users in order to see the other choices they have made[31].

The “*Typicality*”-link next to each basic level choice leads to the individual *prototype analysis pages* (cf. appendix 17 - appendix 19). In the standard view, for each leaf-item of the basic level, three statistical values are displayed, the *average typicality rating*, the variance *population standard deviation*[32] of this rating and its *variance*[33]. On demand, the page can be switched to a more detailed view which shows the individual user ratings (cf. appendix 18). In the case that any user has entered a free text prototype, all manual entries are displayed at the top in the list titled “*Free text prototype entries*” (cf. appendix 19).

[30] candidate-choice ratio = $\frac{number\ of\ different\ choices\ made\ by\ the\ participants}{number\ of\ basic\ level\ candidates}$. A ratio of 1.0 denotes that every basic level candidate has at least been chosen once by a user.

[31] This could be used, for example, if only one user has chosen a term that is obviously not a basic level term. By listing the other choices of this user, it could be discovered that the participant did not take the survey seriously.

[32] population standard deviation = $\sqrt{variance}$

[33] variance = $\frac{\sum(X-m)^2}{N}$, with X=individual typicality ratings, m=average typicality rating, N=number of ratings

In the current version, the data analysis tool uses only the data from the database. It may, however, be interesting to connect the analysis tool to WordNet directly. Using the *lexical item ids* and *synset ids*, which have been stored during the preprocessing, this link could easily be established and might offer a greater potential for further research.

4.3.5 Availability

The software described in sections 4.3.2 to 4.3.4 is included on the CD-ROM that accompanies this thesis. The web applications can be accessed directly under the URLs

http://survey.ferschke.de (for the survey tools)

and

http://survey.ferschke.de/analysis.jsp (for the analysis tool)

In order to being able to use the survey tool, a user id has to be created[34], which can be done via the "Create new login"-link. One must, however, keep in mind that the usage of the survey tool has direct influence on the data displayed in the analysis tool.

4.4 Analysis

In order to test whether the previously described approach really produces satisfactory results, a small-scale study has been carried out with 17 students from the University of Würzburg. As test data, hierarchies (leaf-root paths) with high frequency[35] leaf synsets have been used. They were taken from three different semantic fields[36], *athletics*, *vehicles* and *furniture*. The frequency check was performed without any sense disambiguation, which is why the resulting frequency values are far from being representative for the actual meaning, but they can still be used for rough orientation. Although it would generally be desirable to have sense-specific

[34] User id creation is usually disabled during a study. Only the ids that have been provided to the users may be used as long as the study lasts.

[35] The limitation on high frequency synsets was necessary because of the small scope of the study.

[36] This was done by using the query words *athletics*, *vehicles* and *furniture* in the preprocessor.

frequency information, it is not clear whether the data would be very helpful in this case, because most leaf-concepts in WordNet are very infrequent and specific.

The preprocessing produced 33 basic level questions. The number of the subsequent prototype questions was subject to variation for each individual user, depending on the answers in the basic level part of the survey.

In figure 20, the final basic levels from all three semantic fields are shown, which resulted from the first part of the study. The numerical value next to each basic level term shows the percentage of participants who chose this term as the basic level. The *mult*-flag denotes that there either were multiple items with the same (maximal) percentage (e.g. *{ round of golf , round }*) or that the question for the respective synset appeared twice in the survey and resulted in two *different* basic level terms (e.g. *{ soccer, association football }*). In that case, one of the items was chosen automatically by a procedure similar to the basic level candidate preprocessing. The underlying data for each basic level can be seen in appendix 20 - appendix 30.

	basic level term	%	mult
Semantic field "athletics" *{sport, athletics}*			
{ fight }	**boxing**	**0,47**	
{ rugby , rugby football , rugger }	**football**	**0,35**	x
{ soccer , association football }	**football**	**0,41**	x
{ track , running }	**athletics**	**0,47**	
{ dip , plunge }	**swimming**	**0,35**	
{ round of golf , round }	**golf**	**0,29**	x
{ steal }	**baseball**	**0,76**	
{ cricket }	**badminton**	**0,52**	
{ singles }	**tennis**	**0,47**	
Semantic field "furniture" *{furniture, piece of furniture, article of furniture}*			
{ lower berth , lower }	**bed**	**0,23**	x
{ upper berth , upper }	**berth**	**0,29**	x
{ settle , settee }	**bench**	**0,35**	
{ bench }	**seat**	**0,52**	
{ box , box seat }	**furniture**	**0,35**	
{ secretary , writing table , escritoire , secretaire }	**desk**	**0,47**	
Semantic field "vehicle" *{vehicle}*			
{launch}	**motorboat**	**0,29**	
{ tugboat , tug , towboat , tower }	**boat**	**0,52**	
{ bottom , freighter , merchantman , merchant ship }	**ship**	**0,52**	
{ technical }	**instrumentation**	**0,23**	
{ bobsled , bobsleigh , bob }	**sled**	**0,29**	
{ ordinary , ordinary bicycle }	**bicycle**	**0,29**	
{ flatcar , flatbed , flat }	**transport**	**0,47**	
{ tank car , tank }	**container**	**0,35**	
(Hornby 2000)	**transport**	**0,47**	
{ stagecoach , stage }	**carriage**	**0,47**	
{ bus , jalopy , heap }	**vehicle**	**0,29**	
{ electric , electric automobile , electric car }	**machine**	**0,41**	
{ van , caravan }	**camper**	**0,47**	
{ Caterpillar , cat }	**vehicle**	**0,58**	

figure 20: The final list of identified basic levels

The best results were reached for the semantic field "*athletics*". Nearly all of the identified terms really are basic level objects. Only the term *athletics* could be seen as problematic, because, strictly speaking, it belongs to the synset *{sport, athletics}*, which is not at the basic level. However, since we are not dealing with synsets but with single lexical items, *athletics* could also be regarded as the synonym of "*track and field*" or the German *Leichtathletik*[37], which could be argued to be a basic level term. This would mean a perfect set of basic level terms for the semantic field *athletics*.

For the lexical field *furniture*, the results are not as good as for athletics, but with a 67% success rate still noticeable. The two false identifications can be traced back to two separate phenomena that will be even more important for the analysis of the third semantic field. The synset *{box, box seat}*, which is described as "the driver's seat on a coach" by its explanatory gloss, is a very specific an infrequent term. When we analyze the complete hierarchy for *{box, box seat}* again, we can say that the term *seat* is the only appropriate basic level term, although it has many characteristics of a superordinate. In fact, the percentage of participants that have chosen *seat* is only slightly lower than the percentage for *furniture* (cf. appendix 30). The reason why *furniture* might eventually have gotten more votes is that people who are not really familiar with a term tend to categorize it with a rather generic superordinate term instead of a more specific one. This probably is the case in this study, because all the participants were non-native speakers and therefore not familiar to the required extent with infrequent and very specific terms.

The other false basic level term was identified for the synset *{upper berth, upper}*. This is a good example for a phenomenon in WordNet that might result from the lack of a more differentiated hyponymy relation (cf. chapter 3.3.2.1). When comparing the leaf-synsets in WordNet, it becomes obvious that they are at complete different levels of specificity. *Upper berth* can almost be seen as a meronym of *bunk bed* instead of a hyponym. *Round of golf*, to choose an example from our first semantic field, is part of a game, whereas *cricket* denotes a whole game by itself. This poses a problem, because when the relation of a leaf-concept and its hypernym

[37] This probably is the term the German speaking participants had in mind.

is very similar to meronymy, people[38] tend to choose the *quasi-holonym* as the basic level (cf. *{upper berth, upper}* – *berth*), although the correct basic level might be further up the hierarchy (*{upper berth, upper}* – *bed*). It might be the case that this effect is less present with native speakers.

The case of *{bench}* furthermore shows that polysemy also has an effect on the basic level identification process. Whereas *seat* seems to be the appropriate basic level item in the hierarchy with the leaf-synset *{bench}*[39], the lexical item *bench*[40] also shows up in the hierarchy of *{settle, settee}*, but this time as a basic level term. In the latter case, *seat* names a superordinate level (cf. figure 21). This makes clear that a lexical item that is on the basic level in one hierarchy does not necessarily have to name the basic level in every other hierarchy it occurs in. The basic level characteristic of *bench*, to stay in line with the example, depends on its concrete sense. It is WordNet's handling of polysemous words that allows terms like *seat*, which are usually rather considered to be a superordinate term, to occur at the basic level.

bench (the seat for judges in a courtroom)	**settle, settee (a long wooden bench with a back)**
seat	**bench** -- (a long seat for more than one person)
furniture, piece of furniture, article of furniture	seat
furnishing	furniture, piece of furniture, article of furniture
instrumentality, instrumentation	furnishing
artifact, artefact	instrumentality, instrumentation
whole, unit	artifact, artefact
object, physical object	whole, unit
physical entity	object, physical object
entity	physical entity
	entity

figure 21: Hierarchies for {bench} and {settle, settee}

With a success rate of only 43%, the results for the semantic field *vehicle* are far from good. This is mainly due to the special vocabulary that occurred in the leaf-synsets. The problems here are basically the same as described for *furniture* before.

[38] Especially when they are rather unfamiliar with the terms as it is the case with the non-native speakers in this study.

[39] With the explanatory gloss "the seat for judges in a courtroom".

[40] With the explanatory gloss "a long seat for more than one person".

It remains to be checked whether the results improve when the survey is done with native speakers of English. It would also be possible to try to use a more generic concept as a base for the questions instead of the leaf concept, when the frequency of the leaf-synset is very low. However, to achieve this, sense disambiguation must be possible in order to obtain correct frequency information.

In general, it can be said that the approach is certainly promising. Even with as little as 17 participants, the results in the field of *athletics* were surprisingly good. A greater number of participants would undoubtedly deliver better results, because statistical analyses become more robust with a larger population. Furthermore, the participation of native-speakers would eliminate the problem of deficient command of language and a lack of vocabulary. In combination with further optimizations and adaptations to some peculiarities of WordNet, the results of this approach could certainly enrich the semantic network.

As it has been described before, the second part of the survey dealt with the acquisition of typicality ratings for all leaf-items of each basic level. The items which were rated most typical in respect to their corresponding basic level were expected to be also prototypical for the whole semantic field they belong to. This assumption generally follows the *prototype-as-exemplar* view that regards prototypes as especially typical exemplars of a category rather than being the conceptual centre or a subcategory of the same category. (Taylor 2007, 63–64) However, the understanding differs from Rosch's definition of prototypes, because prototypicality is (here) not determined within the set of co-hyponyms (the horizontal dimension), but within a whole sub-hierarchy which, at least in parts, represents a distinct semantic field in WordNet[41].

Figure 22 illustrates this idea. When we look at the semantic field "*athletics*", for example, it is assumed that for each basic level (e.g. *football)*, the most typical exemplars (e.g. *soccer*) are also the most prototypical leaf-items in the whole

[41] As we have seen in chapter 3.4.3, WordNet does not contain topical information, which implies that we cannot simply identify semantic fields in the traditional sense, because the items belonging together are not necessarily connected. For the lack of a better expression, s*emantic field* is used here more freely. It denotes the set of items which are connected by hyponymy to a specific superordinate that is usually used as a descriptor for a semantic field, e.g. "furniture", "animals" etc. The hyponyms of these superordinates are sure to be members of the same semantic field, as long as the chosen superordinate is not too generic. These superordinates are also used as query words in the preprocessor.

semantic field. With this kind of information, WordNet could deliver the most typical specific instances (on the leaf level) for each sub-hierarchy.

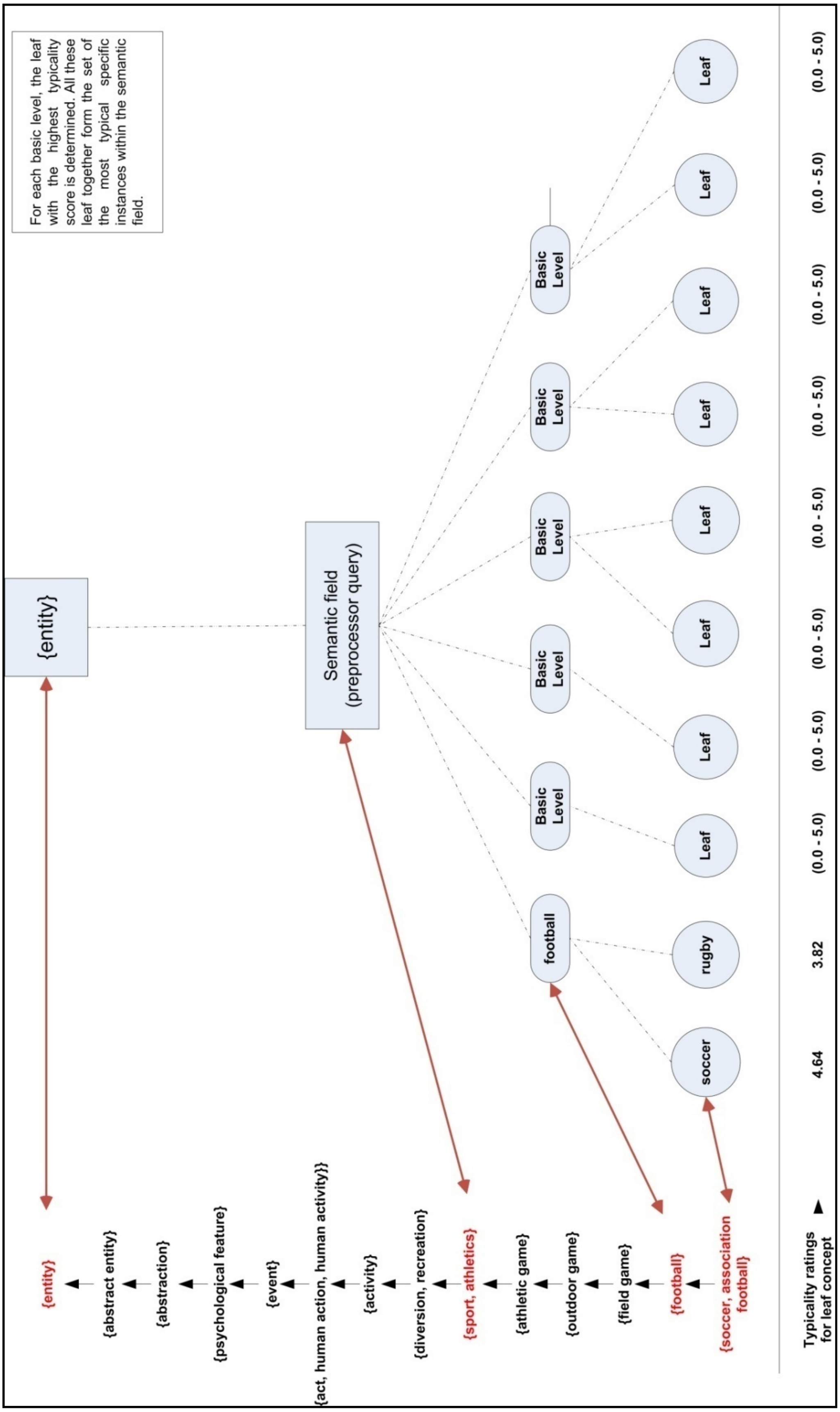


figure 22: Typicality ratings of leaf-nodes in respect to their basic levels

Unfortunately, the study could not prove or disprove the feasibility of this idea due to two main problems. The first problem was the small scale of the study. The basic level questions had to be reduced to a small number, so that the study could be completed within the close time frame. This, however, entails that the number of leaf-nodes had to be restricted, because each leaf node results in a leaf-root path which again results in a new basic level question. Such a restriction of the number of leaf concepts made a reasonable typicality rating process impossible. It would have been necessary to utilize all available leaf-concepts for each basic level concept in this part of the study and not only the ones that had been chosen for the basic level part. This, on the other hand, would have made a direct connection between the survey tool and WordNet necessary, which was not possible because of technical limitations of the web server used in the study.

The second problem arose from the fact that all participants were non-native speakers. Although only high frequency leaf concepts had been chosen, most of them still were very specific. Consequently, most participants were not familiar with many of them. In order to rate the typicality of a lexical item, the participants had to have certain experience with those words. This, however, was not the case. Therefore, no usable typicality data was produced. In the same context, an interesting phenomenon could be observed. Whenever the list of leaf concepts exclusively contained terms that are rather uncommon for non-native speakers, the typicality ratings were pretty uniform on a high level. It seems like the participants were inhibited to give every item in the list a low score. When the list contained at least one item that could be regarded as well-known among non-native speakers, the same item was rated very high whereas the others were rated realistically low.

Despite the unsatisfactory results of the prototype part of this small scale study, it might still be interesting to perform the same procedure with native speakers. Provided that either complete hierarchies are used for the survey instead of only the high frequency branches or by connecting the survey tool to WordNet, this approach might still have the potential to provide results that would greatly complement the basic level information that followed from the first part of this method.

Epilogue

In the previous decades, it was the job of the user to learn a computer language or to memorize formal commands in order to get the information that they desired or to execute the commands that they wanted. At the moment, this situation is about to change. The computer is about to learn to "understand" natural language. For this "understanding", the computer requires semantic information. Emerging technologies like the *semantic web* utilize taxonomies and ontologies to provide a knowledge base for the machines in order that they can process our spoken or written everyday language so that our interaction with them will become more natural. Currently, the research community is full of ideas to exploit semantic information contained in resources like WordNet or corpora like the BNC. In this thesis, I have given a comprehensive account of the semantic relations in these resources. I have described their most important characteristics and how these relations practically hold our lexicon together. Furthermore, it was shown that the greatest value can be gained when combining manually engineered and knowledge based semantic networks with semantic data obtained by an empirical approach based on text evidence.

Language is a product of the human mind. Philosophers say that it is one of the criteria that distinguish the human race from other living beings. In order to fully understand language, which is needed to import this understanding into machines, we must learn how the mind processes language. Because we are not yet at the point that we can fully grasp how our brain functions, approaches which claim to add a cognitive psychological aspect to resources such as WordNet have to rely on human participation. That is the reason why the project that has been described in this paper did not try to identify basic level concepts fully automatically but, instead, used the human brain as input. Even though the trend goes towards an increasing automation, we should not totally lose sight of semi-automatic methods.

However, our future challenge will be to decrease the share of manual labour in these processes and to bring the computer-model of our lexicon to the same level of flexibility and efficiency as our mental lexicon. This would be a big step towards machines which can really think.

Bibliography

Anderson, John Robert, Ralf Graf, and Joachim Grabowski. Kognitive Psychologie. 3., [überarb. und aktualisierte] Aufl. Spektrum-Lehrbuch. Heidelberg: Spektrum Akad. Verl., 2001.

Aston, Guy, and Lou Burnard. The BNC handbook: Exploring the British National Corpus with SARA. Edinburgh textbooks in empirical linguistics. Edinburgh: Edinburgh Univ. Press, 1998.

Barr, Avron, Paul R. Cohen, and Edward A. Feigenbaum. The handbook of artificial intelligence. Vol 1. Stanford Calif.: HeurisTech Press, 1981.

Bublitz, Wolfram. Englische Pragmatik: Eine Einführung. Grundlagen der Anglistik und Amerikanistik 21. Berlin: Schmidt, 2001.

Burnard, Lou. Reference Guide for the British National Corpus (XML Edition). Oxford2007. 18. Aug. 2008. <http://www.natcorp.ox.ac.uk/XMLedition/URG/>.

Chklovski, Timothy, and Patrick Pantel. "VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations". Proceedings of Conference on Empirical Methods in Natural Language Processing. Ed. ACL. Barcelona2004.

“Categories (Aristotle).” Wikipedia, the free encyclopedia. 20. Nov. 2008. <http://en.wikipedia.org/wiki/Categoriae>.

WordNet 3.0 Reference - Morphy. Cognitive Science Lab, Princeton University. 20. Jan. 2009. <http://wordnet.princeton.edu/man/morphy.7WN>.

WordNet 3.0 database statistics. Cognitive Science Lab, Princeton University. 08. Nov. 2008. <http://wordnet.princeton.edu/man/wnstats.7WN>.

Cruse, D. A. Lexical semantics. Cambridge textbooks in linguistics. Cambridge: Cambridge Univ. Press, 1986.

Dörpinghaus, Andreas, Andreas Poenitsch, and Lothar Wigger. Einführung in die Theorie der Bildung. WBG (Wissenschaftliche Buchgesellschaft), 2006.

Erin McKean. "Verbatim" - Erin McKean speaks at Google. 03. Jan. 2009. <http://video.google.com/videoplay?docid=-1588634025806636713>.

Fellbaum, Christiane. "WordNet: Ein semantisches Netz als Bedeutungstheorie". Bedeutung - Konzepte, Bedeutungskonzepte: Theorie und Anwendung in Linguistik und Psychologie. Ed. Joachim Grabowski, Gisela Harras, and Theo Herrmann. Psycholinguistische Studien. Opladen: Westdt. Verl., 1996. 211–230.

Fellbaum, Christiane. "A Semantic Network of English Verbs". WordNet: An electronic lexical database. 2. printing. Ed. Christiane Fellbaum. Language, speech, and communication. Cambridge, Mass.: MIT Press, 1999a. 69–104.

Fellbaum, Christiane. "Introduction". WordNet: An electronic lexical database. 2. printing. Ed. Christiane Fellbaum. Language, speech, and communication. Cambridge, Mass.: MIT Press, 1999b. 1–19.

Fellbaum, Christiane, ed. WordNet: An electronic lexical database. 2. printing. Language, speech, and communication. Cambridge, Mass.: MIT Press, 1999c.

Finlayson, Mark A. MIT Java Wordnet Interface: User's Guide. M.I.T. 08. Nov. 2008.

Grabowski, Joachim; Gisela Harras, and Theo Herrmann, eds. Bedeutung - Konzepte, Bedeutungskonzepte: Theorie und Anwendung in Linguistik und Psychologie. Psycholinguistische Studien. Opladen: Westdt. Verl., 1996.

Graham, Ronald L, Donald Ervin Knuth, and Oren Patashnik. Concrete mathematics: A foundation for computer science. 2. ed. Reading, Mass.: Addison-Wesley, 1990.

Green, Rebecca. "Overview of Relationships in Knowledge Organization". Relationships in the organization of knowledge. Ed. Carol A. Bean, and Rebecca Green. Information science and knowledge management. 2. Dordrecht , Boston , Norwell MA: Kluwer Academic Publishers, 2001a. 3–18.

Green, Rebecca. Vocabulary Alignment via Basic Level Concepts: Final Report 2003 OCLC/ALISE Library and Information Science Research Grant Project. Ed. OCLC Research. Dublin2005b. 13. Jan. 2009.
<http://www.oclc.org/research/grants/reports/green/rg2005.pdf>.

Hearst, Marti. "Automatic Acquisition of Hyponyms from Large Text Corpora". Proceedings of the Fourteenth International Conference on Computational Linguistics (COLING-92). Ed. International Commitee on Computational Linguistics. Nantes1992a. 539–545.

Hearst, Marti A. "Automated Discovery of WordNet Relations". WordNet: An electronic lexical database. 2. printing. Ed. Christiane Fellbaum. Language, speech, and communication. Cambridge, Mass.: MIT Press, 1999b. 131–151.

Herbst, Thomas, Rita Stoll, and Rudolf Westermayr. Terminologie der Sprachbeschreibung: Ein Lernwörterbuch für das Anglistikstudium. 1. Aufl. Forum Sprache. Ismaning: Hueber, 1991.

Heyer, Gerhard, et al. Learning Relations using Collocations. Leipzig University. 22. Aug. 2008. <http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-38/IJCAI_2001_WS_Ontologies_Heyer_etal.pdf>.

Hornby, Andrew S. Oxford advanced learner's dictionary of current English. 5. Aufl. Berlin: Cornelsen & Oxford, 2000.

Knuth, Donald Ervin. Fundamental algorithms. 3rd ed. The art of computer programming Vol. 1. Boston: Addison-Wesley, 2007.

Krenn, B. "Distibutional and Linguistic Implications of Collocation Identification". Proc. Collocations Workshop, DGfS Conference. Ed. Deutsche Gesellschaft für Sprachwissenschaft. Marburg2000.

Labov, William. "The boundaries of words and their meanings". New Ways of Analyzing Variation in English. Ed. Charles-James N. Bailey, and Roger W. Shuy. Washington: Georgetown University Press.

Lakoff, George. Women, fire, and dangerous things: What categories reveal about the mind. [Nachdr.]. Chicago: Chicago Press, 2005.

Leech, Geoffrey. Semantics. Harmondsworth: Penguin Books, 1974.

Löbner, Sebastian. Semantik: Eine Einführung. De-Gruyter-Studienbuch. Berlin: de Gruyter, 2003.

Mihatsch, Wiltrud. Kognitive Grundlagen lexikalischer Hierarchien: Untersucht am Beispiel des Französischen und Spanischen. Linguistische Arbeiten 506. Tübingen: Niemeyer, 2006.

Miller, George A. "Foreword". WordNet: An electronic lexical database. 2. printing. Ed. Christiane Fellbaum. Language, speech, and communication. Cambridge, Mass.: MIT Press, 1999a. xv–xxii.

Miller, George A. "Nouns in WordNet". WordNet: An electronic lexical database. 2. printing. Ed. Christiane Fellbaum. Language, speech, and communication. Cambridge, Mass.: MIT Press, 1999b. 23–46.

Miller, Katherine J. "Modifiers in WordNet". WordNet: An electronic lexical database. 2. printing. Ed. Christiane Fellbaum. Language, speech, and communication. Cambridge, Mass.: MIT Press, 1999c. 47–67.

Nastase, Vivi, Marina Sokolava, and Stan Szpakowicz. Learning Noun-Modifier Semantic Relations with Corpus-based and WordNet-based Features. Ottawa: American Association for Artificial Itelligence, 2006. 22. Aug. 2008. <http://www-etud.iro.umontreal.ca/~sokolovm/compare_contexts_NMRs.pdf>.

Oxford University Computing Services, et al. [oucs] Xaira Page. 03. Jan. 2009. <http://www.oucs.ox.ac.uk/rts/xaira/>.

Pinker, Steven. The language instinct. Harperperennial modern classics. New York NY u.a.: Harper Collins, 2007. how the mind creates language.

Quillian, M. Ross. Semantic memory. Cambridge, Mass.: Bolt, Beranak and Newman, 1966.

Richardson, Chris. POJOs in action: Developing enterprise applications with lightweight frameworks. Java. Greenwich, Conn.: Manning, 2006.

Rosario, Barbara, and Marti Hearst. "Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy". Proceedings of Conference on Empirical Methods in Natural Language Processing. 2001. 82–90.

Rosch, Eleanor. "Principles of Categorization". Cognition and Categorization. Ed. Eleanor Rosch, and Barbara B. Lloyd. Hillsdale, New Jersey: Lawrence Erlabum Assoc., 1978. 27–48.

Russell, Stuart J., and Peter Norvig. Artificial intelligence: A modern approach. 2. ed., Eastern econonmy ed., Indian reprint. Pearson international edition. New Delhi: Prentice-Hall, 2007.

Schulte im Walde, Sabine, and Alexander Koller. Introduction to Corpus-based Computational Semantics: Semantic relations between words. 28. Aug. 2008. <http://www.ims.uni-stuttgart.de/~schulte/Teaching/ESSLLI-07/Slides/sem-sim2.pdf>.

Scriver, Aaron D. Semantic Distance in WordNet: A Simplified and Improved Measure of Semantic Relatedness. 17. Nov. 2008. <http://uwspace.uwaterloo.ca/bitstream/10012/1016/1/adscrive2006.pdf>.

Siemen, Peter Thorben. Hyperonymerkennung: Automatische Generierung von Begriffshierarchien. Berlin2006. 28. Aug. 2008. <http://www2.informatik.hu-berlin.de/~siemen/daten/hagvh.pdf>.

Sinclair, John. Corpus, concordance, collocation. 3. impr. Describing English language. Oxford: Oxford Univ. Press, 1995a.

Sinclair, John. "Lexical Grammar". Naujoji Metodologija. 24 (2000b): 191–203. <http://donelaitis.vdu.lt/publikacijos/sinclair.pdf>.

Sperber, Dan, and Deirdre Wilson. "The mapping between the mental and the public lexicon". UCL Working Papers in Linguistics. . 9. 1997.

Steyvers, Mark, and Joshua B. Tenenbaum. "The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth". Cognitive Science. 29 (2005): 41–78. 05. Nov. 2008. <http://web.mit.edu/cocosci/Papers/03nSteyvers.pdf>.

Taylor, John R. Linguistic categorization. 3. ed., reprint. Oxford textbooks in linguistics. Oxford: Oxford Univ. Press, 2007.

Teubert, Wolfgang, and Anna Cermáková. Corpus linguistics: A short introduction. Repr. London: Continuum, 2008.

“Tree (data structure).” Wikipedia, the free encyclopedia. 20. Nov. 2008. <http://en.wikipedia.org/wiki/Tree_data_structure>.

Trier, Jost. Der deutsche Wortschatz im Sinnbezirk des Verstandes: Die Geschichte eines sprachlichen Feldes. Heidelberg: Winter, 1931.

Truyen, Eddy, et al. "A Generalization and Solution to the Common Ancestor Dilemma Problem in Delegation-Based Object Systems". Proceedings of the 2004 Dynamic Aspects Workshop. Ed. Robert E. Filman, Michael Haupt, Katharina Mehner, and Mira Mezini.

List of figures

List of employed software

The following software products have been used for the software development described in chapter 4. This software must also be installed in order to be able to run the preprocessor.

Princeton *WordNet for Windows 2.1*

Princeton *WordNet 3.0* (database files only)

http://wordnet.princeton.edu

The British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium.

http://www.natcorp.ox.ac.uk/

XAIRA version 1.2.3 (with a bugfixed version of *XairoServer.java*, which is available on the CD accompanying this thesis)

http://www.oucs.ox.ac.uk/rts/xaira/

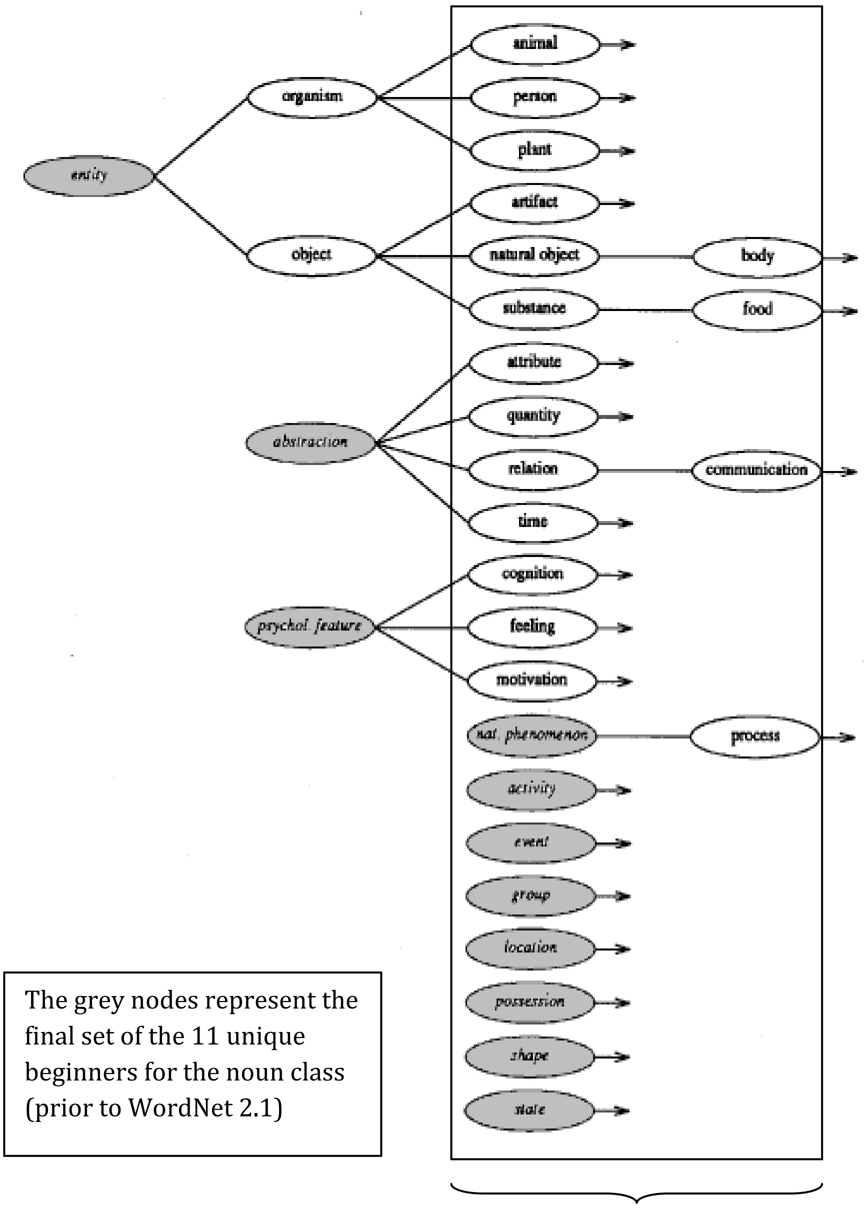
MIT Java Wordnet Interface 2.1.4

http://projects.csail.mit.edu/jwi/

Eclipse Integrated Development Environment 3.4.1

http://www.eclipse.org

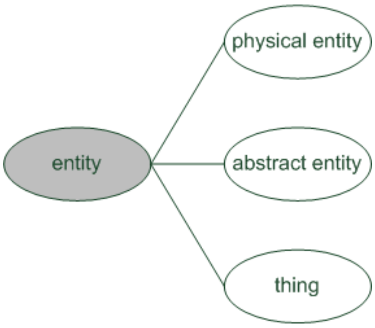
Appendix



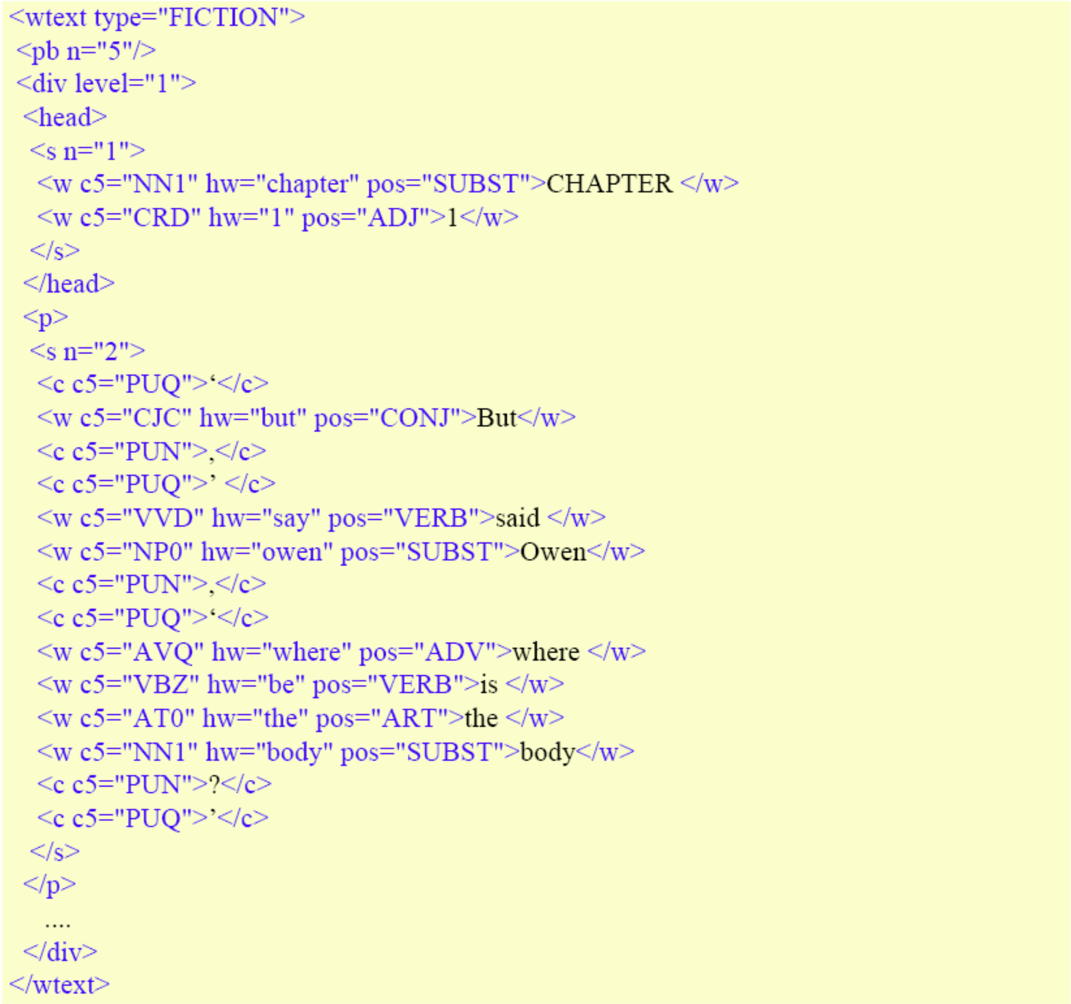
The grey nodes represent the final set of the 11 unique beginners for the noun class (prior to WordNet 2.1)

Initial 25 unique beginners

appendix 1: Diagram of unique beginners for nouns (prior to WordNet 2.1) (cf. Miller 1999b, 30)



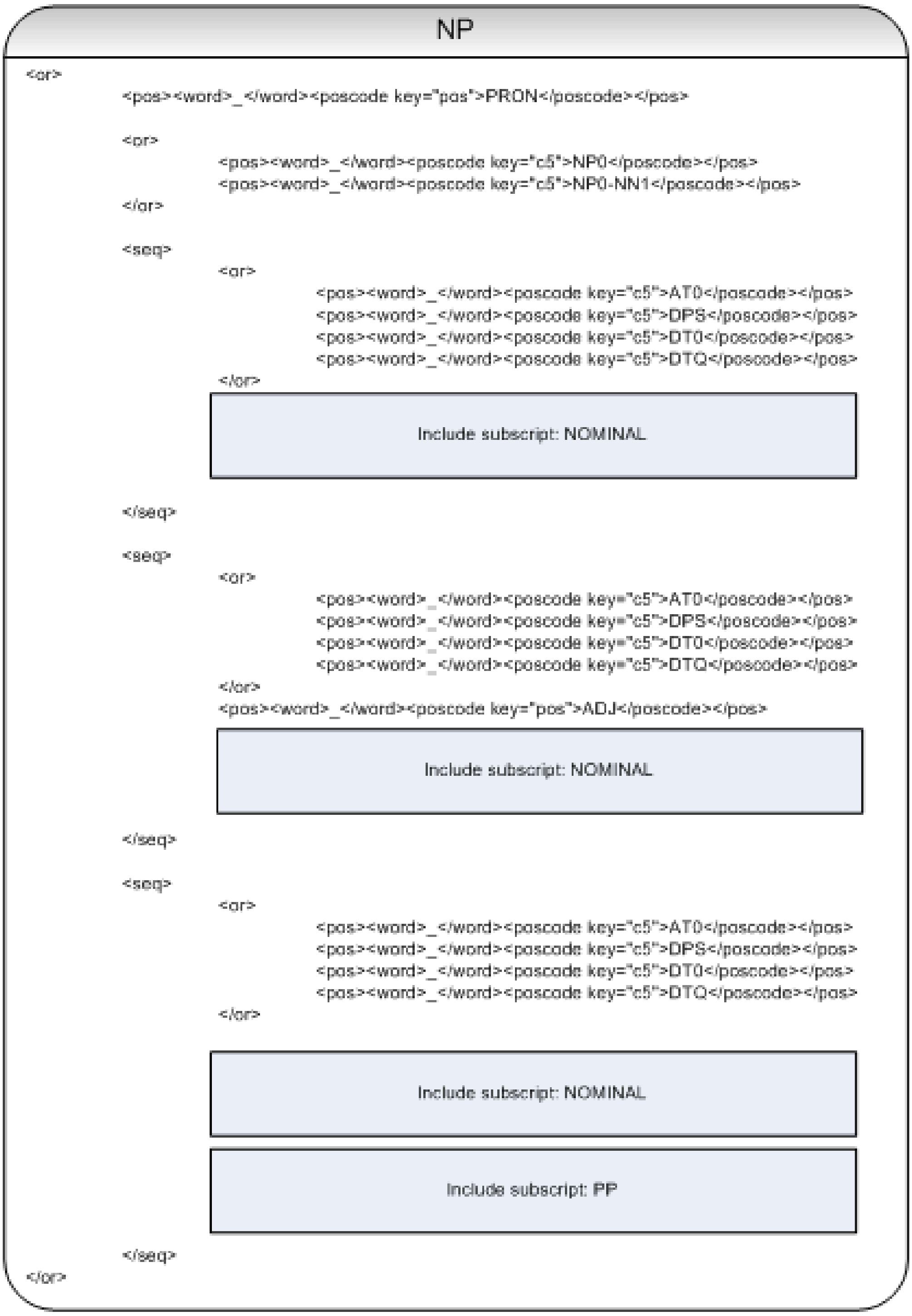
appendix 2: The unique beginner {entity} and its direct hyponyms (as of WordNet 2.1)



```
<wtext type="FICTION">
 <pb n="5"/>
 <div level="1">
  <head>
   <s n="1">
    <w c5="NN1" hw="chapter" pos="SUBST">CHAPTER </w>
    <w c5="CRD" hw="1" pos="ADJ">1</w>
   </s>
  </head>
  <p>
   <s n="2">
    <c c5="PUQ">‘</c>
    <w c5="CJC" hw="but" pos="CONJ">But</w>
    <c c5="PUN">,</c>
    <c c5="PUQ">’ </c>
    <w c5="VVD" hw="say" pos="VERB">said </w>
    <w c5="NP0" hw="owen" pos="SUBST">Owen</w>
    <c c5="PUN">,</c>
    <c c5="PUQ">‘</c>
    <w c5="AVQ" hw="where" pos="ADV">where </w>
    <w c5="VBZ" hw="be" pos="VERB">is </w>
    <w c5="AT0" hw="the" pos="ART">the </w>
    <w c5="NN1" hw="body" pos="SUBST">body</w>
    <c c5="PUN">?</c>
    <c c5="PUQ">’</c>
   </s>
  </p>
    ....
  </div>
</wtext>
```

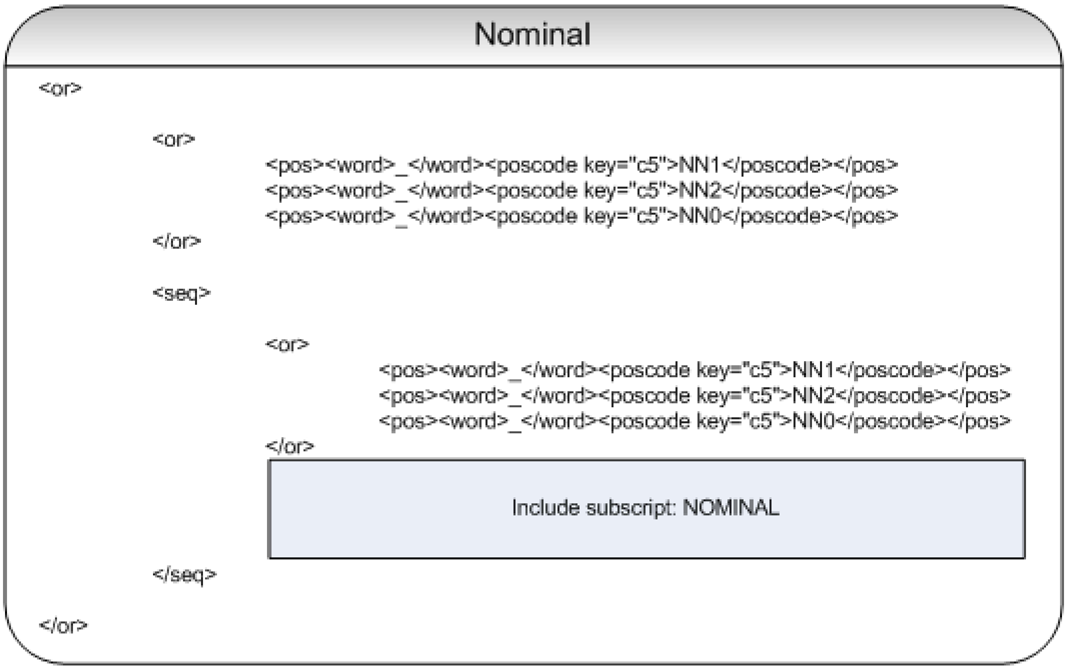
appendix 3: Example for tagged text in the BNC.

Extract from the BNC sources: *CHAPTER 1 'But,'said Owen,'where is the body?'* (Burnard 2007)



```
<or>
        <pos><word>_</word><poscode key="pos">PRON</poscode></pos>

        <or>
                <pos><word>_</word><poscode key="c5">NP0</poscode></pos>
                <pos><word>_</word><poscode key="c5">NP0-NN1</poscode></pos>
        </or>

        <seq>
                <or>
                        <pos><word>_</word><poscode key="c5">AT0</poscode></pos>
                        <pos><word>_</word><poscode key="c5">DPS</poscode></pos>
                        <pos><word>_</word><poscode key="c5">DT0</poscode></pos>
                        <pos><word>_</word><poscode key="c5">DTQ</poscode></pos>
                </or>
                [Include subscript: NOMINAL]

        </seq>

        <seq>
                <or>
                        <pos><word>_</word><poscode key="c5">AT0</poscode></pos>
                        <pos><word>_</word><poscode key="c5">DPS</poscode></pos>
                        <pos><word>_</word><poscode key="c5">DT0</poscode></pos>
                        <pos><word>_</word><poscode key="c5">DTQ</poscode></pos>
                </or>
                <pos><word>_</word><poscode key="pos">ADJ</poscode></pos>
                [Include subscript: NOMINAL]

        </seq>

        <seq>
                <or>
                        <pos><word>_</word><poscode key="c5">AT0</poscode></pos>
                        <pos><word>_</word><poscode key="c5">DPS</poscode></pos>
                        <pos><word>_</word><poscode key="c5">DT0</poscode></pos>
                        <pos><word>_</word><poscode key="c5">DTQ</poscode></pos>
                </or>

                [Include subscript: NOMINAL]
                [Include subscript: PP]
        </seq>
</or>
```

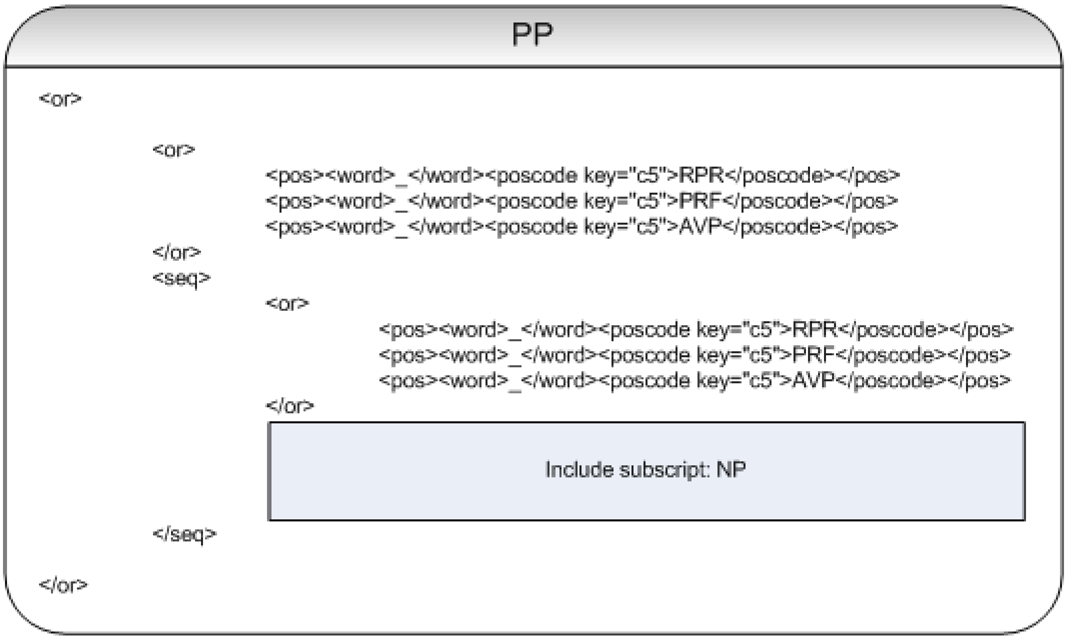
appendix 4: CQL-script for NP detection



Nominal

```
<or>

        <or>
                <pos><word>_</word><poscode key="c5">NN1</poscode></pos>
                <pos><word>_</word><poscode key="c5">NN2</poscode></pos>
                <pos><word>_</word><poscode key="c5">NN0</poscode></pos>
        </or>

        <seq>


                <or>
                        <pos><word>_</word><poscode key="c5">NN1</poscode></pos>
                        <pos><word>_</word><poscode key="c5">NN2</poscode></pos>
                        <pos><word>_</word><poscode key="c5">NN0</poscode></pos>
                </or>
                        Include subscript: NOMINAL

        </seq>

</or>
```

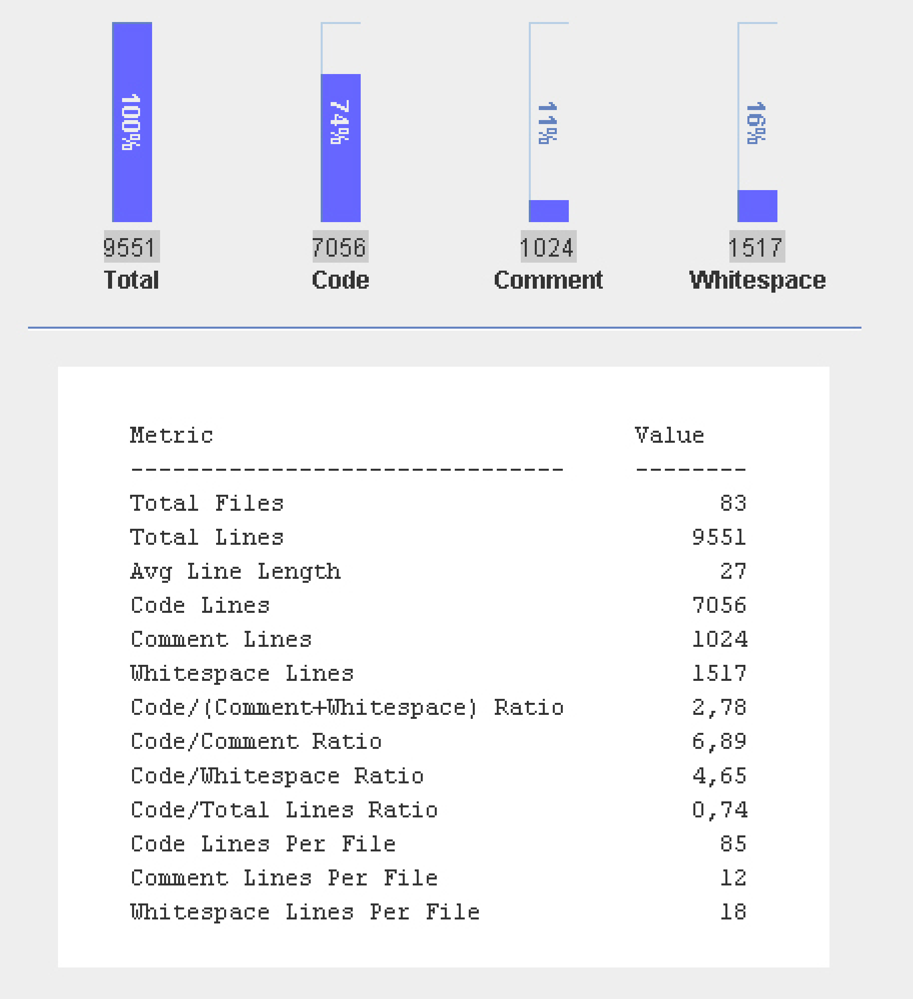
appendix 5: CQL-script for the detection of nominals



PP

```
<or>

        <or>
                <pos><word>_</word><poscode key="c5">RPR</poscode></pos>
                <pos><word>_</word><poscode key="c5">PRF</poscode></pos>
                <pos><word>_</word><poscode key="c5">AVP</poscode></pos>
        </or>
        <seq>
                <or>
                        <pos><word>_</word><poscode key="c5">RPR</poscode></pos>
                        <pos><word>_</word><poscode key="c5">PRF</poscode></pos>
                        <pos><word>_</word><poscode key="c5">AVP</poscode></pos>
                </or>
                        Include subscript: NP

        </seq>

</or>
```

appendix 6: CQL-script for PP detection



```
Metric                              Value
-------------------------------    --------
Total Files                              83
Total Lines                            9551
Avg Line Length                          27
Code Lines                             7056
Comment Lines                          1024
Whitespace Lines                       1517
Code/(Comment+Whitespace) Ratio        2,78
Code/Comment Ratio                     6,89
Code/Whitespace Ratio                  4,65
Code/Total Lines Ratio                 0,74
Code Lines Per File                      85
Comment Lines Per File                   12
Whitespace Lines Per File                18
```

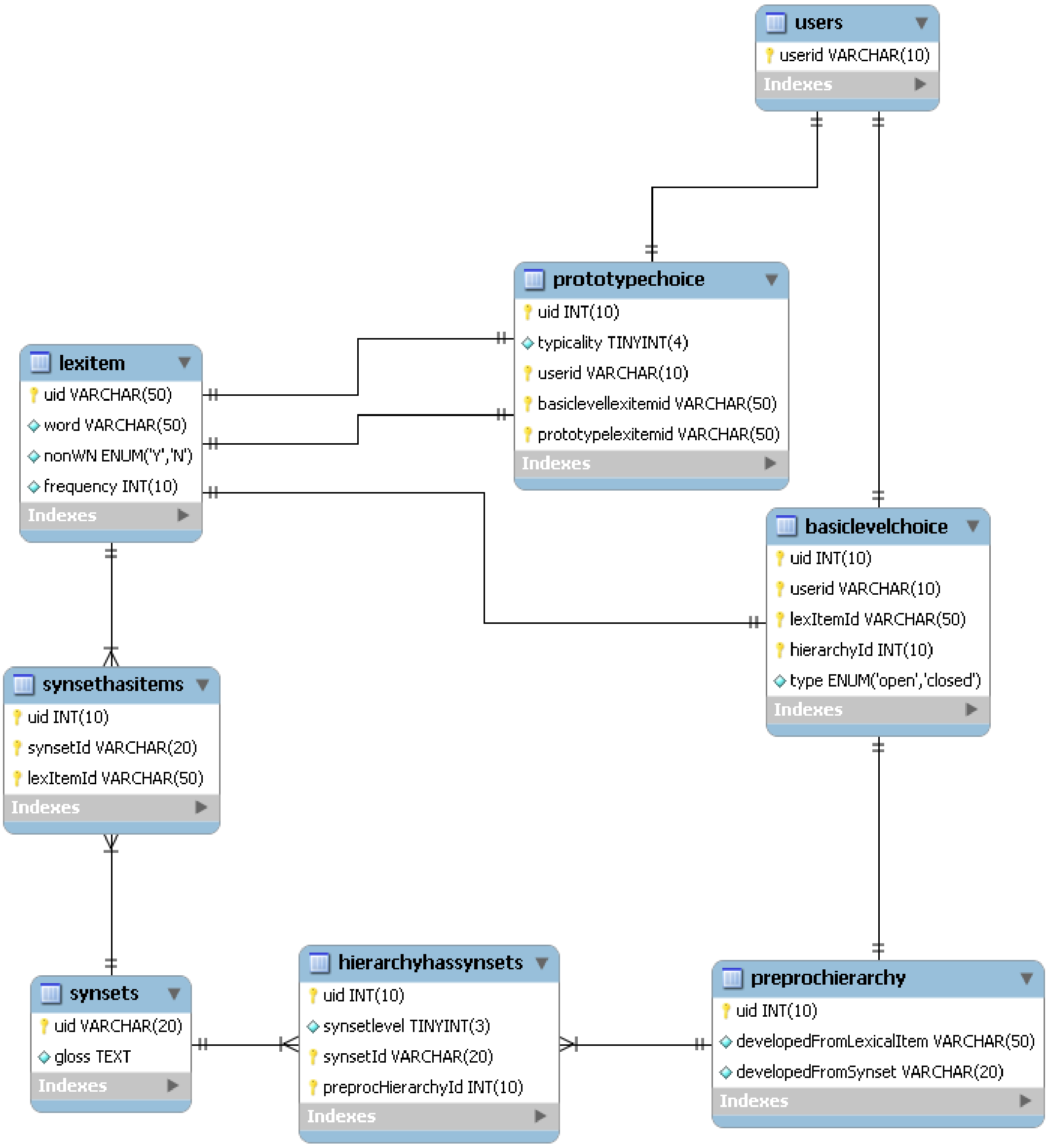
appendix 7: Program source code – statistics

The full source code of the software is available on the enclosed CD-ROM.

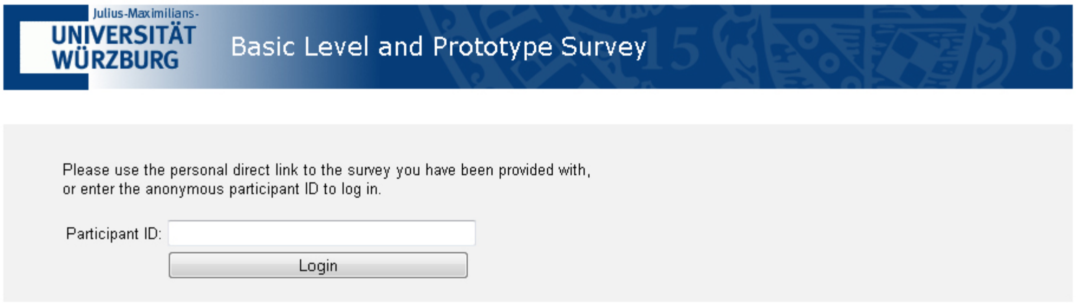
The CD contains a file named README.txt which holds further information about the content.

The software can also be directly accessed under the following URLs

http://survey.ferschke.de	(Leads to the survey application)
http://survey.ferschke.de/analysis.jsp	(Leads to the analysis tool)



appendix 8: Entity-Relationship model of the database

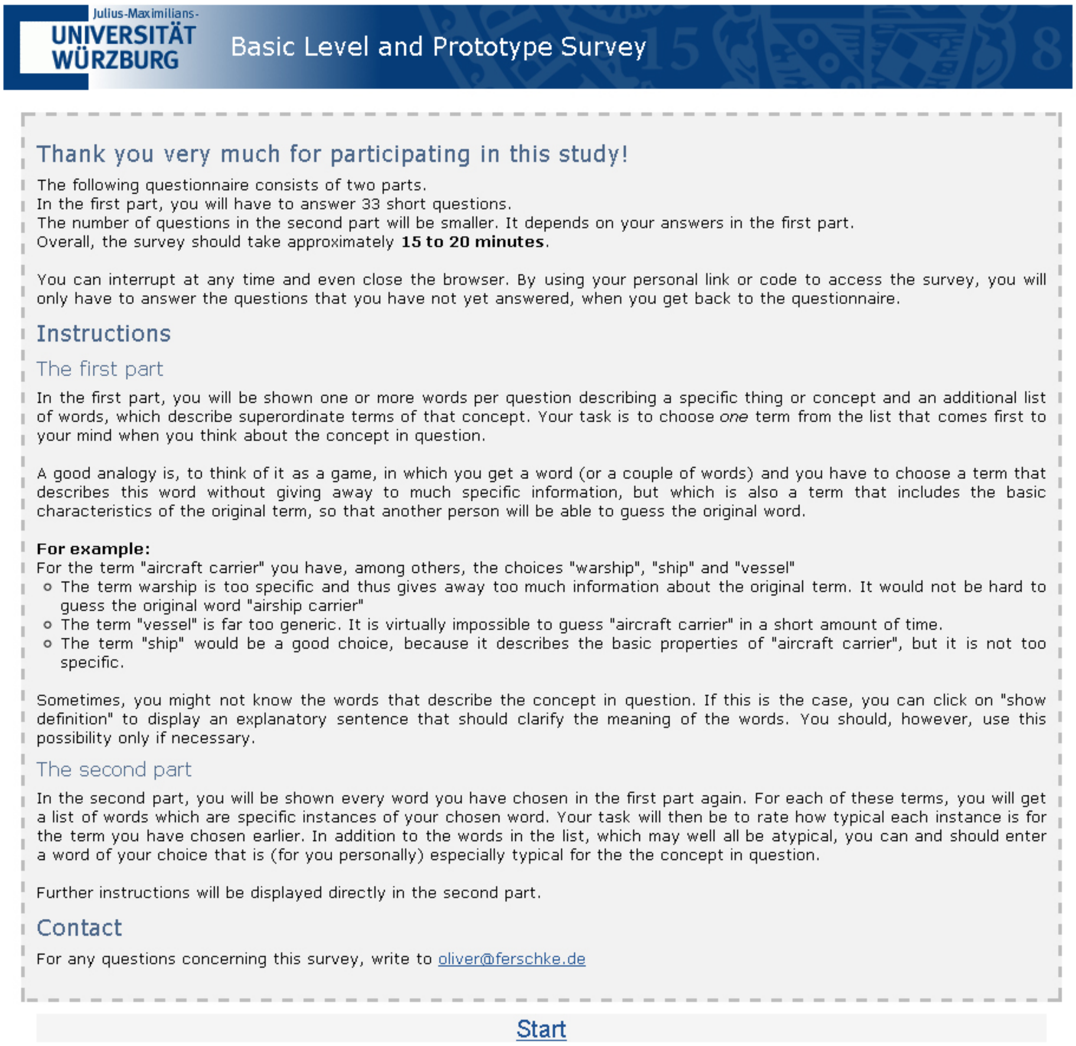


Basic Level and Prototype Survey

Please use the personal direct link to the survey you have been provided with,
or enter the anonymous participant ID to log in.

Participant ID:

appendix 9: Survey login form



Basic Level and Prototype Survey

Thank you very much for participating in this study!

The following questionnaire consists of two parts.
In the first part, you will have to answer 33 short questions.
The number of questions in the second part will be smaller. It depends on your answers in the first part.
Overall, the survey should take approximately **15 to 20 minutes**.

You can interrupt at any time and even close the browser. By using your personal link or code to access the survey, you will only have to answer the questions that you have not yet answered, when you get back to the questionnaire.

Instructions

The first part

In the first part, you will be shown one or more words per question describing a specific thing or concept and an additional list of words, which describe superordinate terms of that concept. Your task is to choose *one* term from the list that comes first to your mind when you think about the concept in question.

A good analogy is, to think of it as a game, in which you get a word (or a couple of words) and you have to choose a term that describes this word without giving away to much specific information, but which is also a term that includes the basic characteristics of the original term, so that another person will be able to guess the original word.

For example:
For the term "aircraft carrier" you have, among others, the choices "warship", "ship" and "vessel"

- The term warship is too specific and thus gives away too much information about the original term. It would not be hard to guess the original word "airship carrier"
- The term "vessel" is far too generic. It is virtually impossible to guess "aircraft carrier" in a short amount of time.
- The term "ship" would be a good choice, because it describes the basic properties of "aircraft carrier", but it is not too specific.

Sometimes, you might not know the words that describe the concept in question. If this is the case, you can click on "show definition" to display an explanatory sentence that should clarify the meaning of the words. You should, however, use this possibility only if necessary.

The second part

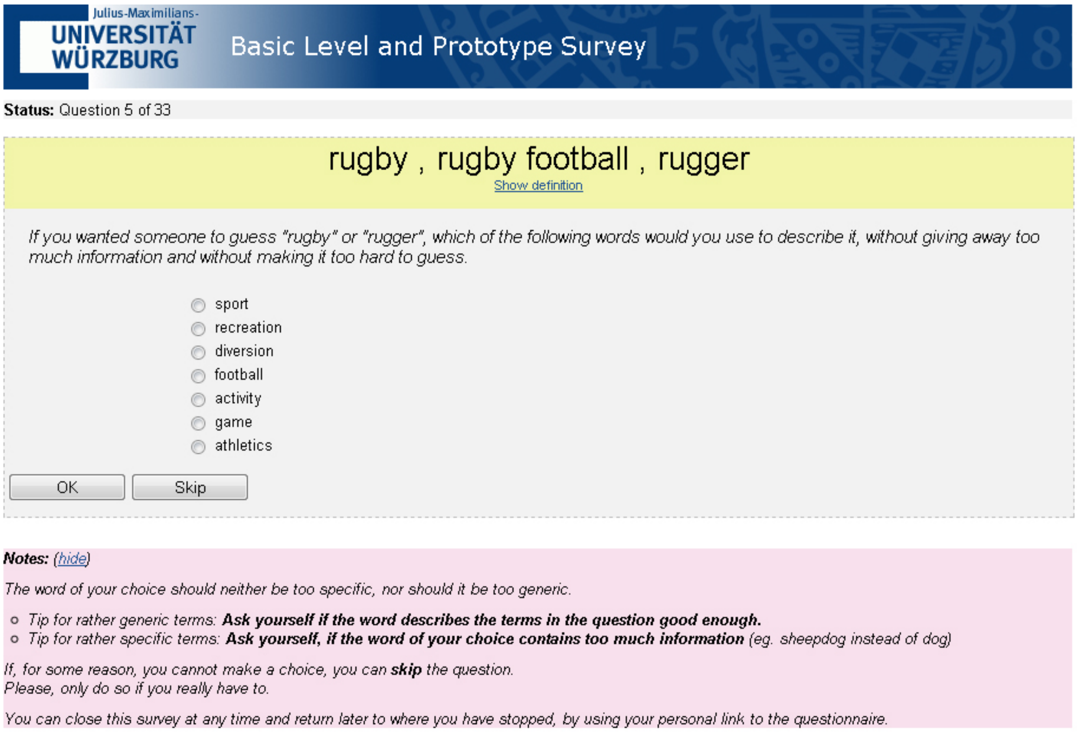
In the second part, you will be shown every word you have chosen in the first part again. For each of these terms, you will get a list of words which are specific instances of your chosen word. Your task will then be to rate how typical each instance is for the term you have chosen earlier. In addition to the words in the list, which may well all be atypical, you can and should enter a word of your choice that is (for you personally) especially typical for the the concept in question.

Further instructions will be displayed directly in the second part.

Contact

For any questions concerning this survey, write to oliver@ferschke.de

appendix 10: Introduction and instructions



Basic Level and Prototype Survey

Status: Question 5 of 33

rugby , rugby football , rugger

Show definition

If you wanted someone to guess "rugby" or "rugger", which of the following words would you use to describe it, without giving away too much information and without making it too hard to guess.

○ sport
○ recreation
○ diversion
○ football
○ activity
○ game
○ athletics

Notes: *(hide)*

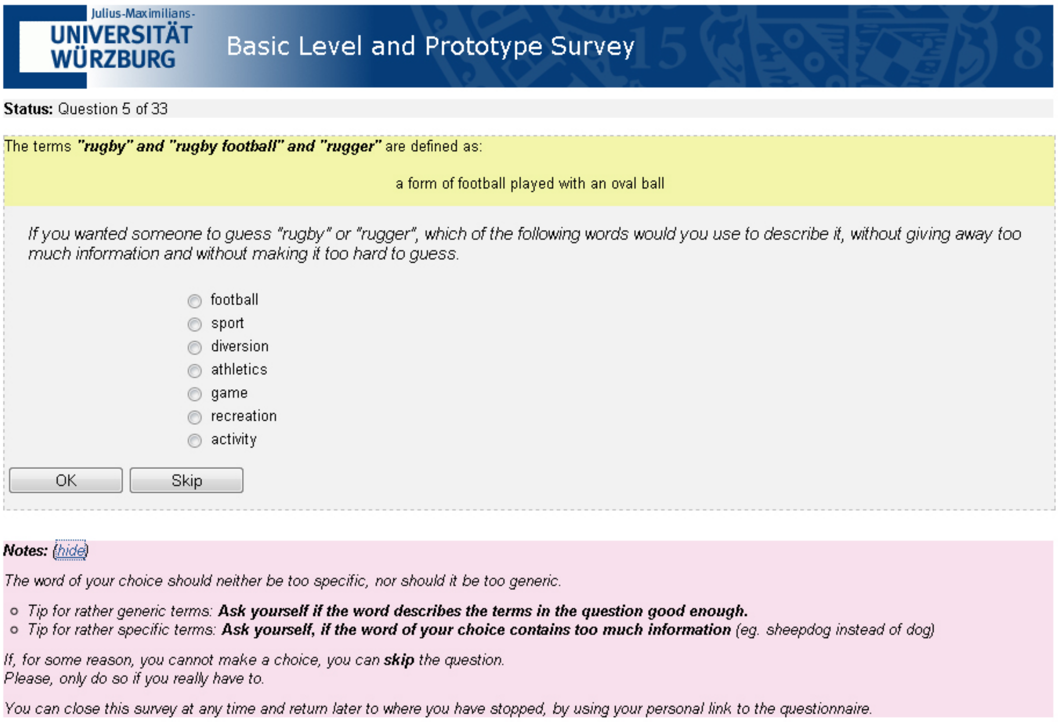
The word of your choice should neither be too specific, nor should it be too generic.

- *Tip for rather generic terms:* ***Ask yourself if the word describes the terms in the question good enough.***
- *Tip for rather specific terms:* ***Ask yourself, if the word of your choice contains too much information*** *(eg. sheepdog instead of dog)*

If, for some reason, you cannot make a choice, you can ***skip*** *the question.*
Please, only do so if you really have to.

You can close this survey at any time and return later to where you have stopped, by using your personal link to the questionnaire.

appendix 11: Basic level form for {rugby, rugby football, rugger}



Basic Level and Prototype Survey

Status: Question 5 of 33

The terms ***"rugby" and "rugby football" and "rugger"*** are defined as:

a form of football played with an oval ball

If you wanted someone to guess "rugby" or "rugger", which of the following words would you use to describe it, without giving away too much information and without making it too hard to guess.

○ football
○ sport
○ diversion
○ athletics
○ game
○ recreation
○ activity

Notes: *(hide)*

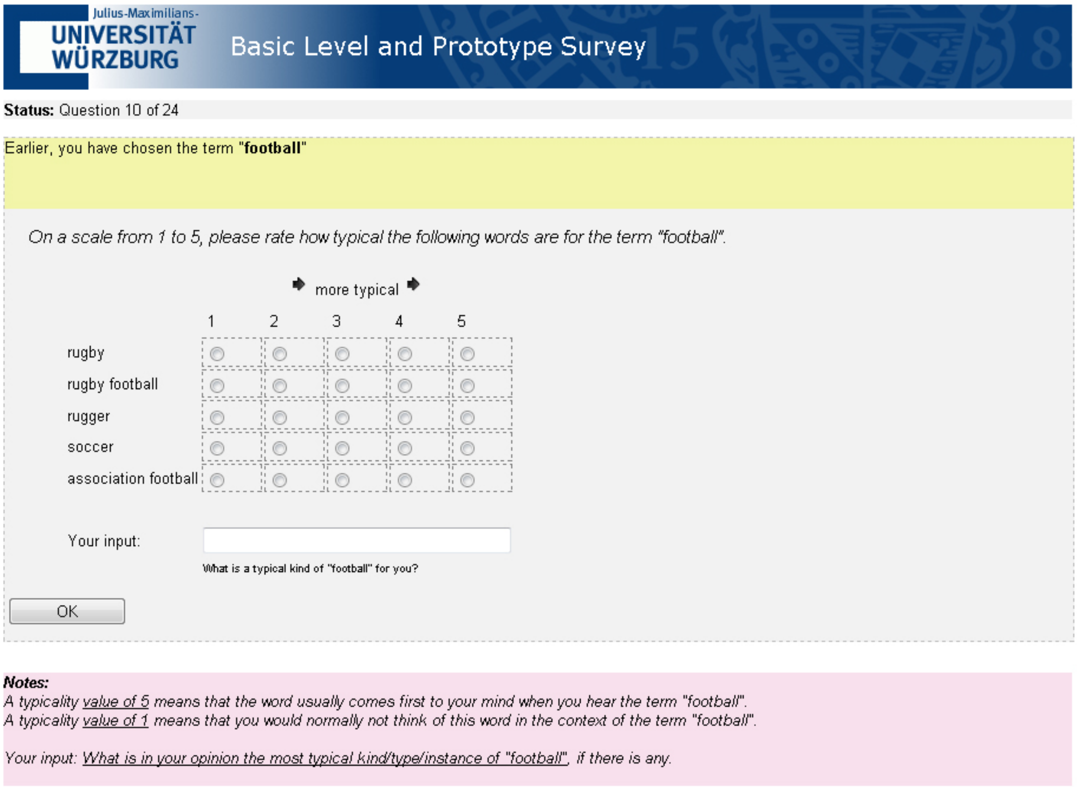
The word of your choice should neither be too specific, nor should it be too generic.

- *Tip for rather generic terms:* ***Ask yourself if the word describes the terms in the question good enough.***
- *Tip for rather specific terms:* ***Ask yourself, if the word of your choice contains too much information*** *(eg. sheepdog instead of dog)*

If, for some reason, you cannot make a choice, you can ***skip*** *the question.*
Please, only do so if you really have to.

You can close this survey at any time and return later to where you have stopped, by using your personal link to the questionnaire.

appendix 12: Basic level form for {rugby, rugby football, rugger} (with definition)



Julius-Maximilians-UNIVERSITÄT WÜRZBURG Basic Level and Prototype Survey

Status: Question 10 of 24

Earlier, you have chosen the term "**football**"

On a scale from 1 to 5, please rate how typical the following words are for the term "football".

	➧ more typical ➧				
	1	2	3	4	5
rugby	○	○	○	○	○
rugby football	○	○	○	○	○
rugger	○	○	○	○	○
soccer	○	○	○	○	○
association football	○	○	○	○	○

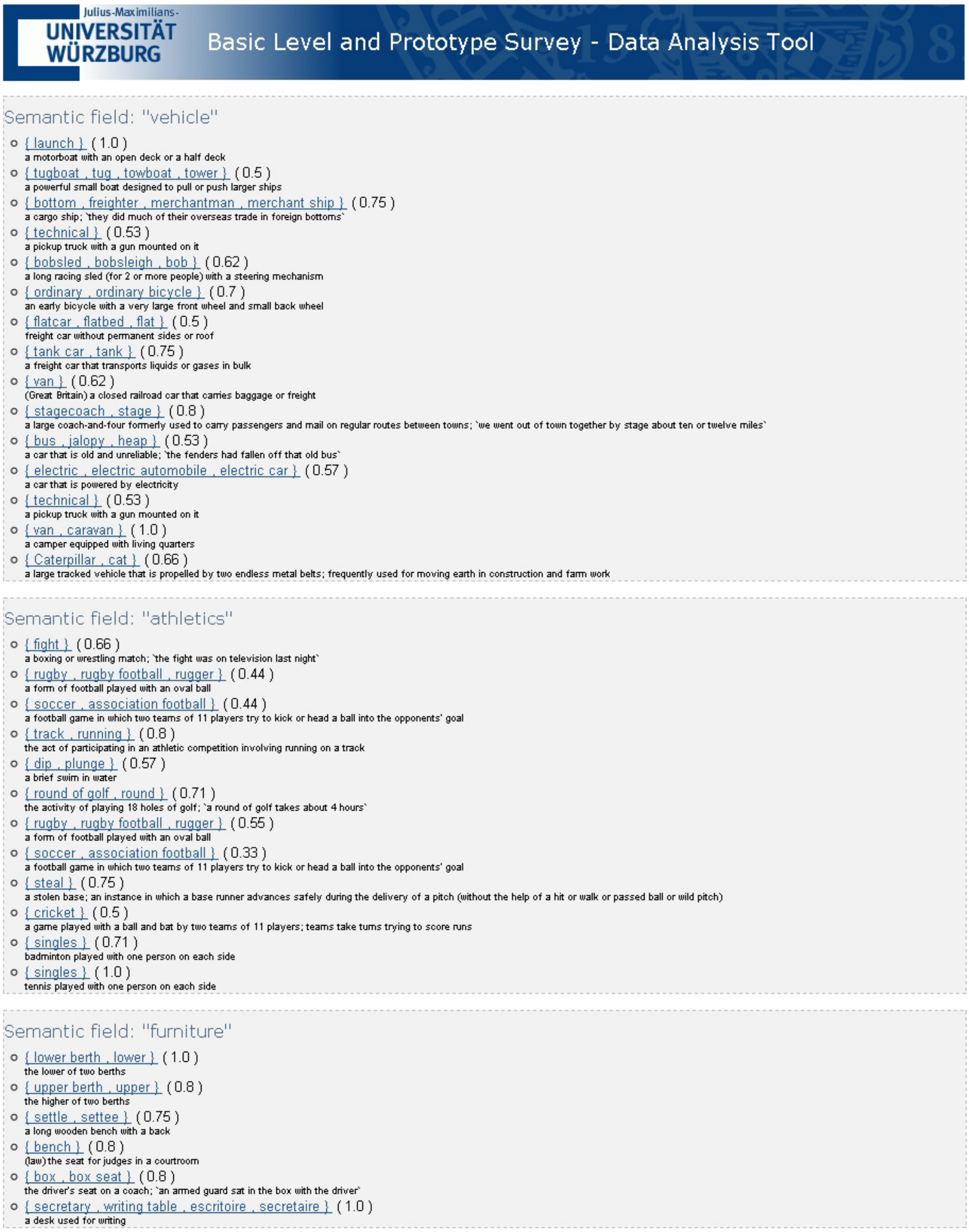
Your input:

What is a typical kind of "football" for you?

Notes:
A typicality value of 5 means that the word usually comes first to your mind when you hear the term "football".
A typicality value of 1 means that you would normally not think of this word in the context of the term "football".

Your input: What is in your opinion the most typical kind/type/instance of "football", if there is any.

appendix 13: Prototype form for the term "football"



Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Basic Level and Prototype Survey - Data Analysis Tool

Semantic field: "vehicle"

- { launch } (1.0)
  a motorboat with an open deck or a half deck
- { tugboat , tug , towboat , tower } (0.5)
  a powerful small boat designed to pull or push larger ships
- { bottom , freighter , merchantman , merchant ship } (0.75)
  a cargo ship; `they did much of their overseas trade in foreign bottoms`
- { technical } (0.53)
  a pickup truck with a gun mounted on it
- { bobsled , bobsleigh , bob } (0.62)
  a long racing sled (for 2 or more people) with a steering mechanism
- { ordinary , ordinary bicycle } (0.7)
  an early bicycle with a very large front wheel and small back wheel
- { flatcar , flatbed , flat } (0.5)
  freight car without permanent sides or roof
- { tank car , tank } (0.75)
  a freight car that transports liquids or gases in bulk
- { van } (0.62)
  (Great Britain) a closed railroad car that carries baggage or freight
- { stagecoach , stage } (0.8)
  a large coach-and-four formerly used to carry passengers and mail on regular routes between towns; `we went out of town together by stage about ten or twelve miles`
- { bus , jalopy , heap } (0.53)
  a car that is old and unreliable; `the fenders had fallen off that old bus`
- { electric , electric automobile , electric car } (0.57)
  a car that is powered by electricity
- { technical } (0.53)
  a pickup truck with a gun mounted on it
- { van , caravan } (1.0)
  a camper equipped with living quarters
- { Caterpillar , cat } (0.66)
  a large tracked vehicle that is propelled by two endless metal belts; frequently used for moving earth in construction and farm work

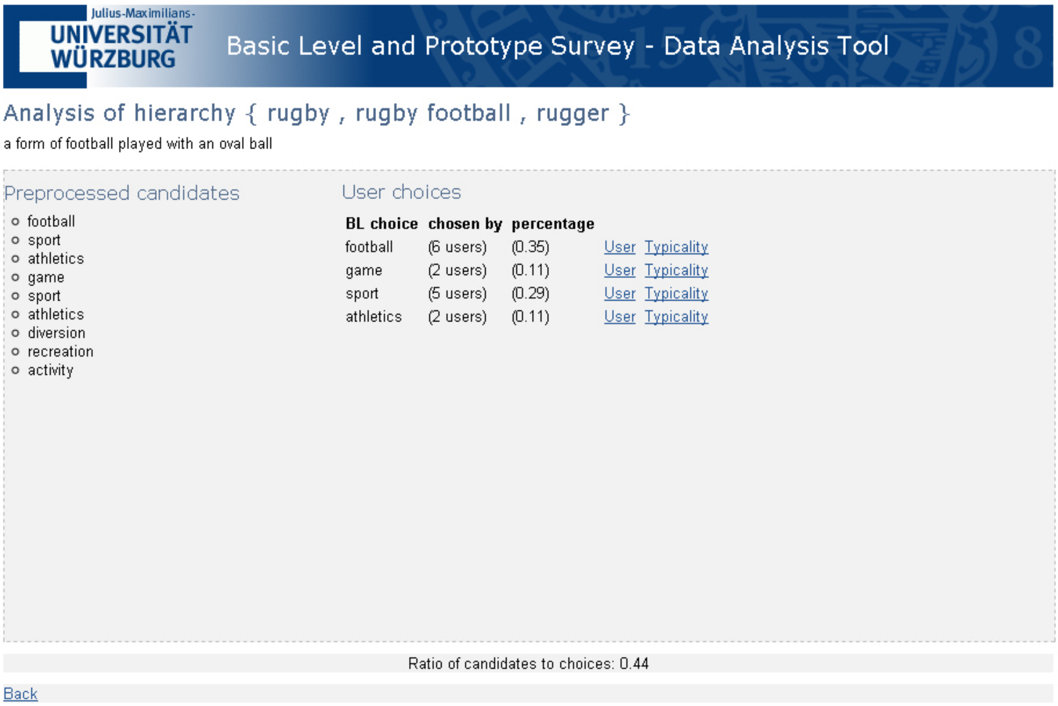
Semantic field: "athletics"

- { fight } (0.66)
  a boxing or wrestling match; `the fight was on television last night`
- { rugby , rugby football , rugger } (0.44)
  a form of football played with an oval ball
- { soccer , association football } (0.44)
  a football game in which two teams of 11 players try to kick or head a ball into the opponents' goal
- { track , running } (0.8)
  the act of participating in an athletic competition involving running on a track
- { dip , plunge } (0.57)
  a brief swim in water
- { round of golf , round } (0.71)
  the activity of playing 18 holes of golf; `a round of golf takes about 4 hours`
- { rugby , rugby football , rugger } (0.55)
  a form of football played with an oval ball
- { soccer , association football } (0.33)
  a football game in which two teams of 11 players try to kick or head a ball into the opponents' goal
- { steal } (0.75)
  a stolen base; an instance in which a base runner advances safely during the delivery of a pitch (without the help of a hit or walk or passed ball or wild pitch)
- { cricket } (0.5)
  a game played with a ball and bat by two teams of 11 players; teams take turns trying to score runs
- { singles } (0.71)
  badminton played with one person on each side
- { singles } (1.0)
  tennis played with one person on each side

Semantic field: "furniture"

- { lower berth , lower } (1.0)
  the lower of two berths
- { upper berth , upper } (0.8)
  the higher of two berths
- { settle , settee } (0.75)
  a long wooden bench with a back
- { bench } (0.8)
  (law) the seat for judges in a courtroom
- { box , box seat } (0.8)
  the driver's seat on a coach; `an armed guard sat in the box with the driver`
- { secretary , writing table , escritoire , secretaire } (1.0)
  a desk used for writing

appendix 14: Analysis tool - Overview



Basic Level and Prototype Survey - Data Analysis Tool

Analysis of hierarchy { rugby , rugby football , rugger }

a form of football played with an oval ball

Preprocessed candidates

- football
- sport
- athletics
- game
- sport
- athletics
- diversion
- recreation
- activity

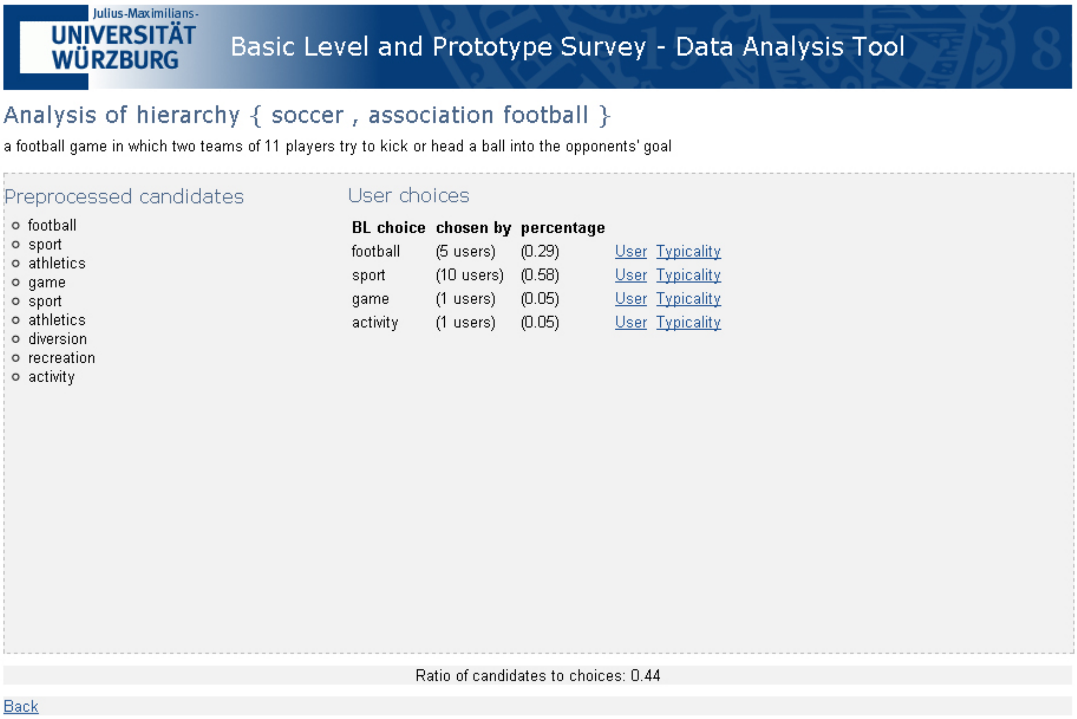
User choices

BL choice	chosen by	percentage	
football	(6 users)	(0.35)	User Typicality
game	(2 users)	(0.11)	User Typicality
sport	(5 users)	(0.29)	User Typicality
athletics	(2 users)	(0.11)	User Typicality

Ratio of candidates to choices: 0.44

Back

appendix 15: Analysis tool - Analysis of {rugby, rugby football, rugger}



Basic Level and Prototype Survey - Data Analysis Tool

Analysis of hierarchy { soccer , association football }

a football game in which two teams of 11 players try to kick or head a ball into the opponents' goal

Preprocessed candidates

- football
- sport
- athletics
- game
- sport
- athletics
- diversion
- recreation
- activity

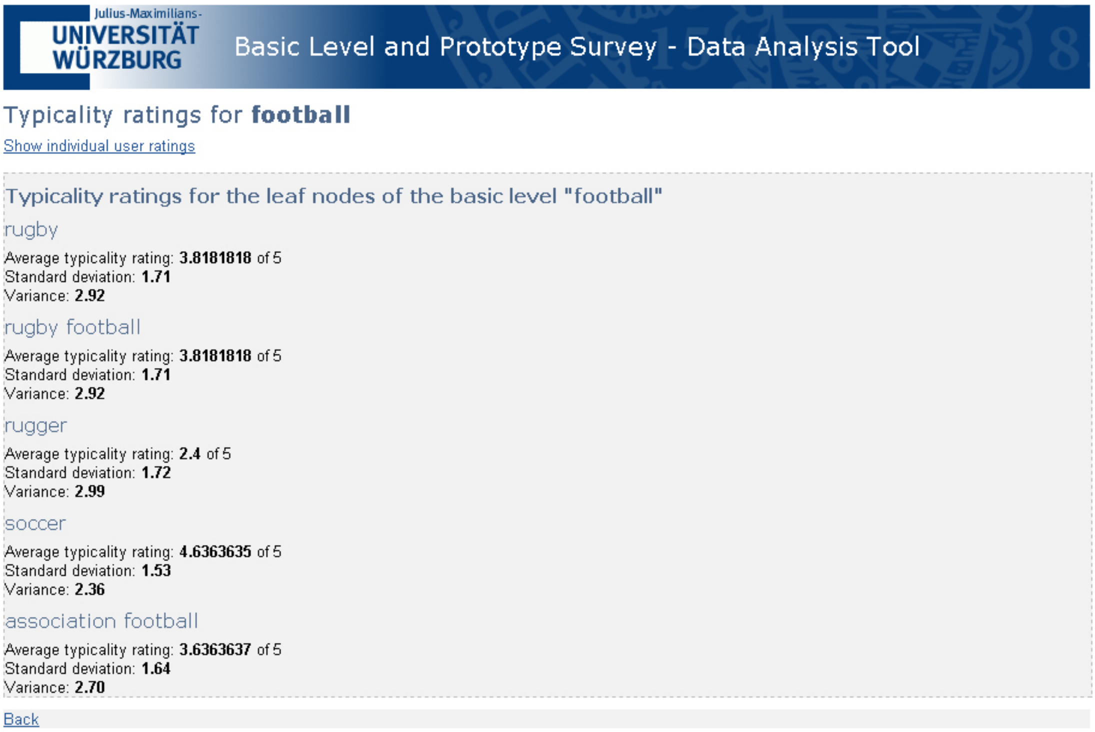
User choices

BL choice	chosen by	percentage	
football	(5 users)	(0.29)	User Typicality
sport	(10 users)	(0.58)	User Typicality
game	(1 users)	(0.05)	User Typicality
activity	(1 users)	(0.05)	User Typicality

Ratio of candidates to choices: 0.44

Back

appendix 16: Analysis tool - Analysis of {soccer, association football}



Typicality ratings for **football**

Show individual user ratings

Typicality ratings for the leaf nodes of the basic level "football"

rugby

Average typicality rating: **3.8181818** of 5
Standard deviation: **1.71**
Variance: **2.92**

rugby football

Average typicality rating: **3.8181818** of 5
Standard deviation: **1.71**
Variance: **2.92**

rugger

Average typicality rating: **2.4** of 5
Standard deviation: **1.72**
Variance: **2.99**

soccer

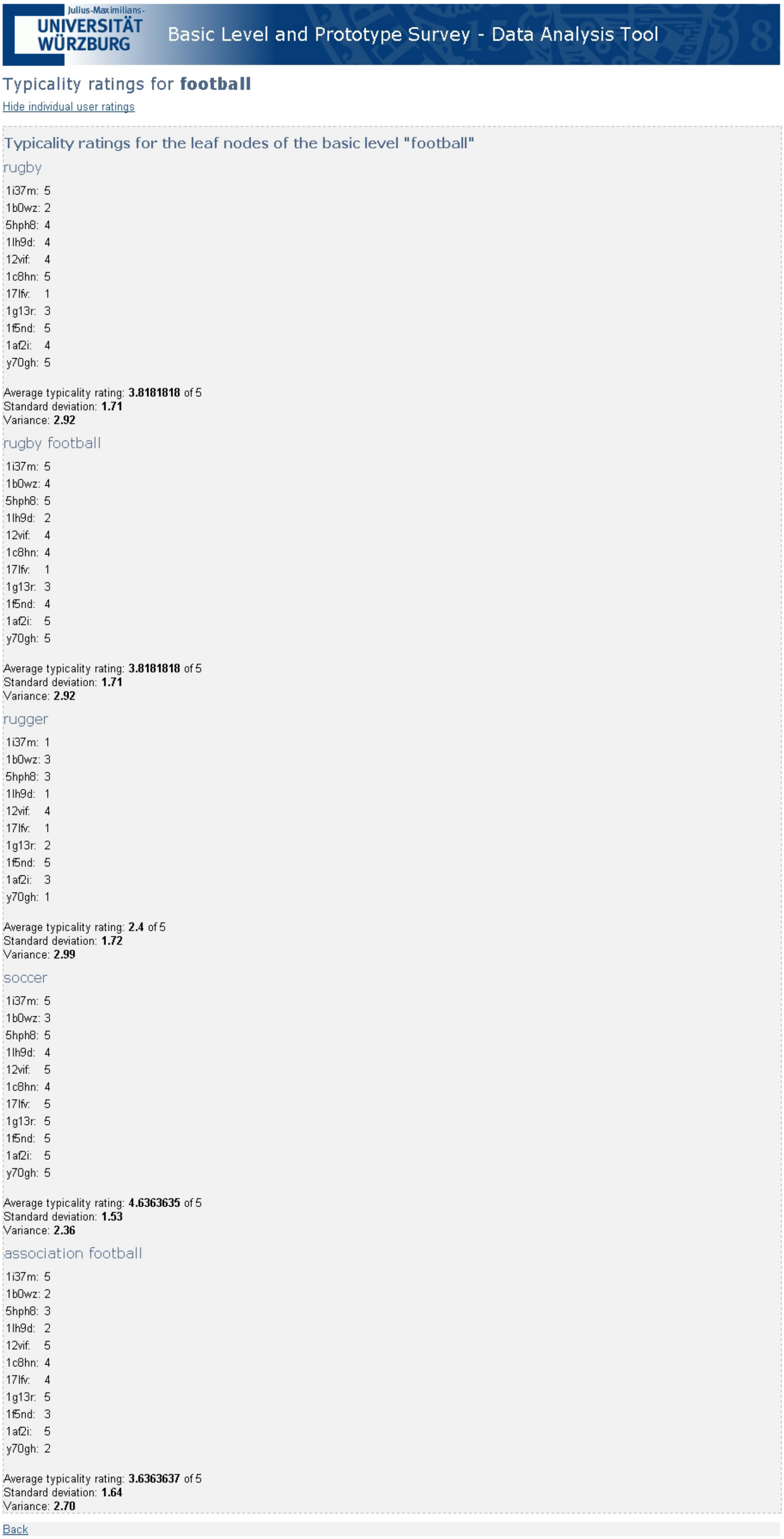
Average typicality rating: **4.6363635** of 5
Standard deviation: **1.53**
Variance: **2.36**

association football

Average typicality rating: **3.6363637** of 5
Standard deviation: **1.64**
Variance: **2.70**

Back

appendix 17: Analysis tool - Prototype analysis for the alleged basic level "football"



Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Basic Level and Prototype Survey - Data Analysis Tool

Typicality ratings for **football**

Hide individual user ratings

Typicality ratings for the leaf nodes of the basic level "football"

rugby

1i37m: 5

1b0wz: 2

5hph8: 4

1lh9d: 4

12vif: 4

1c8hn: 5

17lfv: 1

1g13r: 3

1f5nd: 5

1af2i: 4

y70gh: 5

Average typicality rating: **3.8181818** of 5
Standard deviation: **1.71**
Variance: **2.92**

rugby football

1i37m: 5

1b0wz: 4

5hph8: 5

1lh9d: 2

12vif: 4

1c8hn: 4

17lfv: 1

1g13r: 3

1f5nd: 4

1af2i: 5

y70gh: 5

Average typicality rating: **3.8181818** of 5
Standard deviation: **1.71**
Variance: **2.92**

rugger

1i37m: 1

1b0wz: 3

5hph8: 3

1lh9d: 1

12vif: 4

17lfv: 1

1g13r: 2

1f5nd: 5

1af2i: 3

y70gh: 1

Average typicality rating: **2.4** of 5
Standard deviation: **1.72**
Variance: **2.99**

soccer

1i37m: 5

1b0wz: 3

5hph8: 5

1lh9d: 4

12vif: 5

1c8hn: 4

17lfv: 5

1g13r: 5

1f5nd: 5

1af2i: 5

y70gh: 5

Average typicality rating: **4.6363635** of 5
Standard deviation: **1.53**
Variance: **2.36**

association football

1i37m: 5

1b0wz: 2

5hph8: 3

1lh9d: 2

12vif: 5

1c8hn: 4

17lfv: 4

1g13r: 5

1f5nd: 3

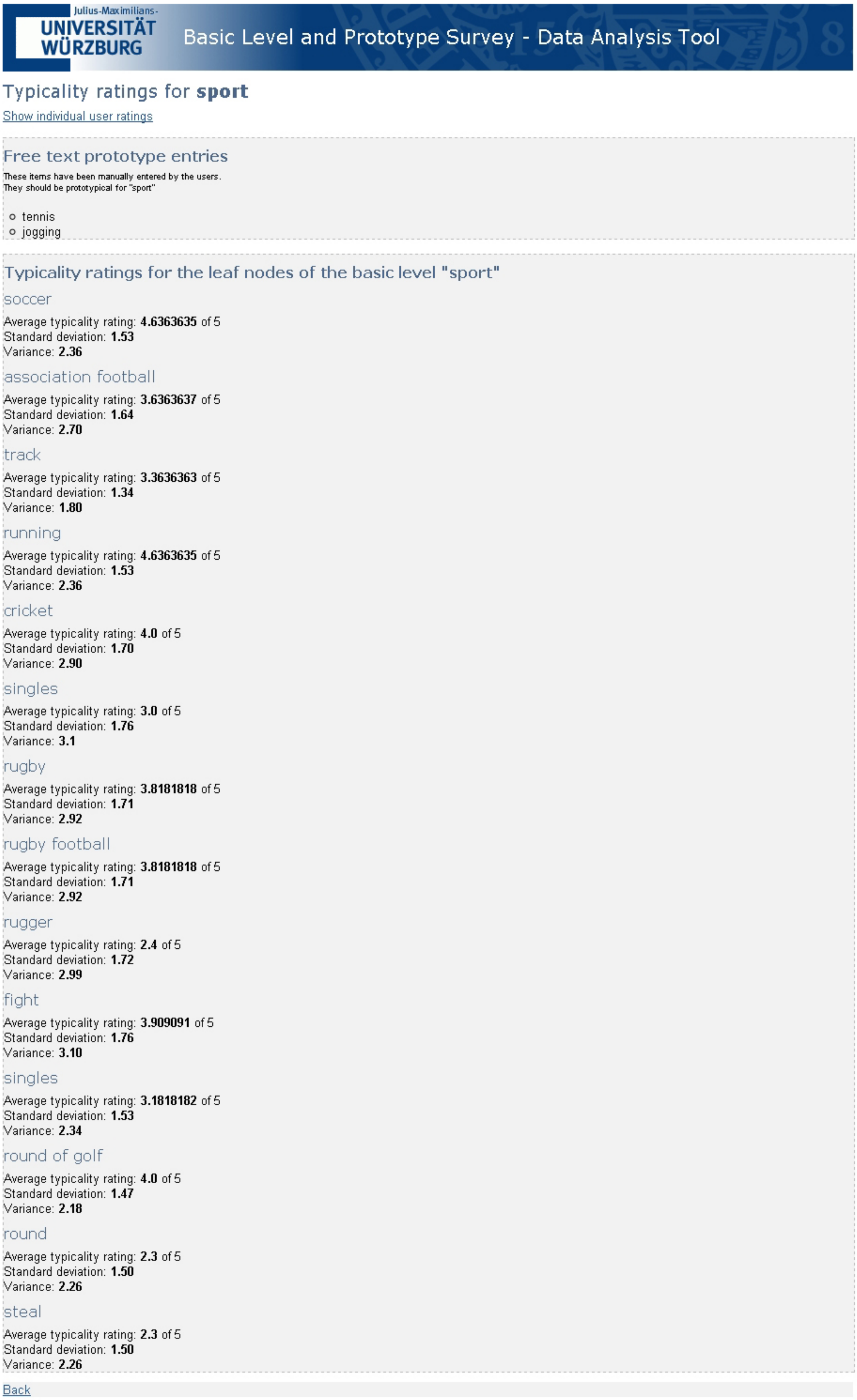
1af2i: 5

y70gh: 2

Average typicality rating: **3.6363637** of 5
Standard deviation: **1.64**
Variance: **2.70**

Back

appendix 18: Analysis tool - Prototype analysis for the alleged basic level "football" (detailed)



Basic Level and Prototype Survey - Data Analysis Tool

Typicality ratings for **sport**

Show individual user ratings

Free text prototype entries

These items have been manually entered by the users.
They should be prototypical for "sport"

- tennis
- jogging

Typicality ratings for the leaf nodes of the basic level "sport"

soccer

Average typicality rating: **4.6363635** of 5
Standard deviation: **1.53**
Variance: **2.36**

association football

Average typicality rating: **3.6363637** of 5
Standard deviation: **1.64**
Variance: **2.70**

track

Average typicality rating: **3.3636363** of 5
Standard deviation: **1.34**
Variance: **1.80**

running

Average typicality rating: **4.6363635** of 5
Standard deviation: **1.53**
Variance: **2.36**

cricket

Average typicality rating: **4.0** of 5
Standard deviation: **1.70**
Variance: **2.90**

singles

Average typicality rating: **3.0** of 5
Standard deviation: **1.76**
Variance: **3.1**

rugby

Average typicality rating: **3.8181818** of 5
Standard deviation: **1.71**
Variance: **2.92**

rugby football

Average typicality rating: **3.8181818** of 5
Standard deviation: **1.71**
Variance: **2.92**

rugger

Average typicality rating: **2.4** of 5
Standard deviation: **1.72**
Variance: **2.99**

fight

Average typicality rating: **3.909091** of 5
Standard deviation: **1.76**
Variance: **3.10**

singles

Average typicality rating: **3.1818182** of 5
Standard deviation: **1.53**
Variance: **2.34**

round of golf

Average typicality rating: **4.0** of 5
Standard deviation: **1.47**
Variance: **2.18**

round

Average typicality rating: **2.3** of 5
Standard deviation: **1.50**
Variance: **2.26**

steal

Average typicality rating: **2.3** of 5
Standard deviation: **1.50**
Variance: **2.26**

Back

appendix 19: Analysis tool - Prototype analysis for the alleged basic level "sport"

Analysis of: {fight}

Synset explanatory gloss: a boxing or wrestling match

Original leaf-root path	Basic level candidates	User choices	Chosen by	Percentage
boxing	boxing	**boxing**	8	0,47
pugilism				
fisticuffs	fisticuffs	**fisticuffs**	3	0,17
contact sport				
sport	sport	**sport**	2	0,11
athletics	athletics			
diversion	diversion	**diversion**	2	0,11
recreation	recreation			
activity				
act				
deed				
human action				
human activity				
event				
psychological feature				
abstraction				
abstract entity				
entity				

appendix 20: Analysis of {fight}

Analysis of: { track,running}

Synset explanatory gloss: the act of participating in an athletic competition involving running on a track

Original leaf-root path(s)	Basic level candidates	User choices	Chosen by	Percentage
track and field				
sport	sport	**sport**	3	0,17
athletics	athletics	**athletics**	8	0,47
diversion	diversion			
recreation	recreation	**recreation**	2	0,11
activity	activity	**activity**	2	0,11
act				
human action				
human activity				
event				
psychological feature				
abstraction				
abstract entity				
entity				

appendix 21: Analysis of {track, running}

Analysis of: { dip, plunge }
Synset explanatory gloss: a brief swim in water

Original leaf-root path(s)	Basic level candidates	User choices	Chosen by	Percent-age
swimming	swimming	**swimming**	**6**	**0,35**
swim	swim	**swim**	**5**	**0,29**
water sport				
aquatics	aquatics	**aquatics**	**2**	**0,11**
sport	sport			
athletics	athletics			
diversion	diversion	**diversion**	**2**	**0,11**
recreation	recreation			
activity				
act				
deed				
human action				
human activity				
event				
psychological feature				
abstraction				
abstract entity				
entity				

appendix 22: Analysis of {dip, plunge}

Analysis of:	{ golf, round of golf }
Synset explanatory gloss:	the activity of playing 18 holes of golf; `a round of golf takes about 4 hours`

Original leaf-root path(s)	**Basic level candidates**	**User choices**	**Chosen by**	**Percent-age**
golf	golf	**golf**	**5**	**0,29**
golf game				
outdoor game				
athletic game				
sport	sport	**sport**	**2**	**0,11**
athletics	athletics			
diversion	diversion			
recreation	recreation	**recreation**	**3**	**0,17**
activity				
act				
deed				
human action				
human activity				
event				
psychological feature				
abstraction				
abstract entity				
entity				
game	game	**game**	**5**	**0,29**
activity	activity	**activity**	**1**	**0,05**
act				
deed				
human action				
human activity				
event				
psychological feature				
abstraction				
abstract entity				
entity				

In this case of a tangled hierarchy, more than one leaf-root path are present (see vertical segmentation of the table)
The preprocessor automatically solves the problem during the processing of the BL candidates.

appendix 23: Analysis of {golf, round of golf}

Analysis of:	{ steal }
Synset explanatory gloss:	a stolen base; an instance in which a base runner advances safely during the delivery of a pitch(without the help of a hit or walk or passed ball or wild pitch)

Original leaf-root path(s)	Basic level candidates	User choices	Chosen by	Percentage
baseball	baseball	**baseball**	**13**	**0,76**
baseball game				
ball game				
ballgame				
field game				
outdoor game				
athletic game				
sport	sport	**sport**	**2**	**0,11**
athletics	athletics			
diversion				
recreation				
activity				
act				
deed				
human action				
human activity				
event				
psychological feature				
abstraction				
abstract entity				
entity				
game	game	**game**	**1**	**0,05**
activity				
act				
deed				
human action				
human activity				
event				
psychological feature				
abstraction				
abstract entity				
entity				

In this case of a tangled hierarchy, more than one leaf-root path are present (see vertical segmentation of the table)
The preprocessor automatically solves the problem during the processing of the BL candidates.

appendix 24: Analysis of {steal}

Analysis of: { singles }
Synset explanatory gloss: badminton played with one person on each side

Original leaf-root path(s)	Basic level candidates	User choices	Chosen by	Percent-age
badminton	badminton	**badminton**	**9**	**0,52**
court game				
athletic game				
sport	sport	**sport**	**3**	**0,17**
athletics	athletics	**athletics**	**1**	**0,05**
diversion	diversion			
recreation	recreation	**recreation**	**1**	**0,05**
activity				
act				
deed				
human action				
human activity				
event				
psychological feature				
abstraction				
abstract entity				
entity				
game	game	**game**	**2**	**0,11**
activity	activity			
act				
deed				
human action				
human activity				
event				
psychological feature				
abstraction				
abstract entity				
entity				

In this case of a tangled hierarchy, more than one leaf-root path are present (see vertical segmentation of the table)
The preprocessor automatically solves the problem during the processing of the BL candidates.

appendix 25: Analysis of {singles} (badminton)

Analysis of: { singles }

Synset explanatory gloss: tennis played with one person on each side

Original leaf-root path(s)	Basic level candidates	User choices	Chosen by	Percent-age
tennis	tennis	**tennis**	**8**	**0,47**
lawn tennis				
court game				
athletic game				
sport	sport	**sport**	**2**	**0,11**
athletics	athletics	**athletics**	**2**	**0,11**
diversion	diversion	**diversion**	**1**	**0,05**
recreation	recreation	**recreation**	**1**	**0,05**
activity				
act				
deed				
human action				
human activity				
event				
psychological feature				
abstraction				
abstract entity				
entity				
game	game	**game**	**2**	**0,11**
activity	activity	**activity**	**1**	**0,05**
act				
deed				
human action				
human activity				
event				
psychological feature				
abstraction				
abstract entity				
entity				

In this case of a tangled hierarchy, more than one leaf-root path are present (see vertical segmentation of the table)
The preprocessor automatically solves the problem during the processing of the BL candidates.

appendix 26: Analysis of {singles} (tennis)

Analysis of: { rugby, rugby football, rugger }

Synset explanatory gloss: a form of football played with an oval ball

Original leaf-root path(s)	**Basic level candidates**	**User choices**	**Chosen by**	**%**	**User choices (2)**	**chosen by (2)**	**% (2)**
football	football	**football**	**6**	**0,35**	**football**	**6**	**0,35**
football game							
field game							
outdoor game							
athletic game							
sport	sport	**sport**	**5**	**0,29**	**sport**	**7**	**0,41**
athletics	athletics	**athletics**	**2**	**0,11**	**athletics**	**1**	**0,05**
diversion							
recreation							
activity							
act							
deed							
human action							
human activity							
event							
psychological feature							
abstraction							
abstract entity							
entity							
game	game	**game**	**2**	**0,11**			
activity							
act							
deed							
human action							
human activity							
event							
psychological feature							
abstraction							
abstract entity							
entity							

⇩ *(to be continued on the next page)*

⇧ *(continuation of* { rugby, rugby football, rugger })

Original leaf-root path(s)	**Basic level candidates**	**User choices**	**Cho-sen by**	**%**	**User choices (2)**	**chosen by (2)**	**% (2)**
contact sport							
sport	sport	*sport*	5	*0,29*			
athletics	athletics	*athlet-ics*	2	*0,11*			
diversion	diversion						
recreation	recreation				**recrea-tion**	**1**	**0,05**
activity	activity				**activity**	**1**	**0,05**
act							
deed							
human action							
human activity							
event							
psychological feature							
abstraction							
abstract entity							
entity							

In this case of a tangled hierarchy, more than one leaf-root path are present (see vertical segmentation of the table)
The preprocessor automatically solves the problem during the processing of the BL candidates.

{ rugby, rugby football, rugger } appeared twice in the questionnaire for validation reasons. Therefore, two choices appear in the table.

appendix 27: Analysis of {rugby, rugby football, rugger}

Analysis of: { soccer, association football }

Synset explanatory gloss: a football game in which two teams of 11 players try to kick or head a ball into the opponents' goal

Original leaf-root path(s)	Basic level candidates	User choices	Chosen by	%	User choices (2)	Chosen by (2)	% (2)
football	football	**football**	**5**	**0,29**	**football**	**7**	**0,41**
football game							
field game							
outdoor game							
athletic game							
sport	sport	**sport**	**10**	**0,58**	**sport**	**7**	**0,41**
athletics	athletics						
diversion							
recreation							
activity					**activity**	**1**	**0,05**
act							
deed							
human action							
human activity							
event							
psychological feature							
abstraction							
abstract entity							
entity							
game	game	**game**	**1**	**0,05**			
activity							
act							
deed							
human action							
human activity							
event							
psychological feature							
abstraction							
abstract entity							
entity							

⇩ *(to be continued on the next page)*

⇧ *(continuation of* { soccer, association football })

Original leaf-root path(s)	**Basic level candidates**	**User choices**	**Chosen by**	**%**	**User choices (2)**	**Chosen by (2)**	**%(2)**
contact sport							
sport	sport	*sport*	*10*	*0,58*			
athletics	athletics						
diversion	diversion						
recreation	recreation						
activity	activity	**activ-ity**	**1**	**0,05**			
act							
deed							
human action							
human activ-ity							
event							
psychological feature							
abstraction							
abstract entity							
entity							

In this case of a tangled hierarchy, more than one leaf-root path are present
(see vertical segmentation of the table)
The preprocessor automatically solves the problem during the processing of the BL candidates.

*{ soccer, association football } appeared twice in the questionnaire for validation reasons.
Therefore, two choices appear in the table.*

appendix 28: Analysis of {soccer, association football}

Analysis of the examples from the sematic field "vehicle"

{launch} a motorboat with an open deck or a half deck

Basic level candidates	**BL choices**		**%**
motorboat	**motorboat**	**5**	**0,29**
powerboat	powerboat	2	0,11
boat	boat	4	0,23
vessel	vessel	2	0,11
craft	craft	1	0,05
vehicle	vehicle	1	0,05

{ tugboat, tug, towboat , tower } **a powerful small boat designed to pull or push larger ships**

boat	**boat**	**9**	**0,52**
vessel			
craft	craft	7	0,41
vehicle			
conveyance			
transport	transport	1	0,05

{ bottom , freighter , merchantman , merchant ship } **a cargo ship**

ship	**ship**	**9**	**0,52**
vessel	vessel	3	0,17
craft	craft	4	0,23
vehicle			

⇩ *(to be continued on the next page)*

{ technical }

a pickup truck with a gun mounted on it

pickup	pickup	1	0,05
truck	truck	3	0,17
vehicle			
container			
vehicle			
conveyance			
transport			
instrumentality	instrumentality	3	0,17
instrumentation	**instrumentation**	**4**	**0,23**
artifact	artifact	1	0,05
artefact			
whole			
unit	unit	2	0,11

{ bobsled , bobsleigh , bob }

a long racing sled (for 2 or more people) with a steering mechanism

sled	**sled**	**5**	**0,29**
sledge	sledge	2	0,11
sleigh	sleigh	3	0,17
vehicle	vehicle	5	0,29
conveyance			
transport	transport	1	0,05
instrumentality			
instrumentation			

{ ordinary , ordinary bicycle }

an early bicycle with a very large front wheel and small back wheel

bicycle	**bicycle**	**5**	**0,29**
bike	bike		
wheel	wheel	2	0,11
cycle	cycle	2	0,11
vehicle	vehicle	3	0,17
conveyance			
transport	transport	1	0,05
container			
instrumentality			
instrumentation			

⇩ *(to be continued on the next page)*

{ flatcar , flatbed , flat } **freight car without permanent sides or roof**

car	car	4	0,23
railcar			
vehicle	vehicle	3	0,17
conveyance	conveyance	2	0,11
transport	**transport**	**8**	**0,47**
container			
instrumentality			
instrumentation			

{ tank car , tank } **a freight car that transports liquids or gases in bulk**

car	car	2	0,11
railcar	railcar	1	0,05
vehicle	vehicle	2	0,11
conveyance	conveyance	2	0,11
transport	transport	3	0,17
container	**container**	**6**	**0,35**
instrumentality			
instrumentation			

(Hornby 2000) **a closed railroad car that carries baggage or freight**

car	car	2	0,11
railcar	railcar	3	0,17
vehicle	vehicle	1	0,05
conveyance	conveyance	2	0,11
transport	**transport**	**8**	**0,47**
container			
instrumentality			
instrumentation			

{ stagecoach , stage } **a large coach-and-four formerly used to carry passengers and mail on regular routes between towns**

coach	coach	5	0,29
carriage	**carriage**	**8**	**0,47**
rig	rig	1	0,05
vehicle	vehicle	2	0,11
container			

⇩ *(to be continued on the next page)*

{ bus , jalopy , heap } **a car that is old and unreliable**

car	car	3	0,17
auto	auto	2	0,11
automobile	automobile	3	0,17
machine			
motorcar	motorcar	1	0,05
vehicle	**vehicle**	**5**	**0,29**
container	container	1	0,05
saying	saying	1	0,05
expression			
locution			
speech			
language			
communication			

{ electric , electric automobile , electric car } **a car that is powered by electricity**

car	car	3	0,17
auto			
automobile	automobile	2	0,11
machine	**machine**	**7**	**0,41**
motorcar			
vehicle	vehicle	3	0,17
container			

{ van , caravan } **a camper equipped with living quarters**

camper	**camper**	**8**	**0,47**
RV	RV	2	0,11
vehicle	vehicle	4	0,23
container	container	1	0,05

{ Caterpillar , cat } **a large tracked vehicle that is propelled by two endless metal belts**

vehicle	**vehicle**	**10**	**0,58**
conveyance			
transport	transport	1	0,05
container			
instrumentality	instrumentality	3	0,17
instrumentation	instrumentation	2	0,11

appendix 29: Analysis of the examples from the sematic field "vehicle"

Analysis of the examples from the semantic field "furniture"

{ lower berth , lower } **the lower of two berths**

Basic level candidates	BL choices		%
berth	**berth**	**4**	**0,23**
bunk	**bunk**	**4**	**0,23**
bed	**bed**	**4**	**0,23**
furniture	furniture	3	0,17
furnishing			

{ upper berth , upper } **the higher of two berths**

berth	**berth**	**5**	**0,29**
bunk	**bunk**	**5**	**0,29**
bed	bed	3	0,17
furniture	furniture	3	0,17
furnishing			

{ settle , settee } **a long wooden bench with a back**

bench	**bench**	**6**	**0,35**
seat	seat	5	0,29
furniture	furniture	5	0,29
furnishing			

{ bench } **(law) the seat for judges in a courtroom**

seat	**seat**	**9**	**0,52**
furniture	furniture	5	0,29
furnishing	furnishing	2	0,11
instrumentality	instrumentality	1	0,05
instrumentation			

{ box , box seat } **the driver's seat on a coach**

seat	seat	5	0,29
furniture	**furniture**	**6**	**0,35**
furnishing	furnishing	3	0,17
instrumentality			
instrumentation	instrumentation	1	0,05

⇩ *(to be continued on the next page)*

{ secretary, writing table,
escritoire, secretaire } **a desk used for writing**

desk	**desk**	**8**	**0,47**
table	table	6	0,35
furniture	furniture	2	0,11
furnishing	furnishing	1	0,05

appendix 30: Analysis of the examples from the semantic field "furniture"

ERKLÄRUNG

Ich erkläre, dass das Thema dieser Arbeit nicht identisch ist mit dem Thema einer von mir bereits für ein anderes Examen eingereichten Arbeit. Ich erkläre weiterhin, dass ich die Arbeit nicht bereits an einer anderen Hochschule zur Erlangung eines akademischen Grades eingereicht habe.

Ich versichere, dass ich die Arbeit selbstständig verfasst und keine anderen als die angegebenen Grundlagen benutzt habe. Die Stellen der Arbeit, die anderen Werken dem Wortlaut oder dem Sinn nach entnommen sind, habe ich unter Angabe der Quellen der Entlehnung kenntlich gemacht. Dies gilt sinngemäß auch für gelieferte Zeichnungen, Skizzen und bildliche Darstellungen u.dgl.

KURZEGEFASSTER LEBENSLAUF

Oliver Ferschke
Randersackerer Straße 42
97072 Würzburg
Tel. 0931 – 20 54 215

Geboren: 10.10.1982 in Würzburg

AUSBILDUNG

Max-Dauthendey-Grundschule, Röntgen-Gymnasium Würzburg

1989 – 2002	Schulbildung Abschluss: Abitur

Universität Würzburg

10/2003 – 9/2004	Studium der Informatik (Diplom)
10/2004 – heute	Studium der Informatik und Anglistik Angestrebter Abschluss: Staatsexamen
04/2006 – heute	Erweiterndes Studium der Geographie Angestrebter Abschluss: Erweiterungsprüfung für das Staatsexamen
10/2007 – heute	Zweitstudium der Fächer Englische Linguistik, Englische Literaturwissenschaft und Informatik Angestrebter Abschluss: Magister Artium

PRAKTISCHE ERFAHRUNGEN

09/2002 – 07/2003	Zivildienst im Behindertenfahrdienst beim Malteser Hilfsdienst in Würzburg.
05/2004 – 06/2004	Praktikum im Computation Systemhaus GmbH Bad Mergentheim
04/2005 – 07/2005	Softwarepraktikum der Universität Würzburg: Entwurf einer Suchmaschine für den Bereich eGovernment der Stadt Würzburg
07/2005 – 10/2006	Dozent im Futurekids Computercenter: Entwicklung und Einführung eines Java-Programmierkurses, Leitung von EDV-Kursen für Kinder
2005	Selbstständige Tätigkeiten im Bereich Webdesign
10/2006 – heute	Studentischer Mitarbeiter im Rechenzentrum der Universität Würzburg: Eigenverantwortliche Durchführung von Softwareschulungen aus den Bereichen Projektmanagement, Groupware- und Kollaborationssoftware, Wissensmanagement. Außerdem Usersupport und Serveradministration.
06/2007 – heute	Studentischer Mitarbeiter am Lehrstuhl für angewandte Informatik und künstliche Intelligenz: Softwareentwicklung im Bereich e-Learning
09/2008 – heute	Aushilfslehrer am Gymnasium St. Ursula im Fach Informatik Leitung des Grundkurses in Informatik, Schuljahr 08/09

Würzburg, 22.01.2009