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

Sabine Schulte im Walde[1] and Alissa Melinger[2]

[1]Institute for Natural Language Processing, University of Stuttgart, Germany
[2]School of Psychology, University of Dundee, Scotland

schulte@ims.uni-stuttgart.de, a.melinger@dundee.ac.uk

Semantic associations, namely words that are called to mind in response to a given stimulus, have been of interest to cognitive scientists for over a century. Over the years, they have come to represent a window into semantic knowledge, 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 investigative the mechanisms underlying semantic memory, giving the researcher insights into the way semantic information is accessed and represented with the behavioural system.

One way to evaluate semantic associations are co-occurrence-based accounts, which address directly the issue of the relationship between the elicited stimulus-response pairs and the context in which they occur in language: The hypothesis that semantic association and textual co-occurrence index the same lexical relationships was developed by Miller (1969) and first tested empirically by Spence and Owens (1990). They tested their hypothesis by searching corpora for the co-occurrence of strongly related semantic associates. Comparing the co-occurrence frequencies of associates to frequency-matched unrelated word pairs, they found significantly higher rates of co-occurrence for the related words than unrelated words. Furthermore, the notion of co-occurrence distributions has 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 and Staab, 2000; Navigli and Velardi, 2004), anaphora resolution (Vieira and Poesio, 2000; Ji *et al.,* 2005), and textual entailment (Geffet and Dagan, 2005; Tatu and Moldovan, 2005). Many researchers within that area have identified the value of human data to their task, in order to evaluate computational models; among them is work that used free association norms as a test-bed for distributional models of semantic relatedness (Church and Hanks, 1990; Rapp, 1996; Rapp, 2002; Lemaire and Denhière, 2006; Schulte im Walde, 2006). The approach we take in this talk is to conduct a descriptive and in depth examination of the distributional properties of stimulus-associate pairs within a co-occurrence context window. Much research has already

addressed this question to varying degrees. We review these contributions but then extend them with our own analyses.

The basis for the investigation is a collection of semantic associates evoked by German verbs. Taking as a starting point the co-occurrence analyses by Spence and Owens (S&O), we replicate those first experiments that founded the co-occurrence assumption for association norms. On the one hand, we break down the analyses into various categories which have – independently of S&O's co-occurrence assumption – previously been identified as distributionally interesting (Deese, 1965; Clark, 1971). On the other hand, we add analyses that question some of the intuitive conclusions from early work on the co-occurrence assumption. Thus, the contributions of our work are three-fold. First, we bring together existing work on association norms and co-occurrence that has previously not necessarily built on each other. Second, we replicate the analyses on a common data set, our collection of associations to German verbs. And third, we identify additional properties of association norms that have not yet been investigated, and add the respective analyses.

More specifically, we address the following questions:

- Does the co-occurrence hypothesis transfer to our association norms?
- To what extent does *corpus size* influence the co-occurrence hypothesis?
- What is the influence of the *window direction*, i.e., distinguishing between a left and a right context?
- Are associates of a certain *part-of-speech* more likely to occur in the corpus, and does their proximity to the stimuli differ?
- Combining insights on the window direction and the part-of-speech analysis, does German free word order allow inferences about *typical argument functions* among semantic associates?
- Do *semantic and empirical properties of the stimuli* (e.g., semantic class, or corpus frequency) influence the distribution of the semantic associates?
- Does *association chaining,* i.e. the tendency for response *n+1* to be related to response *n* rather than to the target word, contaminate later responses?

Bringing the various experiments together, this talk tries to provide a more complete picture of the co-occurrence distributions of semantic associates than has previously been compiled. Furthermore, it contributes both to psycholinguistic discussions – by demonstrating that some long-standing concerns about semantic associates are only partly justified – as well as to computational linguistics research on word associations, such as the automatic induction of multi-word expressions.

References

Kenneth W. Church and Patrick Hanks (1990). Word association norms, mutual information, and lexicography. Computational Linguistics, 16(1).

Herbert H. Clark (1971). Word associations and linguistic theory. Penguin.

James Deese (1965). The structure of associations in language and thought. The John Hopkins Press.

Maavan Geffet and 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.

Heng Ji, David Westbrook, and 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.

Benoit Lemaire and Guy Denhière (2006). Effects of high-order co-occurrences on word semantic similarity. Current Psychology Letters – Behaviour, Brain and Cognition, 18(1).

Alexander Maedche and Steffen Staab (2000). Discovering conceptual relations from text. In: Proceedings of the 14th European Conference on Artificial Intelligence.

George Miller (1969). The organization of lexical memory: Are word associations sufficient? In G.A. Talland and N.C. Waugh, eds., The pathology of memory.

Roberto Navigli and Paola Velardi (2004). Learning domain ontologies from document warehouses and dedicated web sites. Computational Linguistics, 30(2).

Reinhard Rapp (1996). Die Berechnung von Assoziationen. Georg Olms Verlag.

Reinhard Rapp (2002). The computation of word associations: comparing syntagmatic and paradigmatic approaches. In: Proceedings of the 19th International Conference on Computational Linguistics. Taipei, Taiwan.

Sabine Schulte im Walde (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, NY.

Donald P. Spence and Kimberly C. Owens (1990). Lexical co-occurrence and association strength. Journal of Psycholinguistic Research, 19.

Marta Tatu and Dan Moldovan (2005). A semantic approach to recognizing textual entailment. In: Proceedings of the joint Conference on Human Language Technology and Empirical Methods in Natural Language Processing. Vancouver, Canada.

Renata Vieira and Massimo Poesio (2000). An empirically-based system for processing definite descriptions. Computational Linguistics, 26(4).