Word Embedding Distance Does not Predict Word Reading Time

Stefan L. Frank (s.frank@let.ru.nl)
Centre for Language Studies, Radboud University
P.O. Box 9103, 6500 HD Nijmegen, The Netherlands

Abstract

It has been claimed that larger semantic distance between the words of a sentence, as quantified by a distributional semantics model, increases both N400 size and word-reading time. The current study shows that the reading-time effect disappears when word surprisal is factored out, suggesting that the earlier findings were caused by a confound between semantic distance and surprisal. This absence of a behavioural effect of semantic distance (in the presence of a strong neurophysiological effect) may be due to methodological differences between eye-tracking and EEG experiments, but it can also be interpreted as evidence that eye movements are optimized for reading efficiency.

Keywords: reading; eye tracking; N400; distributional semantics; semantic distance; word surprisal

Introduction

An open question in the study of human language processing is to what extent mere semantic similarity among words within a sentence or text affects the comprehension process. Results from controlled experiments are inconclusive. On the one hand, there is ample evidence for effects on the N400 event-related brain potential (ERP) component: Reading a word that is semantically related to words in the preceding context decreases N400 size, relative to when the context words are not meaning related (Camblin, Gordon, & Swaab, 2007; Metusalem et al., 2012; Paczynski & Kuperberg, 2012). A number of behavioural experiments, however, failed to find corresponding effects on word-reading time (Gordon, Hendrick, Johnson, & Lee, 2006; Traxler, Foss, Seely, Kaup, & Morris, 2000). In contrast, two studies that analysed reading times on naturalistic texts (instead of taking a controlled experimental approach) did find that words are read faster when they have stronger semantic relatedness to earlier words in the text (Mitchell, Lapata, Demberg, & Keller, 2010; Pynte, New, & Kennedy, 2008). In those studies, semantic relatedness measures were obtained from a distributional semantics model, which assigns numerical vectors to words on the basis of the words' co-occurrence patterns in large text corpora. These vector representations are known as *word embeddings* in the computational linguistics literature. Words that tends to occur in similar contexts receive similar embeddings. Consequently, distances between the words' embedding vectors correspond to semantic distances between the corresponding words.

If semantically related words tend to co-occur, a word's occurrence can (to some extent) be predicted from the presence of related words. Consequently, if one wants to claim that the reading process on word w_t is affected by the word's semantic relatedness to the preceding words $(w_1, \ldots, w_{t-1})$, it is crucial to factor out any effect of the *predictability* of w_t from its previous context. Otherwise, apparent effects of relatedness could in fact be due to word predictability instead. Frank and Willems (in press) recently showed that N400 effects of semantic distance (as quantified by a distributional semantics model) remain when factoring out the words' (un)predictability as quantified by their surprisal (i.e., $-\log P(w_t|w_1, \ldots, w_{t-1})$), leaving no room for a confound between predictability and semantic distance. The current paper will show that the same is not true for reading times: Effects of semantic similarity on reading times for naturalistic materials, of the type reported by Mitchell et al. (2010) and Pynte et al. (2008), disappear when surprisal is factored out, provided that surprisal is computed by a powerful enough language model. Hence, semantic similarity between the words of a sentence or text affects N400 size but not reading time.

Method

Eye-tracking Data

Word-reading times were extracted from two published sets of eye-tracking data: The UCL corpus (Frank, Monsalve, Thompson, & Vigliocco, 2013) and the English Dundee corpus (Kennedy & Pynte, 2005). The UCL corpus comprises data from 42 native English speakers reading 205 individual sentences sampled from three unpublished novels; the Dundee corpus has 10 participants reading newspaper editorials. Frank and Willems (in press) demonstrated strong N400 effects of semantic distance (over and above the effect of surprisal) for the sentences of the UCL corpus. Mitchell et al. (2010) reported reading-time effects of semantic distance in the Dundee data, and similar results by Pynte et al. (2008) were based on the French part of the Dundee corpus, also comprising newspaper texts.

Four measures of reading time will be investigated: first-fixation duration, first-pass duration (the sum of fixation durations on a word before the first fixation on any other word), right-bounded reading time (the sum of fixation durations on a word before the first fixation on a later word), and go-past reading time (the sum of fixations on *all* words from the first fixation on the current word until the first fixation on a later word). These four measures, in this order, have been argued to reflect increasingly late cognitive processes (Clifton Jr., Staub, & Rayner, 2007; Gordon et al., 2006).

Models

Each content word of the UCL and Dundee corpora was assigned a measure of semantic distance to preceding content words, as well as five estimates of word surprisal. The distributional semantics and surprisal models were trained on

the first slice of the ENCOW14 web corpus (Schäfer, 2015), comprising 644.5M word tokens of 2.81M types.

Semantic Distance Word embeddings were generated by the word2vec skipgram model (Mikolov, Chen, Corrado, & Dean, 2013), which is basically a feedforward neural network with one hidden layer. The network learns to associate each input word w_t to the k words immediately preceding and following (i.e., the sequence $w_{t-k},\ldots,w_{t-1},w_{t+1},\ldots,w_{t+k}$). After training the network, the vector of connection weights from each input unit to the 300-unit hidden layer forms the embedding for the word corresponding to the input unit. The 'window size' parameter was set to $k = 5$ in the current application of the model.

As explained in the Introduction, the distance between two word vectors quantifies the semantic distance between the two words. A common distance measure used in distributional semantics is the cosine of the angle between the vectors. Here, we require a measure for the distance between the current word's embedding $\vec{w}_t$ and its entire previous context (not just a single word). The vector representing the combination of content words from the previous context is defined as simply the sum of the words' individual vectors. Thus, the relevant distance measure becomes

$$\text{semdist}(t) = -\cos\left(\vec{w}_t, \sum_{w\in A_t}\vec{w}\right), \qquad (1)$$

where A_t is a collection of content words that precede w_t in the sentence or text. For the individual sentences of the UCL corpus, A_t contains all content words preceding w_t in the sentence. For the full texts of the Dundee corpus, A_t contains the four content words immediately preceding w_t in the text (if w_t is among the text's first four content words, A_t will contain correspondingly fewer words). If A_t is empty, word w_t has no semantic distance. Semantic distance values on the UCL corpus were identical to those used by Frank and Willems (in press) to analyse N400 ERP effects.

Surprisal Word surprisal was computed by n-gram language models, which simplify the full conditional probability $P(w_t|w_1,\ldots,w_{t-1})$ to $P(w_t|w_{t-n+1},\ldots,w_{t-1})$, that is, only the $n-1$ previous words are taken into account when estimating the occurrence probability of w_t. Model order n was varied from $n = 2$ to $n = 5$, and the model was generated by SRILM (Stolcke, 2002) with modified Kneser-Ney smoothing (Chen & Goodman, 1999).

The semantic distance measure defined above is sensitive to content words beyond the $n-1$ previous words that matter to an n-gram model. If semantic distance correlates with surprisal, this could yield apparent effects of semantic distance that are in fact due to unpredictability resulting from words outside of the n-gram window. To control for this, a 'skip-bigram' language model (SBLM) was used to obtain a fifth set of surprisal values:

$$P_{\text{sblm}}(w_t|A_t) = \frac{1}{|A_t|}\sum_{w_i\in A_t} P(w_t|w_i) = \frac{1}{|A_t|}\sum_{w_i\in A_t}\frac{P(w_i,w_t)}{P(w_i)},$$

with A_t as defined as in Equation 1 and $|A_t|$ the number of words in A_t. $P(w_t|w_i)$ denotes the probability that w_t occurs within a distance of 15 words after occurrence of w_i. That is, the preceding content words $w_i \in A_t$ are taken as independent cues to the occurrence of w_t, whose skip-bigram probability is computed by averaging over these individual cues.

The required word-pair probabilities $P(w_i,w_t)$ are estimated from co-occurrence frequencies in the training corpus, using the Simple Good-Turing smoothing method (Gale & Sampson, 1995) to estimate the total probability of all unseen pairs. This total probability P_0 is divided over the unseen pairs (v,w) in proportion to $P(v)P(w)$, that is, the probability of each particular unseen pair (v,w) is given by:

$$P(v,w) = \frac{P_0 P(v)P(w)}{1-\sum_{(v',w')\in S}P(v')P(w')},$$

where S is the set of all ordered word pairs observed in the training data within a 15-word distance from each other.

Relation between semantic distance and surprisal Table 1 shows there indeed exists a positive confound between surprisal and semantic distance, which grows stronger as the language model is able to use words from further back in the context.

Frank and Willems (in press) interpolate the 5-gram and skip-bigram models to minimize average surprisal over the UCL corpus and show empirically that the semantic distances do not contain information that can be used to further improve this interpolated language model. Hence, if the semantic distances account for variance in human reading difficulty measures over and above what is already explained by the surprisal values, this cannot be attributed to a confound between semantic relatedness and predictability but must be due to the effect of semantic relatedness itself.

Data Analysis

Linear mixed-effects regression models were fitted to the log-transformed reading times using as covariates: word position in the sentence, word length (number of characters), word log-frequency in ENCOW14, and a binary factor indicating whether or not the previous word was fixated. To account for the possibility that reading-time effects appear shortly after the point at which they originate (so-called spillover effects), the previous word's length and log-frequency were also included. All two-way interactions between these six factors were also present.

Table 1: Correlation coefficients between semantic distance and surprisal values.

	Language model				
Data set	2-gram	3-gram	4-gram	5-gram	SBLM
UCL	.19	.26	.27	.27	.29
Dundee	.05	.18	.20	.21	.26

The main factor of interest was the word's semantic distance measure. Separate analyses were run using the current and previous word's semantic distance; the latter capturing potential spillover.[1] In addition to a control condition without any surprisal measure in the regression, five separate analyses were run including n-gram surprisal with $n = 2,3,4,5$, or both 5-gram and SBLM surprisal (always for both the current and previous word).

Random effects in the regression model were the by-subject and by-word intercept, and by-subject slopes of semantic distance and any surprisal measure that was included as a fixed effect.

Regression models were fitted to each of the four reading time measures from both data sets, making a total of 96 analyses: 4 reading time measures $\times$ 2 corpora $\times$ 2 semantic distance measures (of current or previous word) $\times$ (5 surprisal measures + 1 control). Words were excluded from analysis if they were not fixated, were attached to punctuation, contained any non-letter or more than one capital letter, or were the first or last word on a line.

Results

Figure 1 displays the estimated regression coefficient (i.e., effect size) of the semantic distance predictor in each of the 96 fitted regression models. Note that effect sizes cannot be compared between the analyses investigating the current versus previous word's semantic distance. This is because these analyses apply to different sets of words: All content words when the current word's semantic distance is used, but the words directly following content words (including many function words) when the previous word's semantic distance is the variable under investigation. The same holds for the estimated regression coefficients of the surprisal predictors, plotted in Figure 2.[2]

For the UCL corpus, none of the semantic distance effects reach statistical significance. For the Dundee corpus, there is a clear effect of semantic distance in the expected (i.e., positive) direction when surprisal is not factored out, and it remains present for later reading time measures when surprisal takes only very local context into account (i.e., under a bigram model).

As is clear from Figure 2, words with higher surprisal take longer to read, as is well known from the literature (e.g. Monsalve, Frank, & Vigliocco, 2012; Smith & Levy, 2013). Surprisal computed by the novel SBLM language model has an effect over and above 5-gram surprisal, at least for the Dundee corpus, which means that it is not merely the local, 4-word context that is is taken into account when generating expectations about upcoming words. Rather, long-distance co-occurrence patterns between content words matter as well.

There are a few noticeable difference between the results for the UCL and Dundee data sets, which mirror differences in the text materials of these two corpora. Surprisal effects appear to be more reliable in the Dundee data, in that the zero point falls further outside the confidence intervals. This can simply be explained by the Dundee data set being much larger than the UCL data set (134,203 versus 18,178 data points). Interestingly, the UCL corpus results show larger effect sizes (i.e., larger coefficients) which is probably due to these materials having been specifically designed for language model evaluation. Compared to the Dundee corpus texts, the UCL corpus sentences contain fewer low-frequency words (for which surprisal is hard to estimate reliably) and can comprehended more easily without relying on world knowledge (which the language models do not incorporate). Finally, the fact that the SBLM model explained unique variance in reading times from the Dundee corpus only can be explained by the fact that this corpus consists of full texts as opposed to the UCL corpus's individual sentences. Compared to individual sentences, full texts will contain more content words outside of the 5-gram window, making the SBLM model more influential.

Discussion

Results on the Dundee corpus showed significant, positive effects of semantic distance on all four reading time measures when surprisal was not taken into account. However, factoring out surprisal as computed by anything more powerful than a bigram model made the effects of semantic distance disappear. Apparently, these effects were due to a confound between semantic distance and surprisal, that is, a word is less likely to appear if it has weaker semantic relatedness to earlier words. It was actually a word's unpredictability, rather than its semantic content per se, that resulted in increased reading time.

Indeed, the findings by Pynte et al. (2008) and Mitchell et al. (2010), on the French and English Dundee corpus, respectively, can be attributed to confounds between semantic relatedness and predictability. Pynte et al. (2008) did not factor out surprisal (or even simple transitional probabilities between words) in their analysis of the effect of semantic distance. Mitchell and Lapata's (2009) goal was to show that incorporating semantic distance measures from their own 'simple semantic space model' (as well as from a Latent Dirichlet Allocation Topics model; Griffiths, Steyvers, & Tenenbaum, 2007) reduces perplexity of a combined n-gram and probabilistic phrase-structure grammar. That is, taking these semantic measures into account improves the language model. Consequently, the improved fit to reading time could be due merely to more accurate next-word prediction rather than to semantic similarity per se.

The UCL corpus results showed no effect of semantic dis-

[1] If both the current and previous word's semantic distance had been included as factors in a single regression model, this would have greatly reduced the amount of usable data because both adjacent words would have to be content words.

[2] The displayed coefficients for current (previous) surprisal come from the regression model that includes current (previous) semantic distance. Consequently, exactly the same set of words was involved in estimating the coefficients for the surprisal and semantic distance measures, even though surprisal (unlike semantic distance) is also defined for function words.

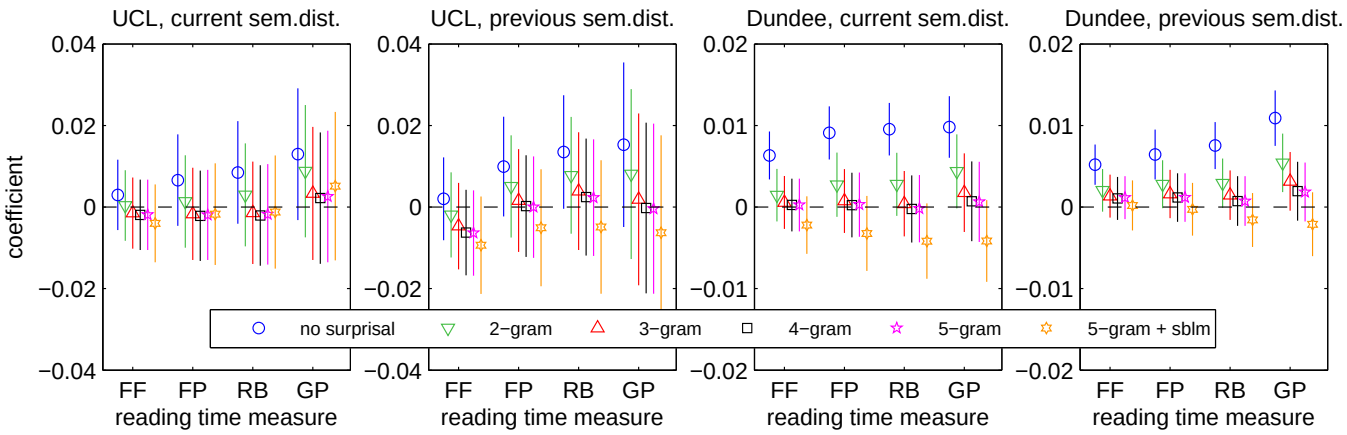

Figure 1: Regression coefficients (with 95% confidence intervals) of semantic distance predictor, when factoring out different measures of surprisal. The leftmost two panels display results on the UCL corpus; the Dundee corpus results are shown in the rightmost panels. The 2nd and 4th panel show the coefficient of the previous word's semantic distance. Reading time measures are indicated by FF (first fixation), FP (first pass), RB (right-bounded), and GP (go-past).

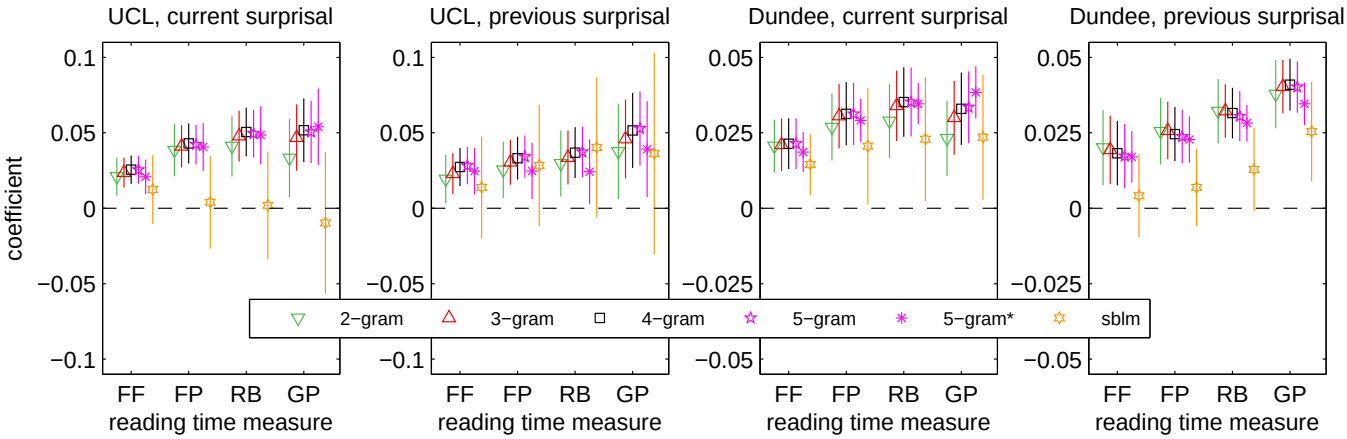

Figure 2: Regression coefficients (with 95% confidence intervals) of surprisal predictor, when factoring out semantic distance. "5-gram*" refers to the effect of 5-gram surprisal when SBLM-surprisal is also included as a regressor, and "sblm" refers to the effect of SBLM-surprisal over and above 5-gram surprisal. The leftmost two panels display results on the UCL corpus; the Dundee corpus results are shown in the rightmost panels. The 2nd and 4th panel show the coefficient of the previous word's semantic distance. Reading time measures are indicated by FF (first fixation), FP (first pass), RB (right-bounded), and GP (go-past).

tance on reading times whatsoever, even when surprisal was not taken into account. This is remarkable considering that Frank and Willems (in press) found that N400 effects of the very same semantic distance values are of similar size as – and independent from – the effect of surprisal as computed by an interpolated 5-gram and skip-bigram language model. This discrepancy between neurophysiological and behavioral effects is consistent with findings from the controlled experimental studies mentioned in the Introduction. But how can it be explained?

One possible cause is the difference in stimuli presentation method. The eye-tracking methodology allows a natural reading processes whereas in most EEG reading studies, words are presented one at a time for an unnaturally long duration. The EEG data used by Frank and Willems (in press) came from a study with word-length dependent presentations durations of at least 627ms (Frank, Otten, Galli, & Vigliocco, 2015), which is much longer than fixation durations in natural reading. Wlotko and Federmeier (2015) showed that using more natural word presentation rates in an ERP reading study can remove particular effects of semantic relatedness on the N400. If semantic distance effects are delayed relative to surprisal effects, this could explain their absence in reading times: By the time they would have appeared, any effect has already been washed out by the processing of several other words. Although Figure 1 indeed shows a trend for the semantic distance effect to be somewhat stronger for the later reading time measures (as was also found by Pynte et al., 2008), the same is true for the surprisal effect (Figure 2) so this cannot explain why reading times are insensitive to semantic distance. Moreover, Frank and Willems (in press) found fMRI effects of semantic distance (as quantified by distributional semantics) during normal speech comprehension, indicating that the presence of a measurable neural response does not rely on unnaturally slow presentation rates.

An alternative, and possibly more interesting explanation of the difference between N400 and reading time effects is that reading is optimized for speed (Smith & Levy, 2013). Being faster on more predictable (i.e., lower surprisal) words increases overall efficiency, whereas there is no reason to be faster on merely semantically related words. Hence, we would expect reading times to display effects of surprisal but not of semantic distance. Other dependent variables from eye-tracking, however, could show sensitivity to semantic distance, and this is exactly what Van den Hoven, Hartung, Burke, and Willems (2016) found in a recent analysis of data from a Dutch narrative text reading eye-tracking study: Semantic distance correlated with saccade distance and regression probability but not with reading time after factoring out trigram surprisal. In contrast, the reason why the N400 shows effects of both surprisal and semantic distance could be that it forms an index of the difficulty of retrieving lexical information from long-term memory (Brouwer, Fitz, & Hoeks, 2012; Kutas & Federmeier, 2000). As Frank and Willems (in press) argue, this difficulty is reduced both by probabilistic word prediction (surprisal) and by semantic similarity to earlier words (word embedding distance).

Conclusion

The current results failed to replicate earlier findings of a positive correlation between reading times on naturalistic data and semantic relatedness between words, as quantified by a distributional semantics model. This apparent effect of semantic relatedness appeared to be due to a confound with word predictability. Of course, it is possible that an effect of semantic distance reappears when using a different distributional semantics model, or a more sophisticated technique for combining single word vectors into a sentence context vector (Equation 1). However, it is equally true that improved surprisal models may undo the work of more sophisticated word embedding models. And crucially, the current distributional semantics modelling choices were appropriate for predicting reading times when surprisal was not taken into account, as well as N400 sizes over and above surprisal, so they should also have sufficed for revealing reading time effects of semantic distance that are independent from surprisal, if there had been any.

Acknowledgements

The work presented here was funded by the Netherlands Organisation for Scientific Research (NWO) Gravitation Grant 024.001.006 to the Language in Interaction Consortium.

References

Brouwer, H., Fitz, H., & Hoeks, J. C. J. (2012). Getting real about semantic illusions: Rethinking the functional role of the P600 in language comprehension. *Brain Research*, *1446*, 127–143.

Camblin, C. C., Gordon, P. C., & Swaab, T. Y. (2007). The interplay of discourse congruence and lexical association during sentence processing: Evidence from ERPs and eye tracking. *Journal of Memory and Language*, *56*, 103–128.

Chen, S. F., & Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, *13*, 359–394.

Clifton Jr., C., Staub, A., & Rayner, K. (2007). Eye movements in reading words and sentences. In R. Van Gompel, M. Fisher, W. Murray, & R. L. Hill (Eds.), *Eye movement research: A window on mind and brain* (pp. 341–372). Elsevier, Oxford, UK.

Frank, S. L., Monsalve, I., Thompson, R. L., & Vigliocco, G. (2013). Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods*, *45*, 1182–1190.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11.

Frank, S. L., & Willems, R. M. (in press). Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Language, Cognition and Neuroscience*.

Gale, W. A., & Sampson, G. (1995). Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics*, *2*, 217–237.

Gordon, P. C., Hendrick, R., Johnson, M., & Lee, Y. (2006). Similarity-based interference during language comprehension: Evidence from eye tracking during reading. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *32*, 1304–1321.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*, 211–244.

Kennedy, A., & Pynte, J. (2005). Parafoveal-on-foveal effects in normal reading. *Vision Research*, *45*, 153–168.

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, *4*, 463–470.

Metusalem, R., Kutas, M., Urbach, T. P., Hare, M., McRae, K., & Elman, J. L. (2012). Generalized event knowledge activation during online sentence comprehension. *Journal of Memory and Language*, *66*, 545–567.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of the ICLR Workshop.*

Mitchell, J., & Lapata, M. (2009). Language models based on semantic composition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (pp. 430–439). Singapore: Association for Computational Linguistics.

Mitchell, J., Lapata, M., Demberg, V., & Keller, F. (2010, July). Syntactic and semantic factors in processing difficulty: An integrated measure. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 196–206). Uppsala, Sweden: Association for Computational Linguistics.

Monsalve, I. F., Frank, S. L., & Vigliocco, G. (2012). Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 398–408). Avignon, France: Association for Computational Linguistics.

Paczynski, M., & Kuperberg, G. R. (2012). Multiple influences of semantic memory on sentence processing: distinct effects of semantic relatedness on violations of real-world event/state knowledge and animacy selection restrictions. *Journal of Memory and Language*, *67*, 426–448.

Pynte, J., New, B., & Kennedy, A. (2008). On-line contextual influences during reading normal text: A multiple-regression analysis. *Vision Research*, *48*, 2172-2183.

Schäfer, R. (2015). Processing and querying large web corpora with the COW14 architecture. In P. Bański, H. Biber, E. Breiteneder, M. Kupietz, H. Lüngen, & A. Witt (Eds.), *Proceedings of the 3rd Workshop on the Challenges in the Management of Large Corpora* (pp. 28–34). Mannheim, Germany: Institut für Deutsche Sprache.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*, 302–319.

Stolcke, A. (2002). SRILM – an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing* (pp. 901–904). Denver, Colorado.

Traxler, M. J., Foss, D. J., Seely, R. E., Kaup, B., & Morris, R. K. (2000). Priming in sentence processing: intralexical spreading activation, schemas, and situation models. *Journal of Psycholinguistic Research*, *29*, 581–595.

Van den Hoven, E., Hartung, F., Burke, M., & Willems, R. M. (2016). Individual differences in sensitivity to style during literary reading: Insights from eye-tracking. *Collabra*, *2*, 25.

Wlotko, E. W., & Federmeier, K. D. (2015). Time for prediction? the effect of presentation rate on predictive sentence comprehension during word-by-word reading. *Cortex*, *68*, 20-32.