Syntactic and Semantic Factors in Processing Difficulty: An Integrated Measure

Jeff Mitchell, Mirella Lapata, Vera Demberg and Frank Keller
University of Edinburgh
Edinburgh, United Kingdom
jeff.mitchell@ed.ac.uk, mlap@inf.ed.ac.uk,
v.demberg@ed.ac.uk, keller@inf.ed.ac.uk

Abstract

The analysis of reading times can provide insights into the processes that underlie language comprehension, with longer reading times indicating greater cognitive load. There is evidence that the language processor is highly predictive, such that prior context allows upcoming linguistic material to be anticipated. Previous work has investigated the contributions of semantic and syntactic contexts in isolation, essentially treating them as independent factors. In this paper we analyze reading times in terms of a single predictive measure which integrates a model of semantic composition with an incremental parser and a language model.

1 Introduction

Psycholinguists have long realized that language comprehension is highly *incremental*, with readers and listeners continuously extracting the meaning of utterances on a word-by-word basis. As soon as they encounter a word in a sentence, they integrate it as fully as possible into a representation of the sentence thus far (Marslen-Wilson 1973; Konieczny 2000; Tanenhaus et al. 1995; Sturt and Lombardo 2005). Recent research suggests that language comprehension can also be highly *predictive*, i.e., comprehenders are able to anticipate upcoming linguistic material. This is beneficial as it gives them more time to keep up with the input, and predictions can be used to compensate for problems with noise or ambiguity.

Two types of prediction have been observed in the literature. The first type is semantic prediction, as evidenced in semantic priming: a word that is preceded by a semantically related prime or a semantically congruous sentence fragment is processed faster (Stanovich and West 1981; van Berkum et al. 1999; Clifton et al. 2007). Another example is argument prediction: listeners are able to launch eye-movements to the predicted argument of a verb before having encountered it, e.g., they will fixate an edible object as soon as they hear the word *eat* (Altmann and Kamide 1999). The second type of prediction is syntactic prediction. Comprehenders are faster at naming words that are syntactically compatible with prior context, even when they bear no semantic relationship to the context (Wright and Garrett 1984). Another instance of syntactic prediction has been reported by Staub and Clifton (2006): following the word *either*, readers predict *or* and the complement that follows it, and process it faster compared to a control condition without *either*.

Thus, human language processing takes advantage of the constraints imposed by the preceding semantic and syntactic context to derive expectations about the upcoming input. Much recent work has focused on developing computational measures of these constraints and expectations. Again, the literature is split into syntactic and semantic models. Probably the best known measure of syntactic expectation is *surprisal* (Hale 2001) which can be coarsely defined as the negative log probability of word w_t given the preceding words, typically computed using a probabilistic context-free grammar.

Modeling work on semantic constraint focuses on the degree to which a word is related to its preceding context. Pynte et al. (2008) use Latent Semantic Analysis (LSA, Landauer and Dumais 1997) to assess the degree of contextual constraint exerted on a word by its context. In this framework, word meanings are represented as vectors in a high dimensional space and distance in this space is interpreted as an index of processing difficulty. Other work (McDonald and Brew 2004) models contextual constraint in information theoretic terms. The assumption is that words carry prior *semantic expectations* which are updated upon seeing the next word. Expectations are represented by a vector of probabilities which reflects the likely location in semantic space of the upcoming word.

The measures discussed above are typically computed automatically on real-language corpora using data-driven methods and their predictions are verified through analysis of eye-movements that people make while reading. Ample evidence

Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 196–206.
Uppsala, 2010.

(Rayner 1998) demonstrates that eye-movements are related to the moment-to-moment cognitive activities of readers. They also provide an accurate temporal record of the on-line processing of natural language, and through the analysis of eye-movement measurements (e.g., the amount of time spent looking at a word) can give insight into the processing difficulty involved in reading.

In this paper, we investigate a model of prediction that is incremental and takes into account syntactic as well as semantic constraint. The model essentially integrates the predictions of an incremental parser (Roark 2001) together with those of a semantic space model (Mitchell and Lapata 2009). The latter creates meaning representations *compositionally*, and therefore builds semantic expectations for word sequences (e.g., phrases, sentences, even documents) rather than isolated words. Some existing models of sentence processing integrate semantic information into a probabilistic parser (Narayanan and Jurafsky 2002; Padó et al. 2009); however, the semantic component of these models is limited to semantic role information, rather than attempting to build a full semantic representation for a sentence. Furthermore, the models of Narayanan and Jurafsky (2002) and Padó et al. (2009) do not explicitly model prediction, but rather focus on accounting for garden path effects. The proposed model simultaneously captures semantic and syntactic effects in a single measure which we empirically show is predictive of processing difficulty as manifested in eye-movements.

2 Models of Processing Difficulty

As described in Section 1, reading times provide an insight into the various cognitive activities that contribute to the overall processing difficulty involved in comprehending a written text. To quantify and understand the overall cognitive load associated with processing a word in context, we will break that load down into a sum of terms representing distinct computational costs (semantic and syntactic). For example, surprisal can be thought of as measuring the cost of dealing with unexpected input. When a word conforms to the language processor's expectations, surprisal is low, and the cognitive load associated with processing that input will also be low. In contrast, unexpected words will have a high surprisal and a high cognitive cost.

However, high-level syntactic and semantic factors are only one source of cognitive costs. A sizable proportion of the variance in reading times is accounted for by costs associated with low-level features of the stimuli, e.g.. relating to orthography and eye-movement control (Rayner 1998). In addition, there may also be costs associated with the integration of new input into an incremental representation. Dependency Locality Theory (DLT, Gibson 2000) is essentially a distance-based measure of the amount of processing effort required when the head of a phrase is integrated with its syntactic dependents. We do not consider integration costs here (as they have not been shown to correlate reliably with reading times; see Demberg and Keller 2008 for details) and instead focus on the costs associated with semantic and syntactic constraint and low-level features, which appear to make the most substantial contributions.

In the following subsections we describe the various features which contribute to the processing costs of a word in context. We begin by looking at the low-level costs and move on to consider the costs associated with syntactic and semantic constraint. For readers unfamiliar with the methodology involved in modeling eye-tracking data, we note that regression analysis (or the more general mixed effects models) is typically used to study the relationship between *dependent* and *independent* variables. The independent variables are the various costs of processing effort and the dependent variables are measurements of eye-movements, three of which are routinely used in the literature: *first fixation duration* (the duration of the first fixation on a word regardless of whether it is the first fixation on a word or the first of multiple fixations on the same word), *first pass duration*, also known as *gaze duration*, (the sum of all fixations made on a word prior to looking at another word), and *total reading time* (the sum of all fixations on a word including refixations after moving on to other words).

2.1 Low-level Costs

Low-level features include word frequency (more frequent words are read faster), word length (shorter words are read faster), and the position of the word in the sentence (later words are read faster). Oculomotor variables have also been found to influence reading times. These include previous fixation (indicating whether or not the previous word has been fixated), launch distance (how many characters intervene between the current fixation and the previous fixation), and landing position (which letter in the word the fixation landed on).

Information about the sequential context of a word can also influence reading times. Mc-

Donald and Shillcock (2003) show that forward and backward transitional probabilities are predictive of first fixation and first pass durations: the higher the transitional probability, the shorter the fixation time. Backward transitional probability is essentially the conditional probability of a word given its immediately preceding word, $P(w_k|w_{k-1})$. Analogously, forward probability is the conditional probability of the current word given the next word, $P(w_k|w_{k+1})$.

2.2 Syntactic Constraint

As mentioned earlier, surprisal (Hale 2001; Levy 2008) is one of the best known models of processing difficulty associated with syntactic constraint, and has been previously applied to the modeling of reading times (Demberg and Keller 2008; Ferrara Boston et al. 2008; Roark et al. 2009; Frank 2009). The basic idea is that the processing costs relating to the expectations of the language processor can be expressed in terms of the probabilities assigned by some form of language model to the input. These processing costs are assumed to arise from the change in the expectations of the language processor as new input arrives. If we express these expectations in terms of a distribution over all possible continuations of the input seen so far, then we can measure the magnitude of this change in terms of the Kullback-Leibler divergence of the old distribution to the updated distribution. This measure of processing cost for an input word, w_{k+1}, given the previous context, $w_1 \dots w_k$, can be expressed straightforwardly in terms of its conditional probability as:

$$S = -\log P(w_{k+1}|w_1 \dots w_k) \tag{1}$$

That is, the processing cost for a word decreases as its probability increases, with zero processing cost incurred for words which must appear in a given context, as these do not result in any change in the expectations of the language processor.

The original formulation of surprisal (Hale 2001) used a probabilistic parser to calculate these probabilities, as the emphasis was on the processing costs incurred when parsing structurally ambiguous garden path sentences.[1] Several variants of calculating surprisal have been developed in the literature since using different parsing strategies (e.g., left-to-right vs. top-down, PCFGs vs dependency parsing) and different degrees of lexicalization (see Roark et al. 2009 for an overview) . For instance, unlexicalized surprisal can be easily derived by substituting the words in Equation (1) with parts of speech (Demberg and Keller 2008). Surprisal could be also defined using a vanilla language model that does not take any structural or grammatical information into account (Frank 2009).

2.3 Semantic Constraint

Distributional models of meaning have been commonly used to quantify the semantic relation between a word and its context in computational studies of lexical processing. These models are based on the idea that words with similar meanings will be found in similar contexts. In putting this idea into practice, the meaning of a word is then represented as a vector in a high dimensional space, with the vector components relating to the strength on occurrence of that word in various types of context. Semantic similarities are then modeled in terms of geometric similarities within the space.

To give a concrete example, Latent Semantic Analysis (LSA, Landauer and Dumais 1997) creates a meaning representation for words by constructing a word-document co-occurrence matrix from a large collection of documents. Each row in the matrix represents a word, each column a document, and each entry the frequency with which the word appeared within that document. Because this matrix tends to be quite large it is often transformed via a singular value decomposition (Berry et al. 1995) into three component matrices: a matrix of word vectors, a matrix of document vectors, and a diagonal matrix containing singular values. Re-multiplying these matrices together using only the initial portions of each (corresponding to the use of a lower dimensional spatial representation) produces a tractable approximation to the original matrix. In this framework, the similarity between two words can be easily quantified, e.g., by measuring the cosine of the angle of the vectors representing them.

As LSA is one the best known semantic space models it comes as no surprise that it has been used to analyze semantic constraint. Pynte et al. (2008) measure the similarity between the next word and its preceding context under the assumption that high similarity indicates high semantic constraint (i.e., the word was expected) and analogously low similarity indicates low semantic constraint (i.e., the word was unexpected). They oper-

[1] While hearing a sentence like *The horse raced past the barn fell* (Bever 1970), English speakers are inclined to interpreted *horse* as the subject of *raced* expecting the sentence to end at the word *barn*. So upon hearing the word *fell* they are forced to revise their analysis of the sentence thus far and adopt a reduced relative reading.

ationalize preceding contexts in two ways, either as the word immediately preceding the next word as the sentence fragment preceding it. Sentence fragments are represented as the average of the words they contain independently of their order. The model takes into account only content words, function words are of little interest here as they can be found in any context.

Pynte et al. (2008) analyze reading times on the French part of the Dundee corpus (Kennedy and Pynte 2005) and find that word-level LSA similarities are predictive of first fixation and first pass durations, whereas sentence-level LSA is only predictive of first pass duration (i.e., for a measure that includes refixation). This latter finding is somewhat counterintuitive, one would expect longer contexts to have an immediate effect as they are presumably more constraining. One reason why sentence-level influences are only visible on first pass duration may be due to LSA itself, which is syntax-blind. Another reason relates to the way sentential context was modeled as vector addition (or averaging). The idea of averaging is not very attractive from a linguistic perspective as it blends the meanings of individual words together. Ideally, the combination of simple elements onto more complex ones must allow the construction of novel meanings which go beyond those of the individual elements (Pinker 1994).

The only other model of semantic constraint we are aware of is Incremental Contextual Distinctiveness (ICD, McDonald 2000; McDonald and Brew 2004). ICD assumes that words carry prior semantic expectations which are updated upon seeing the next word. Context is represented by a vector of probabilities which reflects the likely location in semantic space of the upcoming word. When the latter is observed, the prior expectation is updated using a Bayesian inference mechanism to reflect the newly arrived information. Like LSA, ICD is based on word co-occurrence vectors, however it does not employ singular value decomposition, and constructs a word-word rather than a word-document co-occurrence matrix. Although this model has been shown to successfully simulate single- and multiple-word priming (McDonald and Brew 2004), it failed to predict processing costs in the Embra eye-tracking corpus (McDonald and Shillcock 2003).

In this work we model semantic constraint using the representational framework put forward in Mitchell and Lapata (2008). Their aim is not so much to model processing difficulty, but to construct vector-based meaning representations that go beyond individual words. They introduce a general framework for studying vector *composition*, which they formulate as a function f of two vectors $\mathbf{u}$ and $\mathbf{v}$:

$$\mathbf{h} = f(\mathbf{u}, \mathbf{v}) \tag{2}$$

where $\mathbf{h}$ denotes the composition of $\mathbf{u}$ and $\mathbf{v}$. Different composition models arise, depending on how f is chosen. Assuming that $\mathbf{h}$ is a linear function of the Cartesian product of $\mathbf{u}$ and $\mathbf{v}$ allows to specify *additive* models which are by far the most common method of vector combination in the literature:

$$h_i = u_i + v_i \tag{3}$$

Alternatively, we can assume that $\mathbf{h}$ is a linear function of the tensor product of $\mathbf{u}$ and $\mathbf{v}$, and thus derive models based on *multiplication*:

$$h_i = u_i \cdot v_i \tag{4}$$

Mitchell and Lapata (2008) show that several additive and multiplicative models can be formulated under this framework, including the well-known tensor products (Smolensky 1990) and circular convolution (Plate 1995). Importantly, composition models are not defined with a specific semantic space in mind, they could easily be adapted to LSA, or simple co-occurrence vectors, or more sophisticated semantic representations (e.g., Griffiths et al. 2007), although admittedly some composition functions may be better suited for particular semantic spaces.

Composition models can be straightforwardly used as predictors of processing difficulty, again via measuring the cosine of the angle between a vector $\mathbf{w}$ representing the upcoming word and a vector $\mathbf{h}$ representing the words preceding it:

$$sim(\mathbf{w}, \mathbf{h}) = \frac{\mathbf{w} \cdot \mathbf{h}}{|\mathbf{w}||\mathbf{h}|} \tag{5}$$

where $\mathbf{h}$ is created compositionally, via some (additive or multiplicative) function f.

In this paper we evaluate additive and compositional models in their ability to capture semantic prediction. We also examine the influence of the underlying meaning representations by comparing a simple semantic space similar to McDonald (2000) against Latent Dirichlet Allocation (Blei et al. 2003; Griffiths et al. 2007). Specifically, the simpler space is based on word co-occurrence counts; it constructs the vector representing a given target word, t, by identifying all the tokens of t in a corpus and recording the counts of context words, c_i (within a specific window). The context words, c_i, are limited to a set of the n most common content words and each vector component is given by the ratio of the probability of a c_i given t to the overall probability of c_i.

$$v_i = \frac{p(c_i|t)}{p(c_i)} \tag{6}$$

Despite its simplicity, the above semantic space (and variants thereof) has been used to successfully simulate lexical priming (e.g., McDonald 2000), human judgments of semantic similarity (Bullinaria and Levy 2007), and synonymy tests (Padó and Lapata 2007) such as those included in the Test of English as Foreign Language (TOEFL).

LDA is a probabilistic topic model offering an alternative to spatial semantic representations. It is similar in spirit to LSA, it also operates on a word-document co-occurrence matrix and derives a reduced dimensionality description of words and documents. Whereas in LSA words are represented as points in a multi-dimensional space, LDA represents words using topics. Specifically, each document in a corpus is modeled as a distribution over K topics, which are themselves characterized as distribution over words. The individual words in a document are generated by repeatedly sampling a topic according to the topic distribution and then sampling a single word from the chosen topic. Under this framework, word meaning is represented as a probability distribution over a set of latent topics, essentially a vector whose dimensions correspond to topics and values to the probability of the word given these topics. Topic models have been recently gaining ground as a more structured representation of word meaning (Griffiths et al. 2007; Steyvers and Griffiths 2007). In contrast to more standard semantic space models where word senses are conflated into a single representation, topics have an intuitive correspondence to coarse-grained sense distinctions.

3 Integrating Semantic Constraint into Surprisal

The treatment of semantic and syntactic constraint in models of processing difficulty has been somewhat inconsistent. While surprisal is a theoretically well-motivated measure, formalizing the idea of linguistic processing being highly predictive in terms of probabilistic language models, the measurement of semantic constraint in terms of vector similarities lacks a clear motivation. Moreover, the two approaches, surprisal and similarity, produce mathematically different types of measures. Formally, it would be preferable to have a single approach to capturing constraint and the obvious solution is to derive some form of semantic surprisal rather than sticking with similarity. This can be achieved by turning a vector model of semantic similarity into a probabilistic language model.

There are in fact a number of approaches to deriving language models from distributional models of semantics (e.g., Bellegarda 2000; Coccaro and Jurafsky 1998; Gildea and Hofmann 1999). We focus here on the model of Mitchell and Lapata (2009) which tackles the issue of the composition of semantic vectors and also integrates the output of an incremental parser. The core of their model is based on the product of a trigram model $p(w_n|w_{n-2}^{n-1})$ and a semantic component $\Delta(w_n,h)$ which determines the factor by which this probability should be scaled up or down given the prior semantic context h:

$$p(w_n) = p(w_n|w_{n-2}^{n-1}) \cdot \Delta(w_n,h) \tag{7}$$

The factor $\Delta(w_n,h)$ is essentially based on a comparison between the vector representing the current word w_n and the vector representing the prior history h. Varying the method for constructing word vectors (e.g., using LDA or a simpler semantic space model) and for combining them into a representation of the prior context h (e.g., using additive or multiplicative functions) produces distinct models of semantic composition.

The calculation of Δ is then based on a weighted dot product of the vector representing the upcoming word w, with the vector representing the prior context h:

$$\Delta(w,h) = \sum_i w_i h_i p(c_i) \tag{8}$$

As shown in Equation (7) this semantic factor then modulates the trigram probabilities, to take account of the effect of the semantic content outside the n-gram window.

Mitchell and Lapata (2009) show that a combined semantic-trigram language model derived from this approach and trained on the Wall Street Journal outperforms a baseline trigram model in terms of perplexity on a held out set. They also linearly interpolate this semantic language model with the output of an incremental parser, which computes the following probability:

$$p(w|h) = \lambda p_1(w|h) + (1-\lambda)p_2(w|h) \tag{9}$$

where $p_1(w|h)$ is computed as in Equation (7) and $p_2(w|h)$ is computed by the parser. Their implementation uses Roark's (2001) top-down incremental parser which estimates the probability of the next word based upon the previous words of the sentence. These *prefix* probabilities are calculated from a grammar, by considering the likelihood of seeing the next word given the possible grammatical relations representing the prior context.

Equation (9) essentially defines a language model which combines semantic, syntactic and *n*-gram structure, and Mitchell and Lapata (2009) demonstrate that it improves further upon a semantic language model in terms of perplexity. We argue that the probabilities from this model give us a means to model the incrementally and predictivity of the language processor in a manner that integrates both syntactic and semantic constraints. Converting these probabilities to surprisal should result in a single measure of the processing cost associated with semantic and syntactic expectations.

4 Method

Data The models discussed in the previous section were evaluated against an eye-tracking corpus. Specifically, we used the English portion of the Dundee Corpus (Kennedy and Pynte 2005) which contains 20 texts taken from *The Independent* newspaper. The corpus consists of 51,502 tokens and 9,776 types in total. It is annotated with the eye-movement records of 10 English native speakers, who each read the whole corpus. The eye-tracking data was preprocessed following the methodology described in Demberg and Keller (2008). From this data, we computed total reading time for each word in the corpus. Our statistical analyses were based on actual reading times, and so we only included words that were not skipped. We also excluded words for which the previous word had been skipped, and words on which the normal left-to-right movement of gaze had been interrupted, i.e., by blinks, regressions, etc. Finally, because our focus is the influence of semantic context, we selected only content words whose prior sentential context contained at least two further content words. The resulting data set consisted of 53,704 data points, which is about 10% of the theoretically possible total.[2]

Model Implementation All elements of our model were trained on the BLLIP corpus, a collection of texts from the Wall Street Journal (years 1987–89). The training corpus consisted of 38,521,346 words. We used a development corpus of 50,006 words and a test corpus of similar size. All words were converted to lowercase and numbers were replaced with the symbol $\langle$num$\rangle$. A vocabulary of 20,000 words was chosen and the remaining tokens were replaced with $\langle$unk$\rangle$.

Following Mitchell and Lapata (2009), we constructed a simple semantic space based on co-occurrence statistics from the BLLIP training set. We used the 2,000 most frequent word types as contexts and a symmetric five word window. Vector components were defined as in Equation (6). We also trained the LDA model on BLLIP, using the Gibb's sampling procedure discussed in Griffiths et al. (2007). We experimented with different numbers of topics on the development set (from 10 to 1,000) and report results on the test set with 100 topics. In our experiments, the hyperparameter α was initialized to .5, and the β word probabilities were initialized randomly.

We integrated our compositional models with a trigram model which we also trained on BLLIP. The model was built using the SRILM toolkit (Stolcke 2002) with backoff and Kneser-Ney smoothing. As our incremental parser we used Roark's (2001) parser trained on sections 2–21 of the Penn Treebank containing 936,017 words. The parser produces prefix probabilities for each word of a sentence which we converted to conditional probabilities by dividing each current probability by the previous one.

Statistical Analysis The statistical analyses in this paper were carried out using linear mixed effects models (LME, Pinheiro and Bates 2000). The latter can be thought of as generalization of linear regression that allows the inclusion of random factors (such as participants or items) as well as fixed factors (e.g., word frequency). In our analyses, we treat participant as a random factor, which means that our models contain an intercept term for each participant, representing the individual differences in the rates at which they read.[3]

We evaluated the effect of adding a factor to a model by comparing the likelihoods of the models with and without that factor. If a χ^2 test on the

[2] The total of all words read by all subjects is 515,020. The pre-processing recommended by Demberg and Keller's (2008) results in a data sets containing 436,000 data points. Removing non-content words leaves 205,922 data points. It only makes sense to consider words that were actually fixated (the eye-tracking measures used are not defined on skipped words), which leaves 162,129 data points. Following Pynte et al. (2008), we require that the previous word was fixated, with 70,051 data points remaining. We exclude words on which the normal left to right movement of gaze had been interrupted, e.g., by blinks and regressions, which results in the final total to 53,704 data points.

[3] Other random factors that are appropriate for our analyses are word and sentence; however, due to the large number of instances for these factors (given that the Dundee corpus contains 51,502 tokens), we were not able to include them: the model fitting algorithm we used (implemented in the R package `lme4`) does not converge for such large models.

Factor	Coefficient
Intercept	$-.011$
Word Length	.264
Launch Distance	.109
Landing Position	.612
Word Frequency	$-.010$
Reading Time of Last Word	.151

Table 1: Coefficients of the baseline LME model for total reading time

Model	Composition	Coefficient
SSS	Additive	$-.03820^{***}$
	Multiplicative	$-.00895^{***}$
LDA	Additive	$-.02500^{***}$
	Multiplicative	$-.00262^{***}$

Table 2: Coefficients of LME models including simple semantic space (SSS) or Latent Dirichlet Allocation (LDA) as factors; $^{***}p < .001$

likelihood ratio is significant, then this indicates that the new factor significantly improves model fit. We also experimented with adding random slopes for participant to the model (in addition to the random intercept); however, this either led to non-convergence of the model fitting procedure, or failed to result in an increase in model fit according to the likelihood ratio test. Therefore, all models reported in the rest of this paper contain random intercept of participants as the sole random factor.

Rather than model raw reading times, we model times on the log scale. This is desirable for a number of reasons. Firstly, the raw reading times tend to have a skew distribution and taking logs produces something closer to normal, which is preferable for modeling. Secondly, the regression equation makes more sense on the log scale as the contribution of each term to raw reading time is multiplicative rather than additive. That is, $log(t) = \sum_i \beta_i x_i$ implies $t = \prod_i e^{\beta_i x_i}$. In particular, the intercept term for each participant now represents a multiplicative factor by which that participant is slower or faster.

5 Results

We computed separate mixed effects models for three dependent variables, namely first fixation duration, first pass duration, and total reading time. We report results for total times throughout, as the results of the other two dependent variables are broadly similar. Our strategy was to first construct a baseline model of low-level factors influencing reading time, and then to take the residuals from that model as the dependent variable in subsequent analyses. In this way we removed the effects of low-level factors before investigating the factors associated with syntactic and semantic constraint. This avoids problems with collinearity between low-level factors and the factors we are interested in (e.g., trigram probability is highly correlated with word frequency). The baseline model contained the factors word length, word frequency, launch distance, landing position, and the reading time for the last fixated word, and its parameter estimates are given in Table 1. To further reduce collinearity, we also centered all fixed factors, both in the baseline model, and in the models fitted on the residuals that we report in the following. Note that some intercorrelations remain between the factors, which we will discuss at the end of Section 5.

Before investigating whether an integrated model of semantic and syntactic constraint improves the goodness of fit over the baseline, we examined the influence of semantic constraint alone. This was necessary as compositional models have not been previously used to model processing difficulty. Besides, replicating Pynte et al.'s (2008) finding, we were also interested in assessing whether the underlying semantic representation (simple semantic space or LDA) and composition function (additive versus multiplicative) modulate reading times differentially.

We built an LME model that predicted the residual reading times of the baseline model using the similarity scores from our composition models as factors. We then carried out a χ^2 test on the likelihood ratio of a model only containing the random factor and the intercept, and a model also containing the semantic factor (cosine similarity). The addition of the semantic factor significantly improves model fit for both the simple semantic space and LDA. This result is observed for both additive and multiplicative composition functions. Our results are summarized in Table 2 which reports the coefficients of the four LME models fitted against the residuals of the baseline model, together with the p-values of the χ^2 test.

Before evaluating our integrated surprisal measure, we evaluated its components *individually* in order to tease their contributions apart. For example, it may be the case that syntactic surprisal is an overwhelmingly better predictor of reading time than semantic surprisal, however we would not be able to detect this by simply adding a factor based on Equation (9) to the baseline model. The

	Factor	SSS Coef	LDA Coef
	$-\log(p)$	$.00760^{***}$	$.00760^{***}$
Add	$-\log(\Delta)$	$.03810^{***}$	$.00622^{***}$
	$\log(\lambda+(1-\lambda)\frac{p_2}{p_1})$	$.00953^{***}$	$.00943^{***}$
Mult	$-\log(\Delta)$	$.01110^{***}$	$-.00033$
	$\log(\lambda+(1-\lambda)\frac{p_2}{p_1})$	$.00882^{***}$	$.00133$

Table 3: Coefficients of nested LME models with the components of SSS or LDA surprisal as factors; only the coefficient of the additional factor at each step are shown

integrated surprisal measure can be written as:

$$S = -\log(\lambda p_1 + (1-\lambda)p_2) \qquad (10)$$

Where p_2 is the incremental parser probability and p_1 is the product of the semantic component, Δ, and the trigram probability, p. This can be broken down into the sum of two terms:

$$S = -\log(p_1) - \log(\lambda + (1-\lambda)\frac{p_2}{p_1}) \qquad (11)$$

Since the first term, $-\log(p_1)$ is itself a product it can also be broken down further:

$$S = -\log(p) - \log(\Delta) - \log(\lambda + (1-\lambda)\frac{p_2}{p_1}) \quad (12)$$

Thus, to evaluate the contribution of the three components to the integrated surprisal measure we fitted nested LME models, i.e., we entered these terms one at a time into a mixed effects model and tested the significance of the improvement in model fit for each additional term.

We again start with an LME model that only contains the random factor and the intercept, with the residuals of the baseline models as the dependent variable. Considering the trigram model first, we find that adding this factor to the model gives a significant improvement in fit. Also adding the semantic component ($-\log(\Delta)$) improves fit further, both for additive and multiplicative composition functions using a simple semantic space. Finally, the addition of the parser probabilities ($\log(\lambda + (1-\lambda)\frac{p_2}{p_1})$) again improves model fit significantly. As far as LDA is concerned, the additive model significantly improves model fit, whereas the multiplicative one does not. These results mirror the findings of Mitchell and Lapata (2009), who report that a multiplicative composition function produced the lowest perplexity for the simple semantic space model, whereas an additive function gave the best perplexity for the LDA space. Table 3 lists the coefficients for the nested models for all four variants of our semantic constraint measure.

Model	Composition	Coefficient
SSS	Additive	$.00804^{***}$
	Multiplicative	$.00819^{***}$
LDA	Additive	$.00817^{***}$
	Multiplicative	$.00640^{***}$

Table 4: Coefficients of LME models with integrated surprisal measure (based on SSS or LDA) as factor

Finally, we built a separate LME model where we added the integrated surprisal measure (see Equation (9)) to the model only containing the random factor and the intercept (see Table 4). We did this separately for all four versions of the integrated surprisal measure (SSS, LDA; additive, multiplicative). We find that model fit improved significantly all versions of integrated surprisal.

One technical issue that remains to be discussed is collinearity, i.e., intercorrelations between the factors in a model. The presence of collinearity is problematic, as it can render the model fitting procedure unstable; it can also affect the significance of individual factors. As mentioned in Section 4 we used two techniques to reduce collinearity: residualizing and centering. Table 5 gives an overview of the correlation coefficients for all pairs of factors. It becomes clear that collinearity has mostly been removed; there is a remaining relationship between word length and word frequency, which is expected as shorter words tend to be more frequent. This correlation is not a problem for our analysis, as it is confined to the baseline model. Furthermore, word frequency and trigram probability are highly correlated. Again this is expected, given that the frequencies of unigrams and higher-level n-grams tend to be related. This correlation is taken care of by residualizing, which isolates the two factors: word frequency is part of the baseline model, while trigram probability is part of the separate models that we fit on the residuals. All other correlations are small (with coefficients of .27 or less), with one exception: there is a high correlation between the $-\log(\Delta)$ term and the $\log(\lambda + (1-\lambda)\frac{p_2}{p_1})$ term in the multiplicative LDA model. This collinearity issue may explain the absence of a significant improvement in model fit when these two terms are added to the baseline (see Table 3).

	Factor	Len	Freq	$-\mathrm{l}(p)$	$-\mathrm{l}(\Delta)$
	Frequency	$-.310$			
	$-\log(p)$	$.230$	$-.700$		
SSS Add	$-\log(\Delta)$	$.016$	$-.120$	$.025$	
	$\log(\lambda+(1-\lambda)\frac{p_2}{p_1})$	$.024$	$.036$	$-.270$	$.065$
SSS Mult	$-\log(\Delta)$	$-.015$	$-.110$	$.035$	
	$\log(\lambda+(1-\lambda)\frac{p_2}{p_1})$	$.020$	$.028$	$-.260$	$.160$
LDA Add	$-\log(\Delta)$	$-.024$	$-.130$	$.046$	
	$\log(\lambda+(1-\lambda)\frac{p_2}{p_1})$	$.005$	$.014$	$-.250$	$.030$
LDA Mult	$-\log(\Delta)$	$-.120$	$.006$	$-.046$	
	$\log(\lambda+(1-\lambda)\frac{p_2}{p_1})$	$-.089$	$-.005$	$-.180$	$.740$

Table 5: Intercorrelations between model factors

6 Discussion

In this paper we investigated the contributions of syntactic and semantic constraint in modeling processing difficulty. Our work departs from previous approaches in that we propose a *single* measure which integrates syntactic and semantic factors. Evaluation on an eye-tracking corpus shows that our measure predicts reading time better than a baseline model that captures low-level factors in reading (word length, landing position, etc.). Crucially, we were able to show that the semantic component of our measure improves reading time predictions over and above a model that includes syntactic measures (based on a trigram model and incremental parser). This means that semantic costs are a significant predictor of reading time *in addition* to the well-known syntactic surprisal.

An open issue is whether a single, integrated measure (as evaluated in Table 4) fits the eye-movement data significantly better than separate measures for trigram, syntactic, and semantic surprisal (as evaluated in Table 3. However, we are not able to investigate this hypothesis: our approach to testing the significance of factors requires nested models; the log-likelihood test (see Section 4) is only able to establish whether adding a factor to a model improves its fit; it cannot compare models with disjunct sets of factors (such as a model containing the integrated surprisal measure and one containing the three separate ones). However, we would argue that a single, integrated measure that captures human predictive processing is preferable over a collection of separate measures. It is conceptually simpler (as it is more parsimonious), and is also easier to use in applications (such as readability prediction). Finally, an integrated measure requires less parameters; our definition of surprisal in 12 is simply the sum of the trigram, syntactic, and semantic components. An LME model containing separate factors, on the other hand, requires a coefficient for each of them, and thus has more parameters.

In evaluating our model, we adopted a broad coverage approach using the reading time data from a naturalistic corpus rather than artificially constructed experimental materials. In doing so, we were able to compare different syntactic and semantic costs on the same footing. Previous analyses of semantic constraint have been conducted on different eye-tracking corpora (Dundee and Embra Corpus) and on different languages (English, French). Moreover, comparisons of the individual contributions of syntactic and semantic factors were generally absent from the literature. Our analysis showed that both of these factors can be captured by our integrated surprisal measure which is uniformly probabilistic and thus preferable to modeling semantic and syntactic costs disjointly using a mixture of probabilistic and non-probabilistic measures.

An interesting question is which aspects of semantics our model is able to capture, i.e., why does the combination of LSA or LDA representations with an incremental parser yield a better fit of the behavioral data. In the psycholinguistic literature, various types of semantic information have been investigated: lexical semantics (word senses, selectional restrictions, thematic roles), sentential semantics (scope, binding), and discourse semantics (coreference and coherence); see Keller (2010) of a detailed discussion. We conjecture that our model is mainly capturing lexical semantics (through the vector space representation of words) and sentential semantics (through the multiplication or addition of words). However, discourse coreference effects (such as the ones reported by Altmann and Steedman (1988) and much subsequent work) are probably not amenable to a treatment in terms of vector space semantics; an explicit representation of discourse entities and coreference relations is required (see Dubey 2010 for a model of human sentence processing that can handle coreference).

A key objective for future work will be to investigate models that integrate semantic constraint with syntactic predictions more tightly. For example, we could envisage a parser that uses semantic representations to guide its search, e.g., by pruning syntactic analyses that have a low semantic probability. At the same time, the semantic model should have access to syntactic information, i.e., the composition of word representations should take their syntactic relationships into account, rather than just linear order.

References

ACL. 2010. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala.

Altmann, Gerry T. M. and Yuki Kamide. 1999. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition* 73:247–264.

Altmann, Gerry T. M. and Mark J. Steedman. 1988. Interaction with context during human sentence processing. *Cognition* 30(3):191–238.

Bellegarda, Jerome R. 2000. Exploiting latent semantic information in statistical language modeling. *Proceedings of the IEEE* 88(8):1279–1296.

Berry, Michael W., Susan T. Dumais, and Gavin W. O'Brien. 1995. Using linear algebra for intelligent information retrieval. *SIAM review* 37(4):573–595.

Bever, Thomas G. 1970. The cognitive basis for linguistic strutures. In J. R. Hayes, editor, *Cognition and the Development of Language*, Wiley, New York, pages 279–362.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Bullinaria, John A. and Joseph P. Levy. 2007. Extracting semantic representations from word cooccurrence statistics: A computational study. *Behavior Research Methods* 39:510–526.

Clifton, Charles, Adrian Staub, and Keith Rayner. 2007. Eye movement in reading words and sentences. In R V Gompel, M Fisher, W Murray, and R L Hill, editors, *Eye Movements: A Window in Mind and Brain*, Elsevier, pages 341–372.

Coccaro, Noah and Daniel Jurafsky. 1998. Towards better integration of semantic predictors in satistical language modeling. In *Proceedings of the 5th International Conference on Spoken Language Processing*. Sydney, Australia, pages 2403–2406.

Demberg, Vera and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* 101(2):193–210.

Dubey, Amit. 2010. The influence of discourse on syntax: A psycholinguistic model of sentence processing. In ACL.

Ferrara Boston, Marisa, John Hale, Reinhold Kliegl, Umesh Patil, and Shravan Vasishth. 2008. Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research* 2(1):1–12.

Frank, Stefan L. 2009. Surprisal-based comparison between a symbolic and a connectionist model of sentence processing. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*. Austin, TX, pages 139–1144.

Gibson, Edward. 2000. Dependency locality theory: A distance-dased theory of linguistic complexity. In Alec Marantz, Yasushi Miyashita, and Wayne O'Neil, editors, *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*, MIT Press, Cambridge, MA, pages 95–126.

Gildea, Daniel and Thomas Hofmann. 1999. Topic-based language models using EM. In *Proceedings of the 6th European Conference on Speech Communiation and Technology*. Budapest, Hungary, pages 2167–2170.

Griffiths, Thomas L., Mark Steyvers, and Joshua B. Tenenbaum. 2007. Topics in semantic representation. *Psychological Review* 114(2):211–244.

Hale, John. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the 2nd Conference of the North American Chapter of the Association*. Association for Computational Linguistics, Pittsburgh, PA, volume 2, pages 159–166.

Keller, Frank. 2010. Cognitively plausible models of human language processing. In ACL.

Kennedy, Alan and Joel Pynte. 2005. Parafoveal-on-foveal effects in normal reading. *Vision Research* 45:153–168.

Konieczny, Lars. 2000. Locality and parsing complexity. *Journal of Psycholinguistic Research* 29(6):627–645.

Landauer, Thomas K. and Susan T. Dumais. 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.

Levy, Roger. 2008. Expectation-based syntactic comprehension. *Cognition* 106(3):1126–1177.

Marslen-Wilson, William D. 1973. Linguistic structure and speech shadowing at very short latencies. *Nature* 244:522–523.

McDonald, Scott. 2000. *Environmental Determinants of Lexical Processing Effort*. Ph.D. thesis, University of Edinburgh.

McDonald, Scott and Chris Brew. 2004. A distributional model of semantic context effects in lexical processing. In *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics*. Barcelona, Spain, pages 17–24.

McDonald, Scott A. and Richard C. Shillcock. 2003. Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. *Vision Research* 43:1735–1751.

Mitchell, Jeff and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*. Columbus, OH, pages 236–244.

Mitchell, Jeff and Mirella Lapata. 2009. Language models based on semantic composition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore, pages 430–439.

Narayanan, Srini and Daniel Jurafsky. 2002. A Bayesian model predicts human parse preference and reading time in sentence processing. In Thomas G. Dietterich, Sue Becker, and Zoubin Ghahramani, editors, *Advances in Neural Information Processing Systems 14*. MIT Press, Cambridge, MA, pages 59–65.

Padó, Sebastian and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics* 33(2):161–199.

Padó, Ulrike, Matthew W. Crocker, and Frank Keller. 2009. A probabilistic model of semantic plausibility in sentence processing. *Cognitive Science* 33(5):794–838.

Pinheiro, Jose C. and Douglas M. Bates. 2000. *Mixed-effects Models in S and S-PLUS*. Springer, New York.

Pinker, Steven. 1994. *The Language Instinct: How the Mind Creates Language*. HarperCollins, New York.

Plate, Tony A. 1995. Holographic reduced representations. *IEEE Transactions on Neural Networks* 6(3):623–641.

Pynte, Joel, Boris New, and Alan Kennedy. 2008. On-line contextual influences during reading normal text: A multiple-regression analysis. *Vision Research* 48:2172–2183.

Rayner, Keith. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin* 124(3):372–422.

Roark, Brian. 2001. Probabilistic top-down parsing and language modeling. *Computational Linguistics* 27(2):249–276.

Roark, Brian, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, pages 324–333.

Smolensky, Paul. 1990. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence* 46:159–216.

Stanovich, Kieth E. and Richard F. West. 1981. The effect of sentence context on ongoing word recognition: Tests of a two-pricess theory. *Journal of Experimental Psychology: Human Perception and Performance* 7:658–672.

Staub, Adrian and Charles Clifton. 2006. Syntactic prediction in language comprehension: Evidence from either . . . or. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 32:425–436.

Steyvers, Mark and Tom Griffiths. 2007. Probabilistic topic models. In T. Landauer, D. McNamara, S Dennis, and W Kintsch, editors, *A Handbook of Latent Semantic Analysis*, Psychology Press.

Stolcke, Andreas. 2002. Srilm - an extensible language modeling toolkit. In *Proceedings of the Internatinal Conference on Spoken Language Processing*. Denver, Colorado.

Sturt, Patrick and Vincenzo Lombardo. 2005. Processing coordinated structures: Incrementality and connectedness. *Cognitive Science* 29(2):291–305.

Tanenhaus, Michael K., Michael J. Spivey-Knowlton, Kathleen M. Eberhard, and Julie C. Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science* 268:1632–1634.

van Berkum, Jos J. A., Colin M. Brown, and Peter Hagoort. 1999. Early referential context effects in sentence processing: Evidence from event-related brain potentials. *Journal of Memory and Language* 41:147–182.

Wright, Barton and Merrill F. Garrett. 1984. Lexical decision in sentences: Effects of syntactic structure. *Memory and Cognition* 12:31–45.