The Effect of Semantic Similarity is a Function of Contextual Constraint

Hongoak Yun (hyun3@buffalo.edu)
Gail Mauner (mauner@buffalo.edu)
Department of Psychology, 204 Park Hall
Buffalo, NY14260 USA

Douglas Roland (droland@buffalo.edu)
Jean-Pierre Koenig (jpkoenig@buffalo.edu)
Department of Linguistics, 609 Baldy Hall
Buffalo, NY 14260 USA

Abstract

We investigate how the degree to which a context constrains the words that could occur in a sentence affects the processing of the word that does occur. Roland et al. (2012) found that processing was facilitated when target words were more semantically similar to word alternatives that could have appeared. Because this effect is independent of word predictability, it suggests that comprehenders may have separate expectations for words and more general semantic features. We show that the semantic similarity effect is modulated by the degree of contextual constraint. We found that facilitation due to semantic similarity was greater when contexts were less constraining, and lower when contexts were more constraining, independent of word predictability. We interpret these results as suggesting that in highly constraining contexts, comprehenders may expect specific words, and face difficulties when these expectations are violated, while in less constraining contexts, they may have more general expectations for semantic properties shared between the words that could occur.

Keywords: sentence processing; semantic similarity; predictability; entropy; contextual constraint; expectation-based language comprehension

Introduction

In expectation-based models of sentence comprehension, contextual information has an enormous effect on how words are integrated into sentences. These models predict that the degree of difficulty a reader encounters in integrating a new word into a sentence is either entirely or in large measure a function of how predictable that word is given prior context (e.g., Levy, 2008). Presumably this is because predicted words are activated by context in advance of when they are encountered, making them easier to retrieve from memory or because predictable words are easier to integrate into the representations being constructed during comprehension. The effect of predictability on processing time has been observed in many studies (e.g., Bicknell, Elman, Hare, McRae, & Kutas, 2010; Ehrlich & Rayner, 1981; Frisson, Rayner, & Pickering, 2005; Staub, 2011; DeLong, Urbach, & Kutas, 2005; Federmeier, Wlotko, De Ochoa-Dewald, & Kutas, 2007; Otten & Van Berkum, 2008; Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005).

The relationship between a word's predictability and the amount of effort required to process it has been formalized in a number of computational models of language processing known as surprisal models (e.g., Boston, Hale, Patil, Kliegl, & Vasishth, 2008; Hale, 2001; Levy, 2008; Padó, Crocker, & Keller, 2009). In these models, the amount of cognitive effort required to integrate a word into a sentence depends on the negative log probability of that word given its preceding context. Surprisal models have had considerable success in predicting differences in reading times based on a word's predictability given its preceding context.

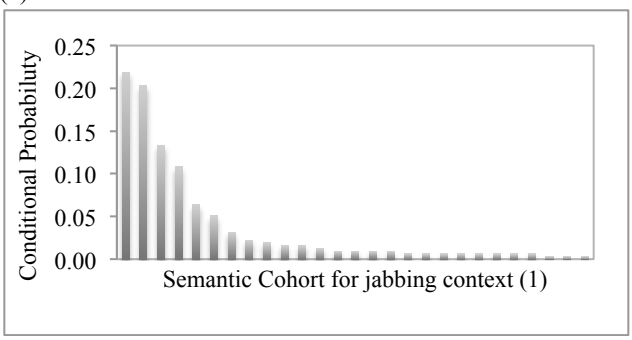
As it turns out, the amount of processing effort associated with integrating a word into a sentence cannot be entirely reduced to its predictability given its preceding context. Roland, Yun, Koenig, and Mauner (2012) examined the effects of the semantic cohort of a target word (i.e., the other words that could appear in the same position/context as the target word) on the processing of a target word. They found that words that are more semantically similar to their semantic cohort are easier to process when word predictability and other factors are controlled for. This result is important because it points to a limitation in expectation-based computational accounts of sentence processing that claim that a word's probability given its context is the sole predictor of processing effort (e.g., Levy, 2008). The results we present further constrain expectation-based accounts of sentence processing by showing that the effect of semantic similarity between a target word and its semantic cohort interacts with the degree of constraint provided by context.

The findings of Schwanenfluegel and LaCount (1988) motivate the possibility that the benefits of semantic similarity on word integration might be modulated by contextual constraint. Schwanenflugel and LaCount found that unpredictable words that were semantically related to the most predictable word that could occur in the same position were processed faster than other equally unpredictable words that were also semantically unrelated to that most predictable word. What is crucial for this discussion is that the benefit of shared semantic information was not consistently observed for all unpredictable words. Shared semantic information only facilitated the processing of unpredictable words when contexts were weakly constraining.

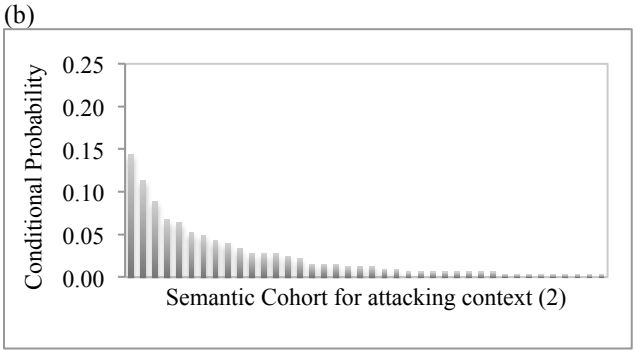
To illustrate why the benefit of shared semantic similarity to other words activated by the preceding context would be greatest when a target word is unpredictable and its context is only weakly constraining, consider the sentence contexts in examples (1) and (2), for which we have obtained word completions (this study will be described in greater detail later).

(1) The gladiator jabbed the African tiger with
(2) The aborigine attacked the angry lion with

(a)



(b)



Figures 1a-b: Probability distributions for semantic cohorts for Examples (1) and (2).

In both contexts, an instrument noun is most likely to be the next word. However, the types of instruments in the semantic cohort differed across contexts. For context (1), instruments like *sword, spear, stick, knife*, and *spike* were mentioned. These instruments share a typical property, i.e., all can be used as "pokers". In contrast, instruments like *sword, spear, knife, stick*, *fire, net, whip*, and *rock*, which were mentioned for context (2), have few salient characteristics that are common to instruments of attacking. This difference in the degree of shared characteristics suggests that context (1) places greater restrictions on the range of possible instruments than context (2). Using responses obtained from a completion study, we illustrate the distribution pattern of the probabilities of possible instruments for each context. In comparing Figure 1a to Figure 1b, two things become apparent. First, the most probable instruments for jabbing are more likely than the most probable instruments for attacking. Second, the probabilities of the jabbing semantic cohort drop more sharply than do the probabilities of the attacking semantic cohort. One way of quantifying the greater degree of constraint provided by the jabbing context is to note that the top three items have a combined probability of .55. In the attacking context, even the first 6 items do not match that combined probability.

Hypotheses and Prediction

Based on the findings of Schwanenflugel and LaCount (1988), we predict that the semantic similarity effect found by Roland et al. (2012) will be stronger in more weakly constraining contexts and weaker in highly constraining contexts. While Schwanenflugel and LaCount only examined the processing of unpredictable words, we do not expect interactions with word predictability, since Roland et al. found no interaction between similarity and predictability.

Entropy as a measure of contextual constraint

In order to measure the effects of contextual constraint, we need a measure to quantify the degree of contextual constraint. Recall that in a more constraining context, there are larger differences in the probabilities of cohort members, because a small subset of the possible words is more likely, while the others are unexpected. Alternatively, in a less constraining context, there are a larger number of words that are more or less equally likely. We will use Entropy (H), a standard measure from information theory shown in Equation 1, to reflect these differences in the distributions of the probabilities the cohort members. Entropy is higher when the choices are more similar in probability, as in low constraint contexts, and is lower when choices are less similar in probability, as in more highly constraining contexts.

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) \qquad \text{Equation 1}$$

Experiment to Generate Reading Times

Participants One hundred thirty native English-speaking undergraduates from the University at Buffalo received partial course credit for participation.

Materials We constructed 3 sets of 60 active declarative sentences with optional prepositional phrases similar to those in Example (3). Sets were differentiated by having an instrument noun that was highly likely (e.g., *sword*), moderately likely (e.g., *spear*), or unlikely (e.g., *spike*). To avoid wrap-up effects on instrument reading times (Just & Carpenter, 1980), all sentences included sentence-final phrases like *in the Colosseum*. Presentation regions are indicated in example (3) by vertical lines (|).

(3) The gladiator |jabbed |the African tiger |with |a sword/spear/spike |in |the Colosseum.

Selection of target instruments was based on responses from a listing study in which 42 participants produced five instruments for sentence fragments like (1) and (2). Cloze probabilities for highly likely, moderately likely and unlikely instruments were M = .23, $S.D.$ = .06, M = .10, $S.D.$ = .04 and M = .02, $S.D.$ = .01, respectively. Results of plausibility rating revealed that all instruments were plausible. Co-occurrence frequencies between target verbs and instrument prepositional phrases, counted using the British National Corpus (BNC) (Burnard, 1995), were very low (M = .03, $S.D.$= .03), and separate modeling showed that the frequency with which each verb occurred with an instrument phrase played no role in our results.

Experimental sentences were counterbalanced across six presentation lists, each consisting of 10 experimental sentences for each level of predictability. To obscure systematicities, these sentences were intermixed with 90 distractor sentences with varied syntactic structures (e.g., subordinate clause, adverbial phrase, or relative-clause sentences) and prepositional phrases with different prepositions (e.g., *on*, *in*, or *from*). Finally, because participants judged whether each sentence made sense, 33% of the total number of trials were designed not to make sense.

Procedure Participant-paced, region-by-region reading was accompanied by a secondary make-sense judgment task. This task was used to increase sensitivity to subtle semantic effects that might not be observed in a straight reading paradigm. Trials were divided into two blocks with a two-minute break between blocks to lessen fatigue.

Dependent Variables While the primary dependent variable was the reading times for sentences that participants continued to judge acceptable, we examined "No" judgments to ensure that they did not differ as a function of instrument likelihood. Across conditions, percentages of "No" responses adjusted for remaining chances to say "No" (see Boland, Tanenhaus & Garnsey, 1990) were low (under 5% in all conditions) and their variances were small. "Yes" reading times for instrument noun phrases were filtered for outliers such that reading times greater than 4,000 ms or less than 200 ms were omitted. Filtering resulted in the removal of 27 of 3723 (0.7%) reading times.

Measuring Effects of Predictability, Semantic Similarity, and Contextual Constraint

The goal of the modeling was to investigate how contextual constraints modulated the effect of semantic similarity. Reading times were submitted to a linear mixed-effects model for analysis using the R statistics program (version 2.14.0, R Development Core Team, 2011) using lme4 (version 0.999375-42, Bates & Maechle, 2011). Fixed factors consisting of Predictability, Similarity, Constraint, Length, and Frequency, described in more detail below, were used to predict reading time variances. Participants and items were random factors. Fixed effects terms that did not contribute significantly to the fit of the model, including all 4-way and 5-way interactions, were removed. We simplified the initial fully crossed and fully specified random effects structure to yield the maximally justified random effect structure, as discussed by Jaeger (2009) and Baayen, Davidson, and Bates (2008). Outliers with a standardized residual at a distance greater than 2.5 standard deviations from zero were removed (Baayen, 2008).

Model Predictors

Predictability We used log-transformed cloze probabilities from the above-mentioned listing study to estimate predictability. Each of the five responses was weighted by its order of mention. If an instrument was a participant's first choice, it was weighted 5, if it was the second choice, it was weighted 4, and so on.

Similarity We measured the degree of semantic similarity between each target instrument and each member of its semantic cohort (i.e., the other words produced in the above-mentioned listing study) using Latent Semantic Analysis (LSA) cosines (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990) with a semantic space created from the BNC. LSA cosines were weighted by their cohort-frequencies to determine the average semantic similarity of a target instrument with its semantic cohort. Average LSA cosines between targets and their semantic cohorts ranged from .08 for the lowest similarity to .54 for the highest similarity. Our measure of semantic similarity differs from that used by Schwanenflugel and LaCount (1988), in that they compared the target word with only the most likely word, rather than with all of the words in the semantic cohort. In addition, they used human similarity judgments, while we used LSA cosines as a measure of similarity.

Constraint We used the entropy of the probability distribution of all possible instruments for a context to measure the degree of constraint provided by the preceding context. Entropy values ranged between 2.55 for the most constraining contexts and 5.02 for the least constraining contexts, with a mean of 3.88.

Length The lengths of instrument noun phrases were included as an additional factor to control for any potential reading time differences which might be due to this perceptual factor. Length was measured in number of characters, including spaces. Lengths ranged from 5 to 16 characters, with a mean of 8.36 characters.

Frequency We log-transformed the raw frequencies of the head nouns of the instrument noun phrases, which were obtained from the BNC. Base 10 log-transformed frequencies ranged from 0 to 4.56 (i.e., occurring between 1 and ~36K times in the BNC), with a mean of 2.85. Because frequency was correlated with Length (r = .65) and

Predictability (r = .31), we residualized Frequency for Length and Predictability, so that the predictors would only reflect the component of frequency that did not overlap with length and predictability. All other predictors had correlations of less than 0.30.

Model Results and Discussion

We provide a summary of the linear mixed-effect regression model in Table 1 and a graphical representation of the interaction between Similarity and Constraint in Figure 2.

Length Longer words took longer to read. This is consistent with previous findings showing the effects of length (e.g., Juhasz & Rayner, 2003). Besides the interactions discussed below, there was a 3-way interaction between Length, Frequency, and Constraint. This was due to length effects being larger for low constraint, low frequency items and high constraint, high frequency items, and smaller for low constraint, high frequency items and high constraint, low frequency items. This possibly due to idiosyncrasies within our items, since we did not attempt to make sure that the same range of target word lengths were found in all conditions.

Frequency Unsurprisingly, more frequent words were read faster than less frequent words. This too is consistent with previous studies (e.g., Ashby et al., 2005; Juhasz & Rayner, 2003; Kliegl, Grabner, Rolfs, & Engbert, 2004; Staub, 2011). Frequency interacted with a number of other predictors as discussed below.

Predictability Consistent with many previous studies (Ashby, Rayner, & Clifton, 2005; Bicknell et al., 2010; DeLong et al., 2005; Ehrlich & Rayner, 1981; Federmeier et al., 2007; Frisson et al., 2005; Otten & Van Berkum, 2008; Rayner & Well, 1996; Staub, 2011; Van Berkum et al., 2005)**,** more predictable instruments were processed more quickly.

There was a 3-way interaction between Predictability, Frequency, and Constraint, as well as a 2-way interaction between Predictability and Frequency, and a marginally significant 2-way interaction between Predictability and Constraint. These are due to low frequency unpredictable words taking longer to read in low constraint contexts than would be expected from the simple effects of frequency and predictability (i.e., when all factors combine to give the comprehender the least amount of help in predicting the word). This may have resulted in the model underestimating the reading times for low frequency unpredictable words in low constraint contexts, giving the appearance of a lack of a frequency effect for highly predictable words in low constraint contexts.

Table 1: Summary of fixed factors from the linear mixed-effect regression model, when the effects of random variables were maximized, for predicting reading times of that target noun.

	Estimated Coefficient	S.E.	*t*-value
Intercept	713.72 (713.64)	17.51	40.75
Predictability	-59.63 (-33.22)	5.49	-6.05
Similarity	-271.95 (-28.94)	7.67	-3.77
Constraint	17.25 (8.99)	10.31	0.87
Length	15.35 (33.28)	7.43	4.48
Frequency	-31.05 (-31.22)	7.10	-4.40
Predictability x Similarity	-57.68 (-3.59)	5.44	-0.66
Predictability x Constraint	-36.97 (-10.89)	5.61	-1.94
Predictability x Length	-3.92 (-4.60)	5.84	-0.79
Predictability x Frequency	23.89 (13.56)	5.31	2.55
Similarity x Constraint	-453.57 (-24.82)	6.55	-3.79
Similarity x Length	-52.25 (-12.09)	5.92	-2.04
Similarity x Frequency	35.14 (3.85)	6.57	0.59
Constraint x Length	-8.38 (-9.45)	8.10	-1.17
Constraint x Frequency	11.42 (6.18)	6.42	0.96
Length x Frequency	1.51 (3.44)	5.02	0.69
Predictability x Frequency x Constraint	44.19 (12.57)	10.66	2.29
Similarity x Constraint x Length	-138.28 (-16.18)	6.13	-2.37
Similarity x Frequency x Length	-55.33 (-12.90)	2.54	-2.35
Constraint x Frequency x Length	-26.41 (-29.67)	2.84	-4.82

Note: All predictors are centered. Parenthetical values below the coefficients are standardized coefficients from an alternate version of the model with standardized predictors. *t*-values with an absolute value greater than 2 are significant at an alpha level of .05 (Gelman & Hill, 2007).

Similarity Instruments were read faster when they were more similar to the members of their semantic cohort than when they were less similar. This result replicates Roland et al.'s (2012) results. There was also a 3-way interaction between Similarity, Frequency, and Length and a 2-way interaction between Similarity and Length. These interactions were due to the effect of similarity being larger for longer words and smallest for short, high frequency words. These are both consistent with the notion that similarity effects are due to spreading activation during processing, as the slower reading times for longer words provide more chance for activation to spread between pre-activated words, and short, fast words, being read quickly, provide the least time.

Constraint There was no main effect of contextual constraint. Importantly however, Constraint interacted with Semantic Similarity, just as hypothesized. We analyzed this interaction by performing separate analyses on the data where one standard deviation was either added or subtracted from the values for each of the predictors in the interaction to create models reflecting low and high conditions for each predictor, respectively (Aiken and West 1991). There was an effect of Semantic Similarity when Entropy was high (i.e., low constraint contexts) (Estimated coefficient = -500.28, S.E. = 103.43, t-value = -4.84), but no effect of Semantic Similarity when entropy was low (i.e., high constraint contexts) (Estimated coefficient = -34.14, S.E. = 85.70, t-value = -0.40). The estimated high and low reading times are shown in Figure 2. The fact that Semantic Similarity did not facilitate the integration of instruments in strongly-constraining contexts is consistent with Schwanenflugel and LaCount's (1988) results. In addition, there was a 3-way interaction between Similarity, Constraint, and Length, with the similarity effects in the low constraint conditions being larger for longer words than for shorter words. Again, this is consistent with the notion that the added reading times of longer words allows more time for activation to spread between pre-activated words.

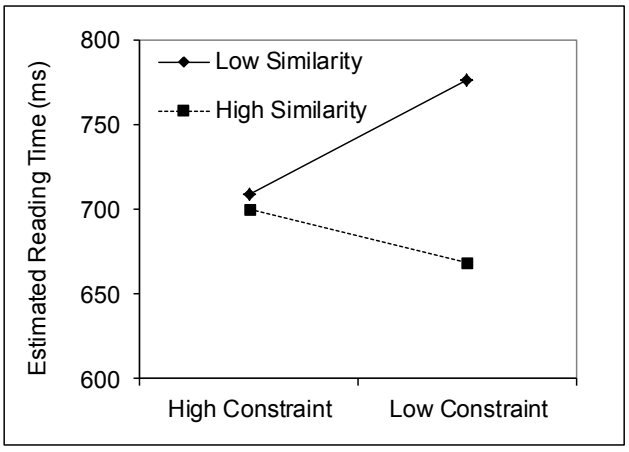


Figure 2: Interaction of Contextual Constraint and Similarity using standardized coefficients.

General Discussion

We found that semantic similarity between a target word and its semantic cohort has a stronger effect on processing when the context provides fewer constraints on what may appear in the target position. Alternatively, the effects of semantic similarity become weaker as the context becomes more constraining. The effect of contextual constraint on the degree to which semantic similarity affects processing has important implications for models of processing. Roland et al. (2012) suggested two possible causes for the semantic similarity effect: spreading activation between the representations for the words that comprehenders were anticipating, and the possibility that expectations for words and expectations for semantic features could have independent effects on comprehension difficulty. Our results suggest that the nature of comprehenders' expectations may vary with the degree of contextual constraint. In a highly constraining context (i.e., low entropy), there is no effect of semantic similarity, and comprehension difficulty appears to be primarily determined by the predictability of the target word. If the target word is expected, it is easy to process. If the target word is unexpected, it is difficult to process.

On the other hand, in a less constraining context, semantic similarity and predictability both influence processing. Not only are more predictable words easier to process, but so are words that are more similar to the other members of the semantic cohort. Words are most difficult to process when they are both unexpected and semantically distant from their semantic cohort.

One possible explanation for why contextual constraint modulates the influence of semantic similarity for unpredictable words is that in a highly constraining context, comprehenders may be expecting specific words, and face difficulty when the expectations turn out to be wrong. In a less constraining context, comprehenders may have less specific expectations – anticipating semantic features in common between a set of possible words (in addition to, or as an alternative to anticipating specific words). Thus, they face less difficulty when the target word turns out to be something other than the most likely word – as long as the target word shares some level of semantic similarity with the other likely possible words. Overall, our data suggests that word predictability, semantic similarity, and contextual constraint all have an impact on language comprehension.

References

Aiken, L.S., & West, S.G. (1991). *Multiple Regression:Testing and Interpreting Interactions*. Newbury Park, CA: Sage.

Ashby, J., Rayner, K., & Clifton, C., Jr. (2005). Eye movements of highly skilled and average readers: Differential effects of frequency and predictability. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 58A*(6), 1065-1086.

Baayen, R. H. (2008). Analyzing linguistic data: A practical introduction to statistics using R. Cambridge University

Press.
Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language, 59(4), 390–412.
Bates, D., Maechler, M. & Bolker, B. (2011). lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-42. http://CRAN.R-project.org.package=lme4
Bicknell, K., Elman, J. L., Hare, M., McRae, K., & Kutas, M. (2010). Effects of event knowledge in processing verbal arguments. *Journal of Memory and Language, 63*, 489-505.
Boland, J. E., Tanenhaus, M. K., & Garnsey, S. M. (1990). Evidence for the immediate use of verb control information in sentence processing. Journal of Memory and Language, 29(4), 413–432.
Boston, M. F., Hale, J., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. Journal of Eye Movement Research, 2(1). 1, 1–12.
Burnard, L. (1995). Users reference guide for the British National Corpus. Oxford: Oxford University Computing Services.
Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society For Information Science, 41*, 391-407.
DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, 8, 1117–1121.
Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior, 20,* 641–655.
Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research, 1146*, 75-84.
Frisson, S., Rayner, K., & Pickering, M. (2005). Effects of contextual predictability and transitional probability on eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31 (5),* 862-877.
Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.
Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics* (pp. 1–8). Pittsburgh, PA: Carnegie Mellon University.
Jaeger, T. F. (2009, May 14). Random effect: Should I stay or should I go? [Web log post]. http://hlplab.wordpress.com/2009/05/14/random- effect-structure/> Retrieved 24.07.11.
Juhasz, B. J., & Rayner, K. (2003). Investigating the Effects of a Set of Intercorrelated Variables on Eye Fixation Durations in Reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(6), 1312-1318.
Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. Psychological Review, 87(4), 329–354.
Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology, 16*(1-2), 262-284.
Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106, 1126-1177.
Otten, M., & Van Berkum, J. J. A. (2008). Discourse-based lexical anticipation: prediction or priming? *Discourse Processes*, 45(6), 464-496.
Padó, U.,Crocker, M., & Keller, F. (2009). A Probabilistic Model of Semantic Plausibility in Sentence Processing. *Cognitive Science*, 33(5):794-838.
R Development Core Team. (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.
Rayner, K., & Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review, 3*(4), 504-509.
Roland, D.,Yun, H., Koenig, J.-P., & Mauner, G. (2012). Semantic similarity, predictability, and models of sentence processing. *Cognition*, 122, 267-279.
Schwanenflugel, P. J., & LaCount, K. L. (1988). Semantic relatedness and the scope of facilitation for upcoming words in sentences. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 14,* 344–354.
Staub, A. (2011). The effect of lexical predictability on distributions of eye fixation durations. *Psychonomic Bulletin & Review*, *18*, 371-376.
Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 443–467.