Word Association Strength, Mutual Information and Semantic Similarity

Anat Prior (anatprior@huji.ac.il)
Department of Psychology, The Hebrew University of Jerusalem
Mount Scopus, Jerusalem 91905 Israel

Ma'ayan Geffet (mary@cs.huji.ac.il)
Department of Computer Science, The Hebrew University of Jerusalem
Giv'at Ram, Jerusalem 91904 Israel

Abstract

Associated word pairs differ in their degree of Association Strength, i.e., how commonly the target is given by subjects as a response to the cue. In the present work we investigated the degree to which Association Strength can be predicted by a measure of Mutual Information of the word pair, by the Semantic Similarity of the words, or by both factors jointly. We examined this issue in two compilations of free association norms, one in Hebrew and one in English, and analyzed circa 6,000 associated pairs in total. Further, English associated pairs were classified as *syntagmatic* or *paradigmatic* associates, while Hebrew noun pairs were classified according to the semantic relation between them (i.e., *synonyms* and *antonyms*). We found both Mutual Information and Semantic Similarity were significantly correlated with Association Strength for the English associates of both types. For the Hebrew associates, Mutual Information was significantly correlated with Association Strength for noun pairs related *idiomatically, functionally* and *hierarchically*, while Semantic Similarity was found to reliably predict Association Strength only for *antonyms*. The importance of these results for understanding the cognitive operations underlying the free association task is discussed.

Introduction

Free association data are widely used for stimulus selection in psycholinguistic and memory research. Various attempts have been made to investigate the mechanisms underlying the associative structures that are expressed in free association norms. Classic laws of association (Hume, 1738/1962) would lead us to expect the patterns of word co-occurrence in language to determine associative strength to a large degree. That is, words tending to co-occur more frequently in spoken and written language would often be given as responses in a free association task.

Even a cursory examination of word-association norms leads us to formulate the hypothesis that similarity in meaning plays an important role in determining associative structure as well. Indeed, McDonald & Lowe (1998) have found word pairs that are both semantically related and associated are more similar (on a corpus derived measure) than semantically related pairs that are not normatively associated, though others have found no such relation (Lund, Burgess & Audet, 1996).

Recent explorations of this issue have utilized currently available computing power and large language corpora. Co-occurrence rates of words can be calculated from these corpora, and various measures of similarity have also been developed. The general idea is that words similar in meaning appear in similar contexts (Miller & Charles, 1991). Thus, representations of words based on their co-occurrence with other words can capture the Semantic Similarity between words.

Corpus derived measures can, therefore, allow us to examine both local co-occurrence patterns (words that tend to appear close together in text) and words similar in their global co-occurrence patterns (words that tend to appear in similar environments throughout the text), the latter being a measure of Semantic Similarity.

Previous work has found Association Strength to be significantly correlated with frequency of co-occurrence (Plaut, 1995; Spence & Owens, 1990). Further, words that were both semantically related and normatively associated were found to co-occur more frequently than words that were only semantically related (McDonald & Lowe, 1998). Others, however, have found co-occurrence to predict Association Strength only for pairs which were semantically similar (Lund et al., 1996). Thus, it seems that the relations obtaining among Semantic Similarity, textual co-occurrence and Association Strength call for further investigation.

Besides certain inconsistencies in their results, the works cited above suffer from several other limitations. The number of associated pairs examined was usually quite low (ranging from less than 50 to at most 400). In the study reported here we examined 2000 pairs of associated Hebrew words and 4000 pairs of associated English words. This allows us to extend any previous results and test them on a larger database.

An additional point worth noting is that different studies used various co-occurrence measures: The raw frequency of co-occurrence (Lund et al., 1996); Co-occurrence normalized by the basic frequency of the target (McDonald & Lowe, 1998); or co-occurrence of the pair normalized by subtracting the co-occurrence of the cue with an unrelated target (Spence & Owens, 1990). In the computational linguistics literature, other co-occurrence measures have been suggested. Church & Hanks (1990) proposed a measure of Mutual Information as an objective way of estimating Association Strength. However, in their work they do not systematically compare the proposed measure with human word-association norms, in order to test the validity of this measure. Wettler & Rapp (1993) present such a comparison, though once again, they examine 100 only stimulus words, and their procedure consists of predicting the first associate and does not look at overall correlations between Co-occurrence and Association Strength for the full spectrum of Association Strength. (The results reported, further, do not seem to be very accurate in their ability to correctly predict responses given by human subjects – only 17 of the first associates predicted by the model corresponded to those given by human subjects, and 35 predicted first associates were not given by any human subjects).

The Current Work

In the work reported here, we set out to examine the questions outlined above on a much larger database of associated word-pairs than used previously (in total close to 6000 word pairs were used). We actually used two such databases – one in the English language (Nelson, McEvoy & Schreiber, 1998) and one in the Hebrew language (Ben Gurion Association Norms). Therefore, any results found in both association databases will be based on a large number of stimuli across two languages. We utilized a more sophisticated measure of textual co-occurrence than those used previously, one based on Mutual Information, which better controls for effects of single word frequency. Subsequently, we joined Mutual Information with a measure of Semantic Similarity (based on global co-occurrence patterns) and examined how well a combination of both these factors predicts Association Strength.

Aside from attempting to solidify previous findings by using larger data sets, we also examined two novel questions: In the Hebrew association norms, we classified each pair of nouns according to the semantic relation obtaining between the words. We defined 10 semantic relations: *synonymy, antonymy, meronomy* (part-whole and whole-part), *hierarchic* relations (category-exemplar, exemplar-category and category coordinates), *idiomatic, functional* relations and not otherwise specified (Cruse, 1986). The semantic classification of Hebrew pairs was singular, though certain pairs might embody more than one relation (*antonyms*, for example, are almost by definition also *category coordinates* – 'black' and 'white' both being colors). Our decisions in these cases might have influenced the final results.

We expected the two factors examined to have different influences on Association Strength in the various semantic categories. Thus, we predicted that for *synonyms* and *antonyms* Semantic Similarity would have a greater weight in determining Association Strength than for other categories, while Mutual Information as a measure of co-occurrence would have a lesser influence, since these types of words do not tend to co-occur frequently in language. Conversely, Mutual Information was expected to be the important factor determining Association Strength for noun pairs related *functionally* and *idiomatically*, relations that to a great extent are manifest in usage patterns. All analyses were performed separately for each type of semantic relation, allowing us to compare the role of the two factors.

The English association norms include words from different parts of speech. We made a distinction between pairs in which both words came from the same part of speech and pairs in which words came from different parts of speech. This distinction is reminiscent of the well-established classification of associations as being *paradigmatic* or *syntagmatic*, respectively (e.g. Nelson, 1977). Again, we expected the two predicting factors to have differential influences on Association strength in these two categories. For the *paradigmatic* pairs, a more dominant role for Semantic Similarity was expected, while for the *syntagmatic* pairs we expected to see a stronger relation between Mutual Information and Association Strength.

The Word Association Data

In Hebrew, we used 1700 noun-noun pairs, and in English we used 4000 word pairs, from all parts of speech. Association Strength ranged from 5% to 94%. All associations were collected by the method of a single response to each cue, to avoid chaining of associations (Nelson, McEvoy & Dennis, 2000). All Hebrew nouns used appeared in the text corpus at least 50 times. All English words had a frequency of at least 10 per million (Kucera & Francis, 1967). These criteria were adopted in order to limit the influence of spurious associations and unreliable appearance patterns on the analysis.

Two independent judges classified the noun pairs according to semantic relation, with a concordance rate of 89%. All remaining pairs were examined jointly, and an agreement was reached as to their classification. English pairs were classified as *paradigmatic* – when the two words were the same part of speech, or *syntagmatic* otherwise.

The Text Corpora

The Hebrew corpus is a compilation of newspaper text, while the English corpus is language gathered from UseNet newsgroups. Both corpora are of similar size - around 128,000,000 words each. Even the most representative of texts does not capture the full world knowledge of the reader. For example, entities that are closely related in the world will not necessarily co-occur frequently in the text if the presence of one of them might allow the average reader to implicitly derive the existence of the other. An additional possibility is that our choice to limit ourselves to co-occurrence within sentence boundaries led to the omission of relations extending beyond these boundaries. These problems are exacerbated by the fact that our text corpus was fairly constrained. Unfortunately, the corpus of newspaper text in Hebrew is quite topically limited, with the bulk dealing mainly with economy, politics, sports and cultural events. Thus, the most Semantically Similar Hebrew pairs are those most characteristically found in newspapers, e.g. 'prime minister' and 'government'.

Textual Co-occurrence – Mutual Information

Previous work found no significant relation between Association Strength and co-location separation – the distance of the two words (Lund et al., 1996; though see Spence & Owens, 1990). We therefore decided to define two words as co-occurring if they appeared within the same sentence, in any order. This decision was further motivated by our belief that co-occurrence within a sentence is more indicative of joint processing of the two words (leading to association) (Prior & Bentin, in press) than co-occurrence within a random textual window, which often transcends sentential boundaries. The number of co-occurrences was counted for each pair throughout the relevant text corpus, and the Mutual Information of each pair was then computed using the following formula (Fano, 1961):

$$I(x,y) = \log \frac{P(x,y)}{P(x)P(y)}$$

where P(x,y) denotes the probability of co-occurrence, and P(x), P(y) denote the probability of appearance of each single word. Intuitively, Mutual Information compares the probability of observing the two items together with the probability of observing each item independently (their probability of co-occurring by chance). This measure increases as two words systematically appear together above and beyond the chance probability.

Semantic Similarity/Distance

The Semantic Distance of word pairs in English was defined as the Euclidean distance between the HAL representations of the two words, based on the 3.8 million words CHILDES database. Thus, for English pairs, we predicted a negative relation between Semantic Distance and Association Strength – the closer the two words, the stronger the association.

The Semantic Similarity of Hebrew words was defined using distributional vectors extracted from the present corpus. We built a distributional vector for each word, which contains frequencies of words (attributes) co-occurring with it within sentence boundaries. Semantic Similarity was computed by applying the Min/Max similarity metric, shown to be a superior metric for this purpose (Lotan, 1998):

$$sim(u,v) = \frac{\sum_{att \in Both(u,v)} min(assoc(u,att), assoc(v,att))}{\sum_{att \in Either(u,v)} max(assoc(u,att), assoc(v,att))}$$

where

$$assoc(u,att) = \log(1 + freq(u,att))$$

Using a logarithmic transformation of the raw frequency counts decreases the influence of highly frequent attributes. The upper bound of this metric is 1, for two identical vectors.

Note that both Mutual Information and Semantic Similarity were defined as symmetric measures, though we are aware that others have occasionally chose to define them otherwise (e.g. Church & Hanks, 1990, for an asymmetric definition of Mutual Information)

Analysis and Results

Of 1730 Hebrew associated pairs analyzed, only 112 were classified as *other*. The great majority of noun pairs adhered to one of the semantic relations defined, albeit the largest category is that of *functional* relations, which probably has the least rigid formal definition.

Figure 1 presents the mean Association strength, Mutual Information and Semantic Similarity for each of the associative categories. All three factors were found to differ significantly for the semantic categories defined (Using one-way ANOVA, with 9 df, all p<0.001). Therefore, a separate analysis of the semantic categories is called for.

The correlation of each of the factors (Mutual Information and Semantic Similarity) with Association Strength (following a logarithmic transformation, introduced since the original distribution was highly skewed towards low association values) was calculated. Results are presented in Table 1. As a second step, both factors were entered into a multiple regression, to see how well they predict Association Strength. The R^2 is reported (corresponding to the percentage of variance in association strength attributed to both predicting factors).

Overall, Mutual Information had a significant correlation with Association Strength (r=0.148). As is evident from the breakdown of noun pairs according to semantic relation, this correlation was significant in only 4 categories: *part-whole*, *category coordinates*, *functional* relations and *idiomatic* relations. Our predictions, therefore, were partially successful: In addition to finding a significant role for Mutual Information in determining Association Strength for *functional* and *idiomatic* pairs, as expected, we found Mutual Information to play a significant role for two out of three *hierarchic* relations defined.

Semantic Similarity, on the other hand, was not significantly correlated with Association Strength overall, but there was a significant correlation between these two factors for the subgroup of *antonyms.* Once again, our expectations are fulfilled to a certain extent, though the evident lack of correlation between Semantic Similarity and Association Strength for *synonyms* is quite surprising. Curiously, we also found significant negative correlations between Semantic Similarity and Association Strength for the *part-whole* and *idiomatic* categories.

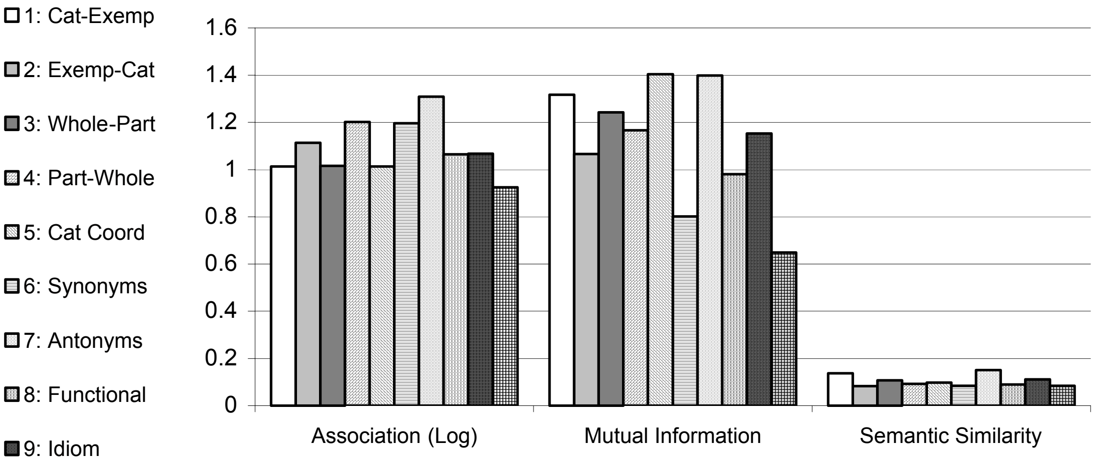

Figure 1: Mean Association Strength (log), Mutual Information and Semantic Similarity for Hebrew associated pairs of different semantic relations.

Table 1: Correlations of Mutual Information and Semantic Similarity with Association Strength (log)

Semantic Relation (N)	Mutual Information	Semantic Similarity	R^2 Joint Prediction
All Pairs (1730)	0.148‡	-0.008	0.022‡
Category - Exemplar (59)	0.208	0.118	0.053
Exemplar - Category (124)	0.289	0.075	0.092†
Whole-Part (86)	-0.017	0.012	0.00
Part-Whole (81)	0.297†	-0.231§	0.145†
Category Coordinates (278)	0.187†	0.071	0.038†
Synonyms (147)	0.136	0.037	0.021
Antonyms (43)	0.242	0.354§	0.187§
Functional (682)	0.169‡	-0.052	0.034‡
Idiomatic (119)	0.208§	-0.222§	0.044
Other (112)	0.055	-0.022	0.057§

§p<0.05 †p<0.01 ‡p<0.001

The English associated pairs analyzed, were classified as either *paradigmatic* or *syntagmatic*. As expected from a college-age sample, there is a majority of paradigmatic responses (65%) (Nelson, 1977). The means of all three variables, for both categories, are presented in Figure 2. The differences between the categories, for all variables, are statistically reliable (Using a one-way A]

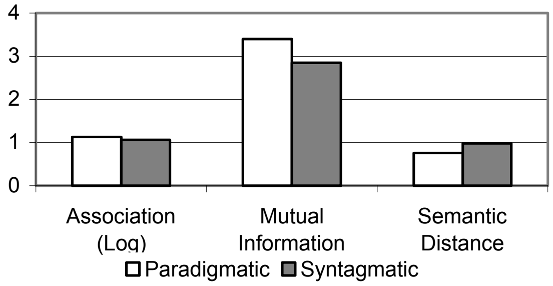

Figure 2: Mean Association Strength (log), Mutual Information and Semantic Distance for *Syntagmatic* and *Paradigmatic* associated English word pairs.

The correlations between each of the predicting variables, namely Mutual Information and Semantic Distance, and Association strength (following a logarithmic transformation) were calculated, and are presented in Table 2.

Overall, a significant correlation was found between Association strength and both factors – a positive relation with Mutual Information, and an expected negative relation with Semantic Distance. We then calculated these correlations for each type of association separately. For both cases, the correlations remained significant. Further, no difference is evident between the patterns emerging for the two types of associations. Our prediction that Semantic Distance would have a greater influence on Association Strength for *paradigmatic* associates, while the same would be true of Mutual Information for *syntagmatic* associates, is not borne out by the data.

Table 2: Correlations of Mutual Information and Semantic Distance with Association Strength (log)

Association Type (N)	Mutual Information	Semantic Distance	R^2 Joint Prediction
All Pairs (4014)	0.199‡	-0.169‡	0.057‡
Paradigmatic (2633)	0.175‡	-0.129‡	0.040‡
Syntagmatic (1381)	0.212‡	-0.161‡	0.066‡

‡p< 0.001

Discussion

The study described above had two main goals. First, we sought to settle several inconsistencies apparent in the literature regarding the relations obtaining between Association Strength and various corpus-derived measures (of both local and global co-occurrence patterns). In order to do so, we analyzed a large number of associated pairs, 6000 in total, from a wide range of association strengths (starting at 5%), and in two languages (English and Hebrew). We believe that conclusions based on such a sample are highly reliable, and thus allow us a high level of confidence in their validity. Our second goal was to investigate whether these relations were modulated by association type. To this end, all Hebrew noun-noun associated pairs were classified by the semantic relation existing between the words, whereas all English associated pairs were classified as *paradigmatic* or *syntagmatic*. Each association type was then analyzed separately.

We found a highly significant positive correlation between Association Strength and Mutual Information, for both Hebrew and English associates, (r=0.148, and r=0.199, respectively). As expounded above, we chose to use Mutual Information instead of more raw counts of co-occurrence used previously, since it allows better control for effects arising from the frequency of single words (Church & Hanks, 1990).

Our results are reminiscent of those reported by others (Lund et al., 1996; McDonald & Lowe, 1998; Spence and Owens, 1990; Wettler & Rapp, 1993), though the range of correlations reported in these studies is wide, and various measures of co-occurrence were used. The current study was the first to use a measure of Mutual Information, and the results lend support to the claim that it is indeed related to Association Strength, a psycholinguistic variable collected from subjects performing a free-association task. However, we do not feel that the magnitude of the correlation found (both r<0.2) allows one to endorse the suggestion that Mutual Information replace Association Strength as a basis for stimuli selection for psychological experimentation (Church & Hanks, 1990). Further, it appears that co-occurrence patterns cannot be the sole determiner of Association Strength, and that other factors should be taken into account. From a cognitive perspective, the free association task probably does not tap representations, or lexical networks, constructed only on the basis of usage patterns, but relies on semantic networks as well.

With this in mind, we turned to examine the correlation between Association Strength and Semantic Similarity. For the Hebrew associated pairs, we found no systematic relation overall, but we did find a strong correlation between the two factors for the subgroup of *antonyms* (r=0.354). This is especially paradoxical as over 90% of associated pairs were successfully classified as adhering to one of the formal semantic relations, hinting to the importance of semantic factors in free association data. On the other hand, for the English data, a significant correlation emerged (r=-0.169): the more similar the words, the stronger the association between them. These results reflect the mixed findings reported previously. Lund et al. (1996) found no relation between Association Strength and Semantic Distance, while McDonald & Lowe (1998) reported semantically related associated pairs to be more semantically similar than non-associated pairs. The contrast with Lund et al. (1996) is particularly striking, since the very same similarity measure was used. Perhaps the limited range of Association Strengths, and the smaller sample size used in their study (less than 400 pairs, compared with more than 4000 in the present study) did not allow the relation between the two factors to emerge. We feel confident that the correlation we found between Semantic Similarity and Association Strength reflects a true phenomenon, since it is not only statistically significant (due to the large sample size), but is of the same magnitude of the correlation between Mutual Information and Association Strength, and therefore cannot be easily dismissed.

The divergence in our findings regarding the Hebrew and English data may be attributed to the use of different similarity matrices. Specifically, it seems that the method utilized by Lotan (1998) for testing the Min/Max similarity index against human similarity judgments is unsatisfactory, possibly leading to our somewhat puzzling results. Alternatively, our failure to find a significant relation between the two factors in Hebrew may be due to the fact that the Hebrew word vectors were based on a newspaper text corpus, while English word vectors were based on a more representative language sample. Possibly the Hebrew representations did not capture the full meaning of the words, due to context limitation, and therefore failed to show any correlation with Association Strength. It remains to be tested whether a more representative text corpus, joined with a different Semantic Similarity metric, might yield different results.

The similar pattern of correlations between Association strength and both Mutual Information and Semantic Distance for *syntagmatic* and *paradigmatic* English associates was contrary to our expectation. A possible explanation for this finding lies in the analysis of Deese (1965), who claims the two types of associates to be more similar than commonly believed. Specifically, many associates defined as *syntagmatic*, owing to the words' belonging to different parts of speech, may actually reflect semantic features and do not necessarily arise from patterns of language use.

Finally, our use of a large database of word-associates allowed us to bring forth the role of significant factors predicting Association Strength. For both English and Hebrew associates, Mutual Information was found to significantly correlate with Association Strength, though a more detailed analysis of Hebrew associates showed this correlation to be significant only for pairs maintaining certain types of semantic relations. Semantic Similarity was also found to be a significant predictor of Association Strength, at least for the English associates. Regression models utilizing both factors were found to significantly predict Association Strength, though for most cases the percentage of explained variance was less than 10%. Thus, despite our significant findings, many factors governing Association Strength remain unaccounted for.

The additional factors not addressed in the current study may include density of semantic neighborhood, category typicality, word frequency and concreteness, just to name a few. As with the factors examined above, it stands to reason that the influence of these factors probably varies greatly for different types of associated pairs. Our conclusion is, therefore, that the free association task is best conceived as an amalgamation of several processes operating jointly to produce the norms we are familiar with. Thus, any model of free association must include the influence of representations and connections at both lexical and semantic levels. The present work demonstrates the importance of Mutual Information (as a measure of textual co-occurrence) and of Semantic Similarity (as arising from a textually derived semantic representation) in accounting for the association patterns found in the free associations of human subjects. However, any additional factors influencing Association Strength remain as yet unidentified.

Acknowledgments

We are grateful to Prof. Shlomo Bentin for useful comments, and to Yaron Silberman and Ayelet Landau for invaluable assistance. We thank Drs. David Plaut and Douglas Rohde for making the English language corpus available to us, and are grateful to Drs. Ping Li and Curt Burgess for providing the HAL vectors. This research was supported by a doctoral research stipend to Anat Prior from the Israeli Foundation Trustees.

Bibliography

Ben Gurion Association Norms, Unpublished.

Church, K. & Hanks, P. (1991). Word association norms, mutual information and lexicography. *Computational Linguistics*, *16 (1)*, 22-29.

Cruse, D.A. (1986). *Lexical Semantics (Cambridge Textbooks in Linguistics)*. Great Britain: Cambridge University Press.

Deese, J.E. (1965). *The Structure of Associations in Language and Thought*. Baltimore: Johns Hopkins Press.

Fano, R. (1961). *Transmission of Information*, Cambridge, Massachusetts: MIT Press.

Hume, D. (1738/1962). *A Treatise of Human Nature*. London: J.M. Dent & Sons, Ltd.

Kucera, H., & Francis, W. N. (1967). Computational Analysis of Present-day American English. Providence, RI : Brown University Press.

Lotan, E. (1998). *Automatic construction of a statistical thesaurus*. M.Sc. Thesis, Department of Mathematics and Computer Science, Bar Ilan University, Israel.

Lund, K., Burgess, C. & Audet, C. (1996). Dissociating semantic and associative word relationships using high-dimensional semantic space. In Proceedings of the 18th Annual Conference of the Cognitive Science Society, (pp.603-608). Mahwah, NJ: Lawrence Erlbaum Associates.

McDonald, S. & Lowe, W. (1998). Modeling functional priming and the associative boost. In *Proceedings of the 20th Annual Conference of the Cognitive Science Society*, (pp.675-680). Hillsdale, NJ: Lawrence Erlbaum Associates

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

Nelson, D.L., McEvoy, C.L. & Dennis, S. (2000). What is free association and what does it measure? *Memory and Cognition, 28(6),* 887-899.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. http://www.usf.edu/FreeAssociation/.

Nelson, K. (1977). The syntagmatic-paradigmatic shift revisited: A review of research and theory. *Psychological Bulletin, 84 (1)*, 93-116.

Plaut, D.C. (1995). Semantic and associative priming in a distributed attractor network. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, (pp. 37-42). Mahwah, NJ: Lawrence Erlbaum Associates.

Prior, A. & Bentin, S. (In press). Incidental formation of episodic associations: the importance of sentential context. *Memory and Cognition.*

Spence, D.P. & Owens, K.C. (1990). Lexical co-occurrence and association strength. *Journal of Psycholinguistic Research, 19 (5)*, 317-330.

Wettler, M. & Rapp, R. (1993). Computation of word associations based on the co-occurrences of words in large corpora. In: *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives* (pp. 84-93). Columbus, Ohio.