Research on Language and Computation **00:** 1–36, 2008.

Human Associations and the Choice of Features for Semantic Verb Classification**

SABINE SCHULTE IM WALDE (schulte@ims.uni-stuttgart.de)
Institute for Natural Language Processing,
University of Stuttgart, Germany

February 8, 2008

Abstract. This article investigates whether human associations to verbs as collected in a web experiment can help us to identify salient features for semantic verb classes. Starting from the assumption that the associations, i.e., the words that are called to mind by the stimulus verbs, reflect highly salient linguistic and conceptual features of the verbs, we apply a cluster analysis to the verbs, based on the associations, and validate the resulting verb classes against standard approaches to semantic verb classes. Then, we perform various clusterings on the same verbs using standard corpus-based feature types, and evaluate them against the association-based clustering as well as GermaNet and FrameNet classes. Comparing the cluster analyses provides an insight into the usefulness of standard feature types in verb clustering, and assesses shallow vs. deep syntactic features, and the role of corpus frequency. We show that (a) there is no significant preference for using a specific syntactic relationship (such as direct objects) as nominal features in clustering; (b) that simple window co-occurrence features are not significantly worse (and in some cases even better) than selected grammar-based functions; and (c) that a restricted feature choice disregarding high- and low-frequency features is sufficient. Finally, by applying the feature choices to GermaNet and FrameNet verbs and classes, we address the question of whether the same types of features are salient for different types of semantic verb classes. The variation of the gold standard classifications demonstrates that the clustering results are significantly different, even when relying on the same features.

Key words: semantic verb classes, distributional features, clustering experiments, association norms

1. Motivation

In recent years, the computational linguistics community has developed an impressive number of *semantic verb classifications*, i.e., classifications that generalise over verbs according to their semantic properties. Intuitive examples of such classifications are the MOTION WITH A VEHICLE class, including verbs such as *drive*, *fly*, *row*, etc., or the BREAK A SOLID SURFACE WITH AN INSTRUMENT class, including verbs such as *break*, *crush*, *fracture*, *smash*, etc. Semantic verb classifications are of great interest to computational linguistics, specifically regarding the pervasive problem of data sparseness in the processing of natural language. Up to now, such classifications have been used in applications such as word sense disambiguation (Dorr and Jones, 1996; Kohomban and Lee, 2005),

** The original publication is available at www.springerlink.com, doi 10.1007/s11168-008-9044-8.

machine translation (Dorr, 1997; Prescher et al., 2000; Koehn and Hoang, 2007), document classification (Klavans and Kan, 1998), statistical lexical acquisition in general (Rooth et al., 1999; Merlo and Stevenson, 2001; Korhonen, 2002; Schulte im Walde, 2006b), and also in psycholinguistic models of human sentence processing (Padó et al., 2006).

Given that the creation of semantic verb classifications is not an end task in itself, but depends on the application scenario of the classification, it is obvious that the goals, the strategies and, accordingly, the results of the creation process vary to a large degree. Taking English as an example, major frameworks are the Levin classes (Levin, 1993), *WordNet* (Fellbaum, 1998), and *FrameNet* (Fillmore et al., 2003), which embody different instantiations of semantic groupings: Levin's classification refers to verb similarity with respect to the verbs' syntax-semantic alternation behaviour, WordNet uses synonymy, and FrameNet relies on situation-based agreement as defined by Fillmore's frame semantics (Fillmore, 1982). These various instantiations of semantic relatedness naturally lead to different semantic verb classes, as the following example illustrates. In the Levin class of GET verbs, a sub-class of OBTAINING verbs, the English verb *buy* is assigned to the same class as the verbs *catch, earn, find, steal,* etc., mainly because all verbs participate in the benefactive alternation. WordNet assigns this sense of the verb *buy* to the same class as its near-synonyms *purchase* and *take*, as a sub-class of GET/ACQUIRE verbs. And FrameNet assigns the verb *buy* to the frame COMMERCE/BUY, also together with the verb *purchase*, because both verbs describe a commercial transaction involving a buyer and a seller exchanging money and goods.

As an alternative to resource-intensive manual classifications, automatic methods such as classification and clustering approaches have been applied to induce semantic verb classes from corpus data, e.g., Siegel and McKeown (2000), Merlo and Stevenson (2001), Korhonen et al. (2003), Ferrer (2004), Schulte im Walde (2006b), Joanis et al. (2008). Depending on the types of verb classes to be induced, the techniques vary their choice of verbs and classification/clustering algorithm. However, another central parameter for the automatic induction of semantic verb classes is the selection of verb features. A priori (i.e., without any kind of semantic pre-processing), the lexical acquisition of semantic features from corpus data is not trivial, and few resources are semantically annotated and provide semantic information off-the-shelf (such as *FrameNet* (Fillmore et al., 2003) and *PropBank* (Palmer et al., 2005)). Therefore, the automatic construction of semantic classes typically benefits from a long-standing linguistic hypothesis which asserts a tight connection between the lexical meaning of a verb and its distributional behaviour (Pinker, 1989; Levin, 1993; Dorr and Jones, 1996; Siegel and McKeown, 2000; Merlo and Stevenson, 2001; Lapata and Brew, 2004; Schulte im Walde, 2006b; Joanis et al., 2008). Even though the meaning-behaviour relationship is not perfect, various automatic approaches have demonstrated that a classification based on verb behaviour actually shows substantial agreement with a semantic

classification. The verb behaviour itself is commonly captured by 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). The specific features within the distributional descriptions vary according to the target classification, ranging over e.g. morpho-syntactic cues such as active vs. passive constructions, subcategorisation frames and selectional preferences, word co-occurrence (typically with respect to the syntactic structure, such as direct objects), and corpus-based linguistic heuristics (such as the usage of personal pronouns as indicators of agentivity in semantic roles). As target classifications, the automatic approaches commonly use existing classifications (Dorr and Jones, 1996; Korhonen et al., 2003; Lapata and Brew, 2004; Joanis et al., 2008); or, alternatively, they rely on their own gold standard classes or criteria (Siegel and McKeown, 2000; Merlo and Stevenson, 2001; Schulte im Walde, 2006b).

In sum, corpus-based, distributional models of verb behaviour have proven useful within classification and clustering approaches towards semantic verb classes. Nevertheless, this article suggests that there are (at least) two issues which require further consideration.

1. The first issue concerns the *types of distributional features that are considered in automatic approaches to semantic verb classifications*. Any selection of features is expected to refer to some aspects of verb meaning. With respect to very specific types of verb classifications,[1] the feature choice is straightforward to a certain extent. However, when it comes to larger-scale classifications with several hundreds of verbs and a variety of semantic classes, e.g., Korhonen et al. (2003), Schulte im Walde (2006b), Joanis et al. (2008), an appropriate choice of verb features that correlate with aspects of verb meaning seems less obvious. Some features (such as subcategorisation frame types) have proven useful and some features (such as selectional preferences) have proven unreliable across various target classifications. What is missing is a general instrument to suggest and evaluate the semantic appropriateness of features. This article proposes *association norms* as one such instrument: Association norms – collections of words that are called to mind by a set of stimulus words – have a long tradition in psycholinguistic research, where they have been used for more than 30 years to investigate semantic memory, making use of the implicit notion that associates reflect aspects of word meaning (Tanenhaus et al., 1979; McKoon and Ratcliff, 1992; Plaut, 1995; McRae and Boisvert, 1998, among others). Given that the meaning aspects of verbs are exactly what underlies any semantic classification of verbs, we take advantage of this long-standing notion: We exploit a collection of associations to check the salience of previously suggested features.[2] The first question posed in this article is therefore: *Do human associations help identify salient features for inducing semantic verb classes?* Of course, we do not assume that there is an overall optimal set of verb features in automatic semantic verb classification (which would in fact be counter-intuitive to our second question, to

follow). The goal is rather to determine whether association norms represent an appropriate source of information for aspects of meaning that are generally applicable to semantic verb classification.

2. The second issue concerns the *relationship between the target classification and the chosen features*: the choice of features to model verb meaning depends on the type of target classification. For example, if the target classification of the automatic induction process is Levin-style classes, the verb features should refer to aspects of alternation behaviour; if the target classification is FrameNet-style classes, the verb features need to capture various kinds of situation-based relatedness, such as synonymy, converses (i.e., a sub-type of antonymy), causality, etc. With exceptions (cf. footnote 1), though, features have been developed on a general basis. So far, no previous work has specified these general features with respect to various target classifications, or evaluated an induced classification result against various gold standard classifications, rather than against one individual, pre-determined gold standard. The second question posed in this article is therefore: *Are the same types of features salient for different types of semantic verb classes?*

Guided by our two questions, this article is organised as follows. As the basis for this investigation, we present a collection of semantic associations to German verbs (Section 2), complemented by various analyses of their empirical properties. In a preparatory step, we perform an unsupervised clustering on the experiment stimulus verbs, based on the verb associations, and we validate the resulting verb classes as a reference set of semantic classes for German verbs by demonstrating that they show considerable overlap with standard approaches to semantic verb classes, i.e., GermaNet and FrameNet (Section 3). In the main body of this work, we perform an analysis of the empirical properties of the verb associations, and apply these insights to the selection of feature types for semantic verb classifications (Section 4). The analysis allows insights into the usefulness of standard feature types in verb clustering (such as direct objects), and an assessment of shallow window co-occurrence features vs. deeper information using syntactic frame fillers. In addition, we vary the corpus-based features with respect to their corpus frequency to determine the influence of the feature frequency within the cluster analyses. Finally, by applying the feature choices not only to our association-based reference set but also to GermaNet and FrameNet, we address the question of whether the same types of features are salient for different types of semantic verb classes.

2. Human Verb Associations

In general, association norms are collected by presenting *target stimuli* to the participants in an experiment, who then provide *associate responses*, i.e., words that are called to mind by the stimulus words. As introduced in the previous section, association norms have a long tradition in psycholinguistic research. One of the first

collections of word association norms was done by Palermo and Jenkins (1964), comprising associations for 200 English words. The *Edinburgh Association Thesaurus* (Kiss et al., 1973) was a first attempt to collect association norms on a larger scale, and also to create a network of stimuli and associates, starting from a small set of stimuli derived from the Palermo and Jenkins norms. Researchers at the University of South Florida compiled association norms over the course of more than 20 years, from 1973 (Nelson et al., 1998). Their goal was to obtain the "largest database of free associations ever collected in the United States available to interested researchers and scholars". Smaller sets of association norms have also been collected for example in Dutch (Lauteslager et al., 1986), French (Ferrand and Alario, 1998) and Spanish (Fernández et al., 2004) as well as for different populations of speakers, such as adults vs. children (Hirsh and Tree, 2001). Last but not least, there is a small-scale collection for German (Russell and Meseck, 1959; Russell, 1970), based on 100 stimulus words across part-of-speech. The collection has been used in closely related work to ours, by Reinhard Rapp, Manfred Wettler and colleagues (see details in Section 5).

This section introduces the association norms that are used in the course of this article. The data collection was performed as a web experiment,[3] which asked native speakers to provide associations to German verbs. Details of the method for collecting the associations are described in Section 2.1, and a series of empirical linguistic analyses of the data are described in Section 2.2.

2.1. DATA COLLECTION

2.1.1. *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. Selecting verbs from different categories was only intended 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. Appendix A provides two example classes, accompanied by their choice of verbs.

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.

2.1.2. *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.[4] Participants had 30 seconds per verb to type as many associations as they could. After this time limit, the program automatically advanced to the next verb.

2.1.3. *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 79,480 associations from 16,445 trials; each trial elicited an average of 5.16 associate responses with a range of 0-16.

2.1.4. *Data Preparation*

Each completed data set contains the background information of the participant, followed by the list of stimulus verbs. Each stimulus is paired with a list of associations in the order in which the participant provided them. For the analyses to follow, we pre-processed all data sets in the following way: For each stimulus verb, we quantified over all responses in the experiment, disregarding the participant's background and the order of the associates. Table I lists the 10 most frequent responses for the polysemous verb *klagen* 'complain, moan, sue'. The verb responses were not distinguished according to polysemic senses of the verbs.

2.2. EMPIRICAL ANALYSES OF VERB ASSOCIATIONS

The associations to the verbs were investigated on several linguistic dimensions (Schulte im Walde and Melinger, 2005). In this section we only repeat those analyses which we consider to be relevant with respect to an automatic semantic classification:

1. The associations were distinguished with respect to the major parts-of-speech: nouns, verbs, adjectives, adverbs.

2. For each noun associate, we investigated the kinds of linguistic functions that were realised by the noun with respect to the stimulus verb (e.g., subject, direct objects, etc.).

Table I. Association frequencies for example stimulus.

Stimulus: *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

3. The co-occurrence strengths of the stimulus verbs and their associations were determined using a 200 million word corpus of German newspaper text.

After a brief introduction of the empirical grammar model which underlies a part of the analyses, Sections 2.2.2 to 2.2.4 describe the motivations for these three analyses in more detail, and then present the actual analyses.

2.2.1. *Excursus: Empirical Grammar Model*

Some of the quantitative data in the analyses to follow were derived from an empirical grammar model (Schulte im Walde, 2003, chapter 3): we developed a German context-free grammar paying specific attention to 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, part-of-speech tags and lemmas, and quantitative information on lexicalised rules and syntax-semantics head-head co-occurrences.

2.2.2. *Morpho-Syntactic Analysis*

In the morpho-syntactic analysis, each association of the stimulus verbs was assigned its – possibly ambiguous – part-of-speech by our empirical grammar dictionary, cf. Section 2.2.1. 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). Ambiguities between these categories arose 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 were ambiguous between multiple part-of-speech tags.

Having assigned part-of-speech tags to the associations, we were able to distinguish and quantify the morpho-syntactic categories of the responses. In non-ambiguous situations, the unique part-of-speech received the total stimulus-response frequency; in ambiguous situations, the stimulus-response frequency was split uniformly over the possible part-of-speech tags. As the result of this first analysis, we could specify the frequency distributions of the part-of-speech tags for each verb individually, and also as a sum over all verbs. Table II 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%).[5] The table also shows that the part-of-speech 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'.

Table II. Part-of-speech tag distributions.

	V	N	ADJ	ADV
TOTAL FREQ	19,863	48,905	8,510	1,268
TOTAL PROP	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%

2.2.3. *Syntax-Semantic Noun Functions*

In a second step, we investigated the kinds of linguistic functions that were realised by noun associates in response to stimulus verbs. For this analysis, we assume that the noun responses to verb 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, based on our empirical grammar model, cf. Section 2.2.1. The motivation for this analysis was to identify those nominal functions that might be relevant verb features within a distributional description of verb properties. Most previous work on the automatic induction of semantic verb classes – and on distributional similarity in more general terms – that relied on nominal features as distributional verb properties has either focused on a specific word-word relation (such as Pereira et al. (1993), Rooth et al. (1999) referring to a direct object noun for describing verbs), or used any dependency relation detected by the chunker or parser (such as Lin (1998), McCarthy et al. (2003), Korhonen et al. (2003), Schulte im Walde (2006b)). Little effort has been spent on investigating the salience of the various nominal types of verb features.

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', *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 could determine which linguistic functions were activated by how many noun associates. For example, the most prominent functions for the inchoative-causative verb *backen* 'bake' were the transitive direct object (8%), the intransitive subject (7%) and the transitive subject (4%); for the object-drop *schreiben* 'write' we found 11% for the direct object, 3% and 4% for

the intransitive and the transitive subject, respectively, and evidence for the writing instrument (the PP headed by *mit* 'with' in various frames with a total of 10%).

By generalising over all verbs, we discovered that only 10 frame-slot combinations were activated by at least 1% of the noun tokens: subjects in the intransitive frame and the transitive frame (with accusative/dative object, or prepositional phrase); the accusative object slot in the transitive, the ditransitive frame and the direct object plus PP frame; the dative object in a transitive and ditransitive frame, and the prepositional phrase headed by *Dat:in*, dative (locative) 'in'. The frequencies and proportions are illustrated in Table III; 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 object, 'DO' a dative object, and 'PP' a prepositional phrase. Although accusative 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 found 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, $r(592)=.87$, $p<.001$. Thus, the accusative object and subject roles are not over-represented in our data; they are represented proportionate to their frequency in the grammar. Therefore, the tables do not allow us to conclude that specific functions within distributional representations are dominant.

Table III. Associates as slot fillers.

	Function	Freq	Prop
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%

In total, only 28% of all noun associates were identified by the statistical grammar as frame-slots fillers. The majority of noun responses were not found as slot

fillers: 22% of the associates (marked as 'unknown verb or noun' in Table III) 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) domain of the training corpus, which underrepresented slang responses like *Grufties* 'old people', dialect expressions such as *Ausstecherle* 'cookie-cutter' as well as technical expressions such as *Plosiv* 'plosive'; and (iii) size of the corpus data: the whole newspaper corpus of 200 million words contained 99.4% of the noun association tokens, but the 35 million word partition on which the grammar model was trained contained only 78% of them. The remaining 50% of the nouns (marked as 'unknown function' in Table III) were present in the grammar but did not fill subcategorised-for linguistic functions with respect to the stimulus verbs; clearly the conceptual roles of the noun associates were not restricted to the subcategorisation of the stimulus 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 adverbials such as *Aufmerksamkeit* 'attention' for *bemerken* 'notice' and *Musik* 'music' for *feiern* 'celebrate' were not found. These associates were not captured by the subcategorisation information in the grammar model.

2.2.4. *Co-Occurrence Analysis*

In a third analysis, we determined the co-occurrence strength between the stimulus verbs and their associations. The motivation for this analysis partly came from our syntax-semantics analysis in the previous section, which demonstrated that there were verb-association pairs in local contexts even if they were not related by a subcategorisation function. In addition, 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 analysis examined whether the co-occurrence assumption holds for our German association data, i.e., which proportion of the associations were found in co-occurrence with the stimulus verbs. The analysis used our complete newspaper corpus, 200 million words, and checked whether the response verbs occurred in a window of 20 words to the left or to the right of the relevant stimulus word.[6]

Table IV presents the results of the co-occurrence analysis. The 'all' row shows the percentage of associations 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. In comparison, the co-occurrence proportions of the same verbs with unrelated words (with parts-of-speech and corpus frequencies identical to those of the associations) are 30-40% below the values in Table IV. See Schulte im Walde and Melinger (2008) for an in-depth look into the interpretation of stimulus-associate co-occurrence conditions and interpretations.

Table IV. Verb-association co-occurrence in 20-word window.

	Co-Occurrence Strength						
	1	2	3	5	10	20	50
all	77	70	66	59	50	40	27
N	76	70	66	59	50	40	27
ADV	91	88	85	80	72	62	50

The 'N' row shows the same information as the 'all' row, but is specified for the noun associations. The proportions of noun associations which were found in co-occurrence with their stimulus verbs are almost identical to the overall proportions. Comparing these numbers with the 28% of the nouns that were found as subcategorised by the respective verbs (cf. Table III) demonstrates once more that verb subcategorisation accounts only for a part of the noun associations.[7] Examples of associations that do not appear in co-occurrence with the respective stimulus verbs are *Wasser* 'water' for *auftauen* 'defrost', *Freude* 'joy' for *überraschen* 'surprise', or *Verantwortung* 'responsibility' for *leiten* 'guide'. These associations reflect world knowledge and are therefore not expected to be found in the immediate context of the verbs at all.

Finally, the 'ADV' row in Table IV lists the co-occurrence values of the stimulus verbs and the response adverbs. Even though the adverbs represent only a proportion of 2% of all response tokens, the co-occurrence analysis shows that they play a major role in the corpus proximity. One should keep in mind, though, that there is a high prior probability of finding one or more adverbs in the vicinity of a verb, and that adverbs that appear in a large corpus distance from a verb are not very likely to contribute to the meaning of the verb, but rather to the meaning of the verb in the respective clause.

2.3. SUMMARY

In this section we presented a choice of analyses of the human verb associations that we consider to be potentially helpful in providing an insight into the linguistic

and conceptual features of distributional verb descriptions in semantic verb classification. The morpho-syntactic analysis demonstrated that nouns play a major role among the associates. 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 underrepresented and could therefore be considered a prominent representative of conceptual nominal roles for verbs. The analysis also illustrated that the noun associations are not restricted to verb subcategorisation role fillers, and that clause-internal adjuncts as well as clause-external information might also play a role as verb features. The co-occurrence analysis confirmed this assumption; a context window of 20 words captured two thirds of all noun associations with a co-occurrence strength of 3. These results generalise over the part-of-speech types; for adverbs we even find co-occurrence values up to 90%. With respect to a distributional feature description of verbs, this latter analysis suggests that window-based word features contribute to verb descriptions. This is interesting, since the window approach has largely been disregarded in recent years, in comparison to using syntactic functions. Furthermore, adverbs – which have rarely been used in distributional verb description – should be included.

We close this section with a number of remarks on the analyses. The remarks are not necessary for the reader to understand the remainder of this article, but rather to comment on obvious questions that could arise from the analyses.

1. There are, of course, more aspects of the verb associations than those covered by our analyses, and there are more resources that could be used for such analyses. Our choice of resources and analyses was related to a) which features were taken into account in existing work on semantic verb classes, and b) how these features could be improved.

2. As mentioned in Section 2.2.2, the results of the analyses vary with respect to the individual verbs, the corpus frequencies of the verbs, and the semantic classes of the verbs. For example, the part-of-speech distribution for response words was correlated with stimulus verb frequency. The rate of verb and adverb responses was positively correlated with stimulus 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 was inversely correlated with verb frequency, Pearson's $r(328)=-.155$, $p<.005$ for nouns and $r(328)=.114$, $p<.05$ for adjectives. With respect to the semantic classes of verbs, 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'.

   Similar correlations appear in the other analyses. Therefore, generalising the analysis results over all verbs represents an average over the individual results. If one is interested in semantic features of individual verb classes, the respective analyses should be performed on a per-class basis.

3. Finally, the reader should note that our analyses were strongly influenced by the corpus properties and the properties of the grammar model. For example, the syntax-semantics function analysis could only match verb-association pairs to verb-noun functions in the grammar model if the words were in the corpus and the functions in the grammar. However, we believe that our analyses are sufficiently general for an investigation and comparison of features in distributional verb descriptions.

3. Association-based Verb Classes

This section is closely connected to the central assumption of this article.[8] Recall from our motivation that – based on the respective work in psycholinguistics – we assume that human associations to verbs model salient aspects of the verbs' meaning, and that human associations should therefore represent an excellent choice of features for semantic verb classes. Relying on these assumptions, we perform a cluster analysis of the 330 German verbs from the web experiments, based on their associations, in Section 3.1. The result is suggested as a reference classification of the German verbs, with respect to the feature exploration and variation in the clustering experiments to follow in Section 4. In order to justify the association-based clustering as a reference set, Section 3.2 validates the classification against standard approaches to semantic verb classes, i.e., *GermaNet* as the German WordNet (Kunze, 2000), and the German counterpart of FrameNet in the *Salsa* project (Erk et al., 2003).

3.1. ASSOCIATION-BASED CLUSTERING

Using the associations as verb features within the clustering process assumes that the associations point to meaning aspects of the verbs. Thus, verbs which are semantically related to each other tend to have similar associations, and are therefore expected to be assigned to common classes. Table V illustrates the similarity of associations for two example verbs, the polysemous verb *klagen*, and a near-synonym of one of its senses, *jammern* 'moan'. The table is an extract of all overlapping associations, listing those associations which were given at least twice for each verb, and the response frequencies with respect to the two stimulus verbs. The total overlap of these two verbs is 35 association types.

Considering the associations as verb features, we calculated probability distributions for each of the 330 experiment stimulus verbs over the association types, and performed a standard clustering: The verbs and their features were taken as input to agglomerative (bottom-up) hierarchical clustering. As similarity measure in the clustering procedure (i.e., to determine the distance/similarity for two verbs), we used the standard measure *skew divergence*, cf. Equation (2), a smoothed variant of the *Kullback-Leibler divergence*, cf. Equation (1), which measures the difference between two probability distributions p and q. The weight w was set to 0.9.

Table V. Association overlap for stimulus verbs.

klagen/jammern 'moan'		
Frauen	'women'	2/3
Leid	'suffering'	6/3
Schmerz	'pain'	3/7
Trauer	'mourning'	6/2
bedauern	'regret'	2/2
beklagen	'bemoan'	4/3
heulen	'cry'	2/3
nervig	'annoying'	2/2
nölen	'moan'	2/3
traurig	'sad'	2/5
weinen	'cry'	13/9

The measure has proven effective for distributional similarity in Natural Language Processing (Lee, 2001; Schulte im Walde, 2006b). *Ward's method* (minimising the sum-of-squares) was used as criterion for merging clusters. The goal of the clustering was not to explore the optimal feature combination; thus, we relied on previous clustering experiments and parameter settings (Schulte im Walde, 2006b). Furthermore, we are aware that a hard clustering is sub-optimal for the polysemous data; this article does not approach polysemy in verb classes but rather postpones the issue to future work. For details on the clustering method see e.g. Kaufman and Rousseeuw (1990).

$$KL(p \mid\mid q) = \sum_i p_i \, log \, \frac{p_i}{q_i} \tag{1}$$

$$skew(p,q) = KL(p \mid\mid w * q + (1-w) * p) \tag{2}$$

The hierarchical clustering was cut at a hierarchy level with 100 verb classes, i.e., the classes contain an average of 3.3 verbs. This cut was not empirically verified; we argue that the exact level in the hierarchical clustering is not critical for the analyses to follow. The obtained classes are characterised by a) the verbs in the classes, and b) associations which underlie the respective classes. Table VI shows two example classes from the 100-class analysis, listing the verbs and the most distinctive features of the example classes.[9] The following section validates whether the classes in the hierarchical clustering might be useful as a reference set for semantic verb classification.

Table VI. Examples of association-based classes.

Verbs	*Associations*
bedauern 'regret', *heulen* 'cry', *jammern* 'moan', *klagen* 'complain, moan, sue', *verzweifeln* 'become desperate', *weinen* 'cry'	*Trauer* 'mourning', *weinen* 'cry', *traurig* 'sad', *Tränen* 'tears', *jammern* 'moan', *Angst* 'fear', *Mitleid* 'pity', *Schmerz* 'pain', etc.
abnehmen, *abspecken* (both: 'lose weight'), *zunehmen* 'gain weight'	*Diät* 'diet', *Gewicht* 'weight', *dick* 'fat', *abnehmen* 'lose weight', *Waage* 'scale', *Essen* 'food',*essen* 'eat', *Sport* 'sports', *dünn* 'thin', *Fett* 'fat', etc.

3.2. VALIDATION

Our claim is that the hierarchical verb classes and their underlying features (i.e., the associations to the verbs) represent a coherent semantic classification of the verbs, which is not restricted by a specific framework underlying the class creation. An intuitive inspection of the cluster analysis has confirmed this assumption. To support this claim on a more objective and general basis, we validated the association-based classes against standard approaches to semantic verb classes, i.e., *GermaNet* as the German WordNet (Kunze, 2000), and the German counterpart of FrameNet in the *Salsa* project (Erk et al., 2003).

We could not directly compare the association-based classes against the GermaNet/FrameNet classes, since not all of our 330 experiment verbs were covered by the two resources. Thus, we needed a workaround that adjusted our association-based classes to the respective verbs in the resources. We replicated the above cluster experiment for the verbs that were actually covered by the manual classifications. First, we extracted those classes from the resources which contained any of our 330 verbs; other verbs, light verbs and classes not containing any of our verbs were disregarded. This left us with 33 classes from GermaNet, and 38 classes from FrameNet, containing only verbs from our association experiment. These remaining classifications were polysemous: The 33 GermaNet classes contained 71 verb senses which distributed over 56 verbs, and the 38 FrameNet classes contained 145 verb senses which distributed over 91 verbs. Based on the 56/91 verbs in the two gold standard resources, we performed two cluster analyses replicating our original procedure in Section 3.1, one for the GermaNet verbs, and one for the FrameNet verbs. As for the complete set of experiment verbs, we performed a hierarchical clustering on the respective subsets of the experiment verbs, again using their associations as verb features. The actual validation procedure then used the reduced classifications: The resulting analyses were evaluated against the respective resource classes on each level in the hierarchies, i.e., from 56/91 classes to 1 class. As an evaluation measure, we used a pair-wise measure which calculates

precision, recall and a harmonic f-score as follows: Each verb pair in the cluster analysis was compared to the verb pairs in the gold standard classes, and evaluated as true or false positive (Hatzivassiloglou and McKeown, 1993).

Figures 1 and 2 present the precision, recall and f-score values of the cluster analyses for the GermaNet and FrameNet verbs, respectively. The x-axis shows the number of clusters (ranging from 56/91 to 1), and the y-axis shows the P/R/F percentages. The precision starts at 100% and then decreases with the bottom-up clustering, and the recall increases. For the FrameNet verbs, the decrease of the precision happens faster, and the increase of the recall happens slower than for the GermaNet verbs. This resulted in a lower maximum value for the f-scores (62.69% for GermaNet and 34.68% for FrameNet) and also in a smaller number of clusters in the optimal analyses (32 clusters for GermaNet and 10 clusters for FrameNet). In comparison, an uninformed baseline, where the 56/91 verbs were hierarchically clustered by a random choice of pairing two clusters in each step, reached an f-score of 6.19% for GermaNet (on 4 clusters), and an f-score of 8.23% for FrameNet (on 8 clusters).

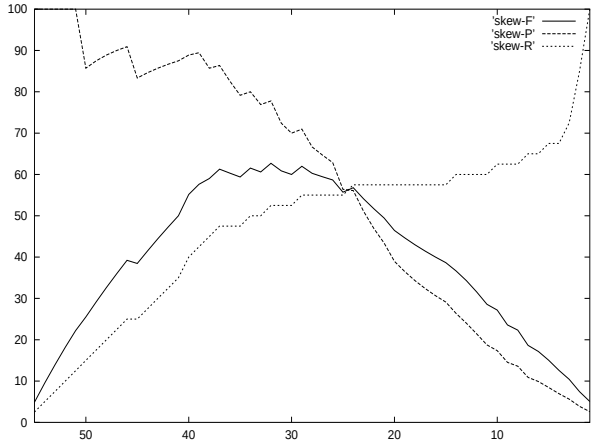

Figure 1. P/R/F for GermaNet clustering.

Comparing the maximum f-scores with the corresponding upper bounds demonstrates that the overlap of the association-based GermaNet/FrameNet clusters with the respective gold standard resources is quite impressive. The upper bounds for both GermaNet and FrameNet are below 100% (82.35% for GermaNet and 60.31% for FrameNet), because the hierarchical clustering assigns a verb to only one cluster, but the lexical resources contain polysemy. To calculate the upper bounds, we therefore created a hard version of the lexical resource classes where we randomly chose one sense of each polysemous verb,[10] and calculated the upper bounds by evaluating the hard versions against the soft versions. In relation to the upper bounds, there is considerable overlap between our association-based classes and existing semantic classes. The different results for the two resources are due to their

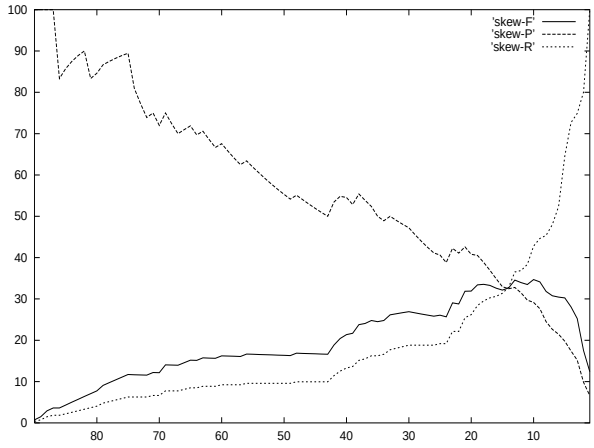

Figure 2. P/R/F for FrameNet clustering.

semantic background (i.e., capturing synonymy vs. situation-based agreement), the numbers of verbs, and the degrees of ambiguity (an average of 1.6 senses per verb in FrameNet, as compared to 1.3 senses in GermaNet), cf. Schulte im Walde (2006c) for more details.

The purpose of the validation against semantic resources was to demonstrate that, in addition to an intuitive approval, a clustering based on the verb associations and a standard clustering setting compares well with existing semantic classes. We take the positive validation results as justification for using the association-based classes as a source for cluster information, i.e., with respect to the verbs in a common association-based class, and the features which are relevant for the respective class, cf. Table VI.

4. Corpus-based Verb Classes

Our hope is that the features underlying the association-based classes will help us guide the feature selection process in future clustering experiments, because the cluster analysis tells us which semantic classes are based on which associations/features. This section actually investigates the potential of the associations, and answers our first question, *whether human associations help identify salient features for inducing semantic verb classes*. We use various corpus-based features to cluster our 330 German verbs, and the results are compared with the association-based classes from the previous section. The comparisons allow insights into the usefulness of standard feature types in verb clustering (such as direct objects), and an assessment of shallow window co-occurrence features vs. deeper syntactic features. In addition, we vary the corpus-based features with respect to their corpus frequency, to determine the influence of the feature frequency within the cluster analyses. Finally, by applying the feature choices not only to our association-

based classes but also to GermaNet and FrameNet, we address our second question *whether the same types of features are salient for different types of semantic verb classes*.

4.1. EXPLORING CORPUS-BASED FEATURES

In the first step, we relied on the association-based classes in the 100-class analysis of the hierarchical clustering[11] and features which exist for at least two verbs in a common class (and therefore hint to a minimum of verb similarity), and compared the associations underlying the association-based classes with standard corpus-based feature types: We examined how many of the association features we found among the corpus-based features, such as adverbs, direct object nouns, etc. Note that these association features were slightly different to the verb-association types collected in the web experiment: first, because we considered only a subset of the associations per verb class (because we only considered associations that were provided for at least two verbs per class); and second, these chosen associations were assumed to indicate common features of all verbs in the respective class and were therefore generalised to all verbs in the class, so that *unseen* verb-association pairs were created in addition to verb-association types from the experiment data. The check on association features against corpus-based feature types enabled us to determine whether the overlap of verb associations and corpus-based feature types correlated with the clustering success of the respective feature types.

There are various ways of determining corpus-based features that potentially cover the associations; we decided in favour of feature types that have been suggested in related work, and feature types that came out of the association analyses:

(a) Grammar-based relations:

As mentioned before, previous work on distributional similarity has focused either on a specific word-word relation (such as Pereira et al. (1993) and Rooth et al. (1999) referring to a direct object noun for describing verbs), or used any syntactic relationship detected by a chunker or a parser (such as Lin (1998) and McCarthy et al. (2003)). We used the statistical grammar from Section 2.2.1 to filter all verb-noun pairs where the nouns represented nominal heads in NPs or PPs in syntactic relation to the verb (subject, object, adverbial function, etc.), and to filter all verb-adverb pairs where the adverbs modified the verbs. The result is a pool of features whose various portions are used as feature sets.

(b) Co-occurrence window:

The findings in the analyses of our association data (cf. Section 2.2) suggested a co-occurrence window as an alternative source for shallow verb features, as opposed to specific syntactic relations. We therefore determined the co-occurring words for all experiment verbs in a 20-word window (i.e., 20 words

preceding and following the verb), irrespective of the part-of-speech of the co-occurring words, and used the resulting co-occurrence vectors as our second pool of features.

Relying on the verb information extracted for (a) and (b), we examined for each verb-association pair whether it occurred among the grammar or window pairs. Table VII illustrates which proportions of the associations we found in the two resource types. For the grammar-based relations, we looked at argument NPs and PPs (as separate sets and together), and in addition we used verb-noun pairs in the most common specific NP functions: *n* refers to the (nominative) intransitive subject, *na* to the transitive subject, and *na* to the transitive (accusative) object. For the windows, *all* examines co-occurrence of verbs and associations in the whole 200-million word corpus. *cut* also queries the whole corpus, but disregards the most and least frequently co-occurring words: verb-word pairs were only considered if the sum of co-occurrence frequencies of the word over all verbs was above 100 (disregarding low frequency pairs) and below 200,000 (disregarding high frequency pairs). Using the cut-offs, we can distinguish the relevance of high- and low-frequency features. Finally, *ADJ, ADV, N, V* perform co-occurrence checks for the whole corpus, but break down the *all* results with respect to the association part-of-speech.

Table VII. Coverage of verb association features by grammar and window resources.

	grammar relations						
	n	na	na	NP	PP	NP&PP	ADV
Coverage (%)	3.82	4.32	6.93	12.23	5.36	14.08	3.63

	co-occurrence: window-20					
	all	cut	ADJ	ADV	N	V
Coverage (%)	66.15	57.79	9.13	1.72	39.27	15.51

As one would have expected, most of the associations (66%) were found in the 20-word co-occurrence window, because the window was neither restricted to a certain part-of-speech, nor to a certain grammar relation; in addition, the window was potentially larger than a sentence. Applying the frequency cut-offs reduced the overlap of association types and co-occurring words to 58%. Specifying the window results for the part-of-speech types once more illustrated that the nouns play the most important role in describing verb meaning.[12]

The proportions of the nouns with a specific grammar relationship to the verbs were all below 10%. Looking at all NPs and/or PPs, we found that the proportions

increased for the NPs, and that the NPs played a more important role than the PPs. Of the adverb associations, we only found a small proportion among the parsed adverbs. All in all, the proportions of association types among the nouns/adverbs with a syntactic relationship to the verbs were rather low.

4.2. CORPUS-BASED CLUSTERING

In the second step, we applied the corpus-based feature types to clusterings. The goal of this step was to determine whether the feature exploration helped to identify salient verb features, in which case we would expect some correlation between the feature exploration results and the clustering results. The clustering experiments were as follows: The 330 experiment verbs were instantiated by the feature types we explored in the previous section. As for the association-based classes, we then performed an agglomerative hierarchical clustering. We cut the hierarchy at a level of 100 clusters, and evaluated the clustering against the 100-class analysis of the original association-based classes.

In addition, we applied the corpus-based features to GermaNet and FrameNet classes, in order to assess the cluster analyses against different semantic classification types. To ensure that the various gold standard classifications were comparable, we created two sub-classifications of the GermaNet and FrameNet resources:

- **GermaNet:** We randomly extracted 100 verb classes from all GermaNet synsets, and created a hard classification for these classes, by randomly deleting additional senses of a verb so as to leave only one sense for each verb. This selection made the GermaNet classes comparable to the association-based classes in size and polysemy. The 100 classes contained 233 verbs. Again, we performed an agglomerative hierarchical clustering on the verbs (as modelled by the different feature types). We cut the hierarchy at a level of 100 clusters, which corresponds to the number of GermaNet classes, and evaluated against the GermaNet classes.

- **FrameNet:** In a pre-release version from May 2005, there were 484 verbs in 214 German FrameNet classes. We disregarded the high-frequency verbs *gehen, geben, sehen, kommen, bringen* which were assigned to classes mostly on the basis of multi-word expressions they are part of. In addition, we disregarded two large classes which contained mostly support verbs, and we disregarded singletons. Finally, we created a hard classification of the classes, by randomly deleting additional senses of a verb so as to leave only one sense for each verb. The classification then contained 77 classes with 406 verbs. Again, we performed an agglomerative hierarchical clustering on the verbs (as modelled by the different feature types). We cut the hierarchy at a level of 77 clusters, which corresponded to the number of FrameNet classes, and evaluated against the FrameNet classes.

Table VIII. Accuracy for induced verb classes.

	grammar relations						
	n	na	na	NP	PP	NP&PP	ADV
Assoc	35.90	37.18	39.25	39.14	37.97	**41.28**	38.53
GN	**58.01**	53.37	51.90	53.10	54.21	51.77	51.82
FN	29.46	30.13	32.74	34.16	28.72	33.91	**35.24**

	co-occurrence: window-20					
	all	cut	ADJ	ADV	N	V
Assoc	39.33	**39.45**	37.31	36.89	39.33	38.84
GN	51.53	52.42	50.88	47.79	**52.86**	49.12
FN	32.01	32.84	31.08	31.00	**34.24**	31.75

For the evaluation of the clustering results, we calculated the *accuracy* of the clusters, a cluster similarity measure that has been applied before, cf. Stevenson and Joanis (2003), Korhonen et al. (2003). Note that we can use *accuracy* for the evaluation because we have a fixed cut in the hierarchy based on the respective gold standard, as opposed to the evaluation in Section 3.2 where we explored the optimal cut level. Accuracy is determined in two steps:

1. For each class in the cluster analysis, the gold standard class with the largest intersection of verbs is determined. The number of verbs in the intersection ranges from one verb only (where all clustered verbs are in different classes in the gold standard) to the total number of verbs in a cluster (where all clustered verbs are in the same gold standard class).

2. Accuracy is calculated as the proportion of the verbs in the clusters covered by the same gold standard classes, divided by the total number of verbs in the clusters. The upper bound of the accuracy measure is 1.

Table VIII shows the accuracy results for the three types of classifications (association-based classes, GermaNet, FrameNet), and the grammar-based and window-based features. The best result per row is highlighted in bold.

The strongest hypothesis we can think of with respect to the result table and the main question of this article whether ”human verb associations help identify salient features for semantic verb classification” could be formulated as follows. Assuming that the associations are salient features for verb clustering, the better we

model the associations with grammar-based or window-based features, the better the clustering. However, this hypothesis is not supported by the result table: there is no correlation between the overlap of associations and feature types in Table VII on the one hand and the clustering results based on the feature types in Table VIII on the other hand (Pearson's correlation, $p>.1$), neither for the association-based classes nor the GermaNet or FrameNet classes. But even though we did not find support for the strong correlation hypothesis, the associations did provide interesting insights into various aspects of feature selection. In the following, the missing correlations as well as the positive insights are described in some detail.

Firstly, we only found corresponding patterns in some specific cases; for example, the clustering results for the intransitive and transitive subject and the transitive object corresponded to the overlap values for the association-based classes and FrameNet: n $<$ na $<$ na. Interestingly, the GermaNet clusterings behaved in the opposite way.

Comparing the grammar-based relations with each other shows that for the association-based classes using all NPs was better than restricting the NPs to (subject) functions, and using both NPs and PPs was best; similarly for the FrameNet classes where using all NPs was the second best result (after adverbs). On the other hand, for the GermaNet classes the specific function of intransitive subjects outperformed the more general feature types, and the PPs were still better than the NPs. We conclude that not only there is no correlation between the association overlap and feature types, but in addition the most successful feature types vary hugely with respect to the gold standard. None of the differences within the feature groups (n/na/na and NP/PP/NP&PP) were significant ($\chi^2, df = 1, \alpha = 0.05$). The adverbial features were surprisingly successful in all three clusterings, in some cases even outperforming the noun-based features.

For both gold standards and the reference set, the best window-based clustering results were below the best grammar-based results. However, it is interesting that the clusterings based on window co-occurrence were not significantly worse ($\chi^2, df = 1, \alpha = 0.05$) and in some cases even better than the clusterings based on selected grammar-based functions. This means that a careful choice and extraction of specific relationships for verb features did not have a significant impact on semantic classes.

Comparing the window-based features against each other shows that even though we discovered a much larger proportion of association types in an unrestricted window *all* than elsewhere, the results in the clusterings did not differ accordingly. Applying the frequency cut-offs had almost no impact on the clustering results, which means that it did no harm to leave out the rather unpredictable features. Somehow expected but nevertheless impressive is the fact that only considering nouns as co-occurring words was as successful as considering all words independent of the part-of-speech. These insights might have an impact on the complexity of comparable clustering approaches, because using the *cut* version of the features

instead of the *all* version means – with respect to our corpus data – cutting down the number of features from 934,000 to 100,000.

Finally, the overall accuracy values were much better for the GermaNet clusterings than for the experiment-based and the FrameNet clusterings. The differences were all significant ($\chi^2, df = 1, \alpha = 0.05$). The reason for these large differences could be either (a) that the clustering task was easier for the GermaNet verbs, or (b) that the differences were caused by the underlying semantics. We argue against case (a) since we deliberately chose the same number of classes (100) as for the association-based reference set. However, Table IX demonstrates the results of a post-check on the empirical properties of the chosen verbs; there were empirical differences in the three original verb classifications, which might have influenced the clustering result: The verbs-per-class ratio for GermaNet vs. the association-based classes and the FrameNet classes was different (2.33 vs. 3.30/5.27) and we cannot be sure what influence this had. In addition, the average verb frequencies in the GermaNet classes (calculated from the 35 million word newspaper corpus) were clearly below those in the other two classifications (1,040 as compared to 2,465 and 1,876), and there were more low-frequency verbs (98 out of 233 verbs (42%) have a corpus frequency below 50, as compared to 41 out of 330 (12%) and 54 out of 406 (13%)). To our knowledge there is, as yet, no existing work that investigates the influence of such parameters is detail, so there is potential for future investigations. In the case of (b), the difference in the semantic class types was modelling synonyms with GermaNet as opposed to situation-based agreement in FrameNet. The association-based class semantics was similar to FrameNet, because the associations were unrestricted in their semantic relation to the experiment verb (Schulte im Walde and Melinger, 2005). A more detailed analysis of which types of semantic verb classifications rely on exactly which types of features is therefore also an interesting question for future research.

Table IX. Properties of verb classifications.

GS	classes	verbs	verbs/class	avg. v-freq	v-freq<50/20/10		
Assoc	100	330	3.30	2,465	41	16	8
GN	100	233	2.33	1,040	98	65	40
FN	77	406	5.27	1,876	54	16	11

5. Related Work

This article is concerned with interdisciplinary research that touches various fields in psycholinguistics and computational linguistics. We therefore sub-divide related

work into several areas, presenting previous collections and investigations of human data on semantic issues, and previous approaches to the automatic induction of semantic relations and semantic classes.

5.1. COLLECTIONS OF HUMAN DATA ON SEMANTIC RELATEDNESS

This article relies on the fact that association norms have a long tradition in psycholinguistic research. Consequently, one finds association norms for various languages and for various domains, as Section 2 has already introduced. These data have been investigated for psycholinguistic reasons as well as for purposes in Natural Language Processing, as Section 5.2 will describe.

In addition to the "classical association norms" and turning towards computational linguistics work, there is an enormous number of approaches that have collected human judgements on semantic relatedness for the development and/or the assessment of linguistic resources and methods. It is impossible to cover the wealth of methods and data, so we just pick two examples: McCarthy et al. (2003) collected human rankings on the semantic relatedness of word pairs, because they were interested in the semantic similarity of particle verbs with respect to their base verbs, to evaluate models of particle verb compositionality. Similarly, Gurevych et al. (2007) collected human rankings across part-of-speech word pairs, and used them as gold standard semantic relatedness data within Information Retrieval experiments.

On a more complex level beyond ranking judgements, and more similar to our data, Morris and Hirst (2004) performed a study on lexical semantic relations that ensure text cohesion. Their work relied on human labels of semantic text relations. Beigman Klebanov and Shamir (2006) investigated how well readers agree on which items in a text are lexically cohesive, and why (i.e., based on which semantic relations); Beigman Klebanov (2006) continued this work, investigated form-based clues to lexical cohesion in text, and modelled the text relations by various WordNet similarity measures. Boyd-Graber et al. (2006) performed a large-scale study on evocation, a semantic relation similar to association, to enhance WordNet.

5.2. INVESTIGATIONS OF ASSOCIATION DATA

In early work on association norms, Clark (1971) identified potential relations between stimulus words and their associations on a theoretical basis. He categorised stimulus-association relations into sub-categories of paradigmatic and syntagmatic relations, such as synonymy and antonymy, selectional preferences, etc. Heringer (1986) performed an actual study of association norms, concentrated on syntagmatic associations to a small selection of 20 German verbs. He asked his subjects to provide question words as associations (e.g., *wer* 'who', *warum* 'why'), in order to investigate the valency behaviour of the verbs. Spence and Owens (1990), as mentioned before, showed that associative strength and word co-occurrence are

correlated. Their investigation was based on 47 pairs of semantically related concrete nouns, as taken from the *Word Association Norms* (Palermo and Jenkins, 1964), and their co-occurrence counts in a window of 250 characters in the 1-million-word Brown corpus. Church and Hanks (1989) were the first to apply information-theoretic measures to corpus data in order to predict word association norms. However, they did not rely on or evaluate against existing association data, but rather concentrated on the usage of the measure for lexicographic purposes. Their paper can be considered as a milestone within the automatic acquisition of distributional semantic similarity.

Further work in that direction was conducted by Reinhard Rapp, Manfred Wettler and colleagues, which is in some respects closely related to our work. They also relied on the co-occurrence assumption that there are correlations between associative strength in association norms and word co-occurrence in language corpora, and exploited this assumption for purposes in computational linguistics. Wettler and Rapp (1993) defined a statistical model that predicted stimulus-associate pairs in English and German association norms. An evaluation of the model was carried out by comparing the predicted associations with the associations in the norms. Subsequent work presented various extensions of their basic model and application scenarios in a series of conference papers, which are summarised to a large extent in Rapp's PhD thesis (Rapp, 1996). Example applications of their model are the generation of search terms in Information Retrieval, and the prediction of marketing effects caused by word usage in advertisements. Our work is similar to their work in that we also show a relationship between association norms and word co-occurrence, and that we exploit this fact for issues in language processing. Differently to their work, though, we did not develop a statistical model for this relationship; for our purposes, it was sufficient to observe the relationship with respect to our association data, in order to formulate hypotheses concerning salient verb features.

Work by Christiane Fellbaum and colleagues in the 1990s focused on the semantic relationships between verbs. Similarly to our association experiment, Fellbaum and Chaffin (1990) asked participants in an experiment to provide associations to verbs. However, their work concentrated on verb-verb relations and therefore explicitly required verb responses to the verb stimuli. Also different from our work, they restricted their stimuli to only 28 verbs; the resulting verb-verb pairs were manually classified into five pre-defined semantic relations. Fellbaum (1995) investigated the relatedness between antonymous verbs and nouns and their co-occurrence behaviour. Within that work, she searched the Brown corpus for antonymous word pairs in the same sentence, and found that regardless of the syntactic category, antonyms occur in the same sentence with much higher-than-chance frequencies. Finally, the WordNet organisation of the various parts-of-speech does rely on psycholinguistic evidence to a large extent (Fellbaum, 1998).

Last but not least, most closely related to this article is our own work on collecting and investigating human associations. Schulte im Walde and Melinger (2005)

presented a more extensive investigation of the associations to German verbs than what was described in Section 2. In addition to the analyses that were repeated in this article, we also analysed the semantic relations between the stimulus verbs and their verb responses using WordNet relations, and performed a more detailed analysis of corpus co-occurrences. The co-occurrence distributions of semantic associations were also the focus of an in-depth investigation by Schulte im Walde and Melinger (2008). Roth (2006) used similar lexical resources as Schulte im Walde and Melinger (2005), i.e., the statistical grammar from Section 2.2.1, WordNet, and an online-dictionary, for an empirical analysis of German noun associations, cf. Melinger and Weber (2006). Finally, Melinger et al. (2006) took the noun associations as input to a soft clustering approach, in order to determine the various noun senses of ambiguous nouns.

5.3. AUTOMATIC INDUCTION OF SEMANTIC CLASSES

Turning towards the motivating application of this article, there is related work with respect to an automatic acquisition of verb (and other part-of-speech) semantic classes. Schulte im Walde (2008) provides an overview of state-of-the-art automatic verb classifications; we therefore restrict ourselves to a few example approaches.

The first set of examples concerns approaches with a similar target classification as this article. As mentioned before, Merlo and Stevenson (2001) investigated three verb classes (unergative, unaccusative, and object-drop verbs) and defined verb features that rely on linguistic heuristics to describe the thematic roles of subjects and objects in transitive and intransitive verb usage. The features included heuristics for transitivity, causativity, animacy, and syntactic features. Joanis and Stevenson (2003) presented an extension of their work that approached 14 Levin classes.[13] They defined an extensive feature space including part-of-speech, auxiliary frequency, syntactic categories, and animacy, plus selectional preference features taken from WordNet. Stevenson and Joanis (2003) then applied various approaches to automatic feature selection in order to reduce the feature set to the relevant features, addressing the problem of too many irrelevant features. They reported a semi-supervised chosen set of features based on seed verbs (i.e., representative verbs for the verb classes) as the most reliable choice. The work by Korhonen et al. (2003) is one out of only a few approaches that used a soft-clustering method, the Information Bottleneck, to cluster verbs with possible multiple senses. They relied on subcategorisation frames as verb features, to produce Levin-style English verb classes. Schulte im Walde (2000; 2006b) described English/German verbs by probabilities for subcategorisation frames including prepositional phrase types, plus selectional preferences referring to the WordNet/GermaNet top-level synsets. The classification target was semantic verb classes such as *manner of motion, desire, observation*.

The second set of examples concerns approaches that were already mentioned with respect to their feature selection. The target classifications were also word classes, but of slightly different style than in the previous examples. Rooth et al. (1999) used verb-noun pairs with a direct-object-relationship and produced soft semantic clusters for English which at the same time represented a classification of verbs as well as of nouns. The conditioning of the verbs and the nouns on each other was done using hidden classes and the joint probabilities of classes. Verbs and nouns were trained by the Expectation-Maximisation algorithm. The resulting model defined conditional membership probabilities of each verb and noun in each class. Earlier work by Pereira et al. (1993) focused on a similar task of creating soft clusters of verbs and (their direct object) nouns, but produced a hierarchical clustering, using a deterministic annealing procedure. Lin (1998) used verb-noun pairs from a dependency parser (not restricted to a specific syntactic relationship) and various similarity measures. His goal was to create thesaurus entries for all words in the corpus. Lin (1999) and McCarthy et al. (2003) are two examples of approaches that applied the same method as Lin (1998) to extract distributional features, both for the judgement of the compositionality of multi-word expressions.

5.4. AUTOMATIC INDUCTION OF SEMANTIC RELATIONS

Closely related to the automatic induction of semantic classes is the automatic induction of semantic relations: words are supposed to be assigned to common semantic classes because of some underlying semantic relatedness between the words. Consequently, the methods of automatic approaches that aim to induce word pairs according to pre-specified semantic relations are to some extent similar to those for automatic class induction. They often rely on co-occurrence and syntactic functions as word features, and use standard similarity measures, similar to work on inducing word classes. In addition, some approaches make use of morpho-syntactic corpus patterns, or knowledge obtained from existing resources. Example approaches that addressed the automatic induction of semantic relations refer to noun-noun relations, such as hypernymy (Hearst, 1998), causal relation (Girju, 2003), part-whole relation (Berland and Charniak, 1999; Girju et al., 2006), various relations between nouns in general (Navigli and Velardi, 2004), or specifically for noun compounds (Rosario and Hearst, 2001; Girju et al., 2005); work on verb-verb relations is more rare, one example being Chklovski and Pantel (2004). Other approaches concentrate on the distinction between syntagmatic and paradigmatic approaches (Rapp, 2002; Biemann et al., 2004; Sahlgren, 2006), or focus on semantic relations that are relevant for creating ontologies (Maedche and Staab, 2000; Navigli and Velardi, 2004; Kavalek and Svatek, 2005).

6. Summary and Outlook

The questions we posed in the beginning of this article were (i) whether human associations help identify salient features for inducing semantic verb classes, and (ii) whether the same types of features are salient for different types of semantic verb classes. A series of analyses of human association data and, in addition, an association-based clustering with 100 classes served as a source for identifying a set of potentially salient verb features, and a comparison with standard corpus-based features determined proportions of feature overlap. Applying the standard feature choices to verbs underlying three verb classifications showed that there was no correlation between the overlap of associations and feature types and the respective clustering results. The associations therefore did not provide any direct help in the specific choice of corpus-based features, as we had hoped. However, the human associations nevertheless provided insight into aspects of feature types that might prove useful in future clustering experiments: (a) There is no significant preference for using a specific syntactic relationship (such as intransitive subjects vs. transitive subjects vs. direct objects) as nominal features in clustering, as has often been employed in previous work. (b) Related to this insight, the assumption that window-based features do contribute to semantic verb classes – this assumption came out of an analysis of the associations – was confirmed: simple window-based features were not significantly worse (and in some cases even better) than selected grammar-based functions. This finding is interesting because window-based features have often been considered too simple for semantic similarity, as opposed to syntax-based features. (c) Adverbs as features in verb descriptions were surprisingly successful in all three clusterings, in some cases even outperforming the noun-based features. This finding might also be of importance to related work, since adverbs have rarely been exploited as distributional features, even though they have the potential to point to aspectual properties of verbs, and moreover are easy to induce from corpus data. (d) In addition, it is not necessary to consider all features that are available from the window co-occurrences; a feature choice disregarding high- and low-frequency features was sufficient, which might have an impact on the complexity of clustering approaches relying on similar features as in our work. (e) Concerning our second question in this article, the clustering results were significantly better for the GermaNet clusterings than for the association-based and the FrameNet clusterings, so the chosen feature sets might be more appropriate for the synonymy-based than the situation-based classifications. The resulting question is: which types of semantic verb classifications rely on exactly which types of features? Our clustering experiments demonstrated that there is no overall optimal set of verb features in automatic semantic verb classification: the clustering results were different and even contradictory with respect to our chosen feature types and our chosen classifications. However, we did not focus on identifying feature types that are discriminative for specific semantic properties of the verb classifications, which could be a concern of future work. Furthermore, a

quick study of the empirical properties of the verbs in the classifications illustrated the common knowledge that classification parameters such as the sizes of the verb classes, the ambiguity of the verbs, and verb frequencies strongly influenced the clustering results. Nevertheless, to our knowledge there is, as yet, no existing work that investigates the influence of such parameters is detail.

Last but not least, we believe that the human association data provides further potential with respect to learning how to model or select features that are useful in automatic semantic verb classification, or related tasks that rely on the lexical-semantic features of verbs. For example, the associations might provide an insight into aspects of polysemy: if it is possible to automatically distinguish associations with respect to the multiple senses of the stimulus words, is it possible to induce feature types or empirical properties of features with respect to polysemy in corpus data? And finally, do the associations provide a means to learning how to model world knowledge?

Acknowledgements

Many thanks to Aoife Cahill, Christian Hying, Alissa Melinger, Sebastian Padó, and three anonymous reviewers for their valuable comments on previous versions of this article.

Appendix

A. Experiment Classes and Verbs

The 330 German verbs that were selected for the association experiment were drawn from a variety of semantic classes. Selecting verbs from different categories was only intended 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 but also on intuitive grounds. In total, we grossly defined a classification with 12 semantic classes, that were sub-divided into 48 classes. The 12 semantic classes were chosen as follows: MOTION, COMMERCE, GIVE & TAKE, ASPECT & EXISTENCE, SHOWING, CAUSING, EXPERIENCING, COGNITION, COMMUNICATION, POSITION, BODY, WEATHER. In order to provide a general idea of the semantic categories, Table X lists two example classes, accompanied by their sub-classes and choices of verbs. The class labels are given in English; the verbs are listed in German, with their English translations (in the case of polysemous verbs, the translation is provided with respect to the semantic class), and their corpus frequencies (determined by a 35 million word newspaper corpus).

Table X. Example classes and verbs.

Class: CAUSING	
DESTROY	*verbrennen* 'burn' (588), *verteilen* 'distribute' (1,966), *zerbrechen* 'break' (316), *zerreißen* 'tear' (183), *zerstören* 'destroy' (1,988)
CREATE	*backen* 'bake' (272), *basteln* 'do handicrafts' (431), *bauen* 'build' (3,878), *bemalen* 'paint (96) , *bilden* 'compose' (3,159), *entwickeln* 'develop' (3,425), *gründen* 'found' (2,465), *kochen* 'cook' (697), *malen* 'paint' (1,748)
QUANTUM CHANGE	*beladen* 'load up' (67), *laden* 'load' (979), *reduzieren* 'reduce' (1,395), *senken* 'decrease' (812), *steigern* 'increase' (808), *verändern* 'change' (2,616)
CHANGE FORM	*aushaken* 'unhook' (1), *beugen* 'bend' (304), *biegen* 'bend' (80), *drücken* 'squeeze' (976), *falten* 'fold' (31), *formen* 'form' (237), *kneten* 'knead' (38), *mischen* 'merge' (509), *schneiden* 'cut' (284), *trennen* 'separate' (1,204)
CHANGE STATE	*auftauen* 'defrost' (34), *aufweichen* 'soften' (58), *einfrieren* 'freeze' (131), *erhitzen* 'heat' (92), *härten* 'harden' (19), *schmelzen* 'melt' (108), *trocknen* 'dry' (52)
ACTIVE CAUSE	*arbeiten* 'work' (8,761), *lesen* 'read' (3,592), *rammen* 'drive against' (193), *schlagen* 'beat' (3,038), *schreiben* 'write' (6,649), *singen* 'sing' (1,875), *treten* 'kick' (2,734), *waschen* 'wash' (299), *wenden* 'turn' (1,780)
Class: EXPERIENCING	
EMOTION	*ärgern* 'annoy' (627), *bedauern* 'regret' (945), *ekeln* 'disgust' (31), *fürchten* 'fear' (2,003), *freuen* 'be happy' (2,478) , *grauen* 'dread' (131), *lachen* 'laugh' (1,428), *vergnügen* 'entertain' (86), *verzweifeln* 'despair' (99), *weinen* 'cry' (452), *wundern* 'be amazed' (707)
LOVE & HATE	*achten* 'respect' (579), *gedenken* 'commemorate' (699), *gefallen* 'like' (1,849), *hassen* 'hate' (409), *lieben* 'love' (2,187), *mögen* 'like' (3,175)
DESIRE	*brauchen* 'need' (10,075), *erhoffen* 'hope' (680), *gelüsten* 'be overcome by desire' (8), *hoffen* 'hope' (4,185), *wollen* 'want' (21,464), *wünschen* 'wish' (2,534)
PERCEPTION	*hören* 'hear' (5,040), *schmecken* 'taste' (427), *sehen* 'see' (24,862), *spüren* 'sense' (1,706), *wahrnehmen* 'perceive' (824)
EXPERIENCE	*amüsieren* 'amuse' (179), *aufregen* 'upset' (214), *bedrohen* 'threaten' (1,138), *begeistern* 'enthuse' (573), *ekeln* 'disgust' (31), *erschrecken* 'scare' (230), *schockieren* 'shock' (106), *staunen* 'be astonished' (239), *überraschen* 'surprise' (972), *verblüffen* 'amaze' (89) , *vergessen* 'forget' (2,187) , *verwirren* 'confuse' (129)
ATTEMPT	*hadern* 'quarrel' (64), *testen* 'test' (452), *versuchen* 'try' (7,144)

Notes

[1]For example, the two-class division by Siegel and McKeown (2000) distinguishes event verbs from stative verbs and consequently uses distributional indicators such as manner adverb, duration *in*-PP, past tense, perfect tense, etc. The three-class division by Merlo and Stevenson (2001), which divides transitive verbs into classes according to their alternation behaviour, relies on distributional indicators of thematic roles.

[2]This article is not the first to use association norms for analyses regarding natural language processing issues, cf. Section 5 on related work.

[3]The web experiment was conducted in collaboration with two colleagues from Saarland University (Saarbrücken, Germany), Katrin Erk and Alissa Melinger.

[4]Despite these instructions, some participants failed to use capitalisation, leading to some ambiguity. For example, the associate *wärme* represents a morphologically plausible imperative of the verb *wärmen* (and is analysed as such in the morphological analysis in Section 2.2.2). However, it is rather unlikely that the experiment participant intended to provide an imperative verb; he/she most probably wanted to refer to the noun *Wärme*, but did not use the appropriate capitalisation.

[5]All of our analyses reported in this article were based on response tokens; however, the type analyses showed the same overall pictures.

[6]The original analyses in Schulte im Walde and Melinger (2005) used three window sizes: 5, 20 and 50, to also cover more extreme window sizes; the 20-word window is considered to be appropriate for covering a local context that goes beyond the clause boundaries.

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

[8]Major parts of this section have been published in Schulte im Walde (2006a).

[9]The most distinctive features for a class were identified as those associations which accumulated the most probability mass, summed over all verbs in the class.

[10]The reader might wonder why we did not use the predominant senses of the GermaNet verbs, following a common standard, instead of randomly selecting a verb sense. The reason is that we wanted to keep the creation procedures of the two classifications as similar as possible, and since FrameNet does not define a predominant sense, we settled on the random selection.

[11]The exact number of classes or the verb-per-class ratio are not relevant for investigating the use of associations.

[12]Caveat: These numbers correlate with the part-of-speech types of all associate responses, cf. Section 2.2.2.

[13]Joanis et al. (2008) provide a more recent, extended version of this work.

References

Beigman Klebanov, B.: 2006, 'Measuring Semantic Relatedness Using People and WordNet'. In: *Proceedings of the joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics*. New York City, NY, pp. 13–17.

Beigman Klebanov, B. and E. Shamir: 2006, 'Reader-based Exploration of Lexical Cohesion'. *Language Resources and Evaluation* **40**(2), 109–126.

Berland, M. and E. Charniak: 1999, 'Finding Parts in Very Large Corpora'. In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. Maryland, MD, pp. 57–64.

Biemann, C., S. Bordag, and U. Quasthoff: 2004, 'Automatic Acquisition of Paradigmatic Relations using Iterated Co-Occurrences'. In: *Proceedings of the 4th International Conference on Language Resources and Evaluation*. Lisbon, Portugal.

Boyd-Graber, J., C. Fellbaum, D. Osherson, and R. Schapire: 2006, 'Adding Dense, Weighted Connections to WordNet'. In: *Proceedings of the Third Global WordNet Meeting*. Jeju Island, Korea.

Chklovski, T. and P. Pantel: 2004, 'VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations'. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Barcelona, Spain.

Church, K. W. and P. Hanks: 1989, 'Word Association Norms, Mutual Information, and Lexicography'. In: *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*. Vancouver, Canada, pp. 76–83.

Clark, H. H.: 1971, 'Word Associations and Linguistic Theory'. In: J. Lyons (ed.): *New Horizon in Linguistics*. Penguin, Chapt. 15, pp. 271–286.

Dorr, B. J.: 1997, 'Large-Scale Dictionary Construction for Foreign Language Tutoring and Interlingual Machine Translation'. *Machine Translation* **12**(4), 271–322.

Dorr, B. J. and D. Jones: 1996, 'Role of Word Sense Disambiguation in Lexical Acquisition: Predicting Semantics from Syntactic Cues'. In: *Proceedings of the 16th International Conference on Computational Linguistics*. Copenhagen, Denmark, pp. 322–327.

Erk, K., A. Kowalski, S. Padó, and M. Pinkal: 2003, 'Towards a Resource for Lexical Semantics: A Large German Corpus with Extensive Semantic Annotation'. In: *Proceedings of the 41st Annual Metting of the Association for Computational Linguistics*. Sapporo, Japan, pp. 537–544.

Fellbaum, C.: 1995, 'Co-Occurrence and Antonymy'. *Lexicography* **8**(4).

Fellbaum, C. (ed.): 1998, *WordNet – An Electronic Lexical Database*, Language, Speech, and Communication. Cambridge, MA: MIT Press.

Fellbaum, C. and R. Chaffin: 1990, 'Some Principles of the Organization of Verbs in the Mental Lexicon'. In: *Proceedings of the 12th Annual Conference of the Cognitive Science Society of America*.

Fernández, A., E. Diez, M. A. Alonso, and M. S. Beato: 2004, 'Free-Association Norms for the Spanisch Names of the Snodgrass and Vanderwart Pictures'. *Behavior Research Methods, Instruments and Computers* **36**(3), 577–583.

Ferrand, L. and F.-X. Alario: 1998, 'French Word Association Norms for 366 Names of Objects'. *L'Annee Psychologique* **98**(4), 659–709.

Ferrer, E. E.: 2004, 'Towards a Semantic Classification of Spanish Verbs based on Subcategorisation Information'. In: *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*. Barcelona, Spain.

Fillmore, C. J.: 1982, 'Frame Semantics'. *Linguistics in the Morning Calm* pp. 111–137.

Fillmore, C. J., C. R. Johnson, and M. R. Petruck: 2003, 'Background to FrameNet'. *International Journal of Lexicography* **16**, 235–250.

Girju, R.: 2003, 'Automatic Detection of Causal Relations for Question Answering'. In: *Proceedings of the ACL Workshop on Multilingual Summarization and Question Answering – Machine Learning and Beyond*. Sapporo, Japan.

Girju, R., A. Badulescu, and D. Moldovan: 2006, 'Automatic Discovery of Part-Whole Relations'. *Computational Linguistics* **32**(1), 83–135.

Girju, R., D. Moldovan, M. Tatu, and D. Antohe: 2005, 'On the Semantics of Noun Compounds'. *Journal of Computer Speech and Language* **19**(4). Special Issue on Multiword Expressions.

Gurevych, I., C. Müller, and T. Zesch: 2007, 'Electronic Career Guidance based on Semantic Relatedness'. In: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*. Prague, Czech Republic.

Harris, Z.: 1968, 'Distributional Structure'. In: J. J. Katz (ed.): *The Philosophy of Linguistics*, Oxford Readings in Philosophy. Oxford University Press, pp. 26–47.

Hatzivassiloglou, V. and K. R. McKeown: 1993, 'Towards the Automatic Identification of Adjectival Scales: Clustering Adjectives According to Meaning'. In: *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*. Columbus, Ohio, pp. 172–182.

Hearst, M.: 1998, 'Automated Discovery of WordNet Relations'. In (Fellbaum, 1998).
Heringer, H. J.: 1986, 'The Verb and its Semantic Power: Association as the Basis for Valence'. *Journal of Semantics* **4**, 79–99.
Hirsh, K. W. and J. Tree: 2001, 'Word Association Norms for two Cohorts of British Adults'. *Journal of Neurolinguistics* **14**(1), 1–44.
Joanis, E. and S. Stevenson: 2003, 'A General Feature Space for Automatic Verb Classification'. In: *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*. Budapest, Hungary.
Joanis, E., S. Stevenson, and D. James: 2008?, 'A General Feature Space for Automatic Verb Classification'. *Natural Language Engineering.* To appear.
Kaufman, L. and P. J. Rousseeuw: 1990, *Finding Groups in Data – An Introduction to Cluster Analysis*, Probability and Mathematical Statistics. New York: John Wiley & Sons, Inc.
Kavalek, M. and V. Svatek: 2005, 'A Study on Automated Relation Labelling in Ontology Learning'. In: P. Buitelaar, P. Cimiano, and B. Magnini (eds.): *Ontology Learning and Population*, Vol. 123 of *Frontiers in Artificial Intelligence*. IOS Press.
Kiss, G., C. Armstrong, R. Milroy, and J. Piper: 1973, 'An Associative Thesaurus of English and its Computer Analysis'. In: *The Computer and Literary Studies*. Edinburgh University Press.
Klavans, J. L. and M.-Y. Kan: 1998, 'The Role of Verbs in Document Analysis'. In: *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics*. Montreal, Canada, pp. 680–686.
Koehn, P. and H. Hoang: 2007, 'Factored Translation Models'. In: *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Prague, Czech Republic, pp. 868–876.
Kohomban, U. S. and W. S. Lee: 2005, 'Learning Semantic Classes for Word Sense Disambiguation'. In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Ann Arbor, MI, pp. 34–41.
Korhonen, A.: 2002, 'Subcategorization Acquisition'. Ph.D. thesis, University of Cambridge, Computer Laboratory. Technical Report UCAM-CL-TR-530.
Korhonen, A., Y. Krymolowski, and Z. Marx: 2003, 'Clustering Polysemic Subcategorization Frame Distributions Semantically'. In: *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. Sapporo, Japan, pp. 64–71.
Kunze, C.: 2000, 'Extension and Use of GermaNet, a Lexical-Semantic Database'. In: *Proceedings of the 2nd International Conference on Language Resources and Evaluation*. Athens, Greece, pp. 999–1002.
Lapata, M. and C. Brew: 2004, 'Verb Class Disambiguation using Informative Priors'. *Computational Linguistics* **30**(1), 45–73.
Lauteslager, M., T. Schaap, and D. Schievels: 1986, *Schriftelijke Woordassociatienormen voor 549 Nederlandse Zelfstandige Naamworden.* Swets and Zeitlinger.
Lee, L.: 2001, 'On the Effectiveness of the Skew Divergence for Statistical Language Analysis'. *Artificial Intelligence and Statistics* pp. 65–72.
Levin, B.: 1993, *English Verb Classes and Alternations.* The University of Chicago Press.
Lin, D.: 1998, 'Automatic Retrieval and Clustering of Similar Words'. In: *Proceedings of the 17th International Conference on Computational Linguistics*. Montreal, Canada.
Lin, D.: 1999, 'Automatic Identification of Non-compositional Phrases'. In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. Maryland, MD, pp. 317–324.
Maedche, A. and S. Staab: 2000, 'Discovering Conceptual Relations from Text'. In: *Proceedings of the 14th European Conference on Artificial Intelligence*. Berlin, Germany.
McCarthy, D., B. Keller, and J. 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.

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

McRae, K. and S. Boisvert: 1998, 'Automatic Semantic Similarity Priming'. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **24**(3), 558–572.

Melinger, A., S. Schulte im Walde, and A. Weber: 2006, 'Characterizing Response Types and Revealing Noun Ambiguity in German Association Norms'. In: *Proceedings of the EACL Workshop 'Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics together'*. Trento, Italy, pp. 41–48.

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

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

Morris, J. and G. Hirst: 2004, 'Non-Classical Lexical Semantic Relations'. In: *Proceedings of the HLT Workshop on Computational Lexical Semantics*. Boston, MA.

Navigli, R. and P. Velardi: 2004, 'Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites'. *Computational Linguistics* **30**(2), 151–179.

Nelson, D., C. McEvoy, and T. Schreiber: 1998, 'The University of South Florida Word Association, Rhyme, and Word Fragment Norms'.

Padó, U., M. Crocker, and F. Keller: 2006, 'Modelling Semantic Role Plausibility in Human Sentence Processing'. In: *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*. Trento, Italy.

Palermo, D. and J. Jenkins: 1964, *Word Association Norms: Grade School through College*. Minneapolis: University of Minnesota Press.

Palmer, M., D. Gildea, and P. Kingsbury: 2005, 'The Proposition Bank: An annotated Resource of Semantic Roles'. *Computational Linguistics* **31**(1), 71–106.

Pereira, F., N. Tishby, and L. Lee: 1993, 'Distributional Clustering of English Words'. In: *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*. Columbus, OH, pp. 183–190.

Pinker, S.: 1989, *Learnability and Cognition: The Acquisition of Argument Structure*. Cambridge, MA: MIT Press.

Plaut, D. C.: 1995, 'Semantic and Associative Priming in a Distributed Attractor Network'. In: *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, Vol. 17. pp. 37–42.

Prescher, D., S. Riezler, and M. Rooth: 2000, 'Using a Probabilistic Class-Based Lexicon for Lexical Ambiguity Resolution'. In: *Proceedings of the 18th International Conference on Computational Linguistics*.

Rapp, R.: 1996, *Die Berechnung von Assoziationen*, Vol. 16 of *Sprache und Computer*. Georg Olms Verlag.

Rapp, R.: 2002, 'The Computation of Word Associations: Comparing Syntagmatic and Paradigmatic Approaches'. In: *Proceedings of the 19th International Conference on Computational Linguistics*. Taipei, Taiwan.

Rooth, M., S. Riezler, D. Prescher, G. Carroll, and F. 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.

Rosario, B. and M. Hearst: 2001, 'Classifying the Semantic Relations in Noun Compounds via a Domain-Specific Lexical Hierarchy'. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Pittsburgh, PA.

Roth, M.: 2006, 'Relationen zwischen Nomen und ihren Assoziationen'. Studienarbeit. Institut für Computerlinguistik und Phonetik, Universität des Saarlandes.

Russell, W. A.: 1970, 'The complete German Language Norms for Responses to 100 Words from the Kent-Rosanoff Word Association Test'. In: L. Postman and G. Keppel (eds.): *Norms of Word Association*. New York: Academic Press, pp. 53–94.

Russell, W. A. and O. Meseck: 1959, 'Der Einfluss der Assoziation auf das Erinnern von Worten in der deutschen, französischen und englischen Sprache'. *Zeitschrift für Experimentelle und Angewandte Psychologie* **6**, 191–211.

Sahlgren, M.: 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.

Schulte im Walde, S.: 2000, 'Clustering Verbs Semantically According to their Alternation Behaviour'. In: *Proceedings of the 18th International Conference on Computational Linguistics*. Saarbrücken, Germany, pp. 747–753.

Schulte im Walde, S.: 2003, 'Experiments on the Automatic Induction of German Semantic Verb Classes'. Ph.D. thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart. Published as AIMS Report 9(2).

Schulte im Walde, S.: 2006a, 'Can Human Verb Associations help identify Salient Features for Semantic Verb Classification?'. In: *Proceedings of the 10th Conference on Computational Natural Language Learning*. New York City, NY, pp. 69–76.

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

Schulte im Walde, S.: 2006c, 'Human Verb Associations as the Basis for Gold Standard Verb Classes: Validation against GermaNet and FrameNet'. In: *Proceedings of the 5th Conference on Language Resources and Evaluation*. Genoa, Italy, pp. 825–830.

Schulte im Walde, S.: 2008, 'The Induction of Verb Frames and Verb Classes from Corpora'. In: A. Lüdeling and M. Kytö (eds.): *Corpus Linguistics. An International Handbook.*, Handbooks of Linguistics and Communication Science. Berlin: Mouton de Gruyter, Chapt. 61. To appear.

Schulte im Walde, S. and A. Melinger: 2005, 'Identifying Semantic Relations and Functional Properties of Human Verb Associations'. In: *Proceedings of the joint Conference on Human Language Technology and Empirial Methods in Natural Language Processing*. Vancouver, Canada, pp. 612–619.

Schulte im Walde, S. and A. Melinger: 2008, 'An In-Depth Look into the Co-Occurrence Distribution of Semantic Associates'. *Italian Journal of Linguistics. Special Issue on "From Context to Meaning: Distributional Models of the Lexicon in Linguistics and Cognitive Science"*. To appear.

Siegel, E. V. and K. R. McKeown: 2000, 'Learning Methods to Combine Linguistic Indicators: Improving Aspectual Classification and Revealing Linguistic Insights'. *Computational Linguistics* **26**(4), 595–628.

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

Stevenson, S. and E. Joanis: 2003, 'Semi-supervised Verb Class Discovery Using Noisy Features'. In: *Proceedings of the 7th Conference on Natural Language Learning*. Edmonton, Canada, pp. 71–78.

Tanenhaus, M. K., J. M. Leiman, and M. S. Seidenberg: 1979, 'Evidence for Multiple Stages in the Processing of Ambiguous Words in Syntactic Contexts'. *Journal of Verbal Learning and Verbal Behavior* **18**, 427–440.

Wettler, M. and R. Rapp: 1993, 'Computation of Word Associations based on the Co-Occurrence of Words in Large Corpora'. In: *Proceedings of the Workshop on Very Large Corpora*. Columbus, OH, pp. 84–93.