Semantic Similarity Computation for Abstract and Concrete Nouns Using Network-based Distributional Semantic Models

Elias Iosif*, Alexandros Potamianos*, Maria Giannoudaki*, Kalliopi Zervanou†
* Dept. of Electronics & Computer Engineering, Technical University of Crete, Greece
{iosife,potam,maria}@telecom.tuc.gr
†Centre for Language Studies, Radboud University Nijmegen, The Netherlands
k.zervanou@let.ru.nl

Abstract

Motivated by cognitive lexical models, network-based distributional semantic models (DSMs) were proposed in [Iosif and Potamianos (2013)] and were shown to achieve state-of-the-art performance on semantic similarity tasks. Based on evidence for cognitive organization of concepts based on degree of concreteness, we investigate the performance and organization of network DSMs for abstract vs. concrete nouns. Results show a "concreteness effect" for semantic similarity estimation. Network DSMs that implement the maximum sense similarity assumption perform best for concrete nouns, while attributional network DSMs perform best for abstract nouns. The performance of metrics is evaluated against human similarity ratings on an English and a Greek corpus.

1 Introduction

Semantic similarity is the building block for numerous applications of natural language processing (NLP), such as grammar induction [Meng and Siu (2002)] and affective text categorization [Malandrakis et al. (2011)]. Distributional semantic models (DSMs) [Baroni and Lenci (2010)] are based on the distributional hypothesis of meaning [Harris (1954)] assuming that semantic similarity between words is a function of the overlap of their linguistic contexts. DSMs are typically constructed from co-occurrence statistics of word tuples that are extracted from a text corpus or from data harvested from the web. A wide range of contextual features are also used by DSM exploiting lexical, syntactic, semantic, and pragmatic information. DSMs have been successfully applied to the problem of semantic similarity computation. Recently [Iosif and Potamianos (2013)] proposed *network-based DSMs* motivated by the organization of words, attributes and concepts in human cognition. The proposed semantic networks can operate under either the *attributional similarity* or the *maximum sense similarity* assumptions of lexical semantics. According to attributional similarity [Turney (2006)], semantic similarity between words is based on the commonality of their sense attributes. Following the maximum sense similarity hypothesis, the semantic similarity of two words can be estimated as the similarity of their two closest senses [Resnik (1995)]. Network-based DSMs have been shown to achieve state-of-the-art performance for semantic similarity tasks.

Typically, the *degree of semantic concreteness* of a word is not taken into account in distributional models. However, evidence from neuro- and psycho-linguistics demonstrates significant differences in the cognitive organization of abstract and concrete nouns. For example, Kiehl et al. (1999) and Noppeney and Price (2004) show that concrete concepts are processed more efficiently than abstract ones (aka "the concreteness effect"), i.e., participants in lexical decision tasks recall concrete stimuli faster than abstract. According to dual code theory [Paivio (1971)], the stored semantic information for concrete concepts is both verbal and visual, while for abstract concepts stored information is only verbal. Neuropsychological studies show that people with acquired dyslexia (deep dyslexia) face problems in reading abstract nouns aloud [Coltheart (2000)], *verifying that concrete and abstract concepts are stored in different regions of the human brain anatomy* [Kiehl et al. (1999)]. The reversal concreteness effect is also reported for people with semantic dementia with a striking impairment in semantic memory [Papagno et al. (2009)].

Motivated by this evidence, we study the semantic network organization and performance of DSMs for estimating the semantic similarity of abstract vs. concrete nouns. Specifically, we investigate the validity of the maximum sense and attributional similarity assumptions in network-based DSMs for abstract and concrete nouns (for both English and Greek).

2 Related Work

Semantic similarity metrics can be divided into two broad categories: (i) metrics that rely on knowledge resources, and (ii) corpus-based metrics. A representative example of the first category are metrics that exploit the WordNet ontology [Miller (1990)]. Corpus-based metrics are formalized as DSM [Baroni and Lenci (2010)] and are based on the distributional hypothesis of meaning [Harris (1954)]. DSM can be categorized into unstructured (unsupervised) that employ a bag-of-words model [Agirre et al. (2009)] and structured that rely on syntactic relationships between words [Pado and Lapata (2007)]. Recently, motivated by the graph theory, several aspects of the human languages have been modeled using network-based methods. In [Mihalcea and Radev (2011)], an overview of network-based approaches is presented for a number of NLP problems. Different types of language units can be regarded as vertices of such networks, spanning from single words to sentences. Typically, network edges represent the relations of such units capturing phenomena such as co-occurrence, syntactic dependencies, and lexical similarity. An example of a large co-occurrence network is presented in [Widdows and Dorow (2002)] for the automatic creation of semantic classes. In [Iosif and Potamianos (2013)], a new paradigm for implementing DSMs is proposed: a two tier system in which corpus statistics are parsimoniously encoded in a network, while the task of similarity computation is shifted (from corpus-based techniques) to operations over network neighborhoods.

3 Corpus-Based Baseline Similarity Metrics

Co-occurrence-based: The underlying assumption of co-occurrence-based metrics is that the co-existence of words in a specified contextual environment indicates semantic relatedness. In this work, we employ a widely-used co-occurrence-based metric, namely, Dice coefficient [Iosif and Potamianos (2010)]. The Dice coefficient between words w_i and w_j is defined as follows: $D(w_i,w_j)=\frac{2f(w_i,w_j)}{f(w_i)+f(w_j)}$, where $f(.)$ denotes the frequency of word occurrence/co-occurrence. Here, the word co-occurrence is considered at the sentential level, while D can be also defined with respect to broader contextual environments, e.g., at the paragraph level [Véronis (2004)].
Context-based: The fundamental assumption behind context-based metrics is that *similarity of context implies similarity of meaning* [Harris (1954)]. A contextual window of size $2H+1$ words is centered on the word of interest w_i and lexical features are extracted. For every instance of w_i in the corpus the H words left and right of w_i formulate a feature vector v_i. For a given value of H the context-based semantic similarity between two words, w_i and w_j, is computed as the cosine of their feature vectors: $Q^H(w_i,w_j)=\frac{v_i.v_j}{||v_i||\ ||v_j||}$. The elements of feature vectors can be weighted according various schemes [Iosif and Potamianos (2010)], while, here we use a binary scheme.

4 Network-based Distributional Semantic Models

Here, we summarize the main ideas of network-based DSMs as proposed in [Iosif and Potamianos (2013)]. The network is defined as an undirected (under a symmetric similarity metric) graph $F=(V,E)$ whose the set of vertices V are all words in our lexicon L, and the set of edges E contains the links between the vertices. The links (edges) between words in the network are determined and weighted according to the pairwise semantic similarity of the vertices. The network is a parsimonious representation of corpus statistics as they pertain to the estimation of semantic similarities between word-pairs in the lexicon. In addition, the network can be used to *discover relations that are not directly observable in the data*; such relations emerge via the systematic covariation of similarity metrics. For each word (reference word) that is included in the lexicon, $w_i \in L$, we consider a sub-graph of F, $F_i=(N_i,E_i)$, where the set of vertices N_i includes in total n members of L, which are linked with w_i via edges E_i. The F_i sub-graph is referred to as the semantic neighborhood of w_i. The members of N_i (neighbors of w_i) are selected according to a semantic similarity metric (co-occurrence-based D or context-based Q^H defined in Section 3) with respect to w_i, i.e., the n most similar words to w_i are selected. Next, we present two semantic similarity metrics that utilize the notion of semantic neighborhood [Iosif and Potamianos (2013)].

4.1 Maximum Similarity of Neighborhoods

This metric is based on the hypothesis that the similarity of two words, w_i and w_j, can be estimated by *the maximum similarity of their respective sets of neighbors*, defined as follows:

$$M_n(w_i,w_j)=\max\{\alpha_{ij},\alpha_{ji}\}, \quad \text{where } \alpha_{ij}=\max_{x\,\in\,N_j} S(w_i,x),\ \ \alpha_{ji}=\max_{y\,\in\,N_i} S(w_j,y). \tag{1}$$

α_{ij} (or α_{ji}) denotes the maximum similarity between w_i (or w_j) and the neighbors of w_j (or w_i) that is computed according to a similarity metric S: in this work either D or Q^H. N_i and N_j are the set of neighbors for w_i and w_j, respectively. The definition of M_n is motivated by the maximum sense similarity assumption. Here the underlying assumption is that the most salient information in the neighbors of a word are semantic features denoting senses of this word.

4.2 Attributional Neighborhood Similarity

The similarity between w_i and w_j is defined as follows:

$$R_n(w_i, w_j) = \max\{\beta_{ij}, \beta_{ji}\}, \quad \text{where } \beta_{ij} = \rho(C_i^{N_i}, C_j^{N_i}), \;\; \beta_{ji} = \rho(C_i^{N_j}, C_j^{N_j}) \qquad (2)$$

$$\text{where } C_i^{N_i} = (S(w_i, x_1), S(w_i, x_2), \ldots, S(w_i, x_n)), \quad \text{and } N_i = \{x_1, x_2, \ldots, x_n\}.$$

Note that $C_j^{N_i}$, $C_i^{N_j}$, and $C_j^{N_j}$ are defined similarly as $C_i^{N_i}$. The ρ function stands for the Pearson's correlation coefficient, N_i is the set of neighbors of word w_i, and S is a similarity metric (D or Q^H). Here, we aim to exploit the entire semantic neighborhoods for the computation of semantic similarity, as opposed to M_n where a single neighbor is utilized. The motivation behind this metric is attributional similarity, i.e., we assume that semantic neighborhoods encode attributes (or features) of a word. Neighborhood correlation similarity in essence compares the distribution of semantic similarities of the two words on their semantic neighborhoods. The ρ function incorporates the covariation of the similarities of w_i and w_j with respect to the members of their semantic neighborhoods.

5 Experimental Procedure

Lexica and corpora creation: For English we used a lexicon consisting of $8,752$ English nouns taken from the SemCor3[1] corpus. In addition, this lexicon was translated into Greek using Google Translate[2], while it was further augmented resulting into a set of $9,324$ entries. For each noun an individual query was formulated and the $1,000$ top ranked results (document snippets) were retrieved using the Yahoo! search engine[3]. A corpus was created for each language by aggregating the snippets for all nouns of the lexicon.
Network creation: For each language the semantic neighborhoods of lexicon noun pairs were computed following the procedure described in Section 4 using either co-occurrence D or context-based $Q^{H=1}$ metrics [4].
Network-based similarity computation: For each language, the semantic similarity between noun pairs was computed applying either the max-sense M_n or the attributional R_n network-based metric. The underlying semantic similarity metric (the S metric in (1) and (2)) can be either D or Q^H. Given that for both neighborhood creation and network-based semantic similarity estimation we have the option of D or Q^H, a total of four combinations emerge for this two-phase process: (i) D/D, i.e., use co-occurence metric D for both neighborhood selection and network-based similarity estimation, (ii) D/Q^H, (iii) Q^H/D, and (iv) Q^H/Q^H.

6 Evaluation Datasets

The performance of network-based similarity metrics was evaluated for the task of semantic similarity between nouns. The Pearson's correlation coefficient was used as evaluation metric to compare estimated similarities against the ground truth (human ratings). The following datasets were used:
English (WS353): Subset of WS353 dataset [Finkelstein et al. (2002)] consisting of 272 noun pairs (that are also included in the SemCor3 corpus).
Greek (GIP): In total, 82 native speakers of modern Greek were asked to score the similarity of the noun pairs in a range from 0 (dissimilar) to 4 (similar). The resulting dataset consists of 99 nouns pairs (a subset of pairs translated from WS353) and is freely available [5].
Abstract vs. Concrete: From each of the above datasets two subsets of pairs were selected, where both nouns in the pair are either abstract or concrete, i.e., pairs consisting of one abstract and one concrete nouns were ruled out. More specifically, 74 abstract and 74 concrete noun pairs were selected from WS353, for a total of 148 pairs. Regarding GIP, 18 abstract and 18 concrete noun pairs were selected, for a total of 36 pairs.

[1] http://www.cse.unt.edu/~rada/downloads.html

[2] http://translate.google.com/

[3] http://www.yahoo.com//

[4] We have also experimented with other values of context window H not reported here for the sake of space. However, the highest performance was achieved for $H = 1$.

[5] http://www.telecom.tuc.gr/~iosife/downloads.html

7 Results

The performance of the two proposed network-based metrics, M_n and R_n, for neighborhood size of 100, is presented in Table 1 with respect to the English (WS353) and Greek (GIP) datasets. Baseline performance (i.e., no use of the network) is also shown for co-occurrence-based metric D and context-based metric Q^H. For the max-sense similarity $M_{n=100}$ metric, the use of the co-occurrence metric D for neighbor selection yields the best correlation performance for both languages. For the attributional similarity $R_{n=100}$ metric, best performance is achieved when using the context-based metric D for the selection of neighbors in the network. As explained in [Iosif and Potamianos (2013)], the neighborhoods selected by the D metrics tend to include words that denote word senses (yielding best results for similarity), while neighborhoods computed using the Q^H metric are semantically broader including word attributes (yielding best results for attributional similarity). The network-based DSM results are also significantly higher compared to the baseline for both languages. The best results achieved by D/Q^H for the $M_{n=100}$, and Q^H/D for the $R_{n=100}$ are consistent with the results reported in [Iosif and Potamianos (2013)] for English. The best performing metric for English is $M_{n=100}$ (max-sense) while for Greek $R_{n=100}$ (attributional). Overall, utilizing network neighborhoods for estimating semantic similarity can achieve good performance[6], and the type of metric (feature) used to select the neighborhood is a key performance factor.

Language:	Number of	Baseline		Network	Neighbor selection / Similarity computation			
dataset	pairs	D	Q^H	metric	D/D	D/Q^H	Q^H/D	Q^H/Q^H
English:	272	0.22	0.30	$M_{n=100}$	**0.64**	**0.64**	0.47	0.46
WS353				$R_{n=100}$	0.50	0.14	**0.56**	**0.57**
Greek:	99	0.25	0.13	$M_{n=100}$	**0.51**	**0.51**	0.04	0.04
GIP				$R_{n=100}$	-0.11	0.03	**0.66**	0.11

Table 1: Pearson correlation with human ratings for neighborhood-based metrics for English and Greek datasets. Four combinations of the co-occurrence-based metric D and the context-based metric Q^H were used for the definition of semantic neighborhoods and the computation of similarity scores. Baseline performance is also shown.

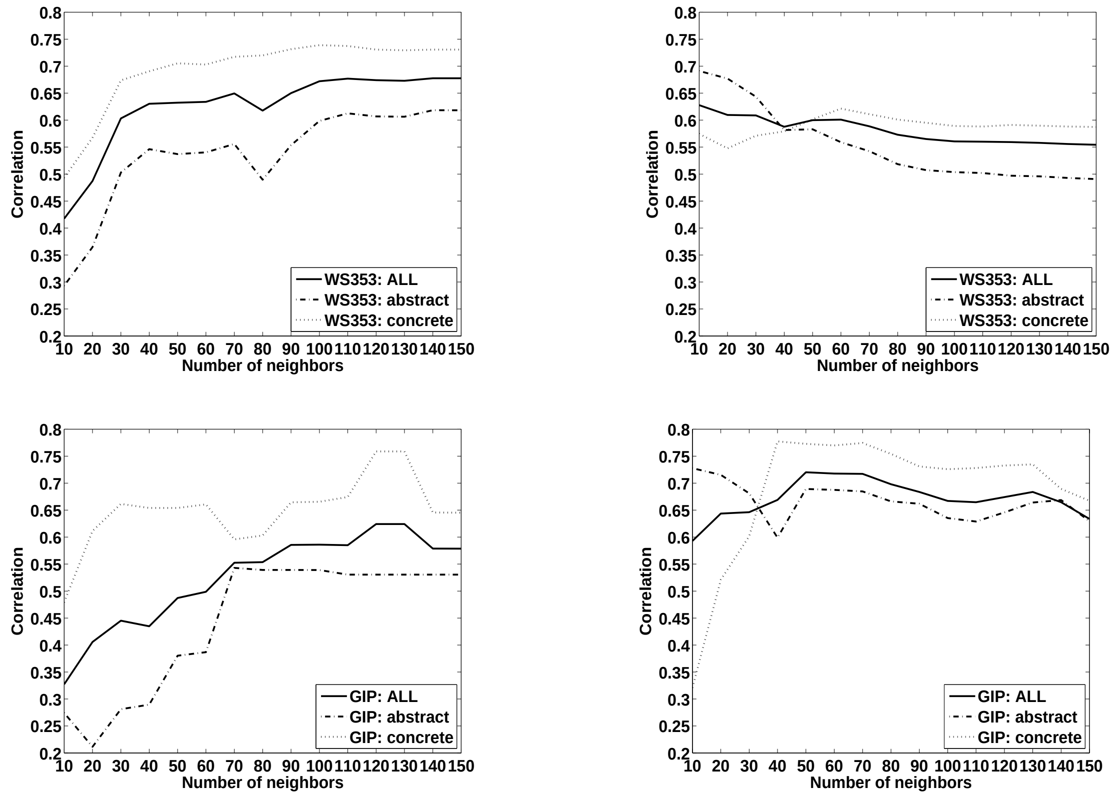
Next, we investigate the performance of the network metrics with respect to the neighborhood size n for the abstract and concrete noun pairs included in English and Greek datasets. The performance of the max-sense M_n (D/Q^H) metric is shown in Fig. 1(a),(c) for the (subsets of) WS353 and GIP, respectively. The performance over the whole (abstract and concrete) dataset is shown with a solid line. Similarly the results for the attributional R_n (Q^H/D) metric are shown in Fig. 1(b),(d). The main conclusions for these experiments (for both languages) are: 1) The correlation performance for concrete noun pairs is higher than for abstract noun pairs. 2) For concrete nouns the max-sense M_n metric achieves best performance, while for abstract nouns the attributional R_n metric is the top performer. 3) For the R_n network metric, very good performance is achieved for abstract noun pairs for a small neighborhood size n (around 10), while for concrete nouns larger neighborhoods are needed (up to 40 and 30 neighbors, for English and Greek, respectively).

Neighbor	Number of	Type of	Type of neighbors (abstract/concrete)			
			English (WS353)		Greek (GIP)	
selection metric	reference nouns	reference nouns	abstract	concrete	abstract	concrete
D	15	abstract	**76%**	24%	**82%**	18%
D	15	concrete	36%	**64%**	23%	**77%**
Q^H	15	abstract	**82%**	18%	**91%**	9%
Q^H	15	concrete	31%	**69%**	31%	**69%**

Table 2: Distribution of abstract vs. concrete nouns in (abstract/concrete noun) neighbourhoods.

In order to further investigate the network organization for abstract vs. concrete nouns, we manually inspected the top twenty neighbors of 30 randomly selected nouns (15 abstract and 15 concrete) and classified each neighbor as either abstract or concrete. The distributions of abstract/concrete neighbors are shown in Table 2 as a function of neighbor selection metric (D vs. Q^H) and reference noun category. It is clear, that the neighborhoods of abstract nouns contain mostly abstract concepts, especially for the Q^H neighbor selection metric (similarly the neighborhoods of concrete nouns contain mainly concrete concepts). The neighbors of concrete nouns mainly belong to the same semantic class (e.g., “vehicle”, “bus” for “car”) often corresponding to relevant senses. The

[6] The best correlation score for the WS353 dataset does not exceed the top performance (0.68) of unsupervised DSMs [Agirre et al. (2006)]. However, we have found that the proposed network metrics obtain state-of-the-art results for other standard datasets, e.g., 0.87 for [Rubenstein and Goodenough (1965)] and 0.91 for [Miller and Charles (1998)].



(a) (b)

(c) (d)

Figure 1: Correlation as a function of number of neighbors for network-based metrics. Max-sense M_n (D/Q^H) for datasets: (a) English and (c) Greek. Attributional R_n (Q^H/D) for datasets: (b) English and (d) Greek.

neighbors of the abstract nouns have an attributive function, reflecting relative attributes and/or aspects of the referent nouns (e.g., “religion”, “justice” for “morality”).

8 Discussion

We investigated the performance of network-based DSMs for semantic similarity estimation for abstract and concrete noun pairs of English and Greek. We observed a “concreteness effect”, i.e., performance for concrete nouns was better than for abstract noun pairs. The assumption of maximum sense similarity as encoded by the M_n metric consistently yielded higher performance for the case of concrete nouns, while the semantic *similarity of abstract nouns was better estimated via the attributional similarity assumption* as implemented by the R_n metric. The results are consistent with the initial hypothesis that differences in cognitive organization may warrant different network organization in DSMs. In addition, abstract concepts were best modeled using an attributional network DSM with small semantic neighborhoods. This is a first step towards the better understanding of the network organization of DSMs for different categories of concepts. In terms of computation algorithms of semantic similarity, it might prove advantageous to define a metric that combines the maximum sense and attributional assumptions based on the semantic concreteness of the words under investigation. Further research on more data and languages is needed to verify the universality of the findings.

9 Acknowledgements

Elias Iosif, Alexandros Potamianos, and Maria Giannoudaki were partially funded by the PortDial project (“Language Resources for Portable Multilingual Spoken Dialogue Systems”) supported by the EU Seventh Framework Programme (FP7), grant number 296170.

References

Agirre, E., E. Alfonseca, K. Hall, J. Kravalova, M. Pasca, and A. Soroa (2009). A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proc. of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 19–27.

Agirre, E., D. Martínez, O. L. de Lacalle, and A. Soroa (2006). Two graph-based algorithms for state-of-the-art WSD. In *Proc. of Conference on Empirical Methods in Natural Language Processing*, pp. 585–593.

Baroni, M. and A. Lenci (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics 36*(4), 673–721.

Coltheart, M. (2000). Deep dyslexia and right-hemisphere reading. *Brain and Language 71*, 299–309.

Finkelstein, L., E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems 20*(1), 116–131.

Harris, Z. (1954). Distributional structure. *Word 10*(23), 146–162.

Iosif, E. and A. Potamianos (2010). Unsupervised semantic similarity computation between terms using web documents. *IEEE Transactions on Knowledge and Data Engineering 22*(11), 1637–1647.

Iosif, E. and A. Potamianos (2013). Similarity Computation Using Semantic Networks Created From Web-Harvested Data. *Natural Language Engineering (submitted)*.

Kiehl, K. A., P. F. Liddle, A. M. Smith, A. Mendrek, B. B. Forster, and R. D. Hare (1999). Neural pathways involved in the processing of concrete and abstract nouns. *Human Brain Mapping 7*, 225–233.

Malandrakis, N., A. Potamianos, E. Iosif, and S. Narayanan (2011). Kernel models for affective lexicon creation. In *Proc. Interspeech*, pp. 2977–2980.

Meng, H. and K.-C. Siu (2002). Semi-automatic acquisition of semantic structures for understanding domain-specific natural language queries. *IEEE Transactions on Knowledge and Data Engineering 14*(1), 172–181.

Mihalcea, R. and D. Radev (2011). *Graph-Based Natural Language Processing and Information Retrieval*. Cambridge University Press.

Miller, G. (1990). Wordnet: An on-line lexical database. *International Journal of Lexicography 3*(4), 235–312.

Miller, G. and W. Charles (1998). Contextual correlates of semantic similarity. *Language and Cognitive Processes 6*(1), 1–28.

Noppeney, U. and C. J. Price (2004). Retrieval of abstract semantics. *NeuroImage 22*, 164–170.

Pado, S. and M. Lapata (2007). Dependency-based construction of semantic space models. *Computational Linguistics 33*(2), 161–199.

Paivio, A. (1971). *Imagery and Verbal Processes*. New York, Holt, Rinehart and Winston.

Papagno, C., R. Capasso, and G. Miceli (2009). Reversed concreteness effect for nouns in a subject with semantic dementia. *Neuropsychologia 47*(4), 1138–1148.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxanomy. In *Proc. of International Joint Conference for Artificial Intelligence*, pp. 448–453.

Rubenstein, H. and J. B. Goodenough (1965). Contextual correlates of synonymy. *Communications of the ACM 8*(10), 627–633.

Turney, P. (2006). Similarity of semantic relations. *Computational Linguistics 32*(3), 379–416.

Véronis, J. (2004). Hyperlex: Lexical cartography for information retrieval. *Computer Speech and Language 18*(3), 223–252.

Widdows, D. and B. Dorow (2002). A graph model for unsupervised lexical acquisition. In *Proc. of the 19th International Conference on Computational Linguistics*, pp. 1093–1099.