Cognition 122 (2012) 267–279

Contents lists available at SciVerse ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Semantic similarity, predictability, and models of sentence processing

Douglas Roland [a,*], Hongoak Yun [b], Jean-Pierre Koenig [a,b], Gail Mauner [b]

[a] Department of Linguistics, University at Buffalo, The State University of New York, United States
[b] Department of Psychology, University at Buffalo, The State University of New York, United States

ARTICLE INFO

Article history:
Received 25 October 2010
Revised 22 November 2011
Accepted 23 November 2011
Available online 23 December 2011

Keywords:
Sentence processing
Prediction
Parsing
Surprisal
Semantic similarity
Spreading activation

ABSTRACT

The effects of word predictability and shared semantic similarity between a target word and other words that could have taken its place in a sentence on language comprehension are investigated using data from a reading time study, a sentence completion study, and linear mixed-effects regression modeling. We find that processing is facilitated if the different possible words that could occur in a given context are semantically similar to each other, meaning that processing is affected not only by the nature of the words that do occur, but also the relationships between the words that do occur and those that could have occurred. We discuss possible causes of the semantic similarity effect and point to possible limitations of using probability as a model of cognitive effort.

1. Introduction

It is a (true) cliché of psycholinguistics that the accuracy of human sentence processing is something of a feat, as words must be processed and integrated very quickly, given the continuous nature of the input stream. A popular, partial explanation for this feat is that, when processing sentences, we use all kinds of information to predict what is coming up next and that preactivation of the upcoming material makes integrating it easier.

Because of the widespread belief in the importance of predictability in sentence comprehension, much work has been done to enumerate the factors that comprehenders use to make their predictions about upcoming linguistic material. Factors that have been proposed to influence comprehension include verb subcategorization biases (e.g., Trueswell, Tanenhaus, & Kello, 1993), thematic fit of noun phrases (e.g., McRae, de Sa, & Seidenberg, 1997; Tanenhaus, Carlson, & Trueswell, 1989), the likelihood of different agents carrying out different actions (e.g., Kamide, Altmann, & Haywood, 2003), and various discourse factors (e.g., Binder, Duffy, & Rayner, 2001; Hare, McRae, & Elman, 2004).

Comprehenders appear to use contextual information to make predictions about upcoming material. DeLong, Urbach, and Kutas (2005) found an N400 response to indefinite determiners in English (*a*, *an*) that did not correspond to the noun that was most likely to occur next given the context. Similarly, Van Berkum, Brown, Zwitserlood, Kooijman, and Hagoort (2005) found an ERP response when Dutch determiners did not match the anticipated following noun in grammatical gender. Both of these results suggest that comprehenders have formed expectations for specific words to occur in advance of the point at which the words actually occur.

The linking assumption between predictability and cognitive effort is that the cognitive representations for expected words (or phonemes, syntactic structures, etc.) are presumed to be more highly activated than those for less expected ones. Consequently, they are presumed to be easier to retrieve from memory, and require less additional activation to incrementally update the set of representations created during the comprehension of the utterance. In a sentence like *The poor student ate macaroni*

* Corresponding author. Address: Department of Linguistics, University at Buffalo, Buffalo, NY 14260, United States.

E-mail address: droland@buffalo.edu (D. Roland).

0010-0277/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2011.11.011

and cheese, the word *cheese* is highly predictable. As a consequence, processing the word *cheese* results in only minor changes to the overall set of cognitive representations involved in comprehending the sentence. If the word *cheese* were replaced with a less predictable word, such as in *The poor student ate macaroni and caviar*, the processing of the sentence at the word *caviar* would require a larger change to the overall set of activated cognitive representations, and thus more cognitive effort.

The expectations for a particular word also reflect the expectations for structures at levels besides the word level. In a sentence like *The horse raced past the barn fell*, the reduced relative structure is very unexpected, as is the word *fell*. Consequently, processing the word *fell* results in major changes to the set of cognitive representations involved in comprehending the sentence. We will discuss the issue of exactly how expectations at different levels of representation are related to expectations at the word level in the final discussion section of this paper, but informally we assume the inclusion of expectations at all levels when we refer to word predictability throughout the paper.

The relationship between word probability and cognitive effort has been formalized in theories such as the surprisal theory (Hale, 2001; Levy, 2008), which relies on Eq. (1) to make predictions of cognitive effort. This equation indicates that the degree of cognitive effort required to process a word is dependent on the negative log probability of that word, given the preceding context. This measure has been described in several ways that are mathematically equivalent, but which emphasize different aspects of possible cognitive interpretations of the measure. Levy (2008) characterizes the measure in terms of the degree of difference between the probability distributions of the possible interpretations of the message before seeing the word and after seeing the word. Jurafsky (2003) characterizes it in terms of the amount of information conveyed by the word. Hale (2001) characterizes it in terms of the probability mass of the interpretations that are disconfirmed upon hearing a word.

$$\text{difficulty} \propto -\log p(w_i|w_{1\ldots i-1}\text{CONTEXT}) \qquad (1)$$

One commonality across discussions of expectations in comprehension is that the degree of cognitive effort needed to process a particular message tends to be cast in terms of how likely a particular word, structure, or message is, relative to another word, structure, or message. Aside from their relative probabilities, little attention is paid to potential relationships between the various possible words or structures. The other words that could have occurred in that position are only relevant in that, if a particular word is very likely, other possible words must necessarily be unlikely. This indirect relationship arises because the probabilities of all possible words must sum to 1. Importantly, it is assumed that the nature of the other words that could have occurred has no other bearing than this indirect relationship on the level of difficulty faced in processing the target word itself.

We challenge this often implicit assumption that the degree of cognitive effort is determined solely by the properties of the material that actually occurs by providing evidence for our *Semantic Similarity Hypothesis*, which predicts that processing will be facilitated to the degree that the different possible choices that could occur in a given context are semantically similar to each other. One possible cause for the predicted processing facilitation is that activation may spread between the representations of the different possible choices that are being activated during processing (e.g., McRae, Ferretti, & Amyote, 1997). In this view, greater semantic similarity between the possible word choices would result in greater activation of this set of words, and thus greater facilitation in processing. Alternate possible causes of a semantic similarity effect will be addressed in the final discussion section.

To better understand our *Semantic Similarity Hypothesis*, consider the sets of possible instruments that could occur in the sentential contexts shown in (1) and (2). Based on the hypothetical distributions of possible instruments shown in Fig. 1 for these contexts, probabilistic theories of language comprehension would predict that instruments such as *spear* and *sword* would be easier to process than instruments such as *machete* and *rock,* due to their greater degrees of anticipatory activation. This prediction is consistent with a long history of experimental results showing that the degree to which material is predictable from the context affects comprehension processes, as reflected in measures such as reading times (e.g., Rayner & Well, 1996), electrophysiological response (e.g., Federmeier, Wlotko, De Ochoa-Dewald, & Kutas, 2007), and the ability to comprehend degraded input (e.g., Obleser & Kotz, 2010). Probability-based accounts such as the surprisal theory (Hale, 2001; Levy, 2008) and the SynSem Integration Model (Padó, Crocker, & Keller, 2009) have had good success at modeling such differences in relative cognitive loads during language comprehension across a wide variety of psycholinguistic phenomena based solely on knowing how likely a target word is, given its context.

(1) The aboriginal man jabbed the angry lion with a/an —.

(2) The aboriginal man attacked the angry lion with a/an —.

Models which base their predictions only on the probability of target words (e.g., Hale, 2001; Levy, 2008) necessarily also make the following predictions for the contexts shown in (1) and (2), given the distributions of possible instruments shown in Fig. 1. First, because the probability of *spear* is the same in both contexts, *spear* should have the same degree of difficulty in either context. Second, because *machete* has the same probability in context (1) as *rock* has in context (2), *machete* and *rock* should also have the same respective degree of difficulty, once other factors such as length and frequency are taken into account. However, in the examples shown in Fig. 1, there is a difference between the distributions of possible instruments for these two contexts. The set of likely instruments for the *jab* context are typically all sharp, pointy objects. Several of the possible instruments for *attack* also share these properties, but many of the less likely instruments for *attack*, including *rock*, do not. If the representations of the various possible instruments are initially activated based on their respective probabilities, activation may

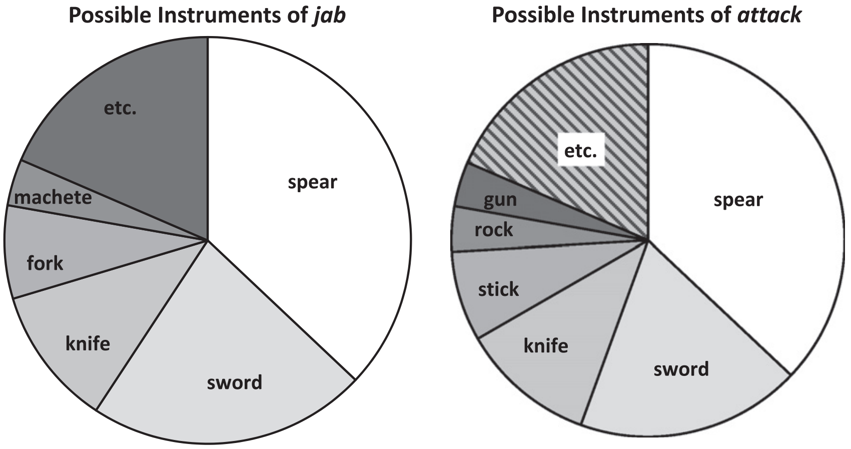

Fig. 1. Hypothetical distribution of possible instruments of jab and attack in examples (1) and (2).

spread between representations based on their degree of shared semantic similarity. Fig. 2 shows the relative locations of these possible instruments in a hypothetical semantic space. Notice that the other possible instruments in the *jab* context are all semantically similar to *spear*, while in the *attack* context, some of the possible instruments such as *rock* and *gun* are less similar to *spear*.

Because more of the possible instruments in context (1) have properties in common with *spear* than those in context (2), our *Semantic Similarity Hypothesis* would predict that *spear* would be processed more quickly in context (1). Similarly, because *machete* has more in common with the other instruments in context (1) than *rock* does in context (2), we would predict a processing advantage for *machete*, even though there is no difference in probability.

The main focus of our investigation is whether the semantic similarity between words that could occur in a context has an additional influence on processing beyond the influence of the predictability of the word that does occur. To do this, we first conducted a completion/listing study to establish the distribution and likelihoods of possible words that could appear in a set of contexts. Then we conducted a reading time study to establish the degree of cognitive effort required to processes specific possible words. We then used Latent Semantic Analysis (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990) to measure the degree of semantic similarity between the possible words that could have occurred in each of the sentences (as determined by the sentence completion/listing study) and the word that does occur in the sentence. Finally, we used a linear mixed-effects regression model to demonstrate that the degree of shared similarity between the actual target word and the other possible words plays a role in predicting processing difficulty above and beyond the effect of the probability of the word that actually occurred.

2. Preparatory studies

2.1. Establishing the distribution of possible words given a context

In order to investigate the roles of predictability and semantic similarity in processing, we need a set of human performance observations based on target words appearing in contexts where the set of possible alternative words is easily characterizable, and yet has a range of differing probabilities and degrees of similarity with other possible fillers of the target word slot. Instrument phrases, such as *with a spear*, in the *jab* and *attack* examples above make an ideal target for investigation as the distribution of possible instruments that are likely to be used in the situation described in each sentence is fairly well characterizable.

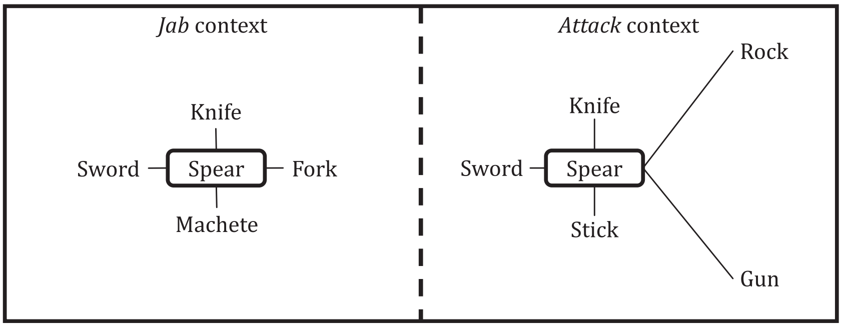

Fig. 2. Semantic similarity between spear and the other possible instruments in contexts (1) and (2).

Importantly, we can create sentences describing different scenarios that exhibit a range of degrees of semantic similarity between the sets of possible instruments.

We prepared 56 partial sentences ending with the word *with*, such that each sentence was likely to be completed with an instrument noun phrase. An example is shown in (3) and the complete set of materials for this completion study and for the reading time study can be found in the Appendix. We asked forty native English-speaking undergraduates from the University at Buffalo, who received partial course credit for their participation, to provide possible instruments for each prompt. Participants were told to produce five possible instruments that they might use in the given event, and to list them from 1 to 5 in the order in which they came to mind. We used a listing task instead of a simple completion task to increase the amount of data gathered per participant. In pilot experiments, we found no differences in our results with respect to the conclusions we draw between versions where we asked participants for a single response or five responses (or whether we used only the first of the five responses or all five responses from the five response version).

(3) The gladiator jabbed the African tiger with a/an —.

The responses collected in this sentence completion/listing task were used to determine the likelihoods of each of the possible instruments given their respective contexts. In analyzing the data, we faced the issue of whether to collapse across similar responses or not. For example, should a *sharp knife* be treated as being equivalent to a *knife*? We used the following criteria to make such decisions: Single word responses were all treated as forming their own instrument type. Multi-word responses were either treated as being separate types or as being part of one of the single word types, depending on the nature of the multi-word response. If the multi-word response consisted of an instrument plus a modifying adjective, (e.g., sharp knife) and the adjectives simply described physical properties of the instruments (e.g., sharp, dull, large, blue), the responses were treated as mentions of the single word instrument. When the multi-word response consisted of a compound noun that seemed to refer to a different type of instrument than the single word version (e.g., *butcher knife* seems to be a separate type of knife), they were treated as being separate types. This decision was made based on whether the compound-noun form of the instrument had a separate entry in the WordNet dictionary. Nearly all such compounds occurred in WordNet. For example, because *butcher knife* is an entry in WordNet, it was treated separately from *knife*. However, two compounds, *taser gun*, which occurred nine times (out of 6562 responses total), and *meat tenderizer*, which occurred once, were not in WordNet, and were treated as instances of *gun* and *tenderizer*, respectively. Finally, when two instruments were produced in a coordinate form (e.g., needle and thread), the response was treated as an instance of each of the separate nouns (e.g., *needle and thread* was treated as one instance of *needle*, and one instance of *thread*).

Another issue we faced in determining the likelihood of each possible instrument was whether instances of an instrument produced as a participant's top-ranked response were equivalent in weight to the same instrument being produced as a low ranking response by another participant (recall each participant had to provide five possible instruments). We chose to use a weighted system[1] to determine the likelihood of each instrument. If an instrument was a participant's first choice, it was given a weighting of 5, if it was the second choice, it was given a rating of 4, and so on. We then computed each instrument's weighted probability by dividing its weighted score by the sum of the weighted scores of all possible instrument fillers for that item. Thus, the probabilities for each individual instrument ranged from zero to one, and the probabilities of all possible instruments given a specific context summed to one, with the difference between our weighted data and the original unweighted data being that the instruments that were more likely to be named as higher ranking choices by participants were treated as being more probable than those named as lower ranking choices.

The responses collected in this sentence completion/listing task were also used to select target instruments for use in the reading time study and to establish the degree of semantic similarity between each of the target instruments and the other possible instruments for each context. The details of these procedures are described in further detail below.

2.2. Generating reading time data to be modeled

To investigate the effects of predictability and semantic similarity on processing difficulty, we conducted a sentence reading time study. The purpose of this study was to obtain reading times for the instrument nouns for which we had obtained probability estimates and for which we could characterize the degree of semantic similarity between the instrument target words and the other possible instrument words that were likely to have been anticipated.

2.2.1. Method

2.2.1.1. Participants. Eighty-four native English-speaking undergraduates from the University at Buffalo, who were not part of the preceding sentence completion study, received partial course credit for their participation in this study.

2.2.1.2. Materials. Fifty-six pairs of declarative sentences with syntactically optional prepositional phrases were constructed, as shown in examples (4) and (5). Each member of the pair of sentences differed only in whether its instrument filler was highly likely, such as *sword* in (4), or was less likely, such as *spike* in (5). We used both highly likely and less likely instruments to increase the range of probabilities reflected in our reading time data and to maximize accuracy in fitting our model for predicting reading times from word probabilities. Two instruments for each item pair (e.g., *sword*, *spike*) were selected based on data

[1] We report results based on the weighted probabilities due to slightly better model fits, but similar results were obtained using unweighted likelihoods.

gathered in the completion/listing study described below. We choose the most likely instrument[2] for each context as the first member of the pair (*Mean probability of most likely instrument* = .24, *SD* = .06, *Range* = .08–.35), and one of the less likely (but still plausible) instruments for the other member of the pair (*Mean probability of less likely instrument* = .02, *SD* = .01, *Range* = .003–.05). The less likely instruments were chosen to be plausible to prevent extreme reactions to our materials. To avoid confounding sentence final wrap-up-effects with the timing of instrument processing in critical regions (Just & Carpenter, 1980), all sentences included sentence-final phrases (e.g., *in the Colosseum*). All sentences were presented region by region. Regions are indicated by vertical lines (|) in Examples 4–5

(4) The gladiator | jabbed | the African tiger | with | a sword | in | the Colosseum.
(5) The gladiator | jabbed | the African tiger | with | a spike | in | the Colosseum.

The experimental sentences had to satisfy two additional conditions. First, because information about the thematic fit encoded in an event is rapidly used during on-line sentence processing (McRae et al., 1997; Trueswell, Tanenhaus, & Garnsey, 1994), a plausibility norming study was conducted to ensure that instrument nouns were plausible. The instrument fillers that were generated from the listing study, which included the selected target instrument fillers, were shown to forty participants who rated, on a 7-point Likert scale, how likely each instrument that had been generated for a sentence frame was as a filler of a sentence fragment like (3), with 7 referring to "highly plausible" and 1 referring to "not-at-all plausible". The mean plausibility ratings for highly-likely and less likely instrument fillers were 6.1 (*SD* = .66, *Range* = 3.4–6.9), and 4.8 (*SD* = .74, *Range* = 3.4–6.6) respectively. Importantly, the plausibility ratings of all target instrument fillers were over 3.4, suggesting that the experimental sentences were considered plausible. Consequently, the processing difficulty associated with instrument fillers in this on-line study is unlikely to be due to the low plausibility of the nouns as instrument role fillers. This measure of plausibility did not significantly predict reading times in models in which predictability was included as a factor.

The second condition that we placed on the experimental sentences was that when action verbs co-occurred with the preposition *with*, readers would be highly likely to expect the *with* prepositional phrase to continue with an instrument interpretation (e.g., *The gladiator jabbed the African lion with a sword*) (Spivey-Knowlton & Sedivy, 1995), instead of a manner (e.g., *The gladiator jabbed the African lion with ferocity*) or comitative interpretation (e.g., *The gladiator jabbed the African lion with the other gladiator*). A completion norming study was conducted to confirm that all items had a bias towards the instrument use of *with*. Forty participants were given sentence fragments like (3) and asked to complete sentences with the first response that came to mind. In this way, it was possible to establish the most natural continuation following the preposition *with*. The "use test" proposed by Koenig, Mauner, Bienvenue, and Conklin (2008) was employed as a criterion to decide whether the noun phrases that were produced were instruments or not. For example, supposing that *sword* was produced as a response to (3), the sentence could be paraphrased as *The gladiator used a sword to jab the African lion*. As long as the paraphrased sentence had the same meaning as the original sentence, *sword* was considered to be an instrument. The percentages of instrument continuations were high across all items (*M* = 94.2, *SD* = 8.6, *Range* = 60–100). The results of this norming minimize the concern that any difficulty in processing instruments might be due to the possibility that instruments were unexpected.

For the actual reading time experiment, 56 pairs of experimental sentences were counterbalanced across four[3] presentation lists (i.e., 28 sentences per list) so that no more than one item from each pair appeared on any given list. Experimental sentences were pseudo-randomly intermixed with 77 distractor sentences, so that no two experimental items appeared consecutively. To obscure any systematicities in experimental materials, distractor sentences included a variety of prepositional phrases headed by a number of different prepositions (e.g., *on*, *in*, or *from*) and other syntactic structures (e.g., subordinate clauses or adverbial phrases). Finally, because participants were asked to judge whether each sentence they read made sense, 30% of the total number of trials were designed to not make sense. Nonsensical sentences could be rejected for a number of reasons, including syntactic violations (e.g., **People went to the west to make many money*), tense violations (e.g., **The FBI has set up a tip line tomorrow to collect information*), or semantic anomalies (e.g., **The table swore to go on a diet last night*). Participants were given a 1–2 min break after they had completed half of the experimental trials.

2.2.1.3. Procedure. Participant-paced, region-by-region reading was accompanied by an incremental judgment task. This secondary judgment task was used to increase the sensitivity to subtle semantic differences and ensure that participants were processing sentences for meaning (see Mauner, Tanenhaus, & Carlson, 1995). Participants first saw a row of dashes and white spaces displayed on a CRT monitor. The dashes corresponded to all of the non-white-space characters of each stimulus sentence. Participants pressed a "Yes" key marked on a computer keyboard to reveal the first sentence region. This caused the dashes corresponding to this region to be replaced by words. To reveal the next region, participants again pressed the "Yes" key. This second press caused the first region to revert to dashes and reveal the next region.

[2] The selection of the most probable instrument was based on the weighted probabilities described in the upcoming modeling section. The decision to use weighted probabilities instead of unweighted probabilities affected only 3 out of 56 items. For two of these items, the instrument we chose would have been the second highest instrument using an unweighted scheme, and for the third, the instrument we chose would have been the third highest ranking instrument. The highest ranked instrument remained the same for the other 53 out of 56 items.

[3] There were four lists instead of two due to the inclusion of an additional factor (there were two subtypes of instrument taking verbs in our items) which we do not analyze in this paper.

Participants continued pressing the "Yes" key to read each subsequent region as long as the sentence they were reading continued to make sense to them syntactically, semantically, and pragmatically. If at any time a sentence stopped making sense, participants pressed a "No" key. The "No" response immediately terminated the current trial and initiated the next trial. "Yes" Reading times and "No" makes-sense judgments were collected for each region. Before the experiment began, participants were asked to read the instructions that described the task with some examples. After reading the instructions, they completed six makes-sense trials and six nonsensical practice trials to familiarize themselves with the task and the response keys.

2.2.2. Dependent variables

The self-paced reading paradigm with a judgment task yielded two dependent variables: the "No" judgments and the reading times for each segmented region to which participants responded "Yes". The "No" judgments were used as an on-line check of the acceptability of the instrument role fillers. Given the results of the plausibility norming, it was predicted that there would be few "No" judgments and that they would not differ across conditions at any region. Consequently, the main dependent variable of interest in this study was the "Yes" reading times.

2.2.2.1. Judgments. For each participant, the adjusted percentage of "No" judgments was tabulated for each region of a stimulus sentence using the procedure outlined in Boland, Tanenhaus, and Garnsey (1990). This was done in order to control for the fact that participants who had rejected sentences at earlier regions did not have the possibility of rejecting the same sentence at later regions. Adjusted percentages for each sentence trial were computed by dividing the number of "No" judgments at a given region by the number of remaining opportunities that a participant had for responding "No" in that sentence. Mean adjusted percentages were computed within condition and region for each participant.

Descriptively, the adjusted percentages of "No" responses across participants at the prepositional noun region for both highly likely and unlikely instrument sentences were 1.33% when target instrument nouns were highly likely and 1.88% when they were unlikely. This suggests that the experimental sentences were highly acceptable in both conditions. More importantly, "No" judgments to highly likely and unlikely instrument sentences did not differ (t (169) = −.94, p = .35).

2.2.2.2. Reading times. Filtering of reading times for outliers was conducted only for sentences that participants continued to judge acceptable. "Yes" reading times greater than 4000 ms and lower than 200 ms were omitted, resulting in the removal of ten scores (less than 0.4% of the total data).

3. Measuring the effects of predictability and semantic similarity

The goal of the modeling we conducted was to determine whether processing time was influenced by the semantic similarity between a target instrument word and other possible instrument words that could have occurred in the target sentence frame. Reading time data were submitted to a linear mixed-effects regression model for analysis. Analyses were conducted using the lme4 (version 0.999375-33, Bates & Maechler, 2010) and languageR libraries (version 1.0, Baayen, 2010) for the R statistics program (R Development Core Team, 2010). Along with a measure of semantic similarity that was derived from the distributions of possible instruments generated in our completion listing study, additional fixed factors included measures of other variables known to influence reading times. These additional fixed factors were measures of word length, log frequency, and predictability. All interactions between the fixed factors were included in an initial model. Terms which did not result in a significant improvement in model fit were removed from the reduced model reported in the paper. Participants and items were included as random factors in the model. All fixed factors were centered. We residualized predictors that were highly correlated to avoid co-linearity effects[4] (Jaeger, 2010). Finally, we simplified the initial fully crossed and fully specified random effects structure to yield the maximal random effect structure justified by model comparison following the procedures discussed by Jaeger (2009) and Baayen, Davidson, and Bates (2008). In what follows, we describe in greater detail how we derived values for our predictor variables.

3.1. Model predictors

3.1.1. Predictability of the target instrument given the context

The predictability of a word given its context is known to be correlated with the reading time for that word (Boston, Hale, Kliegl, Patil, & Vasishth, 2008; Levy, 2008; Padó et al., 2009). We used the data generated in the sentence completion/listing task described above to establish the probabilities for each of our target instruments given their preceding contexts. We used the log transformed probabilities of each instrument as a predictor in our model, as theories such as the surprisal theory (Hale, 2001; Levy, 2008) predict that the log of the probability best predicts cognitive effort.

3.1.2. Semantic similarity between the target instrument and the other possible instruments

The goal of this predictor was to measure the degree of semantic similarity between the target instrument and the other possible instruments that could have occurred in the same context. Given the examples shown in Fig. 2, the prediction of the Semantic Similarity Hypothesis is that reading times would be faster for *spear* in the *jab* context because the other possible instruments were semantically more closely related to *spear* than many of the possible instruments in the *attack* context. More generally, it predicts that target instruments having a greater degree of similarity to the other possible instruments for a given

[4] The correlations between predictors in the final model were all under .35, kappa (condition index) was 1.80, and the variance inflation factor was 1.29, suggesting that colinearity is not an issue in our data.

context will be processed more readily than those with less similarity. To generate a measure of the degree of semantic similarity between the target instrument and the other possible instruments for that context, we used Latent Semantic Analysis (LSA) (Deerwester et al., 1990). LSA is a technique for representing words in a high dimensionality semantic space. Using a semantic space that we generated from the British National Corpus, we computed the cosine between the target instrument and each of the other possible instruments that had been named in the sentence completion/listing task. Potential instruments which were more similar (e.g., *spoon–whisk*, *screwdriver–pliers*, *screwdriver–drill*) had higher cosines (.89, .85, and .72 respectively), while those which were less similar (e.g., *extinguisher–baking soda*, *binoculars–gun*, *hammer–football*) had lower cosines (.16, .19, and .16 respectively). The similarity measure for each target instrument was calculated using the average of the pairwise cosines between the target instrument and each of the other instruments (i.e., excluding the target instrument itself) produced in the sentence completion/listing study. Thus, if the distribution of possible instruments from the completion study for a particular item was the set $\{A, A, A, B, B, C, T, T, T, T\}$, where T is the target instrument, and A, B, and C are other possible instruments, the similarity metric would be equal to (3 cosine(A,T) + 2 cosine(B,T) + cosine(C,T))/6. Note that the similarity between the target instrument and the other possible instruments is not affected by the number of times the target instrument was produced in the completion study.

3.1.3. Length

Longer words have been shown to take longer to read (e.g., Juhasz & Rayner, 2003; Kennedy & Pidcock, 1981; Kliegl, Grabner, Rolfs, & Engbert, 2004; Mikk, 1972). Consequently, the length of the target instrument noun phrase (in characters) was included as a predictor.

3.1.4. Word frequency

More frequent words have been shown to take less time to read than less frequent words (e.g., Juhasz & Rayner, 2003; Kliegl et al., 2004; Raney & Rayner, 1995; Staub, White, Drieghe, Hollway, & Rayner, 2010), with reading times being linearly related to the logarithm of word frequency. Thus, we used log-transformed frequencies of the head nouns of the instrument noun phrases as a predictor. The raw frequencies of target instrument nouns were obtained from the British National Corpus (BNC) (Burnard, 1995). Because word frequency is correlated with both word length (more frequent words are shorter, e.g., Miller, Newman, & Friedman, 1958; Zipf, 1935) and predictability (more frequent words are more likely to be named in the listing task), word frequency was residualized against both word length and word predictability. As a result of this residualization, this predictor reflects the influence of word frequency on reading times once word length and word predictability are taken into account. If cognitive effort were based solely on word predictability, as predicted, for instance, by the surprisal theory, one would not expect to see an independent effect of frequency on reading time. However, there are previous reports of independent effects of frequency and predictability. Dambacher, Kliegl, Hofmann, and Jacobs (2006) found that lexical frequency affected the P200 component while contextual predictability affected the N400 component during word-by-word sentence reading. Rayner, Ashby, Pollatsek, and Reichle (2004) found separate effects of frequency and predictability on reading times during sentence reading in an eye-tracking task. Independent effects of frequency presumably reflect cognitive processes that are sensitive to the degree of long-term exposure to a word, but not the actual context in which the word appeared – for example, non-context sensitive lexical access processes.

3.2. Model results and discussion

Following the procedure outlined in Baayen (2008), we performed an initial fit for our model, and then removed all data points with residuals greater than 2.5 standard deviations from the mean (75 data points, approximately 3.3% of our overall data) before performing the final fit for our model, which is reported below. This procedure removes outliers[5] (in our data, exclusively long reading times) that were likely to have been caused by factors other than those included in our model. Table 1 shows the estimated coefficients, the standard error for the estimated coefficients, and their respective t values.[6] If the absolute t value for a fixed factor was over 2, the effect of the fixed factor was considered to be significant at $\alpha = .05$ (Gelman & Hill, 2007). In addition, a second model was fit using standardized predictors following the same procedure. To facilitate comparison of effect sizes, the standardized estimated coefficients from this second model are provided in parentheses along with the original coefficients in Table 1.

3.2.1. Predictability

We found a significant effect of word predictability, with more predictable words being read more quickly than less expected words. This finding is consistent with a wide variety of findings showing that contextual expectations about upcoming linguistic material affect the cognitive effort needed to process that material, such as Boston et al.'s (2008) findings of an effect of predictability on reading times. As such, it provides support for probabilistic accounts such as the surprisal theory. In addition, we found an interaction between predictability and frequency, which we will discuss in the section on frequency.

In stating that our finding of an effect of predictability on reading time provides support for the surprisal theory, we note that there is some ambiguity in the literature

[5] Because of reviewer concern that the outlier removal procedure in Baayen (2008) relies on a model containing the factor of interest (i.e., similarity), we performed an alternate analysis where outlier removal was based on a model excluding similarity as a predictor. This alternate analysis resulted in the removal of 77 data points having residuals greater than 2.5 standard deviations from the mean, including the entire set of 75 data points removed when using a model including similarity as a predictor, as well as three additional data points. This difference resulted in minor changes to the values of the estimated coefficients, with the same set of predictors being significant in both analyses.

[6] We are not able to provide p values, because MCMC sampling for models with random slopes and intercepts has not yet been implemented.

Table 1
Summary of fixed effect predictors from the linear mixed-effects regression model for predicting reading times of the target word.

		Estimated coefficient	Standard error	*t* Value
(Intercept)	714.70	(714.70)	23.57	30.32
Predictability	−60.23	(−39.49)	10.74	−5.61
Similarity	−179.33	(−18.77)	79.75	−2.25
Length	21.52	(48.64)	3.28	6.57
Frequency	−14.55	(−14.55)	9.42	−1.54
Predictability × Frequency	35.39	(23.20)	10.41	3.40

Note: All predictors are centered, frequency predictor is residualized for length and predictability. Parenthetical values following the estimated coefficients are standardized coefficients from an alternate version of the model with standardized predictors.

about the terms *predictability* and *surprisal*. For example, Boston et al., 2008 report results for separate measures of predictability and surprisal (the former based on the probability of the word, and the later based on structural probabilities). Although most implementations of surprisal measures have relied solely on structural probabilities (e.g., Hale, 2001), surprisal is defined as the negative log of the probability of a word occurring given the context, regardless of how the probability is calculated (see Levy, 2008). Thus, Boston et al.'s *frequency*, *bigram*, *predictability* and *surprisal* measures are all measures of surprisal. Our predictability result supports the surprisal theory, but is based on cloze probabilities – the source of information used in Boston et al.'s *predictability* predictor. Because our materials all have the same syntactic structure, and thus the same non-lexicalized structural predictability, it would not be meaningful to add a measure of non-lexicalized structural predictability (such as Boston et al.'s surprisal measure) to our model.

3.2.2. Semantic similarity

Our key finding is an independent effect of semantic similarity. Words were read more quickly when they shared more semantic similarity with the other words that could have occurred in the same context. This result supports our Semantic Similarity Hypothesis. It also indicates that the cognitive effort required to process a word depends not only on the properties of the word being processed, but on the properties of other words that could have appeared in the same context. Thus, our data also suggests that models of processing need to incorporate knowledge of how the words that appear relate to the other words that could have appeared.

Our results showing separate effects of predictability and semantic similarity support the conclusions drawn by Federmeier and Kutas (1999), while controlling for a potential issue in the design of their materials. Given contexts such as *They wanted to make the hotel look more like a tropical resort. So along the driveway they planted rows of…*, they found larger N400 responses relative to an expected word (e.g., *palms*) for words that were both unexpected and of a different category than the expected target word (e.g., *tulips*), than to words that were unexpected, but from the same semantic category as the target (e.g., *pines*). They interpret their results as providing evidence for "an influence of semantic feature overlap (as reflected in taxonomic semantic categories) that is independent of the fit of that word to the specific sentence context" (Federmeier & Kutas, 1999, p. 490). However, due to the nature of their materials, it is possible that their results were entirely due to the predictability of their target words, rather than semantic similarity/category membership. They measured the (un)expectedness of the target words using a cloze task, reporting a mean cloze probability of 0.004 for the within-category violations and 0.001 for the between-category violations. These probabilities suggest that the within-category and between-category unexpected words were similarly unlikely, and that their results were due to the within-category/between-category difference. However, it appears that most of the words that they used as unexpected words were actually never produced by participants in their cloze norming study. Given that there were either 59 or 56 participants in their cloze study (depending on which list the item appeared in), if each unexpected word had been produced by at least one participant, one would expect mean cloze probabilities greater than .016. If neither *tulips* nor *pines* were produced by their participants, but *tulips* was much less likely than *pines*, *pines* might possibly have been produced once in a sample of 100 participants (yielding a cloze probability of 0.01), while tulips might have required a sample of 10,000 participants before it would have been produced (yielding a cloze probability of 0.0001). Thus, given the uncertainty about the probability of words that were not produced in the cloze task, it is possible that probability differences between their between-category unexpected words and within-category unexpected words was much larger than their mean cloze probabilities suggest. Because of the greater certainty about the probability of the unexpected items in our experiment (all of the target words were produced by participants in our norming task), our results provide stronger evidence for separate effects of predictability and semantic similarity/category membership than the Federmeier and Kutas study.

3.2.3. Length

As expected, we found that word length did predict processing difficulty. Longer words took longer to read. This is consistent with previous findings showing effects of length (e.g., Juhasz & Rayner, 2003; Kennedy & Pidcock, 1981; Kliegl et al., 2004; Mikk, 1972), and with models of reading, such as EZ-Reader (Pollatsek, Reichle, & Rayner, 2006;

Reichle, Rayner, & Pollatsek, 2003), that include word length as a predictor of reading time.

3.2.4. Frequency

We did not find a main effect of frequency[7]. However, we found an interaction between frequency and predictability, indicating that frequency affected reading time (frequent words were faster) when the target instruments were less predictable, but that the frequency had less of an impact on more predictable target instruments. This is somewhat different from previous studies, where independent effects of frequency and predictability have been found. In a study of event-related potentials, Dambacher et al. (2006) manipulated word frequency and predictability in a word-by-word reading task. They observed a P200 component reflecting word frequency, and an N400 component reflecting word predictability, suggesting that processing involves both frequency and predictability-related processes. In an eye-tracking study, Ashby, Rayner, and Clifton (2005), found independent effects of frequency and predictability for highly skilled readers, but no effect of frequency on gaze duration for average readers. Based on their overall pattern of results, they conclude that the lack of an effect of frequency on gaze duration is due to average readers not completing lexical access while fixating on low frequency words. This interpretation was supported by the fact that delayed frequency effects appeared in other eye-tracking measures. Other eye-tracking studies that have found independent effects of frequency and predictability include Dambacher et al. (2006), Hand, Miellet, O'Donnell, and Sereno (2010), Rayner et al. (2004), and Staub (2011). Boston et al. (2008) also found separate effects of predictability and frequency in their analysis of an eye-tracking corpus. They used linear mixed-effects models with frequency, length, and several measures of predictability (bigram probabilities, cloze predictability, and surprisal based on syntactic structures) as predictors, and six different measures of fixation duration, including both early and late measures of eye movement as dependent variables. Frequency was found to be a significant predictor for all six eye movement measures.

The differences between our frequency results and previous results is likely to be due to two differences between this study and previous studies. Our experiment was not specifically designed to test frequency, and our materials had a restricted range of frequencies compared to studies such as the Ashby et al. study. The mean frequency of their high frequency items was 160/million, and the mean frequency of their low frequency items was 4/million. By contrast, if our set of instruments is split into two bins using a median split, the high frequency bin has a mean frequency of only 52/million, while the low frequency bin has a mean frequency of 2 per million. This suggests that our materials lacked the sort of high frequency words found in experiments designed to investigate frequency effects. In fact, only 8 of the 112 instruments in our study had a frequency higher than 100/million. It is possible that we could have found a frequency effect if we had used materials with a greater range of frequencies, although this would still not account for the presence of an interaction between frequency and predictability in our results.

Besides the difference in frequency range, we also used a different paradigm than the previous studies. We used a stop making sense region by region reading paradigm, while the previous researchers who found independent effects of frequency and predictability used eye-tracking paradigms. These methodologies may be more sensitive to effects of earlier aspects of processing, while stop making sense paradigms require an explicit determination that the sentence makes sense, and thus the results may be more strongly driven by later aspects of processing.

Independent effects of predictability and frequency would not be expected (nor would interactions between frequency and predictability) under a probability only view of processing, where the effects of frequency (i.e., the probability of a word given no knowledge of the context) would be subsumed by the effects of predictability (i.e., the probability of a word given knowledge of the context). Because frequency reflects long-term exposure to the word, but not how likely the word is in a particular context, it seems likely that frequency and predictability effects are due to different aspects of the comprehension process, with frequency reflecting earlier, and most likely, more perceptual aspects of lexical access, and predictability reflecting the integration of the new word with the rest of the sentence. The issue of how predictability and frequency effects might be caused by different stages of processing are further discussed in Staub (2011), as are the resulting implications for probability-only models.

4. General discussion and conclusions

Our goal in this paper was to determine whether there was any influence on the processing of a particular word from other possible words that could have occurred in the same context. We found that this was indeed the case. Processing of the target word was facilitated when the other possible words were semantically similar to the target word in comparison with when the other possible words were less semantically similar. We also found expected effects of predictability and word length on reading time. However, we found an interaction between frequency and predictability, rather than the independent effects of frequency and predictability reported in previous literature.

By showing that factors other than word predictability contribute to the cognitive effort required to process the word, we demonstrate the inadequacy of models that base their predictions solely on word probability. Probabilistic models of language comprehension vary in the extent to which they can be considered probability only models. Levy (2008) clearly intends the probability of a word to be the exclusive predictor of cognitive difficulty (this is discussed extensively in his paper in Section 2.3 *surprisal as a*

[7] In a separate analysis where interactions were not included, we found significant main effects of all four of our predictors – frequency, predictability, similarity, and length. However, this main effect of frequency (when interactions are not included) has occurred inconsistently in other similar experiments performed by the authors. Predictability, similarity, and length effects have occurred consistently across experiments.

causal bottleneck). Other papers have also modeled cognitive effort using only probability (e.g., Hale, 2001; Padó et al., 2009) and can implicitly be considered to be probability only models. Alternatively, other authors have explicitly included other factors along with probability in modeling language comprehension (e.g., Boston, Hale, Vasishth, & Kliegl, 2011; Demberg & Keller, 2008).

We are not the first authors to show independent effects of predictability and other factors on processing time. Ashby et al. (2005), Dambacher et al. (2006), Hand et al. (2010), and Rayner et al. (2004) found independent effects of word frequency and word predictability, while Boston et al. (2011) found independent effects of retrieval and predictability. It is difficult to determine how much of a challenge each of these findings poses to the notion of probability as a model of the comprehension process, given that various authors seem to view the role of probability in comprehension in different ways. We presume that all of the authors who have used predictability to model processing difficulty were intending to model only a portion of the entire set of processes involved in determining the reading times of words in sentences. For instance, they may have intended to model only the portion of comprehension related to integrating the knowledge gained from the new word into the representations they had constructed for the utterance/discourse up to that point. If so, we assume that they would model factors such as low level lexical access, eye-movements, and button-pressing separately. To the extent that the effects of a factor such as word frequency can be attributed to low level processes, the independent effects of the factor may not pose a challenge to the notion of predictability as a model of higher level processes of comprehension. To the extent that a factor such as semantic similarity may be affecting the same processes as word predictability, it provides a more direct challenge to the notion of word probability as a complete model of comprehension.

There are a variety of possible causes for the semantic similarity effects that we observed, two of which we will outline here – the *spreading activation* account and the *independent activation* account. One possible explanation for why words that do not occur affect the processing of words that do occur can be found in a spreading activation model (e.g., McRae et al., 1997). In the same manner that one would expect in any probability-only model of processing, the representations relevant for comprehending different possible upcoming words are initially activated in proportion to the degree of expectation for each of those words. This pre-activation of words before they are encountered during processing is consistent with ERP evidence reported by Delong et al. (2005). If this sort of probability driven activation were the only contribution to processing delays, then we would expect measures of cognitive effort such as reading times to align perfectly with measures of probability. However, we suggest that in addition to the initial probability-driven activation of the representations needed to process upcoming material, there is additional activation (or inhibition) of representations that takes place based on how similar those representations are to other representations that are also being activated. One possible way in which this activation spreading takes place is via shared semantic features of the words, as suggested in McRae et al. (1997). More generally, the mechanisms that we suggest are involved are likely to be the same as those involved in phenomena such as semantic priming. In our specific example of the processing of the sentence starting with *the aboriginal man jabbed the angry lion with*, we suggest that different possible instruments such as *spear*, *sword*, *machete*, and *rock* are initially activated in proportion to their respective probabilities. However, because *spear*, *sword*, and *machete* share certain semantic similarities (e.g., all are sharp pointy objects, all have handles), their representations become more strongly activated than probability alone would predict, while the representation of *rock* would ultimately be less strongly activated than probability alone would predict. If spreading activation is the cause of the semantic similarity effects that we observe, then our results serve to highlight the fact that purely probabilistic models do not seem to be able to handle the types of phenomena that are typically attributed to spreading activation.

The independent activation account provides an alternate explanation for our observations of semantic similarity effects. Upon hearing a phrase such as *the aboriginal man jabbed the angry lion with*, comprehenders may have expectations for specific words, such as *spear* or *sword*, but they may also have expectations based on semantic features such as *sharp pointy object* or *has a handle*. One possibility is that these expectations do not have independent effects on comprehension processes. While the relationship between featural expectations and word level expectations has not been extensively discussed in most work on probabilistic models, models which base their predictions only on word probabilities (Levy, 2008, explicitly bases predictions on word probabilities only; models such as Hale, 2001, and Padó et al., 2009, implicitly base their predictions on word probabilities only) inherently assume that expectations at different levels do not have independent effects on comprehension. In a word probability only view, knowing the likelihood of a sharp pointy object (based on expectations for that feature) is equivalent to knowing the likelihoods of each of the individual words that share the feature *sharp pointy objec*t. In other words, expectations at the level of semantic features do not make a contribution to the comprehension process that is independent of expectations at the word level. When processing the word *spear*, knowing the likelihood of *spear* is all you need to know about the likelihood of sharp pointy objects.

Levy (2008), in his discussion of surprisal as a causal bottleneck, views this as a benefit of word probability only models, since one only needs to know the probability of each word, and not the details of the underlying representations, to make predictions. In Levy's account, processing effort associated with the low probability of *fell* following *the horse raced past the barn*... represents the change in probabilities of all of the representations involved in processing the sentence (ranging from the probabilities of a main verb interpretation or a reduced relative clause interpretation, to the probability that the next word would be *fell*, to the probability that the next word would be an action verb), and thus the changes in the levels of activation of all of these representations.

Within an independent activation account, our results suggest that expectations at different levels of representation can have independent effects on comprehension. Given the angry lion context, a comprehender may have a weak expectation for the word *machete*, but have a strong expectation for a sharp pointy object. In this view, the process of activating representations during comprehension does not necessarily involve the spreading of activation between the representation for *sharp pointy object* and *machete*, so that machete ends up with a higher level of activation than would be expected based on its probability (as in the spreading activation account). It simply requires that the semantic feature sharp pointy object has an independent effect on processing, so that the higher activation of for *sharp pointy object* makes up for the lower activation of *machete*.

While probabilistic models have been remarkably successful in modeling language comprehension, our results point to possible limitations of probability as a model of cognitive effort. If the spreading activation account of our semantic similarity findings is correct, then probabilistic models at best serve as a starting point for more complex models of spreading activation. If the independent activation account is correct, then the number of individual probabilities needed to model cognitive effort becomes quite large – suggesting simplicity may no longer be one of the appeals of probabilistic models. Additionally, probabilities are by definition normalized, and always sum to one, while levels of activation in the brain[8] do not necessarily have this restriction. This suggests that the relationship between probability and activation may not be as simple as probabilistic models tend to suggest.

Whether one chooses to explain our results via a spreading activation account or an independent activation account, they suggest that probabilistic models cannot ignore the issue of underlying representations. In a spreading activation view, models of comprehension need to incorporate information about the properties of the complete set of words being anticipated and their relationships with the target word that actually occurs, in order to appropriately model the effects of spreading activation. Modeling the processing of the word *machete* requires knowing about features such as *sharp pointy object* and how likely the word *spear* was. Within the independent activation account, models of comprehension also need to take the nature of the underlying representations into account. Modeling the processing of the word *machete* requires knowing how likely *machete* was, but also how likely *sharp pointy object* was, and how these expectations combine to affect the overall level of effort needed to process *machete*. Either way, a complete model of comprehension cannot ignore the nature of the mental representations called upon during comprehension.

Acknowledgements

We wish to acknowledge the help of the members of the Psycholinguistics Lab and the Computational Psycholinguistics Lab at the University at Buffalo. We also wish to acknowledge Stephani Foraker for helpful feedback.

Appendix A

Experimental stimuli from the reading-time study. Target instrument fillers are italicized, with the highly likely instrument listed first, and the less likely instrument following the slash.

1. Jen's father built the garden shed with a hammer/a drill last Saturday.
2. The gladiator jabbed the African tiger with a sword/a spike in the Colosseum.
3. The housewife covered the ugly table with a tablecloth/a towel before the guests arrived.
4. The malicious neighbor punctured the truck's tire with a knife/a switchblade late last night.
5. The cleaning lady wiped the shower stall with a sponge/a squeegee this afternoon.
6. The FBI interrogator tied up the suspected terrorist with a rope/a cord and took him to jail.
7. The gardener pruned the tree limbs with scissors/a blade last week.
8. Mark's son scrubbed the tire rims with a sponge/a toothbrush to remove the dirt.
9. The child cored some green apples with a knife/a peeler before eating them.
10. The bartender flavored the client's martini with an olive/vermouth because he liked it.
11. The biology students dissected the fetal pigs with a scalpel/forceps for a lab quiz.
12. The farmer prodded the bales of hay with a pitchfork/a foot in the barn.
13. The park ranger marked the new trails with a sign/stones last month.
14. The chef stirred the savory stew with a spoon/a whisk in the kitchen.
15. The young man waxed his new car with a sponge/a washcloth to make it shine.
16. The cook sliced the giant pumpkin with a knife/a blade before baking it.
17. The firefighter smothered the kitchen fire with an extinguisher/a lid after he'd arrived.
18. The gardener dug up the flowering shrubs with a shovel/a tiller and planted them nearby.
19. The carpenter chopped the tree trunk with an axe/a hatchet in the woods.
20. The hair stylist trimmed the man's mustache with scissors/shears for a more groomed look.
21. The old man honed the dull knife with a sharpener/a grinder before he used it.
22. The apprentice carved the wooden statue with a knife/a chainsaw for an exhibit.
23. The traitor stabbed the old king with a knife/a machete around midnight.

[8] In artificial neural networks, levels of activation of output nodes may or may not be normalized, depending on the details of the implementation.

24. The army surgeon amputated the soldier's leg with a saw/a machete to save his life.
25. The electrician cut the long wire with scissors/clippers because it was too long.
26. The teacher underlined the students' errors with a pen/a crayon after the exam.
27. The busboy swept the messy floor with a broom/a duster after the restaurant closed.
28. The angry spouse hit the car windshield with a bat/a fist in the street.
29. John's father assembled the computer desk with a screwdriver/blots last Sunday.
30. The aborigine attacked the angry lion with a knife/a club in the field.
31. The knight killed the sleeping dragon with a sword/a spear before dawn.
32. The vandal destroyed the clay statue with a rock/a mallet to receive attention.
33. The child popped the big balloon with a pin/a screwdriver at the fair.
34. Ted's mother dried the baby's hands with a towel/a rag after washing them.
35. The waiter ruined the diner's dress with wine/mustard by accident.
36. The Mafia tortured the police informant with a knife/a chain in an abandoned building.
37. The guerrilla blocked the railroad tracks with a car/barricades yesterday afternoon.
38. The bully damaged his victim's car with a bat/a crowbar to intimidate him.
39. Toni's grandmother flipped the strawberry crêpe with a spatula/tongs before it burnt.
40. The forensic scientist examined the dead body with a scalpel/a flashlight to look for the cause of death.
41. The boy killed the huge cockroach with a shoe/a stone upon seeing it.
42. The thief opened the bank's safe with a pick/a sledgehammer last night.
43. The boy broke the wooden chair with a hammer/a crowbar before throwing it away.
44. The policemen subdued the violent protesters with guns/batons because they were so loud.
45. The students washed the dirty cars with soap/washcloths for a fundraiser.
46. The children destroyed the sand castles with water/sticks after sunset.
47. Some volunteers assisted the homeless Haitians with food/donations during the month of January.
48. The street vendor cracked the fresh coconut with a knife/a machete and sampled it.
49. The kids shattered the picture window with a rock/bricks and ran away.
50. The detective examined the crime scene with gloves/powder for several hours.
51. The workers moved the concrete panels with a truck/a dolly before lunch.
52. The terrorists attacked the villagers' houses with bombs/explosives last night.
53. The zoo assistant bathed the baby monkeys with soap/a towel early in the morning.
54. The hunter spotted the baby deer with binoculars/a lens in the forest.
55. The soldiers frightened the fleeing villagers with guns/flares after taking over the village.
56. The teenagers damaged the building's murals with paint/eggs last night.

References

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. (2010). languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics". R package version 1.0. <http://CRAN.R-project.org/package=languageR>.

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. (2010). lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-33. <http://CRAN.R-project.org/package=lme4>.

Binder, K. S., Duffy, S. A., & Rayner, K. (2001). The effects of thematic fit and discourse context on syntactic ambiguity resolution. *Journal of Memory and Language, 44*(2), 297–324.

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.

Boston, M. F., Hale, J. T., Vasishth, S., & Kliegl, R. (2011). Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes, 26*(3), 301–349.

Burnard, L. (1995). *Users reference guide for the British National Corpus*. Oxford: Oxford University Computing Services.

Dambacher, M., Kliegl, R., Hofmann, M., & Jacobs, A. M. (2006). Frequency and predictability effects on event-related potentials during reading. *Brain Research, 1084*(1), 89–103.

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 preactivation during language comprehension inferred from electrical brain activity. *Nature Neuroscience, 8*(8), 1117–1121.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition, 109*(2), 193–210.

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language, 41*(4), 469–495.

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.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

Hale, J. (2001). A probabilistic early parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Hand, C. J., Miellet, S., O'Donnell, P. J., & Sereno, S. C. (2010). The frequency–predictability interaction in reading: It depends where you're coming from. *Journal of Experimental Psychology: Human Perception and Performance, 36*(5), 1294–1313.

Hare, M., McRae, K., & Elman, J. L. (2004). Admitting that admitting verb sense into corpus analyses makes sense. *Language & Cognitive Processes, 19*(2), 181–224.

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.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology, 61*(1), 23–62.

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.

Jurafsky, D. (2003). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic linguistics* (pp. 39–96). Cambridge, MA: MIT Press.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87*(4), 329–354.

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory & Language, 49*(1), 133–156.

Kennedy, A., & Pidcock, B. (1981). Eye movements and variations in reading time. *Psychological Research, 43*(1), 69–79.

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.

Koenig, J.-P., Mauner, G., Bienvenue, B., & Conklin, K. (2008). What with? The anatomy of a (proto)-role. *Journal of Semantics, 25*, 175–220.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*, 1126–1177.

Mauner, G., Tanenhaus, M. K., & Carlson, G. N. (1995). Implicit arguments in sentence processing. *Journal of Memory and Language, 34*(3), 357–382.

McRae, K., de Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General, 126*(2), 99–130.

McRae, K., Ferretti, T. R., & Amyote, L. (1997). Thematic roles as verb-specific concepts. *Language and Cognitive Processes, 12*(2 & 3), 137–176.

Mikk, Y. A. (1972). Factors determining the reading time of words in a context. *Voprosy Psychologii, 18*(3), 125–128.

Miller, G. A., Newman, E. B., & Friedman, E. A. (1958). Length–frequency statistics for written English. *Information and Control, 1*(4), 370–389.

Obleser, J., & Kotz, S. A. (2010). Expectancy constraints in degraded speech modulate the language comprehension network. *Cerebral Cortex, 20*(3), 633–640.

Padó, U., Crocker, M., & Keller, F. (2009). A probabilistic model of semantic plausibility in sentence processing. *Cognitive Science, 33*(5), 794–838.

Pollatsek, A., Reichle, E. D., & Rayner, K. (2006). Tests of the E-Z reader model: Exploring the interface between cognition and eye-movement control. *Cognitive Psychology, 52*(1), 1–56.

R Development Core Team (2010). R: A Language and environment for statistical computing: R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. <http://www.R-project.org>.

Raney, G. E., & Rayner, K. (1995). Word frequency effects and eye movements during two readings of a text. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 49*(2), 151–173.

Rayner, K., Ashby, J., Pollatsek, A., & Reichle, E. D. (2004). The effects of frequency and predictability on eye fixations in reading: Implications for the E-Z reader model. *Journal of Experimental Psychology: Human Perception and Performance, 30*(4), 720–732.

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.

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The E-Z Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences, 26*(4), 445–526.

Spivey-Knowlton, M., & Sedivy, J. C. (1995). Resolving attachment ambiguities with multiple constraints. *Cognition, 55*(3), 227–267.

Staub, A. (2011). Word recognition and syntactic attachment in reading: Evidence for a staged architecture. *Journal of Experimental Psychology: General, 140*(3), 407–433.

Staub, A., White, S. J., Drieghe, D., Hollway, E. C., & Rayner, K. (2010). Distributional effects of word frequency on eye fixation durations. *Journal of Experimental Psychology: Human Perception and Performance, 36*(5), 1280–1293.

Tanenhaus, M. K., Carlson, G., & Trueswell, J. C. (1989). The role of thematic structures in interpretation and parsing. *Language & Cognitive Processes Special Issue: Parsing and interpretation, 4*(3-4), SI211–SI234.

Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. M. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory & Language, 33*(3), 285–318.

Trueswell, J. C., Tanenhaus, M. K., & Kello, C. (1993). Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory, Cognition, 19*(3), 528–553.

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*(3), 443–467.

Zipf, G. K. (1935). *The psycho-biology of language: An introduction to dynamic philology*. Boston: Houghton Mifflin Company.