Identifying Semantic Relations and Functional Properties of Human Verb Associations

Sabine Schulte im Walde and Alissa Melinger
Computational Linguistics and Psycholinguistics
Saarland University
Saarbrücken, Germany
`{schulte,melinger}@coli.uni-sb.de`

Abstract

This paper uses human verb associations as the basis for an investigation of verb properties, focusing on semantic verb relations and prominent nominal features. First, the lexical semantic taxonymy GermaNet is checked on the types of classic semantic relations in our data; verb-verb pairs not covered by GermaNet can help to detect missing links in the taxonomy, and provide a useful basis for defining non-classical relations. Second, a statistical grammar is used for determining the conceptual roles of the noun responses. We present prominent syntax-semantic roles and evidence for the usefulness of co-occurrence information in distributional verb descriptions.

1 Introduction

This paper presents an examination of a collection of semantic associates evoked by German verbs in a web experiment. We define semantic associates here as those concepts spontaneously called to mind by a stimulus word. In the current investigation, we assume that these evoked concepts reflect highly salient linguistic and conceptual features of the stimulus word. Given this assumption, identifying the types of information provided by speakers and distinguishing and quantifying the relationships between stimulus and response can serve a number of purposes for NLP applications.

First, the notion of semantic verb relations is crucial for many NLP tasks and applications such as verb clustering (Pereira et al., 1993; Merlo and Stevenson, 2001; Lin, 1998; Schulte im Walde, 2003), thesaurus extraction (Lin, 1999; McCarthy et al., 2003), word sense discrimination (Schütze, 1998), text indexing (Deerwester et al., 1990), and summarisation (Barzilay et al., 2002). Different applications incorporate different semantic verb relations, varying with respect to their demands. To date, limited effort has been spent on specifying the range of verb-verb relations. Morris and Hirst (2004) perform a study on lexical semantic relations which ensure text cohesion. Their relations are not specific to verb-verb pairs, but include e.g. descriptive noun-adjective pairs (such as *professors/brilliant*), or stereotypical relations (such as *homeless/drunk*). Chklovski and Pantel (2004) address the automatic acquisition of verb-verb pairs and their relations from the web. They define syntagmatic patterns to cover strength, enablement and temporal relations in addition to synonymy and antonymy, but they do not perform an exhaustive study. We suggest that an analysis of human verb-verb associations may identify the range of semantic relations which are crucial in NLP applications. We present a preparatory study where the lexical semantic taxonymy GermaNet (Kunze, 2000; Kunze, 2004) is checked on the types of classical semantic verb relations[1] in our data; verb-verb pairs not covered by GermaNet can help to detect missing links in the taxonomy, and provide an empirical basis for defining non-classical relations.

---

[1]We follow Morris and Hirst (2004) and refer to the paradigmatic WordNet relations as the "classical" relations.

Second, in data-intensive lexical semantics, words are commonly modelled by distributional vectors, and the relatedness of words is measured by vector similarity. The features in the distributional descriptions can be varied in nature: words co-occurring in a document, in a context window, or with respect to a word-word relationship, such as syntactic structure, syntactic and semantic valency, etc. Most previous work on distributional similarity has either focused on a specific word-word relation (such as Pereira et al. (1993) referring to a direct object noun for describing verbs), or used any dependency relation detected by the chunker or parser (such as Lin (1999; 1998), and McCarthy et al. (2003)). Little effort has been spent on varying the (mostly nominal) types of verb features. We assume that the noun associates in our verb experiment are related to conceptual roles of the respective verbs, and investigate the linguistic functions that are realised by the response nouns with respect to the target verb, based on an empirical grammar model (Schulte im Walde, 2003). Even though the usage of the distributional features depends on the respective application, we present prominent roles and evidence for the usefulness of co-occurrence information in distributional verb descriptions.

2 Web Experiment

This section introduces our web experiment, as the data source for the explorations to follow. The web experiment asked native speakers to provide associations to German verbs.

2.1 Experiment Method

Material: 330 verbs were selected for the experiment. They were drawn from a variety of semantic classes including verbs of self-motion (e.g. *gehen* 'walk', *schwimmen* 'swim'), transfer of possession (e.g. *kaufen* 'buy', *kriegen* 'receive'), cause (e.g. *verbrennen* 'burn', *reduzieren* 'reduce'), experiencing (e.g. *hassen* 'hate', *überraschen* 'surprise'), communication (e.g. *reden* 'talk', *beneiden* 'envy'), etc. Drawing verbs from different categories was intended only to ensure that the experiment covered a wide variety of verb types; the inclusion of any verb in any particular verb class was achieved in part with reference to prior verb classification work (e.g.

Levin (1993)) but also on intuitive grounds. It is not critical for the subsequent analyses. The target verbs were divided randomly into 6 separate experimental lists of 55 verbs each. The lists were balanced for class affiliation and frequency ranges (0, 100, 500, 1000, 5000), such that each list contained verbs from each grossly defined semantic class, and had equivalent overall verb frequency distributions. The frequencies of the verbs were determined by a 35 million word newspaper corpus; the verbs showed corpus frequencies between 1 and 71,604.

Procedure: The experiment was administered over the Internet. When participants loaded the experimental page, they were first asked for their biographical information, such as linguistic expertise, age and regional dialect. Next, the participant was presented with the written instructions for the experiment and an example item with potential responses. In the actual experiment, each trial consisted of a verb presented in a box at the top of the screen. All stimulus verbs were presented in the infinitive. Below the verb was a series of data input lines where participants could type their associations. They were instructed to type at most one word per line and, following German grammar, to distinguish nouns from other parts of speech with capitalisation. [2] Participants had 30 sec. per verb to type as many associations as they could. After this time limit, the program automatically advanced to the next trial.

Participants and Data: 299 native German speakers participated in the experiment, between 44 and 54 for each data set. 132 of the individuals identified themselves as having had a linguistics education and 166 rated themselves as linguistic novices. In total, we collected 81,373 associations from 16,445 trials; each trial elicited an average of 5.16 associate responses with a range of 0-16.

2.2 Data Preparation

Each completed data set contains the background information of the participant, followed by the list of target verbs. Each target verb is paired with a list of associations in the order in which the participant provided them. For the analyses to follow, we preprocessed all data sets in the following way: For each target verb, we quantified over all responses in

---

[2]Despite these instructions, some participants failed to use capitalisation, leading to some ambiguity.

the experiment, disregarding the participant's background and the order of the associates. Table 1 lists the 10 most frequent responses for the verb *klagen* 'complain, moan, sue'. 64% of all responses were provided more than once for a target verb, and 36% were idiosyncratic, i.e. given only once. The verb responses were not distinguished according to polysemic senses of the verbs.

klagen 'complain, moan, sue'		
Gericht	'court'	19
jammern	'moan'	18
weinen	'cry'	13
Anwalt	'lawyer'	11
Richter	'judge'	9
Klage	'complaint'	7
Leid	'suffering'	6
Trauer	'mourning'	6
Klagemauer	'Wailing Wall'	5
laut	'noisy'	5

Table 1: Association frequencies for target verb.

3 Linguistic Analyses of Experiment Data

The verb associations are investigated on three linguistic dimensions:

1. In a preparatory step, we distinguish the responses with respect to the major part-of-speech tags: nouns, verbs, adjectives, adverbs.

2. For each verb associate, we look up the semantic relation between the target and response verbs using the lexical taxonomy GermaNet.

3. For each noun associate, we investigate the kinds of linguistic functions that are realised by the noun with respect to the target verb. The analysis is based on an empirical grammar.

For expository purposes, the paper is organised into three analysis sections, with discussions following each analysis.

3.1 *Excursus:* Empirical Grammar Model

The quantitative data in the analyses to follow are derived from an empirical grammar model (Schulte im Walde, 2003): a German context-free grammar was developed with specific attention towards verb subcategorisation. The grammar was lexicalised, and the parameters of the probabilistic version were estimated in an unsupervised training procedure,

using 35 million words of a large German newspaper corpus from the 1990s. The trained grammar model provides empirical frequencies for word forms, parts-of-speech tags and lemmas, and quantitative information on lexicalised rules and syntax-semantics head-head co-occurrences.

3.2 Morpho-Syntactic Analysis

The morpho-syntactic analysis is a preparatory step for the analyses to follow. Each associate of the target verb is assigned its (possibly ambiguous) part-of-speech by our empirical grammar dictionary. Originally, the dictionary distinguished approx. 50 morpho-syntactic categories, but we disregard fine-grained distinctions such as case, number and gender features and consider only the major categories verb (V), noun (N), adjective (ADJ) and adverb (ADV). Ambiguities between these categories arise e.g. in the case of nominalised verbs (such as *Rauchen* 'smoke', *Vergnügen* 'please/pleasure'), where the experiment participant could have been referring either to a verb or a noun, or in the case of past participles (such as *verschlafen*) and infinitives (such as *überlegen*), where the participant could have been referring either to a verb ('sleep' or 'think about', for the two examples respectively) or an adjective ('drowsy' or 'superior', respectively). In total, 4% of all response types are ambiguous between multiple part-of-speech tags.

Having assigned part-of-speech tags to the associates, we can distinguish and quantify the morpho-syntactic categories of the responses. In non-ambiguous situations, the unique part-of-speech receives the total target-response frequency; in ambiguous situations, the target-response frequency is split over the possible part-of-speech tags. As the result of this first analysis, we can specify the frequency distributions of the part-of-speech tags for each verb, and also in total. Table 2 presents the total numbers and specific verb examples. Participants provided noun associates in the clear majority of token instances, 62%; verbs were given in 25% of the responses, adjectives in 11%, adverbs almost never (2%).[3] The part-of-speech distribution for response words is correlated with target verb frequency. The rate of verb and adverb

---

[3]All of our analyses reported in this paper are based on response tokens; the type analyses show the same overall pictures.

responses is positively correlated with target verb frequency, Pearson's $r(328)=.294$, $p<.001$ for verbs and $r(328)=.229$, $p<.001$ for adverbs, while the rate of noun and adjective responses is inversely correlated with verb frequency, Pearson's $r(328)=-.155$, $p<.005$ for nouns and $r(328)=.114$, $p<.05$ for adjectives. The distribution of responses over part-of-speech also varies across verb classes. For example, aspectual verbs, such as *aufhören* 'stop', received more verb responses, $t(12)=3.11$, $p<.01$, and fewer noun responses, $t(12)=3.84$, $p<.002$, than creation verbs, such as *backen* 'bake', although the verb sets have comparable frequencies, $t(12)=1.1$, $p<.2$.

	V	N	ADJ	ADV
Total Freq	19,863	48,905	8,510	1,268
Total Prob	25%	62%	11%	2%
aufhören 'stop'	49%	39%	4%	6%
aufregen 'be upset'	22%	54%	21%	0%
backen 'bake'	7%	86%	6%	1%
bemerken 'realise'	52%	31%	12%	2%
dünken 'seem'	46%	30%	18%	1%
flüstern 'whisper'	19%	43%	37%	0%
nehmen 'take'	60%	31%	3%	2%
radeln 'bike'	8%	84%	6%	2%
schreiben 'write'	14%	81%	4%	1%

Table 2: Part-of-speech tags.

3.3 Semantic Verb Relations

For each verb associate, we look up the semantic relation between the target and response verbs using the lexical semantic taxonomy GermaNet (Kunze, 2000; Kunze, 2004), the German counterpart to WordNet (Fellbaum, 1998). The lexical database is inspired by psycholinguistic research on human lexical memory. It organises nouns, verbs, adjectives and adverbs into classes of synonyms (synsets), which are connected by lexical and conceptual relations. The GermaNet version from October 2001 contains 6,904 verbs and defines the paradigmatic semantic relations *synonymy, antonymy, hypernymy/hyponymy* as well as the non-paradigmatic relations *entailment, cause,* and *also see* between verbs or verb synsets. (*Also see* is an underspecified relation, which captures relationships other than the preceding ones. For example, *sparen* 'save' is related to *haushalten* 'budget' by *also see*.) Words with several senses are assigned to multiple synsets.

Based on the GermaNet relations, we can distinguish between the different kinds of verb associations elicited from speakers. Our analysis proceeds as follows. For each pair of target and response verbs, we look up whether any kind of semantic relation is defined between any of the synsets the verbs belong to. For example, if the target verb *rennen* 'run' is in synsets a and b, and the response verb *bewegen* 'move' is in synsets c and d, we determine whether there is any semantic relation between the synsets a and c, a and d, b and c, b and d. Two verbs belonging to the same synset are synonymous. The semantic relations are quantified by the target-response verb frequencies, e.g. if 12 participants provided the association *bewegen* for *rennen*, the hypernymy relation is quantified by the frequency 12. If the target and the response verb are both in GermaNet, but there is no relation between their synsets, then the verbs do not bear any kind of semantic relation, according to GermaNet's current status. If either of them is not in GermaNet, we cannot make any statement about the verb-verb relationship. Table 3 shows the number of semantic relations encoded in our GermaNet version, and the frequencies and probabilities of our response tokens found among them.[4] For example, there are 9,275 verb-verb instances where GermaNet defines a hypernymy-hyponymy relation between their synsets; for 2,807 of our verb-verb pairs (verb response tokens with respect to target verbs) we found a hypernymy relation among the GermaNet definitions, which accounts for 14% of all our verb responses.

The distribution of target-response relations is also correlated with target verb frequency. The proportion of associate responses captured by the respective relations of synonym, antonym and hyponym increases as a function of target verb frequency, $r(323)=.147$ for synonymy, $r(328)=.341$ for antonymy and $r(328)=.243$ for hyponymy (all $p<.01$); the proportion of hypernym relations is not correlated with verb frequency. The distribution of relations also varies by verb class. For example, aspectual target verbs like *aufhören* 'stop' received significantly more antonymic responses like *anfangen* 'begin' or *weitermachen* 'go on' than creation verbs such as *backen* 'bake', $t(12)=3.44$, $p<.05$.

---

[4] Note that the number of encoded relations in GermaNet differs strongly, which influences the number of verb-verb tokens that can potentially be found.

	GermaNet	Freq	Prob
Synonymy	4,633	1,194	6%
Antonymy	226	252	1%
Hypernymy	9,275	2,807	14%
Hyponymy	9,275	3,016	16%
Cause	95	49	0%
Entailment	8	0	0%
Also see	1	0	0%
No relation	-	10,509	54%
Unknown cases	-	1,726	9%

Table 3: Semantic relations.

An interesting piece of information is provided by the verb-verb pairs for which we do not find a relationship in GermaNet. The minority of such cases (9%) is due to part-of-speech confusion based on capitalisation errors by the participants, cf. footnote 2; e.g. the non-capitalised noun *wärme* 'warmth' was classified as a verb because it is the imperative of the verb *wärmen* 'warm'. A remarkable number of verb-verb associations (54%) do not show any kind of semantic relation according to GermaNet despite both verbs appearing in the taxonomy. On the one hand, this is partly due to the GermaNet taxonomy not being finished yet; we find verb associations such as *weglaufen* 'run away' for *abhauen* 'walk off' (12 times), or *untersuchen* 'examine' for *analysieren* 'analyse' (8 times) where we assume (near) synonymy not yet coded in GermaNet; or *weggehen* 'leave' for *ankommen* 'arrive' (6 times), and *frieren* 'be cold' for *schwitzen* 'sweat' (2 times) where we assume antonymy not yet coded in GermaNet. For those cases, our association data provides a useful basis for detecting missing links in GermaNet, which can be used to enhance the taxonomy. However, a large proportion of the "no relation" associations represent instances of verb-verb relations not targeted by GermaNet. For example, *adressieren* 'address' was associated with the temporally preceding *schreiben* 'write' (15 times) and the temporally following *schicken* 'send' (6 times); *schwitzen* 'sweat' was associated with a consequence *stinken* 'stink' (8 times) and with a cause *laufen* 'run' (5 times); *setzen* 'seat' was associated with the implication *sitzen* 'sit' (2 times), *erfahren* 'get to know' with the implication *wissen* 'know' (8 times). Those examples represent instantiations of non-classical verb relations and could be subsumed under *also see* relations in GermaNet, but it is obvi-

ous that one would prefer more fine-grained distinctions. We are specifically interested in those cases, because we expect that human associations cover the range of possible semantic verb relations to a large extent, and we believe that they represent an excellent basis for defining an exhaustive set, as alternative to e.g. text-based relations (Morris and Hirst, 2004). Again, the diversity of semantic verb relations is a crucial ingredient for NLP tasks such as thesaurus extraction, summarisation, and question answering.

Window Look-up We have argued above that an investigation into the types of semantic relations instantiated by verb-verb associations could be relevant in NLP. Thus, we are interested in whether paradigmatically related verb-verb pairs co-occur in texts. To evaluate this point, we perform a window look-up, in order to determine the distance between two associated verbs. We use our complete newspaper corpus, 200 million words, and check whether the response verbs occur in a window of 5/20/50 words to the left or to the right of the relevant target word. For paradigmatically related verb pairs, namely those whose relation we could determine with GermaNet (37%), we find 85/95/97% in the respective windows. For those whose relation is unspecified in GermaNet (63%), we find lower co-occurrence rates, 61/74/79%. The fact that the distances between verbs and the co-occurrence rates differ with respect to the category of semantic relation, e.g. paradigmatic or not, is useful for NLP applications such as summarisation, where both the distances between salient words and their semantic relations are crucial. More precise conditions (e.g. different window sizes, structural sentence/paragraph distinctions, quantification of window matches by their frequencies) shall be specified in future work.

3.4 Syntax-Semantic Noun Functions

In a third step, we investigate the kinds of linguistic functions that are realised by noun associates of the target verbs. Our hypothesis is that the properties of the elicited noun concepts provide insight into conceptual features for distributional verb descriptions.

The analysis utilises the empirical grammar model, cf. Section 3.1. With respect to verb sub-

categorisation, the grammar defines frequency distributions of verbs for 178 subcategorisation frame types, including prepositional phrase information, and frequency distributions of verbs for nominal argument fillers. For example, the verb *backen* 'bake' appeared 240 times in our training corpus. In 80 of these instances it was parsed as intransitive, and in 109 instances it was parsed as transitive subcategorising for a direct object. The most frequent nouns subcategorised for as direct objects are *Brötchen* 'rolls', *Brot* 'bread', *Kuchen* 'cake', *Plätzchen* 'cookies', *Waffel* 'waffle'.

We use the grammar information to look up the syntactic relationships which exist between a target verb and a response noun. For example, the nouns *Kuchen* 'cake', *Brot* 'bread', *Pizza* and *Mutter* 'mother' were produced in response to the target verb *backen* 'bake'. The grammar look-up tells us that *Kuchen* 'cake' and *Brot* 'bread' appear not only as the verb's direct objects (as illustrated above), but also as intransitive subjects; *Pizza* appears only as a direct object, and *Mutter* 'mother' appears only as transitive subject. The verb-noun relationships which are found in the grammar are quantified by the verb-noun association frequency, divided by the number of different relationships (to account for the ambiguity represented by multiple relationships). For example, the noun *Kuchen* was elicited 45 times in response to *bake*; the grammar contains the noun both as direct object and as intransitive subject for that verb, so both functions are assigned a frequency of 22.5. In a second variant of the analysis, we also distributed the verb-noun association frequencies over multiple relationships according to the empirical proportions of the respective relationships in the grammar, e.g. of the total association frequency of 45 for *Kuchen*, 15 would be assigned to the direct object of *backen*, and 30 to the intransitive subject if the empirical grammar evidence for the respective functions of *backen* were one vs. two thirds.

In a following step, we accumulate the association frequency proportions with respect to a specific relationship, e.g. for the direct objects of *backen* 'bake' we sum over the frequency proportions for *Kuchen*, *Brot*, *Plätzchen*, *Brötchen*, etc. The final result is a frequency distribution over linguistic functions for each target verb, i.e. for each verb we can determine which linguistic functions are acti-

vated by how many noun associates. For example, the most prominent functions for the inchoative-causative verb *backen* 'bake' are the transitive direct object (8%), the intransitive subject (7%) and the transitive subject (4%).

By generalising over all verbs, we discover that only 11 frame-slot combinations are activated by at least 1% of the nouns: subjects in the intransitive frame, the transitive frame (with direct/indirect object, or prepositional phrase) and the ditransitive frame; the direct object slot in the transitive, the ditransitive frame and the direct object plus PP frame; the indirect object in a transitive and ditransitive frame, and the prepositional phrase headed by *Dat:in*, dative (locative) 'in'. The frequency and probability proportions are illustrated by Table 4; the function is indicated by a slot within a frame (with the relevant slot in bold font); 'S' is a subject slot, 'AO' an accusative (direct) object, 'DO' a dative (indirect) object, and 'PP' a prepositional phrase. The activation of the functions differs only slightly with the analysis variant distributing the association frequencies with respect to grammar evidence, see above.

Interestingly, different verb classes are associated to frame slots to varying degrees. For example, verbs of creation like *backen* 'bake' elicited direct object slot fillers significantly more often than aspectual verbs like *aufhören* 'stop', $t(12)=2.24$, $p<.05$.

	Function	Freq	Prob
S	**S** V	1,793	4%
	S V AO	1,065	2%
	S V DO	330	1%
	S V AO DO	344	1%
	S V PP	510	1%
AO	S V **AO**	2,298	5%
	S V **AO** DO	882	2%
	S V **AO** PP	706	1%
DO	S V **DO**	302	1%
	S V AO **DO**	597	1%
PP	S V **PP-Dat:in**	418	1%
Unknown noun		10,663	22%
Unknown function		24,536	50%

Table 4: Associates as slot fillers.

In total, only 28% of all noun associates were identified by the statistical grammar as frame-slots fillers. However, the analysis of noun functions shows that a range of linguistic functions might be considered as prominent, e.g. 11 functions are ac-

tivated by more than 1% of the associates. Our hope is that these frame-role combinations are candidates for defining distributional verb descriptions. As mentioned before, most previous work on distributional similarity has focused either on a specific word-word relation (such as Pereira et al. (1993) referring to a direct object noun for describing verbs), or used any syntactic relationship detected by the chunker or parser (such as Lin (1999; 1998) and McCarthy et al. (2003)). Naturally, the contribution of distributional features depends on the distributional objects and the application, but our results suggest that it is worth determining a task-specific set of prominent features.

The majority of noun responses were not found as slot fillers. 22% of the associates are missing because they do not appear in the grammar model at all. These cases are due to (i) lemmatisation in the empirical grammar dictionary, where noun compounds such as *Autorennen* 'car racing' are lemmatised by their lexical heads, creating a mismatch between the full compound and its head; (ii) domain and size of training corpus, which underrepresents slang responses like *Grufties* 'old people', dialect expressions such as *Ausstecherle* 'cookie-cutter' as well as technical expressions such as *Plosiv* 'plosive'. The remaining 50% of the nouns are represented in the grammar but do not fill subcategorised-for linguistic functions; clearly the conceptual roles of the noun associates are not restricted to the subcategorisation of the target verbs. In part what is or is not covered by the grammar model can be characterised as an argument/adjunct contrast. The grammar model distinguishes argument and adjunct functions, and only arguments are included in the verb subcategorisation and therefore found as linguistic functions. Adjuncts such as the instrument *Pinsel* 'brush' for *bemalen* 'paint' (21 times), *Pfanne* 'pan' for *erhitzen* 'heat' (2), or clause-internal information such as *Aufmerksamkeit* 'attention' for *bemerken* 'notice' (6) and *Musik* 'music' for *feiern* 'celebrate' (10) are not found. These associates fulfill scene-related roles which are not captured by subcategorisation in the grammar model. In addition, we find associates which capture clause-external scene-related information or refer to world knowledge which is not expected to be found in the context at all. For example, the association *Trock-*

ner 'dryer' as the instrument for *trocknen* 'dry' (11 times) is not typically mentioned with the verb; similarly *Wasser* 'water' for *auftauen* 'defrost' (14), *Freude* 'joy' for *überraschen* 'surprise' (24), or *Verantwortung* 'responsibility' for *leiten* 'guide' (4) reflect world knowledge and may not be found in the immediate context of the verb.

Window Look-up Of course, the distinction between arguments, adjuncts, scene-related roles and world knowledge reflects a continuum. As a follow-up experiment, we perform a window look-up on the verb-noun pairs, in order to determine what portion of the nouns co-occur in the context of the verb and what portion is missing. This should provide us with a rough idea of the conceptual roles which are world knowledge and not found in the context. We again use our complete newspaper corpus, 200 million words, and check whether the response nouns are in a window of 5/20/50 words to the left or to the right of the relevant target verb. Naturally, most noun associates which were found as slot fillers in the functional analysis also appear in the window (since they are part of the subcategorisation): 99/99/100%. Of those cases which are not argument slot-fillers in the preceding functional analysis, we find 55/69/75% in our large corpus, i.e. more than half of the 72% missing noun tokens are in a window of 5 words from the verb, three quarters are captured by a large window of 50 words, one quarter is still missing. We conclude that, in addition to the conceptual roles referring to verb subcategorisation roles, the associations point to scene-related information and world knowledge, much of which is not explicitly mentioned in the context of the verb. With respect to a distributional feature description of verbs, we suggest that a set of prominent functions is relevant, but in addition it makes sense to include window-based nouns, which refer to scene information rather than intra-sentential syntactic functions. This is an interesting finding, since the window approach has largely been disregarded in recent years, in comparison to using syntactic functions.

4 Summary

This paper presented a study to identify, distinguish and quantify the various types of semantic associations provided by humans, and to illustrate their us-

age for NLP. For the approx. 20,000 verb associates, we specified classical GermaNet relations for 37% of the verb-verb pairs, and demonstrated that the co-occurrence distance between two verbs varies with respect to their semantic relation. Verb-verb pairs with no relation in GermaNet provide an empirical basis for detecting missing links in the taxonomy. Non-classical verb-verb relations such as temporal order, cause, and consequence are represented in a large proportion of the verb-verb pairs. These data represent an excellent basis for defining an exhaustive set of non-classical relations, a crucial ingredient for NLP applications.

For the approx. 50,000 noun associates, we investigated the kinds of linguistic functions that are realised by the verb-noun pairs. For 28% of the noun tokens, we found prominent frame-role combinations which speakers have in mind; our hope is that these conceptual roles represent features which contribute to distributional verb descriptions. Window-based nouns also contribute to verb descriptions by encoding scene information, rather than intra-sentential functions. This finding supports the integration of window-based approaches into function-based approaches.

Future work will establish a set of non-classical verb-verb relations, and then apply variations of verb feature descriptions in order to find dependencies between feature descriptions and verb relations. Such dependencies would improve the application of distributional verb descriptions significantly, knowing which relations are addressed by which kinds of features. In addition, we assume that the (morphological, syntactic, semantic) kinds of associates provided for a verb are indicators for its semantic class. Further investigations into the varied distributions of associate types across semantic classes will enhance the automatic acquisition of such classes. We plan to investigate this issue in more detail.

References

Regina Barzilay, Noemie Elhadad, and Kathleen R. McKeown. 2002. Inferring Strategies for Sentence Ordering in Multidocument News Summarization. *Journal of Artificial Intelligence Research*.

Timothy Chklovski and Patrick Pantel. 2004. VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*.

Christiane Fellbaum, editor. 1998. *WordNet – An Electronic Lexical Database*. MIT Press.

Claudia Kunze. 2000. Extension and Use of GermaNet, a Lexical-Semantic Database. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*.

Claudia Kunze. 2004. Semantische Relationstypen in GermaNet. In Stefan Langer and Daniel Schnorbusch, editors, *Semantik im Lexikon*. Gunter Narr Verlag.

Beth Levin. 1993. *English Verb Classes and Alternations*. The University of Chicago Press.

Dekang Lin. 1998. Automatic Retrieval and Clustering of Similar Words. In *Proceedings of the 17th International Conference on Computational Linguistics*.

Dekang Lin. 1999. Automatic Identification of Non-compositional Phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a Continuum of Compositionality in Phrasal Verbs. In *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*.

Paola Merlo and Suzanne Stevenson. 2001. Automatic Verb Classification Based on Statistical Distributions of Argument Structure. *Computational Linguistics*.

Jane Morris and Graeme Hirst. 2004. Non-Classical Lexical Semantic Relations. In *Proceedings of the HLT Workshop on Computational Lexical Semantics*.

Fernando Pereira, Naftali Tishby, and Lillian Lee. 1993. Distributional Clustering of English Words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*.

Sabine Schulte im Walde. 2003. *Experiments on the Automatic Induction of German Semantic Verb Classes*. Ph.D. thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Hinrich Schütze. 1998. Automatic Word Sense Discrimination. *Computational Linguistics*.