

Jörn Hurtienne

# Inter-coder reliability of categorising force-dynamic events in human-technology interaction

**Abstract:** Two studies are reported that investigate how readily accessible and applicable ten force-dynamic categories are to novices in describing short episodes of human-technology interaction (Study 1) and that establish a measure of inter-coder reliability when re-classifying these episodes into force-dynamic categories (Study 2). The results of the first study show that people can easily and confidently relate their experiences with technology to the definitions of force-dynamic events (e.g. “The driver released the handbrake” as an example of RESTRAINT REMOVAL). The results of the second study show moderate agreement between four expert coders across all ten force-dynamic categories (Cohen’s kappa = .59) when re-classifying these episodes. Agreement values for single force-dynamic categories ranged between ‘fair’ and ‘almost perfect’, i.e. between kappa = .30 and .95. Agreement with the originally intended classifications of study 1 was higher than the pure inter-coder reliabilities. Single coders achieved an average kappa of .71, indicating substantial agreement. Using more than one coder increased kappas to almost perfect: up to .87 for four coders. A qualitative analysis of the predicted versus the observed number of category confusions revealed that about half of the category disagreement could be predicted from strong overlaps in the definitions of force-dynamic categories. From the quantitative and qualitative results, guidelines are derived to aid the better training of coders in order to increase inter-coder reliability.

**Keywords:** inter-coder reliability, force dynamics, image schemas, human-technology interaction

---

**Jörn Hurtienne:** Julius-Maximilians-Universität Würzburg.  
E-mail: joern.hurtienne@uni-wuerzburg.de

## 1 Introduction

Cognitive content, like image schema categories or conceptual metaphors from linguistic data, is often extracted by a solitary coder (mostly the respective author

of the study). The implicit assumption is that the results are reliable, because other analysts of the same data would come to the same or similar conclusions as the linguistic intuitions of speakers of the same language would be the same or similar. The inherent danger of this assumption is, however, that if it is not correct, the extracted cognitive content is not reliable, i.e. independent coders produce disagreeing results and conclusions drawn from the data are of questionable validity.<sup>1</sup>

Because reliability data for image-schema categorisations were not available in the literature, we conducted two studies to investigate the inter-coder agreement for image-schema categorisations of force-dynamic events. Force-dynamics in language has been studied by Talmy (1988), who develops definitions and a graphical notation scheme for different force-dynamic events. A similar approach is taken by Johnson (1987), who describes a number of FORCE image schemas using a more informal style of notation. From these two sources, ten force-dynamic categories (image schemas) can be derived: ATTRACTION/REPULSION, BALANCE, BLOCKAGE, COMPULSION, COUNTERFORCE, DIVERSION, ENABLEMENT, MOMENTUM, RESISTANCE, RESTRAINT REMOVAL.

To establish a comparison standard, in the first study participants received the definitions and notations of these ten FORCE image schemas. In a brainwriting exercise they produced one-sentence episodes meant to illustrate these relations in a context of human-technology interaction (e.g. *The driver released the hand-brake* as an example of RESTRAINT REMOVAL). In the second study, these episodes were given to four coders familiar with FORCE image schemas. Their task was to assign the episodes to the FORCE image schemas they found most suitable.

Three purposes are served by these studies: (1) to investigate how readily accessible and applicable these force-dynamic categories are to novices who used them to describe force-dynamic events in human-technology interaction (Study 1) and (2) to establish a measure of inter-coder reliability of categorising short descriptions of force-dynamic events (Study 2). From the patterns of results it is (3) possible to derive measures that help to increase inter-coder reliability. For example, the definitions of force-dynamic categories sometimes overlap so that predictions can be made what categories are likely to be confused with each

---

<sup>1</sup> For example, image schemas derived from the language of people using interactive electronic products have been used to understand the subconscious mental models of these users. On this basis new user interfaces were designed that were more intuitive to use than previous versions (see for example the redesign of an invoice verification and posting system, Hurtienne, Weber and Blessing 2008; Hurtienne, Israel and Weber 2008). To be successful in design, image-schema categorisations need to be reliable. Basing user interface design decisions on unreliable data would lead to suboptimal results.

other. Also, because the second study was conducted with several coders, it is possible to estimate the reliability of using one or many coders and thus determine the optimum number of coders required for such analyses.

## 2 Force-dynamic events

This study is part of a larger endeavour to study the inter-coder reliability of coding image-schema instantiations in language, behaviour, and technology (cf. Hurtienne 2011: Ch. 7). Force-dynamic events are of particular interest, because FORCE image schemas seem harder to detect and classify than, for example, SPACE image schemas (e.g. NEAR-FAR, UP-DOWN, CENTRE-PERIPHERY), ATTRIBUTE image schemas (e.g. BIG-SMALL, BRIGHT-DARK, WARM-COLD), or CONTAINMENT image schemas that present themselves more readily in language corpora or graphical representations. Therefore, agreement between coders should be high for these groups of image schemas. SPACE, ATTRIBUTE and CONTAINMENT image schemas are often instantiated by static entities. FORCE image schemas like COMPULSION, MOMENTUM, and DIVERSION, in contrast, are instantiated by the more transient dynamics of two or more interacting forces that may be more difficult to detect and agree upon.

Furthermore, FORCE image schemas can be instantiated in both physical (e.g. blocking the movement of a lever) and abstract ways (e.g. blocking an unauthorised user from accessing a website). The rationale therefore is, that if the inter-coder reliability within this challenging group of image schemas is acceptable, it probably is for other groups of image schemas as well.

Definitions and notations of the ten image schemas used are given in Table 1. The graphic notations are based on Talmy's (1988, 2000) notational system of FORCE image schemas and visualisations by Johnson (1987). In Talmy's (1988) system of "force-dynamics" there is always an Agonist and a stronger or weaker Antagonist. Agonists have either an intrinsic tendency toward rest or toward motion. In Talmy's notational system the Agonist is indicated by a circle and the Antagonist by a concave form (Figure 1). Further, the Agonist's intrinsic force tendency, the resultant of the force interaction, and whether an entity is stronger or weaker than the other is coded in the notation. In the example of the notation of the BLOCKAGE image schema, for example, the Agonist's tendency is towards action, but it is held back by a stronger Antagonist so that the Agonist is kept in place.

**Table 1:** Definitions, notations, and examples of FORCE image schemas

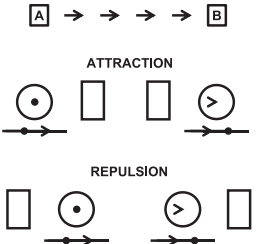

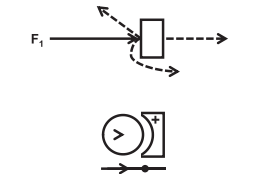
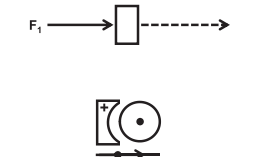
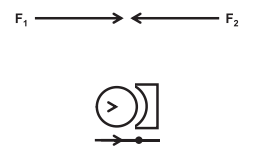
Image Schema	Definition	Notation (after Johnson 1987; Talmy 1988)	Technology Example
ATTRACTION / REPULSION	<p>A FORCE image schema in which a (passive) object exerts a force on another object, either physically or metaphorically, to pull it toward itself (or in the case of REPULSION to repel it), mostly acting from a distance.</p>		<p>If the seatbelt is not fastened in the car, then a beeping sound is activated to alert the driver. (ATTRACTION)</p>
BALANCE	<p>A FORCE image schema that provides an understanding of physical or metaphorical counteracting forces: forces and/or weights counteract/balance off one another. Metaphorically, there is equilibrium, not <i>too much</i> and not <i>not enough</i>.</p>		<p>On both sides of the monitor screen of the cash machine there is the same number of equally sized push buttons. The symmetrical arrangement suggests similar functions of the buttons.</p>
BLOCKAGE	<p>A FORCE image schema in which a force/movement is physically or metaphorically stopped or redirected by an obstacle.</p>		<p>The car driver pulls on the handbrake to prevent inadvertent rolling.</p>
COMPULSION	<p>A FORCE image schema that involves an external force physically or metaphorically causing some passive entity to move.</p>		<p>The car driver steps on the accelerator and the car accelerates.</p>
COUNTERFORCE	<p>A FORCE image schema that involves the active meeting of physically or metaphorically opposing forces that are equally strong. Both forces collide; there is no further movement.</p>		<p>The plane pilot wants to descend, the autopilot to ascend. Both struggle against each other. The plane neither descends nor ascends.</p>

Table 1 (cont.)

Image Schema	Definition	Notation (after Johnson 1987; Talmy 1988)	Technology Example
DIVERSION	A FORCE image schema that involves forces that physically or metaphorically meet and produce a change in direction or force vectors (at least one).		A user is checking her email and finds an interesting link to a website. She follows the link and thereby loses sight of her actual work.
ENABLEMENT	A FORCE image schema that involves having (a) the physical or metaphorical power to perform some act, or a potential force vector and the absence of BLOCKAGE, RESISTANCE, COUNTERFORCE, or COMPULSION; (b) a felt sense of power to perform some action		When the car is taking a bend, the cornering light will actively light into the bend where the driver needs to look. (active ENABLEMENT)
MOMENTUM	A FORCE image schema that involves the tendency of an object to maintain the actual state of motion (or rest) if there is no influence of another agent.		The progress indicator of an mp3-player is moving as long as the song is playing or until it is stopped.
RESISTANCE	A FORCE image schema that involves a force that tends to oppose or retard the motion of another entity.		The shutter release button of a digital camera has a soft stop that triggers the autofocus. When the user presses the button harder, the shutter is finally released.
RESTRAINT REMOVAL	A FORCE image schema that involves the physical or metaphorical removal of a barrier to the action of a force, or absence of a barrier that was potentially present.		The car driver releases the handbrake to move off.

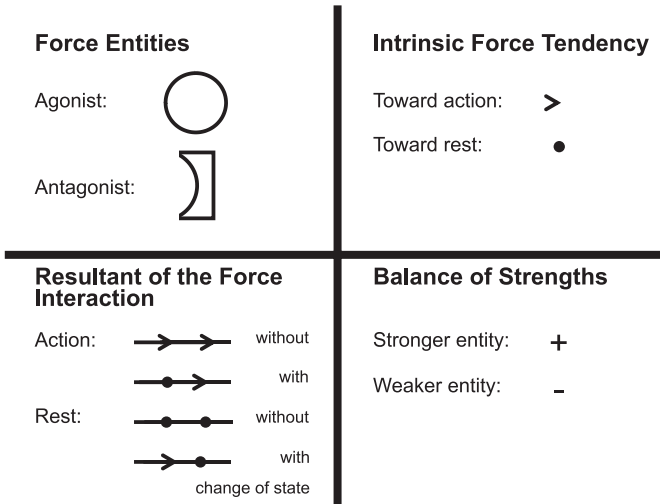


Fig. 1: Elements of Talmy's (1988) notation of force-dynamic elements

### 3 Generation of force-dynamic events (Study 1)

Study 1 was conducted to generate one-sentence episodes of human-technology interaction that could later be used as material for image-schema categorisations. The study had the sub-goal of exploring whether image-schema definitions can easily be understood and whether people are able to relate them to their own experiences, in this case with technology. In a workshop, participants were first introduced to definitions and examples of the ten FORCE image schemas. Then, they were asked to brainstorm examples from their experience with using technology that matched each of these image schemas.

#### 3.1 Method

Eleven researchers from the Engineering Design and Methodology group at TU Berlin took part in this study. Most of the participants were mechanical engineers and had no prior experience with cognitive linguistics or FORCE image schemas, but they were familiar with the specific method of brainwriting used in this study.

The workshop started with 30 minutes of presentation plus 10 minutes of discussion. During the presentation, participants received general information about what image schemas are and how they relate to using technology. Then,

the ten FORCE image schemas were introduced one by one, along with their notations, metaphorical uses, and a discussion of technology examples.

After the presentation, each participant received one questionnaire sheet. Each sheet stated the name, the notation, and the definition of one FORCE image schema (cf. Table 1). Different sheets contained different FORCE image schemas. Participants were instructed to write down examples from their experience with technology that they regarded as instances of the specific image schema on the questionnaire sheet before them. For each example they indicated the direction of the effect, i.e. whether the user influences the technology, or the technology influences the user. They also indicated how confident they were that their example was a proper instance of the specific image schema (coded from 1 = very uncertain to 5 = very certain). To facilitate brainwriting, one example was already printed on the sheet. For instance, the example on the RESTRAINT REMOVAL sheet was “The driver releases the handbrake to move off.”

Participants were instructed to fill in as many examples as they could think of. After two minutes, they had to pass their sheets to their neighbours on the right. They then immediately started working on the sheets they received from their neighbours on the left. After ten of these two-minute cycles, each participant had had the chance to produce examples for all ten image schemas in sequence.

## 3.2 Results and discussion

During the 20-minute brainwriting session the participants produced 146 usage examples in total. This averages to 14.6 examples per image schema and 13.3 examples per participant. Participants were, on average, quite confident in relating their examples to the image schema descriptions ( $M = 3.74$ ,  $SD = 1.19$ ). The highest confidence ratings were given for examples of the image schemas COMPULSION, BLOCKAGE, and ENABLEMENT, the lowest for COUNTERFORCE, DIVERSSION, and MOMENTUM (see Table 2).

Summarising the results it seems that 30 minutes of instruction were enough to induce sufficient confidence in the participants to relate technology experiences to the descriptions of ten FORCE image schemas. Thus, it seems that these force-dynamic categories are readily accessible and applicable even by relative novices. This result is expected, however, if we assume that FORCE image schemas are central to the understanding of force-dynamic events in the world.

The many examples gained were re-used in the second study in which people familiar with image schemas re-assigned image-schema categories to these examples.

**Table 2:** Number of generated usage scenarios and confidence scores in Study 1

Image Schema	Number of usage scenarios generated	Confidence Scores	
		<i>M</i>	<i>SD</i>
ATTRACTION/REPULSION	24	3.83	1.27
ENABLEMENT	18	4.13	1.02
BLOCKAGE	15	4.15	1.46
COMPULSION	14	4.36	1.28
DIVERSION	14	3.29	0.73
MOMENTUM	13	3.31	1.11
RESTRAINT REMOVAL	13	3.62	0.96
RESISTANCE	12	3.75	1.29
COUNTERFORCE	12	3.17	1.27
BALANCE	11	3.60	0.97
Total	146	3.74	1.19

*Note:* Confidence ratings ranged from 1 = very uncertain to 5 = very certain. The number of generated usage scenarios is higher for ATTRACTION/REPULSION because two questionnaire sheets were circulated for this image schema (one each for the other image schemas).

## 4 Image-schema classification (Study 2)

The second study aimed at determining the agreement among people familiar with FORCE image schemas who re-classified the usage examples generated in the first study. This procedure allowed estimating the inter-coder reliabilities of image-schema classifications, both among the coders and as compared to the original classification obtained in Study 1.

As discussed above, we expected force-dynamic events to be hard to assign to image-schema categories. The definitions of FORCE image schemas are not exhausting, not mutually exclusive, not evenly sized, and are not evenly distributed. They are discrete, describe a few ‘typical’ scenarios, overlap in meaning, and are of different scope. FORCE image schemas are not so much about capturing the reality of physical mechanics. They are about the *phenomenology* of physical events – and as such not very systematic. The same physical force can, according to circumstances, be interpreted in different ways, e.g. as RESISTANCE or as BLOCKAGE. This freedom in interpretation could make it difficult to find reliable agreement between coders. More specifically, from the definitions of FORCE image schemas, some confusion between image schema categories should be more likely than others – depending on the similarity or degree of overlap of categories (Table 3).



**Table 3:** Predicted confusions between FORCE image schemas

Image schemas	Proposed similarities
1 MOMENTUM × ENABLEMENT	MOMENTUM focuses on ongoing motion states and is associated to a force-dynamic state of ENABLEMENT. While MOMENTUM focuses on the motion of the Agonist, ENABLEMENT focuses on the absence of Antagonists. ENABLEMENT is a necessary but not sufficient condition for MOMENTUM – a very subtle distinction that could be easily confused.
2 BLOCKAGE × COMPULSION × ATTRACTION/REPULSION	BLOCKAGE and COMPULSION are similar in that they both cause changes in the motion state of the Agonist. In BLOCKAGE change is from motion to rest, in COMPULSION from rest to motion. ATTRACTION/REPULSION is similar to COMPULSION and BLOCKAGE in that the Antagonist can start and stop the motion of the Agonist. ATTRACTION/REPULSION, however, sometimes only affects the direction of the motion PATH.
3 BLOCKAGE × COUNTERFORCE × RESISTANCE	BLOCKAGE, COUNTERFORCE, and RESISTANCE are very similar in that they describe situations in which an Antagonist hinders the movement of the Agonist. The differences lie in the distribution of strength between the Agonist and the Antagonist. In BLOCKAGE, the Antagonist is stronger than the Agonist and movement will cease. In COUNTERFORCE, movement will also cease, but the Antagonist and the Agonist are equally strong. In RESISTANCE, the Antagonist is weaker than the Agonist, so that the motion of the Agonist will be hindered, but will not stop. As it is not always possible to assess the relative strengths of the Agonist and the Antagonists, confusion of the three categories is expected.
4 COUNTERFORCE × BALANCE	As Agonist and Antagonist of the dynamic COUNTERFORCE image schema are equally strong, this could lead to confusion with the non-dynamic BALANCE image schema.
5 RESTRAINT REMOVAL × (BLOCKAGE, COUNTERFORCE, RESISTANCE)	A BLOCKAGE, RESISTANCE, or COUNTERFORCE is the precondition of RESTRAINT REMOVAL. As RESTRAINT REMOVAL always needs to occur together with one of these image schemas, confusions might arise – depending on the focus of the analyst.
6 RESTRAINT REMOVAL × ENABLEMENT	As ENABLEMENT is a direct consequence of RESTRAINT REMOVAL, their co-occurrence might lead to confusion about their categorisation.
7 DIVERSION × (COMPULSION, ATTRACTION/REPULSION, BLOCKAGE, COUNTERFORCE, RESISTANCE)	DIVERSION denotes not the change of motion state, but of the direction of the motion PATH. It is a likely consequence of COMPULSION, ATTRACTION/REPULSION, BLOCKAGE, COUNTERFORCE, and RESISTANCE. If not carefully analysed, it can be easily confused with these.

The more likely confusions an image schema has with other image schemas, the lower the reliability of the classifications into the image schema category. Judging from the predictions in Table 3, lower reliabilities are expected for the image schemas BLOCKAGE, COUNTERFORCE, RESISTANCE, RESTRAINT REMOVAL, and DIVERSION, because they are often confused with other image schemas. Low confusions and high reliabilities are expected for the image schemas MOMENTUM and BALANCE, because they have fewer overlaps with other image schemas.

## 4.1 Method

Four people familiar with FORCE image schemas took part in the study. Two of them had experience in applying the full range of image schemas to the analysis of user interface elements of technology, such as an Airbus cockpit and business software. The other two (project collaborators) had each attended two one-day workshops on the application of FORCE image schemas to the design of technology. None of these participants had participated in or knew about the results of Study 1.

When assembling the material used in this study, the image-schema examples obtained in Study 1 were reviewed and any examples were removed that were not referring to user-technology interaction at the user-interface level (e.g. *driving a car against a tree*). Also, any duplicate examples and ambiguously phrased examples were removed (e.g. a phrase like *measuring devices* that does not point to a specific interaction episode). To further reduce the amount of examples, only those were included in the study that at least had received a confidence score equal or above three (of a maximum of five). This left 80 examples, which were then slightly edited to complete unfinished sentences, define technical terms like *brake power assist unit*, and add some context information like consequences of the described interaction. Care was taken to not give cues away easily as to which image-schema categories apply to the examples. A sentence like *The user blocked the system by doing XY* would be re-formulated to *The user did XY to prevent the system from . . .*

This list of examples was given to the participants, together with a ‘cheat sheet’ containing the ten image-schema definitions, notations, and technology examples (cf. Table 1). Participants were instructed to assign one image schema to each of the 80 examples in the questionnaire. On a five-point scale they indicated for each image-schema assignment how confident they were about their choice (coded from 1 = very uncertain to 5 = very certain).

Different image-schema categories had different prevalence in the list of examples. The categories of ATTRACTION/REPULSION (12 usage examples) and

ENABLEMENT (11 examples) had the highest prevalence. The category of COUNTERFORCE (3 examples) had the lowest prevalence. The other categories had medium prevalence: BLOCKAGE (9), RESTRAINT REMOVAL (9), COMPULSION (9), DIVERSION (8), MOMENTUM (7), BALANCE (6), and RESISTANCE (6).

## 4.2 Inter-coder agreement statistics

Inter-coder agreement is indicated by Cohen's kappa values (Cohen 1960; Eugenio and Glass 2004). In contrast to raw percentage values, kappa takes into account that a proportion of the agreement can occur purely by chance. Kappa values can vary between  $-1$  (complete disagreement) to  $+1$  (complete agreement). A kappa value of 0 indicates chance agreement. Inter-coder agreement can also be computed for single rating categories (here, single FORCE image schemas) and is then called intra-class agreement.

The interpretation of kappa values follows the guidance provided by Landis and Koch (1977). A kappa value of  $\kappa < .00$  indicates poor agreement,  $.00 \leq \kappa \leq .20$  slight agreement,  $.21 \leq \kappa \leq .40$  fair agreement,  $.41 \leq \kappa \leq .60$  moderate agreement,  $.61 \leq \kappa \leq .80$  substantial agreement, and  $.81 \leq \kappa \leq 1.00$  almost perfect agreement.

Note that these guidelines should be interpreted with care as the number and prevalence of categories and rater bias affect the magnitude of the value. The kappa will be lower when there are more categories, when prevalence is not homogenous across categories, and when biases between raters occur (Sim and Wright 2005). As a consequence, kappa values that are prevalence-adjusted and bias-adjusted (*PAK*) are also reported (for the calculation of *PAK*-values see Sim and Wright 2005).

## 4.3 Results

### 4.3.1 Agreement between coders

Four participants classified the 80 usage scenarios into ten image-schema categories. The overall kappa value is  $\kappa = .59$  (Table 4). Using the Landis and Koch (1977) criterion, this is interpreted as a moderate agreement between participants. The kappa values of single image-schema categories are 'almost perfect' in two cases, 'substantial' in two cases, 'moderate' in four cases, and 'fair' in two cases. 'Almost perfect' agreement was obtained for the image schemas ATTRACTION/REPULSION and BALANCE. Only 'fair' agreement was found for the image schemas RESISTANCE and COUNTERFORCE.

**Table 4:** Confidence scores and inter-coder reliabilities for FORCE image schemas

	Confidence scores		Inter-coder reliabilities	
	<i>M</i>	<i>SD</i>	$\kappa$	<i>PAK</i>
Overall	3.40	1.12	.59	.59
ATTRACTION/REPULSION	4.33	0.41	.95	.90
BALANCE	3.47	0.98	.81	.81
DIVERSION	3.46	1.27	.68	.71
ENABLEMENT	2.94	1.06	.61	.50
MOMENTUM	2.50	1.23	.52	.59
BLOCKAGE	3.36	1.18	.52	.48
RESTRAINT REMOVAL	4.22	0.45	.45	.49
COMPULSION	3.83	0.92	.45	.48
COUNTERFORCE	2.58	1.25	.34	.63
RESISTANCE	3.39	0.79	.30	.51

*Note:* Confidence ratings ranged from 1 = very uncertain to 5 = very certain.

Adjusting kappa values for prevalence and bias (*PAK*) had only small effects on most of the ratings, except for COUNTERFORCE, RESISTANCE, and ENABLEMENT. After adjustment, agreement values were much higher for RESISTANCE and COUNTERFORCE, now indicating moderate and substantial agreement. Agreement in the category ENABLEMENT, however, decreased to moderate agreement.

The confidence scores were reasonably high (overall  $M = 3.40$ ) with the lowest scores for COUNTERFORCE, ENABLEMENT, and MOMENTUM, the highest scores for ATTRACTION/REPULSION and RESTRAINT REMOVAL. The correlation between confidence scores and reliabilities is only of medium size ( $r = .39$ ) indicating that confidence scores cannot be used as a perfect predictor for inter-coder agreement.

#### 4.3.2 Agreement with the standard

This measure indicates how strong the agreement of single coders or groups of coders is with the original classifications made in Study 1. The overall kappa value is  $\kappa = .71$ , indicating a ‘substantial’ agreement. The kappa values of single image-schema categories are ‘almost perfect’ in two cases, ‘substantial’ in six cases, ‘moderate’ in one case, and ‘fair’ in one case (Table 5, second column). Again, ‘almost perfect’ agreement was obtained for the image schemas ATTRACTION/REPULSION and BALANCE. Only ‘fair’ agreement was found for the image schema COUNTERFORCE. When using prevalence adjustments the value for COUNTERFORCE changed to a ‘substantial agreement’. None of the other values changed as much.

**Table 5:** Agreement of differently sized groups of coders with the standard classification obtained in Study 1

	Single coders		Coder pairs		Coder triples		Four coders	
	$\kappa$	PAK	$\kappa$	PAK	$\kappa$	PAK	$\kappa$	PAK
Overall	.71	.72	.81	.82	.83	.83	.87	.88
ATTRACTION/REP	.95	.91	.98	.97	.97	.96	.95	.93
BALANCE	.89	.90	.97	.97	1.00	1.00	1.00	1.00
ENABLEMENT	.76	.67	.92	.89	.89	.85	.95	.93
MOMENTUM	.72	.75	.88	.89	.88	.89	.93	.93
DIVERSION	.67	.71	.74	.77	.72	.76	.78	.80
BLOCKAGE	.67	.64	.72	.71	.79	.78	.86	.86
RESTR. REMOVAL	.66	.66	.75	.74	.80	.78	.84	.80
RESISTANCE	.61	.70	.68	.77	.85	.89	1.00	1.00
COMPULSION	.56	.58	.69	.68	.64	.65	.70	.70
COUNTERFORCE	.40	.72	.57	.83	.56	.84	.65	.86

Can the agreement with the standard be enhanced by grouping participants and comparing their combined ratings to the standard? Out of four participants, six possible pairs and four possible triples were formed. Group scores were determined by choosing the image-schema classification that the majority of group members agreed on. Conflicts were resolved by considering the confidence ratings obtained for each classification. If, for example, in a pair of participants one participant classified a usage example as an instance of COUNTERFORCE with a confidence rating of 4 (out of 5) and the other as an instance of BLOCKAGE with a confidence rating of 2, then the classification of the pair was assumed to be COUNTERFORCE and this result was compared with the standard. This procedure served as a simple model for possible negotiation of classifications in a group of coders.

The results show that increasing the number of people classifying examples into image-schema categories enhances agreement with the standard (Table 5). In other words, the more coders participate in the classification of image schemas, the less errors occur. When using pairs of coders, the overall agreement with the standard rises to  $\kappa = .81$ , an ‘almost perfect agreement’ compared to the ‘substantial agreement’ using single coders. Adding a third and a fourth coder further increases the agreement with the standard.

The kappa values of single image schemas show a similar development. With pairs of coders classifying usage examples, four image-schema categories have ‘almost perfect agreement’, five show ‘substantial’ agreement, and one ‘moderate’ agreement with the standard. It is again the image schema COUNTERFORCE that lags behind, while ATTRACTION, BALANCE, ENABLEMENT, and MOMENTUM are

in the top group. A group of four coders delivers even better results: seven categories show ‘almost perfect agreement’, the other three show ‘substantial agreement’. The RESISTANCE category benefitted the most from pooling classifications in a group; the kappa increased from  $\kappa = .61$  with single coders to a maximum of  $\kappa = 1.00$  when a group of four coders was compared against the standard. Again, COUNTERFORCE showed the largest difference when using prevalence adjustments (cf. *PAK*-values in Table 5).

#### 4.3.3 Analysis of category confusions

The agreement-disagreement matrix (Table 6) shows the quantitative result of category confusions. Little more than half of the observed category confusions (56%) matched the predicted confusions (shaded cells in Table 6).

A qualitative analysis of the examples (Table 7) reveals likely sources of category confusions and indicates what can be done to enhance the reliability of image-schema categorisations. In summary, the results show that inter-category confusions occur in a range of typical situations:

**Table 6:** Observed agreement and disagreement between FORCE image schemas

	AT	BA	BL	CF	CP	DI	EN	MO	RE	RR	Total
ATTRACTION	<b>63</b>										63
BALANCE	0	<b>31</b>									31
BLOCKAGE	0	0	<b>35</b>								35
COUNTERFORCE	3	2	5	<b>9</b>							19
COMPULSION	0	2	8	10	<b>25</b>						45
DIVERSION	2	0	6	1	1	<b>26</b>					36
ENABLEMENT	0	3	2	0	12	0	<b>47</b>				64
MOMENTUM	0	3	7	1	10	4	6	<b>22</b>			53
RESISTANCE	1	3	11	8	4	5	2	2	<b>12</b>		48
RESTR. REMOVAL	0	0	11	0	2	3	22	2	9	<b>25</b>	74
Total	69	44	85	29	54	38	77	26	21	25	468

*Note:* The table contains the data of all six possible pairs out of four raters. Numbers in cells are absolute frequencies. Numbers in bold print denote agreement. Shaded cells mark predicted confusions between image-schema classifications.

**Table 7:** Predicted and observed confusions between FORCE image schemas

Predicted confusions	Observed confusions
1 MOMENTUM × ENABLEMENT	This prediction was partly confirmed. There are additional high overlaps of MOMENTUM with BLOCKAGE, DIVERSION, and COMPULSION that were not predicted. The overlap with COMPULSION is due to confusing ongoing motion (MOMENTUM) with motion that has been started by an external force (COMPULSION). The confusion of MOMENTUM with BLOCKAGE occurs with untypical MOMENTUM examples – like ongoing rest. MOMENTUM was sometimes suggested in cases that were generally not easy to assign to a category: MOMENTUM had the lowest confidence score of all categories, $M = 2.5$ (average $M = 3.4$ ).
2 BLOCKAGE × COMPULSION × ATTRACTION/REPULSION	The predicted confusion between BLOCKAGE and COMPULSION was confirmed. ATTRACTION/REPULSION had the lowest confusion rates of all image schemas (between $\kappa = .95$ and $.98$ ). Confusions did not occur with BLOCKAGE or COMPULSION. Few occurred with RESISTANCE and COUNTERFORCE (not expected).
3 BLOCKAGE × COUNTERFORCE × RESISTANCE	The predicted confusions between these three categories occur frequently. There are also unexpected confusions of RESISTANCE or COUNTERFORCE with the COMPULSION category. Often this confusion is based on a different focus on the direction of interaction, particularly if one rater focuses on the Agonist, the other on the Antagonist.
4 COUNTERFORCE × BALANCE	Some of the predicted confusion between COUNTERFORCE and BALANCE can be found in the data. BALANCE also shows some slight overlap with RESISTANCE, ENABLEMENT, and MOMENTUM. Some of the confusion with RESISTANCE is due to an interpretation of RESISTANCE as ‘lack of BALANCE’. BALANCE also seems to convey a sense of ENABLEMENT – there is no misbalance to be taken care of. BALANCE also helps in maintaining MOMENTUM.
5 RESTRAINT REMOVAL × (BLOCKAGE, COUNTERFORCE, RESISTANCE)	These predictions were confirmed by the data.
6 RESTRAINT REMOVAL × ENABLEMENT	This type of confusion is rather frequent. Especially cases of active ENABLEMENT are easily confused with RESTRAINT REMOVAL. Frequent, non-predicted confusions of ENABLEMENT also occur with COMPULSION, often due to a confusion of users enabling themselves to do something when they in fact compelled the system to take action.
7 DIVERSION × (COMPULSION, ATTRACTION/REPULSION, BLOCKAGE, COUNTERFORCE, RESISTANCE)	These types of confusion are rather frequent. Also, not as hypothesised, the image schemas MOMENTUM and RESTRAINT REMOVAL are sometimes confused with DIVERSION in usage scenarios that were ambiguous.

- When two image-schema instances occur together, especially
  - When one image schema describes a direct consequence of another image schema, e.g. *ENABLEMENT* as a consequence of *RESTRAINT REMOVAL*;
  - When one image schema is the precondition of another, e.g. *BLOCKAGE* is the precondition for *RESTRAINT REMOVAL*;
- When the direction of the interaction is not specified, i.e. when it is unclear which entity (here: the user or the technology) should be regarded as the Agonist in the interaction;
- When coders deal with non-typical examples of an image schema, e.g. instances of active (instead of passive) *ENABLEMENT*;
- When coders have not enough information to distinguish similar image-schema categories, e.g. when trying to distinguish *COMPULSION* from *MOMENTUM*, information is needed about the presence (*COMPULSION*) or absence (*MOMENTUM*) of external forces;
- When coders shift their focus between domains, here from human-computer interaction to instances of human-human interaction or system-internal interactions.

Knowing these potential causes of confusion, measures can be taken to increase inter-coder agreement (see below).

## 5 Discussion and conclusion

These studies investigated the inter-coder reliabilities of classifying force-dynamic events into *FORCE*-image-schema categories. In Study 1 eleven people brainstormed examples from their daily interaction with technology for each of ten *FORCE* image schemas. The results showed that people can relate their experiences easily to the image-schema definitions provided and that they did so with high confidence.

These examples were given to four people who were more experienced in dealing with *FORCE* image schemas. Their task was to assign each of 80 examples of usage scenarios to one out of ten *FORCE* image schemas they found most suitable. The overall inter-coder reliability was moderate with an overall kappa value of  $\kappa = .59$ . Kappa values for single image-schema categories ranged between ‘fair’ and ‘almost perfect’, i.e. between  $\kappa = .30$  and  $\kappa = .95$ .

Agreement with the originally intended image-schema classifications was higher than the pure inter-coder reliabilities. Single coders achieved an average kappa value of  $\kappa = .71$ , indicating substantial agreement with the standard. Using



more than one coder increased the agreement with the standard up to  $\kappa = .87$  when using a group of four coders (indicating an ‘almost perfect’ agreement).

The practical recommendation derived from these data could be this: if you strive for maximum agreement with a set standard, use as many coders as possible. Good results, however, can already be achieved with two coders. This makes studies that involve force-dynamic event categorisations economically feasible in practice.

However, a moderate inter-coder reliability of  $\kappa = .59$  still leaves room for improvement. The qualitative analysis leads to the conclusion that a more systematic training of image-schema coders could be beneficial. The training should include definitions and examples of image schemas. Special emphasis when coding force-dynamic events should be put on:

- The subtleties of different image-schema definitions, e.g. of the distinction between passive and active ENABLEMENT;
- The correct identification of the direction of the interaction as it determines the Agonist and Antagonist;
- The distinctions between causes, results, and further consequences of force-dynamic events;
- What information is needed and where this information can be found, e.g. to distinguish BLOCKAGE, RESISTANCE and COUNTERFORCE from each other.

In more general terms, also taking the limitations of this study into account, one may further recommend:

- Use rich information. If information is sparse as in the one-sentence events of this study, then reliabilities may only be moderate.
- Use more than one coder. This study showed that using pairs of coders is best when agreement with a standard needs to be enhanced.
- Provide explicit coding rules. Coders will make up their own rules in rating similar usage situations. If coders applied the same rules, large amounts of disagreement could be removed.
- Check the validity of the rules. Agreed-upon rules can enhance the inter-coder agreement. But it does not follow that they are also valid in a conceptual sense. Rules need to be empirically tested in different contexts of use.
- Train the analysts. Although even novices can apply image schemas directly, coders could still profit from a systematic training. The training should include the presentation of image-schema definitions and rules for assigning them to different components of the context of use, as well as extensive exercise and feedback, especially discussing the sources of disagreement.

Again, it should be noted that reliability is important in establishing image schemas as the vocabulary of concepts and mental models, as is often done in cognitive linguistics. The usefulness of extensive training, using more than one coder, and providing a fixed set of coding rules has already been shown for identifying whether linguistic expressions are metaphoric or not (Steen et al. 2010). In corpora of four registers and two languages, average inter-coder reliabilities of  $\kappa \geq .80$  could be achieved, indicating almost perfect agreement between coders.

Finally, it cannot be the goal for practitioners to achieve complete agreement between coders. Sometimes the most interesting discoveries are made if people disagree on how to categorise events. Disagreement between coders, if acknowledged, can lead to fruitful discussions and eventually to better, more focussed and valid descriptions of force-dynamic events.

Further research needs to probe deeper beyond using single-sentence episodes involving larger corpora, professional linguists as coders, and other cognitive linguistic categories (e.g. metaphors, other image schemas).

## References

- Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1). 37–46.
- Eugenio, Barbara Di & Michael Glass. 2004. The Kappa statistic: A second look. *Computational Linguistics* 30(1). 95–101.
- Hurtienne, Jörn. 2011. *Image schemas and design for intuitive use. Exploring new guidance for user interface design*. Berlin: Technische Universität Berlin doctoral dissertation. Retrieved from [http://opus.kobv.de/tuberlin/volltexte/2011/2970/pdf/hurtienne\\_joern.pdf](http://opus.kobv.de/tuberlin/volltexte/2011/2970/pdf/hurtienne_joern.pdf).
- Hurtienne, Jörn, Johann Habakuk Israel & Katharina Weber. 2008. Cooking up real world business applications combining physicality, digitality, and image schemas. In Albrecht Schmidt, Hans Gellersen, Elsie van den Hoven, Ali Mazalek, Paul Holleis & Nicolas Villar (eds.), *TEI'08. Second International Conference on Tangible and Embedded Interaction*, 239–246. New York: ACM.
- Hurtienne, Jörn, Katharina Weber & Lucienne Blessing. 2008. Prior experience and intuitive use: Image Schemas in user centred design. In Patrick Langdon, P. John Clarkson & Peter Robinson (eds.), *Designing Inclusive Futures*, 107–116. London: Springer.
- Johnson, Mark. 1987. *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. Chicago: University of Chicago Press.
- Landis, J. Richard & Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33. 159–174.
- Sim, Julius & Chris C. Wright. 2005. The Kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, 85(3). 257.
- Steen, Gerard J., Aletta G. Dorst, J. Berenike Herrmann, Anna Kaal, Tina Krennmayr & Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. Amsterdam: John Benjamins.

Talmy, Leonard. 1988. Force dynamics in language and cognition. *Cognitive Science* 12(1). 49–100.

Talmy, Leonard. 2000. *Toward a Cognitive Semantics*. Cambridge, MA: MIT Press.

