An in-depth look into the co-occurrence distribution of semantic associates

Sabine Schulte im Walde & Alissa Melinger

Semantic associations have served as a tool in cognitive science research for decades, and in recent years have also been of importance to computational linguists interested in a corpus-based induction of semantic relatedness. Within both areas, the co-occurrence hypothesis, which holds that free associations reflect co-occurrence in text, has underpinned much of the research that uses associations. However, few investigations have studied the properties of the associates in depth. In this paper, we scrutinise the co-occurrence hypothesis by exploring the distributional characteristics of a large set of stimulus-response pairs in a text corpus. By addressing the question from multiple perspectives we aim to extract more precise information about the generalisability of prior evidence in support of the co-occurrence hypothesis and the validity of many of the assumptions that grow out of the hypothesis.

1. Introduction

Semantic associations, namely words that are called to mind in response to a given stimulus, have been of interest to cognitive scientists for decades. Over the years, they have come to represent a window into knowledge representation, facilitating the development of empirically grounded models of semantic knowledge. Specifically, associations can be used as a tool to evaluate, estimate or describe the meanings of the respective stimuli. They have therefore been used to investigate the mechanisms underlying semantic memory, giving the researcher insights into the way semantic information is accessed and represented by the behavioural system.

Research questions that have capitalised on semantic associates range from memory research (e.g., Nelson et al. 1997; Nelson & Zhang 2000) and word recognition (cf. McNamara 2005) in experimental psychology to the development of semantic networks (e.g., Plaut 1995; Burgess 1998; Steyvers et al. 2004). Many of these research programs build on a linking assumption called the 'co-occurrence hypothesis'

(ricevuto nell'ottobre 2008)

(Miller 1969; Spence & Owens 1990), which holds that semantic association is related to the textual co-occurrence of the stimulus-response pairs. This hypothesis is the focus of the current research, as we explore directly the relationship between a range of elicited stimulus-response pairs and the context in which they occur in language.

The notion of co-occurrence distributions and their semantic interpretation have also been of increasing importance to computational linguists interested in semantic relatedness: for many NLP resources and applications, it is crucial to define and induce semantic relations between words or contexts. These tasks include the creation of ontologies (Maedche & Staab 2000; Navigli & Velardi 2004; Kavalek & Svate 2005), anaphora resolution (Vieira & Poesio 2000; Ji et al. 2005), question answering (Moldovan & Novischi 2002; Girju 2003; Girju et al. 2006), and textual entailment (Geffet & Dagan 2005; Tatu & Moldovan 2005). As the automatic acquisition of semantic knowledge from corpus data is not trivial, a common approach is to rely on distributional cues, thus exploiting the connection between co-occurrence distributions and semantic relatedness. Furthermore, in order to evaluate the computational models, many researchers within that area have identified the value of human data to their task; among them is work that used free association norms as a test-bed for distributional models of semantic relatedness (Church & Hanks 1990; Rapp 1996; Rapp 2002; Lemaire & Denhière 2006; Schulte im Walde 2006).

The approach we take in this article is to conduct a descriptive and in depth examination of the distributional properties of stimulus-associate pairs across context windows, based on a large set of semantic associates elicited to German verbs. Our approach is primarily descriptive in that we do not apply inferential statistics. Reported differences are based on observed numerical patterns but tests to establish the generalisability of differences are not included. We ask a simple empirical question: what proportion of associate responses is observed in the context of their respective stimulus verbs? This is essentially the same question asked by Spence and Owens (1990), who supplied the first empirical support for the co-occurrence hypothesis. However, we engage this question along several dimensions. In addition to replicating the basic experiments by Spence & Owens, we also break the analysis down into various categories which have been independently identified as distributionally interesting (e.g., Deese 1965; Clark 1971; McEvoy & Nelson 1982; Schulte im Walde et al. 2008), such as association strength, corpus frequency of the stimuli, response part-of-speech, etc. Furthermore, we add analyses that question some of the intuitive conclusions from early work on the co-occurrence hypothesis.

The remainder of this article is organised as follows. In Section 2, we start with an overview of association norms, and of previous work that was concerned with co-occurrence analyses of association norms, as a general motivation for the experiments to follow. Having introduced our own association norm, the corpus, and the co-occurrence method in Section 3, Section 4 then presents the experiments, broken down into a basic set of experiments investigating the roles of co-occurrence window size, prior frequencies, and corpus size (Section 4.1); experiments that combine the factors window direction and parts-of-speech of responses (Section 4.2); and experiments shedding new light on the chain effect within association norms (Section 4.3). We conclude with a discussion of the implications of this work as well as some open issues in Section 5.

2. Association Norms and Co-Occurrence Distributions

The main body of this article starts with an overview of previous work related to semantic associations. First, we provide a brief description of association norms in general terms, and then we continue with research that specifically addressed the co-occurrence between stimulus-response pairs. Some studies with direct relevance to our analyses will be further discussed when we motivate each individual experiment.

2.1 Overview of Association Norms

Association norms are created within two steps. The first step captures the collection of associations to stimuli: After a set of target stimuli, also known as cues, has been defined (for specific parts-of-speech or across part-of-speech, controlling for e.g. the number of syllables, corpus frequency, or semantic category, depending on the purpose of the norms, etc.), participants are requested to provide the first word(s) that come to mind when presented with the stimuli [1]. To create the actual norms from the associate responses, a second step then quantifies over the number of response tokens for each stimulus-response (SR) pair. The result is called an association norm.

Association norms have a long tradition in psychological research. Following an idea originally suggested by Francis Galton in 1880, the first association norms were collected by Kent & Rosanoff (1910), for

100 English noun and adjective stimuli. The Kent & Rosanoff stimuli were translated to German, allowing for the collection of parallel association norms in German (Russell & Meseck 1959, Russell 1970). Association norms for 43 taxonomic categories were first published by Cohen et al. (1957). Another well-studied collection was assembled by Palermo and Jenkins (1964), comprising associations for 200 words across various parts-of-speech. 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 & Jenkins norms. Similarly, the association norms from the University of South Florida (Nelson et al. 1998) were compiled over the course of more than 20 years. Their goal was to obtain the "largest database of free associations ever collected in the United States available to interested researchers and scholars". More than 6,000 participants produced nearly three-quarters of a million responses to 5,019 stimulus words. Smaller sets of association norms have also been collected for example for Dutch (Lauteslager et al. 1986), French (Ferrand & Alario 1998), Italian (Peressotti et al. 2002; Guida & Lenci, 2007) and Spanish (Fernandez et al. 2004) as well as for different populations of speakers, such as adults vs. children (Hirsh & Tree 2001).

2.2 Association Norms and Co-Occurrence Distributions

The first empirical support for the co-occurrence hypothesis was supplied by Spence & Owens (1990, hereafter S&O). They tested the hypothesis by searching corpora for the co-occurrence of strongly related semantic associates. S&O used a corpus of 1 million words of English. The stimulus-response pairs were drawn from Palermo & Jenkins (1964) and consisted of 47 concrete noun targets with a frequency of occurrence > 10 per million. As responses, only the most strongly associated noun responses were considered. Thus, their set consisted of noun-noun pairs with relatively high association strengths. Using a co-occurrence window of 250 characters, they found a negative correlation between distance and association strength. They also compared the co-occurrence frequencies of associates to frequency-matched unrelated word pairs and they found significantly higher rates of co-occurrence for the related words than unrelated words. Similarly, Justeson & Katz (1990; see also Charles & Miller 1989) observed that antonymous adjectives like 'dry' and 'wet' co-occurred at least once in the same sentence in the 1 million word Brown corpus. However, they did not evaluate whether antonyms co-occurred more often than other semantically related adjectival pairs such as 'wet' and 'damp' or even unrelated pairs.

S&O's work has stood as strong evidence for the relationship between free association norms and textual co-occurrence. However, while the study was rigorously conducted, its breadth of coverage restricts interpretability. Specifically, their study was limited in the following ways: a) it only considered noun-noun SR pairs, b) it only considered words in a relatively high frequency band, c) it only considered strongly associated SR pairs. Furthermore, they were not interested in the descriptive characteristics of the SR co-occurrence distributions; they did not look into the types of relationship, either syntactic or semantic, expressed by the word pairs. Thus, several questions remain: To what extent do their findings generalise to a wider range of SR pairs; Do other distinctions over the types of responses give additional insights into the relationship between semantic association and textual co-occurrence; What is the co-occurrence distribution of SR pairs across context windows of varying sizes?

The co-occurrence hypothesis has also played a large role in the development of NLP models of semantic relatedness. Church & Hanks (1990) turned the co-occurrence hypothesis on its head by using text co-occurrences to try to predict semantic associations. They proposed that the information-theoretic concept of mutual information could substitute for association strength. Thus, they used co-occurrence frequencies combined with mutual information to identify strongly associated words. Their study was motivated by lexicographic purposes and thus focused on relation types that a) capture pairs found in coordinated structures with a fixed order, e.g., *bread and butter*, b) compounds, e.g., *computer scientist*, c) featurally-similar pairs which would also occur in coordinated structures but without a fixed order, e.g., *man and woman* or *woman and man*, and finally d) grammatically-based word pairs, e.g., phrasal verbs and direct objects of verbs. They demonstrated that mutual information captured the various types of semantic association a-d), and could therefore be useful for a number of NLP tasks. Many subsequent studies have followed on this approach, relying on information-theoretic measures to predict strongly associated word tuples in corpora. Emphasis within this line of research has been on identifying optimal statistical measures (Dunning 1993; Evert 2005), and on the automatic acquisition of semantically strongly linked words (e.g., Lin 1998; 1999). In more general terms, Church and Hanks' work can therefore be regarded as a milestone with respect to corpus-based, distributional models of semantic relatedness, cf. also the various examples in Section 1.

Within a related line of research that explicitly included association norms, Wettler & Rapp (1993) defined a statistical model that predicted stimulus-associate pairs in English and German, and compared the predicted associations against association norms. Subsequent work presented various extensions of their basic model and application scenarios (Rapp 1996). Example applications of their model were the generation of search terms in Information Retrieval, and the prediction of marketing effects caused by word usage in advertisements. Recently, Rapp suggested methods that capture and distinguish paradigmatic vs. syntagmatic associations (Rapp 2002), again evaluating the models against association norms. Similarly to Rapp's model of association, other lines of research have modelled the semantic relatedness within association norms, e.g., by making use of the vector space model (Lund & Burgess 1996; Lowe & McDonald 2000), Latent Semantic Analysis (Griffiths & Steyvers 2003), other higher-order models of co-occurrence (Lemaire & Denhière 2006), and psychological models of learning theories (Seidensticker 2006).

Our own previous work is situated between the psychological and the computational lines of research: We do make use of large-scale computational resources and methods, but not to predict, but rather to analyse the association norms. The insights are also thought to contribute to both cognitive and computational linguistic issues. For example, Schulte im Walde (2006) relied on the verb association norms used in this article to improve the feature choice in verb clustering experiments. Schulte im Walde et al. (2008) used the verb association norms and also noun association norms, both for German, and provided detailed analyses of the grammatical verb-noun functions, and the intra-categorical semantic relationships of the verb-verb and noun-noun pairs within the norms. The insights contribute to NLP questions concerning representations in distributional semantics, and the types of semantic relationships relevant for NLP applications. In a similar vein, Guida (2007) replicated most of our analyses on association norms for Italian verbs.

3. Data Collection, Corpus Resource, and Co-Occurrence Model

3.1 Data Collection

The association norms that are applied within this article were retrieved for German verb stimuli. The data collection of the verb associations was performed as a web experiment, which asked native speakers to provide associations to German verbs.

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., cf. Schulte im Walde (2008: Appendix A) for an overview and example classes. Drawing verbs from different categories was intended primarily to ensure that the experiment covered a wide variety of verb types; the inclusion of any verb into any particular verb class was achieved in part with reference to prior verb classification work (e.g., Levin 1993) but also on intuitive grounds. The stimulus verbs were divided randomly into 6 separate experimental lists of 55 verbs each. The lists were balanced for class affiliation and frequency band (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, a portion of the 200 million word corpus used for our co-occurrence analyses; the verbs showed corpus frequencies between 1 and 71,604.

Procedure

The experiment was administered over the Internet. Participants were first presented with written instructions for the experiment and an example item with potential responses. In the actual experiment, each trial consisted of a verb, presented in the infinitive, displayed in a box at the top of the screen. Below the verb was a series of data input lines where participants could type their associations. They were instructed to type at most one word per line and, following German grammar, to distinguish nouns from other parts-of-speech with capitalisation. Participants had 30 seconds per verb to type as many associations as they could (cf. Battig & Montague 1969). After this time limit, the program automatically advanced to the next verb.

Participants and Data

299 native German speakers participated in the experiment, between 44 and 54 for each data set. In total, we collected 79,480 associates distributed across 39,254 different response types; considering the first per stimulus and participant only, our data comprises 15,788 associates distributed across 7,425 types. Each trial elicited an average of 5.16 associate responses with a range of 0-16. Each com-

pleted data set contains the list of stimulus verbs, paired with a list of associations in the order in which the participant provided them.

Association Norms

To create the actual association norms, we quantified over all first responses from the participants for each stimulus verb. Taking the stimulus verb *klagen* `complain, moan, sue' as an example, 11 participants provided *Gericht* `court' as their first response, 6 provided *jammern* 'moan', another 6 provided *weinen* `cry', etc. The frequencies of the responses over all participants represent the number of stimulus-response tokens (i.e., the 'association strength'*)* of the respective stimulus-response type.

Armed with this single dataset of stimulus verb-response pairs, we now aim to ask the questions: Do semantic associates co-occur with their respective targets? If so, what is the distribution of their co-occurrence across a context window? What types of associate responses are most likely to co-occur closer to their targets and which further away? While many of these questions have been previously addressed in the literature, many of them have not. Furthermore, the individual analyses conducted by different researchers have been applied to different data sets. Thus, one benefit of the present analysis is that all the analyses will be conducted on the same set of data, thus affording maximum comparability. Also, many of the previous investigations used relatively small numbers of SR pairs. Here, we test the entire data set of the 15,788/79,480 SR pairs. However, the use of this particular set of norms has its drawbacks as well. Particularly, the exclusive use of verb stimuli introduces some restrictions that may not be optimal. Also, compared to many other norms, relatively few participants took part. The result is that our data includes a comparatively high proportion of idiosyncratic responses.

3.2 Co-Occurrence Method

In light of the general preference to only consider the first response to a given target, our co-occurrence approach considers the first responses to the stimulus verbs only. The empirical support for this preference is reviewed in Section 4.3 where we also expand the analysis to consider multiple responses in an investigation of association chaining (experiment 6). The basic idea of our co-occurrence analyses is as follows. We address the question 'what proportion of our 15,788 first response tokens is observed in the context of their respective target stimuli'. To answer the question, we used a corpus of approximately 200 million words of German newspaper text. Punctuation did not contribute to the context windows, but all other words (including function words) did contribute. We searched the corpus for each pair of stimulus and response tokens, whether they occurred together within a certain window size of a sliding context window, and how often, throughout the whole corpus. Since we were interested in the contribution of various window sizes to the co-occurrence hypothesis, we used window sizes between 1 and 25 words. We did not distinguish between the direction of the response word in the corpus with respect to the stimulus word (except for experiments 5 and 6), so we used ±1-word up to ±25-word context windows. As a result, we can specify the 'co-occurrence strength' (i.e., how often stimulus and response occurred together) with respect to a specific stimulus-response pair and a specific window size. Finally, we can distinguish between two views on the window size: a) an 'inclusive' view of window size *x,* where every window size ≤*x* contributes to the co-occurrence strength, to answer the question 'which proportion of SR pairs is covered in total by the corpus co-occurrence'; b) an 'exclusive' view of window size *x,* where only the individual window size *x* contributes to the co-occurrence strength, addressing the question of 'which proportion of SR pairs is exactly at a distance of x'. The different views are applied with respect to the specific questions we ask in the various experiments.

Consider the following example. Imagine that *ice* was a common response type to the verb target *freeze*, with an association strength of 12, i.e., provided as a first response by 12 participants. We would search the corpus for all occurrences of *freeze* and determine whether *ice* occurred within ±x words. For instance, the string [*ice* did not *'freeze'* in this warm] is one instance when the response word is found within ±3 words of the target. If this co-occurrence of *freeze* and *ice* were the only instance in the corpus, 12 out of 15,788 tokens would contribute to a co-occurrence strength of 1 in a ±3-word window. In other words, we say that 12 tokens out of 15,788 SR tokens are observed to co-occur in the corpus at least once in a window of ±3-words. With an inclusive view on the window size, this co-occurrence of *freeze* and *ice* also contributes to all window sizes between ±3 and ±25 words, because the smallest window size it appeared in is ±3 words. With an exclusive view on the window size, this co-occurrence only contributes to the window size of ±3 words, thus looking into specific window sizes. We calculated co-occurrence between stimuli and response tokens (rather than types) because we wanted to evaluate the co-occurrence distribution for the entire population of SR pairs, not for SR pair types. Thus, if one very common pair represented 10% of the entire set of SR pairs, and it was the only pair we observed in the corpus, then our coverage would be 10%.

4. Co-Occurrence Experiments

The main section of this article presents a series of co-occurrence experiments, relying on the association norms, the corpus resource, and the co-occurrence method described above. The experiments are organised into sets, starting with a basic experiment setup that is subsequently varied on conditions that refer to corpus criteria and linguistic properties of the stimuli and the responses, which are expected to influence the co-occurrence analyses, or shed light on specific linguistic questions.

4.1 Basic Experiments

Experiment 1: *basic experiment on stimulus-response co-occurrence*

This basic analysis represents a new way of addressing the question originally asked by Spence & Owens (1990). As explained above, we are interested in the proportion of our stimulus-response pairs that co-occurred in the corpus within different window sizes. The basic plots in Figure 1 both show the proportions of SR tokens (y-axis) within a window of ±1-25 words (x-axis) with a corpus co-occurrence strength of at least 1x, 5x, 10x, and 20x (plotted lines). The figures evaluate a) 'what proportion of the data is observed in our corpus in various window sizes' and b) 'how frequently the pairs co-occur'. The lines are cumulative, i.e., the SR pairs that co-occurred at least 5x in the corpus also co-occurred at least 1x, etc. The figure is divided into two panels, with the left panel describing the window sizes 'inclusively', i.e., taking all windows up to a window size *x* into account, and the right panel describing the window sizes 'exclusively', i.e., looking into a window size *x* without considering preceding windows, as explained above.

As can be seen in the left panel of Figure 1, our simplest analysis supports the co-occurrence hypothesis. With a corpus co-occurrence strength threshold of 1, we see that more than 50% of our SR pair tokens are immediately adjacent to each other at least once in the corpus. And even with a higher threshold of 5, almost 30% of the SR pairs are still immediately adjacent to each other. In total, we cover ca. 85% of the SR pairs with a co-occurrence of at least 1, and ca. 70% of the SR pairs with a co-occurrence of at least 5. Even with our strictest co-occurrence strength threshold of 20, more than 50% of the SR pairs are observed within a ±25-word window. The slope of the distribution curve flattens out in the larger window sizes, indicating that few responses are observed in larger context windows that were not already observed in smaller windows. However, since the left panel plot relies on an incrementally growing window, we cannot say how many responses occurred at each individual word position. The right panel of the figure addresses this question. It shows that - for all co-occurrence strengths - more SR pairs are found in smaller windows, and less SR pairs are found in larger windows. The decrease in the proportion of SR pairs observed in proximal compared to distal windows is only 4-7%, thus larger windows also contribute to the link between co-occurrence and free association.

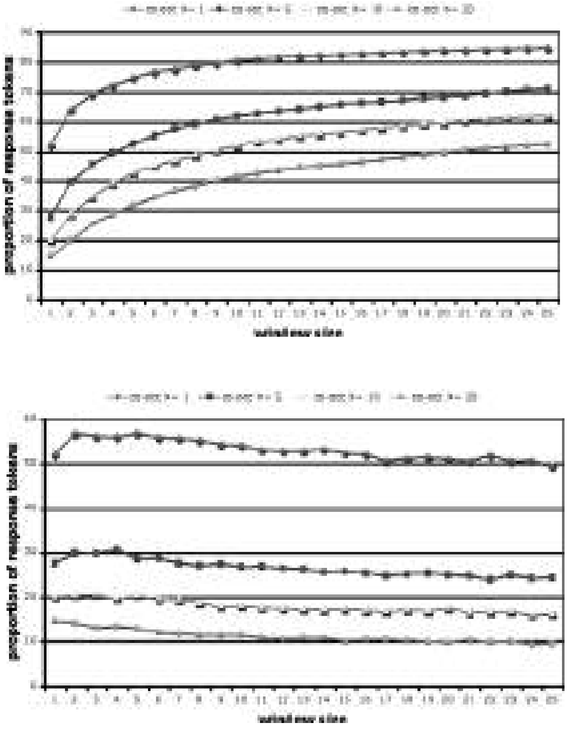


Figure 1. Proportion of SR pairs in ±25-word window co-occurrence, distinguished by co-occurrence strength of at least 1, 5, 10, or 20.

EXPERIMENT 2: *basic experiment, corrected for random co-occurrences*

Experiment 1 demonstrated that SR pairs do co-occur in text at consistently high rates, in close proximity and also in increasing windows. However, there is a missing element to the analysis that must be addressed to fully appreciate the extent of SR textual co-occurrence, and that is to establish a baseline estimated by the co-occurrence rate of unrelated words. So far, we have implicitly assumed that two words co-occur in a corpus because they are semantically related, and that semantically unrelated words do not co-occur. To correct this implicit assumption and estimate the degree of random co-occurrence in our data, a variant of the first experiment takes co-occurrence by chance into account, following an idea by S&O. We created an artificial set of stimulus-response pairs, where for each original stimulus-response pair type, we replaced the response by another word, randomly chosen from the words in our corpus but matched for part-of-speech and corpus frequency. No other filter was placed on the random selection of unrelated responses; for example, we did not artificially skew the set away from random semantic relations. For example, the SR type *abstürzen - Flugzeug* ('crash' - 'airplane') was replaced by *abstürzen - Erkenntnis* ('crash' - 'awareness'), with a corpus frequency of 581 for *Flugzeug,* and 582 for *Erkenntnis.* We then applied the same search criteria as in Experiment 1 to create a baseline for co-occurrence rates. This baseline is presented in the two panels of Figure 2a, and the results for the basic experiment corrected for random co-occurrences (by subtracting the baseline from the original values) are presented in the two panels of Figure 2b. Each of the two figures is arranged in parallel to Figure 1, with the inclusive windows in the left panel, and the exclusive windows in the right panel.

The left panel of Figure 2a shows that the shapes of the plotted SR proportion lines are impressively similar to the results from Experiment 1. However, as one would expect, the coverage of the semantically unrelated words is much lower than the coverage of the semantically related words, with differences ranging from 12-44%. Furthermore, the slopes of the lines as window size increases are slightly steeper than in the original plot. This can be seen more clearly in the right panel of the figure, looking into the SR proportions found at each specific window size: In contrast to the original plot, we find relatively stable rates of co-occurrence across all the windows, with a slight increase in the proportion of co-occurring SR pairs as window size increases, in contrast to the slight decrease seen in

Figure 1. This means that semantically related words tend to co-occur in smaller windows relatively more often than semantically unrelated words.

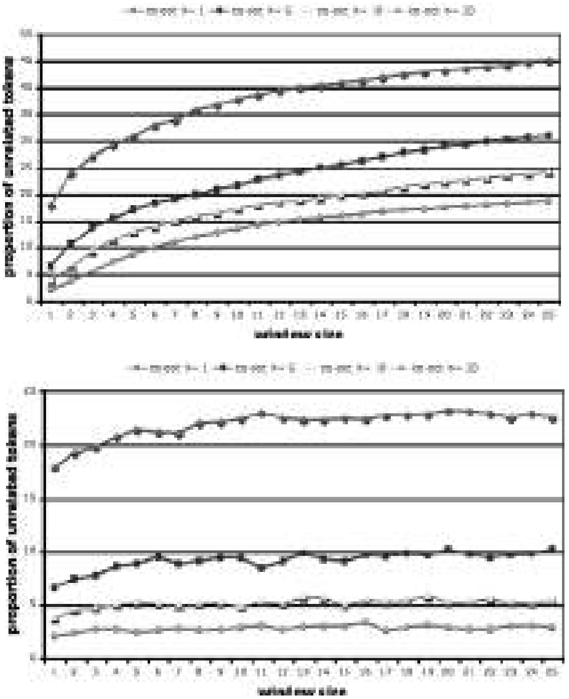
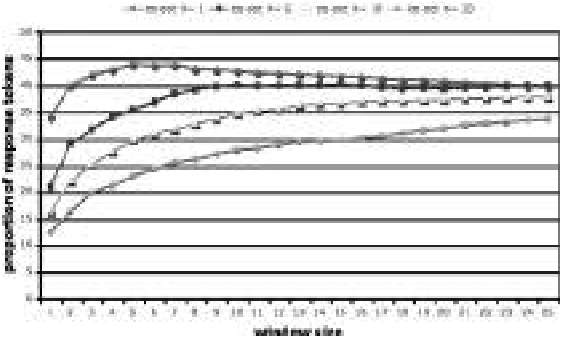


Figure 2a. Proportion of *stimulus-unrelated pairs* in ±25-word window co-occurrence, distinguished by corpus co-occurrence strength.

Using the plots of unrelated responses in Figure 2a to correct the plots of related responses in Figure 1 by subtraction, we can now revisit the window-related properties of SR co-occurrence proportions. The left panel of Figure 2b shows that – taking random co-occurrences into account – 34% of our SR pairs are still immediately adjacent to each other at least once in the corpus, and more than 20% are immediately adjacent to each other at least 5 times. The steep increase in coverage in the small windows is again evident in this graph, especially for the lines reflecting lower co-occurrence thresholds. For thresholds 1x and 5x, the coverage actually decreases as the window size increases, reflecting the result of subtracting the baseline[2]. This observation is more clearly seen in the right panel of Figure 2b. Taking the baseline into account, we observe larger proportions of SR pairs in smaller window sizes, and the proportions decrease with an increasing window size. However, the co-occurrence rates at every window remain above chance and the proportion of coverage drops only moderately across the 25-word window, e.g., with coverage dropping from a peak of 23% to a low of 14% for co-occurrence threshold 5. The peaks of the lines are at proximal windows 1 or 2, depending on the co-occurrence threshold.



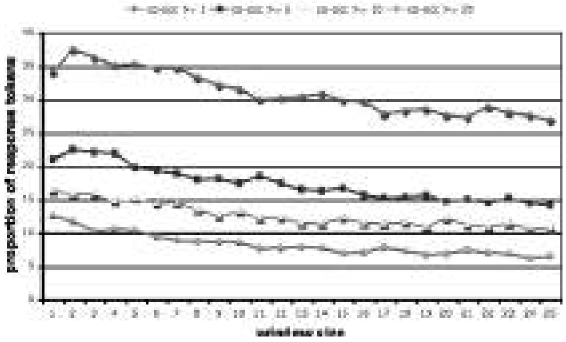


Figure 2b. Proportion of SR pairs corrected by stimulus-unrelated pairs.

Summarising the observations from the basic experiments 1 and 2, the different plots have already addressed the fundamental co-occurrence questions:

- 'Does the co-occurrence hypothesis apply to our association norms?' Figure 1 showed that 85% of our SR tokens were observed at least once in a co-occurrence window of ±25 words, and 71% at least 5 times, which supports the co-occurrence hypothesis for our data.
- 'Is the confirmation of the co-occurrence hypothesis above chance?' Figure 2b showed that - taking random co-occurrences into

account - we still found 44% of our SR tokens at least once and 40% at least 5 times on top of what could be explained purely by chance.

- 'What is the relationship between the co-occurrence hypothesis and window size?' Looking into the individual window sizes, the plots showed that a larger proportion of our SR pairs co-occurred in smaller than in larger windows. This finding complements the negative correlation between association strength and distance observed by S&O.

Having provided support of the basic co-occurrence hypothesis, the first experiments on SR co-occurrence at the same time offer potential for more specific questions, addressing the co-occurrence distributions under various conditions. As outlined earlier, we will continue with analyses that arose from these basic experiments. However, as an intermediate step, we first discuss two analyses that address caveats with respect to our basic findings, and thus have a slightly different status in comparison to the other, hypothesis-based, experiments in this article. We present these caveats because even though they represent common knowledge, they have only partly been taken into account in previous work.

Caveat A: taking target and response frequencies into account

Large co-occurrence thresholds are more likely to be satisfied by SR pairs with high frequencies: clearly the more often two words occur in the corpus the greater the chance that they will co-occur. We have tried to address this point with a frequency-matched unrelated word baseline. But we were also interested in exploring the impact of the prior empirical distributions of the target and/or the response words. Firstly, how does the corpus frequency of the target word influence the co-occurrence distribution of the responses? And likewise, how does the corpus frequency of the response words influence their co-occurrence distribution? We investigated the proportion of co-occurring SR pairs in relation to the target verb frequency [3]. As expected we find that the higher the corpus frequency of the target, the more co-occurring SR pairs are observed. For example, high frequency words satisfy the relatively low threshold of 5x easily; thus they present a steep rise in coverage in the smallest windows and asymptote early. In contrast, low frequency words require a larger window to satisfy the threshold; new responses continue to be observed (and contribute to the threshold) even in the later windows.

The prior corpus frequencies of stimuli and responses thus have a strong influence on the proportions of SR pairs with corpus co-occurrence, which is not surprising because - as mentioned before - the more often a word occurs in a corpus, the higher the chance to find a co-occurrence with another word. This analysis thus highlights the importance of controlling for priors, as we did in experiment 2. Our approach in subsequent experiments will therefore continue to compare the co-occurrence rates of related SR pairs to their frequency-matched unrelated SR pairs, whenever possible. Furthermore, as there are other possibilities than the S&O baseline to take the target and response corpus frequencies into account (cf. Evert 2005, among others), Experiment 3 will add an analysis where corpus frequencies are incorporated into the association strength, relying on the log-likelihood measure.

Caveat B: taking corpus size into account

A corpus co-occurrence analysis is directly dependent on the properties of the corpus that is checked for co-occurrence. This is true not only for the corpus size and the corpus coverage (of the items under consideration), but also with respect to the corpus domain and its homogeneity, (cf. McEnery 2003). This concern is well-known within computational linguistics (cf. Banko & Brill 2001) but has not yet explicitly been illustrated in the context of the co-occurrence hypothesis.

Regarding our corpus resource, the domain of the corpus is newspaper data and therefore it under-represents slang responses like *Grufties* `old people' and *lümmeln* `loll', dialect expressions such as *Ausstecherle* `cookie-cutter' and *heimfahren* `go home', as well as technical expressions such as *Plosiv* `plosive'. Despite the homogeneous domain, we nevertheless find 99% of target and response words at least once in the corpus. Thus, we can assume that, for a co-occurrence threshold of ≥1, we have a theoretical upper limit of 99% co-occurrence coverage for our experiments.

To get a better sense of the influence of corpus size, we explored the contribution of increased corpus size with regard to the coverage of our SR pairs. This analysis inspects the influence of word frequency from another perspective, by manipulating the size of the corpus. We asked what proportion of our SR pairs are covered by 10%, 20%, 30%, etc., of the size of the corpus we have available. We restricted the analysis to SR pairs with corpus co-occurrence frequencies ≥5 in each context window. The results reveal that if we had only used e.g. 20 million words instead of 200 million words, a ±25-word window would have covered less than 40% of all SR tokens, as compared to more than 70% of SR tokens covered by 100% of the corpus.

Unsurprisingly, as the corpus size approaches the full 200 million words, the increments of improved coverage decrease. This suggests that we approach our ceiling for coverage. This is likely due to the restricted and homogenous domain of the corpus. However, a second factor contributing to the ceiling effect may have to do with SR pairs that reflect what Schulte im Walde et al. (2008) characterised as ‘world knowledge’. For example, Schulte im Walde et al. observed SR pairs such as *auftauen - Wasser* (‘defrost’ - ‘water’) which they argued captured aspects of the target word’s meaning that would be unlikely to co-occur in text. Summarising the results of this caveat, we believe that it is important to check the restrictions of the corpus with respect to the data under consideration; in our case, we found the proportions of targets and responses covered by the corpus (both 99% for co-occurrence threshold ≥ 1), and an approximate ceiling of the SR pair co-occurrence (70%).

EXPERIMENT 3: *taking prior corpus frequencies of stimuli and responses into account*

Experiments 1&2 highlighted the importance of controlling for frequency effects caused by priors, thus leading us to establish a baseline for our experiments. This experiment investigates the effect of the prior corpus frequencies of targets and responses. Specifically, we measured the log-likelihood (llh) to gain insights into the idiosyncratic distributional patterns for individual SR pairs. Log-likelihood is one of a large number of association measures that take the prior frequencies of events into account. In essence, it assesses the extent to which two words co-occur in a corpus more or less often than their individual frequencies and the corpus size would predict. If two words occur more often than expected, it suggests a tight semantic link; if they occur less often than expected, it suggests the words are unrelated and drawn from different semantic domains. Llh was first suggested by Dunning (1993) as a suitable association measure for Natural Language Processing tasks, because – among other reasons – it is less vulnerable to the pervasive sparse data problem than e.g. mutual information as used by Church & Hanks (1990). See www.collocations.de/AM/ for an overview of association measures. Given a specific window size, it is possible to compute a standard two-way contingency table for the events *i* (the target verb) and *j* (the associate). Based on the contingency table, we calculated the log-likelihood values according to Evert (2005) as _2 *ij (Oij * log Oij/Eij),_ where *Oij* refer to the observed frequencies of the events *i* and *j* within the contingency table. The expected frequencies *Eij* are calculated as a product of the respective *Oij* marginals (which take the total frequencies of *i* and *j* into account), normalised by the total frequency *N*, the total number of co-occurring events within the respective window size: *N* = *ij Oij*.

Unlike co-occurrence frequencies, log-likelihood values do not monotonically increase by window size. Rather, for each SR type and each window size, the llh value informs us whether the observed co-occurrence frequency of two items is larger than expected by the marginals. Thus, for a specific SR type, the llh value might increase or decrease when increasing the window from ±n to ±n+1 words, as the marginals change as well. The different behaviour of corpus co-occurrence frequencies and llh values is nicely illustrated by Figure 3. On the y-axis, the left panel plots the co-occurrence frequencies of example stimulus-response pairs (not corrected for random co-occurrences), and the right panel plots the llh values. Separate lines correspond to the data for individual SR pairs. In the frequency plot we see a similar pattern for all 4 data points; the co-occurrence frequencies start low and increase as the window size increases, taking new observations into account window by window. The pattern is different for the llh plot on the right. For example, for strong collocations such as *die Wahrheit sagen* 'say the truth', (where the stimulus-response pair is *sagen – Wahrheit*) we find (initially high) llh values that decrease with an increasing window, indicating that the two lexical items are more strongly associated in nearby positions, in this case with a peak at window size ±2. We see a similar pattern for the verb-adverb combination *verteilen gerecht* 'distribute equitably', again suggesting that the adverb appears close to the verb. In comparison, for *basteln* 'do handicrafts' - *malen* 'paint', the llh value also increases for a window size of ±2, but only slightly decreases with an increasing window size, suggesting that the two verbs are associated not only in nearby but also further context positions. Differently to the previous three cases, for *kochen* 'cook' – *essen* 'eat', the llh value increases with an increasing window size, indicating that the two lexical items are associated, but not necessarily constrained to occur in very proximal positions.

Summarising, log-likelihood values for word pairs direct us not only towards strongly associated words (such as our stimulus-response pairs), but when combined with window size, they also indicate some 'typical' distances between the associated words. The distances in return can be used to make hypotheses about the relationships between the words: words near each other tend to be related by some syntactic function, depending on the parts-of-speech of the words under consideration (e.g., typical adverbs of verbs, typical direct object nouns of verbs), whereas larger distances between strong llh-related word pairs might point towards situational or even world knowledge. These insights are partly well-known. As discussed earlier, ever since Church & Hanks, research on the automatic induction of collocations of various types from corpus data has relied on association measures that compared observed co-occurrences with expected co-occurrences in some way, usually with respect to a syntactic word-word relationship of interest. However, to our knowledge, association measures have not generally been used to identify associated word pairs at longer distances, which we consider an interesting contribution towards identifying situational knowledge.

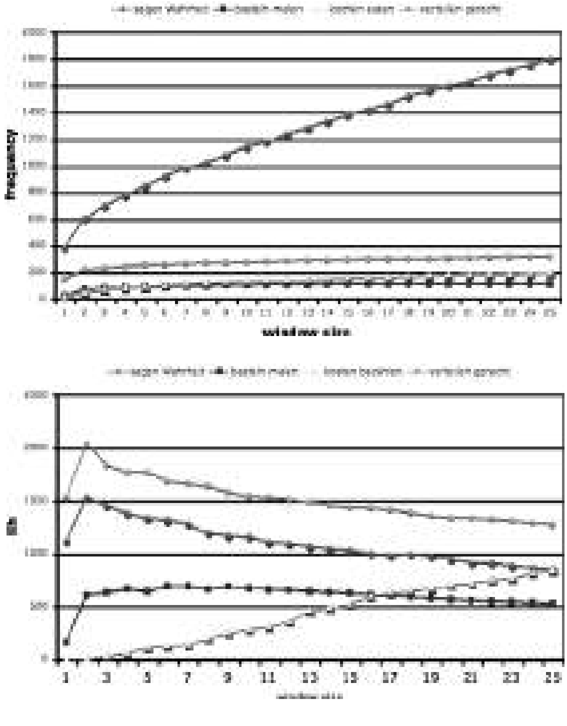


Figure 3. Frequencies and llh values for selected SR examples.

4.2 Experiments on Window Positions, using Direction and Part-of-Speech

The second set of experiments investigates various window positions in more detail. We consider the possibility that certain window positions might be prominent for a particular type of SR relationship, inspired by work concentrating on specific morpho-syntactic relationships (in the vein of Church and Hanks). Without restricting the focus of our investigation to a particular window, we explore whether specific window positions provide insights on functional relationships between stimuli and responses. The following two experiments therefore distinguish the window direction (left vs. right, experiment 4) and the parts-of-speech of the responses (experiment 5), with a final step bringing the two experiments together.

Experiment 4: *taking the window direction into account*

Until now, all our context windows have been plus or minus *n* words of the target stimulus. This means that our context window conflates over responses that precede the target in the context window and those that follow the target. However, some views of semantic association suggest that target stimuli might elicit continuations rather than preceding text. For example, Plaut (1995) defined association as the likelihood of one word to follow another in text (see Moss et al. 1994, for a similar approach). Previous work on co-occurrence has largely ignored the window direction. To our knowledge, only Church & Hanks (1990) explicitly included the search direction into the co-occurrence models, accounting for syntagmatic association pairs that occur in a fixed order (such as 'bread and butter', 'thunder and lightning', or particle verbs such as 'sit on', 'call over', where particle and verb require an order). Hence, in the next analysis, we ask if one direction is more important for SR co-occurrence.

For this analysis we again concentrate only on responses that co-occur with their respective targets at least 5x in each context window, but we distinguish the direction of the co-occurrence. The results of this analysis are reported in Figure 4; both panels report the SR proportions corrected for random co-occurrences. The left panel displays the inclusive proportions across context windows, and the right panel illustrates the proportion of SR pairs that occur in each individual context window exclusively. As can be seen from the left part of Figure 4, more responses preceded their targets rather than followed them, a pattern also observed in the uncorrected distribution. This difference is prominent in the earliest context windows of 1-3 words but is essentially neutralised in larger windows. In fact, the right panel reveals that the difference is largely due to window position 1. Here we see an over-utilisation of the position immediately preceding the target and an under-utilisation of the position immediately following the target.

This pattern runs counter to the hypothesis that targets, as response cues, trigger the production of possible continuations. The fact that we included function words in our analysis may contribute to the difference, as a determiner is likely to occur immediately after a verb while a noun or other part-of-speech can often precede a verb directly. However, even if the difference is fully accounted for by the inclusion of function words, we still see no evidence for right-window dominance in the response distribution. This experiment therefore extends on prior investigations of the co-occurrence distribution of SR pairs by examining the direction of the context window. The addition of this factor allows us to examine predictions about the uni-directionality of associate formation as well as to form and test hypotheses about the contribution of function words to the distribution curves. The second part of Experiment 5 will shed more light on these finding, distinguishing between the parts-of-speech of the responses.

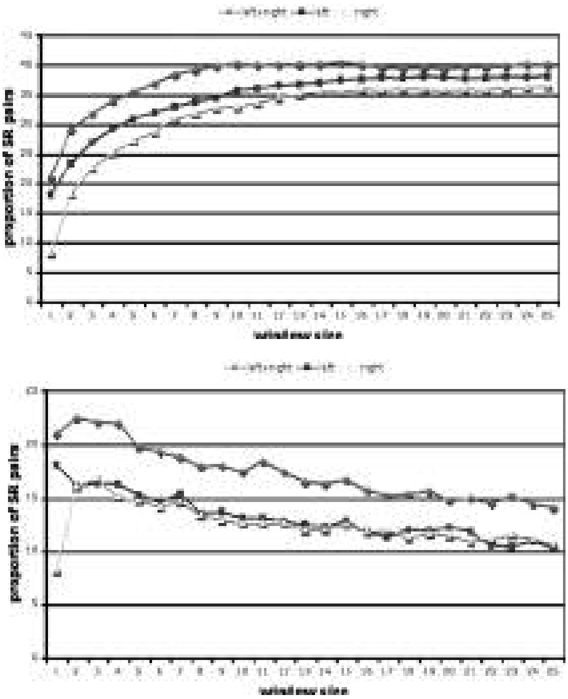


Figure 4. Proportion of SR pairs for left vs. right window.

Experiment 5: *co-occurrence distinguishing the parts-of-speech of responses*

In this analysis, we explore whether SR pairs are more likely to co-occur in the corpus when the response comes from a particular part-of-speech (POS). Additionally, we explore whether differences in response POS lead to changes in the distribution across the various context windows. For example, as all of our target words were verbs, we might expect that the nominal responses reflect arguments of the verb and as such should occur in closer proximity. In contrast, verb responses might occur in close distance from the target verb (in the case of conjunctions, for example), or occur at more distance from the target verb, when they are positioned in separate clauses.

In a preparatory step, each response was automatically assigned its (possibly ambiguous) part-of-speech, relying on a corpus-based empirical dictionary (Schulte im Walde 2003: chapter 3). We considered only the major categories verb (V), noun (N), adjective (ADJ) and adverb (ADV), disregarding fine-grained distinctions such as case and number. The total number of our 15,788 SR pairs distributes over the POS as follows: 8,838 responses (56%) were nouns, 5,355 responses (34%) were verbs, 1,178 responses (7%) were adjectives, and 199 responses (1%) were adverbs. More details about the POS distribution of our responses are reported in Schulte im Walde et al. (2008).

Figure 5a shows the proportions of SR pairs distinguished by the part-of-speech of the responses. As before, we used a co-occurrence threshold of 5x to compare the co-occurrence distributions with respect to their POS. For comparison, we include the line for all POS responses in the graph, which corresponds to the data presented as ≥5 in Figure 2b. The left panel of the figure describes the overall coverage of the responses, corrected for random co-occurrences. Before going into the description of the plot, it is interesting to note that, the uncorrected POS plot (not depicted here for space reasons), revealed a co-occurrence pattern in which: a) adverbs are observed in proximity to their respective targets much more than all other parts-of-speech, b) nouns are also above the average line compared to all POS, and c) adjectives and verbs are below the average line. However, since the uncorrected pattern largely reflects the prior POS distributions in the corpus, the left panel in Figure 5a shows a reversed picture. Thus, taking random co-occurrences into account via our baseline correction, we observe that verbs and adjectives co-occur with the verb targets more often than the average 'all' line, and nouns and adverbs co-occur with the verb targets less often than average. In contrast to prior distributions where the pattern was driven by differences at early window sizes, the differences in POS distributions also emerge at later window sizes.

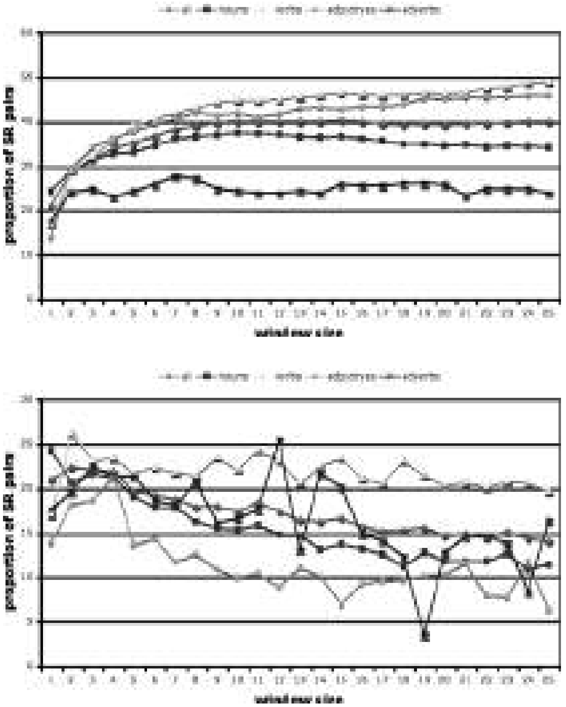


Figure 5a. Proportion of SR pairs, distinguishing parts-of-speech.

The right panel of Figure 5a shows in which individual window positions the POS responses occur. As one might have expected, verbs peak in ±2 words, likely accounting for the target and response verbs which co-occur e.g. in a conjunction (such as *einpacken und verstauen* ‘pack and store’), or in subcategorised verbs (such as *hadern zu streiten* ‘hesitate to argue’). Verb responses are also found in larger windows at fairly consistent rates, potentially accounting for verbs in close, but separate, clauses. Nouns have a peak at ±1 word, hinting towards some adjacency, and their co-occurrence rates drop consistently as distance from the target increases. These observations in window position ±1 account both for nouns that immediately precede the verb and bare nouns that follow. Adjectives co-occur strongest in window positions ±1-4 with a peak at ±4, and less often in larger window positions. These positions can be explained by adjectives within noun phrases preceding or following the target verbs; depending on the complexity of the noun phrases, the adjectives are rarely directly next to the verb but typically in some near position. Adverb responses are distributed across various positions, close and far, with respect to the targets. We believe this distribution is due to the promiscuity of adverbs, i.e., adverbs tend to be high frequency and can modify many verbs as well as whole clauses, and the flexibility of German grammar with regard to adverb placement.

In order to make stronger statements with respect to typical co-occurrence positions and potential functions between the target verbs and the parts-of-speech of the responses, Figure 5b brings experiments 4 and 5a together by distinguishing both the parts-of-speech of the responses and the window direction. As before, the co-occurrence threshold is 5x and corrected for random co-occurrences. As we are interested in specific window positions, the plot is exclusive. The ‘x-axis’ is iconically arranged, i.e., it is centred around the target verb, with preceding windows to the left and subsequent windows to the right. For simplicity, we restricted the figure to ±10 words only, i.e., starting with -10 to -1, and continuing with +1 to +10. The plot confirms our hypothesis that noun responses often occur directly before their respective target verbs, and seldom directly but nevertheless close after. Furthermore, the co-occurrence rates of noun responses decrease in both directions as the window position moves further from the target. We can quite safely assume that this picture is due to NPs directly preceding verbs (thus the head noun of the NP being directly adjacent to the verb, pointing to verb-final clauses in German) vs. NPs directly following verbs (typically though not necessarily with e.g. determiners and adjectives between the verb and the noun, pointing to verb-first and verb-second clauses in German). This distribution suggests that response nouns might often represent argument functions of the target verbs. However, given the free word order in German, we cannot draw any conclusions about the functions of the nouns. Schulte im Walde et al. (2008) analysed the argument potential of response nouns with respect to target verbs in more detail.

With the exception of verb responses, in fact all response types are more frequently observed immediately before their respective targets than after them. The distribution of verb responses has two peaks at -2 and +2 words, which also strengthens our assumption that response verbs might occur e.g. in conjunctions with or subcategorised by target verbs. In addition, the plot shows that the verb arrangement could be of either order. Finally, verbs maintain strong co-occurrence rates across the window positions and in both directions, suggesting that response verbs occur in preceding and following clauses.

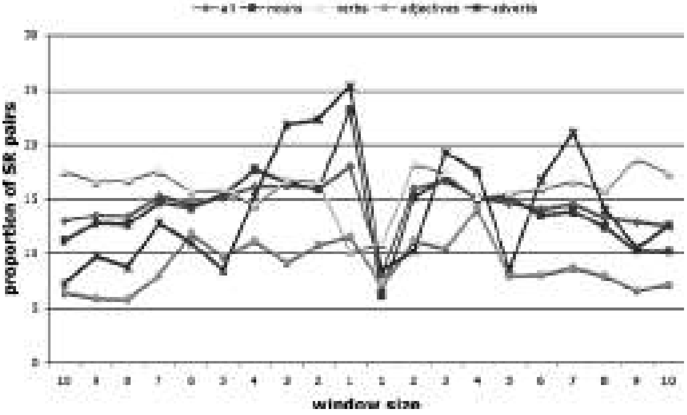


Figure 5b. Proportion of SR pairs, distinguishing parts-of-speech and left vs. right window.

While this analysis dove-tails with the more general picture of co-occurrence that is emerging, it also provides a magnifying glass to examine the details of the distribution. By including both the part-of-speech of the associates and the direction of the context window, we now know more about the variation within the distribution curve as a function of these additional factors.

4.3 The Chain Effect in Association Norms

As mentioned in Section 2, early procedures for eliciting associates allowed participants to supply multiple responses to each stimulus. However, more recent protocols have opted for a discrete elicitation procedure, in which only a single response is provided. The shift towards a discrete procedure was partly due to concerns about association chain effects, i.e., that the *n*th response is associated to the *(n-1)*th response rather than the stimulus, and that association chaining would contaminate the later responses (McEvoy & Nelson 1982). For example, given a target word ‘storm’, a first response could be ‘lightning’ and a second response could be ‘Zeus’, which is arguably more related to ‘lightning’ than it is to the target word ‘storm’. Additionally, investigations into the reliability of associations and the explanatory power of modelling behavioural data have shown that the first response is at least sufficient and possibly superior to subsequent responses (McEvoy & Nelson 1982; Nelson et al. 2000).

Accordingly, the experiments in this article have only considered the first responses to the stimulus verbs. However, the data set we are using includes multiple associations. Thus, we can investigate directly the issue of response chaining. Specifically, we assess the extent to which *n+1* responses are linked to the *n*th response rather than to the target, as indexed by corpus co-occurrence rates. In this section we use the co-occurrence patterns to gauge the degree of semantic relationship between target and response.

Experiment 6: *co-occurrence and association chains*

To address this question, we now include the first 5 responses to each stimulus in the analysis, with each response coded for its ordered position. Because not every participant provided 5+ responses to each target word, the number of responses in each set decreases over the ordered positions. Thus, in comparison to the 15,788 first response tokens, our data provided 15,454 second response tokens, 14,551 third response tokens, 12,504 forth response tokens, and 9,295 fifth response tokens. In Figure 6a, plotted lines represent each ranked response (i.e., rank 1 = first response, rank 2 = second response, etc), making use of the inclusive windows. The top line, ‘target-rank1’, corresponds to the same data plotted in Figure 1 as co-occurrence ≥5. Accordingly, the other lines are also plotted for a co-occurrence strength ≥5; the figure is uncorrected for random co-occurrences, as we are comparing the overall co-occurrence proportions. As expected given prior findings (McEvoy & Nelson 1982; Nelson et al. 2000), the first response exhibits stronger co-occurrence patterns with the target word than any of the later responses. Specifically, focusing on the largest window size (although the pattern is evident in earlier windows as well) the first responses co-occurred with their respective targets 8% more often than the rank 2 responses and 11.5% more often then the rank 3 responses. Rank 2 responses also co-occurred with their respective targets more often than later responses, with an advantage of 3.5% over rank 3 responses. The rank 3+ responses did not differ greatly from one another.

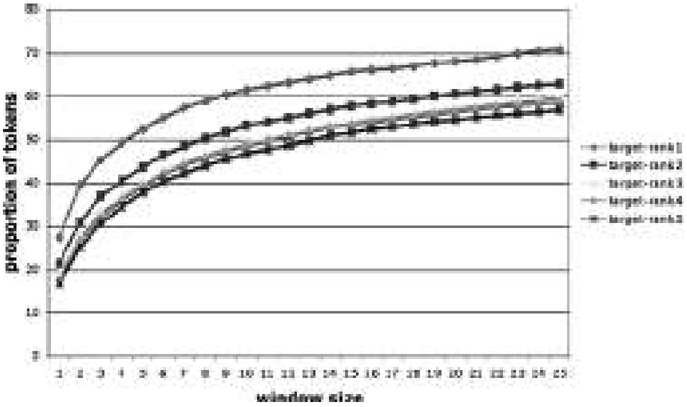


Figure 6a. Proportion of co-occuring SR pairs distinguished by the ranked order of the reponses.

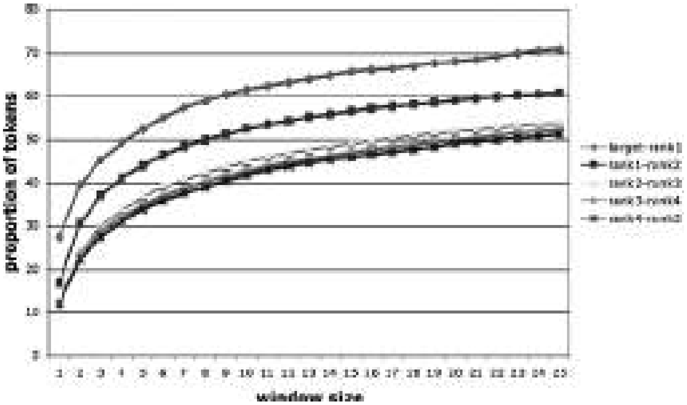


Figure 6b. Proportion of co-occuring pairs of *n* and *n+1* ranked responses.

We can compare Figure 6a with Figure 6b, which plots the proportion of rank *n+1* responses with their respective rank *n* responses. Our data provided 15,012 second response tokens with respect to the first response tokens (ranks 1-2), 14,082 ranks 2-3, 12,092 ranks 3-4, and 9,024 ranks 4-5. The top line in Figure 6b reflects the same data already presented in the top line of Figure 6a, namely the co-occurrence rates for rank 1 responses and their respective targets. The other lines in Figure 6b refer to the co-occurrence rates of rank *n+1* responses with their respective *n*th responses rather than with the target. Interestingly, we see a very similar pattern as in Figure 6a; namely, rank 2 responses co-occur with rank 1 responses less than rank 1 responses co-occur with the target and to a similar extent as they (the rank 2 responses) co-occurred with the targets themselves. Rank 2 responses co-occur with rank 1 responses just 1% less than they co-occur with their respective targets. Also similar to the pattern in Figure 6a, the 3+ responses and their respective preceding ranked responses do not differ greatly from one another. The later responses however are slightly more likely (~5%) to co-occur with their respective targets than with their respective *n-1* responses. These data thus suggest that concerns about response chaining are only partly justified. While it is true that the later responses are related, via co-occurrence, to their *n-1* responses, they are still as related, if not slightly more so, to the target. We would therefore argue that, depending on the goals of the project, multiple responses could provide a richer picture of the semantics of the target by indexing additional meaning components. The fact that they are also related to the preceding responses only highlights the extent to which semantic knowledge is a network of inter-related nodes (cf. Lund & Burgess 1996). If corpus co-occurrence is an index of association response relevance, then it appears that the rank *n+1* responses are as closely related to the targets as they are to their respective rank *n* responses.

4.4 The Influence of Association Strength on Co-Occurrence Distribution

An important construct in the association literature is association strength, namely how many times was a given response provided to a particular stimulus. Spence and Owens (1990) predicted differences in the distributional properties of strongly and weakly associated words. Specifically, they predicted that association strength should be negatively correlated with the distance between stimulus and response. In other words, they predicted that strongly associated word pairs would occur in close proximity in texts while weakly associated pairs would occur further apart, potentially even in different clauses. However, the SR pairs in their study were all comparatively strongly associated (their weakest pair was 107 out of 1000). Also, they only considered a fairly small co-occurrence window of 250 characters. Thus, while they did observe a significant negative correlation (*r2* =.185) it is not clear that this pattern is observed when a greater range in strength, word frequency and window size is examined or when other types of SR pairs than noun-noun pairs are considered. Furthermore, all of the analyses thus far have revealed the overwhelming importance of the proximal context window; we have yet to see evidence for a strong impact of larger windows. Thus, it seems unlikely that very weakly associated words will be revealed to occur further away from the target.

Experiment 7: *taking association strength into account*

As a first step towards analysing the role of association strength within our dataset, we divided the response types into association strength ranges, according to how many tokens were provided per type. In line with most experiments thus far, we considered only the first responses, calculated the co-occurrence strength, and plotted the proportions of SR pairs with a «co-occurrence strength» 5. Different to all previous experiments, we plotted the proportions of SR 'types' and not the proportions of SR 'tokens' in this experiment. This is because the number of tokens per type actually indicates the association strength, and if we took the tokens into account, we would necessarily cover a larger proportion for a larger token-per-type ratio. We defined 5 strength ranges: 1 (an SR pair is idiosyncratic), 2-5, 6-10, 11-20 and >20. Furthermore, we include the mean proportions of co-occurring pairs for our unrelated baseline words (taken from Figure 2a), as an extreme instance of weakly associated word pairs. The results of the proportions of SR types covered within each strength range (including the unrelated pairs) are presented in Figure 7.

The data displayed by the separate lines are 'exclusive'; data contributing to one line does not contribute to any other line. In the left hand panel of this figure we see that, consistent with predictions of S&O, strongly associated SR types co-occur more often than weakly associated types at early windows, and this difference persists inclusively to the end of the window range. This first observation supports the co-occurrence hypothesis, although some measure to identify the tightness of this relationship would aid interpretation. Generally, this pattern fits the correlation reported by S&O between co-occurrence frequency and association strength. To investigate the additional claim that weakly associated words should appear more often in later windows, we additionally calculated the proportions of responses found in each context window exclusively, which is presented in the right frame of the figure. Here we can see that there is no systematic pattern across the different association strengths. All but the unrelated line show a negative relationship between proportion of SR pairs observed and distance from the target, although the decrease is less dramatic as association strength weakens. The stronger associates also show more variance in their distribution, having multiple peaks across the window positions, in a non-systematic matter. Importantly with regard to S&O's prediction, none of the strength ranges co-occur more often further from the target. Thus, this analysis which considers a) many responses from many parts-of-speech, b) different frequency ranges for responses and targets, c) a wider range of association strengths, and d) a larger co-occurrence window, does not find support for S&O's prediction of a negative correlation between association strength and SR co-occurrence distance.

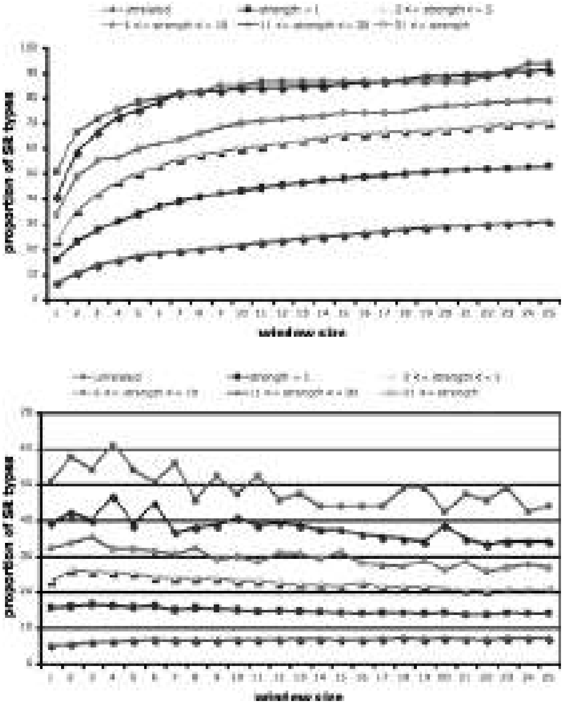


Figure 7. Proportion of SR types, distinguished by association strength.

5. Discussion

The current exploration was largely inspired by previous work into association norms (Schulte im Walde et al. 2008). We observed that, contrary to expectations (e.g., Clark 1971), no single type of relations was over-represented; no semantic relation, such as synonymy or antonymy, nor grammatical function, such as direct object of the stimulus verb, dominated the responses. Rather, we saw a wide variety of responses, including responses that escaped classification by available lexical resources e.g., GermaNet (Kunze 2000; Fellbaum 1998), or a statistical, context-free grammar model for German (Schulte im Walde 2002). Schulte im Walde et al. described many of these 'missing' relations as reflecting general aspects of world knowledge (e.g., *nieseln – nass* 'drizzle – wet', *mampfen – lecker* 'munch – yummy', *auftauen – Wasser* 'defrost – water', *überraschen – Freude* 'surprise – joy', *leiten – Verantwortung* 'guide – responsibility'). Suspecting that it was unlikely that all of such a broad range of responses described by Schulte im Walde et al. were equally likely to occur in the 'immediate' linguistic context of the stimulus – as argued by prior analyses – we decided to investigate more thoroughly the distributional characteristics of associate stimuli-response pair co-occurrence. Thus, the contributions of this article are three-fold. First, we brought together existing work on association norms and co-occurrence that has not been explicitly connected before. Second, we replicated several prior analyses on a common data set, namely, a collection of associate responses to German verbs. And third, we identified additional properties that might modulate the distributional characteristics of SR pairs and explored their influence on co-occurrence patterns. Bringing the various experiments together, this article therefore tried to provide a more complete picture of the co-occurrence distributions of semantic associates than has previously been compiled. While the majority of prior studies investigating semantic associations focussed on noun targets (e.g., S&O), we focus on verb targets. Thus, to the extent that our findings replicate or parallel prior findngs, it suggests robust characteristics of the link between semantic association and co-occurrence. Likewise, differences in observed patterns may be driven in part by features of the respective parts-of-speech. Nevertheless, given the size of the data set analyzed, the insights obtained from this exploration should largely generalise from the specific case of verb associations to other parts-of-speech and other languages.

Given that previous work on the distributional properties of semantic associates had been driven by diverse motivations in psychological and computational linguistic research, our 7 experiments were motivated from different angles. The results therefore contribute to partly disjunctive issues, summarised as follows.

1. Confirmation of the co-occurrence hypothesis

Our basic experiment 1 found high and pervasive co-occurrence patterns for our SR pairs, providing general support for the co-occurrence hypothesis. Its interpretation was further clarified by establishing a baseline to subtract out random co-occurrence. While we have provided more detailed information about the nature of the distributions that underlie the co-occurrence hypothesis, our analysis says nothing about the direction of influence between association and co-occurrence. Specifically, we make no claims about whether textual co-occurrence underpins semantic associates or whether semantic associations underpin textual co-occurrence or, as a third alternative, whether both are underpinned by some third factor, such as a more comprehensive semantic system (cf. the discussion in Glenberg & Mehta, this issue).

2. Caveats concerning co-occurrence distributions

Experiment 2 and discussed caveats A and B demonstrated the importance of controlling for prior frequencies and corpus size. These caveats are well-known, but have not yet been explicitly illustrated in the context of the co-occurrence hypothesis. Most importantly, having established a baseline to estimate random co-occurrence given word frequencies and parts-of-speech, experiment 2 demonstrated the extent to which our SR pairs co-occurred across a comparatively large context window. Here we saw that responses occurred above chance in all context positions, but the proportion of observed responses at all co-occurrence strengths decreased as the window position moved further from the target word.

3. Frequency effects and individual SR pairs

Taking the corpus frequencies of stimuli and responses into account by applying log-likelihood to our target-response pairs, experiment 3 suggested that statistical association measure might not only be useful to detect collocations, as shown in most previous work that used them, but also to identify associated word pairs at longer distances, which we consider an interesting contribution towards identifying situational knowledge.

4. Functional relationships between stimuli and responses

Experiments 4 and 5, taking window direction and response part-of-speech into account, illustrated that specific window positions provide insights into the co-occurrence functions that contribute to the co-occurrence hypothesis. Previous work had only taken positional information into account with respect to specific collocations. We showed more generally that noun, adjective and adverb responses in the association norms are prominently represented in co-occurrence positions immediately preceding the target verb. Unfortunately, the free word order in German does not allow us to draw any strong conclusions about the argument functions of the nouns. Using a chunked or parsed corpus instead of the raw words would tell us more about the functional distributions, as addressed by Schulte im Walde et al. Verbs - differently to the other parts-of-speech - were underrepresented in position 1 but peaked in position ±2, which we interpreted as indicating that many of our verb-verb SR pairs co-occurred e.g. in conjoined VPs, in line with the predictions by Clark (1971). Furthermore, successive peaks in larger window positions were interpreted as occurring in adjacent clauses, thus hinting towards situational agreement between the verbs. Linking this finding to the findings by Schulte im Walde et al., we speculate that the verb responses at larger windows might reflect frame- or scheme-based relations, such as *adressieren - schicken* ‘address - send’, *schwitzen - stinken* ‘sweat - stink’, *erfahren - wissen* ‘get to know - know’.

5. Association chain effect

Comparing the proportions of associations that are found in co-occurrence with the stimuli for associations up to rank 5 with those co-occurrence proportions between the ranks demonstrated that, if corpus co-occurrence is an index of association response relevance, then it appears that the rank *n+1* responses are at least as closely related to the targets as they are to their respective rank *n* responses. These results suggest that the concerns about response chaining are only partly justified. Furthermore, we would suggest that research with the goal of using association norms to describe word meaning, as opposed to modelling behavioural results, might prefer non-discrete elicitation procedures.

6. Association Strength

Taking association strength into account, experiment 7 failed to provide converging support for S&O’s prediction of a negative correlation between association strength and SR co-occurrence distance. Rather, we found that weakly associated responses exhibited similar co-occurrence distributions as strongly associated responses. While there is the chance that our measure of association strength was not sufficiently sensitive, given that only around 50 participants provided responses to each target, we saw no trends to suggest that more sensitivity would have produced a different result.

We have presented 7 experiments which look at the co-occurrence distribution of SR pairs from different angles. We tried to explore the issues that have previously been addressed in the literature, such as association strength and part-of-speech, but this investigation is just the tip of the iceberg. Several outstanding questions remain to be investigated such as distinguishing distributional patterns for different types of semantic relations captured by an SR pair, as characterised by available lexicographic resources (e.g., GermaNet) or for different semantic classes of the target verb. Additionally, space prohibited us from delving into the data in this paper to examine if there is a descriptive pattern in the types of responses observed at proximal vs. distal response windows. Here we could only touch on this very briefly with the POS and window analyses; following the present analyses with a linguistic analysis of the individual responses would add additional support for much of the speculative interpretation presented here. It would also be beneficial for future research to compare or combine the observations reported here using simple co-occurrence with other models that consider paradigmatic relations, second-order associations, etc.

Additionally, throughout the article we alluded to SR pairs that captured what Schulte im Walde et al. referred to as world knowledge, with many of the associations expressing neither a common functional role of the verb (e.g., being a common filler of an argument role) nor a traditional semantic relation like synonymy or hypernymy. Closer examination revealed that many associates seemed to express a variety of meaning characteristics which Schulte im Walde et al. hypothesised would be unlikely to be found regularly in close context windows. Support for this position was not really provided in the current work, since the close window was identified as primary for a majority of responses. Failure to observe certain types of 'non-traditional' relations may in fact support the premise put forth by Glenberg & Mehta (this issue), namely that semantics gives rise to distributional patterns, not the reverse. However, the patterns presented in this article cannot directly address this debate. Future work could be done by classifying the associates on the basis of whether they express 'world knowledge' as intended by Schulte im Walde et al. and comparing their distribution to other types of responses.

Finally, two aspects of this study may have compromised some of our interpretive power. One was our choice of investigating German verbs rather than, for example, English verbs. German has a comparably flexible word order, with verbs found at the beginning (first/second position) and at the end of clauses. Thus, the degrees of freedom for inferring what type of response would likely be occurring immediately before or after a target were much greater for German than they would be for a language with a more rigid word order, such as English. The second was our choice to conduct a large-scale empirical investigation and draw largely descriptive, qualitative, conclusions. On the one hand, this choice allowed us to use thousands of data points and describe the co-occurrence distribution for the entire population of our SR pairs. Since we have not sampled our data, we can rely on the raw numbers and observations. On the other hand, we have no inferential statistics to inform us about how general an observation was across the data set or how representative the patterns in our dataset are of SR pairs in the language more generally. We believe future work that incorporated inferential statistics (via monte carlo sampling) to gain more fine-grained insights into the co-occurrence distribution would be a valuable extension of the present work. We hope that the observations made here will serve as a guidepost to that research, suggesting which comparisons might be worthwhile to test.

In conclusion, our experiments addressed various aspects of the co-occurrence hypothesis, contributing to research questions concerning semantic relatedness in psycholinguistic and computational linguistic research lines. We certainly could not cover all co-occurrence-related issues but we have outlined various avenues for future work to address, for instance, expansion of the current investigation with an application of more powerful statistical methods, an investigation of the properties of co-occurrence distributions with respect to the semantic classes of the stimuli or with respect to the semantic relations between the stimuli and the responses (both suggested by Schulte im Walde et al. 2008; Guida 2007).

Addresses of the Authors
Sabine Schulte im Walde, Institute for Natural Language Processing,
University of Stuttgart, Germany
<schulte@ims.uni-stuttgart.de>

Alissa Melinger, School of Psychology, University of Dundee, Scotland
<a.melinger@dundee.ac.uk>

Notes

[1] While early procedures for eliciting associates allowed participants to supply multiple responses to each stimulus, more recent protocols have opted for a discrete elicitation procedure, in which only a single response is provided. We address the question of one vs. many responses and the related concern about

association chaining (that the *n*th response is associated to the *(n-1)*th response rather than the stimulus) in Section 4.3.

[2] Note that the lower the co-occurrence threshold the easier it is to satisfy the threshold with unrelated words. In fact, the unrelated word pairs co-occurred more in larger windows than in smaller windows while the reverse was true of the related words. This is why the lines in Figure 2a (left panel) flatten out or even decrease as window size increases.

[3] In all subsequent analyses, we use the co-occurrence threshold of ≥5 whenever we compare SR proportions across conditions.

Bibliographical references

BANKO Michelle & Eric BRILL 2001. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics,* Toulouse.

BATTIG William F. & William E. MONTAGUE, 1969. Category norms for verbal items in 56 categories, a replication and extension of the Connecticut category norms. *Journal of Experimental Psychology* 80. 1-46.

BURGESS Curt 1998. From simple associations to the building blocks of language: Modeling meaning in memory with the Hal model. *Behavior Research Methods, Instruments & Computers* 30. 188-198.

CHARLES Walter & George MILLER 1989. Contexts of antonymous adjectives. *Applied Psycholinguistics* 10. 357-75.

CHURCH Kenneth W. & Patrick HANKS 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1). 22-29.

CLARK Herbert H. 1971. Word associations and linguistic theory. In John LYONS (ed.). *New Horizons in Linguistics*. Harmondsworth: Penguin Books Ltd. 271-286.

COHEN B. H., W. A. BOUSFIELD & G. A. WHITMARSH 1957. Cultural norms for verbal items in 43 categories. Technical Report No. 22. University of Connecticut, Contract Nonr. 631(00). Office of Naval Research.

DEESE James 1965. *The Structure of Associations in Language and Thought*. Baltimore: The John Hopkins Press,.

DUNNING Ted 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1). 61-74.

EVERT Stefan 2005. *The Statistics of Word Co-Occurrences: Word Pairs and Collocations*. Institut für Maschinelle Sprachverarbeitung: Universität Stuttgart. PhD dissertation.

FELLBAUM Christiane 1998 (ed.). *WordNet - An Electronic Lexical Database*. Language, Speech, and Communication. Cambridge: MIT Pressi.

FERNÁNDEZ Ana, Emiliano DIEZ, María ANGELES ALONSO & María Soledad BEATO 2004. Free-association norms for the Spanish names of the Snodgrass & Vanderwart pictures. *Behavior Research Methods, Instruments & Computers* 36(3). 577-583.

FERRAND Ludovic & F.-Xavier ALARIO 1998. French word association norms for 366 names of objects. *L'Annee Psychologique* 98(4). 659-709.

GALTON Francis 1880. Psychometric experiments. *Brain* 2. 149-162.

GEFFET Maayan & Ido DAGAN 2005. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, MI.

GIRJU Roxana 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.

GIRJU Roxana, Adriana BADULESCU & Dan MOLDOVAN 2006. Automatic discovery of part-whole relations. *Computational Linguistics* 32(1). 83-135.

GLENBERG Arthur, M. & Sarita MEHTA Constraint on Covariation: It's not Meaning. *This Issue*.

GRIFFITHS Thomas L. & Mark STEYVERS 2003. Prediction and semantic association. *Advances in Neural Information Processing Systems* 15.

GUIDA Annamaria 2007. *The Representation of Verb Meaning within Lexical Semantic Memory: Evidence from Word Associations*. Università degli studi di Pisa. Master's thesis.

GUIDA Annamaria & Alessandro LENCI 2007. Semantic properties of word associations to Italian verbs. *Italia Journal of Linguistics* 19(2).

HIRSH, Katherine W. & Jeremy TREE 2001. Word association norms for two cohorts of British adults. *Journal of Neurolinguistics* 14(1). 1-44.

JI Heng, David WESTBROOK & Ralph GRISHMAN 2005. Using semantic relations to refine coreference decisions. In *Proceedings of the joint Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Vancouver 17-24.

JUSTESON John S. & Slava M. KATZ 1991. Co-occurrence of antonymous adjectives and their contexts. *Computational Linguistics* 17, 1-19.

KAVALEK Martin & Vojtech SVATEK 2005. A study on automated relation labelling in ontology learning. In BUITELAAR et al. (eds.), *Ontology learning from text: methods, evaluations, and applications.* IOS Press.

KENT Grace H. & Aaron J. ROSANOFF 1910. A study of association in insanity. *American Journal of Insanity* 67 47. 37-96.

KISS George R., Christine ARMSTRONG, Robert MILROY & James PIPER 1973. An associative thesaurus of English and its computer analysis. In *The Computer & Literary Studies*. Edinburgh University Press. URL http://www.eat.rl.ac.uk/.

KUNZE Claudia 2000. Extension and use of GermaNet, a lexical-semantic database. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, Athens.

LAUTESLAGER Max, Theo SCHAAP & Dick SCHIEVELS 1986. *Schriftelijke woordassociatienormen voor 549 Nederlandse zelfstandige naamworden*. Swets & Zeitlinger.

LEMAIRE Benoit & Guy DENHIÈRE 2006. Effects of high-order co-occurrences on word semantic similarity. *Current Psychology Letters - Behaviour, Brain & Cognition* 18(1).

LIN Dekang 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational Linguistics*, Montreal.

LIN Dekang 1999. Automatic identification of non-compositional phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, Maryland. 317-324.

LOWE Will & Scott, MCDONALD 2000. The direct route: mediated priming in semantic space. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*.

LUND Kevin & CURT Burgess 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments & Computers* 28(2). 203-208.

MAEDCHE Alexander & Steffen STAAB 2000. Discovering conceptual relations from text. In *Proceedings of the 14th European Conference on Artificial Intelligence*. Berlin.

MCENERY Tony 2003. *Corpus Linguistics.* In MITKOV Ruslan (ed.), *The Oxford Handbook of Computational Linguistics.* 448-463.

MCEVOY Cathy L. & Douglas. L. NELSON, 1982. Category name and instance norms for 106 categories of various sizes. *American Journal of Psychology* 95. 581-634.

MCNAMARA Timothy P. 2005. *Semantic priming: Perspectives from memory and word recognition*. New York: Psychology Press.

MILLER George 1969. The organization of lexical memory: Are word associations sufficient? In G.A. TALLAND & N.C. WAUGH (eds.), *The Pathology of Memory*. New York: Academic Press. 223-237.

MOLDOVAN Dan & Adrian NOVISCHI 2002. Lexical chains for question answering. In *Proceedings of the 19th International Conference on Computational Linguistics*. Taipei.

MOSS Helen E., Mary L. HARE, P. DAY & Lorraine K. TYLER 1994. A distributed memory model of the associative boost in semantic priming. *Connection Science* 6(4). 413-427.

NAVIGLI Roberto & Paola VELARDI 2004. Learning domain ontologies from document warehouses and dedicated web sites. *Computational Linguistics* 30(2). 151-179.

NELSON Douglas L., Cathy L. MCEVOY & Simon DENNIS 2000. What is free association and what does it measure? *Memory & Cognition* 28. 887- 899.

NELSON Douglas, L. & Nan ZHANG 2000. The ties that bind what is known to the recall of what is new. *Psychonomic Bulletin & Review* 7(4). 604-617.

NELSON Douglas L., David J. BENNETT & Todd LEIBERT 1997. One step is not enough: Making better use of association norms to predict cued recall. *Memory & Cognition* 25. 785-796.

NELSON Douglas L., Cathy L. MCEVOY & T.A. SCHREIBER 1998. The University of South Florida word association, rhyme, and word fragment norms. URL http://www.usf.edu/FreeAssociation/.

PALERMO David S. & James J. JENKINS 1964. *Word Association Norms: Grade School through College*. Minneapolis: University of Minnesota Press.

PERESSOTTI Francesca, Francesca PESCIARELLI & Remo JOB 2002. Le associazioni verbali PD-DPSS: norme per 294 parole. *Giornale Italiano di psicologia* 29. 153-170.

PLAUT David C. 1995. Semantic and associative priming in a distributed attractor network. In J.D. MOORE & J.F. LEHMAN (eds.). *Proceedings of* *the XVIIth Annual Conference of the Cognitive Science Society* (August, 2003), Pittsburg Vol 17. 37- 42.

RAPP Reinhard 1996. *Die Berechnung von Assoziationen*, volume 16 of *Sprache und Computer*. Hildesheim-Zürich: Georg Olms Verlag.

RAPP Reinhard 2002. The computation of word associations: Comparing syntagmatic and paradigmatic approaches. In *Proceedings of the 19th International Conference on Computational Linguistics*, Taipei.

RUSSELL Wallace A. 1970. The complete German language norms for responses to 100 words from the Kent-Rosanoff word association test. In Leo POSTMAN Leo & Geoffrey KEPPEL (eds.), *Norms of Word Association*. New York: Academic Press. 53-94.

RUSSELL Wallace A. & O.R. 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.

SCHULTE IM WALDE Sabine 2002. A subcategorisation lexicon for German verbs induced from a lexicalised PCFG. In *Proceedings of the 3rd Conference on Language Resources and Evaluation.* Las Palmas de Gran Canaria. 1351-1357.

SCHULTE IM WALDE Sabine 2003. *Experiments on the Automatic Induction of German Semantic Verb Classes*. Institut für Maschinelle Sprachverarbeitung: Universität Stuttgart. PhD dissertation.

SCHULTE IM WALDE Sabine 2006. 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. 69-76.

SCHULTE IM WALDE Sabine 2008. Human associations and the choice of features for semantic verb classification. *Research on Language and Computation* 6(1). 79-111.

SCHULTE IM WALDE Sabine, Alissa MELINGER, Michael ROTH & Andrea WEBER, 2008. An empirical characterisation of response types in German association norms. *Research on Language and Computation* 4(2). 205-238.

SEIDENSTICKER Petra 2006. *Simulation von Wortassoziationen mit Hilfe von mathematischen Lernmodellen in der Psychologie*. Fakultät der Kulturwissenschaften: Universität Paderborn. PhD dissertation.

SPENCE Donald P. & Kimberly C. OWENS 1990. Lexical co-occurrence and association strength. *Journal of Psycholinguistic Research* 19. 317-330.

STEYVERS Mark, Richard M. SHIFFRIN & Douglas L. NELSON 2004. Word association spaces for predicting semantic similarity effects in episodic memory. In A. HEALY (eds.). *Experimental Cognitive Psychology and its Applications. Festschrift in honor of Lyle Bourne, Walter Kintsch & Thomas Landauer.* Washington, DC: American Psychological Association.

TATU Marta & Dan MOLDOVAN 2005. A semantic approach to recognizing textual entailment. In *Proceedings of the joint Conference on Human Language Technology & Empirial Methods in Natural Language Processing*. Vancouver. 371-378.

VIEIRA Renata & Massimo POESIO 2000. An empirically-based system for processing definite descriptions. *Computational Linguistics* 26(4). 539-593.

WETTLER Manfred & Reinhard 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. 84-93.