Which Distributional Functions are Crucial to Word Meaning: An Investigation of Semantic Associations

Sabine Schulte im Walde
Institute for Natural Language Processing
University of Stuttgart
Germany

Alissa Melinger
School of Psychology
University of Dundee
Scotland, U.K.

Michael Roth
Computational Linguistics
Saarland University
Saarbrücken, Germany

Andrea Weber
Psycholinguistics
Saarland University
Saarbrücken, Germany

Abstract

This article presents a study to distinguish and quantify the various types of semantic associations provided by humans, and to illustrate their usage for NLP purposes. Specifically, we address the task of modelling word meaning by empirical features in data-intensive lexical semantics. Relying on large-scale corpus-based resources, we identify the contextual categories and functions that are activated by the associates and therefore contribute to the salient meaning components of individual words and established across words. As a result, we present prominent conceptual roles and evidence for the usefulness of co-occurrence information in distributional descriptions.

1 Motivation

This article uses a collection of semantic associates as the basis for an empirical characterisation of verb and noun properties. We define *semantic associates* here as those concepts spontaneously called to mind by a stimulus word, and 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 creating NLP resources and defining and applying NLP techniques.

Within this article, we address the task of *modelling word meaning by empirical features*. In order to determine the similarity or dissimilarity between words, sentences, paragraphs, or even documents, approaches to data-intensive lexical semantics must empirically define and induce features that (a) capture the various meaning aspects of the words to be described, and (b) can be obtained automatically from corpus-data. Progressing from the word level to the document level, examples for this task are: clustering of similar words (Pereira et al., 1993; Lin, 1998; Merlo and Stevenson, 2001; Schulte im Walde, 2006), word sense discrimination (Schütze, 1998), the identification of multiword expressions (Lin, 1999) and their decomposability (Baldwin et al., 2003), anaphora resolution (Poesio et al., 2002), and text indexing (Deerwester et al., 1990), among others.

Generally, the necessary semantic features for these tasks are not readily available.[1] Following the *distributional hypothesis*, namely that 'each language can be described in terms of a distributional structure, i.e., in terms of the occurrence of parts relative to other parts' (Harris, 1968), distributional descriptions have been applied to model aspects of word meaning. Specifically, contextual features such as 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. have been used. However, these prior investigations of distributional similarity have either focused on a specific word-word relation to induce features (such as Pereira et al. (1993) and Rooth et al. (1999) referring to a direct object noun for describing verbs, and Curran (2003) referring to subjects and direct objects), or used any dependency relation detected by the chunker or parser (such as Lin (1998) and McCarthy et al. (2003)). Little effort has been spent on investigating the eligibility of the types of features. We assume that se-

[1] Few resources are semantically annotated and provide semantic information off-the-shelf (such as *FrameNet* (Baker et al., 1998) and *PropBank* (Palmer et al., 2005)).

mantic associates provide a useful means to identify the contextual functions that might be relevant to empirical feature descriptions, by examining which functions are activated by the associates and therefore contribute to the salient meaning components of individual words and across words.

The basis for the current investigation is provided by a collection of semantic associates evoked by German verbs and nouns. A series of analyses are performed on this database, to explore the relationships between the stimulus and the response words. Each analysis is motivated by its potential NLP uses, and the analyses are based on available resources with respect to the semantic investigation. As manually linking each stimulus-associate pair to a particular relationship would be time-intensive and subjective, we rely on large-scale lexicographic databases and on empirical, corpus-based resources that have the potential to characterise the associations.

Our work is in the line with recent discussions that relate the computational modelling of language to human data, cf. Daelemans (2006). I.e., we argue that language data as collected from human beings represents an excellent if not optimal source of information about language properties within the computational modelling of language, given that the data are gathered with materials and methods that are appropriate for the respective purpose.

2 Data Collection and Preparation

This section introduces our methods for collecting human associations to German verbs and nouns[2] and a distributional representation of the data as stimulus-associate type frequencies.

Associates of Verb Stimuli The data collection of associates to verb stimuli was performed as a web experiment, which asked native speakers to provide associations to German verbs. 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. The stimulus 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.

The experiment was administered over the Internet. Each trial consisted of a verb presented in a box at the top of the screen. 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.[3] Participants had 30 seconds per verb to type as many associations as they could.

299 native German speakers participated in the experiment, between 44 and 54 for each data set. In total, we collected 79,480 associate responses distributed over 39,254 different response types. Each trial elicited an average of 5.16 associate responses with a range of 0-16. Each completed data set contains the list of stimulus verbs, paired with a list of associations in the order in which the participant provided them.

Associates of Noun Stimuli The data collection of associates of noun stimuli was performed as an offline experiment, which asked native speakers to provide up to three associations to German nouns. 409 German nouns referring to picturable objects were chosen as target stimuli. To ensure broad coverage, target objects represented a variety of semantic classes including animals (e.g. *Affe* 'monkey', *Schwein* 'pig'), plants (e.g. *Tulpe* 'tulip', *Baum* 'tree'), professions (e.g. *Lehrerin* 'teacher', *Jäger* 'hunter'), furniture (e.g. *Stuhl* 'chair', *Bett* 'bed'), vehicles (e.g. *Flugzeug* 'plane', *Zug* 'train'), and tools (e.g. *Hammer* 'hammer', *Besen* 'broom'). The 409 target stimuli were divided randomly into three separate questionnaires consisting of approximately 135 nouns each. Each questionnaire was printed in two formats: target objects were either presented as pictures together with their preferred name (to ensure that associate responses were provided for the desired lexical item), or the name of

[2]The association norms for verbs and nouns were originally collected in independent studies; as a consequence they differ somewhat in the methods used for data collection.

[3]Despite these instructions, some participants failed to use capitalisation, leading to some ambiguity. Similarly, some participants provided multi-word expressions.

the target objects was presented without a representative picture accompanying it. Next to each target stimulus three lines were printed on which participants could write up to three semantic associate responses for the stimulus, one per line. The order of stimulus presentation was individually randomised for each participant. No time limits were given for responding, though participants were told to work swiftly and without interruption. Each version of the questionnaire was filled out by 50 participants, resulting in a maximum of 300 data points for any given target stimulus (50 participants $\times$ 2 presentation modes $\times$ 3 responses).

300 German participants, mostly students from Saarland University, received either course credit or monetary compensation for filling out the questionnaire. In total, we collected 116,714 associate responses distributed over 31,035 different response types. Collected associate responses were entered into a database with the following additional information: For each response type provided by a participant,[4] we coded a) the order of the response, i.e., first, second, third, b) the part-of-speech of the response, c) whether the response was related to the intended, depicted meaning of the stimulus or to an alternative meaning (in cases where the stimulus word was unambiguous) and d) the type of semantic relation between the target stimulus and the response (e.g., part-whole relations such as *car – wheel*, and categorical relationship such as hypernymy, hyponymy, and synonymy). The database is freely accessible (Melinger and Weber, 2006).

Distributional Representation For the analyses to follow, we pre-processed all data sets in the following way: For each stimulus word, we quantified over all responses in the experiment, disregarding the order in which associates were provided and, for noun stimuli, the presentation type of the questionnaire. The result is a frequency distribution for the stimulus words, providing frequencies for each response type. The responses were not distinguished according to polysemic senses of the stimuli. To illustrate the frequency distribution, Table 1 lists the 10 most frequent responses for the polysemous verb *klagen* 'complain, moan, sue' and Table 2 lists the 10 most frequent responses for the polysemous noun *Schloss* 'caste, lock'.

[4]As in the responses to the verb stimuli, there was some ambiguity because not all participants used capitalisation.

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 stimulus verb.

Schloss 'castle, lock'		
Schlüssel	'key'	51
Tür	'door'	15
Prinzessin	'princess'	8
Burg	'castle'	8
sicher	'safe'	7
Fahrrad	'bike'	7
schließen	'close'	7
Keller	'cellar'	7
König	'king'	7
Turm	'tower'	6

Table 2: Association frequencies for stimulus noun.

3 Resources for Data Investigation

This section introduces the manual and empirical resources that contributed to the characterisation of the association norms: a) a German newspaper corpus, and b) a statistical grammar model that was trained on the corpus data.

Corpus Data A German newspaper corpus from the 1990s was used for co-occurrence analyses between verb/noun stimuli and associate responses. The corpus contains approximately 200 million words of newspaper text from *Frankfurter Rundschau*, *Stuttgarter Zeitung*, *VDI-Nachrichten*, *die Tageszeitung*, *German Law Corpus*, *Donaukurier*, and *Computerzeitung*. In addition to the co-occurrence analyses, the corpus was used as training data for the statistical grammar model (see below).

Statistical Grammar Model Some of the quantitative data in the analyses to follow are derived from an empirical grammar model based on a German context-free grammar which paid specific attention to verb subcategorisation (Schulte im Walde, 2002). The grammar was lexicalised, and the parameters of the probabilistic version were estimated in an unsupervised training procedure, using 35 million words of the above German newspaper cor-

pus. The trained grammar model provides empirical frequencies for word forms, part-of-speech tags and lemmas, and quantitative information on lexicalised rules and syntax-semantics head-head co-occurrences.

4 Linguistic Analyses of Association Data

This section represents the main body of the article, providing a series of analyses that investigate step-wise the modelling of word meaning by empirical features: namely, a morpho-syntactic analysis, an analysis of the syntax-semantic functions of the noun (stimuli/associates) with respect to the verb (associates/stimuli), and a co-occurrence analysis of the stimuli-associate pairs. All of our analyses reported in this paper were based on response tokens; however, we also performed the respective type analyses, and they showed the same overall pictures. Each analysis is structured in the same way: first, we introduce the motivation from Natural Language Processing, discussing why the respective analysis is relevant for NLP purposes; second, we present the analyses; third, we interpret the analyses' results.

4.1 Morpho-Syntactic Analysis

The morpho-syntactic analyses of the response tokens distinguish and quantify the part-of-speech categories of the associate responses. On the one hand, this analysis can be considered as a preparatory step for the analyses to follow. In addition, the results will provide insight into the relevance of predominant part-of-speech categories with respect to meaning aspects. This knowledge is important in NLP tasks whenever words are represented by a choice of features that are supposed to model the word meaning, usually with the goal of determining the similarity or dissimilarity of words.

For example, the *vector space model* (Salton et al., 1975) uses words in documents to describe the contents of the respective documents. The model was originally designed for information retrieval (Salton and McGill, 1983), and has been generalised to describe not only documents, but also smaller structural units such as queries in question answering and individual words by co-occurring words. Often, the co-occurring words are restricted to content words, to certain part-of-speech categories, or even to a subset of words from a certain part-of-speech. With respect to a local perspective (i.e., co-occurrence within the near neighbourhood, such as the same sentence, or even the same phrase), the vector space model is related to the above mentioned *distributional hypothesis* and therefore the vector space model forms the basis for distributional descriptions.

Variants of the vector space model have been used in Latent Semantic Analysis for text indexing (Deerwester et al., 1990) and word similarity (Landauer and Dumais, 1997); in NLP tasks and applications including word sense discrimination (Schütze, 1998), anaphora resolution (Poesio et al., 2002), thesaurus extraction (Lin, 1999; McCarthy et al., 2003), and general models of semantic similarity (Lin, 1998; Sahlgren, 2006; Schulte im Walde, 2006; Padó and Lapata, 2007).

Associates of Verb Stimuli Each response to the stimulus verbs was assigned its – possibly ambiguous – part-of-speech *(POS)* by our empirical grammar dictionary. Originally, the dictionary distinguished approx. 50 morpho-syntactic categories, but we disregarded fine-grained distinctions such as case, number and gender features and considered only the major categories verb (V), noun (N), adjective (ADJ) and adverb (ADV). Having assigned part-of-speech tags to the responses, we were able to distinguish and quantify the morpho-syntactic categories of the responses' part-of-speech. The output of this analysis is the frequency distributions of the part-of-speech tags for each verb individually, and also as a sum over all verbs. Table 3 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%). The table also shows that the POS distributions vary across the semantic classes of the verbs. 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'.

Associates of Noun Stimuli In contrast to the analysis of the verb data, the part-of-speech categories of the associate responses to noun stimuli were hand-coded in the association database. The coding distinguished the three major categories verbs (V), nouns (N), adjectives (ADJ), and in addition proper names (PN). A fifth category 'OTHER' comprises all other part-of-speech categories such as particles, interjections (such as *igitt* 'ugh' for food nouns), numbers, and sounds (such as *wau-wau* 'woof-woof' for *Dackel* 'dachshund'). Thus,

	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%
bedrohen 'threaten'	12%	75%	12%	0%
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 3: POS distributions of verb responses.

unlike in the verb analysis, we directly specified the frequency distributions of the part-of-speech tags for each noun individually, and also as a sum over all nouns. Table 4 presents the total numbers and specific noun examples. As for the verb stimuli, participants provided noun associates in the clear majority of token instances, 69%; adjectives were given in 16% of the responses, verbs in 12%, and proper names in 3%. Again, the table also shows that the POS distributions vary with respect to the individual noun stimuli. For example, nouns referring to food or animals enforced a stronger usage of adjectives, such as *Ananas – gelb, süß, lecker* 'pineapple – yellow, sweet, tasty', or *Schildkröte – langsam, alt, grün* 'turtle – slow, old, green' than other nouns $t(407)=51.3$, $p<.001$. Similarly, nouns referring to natural objects evoked more adjectives, $t(407)=46.8$, $p<.001$, and fewer noun responses, $t(407)=6.5$, $p<.02$ than nouns referring to man-made objects.

	ADJ	N	PN	V
TOTAL FREQ	19,075	80,419	3,147	13,905
TOTAL PROB	16%	69%	3%	12%
Ananas 'pineapple'	45%	51%	3%	1%
Daumen 'thumb'	15%	71%	1%	11%
Esel 'donkey'	45%	42%	4%	6%
Löffel 'spoon'	6%	86%	0%	8%
Mund 'mouth'	11%	65%	0%	34%
Schildkröte 'turtle'	50%	44%	3%	3%
Tempel 'temple'	13%	58%	24%	5%
Telefon 'telephone'	4%	53%	2%	41%
Wecker 'alarm clock'	22%	42%	0%	36%
Zwiebel 'onion'	15%	54%	0%	31%

Table 4: POS distributions of noun responses.

Interpretation The morpho-syntactic analyses demonstrate that nouns play a major role among verb and noun features. This insight corresponds to the predominant use of nominal features in distributional descriptions that address the semantic modelling of words for various purposes. However, the analyses also showed that the relevance of the part-of-speech categories with respect to meaning aspects varies according to the semantic class of the word to model. We conclude that nouns are important for distributional descriptions, but other features than nouns should also be relevant in modelling word meaning. This insight should have an impact on the choice of feature categories in distributional representations; restricting the categories to nominal features restricts the feature sets to those features that are relevant for the average of words, but they do not necessarily cover the meaning aspects of all semantic word classes.

4.2 Syntax-Semantic Noun Functions

The analyses in this section continue exploring the eligibility of various types of features for modelling word meaning, now concentrating on the conceptual roles of nouns. As explained in the Introduction, most previous work on distributional similarity that used nominal features within distributional descriptions has either focused on a specific word-word relation to induce features (such as Pereira et al. (1993) and Rooth et al. (1999)), or used any dependency relation detected by the chunker or parser (Lin, 1998; McCarthy et al., 2003; Schulte im Walde, 2006). Little effort has been spent on investigating the eligibility of the various types of nominal features. Even though the use of the distributional features depends on the respective applications, we believe that we can identify prominent roles for distributional verb descriptions by evaluating which functional roles are highlighted by verb-noun pairs. For these analyses, we assume that the noun responses to verb stimuli and verb responses to noun stimuli relate to conceptual roles required by the verbs. Thus, we investigate the linguistic functions that are realised by the response nouns with respect to the stimulus verbs, and by the stimulus nouns with respect to the response verbs. The analyses are based on our empirical grammar model.

Associates of Verb Stimuli With respect to verb subcategorisation, the empirical grammar model offers 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 in the grammar model were *Brötchen* 'rolls', *Brot* 'bread', *Kuchen* 'cake', *Plätzchen* 'cookies', and *Waffel* 'waffle'. We used the grammar information to look up the syntactic relationships which existed between a stimulus verb and a response noun. For example, the nouns *Kuchen* 'cake', *Brot* 'bread', *Pizza* and *Mutter* 'mother' were produced in response to the stimulus verb *backen* 'bake'. The grammar look-up told us that *Kuchen* 'cake' and *Brot* 'bread' appeared not only as the verb's direct objects (as illustrated above), but also as intransitive subjects; *Pizza* only appeared as a direct object, and *Mutter* 'mother' only appeared as transitive subject. The verb-noun relationships which were found in the grammar were quantified by the verb-noun association frequency, taking into account the number and proportions of different relationships (to incorporate the ambiguity represented by multiple relationships). For example, the noun *Kuchen* was elicited 45 times in response to *bake*; the grammar contained the noun both as direct object and as intransitive subject for that verb. 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 accumulated the association frequency proportions with respect to a specific relationship, e.g., for the direct objects of *backen* 'bake' we summed over the frequency proportions for *Kuchen*, *Brot*, *Plätzchen*, *Brötchen*, etc. The final result was a frequency distribution over linguistic functions for each stimulus verb, i.e., for each verb we determined which linguistic functions were activated by how many noun associates. By generalising over all verbs, we discovered that only 10 frame-slot combinations were linked to at least 1% of the noun tokens: subjects in the intransitive frame and the transitive frame (with direct/indirect object, or prepositional phrase); 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 in Table 5; 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.

Function		Freq	Prob
S	**S** V	1,792	4%
	S V AO	1,040	2%
	S V DO	265	1%
	S V PP	575	1%
AO	S V **AO**	3,124	6%
	S V **AO** DO	824	2%
	S V **AO** PP	653	1%
DO	S V **DO**	268	1%
	S V AO **DO**	468	1%
PP	S V **PP-Dat:in**	487	1%
Total (of these 10)		9,496	19%
Total found in grammar		13,527	28%
Unknown verb or noun		10,964	22%
Unknown function		24,250	50%
Total V-N		48,741	100%

Table 5: Associates as nominal slot fillers.

Associates of Noun Stimuli Parallelling the preceding analysis, we checked whether any of the noun-verb relationships were found in our statistical grammar model. In the positive cases, the relationships were quantified by the noun-verb association frequency, again taking into account the number and proportions of the various grammar functions. The most prominent functions are listed in Table 6. The table shows that – to a large extent – the most prominent functions for the noun-verb pairs are the same as for the verb-noun pairs.

Interpretation In total, only 28/41% of all verb-noun pairs were identified by the statistical grammar as a filler for any slot in any of the 178 identified frames (which corresponds to a total of 592 frame-slot combinations). The majority of pairs was not found as slot fillers: 22/11% of the stimulus-associate pairs (marked as 'unknown verb or noun' in Tables 5 and 6) were missing because either the verb or the noun did not appear in the grammar model at all. These cases were due to (i) lemmatisation in the empirical grammar dictionary, where noun compounds such as *Autorennen* 'car racing' were lemmatised by their lexical heads, creating a mismatch between the full compound and its head; (ii) multi-word expressions among the associates, like *Zähne putzen* 'brush teeth' or *frisch machen* 'refresh'; (iii) domain of the training corpus, which underrepresented slang responses like *Grufties* 'old

Function		Freq	Prob
S	**S** V	1,095	8%
	S V AO	300	2%
	S V PP	406	3%
	S V C-2	103	1%
	S V INF	71	1%
AO	S V **AO**	1,480	11%
	S V **AO** DO	206	1%
	S V **AO** PP	218	2%
DO	S V **DO**	144	1%
	S V AO **DO**	99	1%
PP	S V **PP-Dat:auf**	263	2%
	S V **PP-Dat:in**	193	1%
Total (of these 12)		4,578	33%
Total found in grammar		5,661	41%
Unknown verb or noun		1,505	11%
Unknown function		6,712	48%
Total N-V		13,878	100%

Table 6: Stimuli as nominal slot fillers.

people' and *lümmeln* 'loll', dialect expressions such as *Ausstecherle* 'cookie-cutter' and *heimfahren* 'go home', as well as technical expressions such as *Plosiv* 'plosive'; and (iv) size of the corpus data: the whole newspaper corpus of 200 million words contained more than 99% of the stimuli and the associates in the two analyses; the 35 million word partition on which the grammar model was trained contained still more than 99% of the verb stimuli/associates, but only 78% of the noun associates to the verb stimuli, and only 90% of the noun stimuli.

The 50/48% of the nouns/verbs which are marked as 'unknown function' in Tables 5 and 6 were present in the grammar but did not fill subcategorised-for linguistic functions; clearly the conceptual roles of the noun associates were not restricted to the subcategorisation of the stimulus verbs.

Although direct object and subject roles are prominent among the verb-noun relationships, they are also highly frequent in the grammar model as a whole. In fact, across all possible frame-slot combinations, we find an extremely strong correlation between the frequency of a frame-slot combination in the grammar model and the number of responses that link to that frame-slot combination in our data, $r(592)=.925$, $p<.001$ for the noun responses to verbs, and $r(592)=.854$, $p<.001$ for the verb responses to nouns. Thus, the direct object and subject roles are not over-represented in our data; they are represented proportionate to their frequency in the grammar. Therefore, we cannot conclude from the tables that specific functions within distributional representations are dominant and should be recommended.

Furthermore, contrary to our initial assumptions, the majority of nouns in verb-noun pairs did not reflect conceptual roles for the respective verbs. In part what was or was 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 were therefore found as linguistic functions. Adjuncts such as the instrument *Pinsel* 'brush' for *bemalen* 'paint', *Pfanne* 'pan' for *erhitzen* 'heat', or clause-internal information such as *Aufmerksamkeit* 'attention' for *bemerken* 'notice' and *Musik* 'music' for *feiern* 'celebrate' were not found. Similarly, verbs provided as associates for their respective instruments, e.g. *trocknen* 'dry' for *Handtuch* 'towel', *biegen* 'bend' for *Zange* 'pincer', or providing world knowledge, e.g. *streichen* 'paint' for *Klebeband* 'tape', *schlafen* 'sleep' for *kissen* 'cushion', *riechen* 'smell' for *Nase* 'nose' were also not found. These nouns fulfil scene-related roles or represent world knowledge, and were not captured by subcategorisation in the grammar model. The analyses therefore illustrated that the noun stimuli/responses were not restricted to verb subcategorisation role fillers, and that clause-internal adjuncts as well as clause-external, scene-related information or world knowledge should also play a role when using nominal features in distributional descriptions of word meaning.

4.3 Co-Occurrence Analysis

The motivation for the last set of analyses on word meaning features arose from our syntax-semantics analyses in the previous section, which demonstrated that there were verb-noun pairs within the association norms which might co-occur in local contexts even if they were not related by a subcategorisation function. In more general terms, we were interested in the role of co-occurrence information within an empirical distributions description. It is commonly assumed that human associations reflect word co-occurrence probabilities, cf. (McKoon and Ratcliff, 1992; Plaut, 1995); this assumption was supported by observed correlations between associative strength and word co-occurrence in language corpora (Spence and Owens, 1990). Our analyses examined whether the co-occurrence assumption holds for our (much larger) German association

data, i.e., which proportion of the associations were found in co-occurrence with the stimulus words. A positive outcome of these analyses might encourage the use of low-level co-occurrence information in corpus-based word descriptions.

Associates of Verb Stimuli The analysis used our complete newspaper corpus, 200 million words, and checked whether the associate responses occurred in a window of 20 words to the left or to the right of the relevant stimulus word. We determined the co-occurrence strength between the stimulus verbs and their associations. The results are presented in Table 7. The 'all' row shows the percentage of associate responses that were found in co-occurrence with their stimulus verbs just once, or twice, or 3/5/10/20/50 times. The co-occurrence proportions are rather high, especially when taking into account the restricted domain of the corpus. For example, for a co-occurrence strength of 3 we find two thirds of the associations covered by the 20-word window in the corpus data. The following rows are specified for their POS, verbs 'V', nouns 'N', adjectives 'ADJ', and adverbs 'ADV'. The proportions of verb, noun and adjectives responses which were found in co-occurrence with their stimulus verbs are very similar to the overall proportions. The 'ADV' co-occurrence strengths stand out in Table 7: they represent only 2% of all response tokens, but the analysis shows they exhibit a much stronger co-occurrence behaviour to the verbs than the other POS.

	Co-Occurrence Strength						
POS	1	2	3	5	10	20	50
all	77	70	66	59	50	40	27
V	79	71	67	60	50	41	29
N	76	69	66	59	50	40	27
ADJ	77	69	64	57	45	36	22
ADV	91	88	85	80	72	62	50

Table 7: Verb-association co-occurrence.

Associates of Noun Stimuli The co-occurrence analysis for the associates of noun stimuli was conducted exactly as for the verbs. Table 8 presents the results. Again, the proportions of verb, noun and adjectives responses which were found in co-occurrence with their stimulus nouns are very similar to the overall proportions, with the verb proportions slightly above, and the adjective proportions slightly below the overall co-occurrence values. Furthermore, all co-occurrence values are between 6-9% above the co-occurrence values of the verb analysis.

	Co-Occurrence Strength						
POS	1	2	3	5	10	20	50
all	84	77	72	64	52	38	23
V	88	82	77	69	57	44	28
N	84	78	72	65	53	39	23
ADJ	83	76	70	63	50	36	20

Table 8: Noun-association co-occurrence.

Interpretation Our analyses showed that the co-occurrence assumption holds for our German association data, to a large extent: 77/84% of our response tokens were covered at least once in a 20-word window of the stimulus words, approximately two thirds were covered at least three times, and even approximately 40% were covered at least 20 times. These results suggest that co-occurrence information is an integral component for empirical descriptions of word properties, an important insight since co-occurrence information is essentially less expensive (because no high-level pre-processing is necessary) and therefore easier to obtain than annotated data. Thus co-occurrence information could be especially valuable for languages with few NLP resources available.

Furthermore, comparing the co-occurrence strength of nominal responses with the proportions of the nouns that were found as subcategorised by the respective verbs (cf. Tables 5 and 6) demonstrates once more that verb subcategorisation accounts only for a part of the nominal responses, and therefore only for a subset of the verb concepts represented by nouns; but more general scene-related information beyond the clause level is captured by corpus co-occurrence.[5]

Examples of associations that did not appear in co-occurrence with the respective stimulus verbs are *nass* 'wet' for *nieseln* 'drizzle', *lecker* 'yummy' for *mampfen* 'munch', *Trockner* 'dryer' for *trocknen* 'dry', *Wasser* 'water' for *auftauen* 'defrost', *Freude* 'joy' for *überraschen* 'surprise', or *Verantwortung* 'responsibility' for *leiten* 'guide'. Correspondingly, examples of associations that did not appear in co-occurrence with the respective stimulus nouns are *gelb* 'yellow ' for *Ananas* 'pineapple',

[5]Note, however, that the 28/41% subcategorised nouns can only be compared indirectly with the 76/88% co-occurring nouns/verbs, because the former rely on only 35 million of the 200 million word corpus.

kalt 'cold' for *Iglu* 'igloo', *Überraschung* 'surprise' for *Geschenk* 'present', *Weihnachten* 'Christmas' for *Walnuß* 'walnut', *Physik* 'physics' for *Magnet* 'magnet', and *Herbst* 'autumn' for *Drachen* 'kite'. These associations reflect world knowledge rather than clause-internal/-external scene-related information, and are therefore not expected to be found in the immediate context of the stimuli at all. These cases pose an interesting challenge to empirical models of word meaning: It is not surprising that world knowledge is not necessarily represented in corpus data, but the association analyses illustrated that, as a consequence, empirical features that model world knowledge are missing in distributional word meaning descriptions.

Finally, comparing the overall co-occurrence strength of associates with those of specific part-of-speech categories demonstrates that the co-occurrence information for some categories is more easily available than for others. For example, the verb association analysis showed that adverbs play a major role for verbs in the corpus proximity. This is an important insight: adverbs are a closed-class POS and restricted in number, and therefore easy to cover empirically, and at the same time they are successful in capturing verb meaning aspects.

5 Summary and Conclusions

This article presented a study to identify, distinguish and quantify the various types of semantic associations provided by humans, and to illustrate their usage for NLP purposes. We investigated the morpho-syntactic categories and the contextual functions that are represented by the associates with respect to the experiment stimuli. We demonstrated that nouns play a major role among the content word categories; this finding supports the predominant usage of noun features in distributional word representations. In addition, we showed that there is an extremely strong correlation between the frame-slot combinations in a grammar model and frame-slot combinations activated by our data; no linguistic functions are strongly over- or under-represented and could therefore be considered to be prominent to represent conceptual nominal roles for verbs. A final analysis illustrated that clearly the noun stimuli/associations are not restricted to verb subcategorisation role fillers, and that clause-internal adjuncts as well as clause-external, scene-related information or world knowledge should also play a role as features: we showed that the co-occurrence assumption holds for our German association data, to a large extent. These results suggest co-occurrence information for an appropriate usage in empirical descriptions of word properties, an important insight since co-occurrence information is essentially less expensive (because no high-level pre-processing such as parsing is necessary), and therefore easier to obtain – especially in languages with few NLP resources available - than annotated data.

In conclusion, we believe that the association norms have contributed to the understanding of distributional semantic descriptions in computational linguistics. Even though the data represent a collection of word-word associations on a limited scale, they have proven useful to get insight into the computational modelling of words and word features. There is even more potential within the norms, which e.g. will allow us to address representational and distributional requirements with respect to the modelling of polysemy in future work.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics*, pages 86–90, Montreal, Canada.

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An Empirical Model of Multiword Expression Decomposability. In *Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 89–96, Sapporo, Japan.

James Curran. 2003. *From Distributional to Semantic Similarity*. Ph.D. thesis, Institute for Communicating and Collaborative Systems, School of Informatics, University of Edinburgh.

Walter Daelemans. 2006. A Mission for Computation Natural Language Learning. In *Proceedings of the 10th Conference on Computational Natural Language Learning*, pages 1–5, New York City, NY.

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*, 41(6):391–407.

Zellig Harris. 1968. Distributional Structure. In Jerold J. Katz, editor, *The Philosophy of Linguistics*, Oxford Readings in Philosophy, pages 26–47. Oxford University Press.

Thomas K. Landauer and Susan T. Dumais. 1997. A Solution to Plato's Problem: the Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. *Psychological Review*, 104(2):211–240.

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

Dekang Lin. 1999. Automatic Identification of Non-compositional Phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 317–324, Maryland, MD.

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*, Sapporo, Japan.

Gail McKoon and Roger Ratcliff. 1992. Spreading Activation versus Compound Cue Accounts of Priming: Mediated Priming Revisited. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 18:1155–1172.

Alissa Melinger and Andrea Weber. 2006. Database of Noun Associations for German. URL: www.coli.uni-saarland.de/projects/nag/.

Paola Merlo and Suzanne Stevenson. 2001. Automatic Verb Classification Based on Statistical Distributions of Argument Structure. *Computational Linguistics*, 27(3):373–408.

Sebastian Padó and Mirella Lapata. 2007. Dependency-based Construction of Semantic Space Models. *Computational Linguistics*, 33(2). To appear.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated Resource of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

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*, pages 183–190, Columbus, OH.

David C. Plaut. 1995. Semantic and Associative Priming in a Distributed Attractor Network. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, volume 17, pages 37–42.

Massimo Poesio, Tomonori Ishikawa, Sabine Schulte im Walde, and Renata Viera. 2002. Acquiring Lexical Knowledge for Anaphora Resolution. In *Proceedings of the 3rd Conference on Language Resources and Evaluation*, volume IV, pages 1220–1224, Las Palmas de Gran Canaria, Spain.

Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. 1999. Inducing a Semantically Annotated Lexicon via EM-Based Clustering. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, Maryland, MD.

Magnus Sahlgren. 2006. *The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-Dimensional Vector Spaces*. Ph.D. thesis, Stockholm University.

Gerard Salton and Michael McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11):613–620.

Sabine Schulte im Walde. 2002. A Subcategorisation Lexicon for German Verbs induced from a Lexicalised PCFG. In *Proceedings of the 3rd Conference on Language Resources and Evaluation*, volume IV, pages 1351–1357, Las Palmas de Gran Canaria, Spain.

Sabine Schulte im Walde. 2006. Experiments on the Automatic Induction of German Semantic Verb Classes. *Computational Linguistics*, 32(2):159–194.

Hinrich Schütze. 1998. Automatic Word Sense Discrimination. *Computational Linguistics*, pages 97–123. Special Issue on Word Sense Disambiguation.

Donald P. Spence and Kimberly C. Owens. 1990. Lexical Co-Occurrence and Association Strength. *Journal of Psycholinguistic Research*, 19:317–330.