The Plausibility of Semantic Properties Generated by a Distributional Model: Evidence from a Visual World Experiment

Diego Frassinelli (d.frassinelli@sms.ed.ac.uk)
Frank Keller (keller@inf.ed.ac.uk)
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, UK

Abstract

Distributional models of semantics are a popular way of capturing the similarity between words or concepts. More recently, such models have also been used to generate properties associated with a concept; model-generated properties are typically compared against collections of semantic feature norms. In the present paper, we propose a novel way of testing the plausibility of the properties generated by a distributional model using data from a visual world experiment. We show that model-generated properties, when embedded in a sentential context, bias participants' expectations towards a semantically associated target word in real time. This effect is absent in a neutral context that contains no relevant properties.

Keywords: Distributional models of semantics; concepts and properties; context effects; eye movements; visual world.

Introduction

The representation of semantic concepts has been the subject of an intense debate over the last few decades (Murphy, 2002). An emerging consensus is that the internal structure of a concept can be represented as a set of semantic properties (Garrard, Lambon Ralph, Hodges, & Patterson, 2001; Baroni & Lenci, 2008). These properties can be accessed in the form of semantic feature norms elicited from experimental participants (McRae, Cree, Seidenberg, & McNorgan, 2005). In the computational modeling literature, this idea has been taken up by distributional models of semantics. Such models have traditionally been used to compare the similarity of words or concepts. However, recently, a distributional model has been proposed that is able to generate properties associated with a concept (Baroni, Murphy, Barbu, & Poesio, 2010). These properties are computed based on corpus data, and have been shown to overlap with those generated in feature elicitation experiments. Distributional models can therefore be claimed to provide a cognitively plausible representation of concepts in terms of semantic properties.

In the present paper, we propose a novel way of testing this claim using the visual world paradigm (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995), which allows the study of conceptual processing in real time. We embed the properties generated by Baroni et al.'s model for a given target word into a sentential context. If the model-generated properties are cognitively plausible, then they should bias participants' expectations towards a target word, compared to a competitor word not associated with the properties. As a baseline, we also embed the target and competitor in a neutral context; the contextual expectation effect should be absent in this case.

Background

The idea of testing the predictions of distributional models using the visual world paradigm goes back to Huettig, Quinlan, McDonald, and Altmann (2006). They were interested in validating the semantic similarity measures generated by two distributional models: Latent Semantic Analysis (LSA; Landauer & Dumais, 1997) and Contextual Similarity (McDonald, 2000). Huettig et al. (2006) demonstrated that the similarity scores generated by both models are significantly correlated with fixation probabilities in a visual world experiment.

Huettig et al. used a list of 26 target/competitor pairs of semantically related but not strongly associated words. In every pair, one of the words corresponded to a target object depicted in a visual scene (the target word); the other one (the competitor word) was semantically related to the depicted object. For every pair of words, a spoken sentence was recorded that contained either the target or the competitor. Huettig et al. focused on the effect of hearing the target vs. the competitor as critical word. For this reason, the context sentences they used were neutral, providing background information that did not bias the participants towards either the target or the competitor. One of their contexts is given in (1) as an example.

(1) At first, the man laughed loudly, but then he saw the elephant (target)/alligator (competitor) and understood that it was dangerous.

The crucial manipulation in our experiment, however, concerns the context sentence. We run Huettig et al.'s neutral context as a baseline condition, but we add two context conditions: a context containing properties associated with the target, and a context containing properties associated with the competitor. These context sentences were constructed using three properties produced by the distributional model Strudel (Structured Dimension Extraction and Labeling; Baroni et al., 2010). Strudel is a model trained on the lemmatized and part-of-speech tagged version of Ukwac, an English corpus of two billion tokens extracted from the Web (Baroni, Bernardini, Ferraresi, & Zanchetta, 2009).

Strudel shares with other distributional models the assumption that it is possible to represent the meaning of a word in terms of other words that frequently appear in its linguistic context. Unlike traditional distributional approaches, Strudel describes concepts not only in terms of their most frequent

context words, but can also represent a word's internal structure in terms of semantic properties (e.g., visual features, the functions of an artefact). The output of Strudel is a list of properties linked to the corresponding concept through a pattern describing the relation between the concept and the property. An example is the relation *elephant_in_jungle*, in which the concept *elephant* is related to the property *jungle* via the pattern *in*. The set of properties for each concept is computed based on the number of co-occurrences in the corpus, taking into account the number of relevant patterns. The properties that Strudel generates this way are cognitively plausible in the sense that they overlap with human-generated feature norms such as the McRae et al. (2005) norms, as Baroni et al. (2010) demonstrate.

Experiment

This experiment had two main goals. Firstly, we wanted to test Strudel's ability to produce semantic properties for concepts. We evaluated this by using the properties to create sentential contexts, which we predict should bias participants towards the target concept. Secondly, we wanted to establish the effect that such contexts have on the processing of the target concept.

Huettig et al. used a neutral context and found that participants are more likely to fixate a target object when they hear its name, but they also show an increased fixation probability for the name of a semantically associated object. We expect this effect to be modulated by context. More specifically, the processing of properties associated with the target should build up an expectation for the target, and as a consequence, there should be more fixations on the target object when the target word is spoken, compared to the neutral context condition. This effect should be attenuated for the competitor, which is distinct from the target, but semantically related (as in Huettig et al.'s design).

Method

Materials The visual world paradigm requires both visual and linguistic stimuli. We used the same visual scenes as Huettig et al. Each scene contained black and white line drawings of the target object and three distractors; the pictures were extracted from the Snodgrass and Vanderwart (1980) collection. Huettig et al. removed phonological competitors and matched the pictures according to naming and image agreement, familiarity, visual complexity, and word frequency of the correspondent noun. Moreover, they tested the visual similarity between pictures. In our experiment, we used the same scenes used in the original experiment: this allowed us to skip the norming process.

We used the same linguistic materials as Huettig et al. for the neutral context condition. We added to this two context conditions: one for the target concept, and one for the competitor. For each of the 52 concepts (26 competitor/target pairs) in the Huettig et al. materials, we extracted from the output of Strudel the first 20 semantic properties

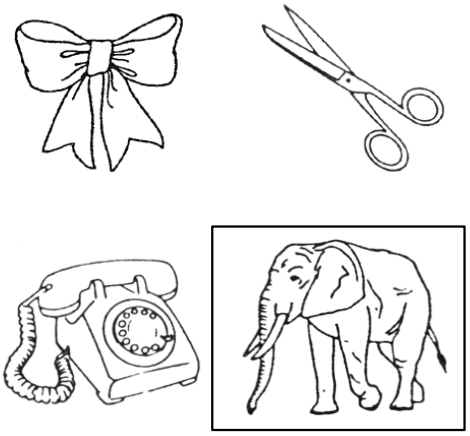


Figure 1: Example scene for the pair elephant (target)/alligator (competitor) in the experiment. The box highlights the target object (not shown to participants).

(nouns, verbs, and adjectives) ordered according to their log-likelihood ratio. We constructed a context sentence for each concept using three of these properties (excluding those associated with words that are part of the same target and competitor pair).

The context sentences had a standard pattern: a temporal subordinate clause introducing the situation followed by the main clause. The target concept is embedded at the end of the main clause and followed by an adverb (which serves as a spill-over region for the analysis). As an example, Figure 1 depicts the scene associated with the pair elephant (target)/alligator (competitor). The sentences associated with this scene are:

(2) `Neutral Context`: At first, the man laughed loudly, but then he saw the **elephant** and understood that it was dangerous.

(3) `Target Context`: While the man was crossing the *jungle*, he saw a *poacher capturing* an **elephant** ferociously.

(4) `Competitor Context`: While the man was crossing the *swamp*, he saw a hippo *attacking* a *gigantic* **elephant** ferociously.

The critical word is given in **bold**; the properties are in *italics*. For every sentence there was also a counterpart that included the competitor word (in this case **alligator**), resulting in six conditions in total.

The quality of the materials was evaluated in two norming studies performed using Amazon Mechanical Turk. In a sentence plausibility judgment task, 33 native English speakers rated the sentences on a scale from 1 (completely implausible) to 7 (completely plausible). The mean rating for the con-

cept in the sentence with the corresponding properties was 5.67 ($SD = 0.63$) and in the opposite sentence, it was 4.70 ($SD = 1.07$); the opposite sentences were created by swapping the critical words across conditions (target for competitor and vice versa). An Anova showed no main effects, but a significant interaction of concept (target or competitor) and sentence (target or competitor) ($F_1(1,35) = 27.86, p < .001$; $F_2(1,32) = 53.81, p < .001$).

In a sentence completion task, we removed the critical words from the sentences and asked 21 participants to complete each of the 52 sentences (two groups of 36 sentences) by typing the most plausible noun. After a process of synonym reduction, we counted the number of occurrences for each word. Good sentences had to elicit primarily the nouns they were associated with and only a small percentage of competitor or unrelated words.

The combination of these two norming studies was used to ensure that a given context was sufficiently associated with the target word, and not with the competitor word. Based on the norming data, we excluded eight pairs of concepts: these were cases in which Strudel had produced properties for a different sense of the word than the one in the Huettig et al. materials, as well as cases in which the target sentences were too different from the competitor ones so that the properties could not be plausibly swapped.

The sentence materials were recorded by a native English speaker at a normal speech rate for presentation in the experiment.

Procedure The entire experiment included 108 sentences: 18 word pairs (36 words in total) embedded in a neutral context and two biasing contexts. We rotated the position of the four objects on the screen to control for order or position effects. In total we therefore obtained 432 distinct items that we split in 24 lists of 18 items. The distribution of items across lists was based on a Latin square design, ensuring that each list included exactly one word from each target/competitor pair. Twenty-five filler items were added and a random presentation order generated for each list.

Twenty-four native English speakers from the University of Edinburgh were paid five pounds for taking part in the experiment. Each participant saw the items of one of the 24 lists, randomly interspersed with nine yes/no questions about the sentence or the scene. The questions were there to ensure that participants paid attention throughout the experiment.

Participants were seated in front of a 21" multi-scan monitor with a resolution of 1024 x 768 pixels and their eye movements were recorded using an EyeLink II head-mounted eye-tracker with a sampling rate of 500 Hz. Only the dominant eye was tracked. At the beginning of the experiment and after every ten trials, the eye-tracker was recalibrated using a nine-point randomized calibration. Before each trial, drift correction was performed. At the beginning of each trial the scene appeared on the screen, and the sentence began to play at the same time; the scene disappeared after 1500 ms after the end of the sentence. The experiment was explained using written instructions and preceded by practice trials. The instructions asked participants to listen carefully to the sentences and look wherever they wanted on the screen. The experiment lasted approximately 30 minutes.

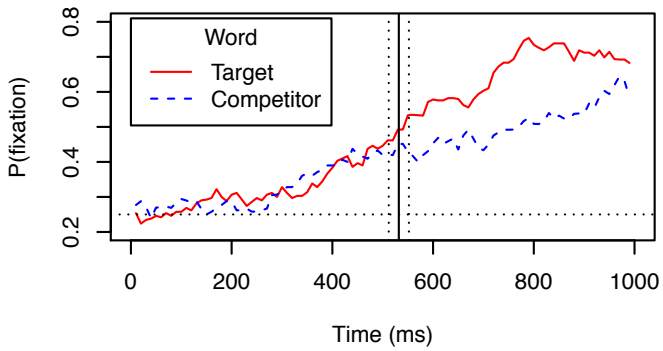
Results and Discussion

Fixation Probabilities Our analysis is based on the fixations on the target object compared to the fixations on the three distractor objects on the display. We excluded out-of-screen fixations and blinks from the analysis. Figure 2 plots the probability of fixating the target object across the three context conditions. The neutral context condition used the sentences of Huettig et al.; the target and competitor conditions used the contextually biased sentences produced based on the Strudel properties. In each plot, 0 ms corresponds to the acoustic onset of the critical word; our analysis takes into account the first 1000 ms after this onset. The vertical line shows the average offset of the critical words, with confidence intervals. The horizontal line at .25 indicates the probability of randomly fixating one of the four objects.

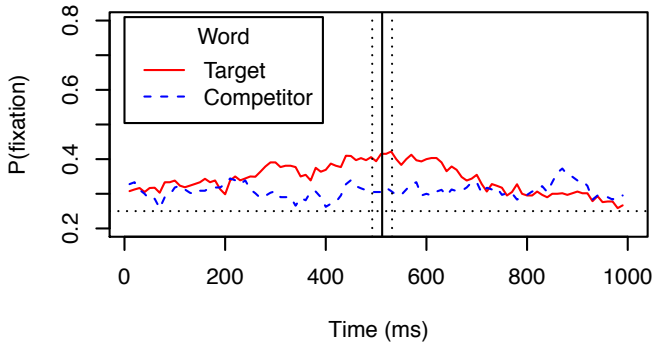
An inspection of the plots reveals a broadly similar trend across the three context conditions. The critical words require some time before they are recognized, which means that the fixation probabilities for the target and the competitor words take between 200 and 500 ms before they diverge. After that, we observe an increase in fixations to the target word compared to the competitor. The point of divergence is about 200 ms later in the neutral context; a semantically related context seems to aid the recognition of the critical word and triggers early fixations to the corresponding object. (Bear in mind that the competitor context is also semantically related to the target, as our norming studies showed.)

In the neutral context condition (Figure 2(a)), we observe a steady increase in fixation probability for both the target and the competitor word, which start to diverge at the offset of the critical word (this is presumably the point at which the critical word has been recognized by the participants). From that point on, we see more fixations on the target than on the competitor. This is in line with what Huettig et al. (2006) found: a competitor word triggers fixations to a semantically related target object, but less fixations than the target word corresponding to the target object. Our neutral context condition therefore provides a replication of Huettig et al.'s results. (The original paper also showed that the difference in fixation probability between target and competitor correlates with their semantic similarity, but we will not test this claim.)

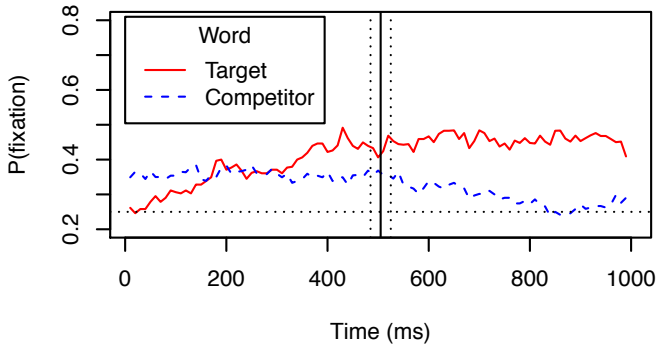
In the target context condition (Figure 2(b)), participants had heard a sentence containing properties of the depicted objects. Presumably this enables them to predict the target word with some accuracy (and our sentence completion study confirmed this). As the target is expected (and hence less interesting) at this point, we only observe a small increase of fixation probability for the target compared to the competitor, which starts early, at around 200 ms. This early start is consistent with the fact that participants are able to predict the critical



(a) Fixation probability in a *neutral context* sentence.



(b) Fixation prob. in a sentence associated with the *target* object.



(c) Fixation prob. in a sentence associated with the *competitor* object.

Figure 2: Fixation probabilities on the target object over time for the target (continuous red line) and competitor (dotted blue line) words. The onset of the critical word is at 0 ms. The vertical lines indicate the mean of the offset of the critical word with confidence interval. The horizontal line shows a probability of .25 (random baseline for four objects).

word in this condition based on the context sentence.

In the competitor context condition (Figure 2(c)), participants had heard a context sentence that is not directly associated with the depicted target object, but is instead associated with the semantically related competitor. In this case, hearing the target word (rather than the contextually appropriate competitor word) is unexpected, i.e., it generates interest and a larger increase in the number of fixations compared to the competitor word. This means that the two conditions diverge more than in the target context condition, and the divergence remains high for the whole period of analysis.

Inferential Statistics To statistically analyze the effect of the experimental manipulation on participants' fixations, we adopted the framework of linear mixed effect models (LME, Baayen, Davidson, & Bates, 2008). As suggested by Barr (2008), the dependent variable was the empirical logit of the fixation probability, calculated for each bin as:

$$\text{emplog} = \log\left(\frac{Y+.5}{N-Y+.5}\right)$$

where Y is the number of fixations on the target object and N is the total number of fixations in the bin.

Our model included the factor `Word` representing the nature of the critical word, coded as Competitor = $-.5$ and Target = .5. To determine context effects, we included two factors in contrast coding: the factor `Context` coded the difference between the neutral context = $-.5$ and the biasing context = .25 conditions; the factor `TargetSentence` differentiated the biasing context sentences further by distinguishing Competitor = $-.5$ and Target = .5. We have also included `Region` as a factor that indicates if the bin is in the critical region (coded as $-.5$) or in the region after the offset of the critical word (coded .5). Finally, the continuous predictor `Time` was discretized into 10 ms bins (range 1–100).

The random effects we included were `Participant` and `Item`, which were intercepts in the model. We also included random slopes for all the main effects (Word, Context, TargetSentence, Region, and Time). We used the model selection procedure of Coco and Keller (2012) to find the minimal model that best fits our data. Table 1 gives the coefficients and significance levels for the minimal model; main effects or interactions not listed in this table were not included in the minimal model by the selection procedure.

Effect of Context The factor `Context` compares fixation probabilities in the neutral context and in the biasing context, collapsing the competitor and the target context in the biasing context condition. We find a significant, positive main effect of this factor, suggesting that participants make more fixations on the target object in the biasing context condition. This is modulated by a negative interaction `Time:Context`, which indicates that fixation probability increases over time in the neutral context condition. This explains the upwards trend in Figure 2(a), but not in the biasing context conditions (Figures 2(b) and 2(c)).

While there is no general effect of whether the context is the competitor or the target sentence (no main effect of `TargetSentence`), we do find a significant positive interaction `Time:TargetSentence`. This confirms that there is a larger increase in fixations to the target object in the target context compared to the competitor context.

Table 1: Coefficients for the mixed effects model for the data in Figure 2.

Predictor	Coefficient
(Intercept)	-1.15^{***}
Time	0.17^{*}
Context	0.80^{*}
Time:Context	-0.64^{***}
TargetSentence	-0.47
Time:TargetSentence	0.11^{**}
Word	-0.06
Time:Word	0.18^{***}
Region	0.09
Region:Context	-0.41^{**}
Region:TargetSentence	-0.61^{***}
Word:TargetSentence	0.84^{***}
Time:Word:TargetSentence	-0.43^{***}
Region:Context:Word	-0.60^{***}

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

Effect of Critical Word While there is no main effect of `Word`, we find a significant positive interaction `Time:Word` that indicates that fixations on the target word increase more quickly than fixations on the competitor word. This is not surprising: when participants hear a word that matches the target object on the screen, they fixate this object more frequently (recall that the target object is depicted in all conditions, the competitor object is never on the screen).

Effect of Region There is no significant main effect of `Region`: whether the fixations are in the critical region (between the onset and the offset of the critical word) or in the post-critical region. However, we find a significant negative interaction `Region:Context`, suggesting that the neutral context sentences receive more fixations in the post-critical region compared to the biasing context sentences. This is compatible with the observation that context facilitates the processing of the critical word, which thus receives fixations earlier in the context condition.

The interaction `Region:TargetSentence` confirms that in the post-critical region participants fixate the target object more in the competitor context, presumably because it conflicts with their contextual expectations in this case. In the target context, however, contextual expectations and target object match, which means there is no reason to fixate the target object more frequently (compare Figures 2(b) and 2(c)).

Interaction of Context and Critical Word The most important interactions with respect to our experimental hypothesis are those involving `Context` and `Word` or `TargetSentence` and `Word`. These interactions demonstrate that context has an effect that is specific to the critical word.

We find a significant positive interaction `Word:TargetSentence`, which demonstrates that the target object receives more fixations when the target word occurs in the target context (rather than in the competitor context). This effect changes over time (significant negative interaction `Time:Word:TargetSentence`): the increase in fixations in the target word condition is larger in the competitor context than in the target context. For the competitor word, the opposite tendency emerges. This confirms the prediction that an expected critical word (i.e., one matching the context) is less interesting, and thus less likely to be fixated.

Finally, we can report a significant negative interaction `Region:Context:Word`, suggesting that the effect of `Word` in the neutral context condition is limited to the post-critical region, while in the biasing condition, it is stronger in the critical region. This corresponds to the observation that the fixation curves for the target and the competitor word diverge earlier for the biasing context conditions (see Figure 2).

General Discussion

First of all, our results replicate the findings of Huettig et al. (2006). In the neutral context condition, we find that participants fixate the target object both when they hear the critical word, and when they hear the semantically related competitor. While we observe less fixations on the target for the competitor word, Figure 2(a) clearly indicates that it is fixated more than chance (corresponding to a probability of .25).

However, the main purpose of our experiment was to test the ability of a distributional model of semantics to generate properties of concepts that are cognitively plausible. We therefore included two context conditions in our experiment, one in which the context sentence contained properties related to the target word, and one in which it contained properties related to the competitor word. In both cases, the properties were created by Strudel, a model of semantic representation.

When we compared these two biasing context conditions to the neutral context condition, we found two main effects. Firstly, a biasing context facilitates the processing of the critical word. Over time, the context builds up an expectation of the critical word, resulting in less fixations to the target object when it is contextually expected. This effect occurs for both types of biasing contexts, which is in line with the fact that the target and the competitor words were semantically related, which presumably implies that their properties are also semantically related. In the neutral context, in contrast, no expectations can be computed, as participants cannot guess the identity of the target word before its onset. The target object is unexpected and hence more interesting and receives more fixations, but these fixations appear later, once the recognition of the target word is complete.

Our second finding is that a biasing context makes it possible to anticipate the critical word: in a target context, we get more fixations to the target during the target word, com-

pared to the competitor word (Figure 2(b)). In the competitor context, we also initially find more fixations during the competitor word than during the target word. However, the pattern reverses after about 200 ms, presumably because of the match between the target word and the target object on the screen, which overrides the contextual expectation of the competitor word. Fixations for the target word remain high, however, compatible with a violation of contextual expectations (Figure 2(c)).

Both effects provide confirmation for the claim that we started out to prove: distributional models of semantics can generate properties that are cognitively plausible. They are plausible in the sense that they can be used to construct contexts that successfully bias participants towards a word that is compatible with the context. This contrasts with a neutral context, in which differences in fixation probabilities are purely driven by the semantic similarity with the target word. We therefore conclude that models like Strudel are a first step towards modeling linguistic context in a distributional way, which contrasts with the single-word approach that most of the distributional semantics literature has taken so far.

Acknowledgments

The work reported here was funded by the European Research Council under award number 203427 "Synchronous Linguistic and Visual Processing". We are grateful to Moreno I. Coco and Christoph Scheepers for their essential support and suggestions and to Desmond Elliott and David Matthews for the help in the production of the linguistic stimuli used during the experiment.

References

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. *Language Resources and Evaluation*, *43*(3), 209–231.

Baroni, M., & Lenci, A. (2008). Concepts and properties in word spaces. *Italian Journal of Linguistics*, *20*(1), 53–86.

Baroni, M., Murphy, B., Barbu, E., & Poesio, M. (2010). Strudel: A corpus-based semantic model based on properties and types. *Cognitive Science*, *34*(2), 222–254.

Barr, D. J. (2008). Analyzing visual world eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, *59*(4), 457–474.

Coco, M. I., & Keller, F. (2012). Scan patterns predict sentence production in the cross-modal processing of visual scenes. *Cognitive Science*, *in press*.

Garrard, P., Lambon Ralph, M. A., Hodges, J. R., & Patterson, K. (2001). Prototypicality, distinctiveness, and intercorrelation: Analyses of the semantic attributes of living and nonliving concepts. *Cognitive Neuropsychology*, *18*(2), 125–174.

Huettig, F., Quinlan, P. T., McDonald, S. A., & Altmann, G. T. M. (2006). Models of high-dimensional semantic space predict language-mediated eye movements in the visual world. *Acta Psychologica*, *121*(1), 65–80.

Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240.

McDonald, S. A. (2000). *Environmental determinants of lexical processing effort*. Unpublished doctoral dissertation, University of Edinburgh.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*(4), 547–59.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology. Human Learning and Memory*, *6*(2), 174–215.

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 1632–34.