Using Distributional Measures to Model Typicality in Categorization

Louise Connell (louise.connell@ucd.ie)
Department of Computer Science, University College Dublin,
Belfield, Dublin 4, Ireland

Michael Ramscar (michael@cogsci.ed.ac.uk)
School of Cognitive Science, University of Edinburgh,
2 Buccleuch Place, Edinburgh, EH8 9LW, Scotland.

Abstract

Typicality effects are ordinarily tied to concepts and conceptualization. The underlying assumption in much of categorization research is that effects such as typicality are reflective of stored conceptual structure. This paper questions this assumption by simulating typicality effects by the use of a co-occurrence model of language, Latent Semantic Analysis (LSA). Despite being a statistical tool based on simple word co-occurrence, LSA successfully simulates subject data relating to typicality effects and the effects of context on categories. Moreover, it does so without any explicit coding of categories or semantic features. The model is then used to successfully predict participants' judgements of typicality in context. In the light of the findings reported here, we question the traditional interpretation of typicality data: are these data reflective of underlying structure in people's concepts, or are they reflective of the distributional properties of the linguistic environments in which they find themselves.

Introduction

The world contains a myriad of objects and events that an intelligent observer could seemingly infinitely partition and generalise from. So how it is that humans can adopt a particular partitioning in the mass of data that confronts them? How do they pick out regularities in the stuff of experience and index them using words? What are these regularities? And how do humans recognise, communicate, learn and reason with them? These questions are central to cognitive science, and traditionally, their close linkage has tempted researchers to seek a unified answer to them: categorization – the act of grouping things in the world – has been commonly linked to the representation of concepts[1], with many researchers assuming that a theory of one provides for the other (Armstrong, Gleitman & Gleitman, 1983; Keil, 1987; Lakoff, 1987).

In much of this work, it is assumed that linguistic behavior (such as naming features associated with a concept, c.f. Rosch, 1973) is determined by, and reflective of, underlying concepts that are grounded in perceptual experience of objects and artifacts themselves. Here, we wish to consider the idea that language itself is part of the environment that determines conceptual behavior. A growing body of research indicates that *distributional information* may play a powerful role in many aspects of human cognition. In particular, it has been proposed that people can exploit statistical regularities in language to accomplish a range of conceptual and perceptual learning tasks. Saffran, Aslin & Newport (1996; see also Saffran, Newport, & Aslin; 1996) have demonstrated that infants and adults are sensitive to simple conditional probability statistics, suggesting one way in which the ability to segment the speech stream into words may be realized. Redington, Chater & Finch (1998) suggest that distributional information may contribute to the acquisition of syntactic knowledge by children. MacDonald & Ramscar (this volume) have shown how information derived from a 100 million word corpus can be used to manipulate subjects' contextual experience with marginally familiar and nonce words, demonstrating that similarity judgements involving these words are affected by the distributional properties of the contexts in which they were read.

The objective of this paper is to examine the extent to which co-occurrence techniques can model human categorization data: What is the relationship between typicality judgements and distributional information? Indeed, are the responses people provide in typicality experiments more reflective of the distributional properties of the linguistic environments in which they find themselves than they are of the underlying structure of people's concepts?

Typicality Effects

The first empirical evidence of typicality effects was provided by Rosch (1973), who found participants judged some category members as more (proto)typical than others. Rosch (1973) gave subjects a category name such as *fruit* with a list of members such as apple,

[1] In the experiments reported, we follow the common assumption (Medin & Smith, 1984; Komatsu, 1992) that categories are classes, concepts are their mental representations and that an instance is a specific example of a category member.

fig, olive, plum, pineapple, strawberry, etc. and asked subjects to rate on a 7-point scale how good an example each member was of its category. The results showed a clear trend of category gradedness – apples are consistently judged a typical *fruit*, while olives are atypical. Further evidence underlines the pervasiveness of typicality (or "goodness of example") and its ability to predict a variety of results. Typicality was found to predict reaction times in sentence verification tasks (Rosch, 1973; McCloskey & Glucksberg, 1979) and order of item output when subjects are asked to name members of a category (Barsalou & Sewell, 1985).

Roth & Shoben (1983) showed that the context a concept appears in affects the typicality of its instances. A typical *bird* in the context-free sense may be a *robin*, but if it appears in the context "The bird walked across the barnyard", then *chicken* would instead be typical. Subject reaction times to sentence verification tasks are faster for the contextually appropriate item (*chicken*) than the normally typical, but contextually inappropriate item (*robin*). Roth and Shoben found that measures of typicality determined in isolation no longer play a predictive role once context has been introduced.

Typicality, Substitutability and LSA

According to Rosch (1978): "The meaning of words is intimately tied to their use in sentences. Prototypicality ratings for members of superordinate categories predict the extent to which the member term is substitutable for the superordinate word in sentences."

This notion of contextual substitutability has a parallel in distributional approaches to word meanings (e.g. Landauer & Dumais, 1997; Burgess & Lund, 1997). In a distributional model of word meaning such as Latent Semantic Analysis (LSA), the corpus analysis calculates a contextual distribution for each lexeme encountered by counting the frequency with which it co-occurs[2] with every other word in the corpus. The contextual distribution of a word can then be summarized by a vector – or point in high-dimensional space – that shows the frequency with which it is associated with the other lexemes in the corpora. In this way, two words that tend to occur in similar linguistic contexts – i.e. are *distributionally* similar – will be positioned closer together in semantic space than two words which do not share as much distributional information. By using the proximity of points in semantic space as a measure of their contextual substitutability, LSA offers a tidy metric of distributional similarity.

Rosch (1973; 1978) held that such substitutability arises as a result of similarities between the underlying structures of the concepts representing the words (although describing these underlying structures has proven elusive, see Komatsu, 1992; Ramscar & Hahn, in submission). However, distributional theories suggest that information about substitutability and word similarity can instead be gleaned from the structure of the *linguistic environment*. Such information is readily – and objectively - obtainable for the purposes of model building and hypothesis testing.

[2] How words are used together within a particular context, such as a paragraph or moving-window.

Experiment 1 – Canonical Typicality

The purpose of this experiment is to examine whether data from typicality studies (Rosch, 1973; Armstrong, Gleitman & Gleitman, 1983; Malt & Smith, 1984) can be modeled using a distributional measure. Specifically, it was predicted that subject typicality scores from previous studies would correlate with a distributional measure (LSA; Landauer & Dumais, 1997) when comparing similarity scores for category members against their superordinate category name.

Materials

Each set of typicality data was divided up according to the taxonomy used in the original study: Set A was taken from Rosch (1973), B from Armstrong, Gleitman & Gleitman (1983), and C from Malt & Smith (1984).

Within these three data sets, 18 sets of typicality ratings existed, across 12 separate categories due to overlap between categories used in Sets A, B and C.

Procedure

For each category in each data set, all items were compared in LSA to the superordinate category name and the similarity scores noted. All scores were calculated in LSA using a corpus whose texts are thought to be representative of reading materials experienced by students up to first year in college[3].

The LSA scores were then scaled from the given [-1, +1] range to fit the standard 7-point typicality scale used in the subject studies, where a score of 1 represents the most typical rating. Malt & Smith used the 7-point scale in reverse order (where 7 represented most typical) so these scores were inverted. LSA score scaling was done by aligning the highest of the LSA scores for each category with the most typical rank on the 7-point scale[4]; i.e. the highest LSA score for a category would be matched to 1, and the other scores falling proportionately towards 7.

[3] Using General Reading up to 1st Year College semantic space, with term-to-term comparison and maximum factors.

[4] The exact formula used is as follows: Where X is the LSA score one wishes to scale and M is the maximum LSA score for this category set:
Scaled LSA score = $M - (M - 1) / (M * X)$.

Results

Spearman's rank correlation (*rho*) was used to compare scaled LSA and subject scores. The global rank correlation between the subject ratings and LSA scores across Sets A, B and C (193 items) was *rho*=0.515 (2-tailed $p<0.001$). Many of the categories that failed to produce greatly significant correlations correlated significantly with the removal of one member, due to it having an extremely high or low LSA score (usually because of its low frequency of occurrence in the corpus). See Table 1 for full LSA results. Also, Set A / Set B correlations for their 4 shared categories of *sport*, *fruit*, *vehicle* and *vegetable* were *rho*=1.0 ($p<0.01$), *rho*=0.943 ($p<0.05$), *rho*=0.886 ($p<0.05$) and *rho*=0.886 ($p<0.05$) respectively.

Table 1: Rank correlation coefficients *rho* (with levels of significance *p*) between LSA and subject scores

Set	Category	Initial category	Adjusted category
A	sport	1.000 ($p<0.01$)	
	fruit	0.886 ($p<0.05$)	
	vehicle	0.829 ($p<0.10$)	1.000 ($p<0.05$)
	crime	0.814 ($p<0.10$)	0.975 ($p<$ (0.10)
	bird	0.714 ($p<0.10$)	0.900 ($p<0.10$)
	science	0.414 (-)	0.675 ($p<0.10$)
	vegetable	0.371 (-)	
B	sport	0.811 ($p<0.01$)	
	vehicle	0.788 ($p<0.01$)	
	vegetable	0.580 ($p<0.10$)	0.745 ($p<0.05$)
	fruit	0.539 ($p<0.10$)	0.748 ($p<0.05$)
	female	0.346 (-)	0.558 ($p<0.10$)
C	trees	0.705 ($p<0.01$)	
	clothing	0.521 ($p<0.05$)	0.676 ($p<0.05$)
	furniture	0.466 ($p<0.05$)	0.609 ($p<0.01$)
	bird	0.375 (-)	0.640 ($p<0.05$)
	fruit	0.157 (-)	
	flowers	-0.499 (-)	

Values (-) represent insignificant correlation of p>0.10

It must be noted that the same rank correlation coefficient results in differing levels of significance within Table 1. This is due to different sizes in categories' data sets (from 5 to 20), where the same score could be significant for one size set and not another; e.g. perfect rank correlation of 1.000 is significant to $p<0.01$ with $N=10$, but only to $p<0.05$ when $N=5$. This high sensitivity to the degrees of freedom from small-sized data sets is why one outlying item was capable of skewing the rank correlation. With small data sets such as these, the power of the tests being used is restricted and they are overly sensitive to individual data points. Larger category data sets are to be found in Sets B and C, where although the rank correlation coefficients may be lower, they are more significant. Thus, it seems reasonable to consider as marginally significant those results where $p<0.10$, given the constraints of the data.

Discussion

In this experiment, LSA similarity scores correlated significantly with subject typicality ratings. Without any hand-coding of category membership or salient features, LSA's semantic space successfully modeled gradients of typicality within categories. Significant global correlation existed between LSA-to-subject typicality ratings at *rho*=0.515 ($p<0.001$). Items that subjects judged typical correlated with those that LSA scored highly in similarity with the category name. The same correlation is true of items that subjects judged to be highly atypical members of their category – these received low similarity scores in LSA. The more closely the ranking of LSA scores mirrored that of the subjects', the higher the correlation, and the closer the level of significance dropped to zero.

Regarding the categories themselves, there were cases where LSA modeled a category's typicality gradient successfully in one data set but not in another. An example of this is the category *fruit*, which was modeled with rank correlation of *rho*=0.886 ($p<0.05$) in Set A and 0.748 ($p<0.05$) in Set B (adjusted), but failed to correlate significantly at all in Set C.

Only one of the 5 category types in Set B came from what Armstrong, Gleitman & Gleitman (1983) term as "well-defined" categories – the category *female*. Despite Armstrong, Gleitman and Gleitman's designation of this category as well-defined, it seems reasonable to regard typicality in *female* as one would any other category examined in this experiment – a measure of contextual substitutability. In this case, the contextual substitutability shown by LSA similarity scores failed to convincingly model the typicality scores for *female*, only reaching correlation of 0.558 ($p<0.10$) when the category was adjusted. We propose the reason for this is that typicality ratings for a category such as *female* are subject to social conditioning in a way other categories such as *fruit* or *sport* are not. For example, the item that LSA scored highest against *female* was *housewife*, which was next followed by *chairwoman*. Although this simply reflects the general contextual substitutability of the words across all of LSA's corpora, it also reflects a ranking that may not be found within a social group. It would be inconsistent for a group of subjects to rate *housewife* as the most typical *female* (a stereotyped sexist attitude), while rating *chairwoman* (a stereotyped politically correct attitude) closely behind. Thus LSA may have failed to convincingly model this category's typicality gradient because it reflects an average of social attitudes across

its corpora, and not just those of one particular group – 1980's Philadelphia undergraduates.

One of the most interesting findings is that in 3 out of 4 cases of shared categories between Set A and Set B, LSA provided as good a fit to Set A typicality ratings as Set B did. When the item *skis* was removed from Set A's *vehicle* category, LSA's correlation bettered that of Set B (with the sole exception of the category *vegetable*). This serves to make an important point and put the data in Table 1 into perspective: it suggests that the difference between subject groups in Rosch's (1973) and Armstrong, Gleitman & Gleitman's (1983) experiments is comparable to the difference between LSA and human subjects. In other words, a co-occurrence model like LSA is as successful at matching the typicality gradients of a subject group as another subject group would be.

Experiment 2 – Contextual Typicality

The first experiment indicates that a co-occurrence model such as LSA can be used to model typicality judgements in canonical (context-free) categories. However, categorization is also subject to linguistic context, whose capacity to skew typicality has been demonstrated by Roth & Shoben (1983).

Having examined canonical typicality in Experiment 1, the purpose of Experiment 2 was to test if LSA could be used to predict subject responses for typicality in context. The hypothesis was that LSA could predict human judgements of exemplar appropriateness (typicality) for given context sentences. LSA similarity scores for each context sentence[5] and respective category members were used to form significantly different clusters of appropriate (high scores / high similarity) and inappropriate (low scores / low similarity) items. It was predicted that subject ratings of typicality in context for these items would fall into the same clusters, and that these clusters would also be significantly different.

Materials

Materials consisted of 7 context sets, each of which consisted of a context sentence and 10 possible members of the category. 3 of the context sentences were taken from Roth & Shoben (1983), the other 4 created for this experiment. Category members were chosen in two ways, to form the appropriate and inappropriate clusters for the context.

First, appropriate items were found by randomly selecting 4-5 high-level category members (e.g. *cow* not *calf* for category *animal*) that appeared in the LSA list of 1500 near neighbors of the context sentence[6]. This list corresponds to the 1500 points in LSA's high-dimensional space that are closest to the context sentence, and would receive the highest similarity scores.

Second, inappropriate items were found by compiling a large list of category members and selecting the 5-6 of those that had the lowest (preferably negative) LSA similarity score against the context sentence.

These materials were split into two sections. Each section consisted of 7 context sets, now with 5 items, selected so that there were at least 2 of both appropriate and inappropriate items in the set and so that each category member appeared only once per section. Subjects received one section apiece, with presentation of section 1 or 2 alternated between subjects. All 35 items within each section were presented in random order, resampled for each subject.

Participants

19 native speakers of English took part in this experiment. All were volunteers who participated by completing an electronic questionnaire.

Procedure

LSA Procedure The scores were calculated in LSA by comparing the context sentence to each item in the list, using the same corpus as for Experiment 1[7].

The LSA scores were then scaled from the given [-1, +1] range to fit the standard 7-point typicality scale used in the subject studies. Due to the presence of very low negative LSA scores, this was done by aligning the extremes of the LSA scores for each category with the opposite extremes of the 7-point scale; i.e. the highest LSA score for a category would be matched to 1, the lowest score to 7, and the intermediate scores falling proportionately in between[8].

Human Procedure Participants read instructions that explained typicality and the 7-point scale as per Rosch (1973) and Armstrong, Gleitman & Gleitman (1983). They were then given this example of a context sentence (not used in experiment) "The girl played the GUITAR while the others sang around the campfire",

[5] The LSA score for a sentence is computed by taking a weighted average of the vectors for each word.

[6] The sentence was processed as a pseudodoc using maximum factors in the same semantic space as used in Experiment 1, from which all words in the corpus with a frequency of less than or equal to 5 had been removed.

[7] Using document-to-term comparison and maximum factors.

[8] The exact formula used is as follows: Where X is the LSA score one wishes to scale, M is the maximum LSA score for this category set and L is the midpoint of the LSA score range for this category set:
Scaled LSA score = $4 - [(L - X) * 3] / (L * M)$.
(4 = midpoint of 7-pt scale; 3 = scale end [7] – midpoint [4]).

and told to consider the appropriateness of the capitalized word in the context given.

Participants were asked not to spend more than 10 seconds deciding on what score to give, and were told that it would not be possible to go back and change an answer (the questionnaire was set up to prevent participants from doing this).

Results

Subjects agreed with LSA's predictions of typicality for 62 of the total 70 items – 10/10 items in 3 context sets, 9/10 items in 3 further context sets, and 5/10 in the remaining context set. Significant difference in clusters, not rank correlation, is the important factor here, because even subject data with low correlation to the LSA score may fall into the two specified clusters (and thus provide support for the main prediction).

Table 2: Wilcoxon's *W* and significance of difference between clusters for each context sentence.

Context Sentence	LSA	Subjects
Stacy volunteered to milk the *animal* whenever she visited the farm *	10 ($p<0.01$)	10 ($p<0.01$)
Fran pleaded with her father to let her ride the *animal* *	15 ($p<0.01$)	15 ($p<0.01$)
The *bird* swooped down on the helpless mouse and carried it off	10 ($p<0.01$)	10 ($p<0.01$)
Jane liked to listen to the *bird* singing in the garden	15 ($p<0.01$)	18 ($p<0.1$) *10 ($p<0.05$) adjusted*
Jimmy loved everything sweet and liked to eat a *fruit* with his lunch every day	15 ($p<0.01$)	18 ($p<0.1$) *10 ($p<0.05$) adjusted*
Sophie was a natural athlete and she enjoyed spending every day at *sport* training	15 ($p<0.01$)	19.5 ($p<0.1$) *10.5($p<0.05$) adjusted*
During the mid morning break the two secretaries gossiped as they drank the *beverage* *	15 ($p<0.01$)	25 ($p<0.7$)**

* Sentences taken from Roth & Shoben (1983)
** Not significant but included for completeness

For all 7 context sets, Mann-Whitney (Wilcoxon Summed Ranks, 2-tailed) tests showed the LSA scores fell into two significantly different clusters. When testing subject scores for difference between the predicted clusters, results varied from three context sets showing significant differences at $p<0.01$ (those at 10/10 agreement), to one set failing to achieve any significant difference at $p=0.69$ (5/10 agreement). Data for clustering in both LSA and subject scores are given in Table 2. Three of the context sets that only produced clusters which were significantly different to $p<0.10$ were those where subjects agreed with LSA-predicted clusters for 9/10 items. With the removal of this lone contentious item, each of these three adjusted subject sets achieved significance of $p<0.025$ (see Table 2).

Discussion

The results support the basic hypothesis that, in the majority of cases, distributional information (in this case modeled in LSA) can predict whether members of a category will be appropriate or inappropriate in a given context. In other words, it can predict human judgements of typicality in context as well as in canonical categories (as demonstrated in Experiment 1). For example, LSA predicted in the context set for *animal* ("Fran pleaded with her father...") that the item *elephant* would be placed in the inappropriate cluster, even though it is entirely possible to ride on an elephant.[9]

In 3 of the 7 context sets, subject typicality scores agreed with LSA predicted clusters for 10/10 items and separated the clusters to a difference significance of $p<0.01$. These sets involved natural kinds as the category for which typicality was taken (*animal*, *bird*). In a further 3 context sets, subjects agreed with LSA's clustering for 9/10 items and separated the clusters to a significant difference of $p<0.05$ when these 9 items were considered. For these sets, two categories were of natural kinds (*bird*, *fruit*) and one was an abstract artifact kind (*sport*). Finally, the context set for which only 5/10 items were agreed to be in the predicted clusters was also for an artifact kind (*beverage*). This suggests that distributional information (or at least, LSA) may perform better in predicting the contextual typicality of natural kinds than artifact kinds. This is perhaps as a result of the vectors for artifact kinds containing a greater degree of contextual variation and thus scoring more unpredictably against the context sentence. Such a theory is compatible with psychological data showing that artifact kinds are processed differently because they may be found in a variety of functional and relational roles, and/or are often polysemous (e.g. Wisniewski & Gentner, 1991).

General Discussion

The success of a distributional measure (LSA) in these modeling experiments suggests interesting possibilities for a theory of categorization based in context, that incorporates information from the structure of language as well as from the structure of the world. Distributional models of language use a representation that is learned from the language alone, assuming that the way words co-occur with one another gives rise to

[9] Although we anticipated a problem with participants' judgements here, the prediction was consistent with the data, where *elephant* received a typicality score of 4.1 and resided in the inappropriate cluster. Is this respect, LSA predictions were sometimes unexpectedly appropriate.

clues about their semantic meaning. Gleitman (1990) has discussed a similar approach with regards to first language acquisition, where this type of representation can be easily learned from an individual's response to their linguistic environment, thus lending a psychologically plausible base to such a theory.

In this respect, the results of these simulations raise interesting questions with regard to people's mental representations of the meanings of words: Do people use distributional information to construct their representation of word meanings? Or are distributional properties of words (which models such as LSA extract) merely an epiphenomenonon; a reflection of the fact that underlying concepts share certain semantic features? By the latter account, the distributional properties associated with words would arise *because* the concepts underlying the words possess certain features, and it is sensitivity to similarities between these concepts that subjects actually manifest. However, MacDonald & Ramscar (in submission) have shown that manipulating the distributional properties of the contexts in which nonce words are read can significantly influence similarity judgements between existing words and nonces. This indicates that not *all* distributional responses can be explained in terms of existing conceptual structure – nonce words won't have an existing conceptual structure. Equally, it seems highly unlikely that the structure of the linguistic environment is entirely unreflective of the structure that people extract from their interactions with the world.

What the results presented here (and other distributional research) seem to indicate is that any proper characterization of conceptual thought will have to consider more than just the information that comes from physical experience and the physical environment. One must also consider experience of language, and the structure of the linguistic environments in which speakers find themselves.

Acknowledgments

We thank Dermot Lynott and Dan Yarlett for many insightful comments on the work reported in this paper.

References

Armstrong, S. L., Gleitman, L. R. & Gleitman, H., (1983). What some concepts might not be. *Cognition, 13*, 263-308.

Barsalou, L. W. and D. R. Sewell (1985). Contrasting the representations of scripts and categories. *Journal of Memory and Language, 24*, 646-665.

Burgess, C. & Lund, K., (1997). Modelling parsing constraints with high-dimensional context space. *Language and Cognitive Processes, 12,* 1-34.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition, 1,* 3-55.

Keil, F.C. (1987). Conceptual Development and Category Structure. In U. Neisser (Ed*.), Concepts and Conceptual Development: Ecological and intellectual Factors in Categorization.* Cambridge:Cambridge University Press.

Komatsu, L., (1992). Recent views of conceptual structure. *Psychological Bulletin, 112,* 500-526.

Lakoff, G., (1987). *Women, Fire and Dangerous Things.* University of Chicago Press.

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.

Malt, B. & Smith, E. (1984). Correlated properties in natural categories. *Journal of Verbal Learning and Verbal Behavior, 23,* 250-269.

McCloskey, M., & Glucksberg, S. (1979). Decision processes in verifying category membership statements: Implications for models of semantic memory. *Cognitive Psychology, 11,* 1-37.

MacDonald, S & Ramscar, M. J. A. (this volume) Testing the distributional hypothesis: the influence of context on judgements of semantic similarity. *This volume*.

Medin, D. & Smith, E. (1984). Concepts and concept formation. *Annual Review of Psychology*, 35, 113-138.

Redington, M., Chater, N. & Finch, S. (1998). Distributional information: a powerful cue for acquiring syntactic categories. *Cognitive Science*, *22*, 425-469.

Ramscar, M. J. A. & Hahn, U. (in submission). *What family resemblances are not: Categorisation and the concept of 'concept'*. Manuscript in submission

Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.) *Cognitive Development and the Acquisition of Language*. New York: Academic Press.

Rosch, E., (1978). Principles of Categorization. In E. Rosch and B. B. Lloyd (Eds.), *Cognition and categorization*. Hillsdale, N.J.: Erlbaum.

Roth, E. M. & Shoben, E. J., (1983). The effect of context on the structure of categories. *Cognitive Psychology, 15,* 346-378.

Saffran, J. R., Aslin, R. N. & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926-1928.

Saffran, J. R., Newport, E. L. & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606-621

Wisniewski, E. J., & Gentner, D. (1991). On the combinatorial semantics of noun pairs: Minor and major adjustments to meaning. In G. B. Simpson (Ed.), *Understanding word and sentence*. Amsterdam: Elsevier.