Sources of Semantic Similarity

Simon De Deyne (simon.dedeyne@psy.kuleuven.be)[a]

Yves Peirsman (yves.peirsman@arts.kuleuven.be)[b]

Gert Storms (gert.storms@psy.kuleuven.be)[a]

University of Leuven, Belgium

[a] Department of Psychology, Tiensestraat 102, B-3000 Leuven, Belgium

[b] QLVL, Department of Linguistics, Blijde-Inkomststraat PO Box 3308, B-3000 Leuven, Belgium

Abstract

Similarity is the key notion underlying many contemporary theories about the representation of meaning through words or concepts. However, these representations are strongly colored by the kind of information captured by various semantic measures. In this paper we present a systematic comparison of human similarity judgments and calculated similarity coefficients from different sources of semantic similarity based on concept features, word associations, word co-occurrence and expert knowledge. We show that these measures capture our semantic representations to a large extent, but also model different aspects of our semantic knowledge, depending on (a) the semantic domain and (b) the range of similarity comparisons.

Keywords: Concepts; similarity; distributional semantics, word associations, WordNet.

Introduction

Semantic knowledge is a multifaceted concept. About bananas, we know that they are a type of fruit just like pineapples, that they are typically eaten by monkeys, and that people can slip on their skin. These various types of information, and much more, are part of our knowledge about the concept *banana*. Yet, studies of semantic knowledge often make use of only one source of information, like the free associations to a word, or its distribution in a large collection of texts (Landauer & Dumais, 1997; Burgess, Livesay, & Lund, 1998; Steyvers, Shiffrin, & Nelson, 2004). At the same time, it is unlikely that these different sources of information all correlate with the same types of semantic knowledge. In this paper, we will therefore focus on just one type of semantic knowledge instantiated by human judgments of semantic similarity and investigate what sources of information are best able to model these judgments.

We explore four sources of semantic information. The first is that of concept features, like <*lays eggs*> and <*can fly*> for the concept *robin* (McRae, Cree, Seidenberg, & McNorgan, 2005). The second is that of free associations, like the words *foot* and *shoelace* for the cue word *shoe* (Nelson, McEvoy, & Dennis, 2000). The third source is that of lexical co-occurrence in text, where we determine what words typically appear in the context of a target word (Sahlgren, 2006). The fourth and final source of semantic information is the Dutch EuroWordNet, a lexical ontology which contains taxonomical expert information about Dutch nouns (Vossen, 1998). It tells us, for instance, that *eagle* IS-A <*bird of prey*>, which in turn IS-A <*bird*>.

It has been suggested before (Maki & Buchanan, 2008) that these different sources of information tap into different types of human semantic knowledge. Still, it remains unclear what precise knowledge they capture. Our first goal is therefore to focus on semantic similarity in particular and measure to what extent our several sources of information are able to mirror people's similarity judgments. A second aim of this paper is to provide more insight into the performance of these techniques on *intra-category* rather than domain or *inter-category* similarity judgments. While earlier studies have mostly focused on inter-category judgments (e.g., Miller & Charles, 1991), far less is known about our ability to model intra-category judgments. These last should not only be harder to model; they can also give us more clues as to where the problems of the particular approaches may emerge — for instance in relation to specific categories of natural kinds or artifacts. Our third contribution lies in the fact that we make use of a controlled collection of experimental data and corpus material. Since many earlier individual studies that focused on one particular model made use of different materials, their results were difficult to compare. Our approach ensures that the results of the several models are perfectly comparable throughout.

Semantic Models

Types of Information

Features The first type of information we will study is that of concept features. Features cover the perceptual, physical, functional, etc. characteristics of a particular entity. They have proved their usefulness in the prediction of a large variety of semantic phenomena such as response times in a speeded categorization task (Storms, De Boeck, & Ruts, 2000), typicality (Hampton, 1997), similarity judgments (Markman & Gentner, 1993) and concept coherence (Sloman, Love, & Ahn, 1998). We used the feature sets presented in De Deyne et al. (2008). These sets contained features for a total of 420 concepts. Each concept belonged to one of 15 different categories. The categories organized by domain are shown in Table 2. The number of exemplars in each category varied from 20 to 33. A first type of feature set consisted of 15 exemplar $\times$ feature matrices, one for each category. These matrices were constructed using a two-phased approach. First, a large group of participants generated features for each exemplar. Next, these features were tallied and

a different group of participants judged the applicability of each feature to each exemplar within its category. The rows of the matrices corresponded to the category exemplars (varying in number from 20 to 33), and the columns of the matrices corresponded to the selected exemplar features for these categories (varying in number from 156 for *Fish* to 382 for *Sports*). The values in each cell indicate how many persons (out of four) judged the feature (e.g., <*lays eggs*>) to be applicable to the exemplar (e.g., *chicken*). In addition, two huge domain exemplar $\times$ feature matrices were constructed: one for the *Animal* domain and one for the *Artifact* domain. For both domains, all exemplars and all features of their member categories were aggregated. The *Animal* matrix contained 129 rows corresponding to animal names that belonged to the categories *Birds*, *Fish*, *Insects*, *Mammals*, and *Reptiles*. The columns corresponded to 765 exemplar features of *Animals*. The *Artifact* matrix contained 166 rows corresponding to object names that belong to the categories *Clothing*, *Kitchen Utensils*, *Musical Instruments*, *Tools*, *Vehicles*, and *Weapons*. The columns of this matrix represented 1,295 exemplar features of *Artifacts*.

Associations Association measures have been successful in the prediction of many semantic memory tasks such the distance effects in free recall and cued recall (Steyvers et al., 2004). We used a set of word associations collected between 2003 and 2006 (De Deyne & Storms, 2008). The experiment asked participants to give three different associations for each cue. In this way, 381,909 responses were collected for a total of 1,424 cues. This amounts to at least 360 association responses to a particular cue. In contrast to other word association databases (e.g., Nelson, McEvoy, & Schreiber, 2004), we collected three associations from each participant in a continuous task, instead of just one. This has two advantages. First, we can collect weak(er) associations, which is especially important for cues with very dominant associations (e.g., *blood* and <*red*>). Second, the resulting representations are denser and therefore more suited for a distributional approach of meaning.

Lexical Context A third source of knowledge that contains information about the meaning of a word is the contexts in which these words are used. This context can be defined as the documents in which a word occurs (LSA) (Landauer & Dumais, 1997), the words in a context window around the target word (Lund & Burgess, 1996), or the syntactic relations in which a word takes part (Pereira, Tishby, & Lee, 1993). Such contextual information has been used in the modelling of semantic priming (Burgess et al., 1998; Landauer & Dumais, 1997) and semantic dyslexia (Buchanan, Burgess, & Lund, 1996), etc. Here we will use a so-called *word-based* approach where the context features are the four words to the left and right of the target in each of its contexts in a corpus[1]. In order to obtain enough contextual information about the target words, we crawled a large corpus from the web. For each of our target words, we collected at least 1,000 documents. The resulting corpus contained 768 million word tokens and about 6 million types.

Semantic relations Our final knowledge source is an ontology. English WordNet (Fellbaum, 1998) and its sibling EuroWordNets (Vossen, 1998) are lexical databases that bring together groups of synonyms (so-called synsets) in large networks, which show the semantic relationships between individual words. Possible relationships are for instance hypernymy (*motor vehicle* – <*car*>) and hyponymy (*taxi* – <*car*>). Semantic distances obtained on the basis of WordNet have been shown to explain human similarity judgments independently of associative strength, lexical co-occurrence or featural similarity (Maki, McKinley, & Thompson, 2004).

Vector models

The first three types of semantic knowledge were subsequently transformed to a word-by-word matrix. The rows of this matrix represent the target words; the columns correspond to the words that describe the targets. For the feature sets, these are the judged features for the *Animal* or *Artifacts* matrices (used in Experiment 1) or the smaller category matrices (used in Experiment 2). For the set of associations, the columns are all associations given by the participants. For the lexical contexts, finally, they are the context words within a window of four words to the left and right of the target word. Because of its extremely large size, we reduced this last context matrix by removing rows with less than 10 elements, and columns with less than 2. Next, the frequencies of the associations, feature and context words were weighted. A first weighting function was performed by dividing the number of times the column word co-occurs with the target by its total frequency in the matrix (Inverse Vector Frequency or *IVF*). We also considered a second weighting function that is often applied in linguistic studies for measuring the association between the co-occurrence of two words. This approach considers the fact that two words can co-occur by chance. We followed a proposal by Church, Gale, Hanks, and Hindle (1991) that uses t-scores to measure if two words are collocates (i.e. it captures the extent to which the occurrence of one word depends on the other).[2]. In contrast to the vector models, the EuroWordNet ontology contains representations in a graph and can be derived more directly. More details follow in the next section.

Hypotheses

Given our knowledge about these four sources of information, we can make some predictions as to which ones should best be able to reflect human similarity judgments. We expect the most valuable information about semantic similarity

[1] Pilot studies showed that the relatively small context windows that we use (four words to either side of the target) result in better models of semantic similarity than approaches based on much larger windows, or on the distribution of words in paragraphs or documents rather than their mutual co-occurrence.

[2] We also tried other weighting schemes, but found these functions to perform consistently better.

to come from the features as well as the expert ontology. Features should give us most insight about the similarities and differences between concepts that belong to the same category (due to the procedure), while the taxonomical structure of EuroWordNet and the subclasses of a category should again correlate very well with similarities between more abstract subsets of concepts (due to the type of relationships commonly encoded in EuroWordNet).

Word associations and lexical co-occurrence information collected from corpora will probably contain information about the more general type of semantic relatedness than semantic similarity due to the large variety in semantic relationships in these sources and the lack of syntax and category-context compared to the concept features.

Next, we evaluated these models in two experiments. The first experiment considers semantic similarity between concepts spanning broad domains (*Artifacts* or *Animals*). This experiment corresponds closely to previous studies where the stimuli cover a wide range of semantic relationships. For example, LSA (Landauer & Dumais, 1997) correlated r = .72 with similarity judgments collected by Rubenstein and Goodenough (1965) and r = .64 with judgments reported by Miller and Charles (1991). These findings indicate we can expect LSA-like models such as our context model to perform reasonably well in this task

In the second study, we investigate how well these models approach detailed semantic representations that measure judgments of similarity within a category such as *Fruit*, *Insects*, *Musical Instruments* or *Professions*. While the similarity structure in these categories has received considerable attention in the categorization literature, hardly anything is known about the ability of distributional approaches based on context words or associations to capture these fine-grained representations.

Experiment 1: Domain Similarity

Method

Participants In total 30 persons, 26 females and 4 males (average age 20 years) participated in the experiment. They were mainly students at the University of Leuven, and were paid the equivalent of $10/h.

Materials and Procedure The stimuli consisted of members belonging to 6 different *Artifact* categories (*Clothing*, *Kitchen Utensils*, *Musical Instruments*, *Tools*, *Vehicles* and *Weapons*) and 5 *Animal* categories (*Birds*, *Fish*, *Insects*, *Mammals*, and *Reptiles*) described in De Deyne et al. (2008). Since it is not feasible to present all pairwise combinations of all exemplars of these categories we selected 5 exemplars from each of the *Artifacts* and *Animals* categories that cover a wide range of typicality. This way some members were central to the category representation (e.g., *sparrow* is a typical bird and thus a central member) while others were not (e.g. *bat* is an atypical member in the periphery of *Mammals*, and closely related to *Birds*). To increase the generalizability of our results, two replications of the above procedure were performed, resulting in a set A and B each consisting of 6 $\times$ 5 *Artifacts* and 5 $\times$ 5 *Animals*. Each participant rated *Animal* pairs or *Artifact* pairs of either set A or B in multiple sessions, with the only restriction that no replication sets were allowed to be rated the same day. All pairwise combinations (435 for the *Artifacts*, 300 for *Animals*) were presented in a random order on a computer screen. The word order in the pairs was randomized. Participants were asked to enter a number between '1' (for totally dissimilar) and '20' (for totally similar). In case one or two words of an exemplar pair were unknown, they had to enter '-1'. They completed the task for a single domain in less than 1 hour.

Experimental Results All exemplar pairs in set A and B were judged by at least 12 and by at most 18 different participants and were known by the majority of the participants. We removed the data from one participant in set B of the *Animals* as these data correlated less than .45 with the average ratings. The resulting ratings were all very reliable with Spearman Brown split-half correlations ranging from .94 to .97.

Model Results Similarity coefficients were obtained for the Feature, Association and Context model by calculating the cosine between the vectors of each concept pair. Traditionally, graph-theoretic measures are used to derive similarity from the EuroWordNet model. The measures we used for this model was the *Inverse Path Length* measure, which simply takes the inverse of the number of steps between two words in the taxonomy (see Budanitsky & Hirst, 2006, for an extensive discussion). Three items (*swan* and *seagull*, and *blackbird*) had to be removed from the *Animals* because they did not occur literally in the EuroWordNet ontology.

First we tested vector spaced models where no weighting procedure was applied to the term frequencies in the word $\times$ feature, word $\times$ association or word $\times$ context matrices. Averaged over both replications, the best results were found for the Feature model ($r = .82$) followed by the Context model ($r = .65$), the Association model ($r = .60$) and EuroWordNet ($r = .51$). However, we found that the unweighted term frequencies might present too harsh a test on these models since the models differ in size and the way their features (associations or context words) are distributed. In order to take the relative importance of frequencies into account, we applied the weighting schemes discussed above. Here we report the t-score weighted results for the Association and Context model and the IVF results for the Feature model[3].

Table 1 shows the correlations of the weighted similarity coefficients from the models with the human similarity ratings for set A and B of the *Artifact* and *Animal* domain. The results are comparable for both replications. The Feature model gives a near perfect account irrespective of the domain. The same holds for the Association model. While the Con-

[3]The choice of a weighting scheme did not change the ranking of the various models in both Experiment 1 and 2, except for word associations, where the unweighted term frequencies performed systematically worse than any other scheme.

text model gives a good account for the *Artifacts*, it is less accurate in predicting the judgments in the *Animal* domain. Finally, the EuroWordNet model (WN-D) performs inconsistently for *Artifacts* and scores poorly in the *Animal* domain.

Table 1: Correlations of the human similarity judgments of Experiment 1 and the different model similarity coefficients.

Domain	Set	Feat.	Asso.	Context	WN-D
Animals	A	.91**	.87**	.53**	.34**
	B	.87**	.82**	.55**	.16*
Artifacts	A	.92**	.76**	.81**	.63**
	B	.85**	.75**	.80**	.39**

Note: * $p < .05$, ** $p < .01$ (two-tailed)

Experiment 2: Intra-Category Similarity

Method

Participants A total of 97 participants, 64 females and 33 males (average age 21 years), mainly students at the University of Leuven, participated in this task. Each was paid the equivalent of $10/h.

Materials and Procedure We collected within-category similarity ratings for 15 categories with 420 exemplars in total. The procedure was identical to Experiment 1. Each participant rated all pairwise combinations of the exemplars of at least one and at most seven categories, with the only restriction that the exemplar pairs of the contrast categories *Fruit* and *Vegetables* were never rated by the same participant. They completed the task in between 1 and 5 hours. They never participated more than 1 hour in a single session and always took a break of at least 2 hours before continuing.

Experimental Results All exemplar pairs of the 15 categories were rated by at least 15 and by at most 22 different participants. We removed 5% of the participants, as they correlated less than .45 with the average ratings. The resulting ratings were all very reliable with Spearman Brown split-half correlations ranging from .85 to .96. Prior to further analysis, three items were removed because they were unknown to most participants (*Komodo dragon*, *iguanodon* and *spanner*). In addition we also removed three concepts that were compounds separated by a blank (e.g., *red cabbage*) and five words with ambiguous meanings (e.g., *golf*, which means both 'wave' and 'sport' in Dutch).

Model Results The same weighting and similarity functions were used as those described in Experiment 1. Table 2 gives the correlations of the similarity coefficients from the models with the human similarity ratings for each category. In addition we also computed the correlations for the domains.

The Feature Model gives the best agreement with the judged similarities. Next are the Association and the Context model. The WN-D ontology model had the worst performance. Looking at the domain averages the best results were found for *Artifacts* (.46) and *Activities* (.37), which were still below the values found for any of the other models.

Table 2 shows that at a category level, the models are not always consistently ranked. Consider the category *Insects* for example. Here the Association model gives a better account than the Feature model. In summary, it is easy to see that the variability across categories indicates that different semantic models capture different semantic content.

The averages for the four domains also indicate that the correlations for some domains differ systematically compared to others. For instance, in the ontology model all Artifact categories except *Clothing* receive relatively high correlations. The *Natural Foods* and *Animals*, by contrast, correspond less with human ratings. The reason for this result becomes clear when we look at the structure of the EuroWordNet categories in more detail. Typically, *Artifacts* display a rather fine-grained structure. With the *Natural Kinds* (*Food* and *Animals*), this detail is rather exceptional: often, all exemplars of a category are listed as immediate daughters of the category name. On the basis of a tree structure, it is then impossible to arrive at reliable similarity figures. In general, the results for *Natural Kinds* are not as good as those from *Artifacts* and *Activities* for all models but the Feature model.

How can we explain these domain differences? One explanation is that with concrete concepts, humans employ a holistic – mostly perceptual – comparison strategy, which might be underestimated by our models. While the Feature model contains many perceptual features, it is not always clear how to weight them independent from the task at hand. For instance, in the case of *Insects*, the distinction between flying insects and non-flying insects has a profound influence on the similarity ratings, even though this feature constitutes a characteristic rather than defining feature (Smith, Shoben, & Rips, 1974) and is just one of the 214 features of our *Insects* matrix. According to this account, perceptual information is important for *Natural Kinds* such as *Fruit*, *Vegetables* and *Animals*, while other types of information such as functional and thematic information is a larger determinant of *Artifacts* and *Activities*.

Discussion

We have investigated whether several sources of semantic information — features, word associations, co-occurrence in context and an ontology structure — are able to model human similarity ratings. Our results suggest that they all have the potential to do so, but that this potential is not always fully realized. As expected, most models do well when similarity is considered covering a wide conceptual space as was the case in Experiment 1.

However, this was not the case when only a small region of this concept space is considered. While a Feature model gave a good account for the fine-grained intra-category similarity judgments, associations and context co-occurrences led to slightly less positive numbers. Furthermore, the ontology

Table 2: Correlations of the model predicted and human similarity judgments of Experiment 2 for *n* pairs and 15 different categories.

Domain	Category	*n*	Feat.	Asso.	Context	WN-D
Natural Food	Fruit	406	.75**	.67**	.22**	.07
	Vegetables	325	.72**	.47**	.31**	.29**
		731	**.74****	**.59****	**.26****	**.16****
Animals	Birds	300	.77**	.53**	.49**	-.01
	Fish	120	.79**	.77**	.42**	.44**
	Insects	253	.55**	.71**	.28**	.08
	Mammals	351	.77**	.53**	.35**	.11*
	Reptiles	78	.89**	.72**	.26*	.49**
		1102	**.73****	**.61****	**.37****	**.13****
Artifacts	Clothing	378	.72**	.57**	.34**	.25**
	Kitchen Utensils	465	.78**	.53**	.50**	.46**
	Musical Instruments	276	.81**	.58**	.46**	.68**
	Tools	325	.70**	.56**	.45**	.50**
	Vehicles	351	.83**	.76**	.63**	.49**
	Weapons	153	.85**	.71**	.69**	.39**
		1948	**.77****	**.60****	**.50****	**.46****
Activities	Professions	3577	.80**	.63**	.63**	.32**
	Sports	105	.86**	.82**	.63**	.53**
		455	**.83****	**.70****	**.64****	**.37****

Note: * $p < .05$, ** $p < .01$ (two-tailed)

model proved least appropriate. To gain a better insight in these differences, we looked at extreme cases in the scatter plots and influence statistics of the human and model-based similarities. This investigation revealed two interesting patterns. First, the Association and Context models capture thematic relation semantics, which might be less important in the similarity judgments. For example when plotting the association coefficients against human judgments, pairs like *judge-lawyer*, *car-bicycle*, *scarf-mittens* or *furnace-apron* are considered more similar according to the model than to human judges. The influence of such thematic or situational properties is also present in the Context model, where it predicts higher similarities for pairs like *bow-sword*, *lama-horse*, or *bus-taxi*. Importantly, such situational information is not as strongly present in the deviations for the feature similarities. In this respect, the similarity rating task might have been somewhat artificial compared to everyday processing of meaning where situational properties are useful cues for understanding our environment. It can be expected that a different task such as judgments of semantic relatedness or priming studies would converge better with the Association and Context models and less with the Feature account. Second, we already mentioned that perceptual similarity could also present an important bias for the similarity judgments for *Natural Kinds*. This can be illustrated by the high similarity judgment for *cucumber-zucchini*. These vegetables are visually very similar but can hardly be considered synonyms.

All these models have their idiosyncrasies and limitations. The weakness of the ontology model is least consistent with our initial hypothesis, which claimed that the structure of the taxonomy should basically reflect the similarities and dissimilarities between concepts. However, the main problem of this approach appears to be the low coverage of Dutch EuroWordNet, and the lack of detail in some of its categories. This is particularly apparent with the *Natural Kinds* categories. The Association and Context models suffer from different problems. On the one hand, their low performance as compared to the Features model in Experiment 2 underpins our initial hypothesis that both of these approaches (and the Association model in particular) generally capture a broader type of semantic relatedness than just similarity. On the other hand, the performance of the Context model fully depends on the corpus that was used to collect the co-occurrence frequencies. While more data could lead to higher correlation figures, we also expect that some aspects of a word's semantics (like its visual characteristics) are expressed in text only rarely.

Despite some of these practical limitations, our results highlight the importance of the correct feature selection in modelling of semantic knowledge. The information contained in the Feature model pertains mostly to properties of the entities itself, such as their appearance or function. There are of course a number of fields where the success of vector-

based models does not directly hinge on these features only. In priming studies, for instance, the priming effect may be a result from a number of semantic relations, and could therefore be modelled with several types of semantic knowledge other than concept features. However, when the goal is to model one specific type of semantic relationship, the choice of features appears to be crucial. It is also here that we anticipate the biggest improvements for our models in particular. For example, we expect a context-based model that starts from syntactic relations instead of lexical co-occurrence to give similarity judgments that correlate better with human similarity ratings. After all, the syntactic relations in which a word takes part are linked directly to the features of its concept, which in their turn influence the similarity between two concepts.

Acknowledgments

This work was supported by Grant 3H080198 of the Leuven University Council and Grant G.0513.08 of the Belgian National Science Foundation amended to the third author and a research grant funded by the Research Foundation - Flanders (FWO) to the first and second author.

References

Buchanan, L., Burgess, C., & Lund, K. (1996). Overcrowding in semantic neighborhoods: Modeling deep dyslexia. *Brain and Cognition*, *32*, 111-114.

Budanitsky, A., & Hirst, G. (2006). Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, *35*, 13-47.

Burgess, C., Livesay, K., & Lund, K. (1998). Explorations in context space: words, sentences, discourse. *Discourse Processes*, *25*, 211-257.

Church, K., Gale, W., Hanks, P., & Hindle, D. (1991). Using statistics in lexical analysis. In U. Zernik (Ed.), *Lexical acquisition: Exploiting on-line resources to build a lexicon* (p. 115-164). Hillsdale, NJ: Lawrence Erlbaum Associates.

De Deyne, S., & Storms, G. (2008). Word associations: Norms for 1,424 Dutch words in a continuous task. *Behavior Research Methods*, *40*, 198-205.

De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M., Voorspoels, W., et al. (2008). Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. *Behavior Research Methods*, *40*, 1030-1048.

Fellbaum, C. (1998). *WordNet: An electronic lexical Database*. Cambridge, MA: MIT Press. (Available at http://www.cogsci.princeton.edu/ wn)

Hampton, J. A. (1997). Associative and similarity-based processes in categorization decisions. *Memory & Cognition*, *25*, 625-640.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's Problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, *104*, 211-240.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, and Computers*, *28*, 203-208.

Maki, W. S., & Buchanan, E. (2008). Latent structure in measures of associative, semantic, and thematic knowledge. *Psychonomic Bulletin & Review*, *15*(3), 598-603.

Maki, W. S., McKinley, L. N., & Thompson, A. G. (2004). Semantic distance norms computed from an electronic dictionary (WordNet). *Behavior Research Methods, Instruments, and Computers*, *36*, 421-431.

Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology*, *25*, 431-467.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*, 547–559.

Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, *6*, 1-28.

Nelson, D. L., McEvoy, C. L., & Dennis, S. (2000). What is free association and what does it measure? *Memory & Cognition*, *28*, 887-899.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, and Computers*, *36*, 402–407.

Pereira, F., Tishby, N., & Lee, L. (1993). Distributional clustering of English words. In *31st annual meeting of the acl* (p. 183-190).

Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, *8*, 627–633.

Sahlgren, M. (2006). *The Word-Space Model. Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Unpublished doctoral dissertation, University of Stockholm.

Sloman, S. A., Love, B. C., & Ahn, W. K. (1998). Feature centrality and conceptual coherence. *Cognitive Science*, *22*, 189-228.

Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, *81*, 214-241.

Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (2004). Experimental Cognitive Psychology and its Applications. In A. Healy (Ed.), (chap. Word association Spaces for Predicting Semantic Similarity Effects in Episodic Memory.). Washington, DC: American Psychological Association.

Storms, G., De Boeck, P., & Ruts, W. (2000). Prototype and exemplar based information in natural language categories. *Journal of Memory and Language*, *42*, 51-73.

Vossen, P. (1998). *EuroWordNet: A multilingual database with lexical semantic networks for european languages*. Dordrecht: Kluwer.