Uncertainty reduction as a measure of cognitive load in sentence comprehension

Stefan L. Frank
s.frank@ucl.ac.uk
Department of Cognitive, Perceptual, and Brain Sciences
University College London
26 Bedford Way, London WC1H 0AP, United Kingdom

Abstract

The entropy-reduction hypothesis claims that the cognitive processing difficulty on a word in sentence context is determined by the word's effect on the uncertainty about the sentence. Here, this hypothesis is tested more thoroughly than has been done before, using a recurrent neural network for estimating entropy and self-paced reading for obtaining measures of cognitive processing load. Results show a positive relation between reading time on a word and the reduction in entropy due to processing that word, supporting the entropy-reduction hypothesis. Although this effect is independent from the effect of word surprisal, we find no evidence that these two measures correspond to cognitively distinct processes.

Keywords: Sentence comprehension, self-paced reading, cognitive load, word information, entropy reduction, surprisal, recurrent neural network

1 Introduction

Although many cognitive scientist believe information processing to be central to cognition, it is not always clear what is meant by 'information' (cf. Piccinini & Scarantino, 2011). Presumably, stimuli that are more 'informative' increase 'cognitive load' and should therefore require more 'mental effort' to process, but unless information content is properly quantified such statements remain somewhat metaphorical. To go beyond the metaphor we need to turn to information theory, which has indeed been brought to bear in explanations of cognitive phenomena. For example, information-theoretic considerations have recently been called upon to explain aspects of language production (Jaeger, 2010) and properties of language itself (Piantadosi, Tily, & Gibson, 2012).

The current paper looks at information-theoretic measures of cognitive processing load in language comprehension. More in particular, it investigates to what extent experimentally obtained word-reading times

are predicted by formally derived measures of the amount of information conveyed by each word in a sentence. In this context, the most common quantification of a word's information content is its *surprisal* (also known as *self-information*), a measure of the extent to which the word came unexpected to the reader or listener. Formally, if word sequence w_1^t (short for $w_1 w_2 \dots w_t$) forms the first t words of a sentence, the surprisal of the next word, w_{t+1}, is defined as $-\log P(w_{t+1}|w_1^t)$. It has been argued that a word's surprisal is an important predictor of the cognitive load experienced on encountering the word (Hale, 2001; Levy, 2008) and indeed it is well established by now that word-reading times correlate positively with surprisal values (e.g., Boston, Hale, Patil, Kliegl, & Vasishth, 2008; Boston, Hale, Vasishth, & Kliegl, 2011; Demberg & Keller, 2008; Frank & Bod, 2011).

Although any sufficiently accurate probabilistic language model[1] generates surprisal estimates that can account for reading times, constructing such an accurate model can be problematic. For this reason, surprisal values are often assigned to the words' part-of-speech (i.e., syntactic category) rather than to the words themselves. Nevertheless, reading-time effects of the surprisal of actual words have also been demonstrated (Brakel, 2009; Fernandez Monsalve, Frank, & Vigliocco, 2012; Fossum & Levy, 2012; Mitchell, Lapata, Demberg, & Keller, 2010; Roark, Bachrach, Cardenas, & Pallier, 2009; Smith & Levy, 2008)

A lesser known alternative operationalization of a word's information content follows from the idea that, at each point during sentence comprehension, the reader or listener experiences some degree of uncertainty about what is being communicated. This uncertainty (usually) decreases with each incoming word, and the extent of this decrease corresponds to the information conveyed by the word in the current sentence context.

Formally, if X denotes the set of all possible sentence structures (or 'interpretations'), the goal of comprehension is to identify which structure $x \in X$ is being communicated. We can view X as a random variable, where each x has an occurrence probability $P(x)$. The uncertainty about x is quantified as the *entropy* (Shannon, 1948) of the probability distribution over X, defined as:

$$H(X) = -\sum_{x \in X} P(x) \log P(x).$$

The larger the entropy, there more uncertainty there is about the value of x. It is easy to show that entropy is maximal when all x have the same probability (which, intuitively, indeed comes down to maximal uncertainty) and that entropy is zero if (and only if) $P(x) = 1$ for one particular x, that is, there is absolute certainty about what is communicated.

When the first t words of a sentence (i.e., w_1^t) have been processed, the probability distribution over X has changed from $P(X)$ to $P(X|w_1^t)$. The corresponding entropy equals:

$$H(X; w_1^t) = -\sum_{x \in X} P(x|w_1^t) \log P(x|w_1^t). \tag{1}$$

The amount of information that the sentence's next word, w_{t+1}, gives about the random variable X is

[1] A language model is, by definition, a probability model that assigns probabilities to sentences. The next-word probabilities $P(w_{t+1}|w_1^t)$ follow directly from these sentence probabilities.

defined as the reduction in entropy due to that word:[2]

$$\Delta H(X; w_{t+1}) = H(X; w_1^t) - H(X; w_1^{t+1}). \tag{2}$$

Hale (2003, 2006, 2011) argues that this entropy reduction should be predictive of the cognitive load experienced when processing w_{t+1} and demonstrates how the relevant entropy values can be computed given a probabilistic grammar (Hale, 2006). Blache and Rauzy (2011) propose a simpler entropy measure that is computed over the probabilities of part-of-speech assignments rather than syntactic tree structures, in effect marginalizing over these structures. However, they do not compare the resulting entropy-reduction values to empirical measures of cognitive processing load. Wu, Bachrach, Cardenas, and Schuler (2010) did find an effect of entropy reduction on word-reading time but their model takes only the possible structures of the sentence so far (rather than complete sentences) into account. Roark et al. (2009) also found an entropy-reduction effect but greatly simplified the computation of entropy by considering the probability of the single next word only.

In order to look further ahead than a single word, Frank (2010) used a much simpler language model that does not assign syntactical categories or structures. Reading times were shown to depend on both surprisal and entropy reduction. However, these two information measures were defined over parts-of-speech rather than words, which has at least two drawbacks: First, the particular choice of syntactic categories is theory dependent and it is not at all clear whether a similar (or, for that matter, any) categorization is cognitively relevant. Second, a word's syntactic category can be ambiguous, in particular at the moment the word appears. Although it is most often possible to unambiguously classify each word after the sentence has been fully interpreted, both the surprisal and entropy-reduction measure are based on the assumption that each word is interpreted immediately, that is, sentence processing is perfectly incremental.

In the study presented here, surprisal and entropy reduction are estimated for actual words rather than just their syntactic categories, allowing for a more thorough investigation into the entropy-reduction hypothesis. Section 2 presents the model that was used for generating the word-information estimates, as well as the text corpus on which it was trained. As in Frank (2010), entropy is not computed over structures but over input sequences. In addition to simplifying the estimation of entropy, this has the advantage that it does not rely on any particular grammar or other assumption about how sentences are interpreted or parsed. Since it is unknown which particular structures people assign to sentences, it may be more appropriate to abstract away from structures altogether and deal only with the sequential input. Note that this is a simplifying assumption that should not be taken as a claim about the cognitive process of sentence comprehension.

Section 3 describes a self-paced reading study in which reading-time data are collected on sentences for which the model estimates information values.[3] As presented in Section 4, an analysis of these word-

[2] Although alternative definitions are possible, the entropy-reduction measure has the important property that it is additive: The information conveyed by w_1^{t+1} equals the information given by w_1^t plus the additional information by w_{t+1} (Blachman, 1968).

[3] The reading times and information values are available as online supplementary material.

information and reading-time measures reveals that surprisal and entropy reduction have independent effects, confirming the entropy-reduction hypothesis. Further analyses, discussed in Section 5, were intended to uncover a cognitively relevant distinction between the surprisal and entropy-reduction information measures. However, no evidence was found that the two measures correspond to cognitively distinct representations, processes, or strategies. As concluded in Section 6, finding such a distinction may require more rigorous, theory-driven experiments.

2 Estimating word information

2.1 Model architecture

Fig. 1 presents the architecture of the recurrent neural network (RNN) that was used as the probabilistic language model for estimating word-surprisal and entropy-reduction values. This network is not proposed as a cognitive model, rather, it serves as a tool for obtaining the required word-information measures; one that has several advantages over alternative models. For one, RNNs process more efficiently than phrase-structure grammars (or other structure-assigning models), which is of particular importance for computing entropy. Also, they can be trained on unannotated sentences (i.e., word strings instead of tree structures). In addition, RNNs have been shown to estimate surprisal values that fit reading times better than do grammar-based surprisal estimates (Frank & Bod, 2011). This was also demonstrated by Fernandez Monsalve et al. (2012), using the very same model and data set as in the current study, which suggests that the model's entropy-reduction estimates, too, may provide a good fit to the data.

Although the use of RNNs as cognitive models of sentence processing has been common ever since the seminal work by Elman (1990), these simulations always deal with hand-crafted toy languages. The current objective, in contrast, was to obtain entropy measures for a realistic sample of English, without resorting to part-of-speech strings (as in Frank, 2010) or a miniature grammar (as in Hale, 2003, 2006). In order to make this task computationally feasible, the vocabulary was limited to 7,754 highly frequent items.[4] This included two special symbols: the comma and an end-of-sentence marker that replaced the period, question mark, and exclamation point.

2.2 Model training

The network training data consisted of all the 702,412 sentences (comprising 7.6 million word tokens) in the written-text part of the British National Corpus (BNC) that contained only words from the vocabulary. The training procedure was divided into three distinct stages corresponding to distinct parts of the network, as shown in Fig. 1 and explained below. Additional technical details can be found Appendix A.

[4]These were the high-frequency content words used by Andrews, Vigliocco, and Vinson (2009) plus the 200 most frequent words of English, most of which are function words.

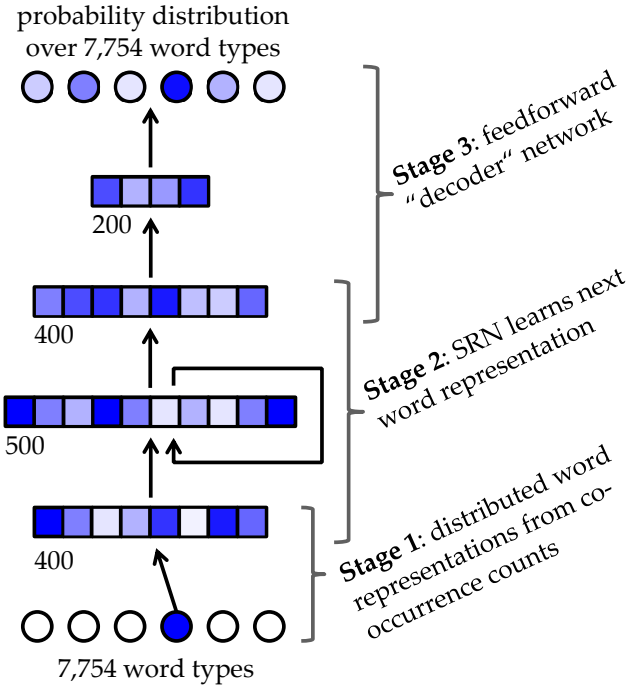

Fig. 1: Architecture of neural network language model, and its three learning stages. Numbers indicate the number of units in each network layer. Reproduced from Fernandez Monsalve et al. (2012).

2.2.1 Stage 1: Developing word representations

Words were represented as vectors in a continuous, high-dimensional space, such that similarities between the representations mirror the distributional similarities between the corresponding words. The vectors were extracted in an efficient and unsupervised manner from word co-occurrences in the selected BNC sentences, using a method akin to that of the Hyperspace Analogue to Language model (Burgess, Livesay, & Lund, 1998): First, a matrix of co-occurrence frequencies was constructed in which each value equals the number of times two word types directly precede or follow each other. After a simple transformation of these values, the 400 matrix columns with the highest variance were selected, resulting in a 400-dimensional word-representation space. As can be seen from Table 1, words cluster by syntactic category and even within categories there is some evidence for semantic clustering (e.g., the three words closest to 'parents' are 'pupils', 'teachers', and 'mothers'). Such clustering of word representations facilitates generalization in RNN training (Frank & Čerňanský, 2008).

Earlier work on distributional semantics in psychology (e.g., Landauer & Dumais, 1997) and, more recently, in computational linguistics (e.g., Baroni & Lenci, 2010) has applied more advanced techniques for

Table 1: Clustering of word representations. Each column corresponds to a syntactic category from which one word (shown in bold font) was chosen at random. That word's ten nearest neighbors are listed in order of increasing Euclidean distance.

noun (singular)	noun (plural)	verb (present)	verb (past)	adjective	adverb	preposition
nightmare	**parents**	**agree**	**stood**	**equal**	**maybe**	**beyond**
verdict	pupils	argue	sat	identical	unfortunately	beneath
realisation	teachers	confirm	stayed	absolute	meanwhile	onto
portrait	mothers	depend	waited	part-time	although	despite
packet	patients	listen	walked	agricultural	wherever	beside
mattress	workers	cope	remained	accurate	instead	amongst
succession	relatives	admit	paused	electrical	fortunately	unlike
skull	clients	fail	lived	enormous	nevertheless	throughout
rifle	employers	respond	stared	informal	perhaps	alongside
plea	children	rely	laughed	artificial	unless	regarding
scrutiny	lawyers	appreciate	rang	adequate	hence	ignoring

extracting lexical semantic information from much larger text copora, but the current goal was not to obtain word representations that are interesting in their own right. Instead, they merely need to reduce learning and processing complexity for the RNN model.

2.2.2 Stage 2: Learning temporal structure

The central part of the model is a standard Simple Recurrent Network (SRN; Elman, 1990). Such a network can incrementally process temporal sequences because (as indicated in Fig. 1) the hidden layer receives not only external input but also its own previous state. That previous state, in turn, depends on even earlier inputs and previous states, so that the processing of the current input depends on the entire input stream so far. The SRN learned to predict, at each point in the training corpus, the next word's vector representation given the sequence of word vectors corresponding to the sentence so far.

2.2.3 Stage 3: Decoding predicted word representations

After processing the input sequence w_1^t, the output of the trained SRN from Stage 2 is a 400-dimensional vector that combines the 7,754 individual word representations, somehow weighted by each word's estimated probability of being w_{t+1}. A feedforward 'decoder' network was trained to disentangle that complex mix of vectors. It took as input the SRN's output vector at each point in the training sentences while it received the actual next word as target output. That target output was encoded as 7,754-dimensional vector consisting of all 0s except for a single 1 at the element corresponding to the target word.

2.3 Obtaining word-information values

The model was used to compute surprisal and entropy reduction values for each word of the experimental sentences (see Section 3). This was done at ten intervals over the course of Stage 3 training: after presenting 2K, 5K, 10K, 20K, 50K, 100K, 200K, and 350K sentences, and after presenting the full training corpus once and twice. In this manner, a range of information values was obtained for each word token.

2.3.1 Surprisal

The decoder network's output activations at each time step are rescaled to constitute a probability distribution over word types. That is, the model's output after processing sentence prefix w_1^t forms an estimate of $P(w_{t+1}|w_1^t)$ for each possible next word w_{t+1}, which translates directly into the surprisal of the actual upcoming word.

2.3.2 Simplified entropy

The RNN does not assign any structure or other kind of interpretation to sentences. Instead, entropy is computed over probabilities of the sentences themselves, that is, the set X of Eq. 1 contains only (and all) word sequences instead of structures. As a consequence, there is no more uncertainty about what has occurred up to the current word w_t: Although the intended structure of w_1^t may be uncertain, the word sequence itself is not. This means that only the upcoming input sequence (i.e., from w_{t+1} onwards) is relevant for entropy. However, the number of sequences is far too large for exact computation of entropy, even if infinite-length sequences are not allowed and some realistic upper bound on sequences length is assumed. Therefore, probabilities are not estimated over complete sentence continuations. Instead, the 'lookahead distance' is restricted to some small value n, that is, only the upcoming n words are considered: Entropy is computed over the distribution $P(w_{t+1}^{t+n}|w_1^t)$. These probabilities can (at least in principle) be computed from the model's output by applying the chain rule: $P(w_{t+1}^{t+n}|w_1^t) = \prod_{i=1}^{n} P(w_{t+i}|w_1^{t+i-1})$. The definition of entropy from Eq. 1 now becomes

$$H_n(W^n; w_1^t) = -\sum_{w_{t+1}^{t+n} \in W^n} P(w_{t+1}^{t+n}|w_1^t) \log P(w_{t+1}^{t+n}|w_1^t),$$

where W^n denotes the set of all sequences of n words (including shorter sequences that end in the end-of-sentence marker). The number of elements in W^n grows exponentially as the lookahead distance n increases, and, consequently, so does computation time. Therefore, n needs to remain very small: The current simulations do not go beyond $n = 4$.

If the vocabulary consisted of a small number of syntactic categories (as in Frank, 2010) no further simplification would be needed. However, a 7,754-word vocabulary is used here, and so for each increase in n by 1, the number of sequences in W^n multiplies by 7,753 (not by 7,754 because one of the word types is the end-of-sentence marker that signal sentence completion). If all possible continuations w_{t+1}^{t+n} are taken

Table 2: Two example sentences and corresponding word-information estimates.

	she	would	soon	be	found	if	she	tried	to	hide
surprisal	3.22	4.16	6.79	0.78	5.13	4.85	2.55	5.69	0.17	7.29
ΔH_4	1.53	0.54	1.50	1.50	0.63	−1.03	2.14	2.42	0.44	1.22
	the	brothers	stood	up	and	came	down	from	the	platform
surprisal	2.26	8.92	5.61	3.02	2.66	5.43	3.19	3.96	1.00	8.69
ΔH_4	2.48	−0.94	0.51	2.09	0.85	−0.84	1.80	0.67	1.63	1.96

into account, even the modest value of $n = 4$ yields over 4×10^{11} sequences. Clearly, further simplification is required. This is accomplished by taking only the 40 most probable words w_{t+i} (i.e., those with highest $P(w_{t+i}|w_1^{t+i-1})$) when expanding to $i+1$ according to the chain rule. Consequently, the number of relevant sequences equals 7754×40^n, which approximates 5×10^8 for $n = 4$.

2.3.3 Simplified entropy reduction

Be reminded that entropy was originally intended to measure the uncertainty about the complete sentence structure $x \in X$ being communicated. The reduction in entropy due to processing word w_{t+1} was expressed by Eq. 2. However, an alternative expression for entropy reduction is needed now that we have simplified entropy to measure the uncertainty about the upcoming n words $w_{t+1}^{t+n} \in W^n$.

Processing word w_{t+1} comes down to identifying the first word of w_{t+1}^{t+n}, so this word drops out of the computation of uncertainty about the upcoming word string. What is left is uncertainty about the $n-1$ words following w_{t+1}. Hence, the relevant entropy at this point is over the probabilities of word strings in W^{n-1} (see Frank, 2010 for a more formal derivation) so that the simplified reduction in entropy due to w_{t+1} becomes:

$$\Delta H_n(W^n; w_{t+1}) = H_n(W^n; w_1^t) - H_{n-1}(W^{n-1}; w_1^{t+1}). \tag{3}$$

Note that positive ΔH_n corresponds to a *decrease* in entropy.

2.3.4 Word-information examples

Table 2 shows two examples of 10-word sentences, with each word's surprisal and ΔH_4 as estimated by the model after complete training. Note that entropy reduction can occasionally be negative (Blachman, 1968). Hale (2006) treats these negative values as 0s, but they were included in the results of Section 4.

3 Collecting reading-time data

Most previous studies on the relation between word information and reading time used data collected over newspaper editorials (e.g., Demberg & Keller, 2008; Fossum & Levy, 2012; Frank, 2010; Frank & Bod,

2011; Smith & Levy, 2008) or narrative texts (e.g., Roark et al., 2009; Wu et al., 2010). Considering the claim that word information is a general predictor of cognitive processing effort, it does make sense to look at reading times over a piece of natural discourse rather than sentence stimuli constructed for specific psycholinguistic experiments. However, there are a few drawbacks to using newspaper texts or narratives. For one, understanding discourse engages a vast amount of background knowledge that the language models cannot access. Also, the language models assume that sentences are independent from one another, whereas word probabilities in texts also depend on material from previous sentences.

For these reasons, the current study uses word-reading times over individual sentences, which were drawn semi-randomly from novels that were freely available on www.free-online-novels.com, a website where aspiring authors can publish their work. Considering that these novels are not published through more traditional channels, it is very unlikely that they were known to the participants. Three novels from different genres were selected,[5] each written in British English spelling. From these novels, 361 sentences were chosen that were at least five words long, could be interpreted without their story context, and contained only words from the 7,754-word vocabulary on which the model was trained. Two of these experimental sentences are shown in Table 2. The average sentence length was 13.7 words, with a maximum of 38. In order to prevent re-occurrence of protagonists' names across sentences, which may lead participants to connect the sentences into a narrative, names were changed such that none occurred more than twice across the stimuli.

A total of 117 students took part in the experiment as part of their first-year Psychology program at University College London. Forty-seven were not native English speakers, so their data were discarded. Because of concerns that the participants would not remain attentive while reading 361 individual sentences, the stimuli were drawn randomly from the total set until 40 minutes had elapsed.[6] Each sentence was preceded by a fixation cross, presented centrally on a computer monitor. As soon as the participant pressed the space bar, the fixation cross was replaced by the sentence's first word in 40-point Courier New font. At each subsequent key press, the current word was replaced by the next, always at the same central location on the display. Just, Carpenter, and Woolley (1982) compared this manner of sentence presentation to the more common moving-window technique and to eye tracking, and found that the obtained reading times are qualitatively similar across methodologies. However, presentation at a fixed display location tended to cause stronger spillover effects, where reading difficulty on word w_t is in fact observed at w_{t+1}. As will be discussed shortly, such an effect was also found in the current reading-time data.

Nearly half the sentences were followed by a yes/no comprehension question. Of the 70 native English speaking participants, 16 scored below 80% correct on the comprehension questions, so their reading times were not analyzed.[7] The remaining 54 participants read between 120 and 349 sentences each, with an average

[5]These were: *Aercu* by Elisabeth Kershaw, *The Unlikely Hero* by Jocelyn Shanks, and *Hamsters! (or: What I Did On My Holidays by Emily Murray)* by Daniel Derrett.

[6]This included an initial, unrelated lexical decision task that took approximately 10 minutes.

[7]The large number of badly performing subjects is probably caused by a lack of motivation, due to the fact that they participated as part of their undergraduate training rather than signing up for the study. Many participants with very high error rates showed unrealistically short response times on each word, indicating that they were simply trying to get the

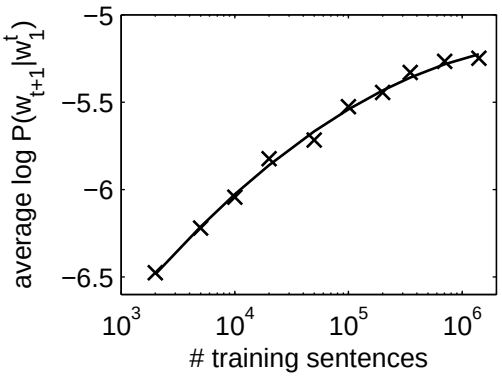

Fig. 2: Accuracy of the language model as a function of the number of sentences presented in Stage 3 training.

of 224. Data on sentence-initial and sentence-final words were removed, as well as words directly preceding or following a comma. Also, reading times below 50 ms were excluded from the analysis, leaving a total of 136,481 data points.

4 Results

4.1 Language model accuracy

An accurate language model is one that captures the linguistic patterns in the training data and generalizes well to the experimental sentences, that is, it assigns high probabilities to the words. More precisely, the model's accuracy was quantified as the average of $\log P(w_{t+1}|w_1^t)$ over the experimental sentences, where each item is weighted by the number of participants for which it took part in the analysis.[8] Fig. 2 shows how the accuracy develops as the decoder network (Stage 3) is trained on more and more sentence tokens. As expected, the network improves its knowledge of the language over the course of training. Importantly, accuracy increases monotonically, showing that the model does not suffer from overfitting.

4.2 Relation between surprisal and entropy reduction

Considering that both surprisal and entropy reduction express the information conveyed by words, one would expect these two values to correlate strongly. This should be the case in particular for $n = 1$, because $\Delta H_1 = -\sum_{w_{t+1}} P(w_{t+1}|w_1^t) \log P(w_{t+1}|w_1^t)$, which equals the expected value of the surprisal of w_{t+1}. Indeed, surprisal and ΔH_1 correlate considerably but, as Fig. 3 shows, this correlation is fairly weak for $n = 3$ and virtually absent for $n = 4$.

experiment over with (they were unaware that it was time-bound).

[8] Alternatively, it would have been possible to measure performance by looking at the output vectors of the SRN of Stage 2, but these do not have a probabilistic interpretation and do not directly lead to word-information values.

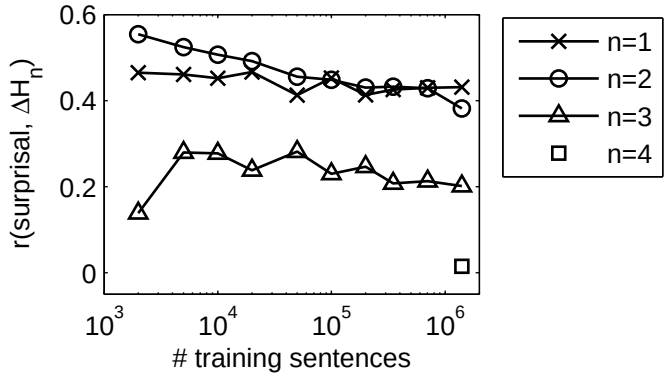

Fig. 3: Coefficient of correlation (where each item is weighted by the number of participants for which it took part in the analysis) between estimates of surprisal and ΔH_n, as a function of n and of the number of sentences on which the network was trained. Because of the very long computation time for ΔH_4, these values are obtained from the fully trained model only.

4.3 Effect of word information

A mixed-effects regression model (Baayen, Davidson, & Bates, 2008) was fitted to the log-transformed reading times[9] (see Appendix B for details). This 'baseline' model did *not* include factors for surprisal or entropy reduction, but did have several well-known predictors of word-reading time: word length, frequency, and forward transitional probability (i.e., the word's probability given the previous word). In order to factor out effects of the previous word, its length, frequency, and forward probability were also included. As recommended by Baayen and Milin (2010), the effect of auto-correlation was reduced by including also the reading time on the previous word as a predictor. In addition, the model had linear and quadratic factors for sentence position in the experiment (capturing practice and fatigue effects) as well as linear and quadratic factors for word position in the sentence (capturing non-linear speed-up or slowdown over the course of a sentence). Finally, significant two-way interactions were included, as were by-subject and by-item random intervals and the three random slopes with the strongest effect.

In order to quantify the effect of one information measure over and above the other (and all the baseline predictors), three extensions of the baseline model were fitted: One with a factor for surprisal, one with a factor for entropy reduction, and one with both. The difference in deviance (i.e., lack of fit) between the model that includes both information measures and a model with only one, quantifies the goodness-of-fit to reading times of the other information measure. This value is the χ^2-statistic of a log-likelihood test for that measure's additional effect (where $\chi^2 > 3.84$ corresponds to statistical significance at the $p = .05$ level). The fit to reading times of surprisal and of entropy reduction (with $n = 3$) were estimated at each of the 10 points during Stage 3 network training. When the network was fully trained, the fit of entropy reduction

[9] An analysis of raw reading times yielded qualitatively very similar results, although all effects were slightly weaker.

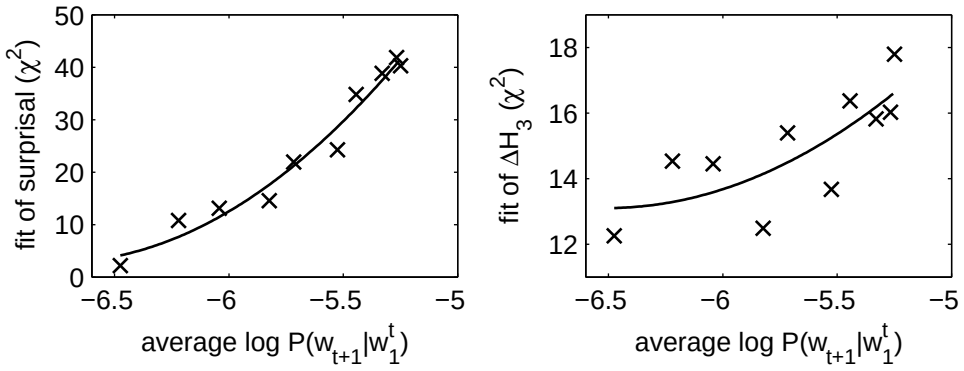

Fig. 4: Goodness-of-fit of surprisal (left) and entropy reduction (right) for $n = 3$, over and above all other predictors, as a function of language model accuracy. Plotted are the estimated χ^2-statistics ($\times$) and fitted second-degree polynomials.

was also estimated for $n = 1, 2$, and 4.

Preliminary analyses revealed a considerable spillover effect: Reading time on a word was predicted much more strongly by surprisal and entropy reduction of the *previous* word than by the word's own information content, although the latter did have a statistically significant effect (cf. Fernandez Monsalve et al., 2012). This does not imply that there is a delay in the integration of the word in its sentence context. More likely, the participants tend to respond too quickly, that is, the key press occurs before the word has been fully processed. Because of the spillover effect, the results presented here show how response time on word w_{t+1} is related to the amount of information conveyed by word w_t.

The extent to which surprisal and entropy reduction predict reading times is displayed in Fig. 4. Either information measure has a statistically significant effect over and above the other and the baseline model predictors (all $\chi^2 > 10.7; p < .002$, except for the effect of surprisal when language model accuracy is still very low). Importantly, all effects are in the correct direction, that is, larger information content corresponds to longer reading time. Also, as expected, the effects grow stronger as the network learns more about the statistics of the language.

Fig. 5 shows how the unique effect of entropy reduction depends on the lookahead distance n, when the network is fully trained. For $n = 1$ or $n = 2$, there is no additional effect of ΔH_n on reading times, but the effect clearly grows with larger n.

5 Discussion

The results presented above clearly support the entropy-reduction hypothesis: Entropy reduction is a significant predictor of reading time, over and above many other factors, including surprisal. Moreover, increasing suffix-length n improves model fit, suggesting that what really matters is uncertainty about the *complete*

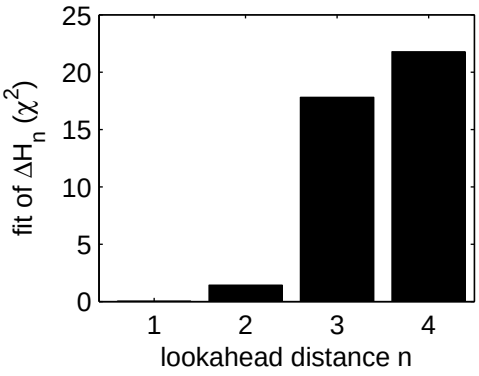

Fig. 5: Goodness-of-fit of entropy reduction as a function of lookahead distance n.

sentence. These findings contribute to a growing body of evidence that the amount of information conveyed by a word in sentence context is indicative of the amount of cognitive effort required for processing, as can be observed from word-reading times. A considerable number of previous studies have shown that surprisal can serve as a cognitively relevant measure for a word's information content. In contrast, the relevance of entropy reduction as a cognitive measure has not been investigated this thoroughly before.

It is tempting to take the independent effects of surprisal and entropy reduction as evidence for the presence of two distinct cognitive representations or processes, one related to surprisal, the other to entropy reduction. If the two information measures are indeed cognitively distinct, it may be possible to discover a corresponding dissociation in the reading-time data. For example, the difference between surprisal and entropy reduction may correspond to a difference in reading strategies, such that some participants show a stronger effect of surprisal and others of entropy reduction. Whether this is the case can be investigated by including in the regression analysis by-subject random slopes, which capture how participants vary in their sensitivity to surprisal or entropy reduction. Crucially, it can be either assumed that there is no correlation between readers' sensitivity to surprisal and to entropy reduction, or that such a correlation does exist. In the first case, the random slopes of surprisal are uncorrelated to those of entropy reduction. In the second case, a correlation coefficient between the two sets of slopes is also estimated. As it turns out, estimating this coefficient increases the regression model's fit ($\chi^2 = 6.62; p < .02$), indicating that the effects of the two information measures are related across participants. However, the correlation coefficient between the sets of slopes is positive ($r = .56$), which means that readers who show a stronger effect of surprisal also show a stronger effect of entropy reduction. That is, readers differed in their overall sensitivity to word information but not in their reading strategy.

The information-theoretic account of sentence comprehension is not only applicable to particular sentence types or syntactic constructions. Surprisal and entropy reduction are general measures of information content so, in principle, they should predict cognitive load on each word of any sentence. This is not to say that they always predict equally well. For instance, Grodner and Gibson (2005) compared reading times on sentences

containing a subject-relative clause (as in 'the reporter who sent the photographer ...') to object-relative sentences ('the reporter who the photographer sent ...') and found that the verb 'sent' is read more slowly in the object-relatives. It has been argued that surprisal cannot account for this effect (Bartek, Lewis, Vasishth, & Smith, 2011; Levy, 2008) whereas Hale (2003) shows that entropy reduction does predict the difference between the two sentence types (at least, under the particular grammar he assumes). Although this does not imply that the two information measures are cognitively distinct, it may provide a starting point for teasing apart their effects and, thereby, their cognitive interpretations. Therefore, it is of interest to investigate whether sentences differ regarding the information measure that best explains the associated reading times. This can be done by including in the regression by-sentence random slopes of surprisal and entropy reduction (just like by-subject slopes were included to investigate differences across subjects). Possibly, reading times on some sentences depend more on surprisal whereas entropy reduction is more important for others. However, this was not the case as there was no significant correlation between the two sets of by-sentence random slopes ($\chi^2 \approx 0; p > .8$).

A third possibility is that reading times on different syntactic categories are affected differentially by the two information measures. To investigate whether this was the case, an automatic tagger (Tsuruoka & Tsujii, 2005) assigned a part-of-speech tag to each word of the sentence stimuli, after which all tags where checked by hand and, if needed, changed to conform to the Penn Treebank guidelines (Santorini, 1991). Next, by-part-of-speech random slopes of surprisal and entropy reduction were included in the regression model.[10] Once again, the two sets of slopes were not significantly correlated ($\chi^2 = 0.04; p > .8$).

In a final attempt to find a dissociation between effects of surprisal and entropy reduction, the words were divided into two broad syntactic classes: content words (i.e., those tagged as nouns, verbs, adjectives, etc.) and function words (tagged as determines, pronouns, prepositions, etc.). This time, the regression model was not extended with additional random slopes but with interactions between each information measure and word class (a categorical predictor). Neither interaction was significant (surprisal $\times$ class: $t = -1.75, p > .08$; $\Delta H_4 \times$ class: $t = 0.75, p > .4$; where a positive t-value indicates a stronger effect for content words) and the difference between the coefficients of the two interactions was only marginally significant ($t = 1.87, p > .06$).[11] Hence, the effects of surprisal and entropy reduction do not systematically differ between function and content words.

To sum up, no evidence was found of a dissociation between surprisal and entropy reduction: The difference between the two effects does not correspond to a difference between subjects, sentences, parts-of-speech, or content versus function words. On the basis of these data and analyses, there is no reason to reject the null hypothesis that surprisal and entropy reduction are cognitively indistinguishable. Hence, the two information measures may merely form complementary formalizations of a single, cognitively relevant (but not necessarily information-theoretically grounded) notion of how much a word contributes to the

[10] Because of the spillover effect on reading times, the part-of-speech of the previous word was used. By-item random effects were not included in this analysis, nor were they in the by-sentence analysis.

[11] Using treatment coding for the variable 'class', the two interaction terms were nearly orthogonal ($r = .07$).

communication. A demonstration of how this may work is provided by one recent model (Frank & Vigliocco, 2011) that simulates sentence comprehension as the incremental and dynamical update of a non-linguistic representation of the state-of-affairs described by the sentence. In this framework, surprisal and entropy reduction are defined with respect to a probabilistic model of the *world*, rather than a model of the *language*: Information quantities do not depend on how well a sentence's form matches the statistical patterns of the language, but on the relation between the sentence's meaning and the statistical patterns of events in the world. As it turns out, the model's word-processing times correlate positively with both surprisal and entropy reduction even though there is nothing in the model itself that can be considered to generate either 'surprisal effects' or 'entropy-reduction effects'. Rather, there is one comprehension mechanism that is responsible for the incremental revision of a semantic representation. Surprisal and entropy reduction form two complementary, imperfect quantifications of the extent of this revision.

6 Conclusion

Words that convey more information take longer to read. Two different measures of information, surprisal and entropy reduction, were shown to have independent effects on reading time, demonstrating that both these information-theoretic quantities have cognitive import. More in general, there is at least some truth to the information-processing metaphor of cognitive science.

The question remains which underlying cognitive mechanism is responsible for longer reading times on words that convey more information. The surprisal and entropy-reduction theories describe the relation between information content and cognitive load at Marr's (1982) computational level only. That is, they do not propose, at the algorithmic level, any sentence-processing mechanism that takes longer to process words that have higher surprisal or lead to greater reduction in sentence entropy. If the two different measures had explained different parts of the reading-time data, this might have provided a clue about the underlying cognitive mechanism. Alas, such a dissociation was not found.

Both Hale (2011) and Levy (2008) present a possible mechanism giving rise to effects of word information, based on the notion that each incoming word rules out some possibilities, be it either parser states (Hale, 2011) or (interpretations of) sentences (Levy, 2008). The more is ruled out by a word, the more information is conveyed and the more processing work needs to be done. Hale (2011) relates this amount of work to entropy reduction whereas Levy (2008) derives surprisal as the relevant value, but considering the findings presented here, what we need is a mechanistic model that generates *both* effects. Such a model should be able to make predictions about which information measure is most important under which circumstances. Testing these predictions in a set of controlled experiments might then, after all, uncover a cognitive distinction between surprisal on the one hand and entropy reduction on the other.

Acknowledgments

I would like to thank Irene Fernandez Monsalve and Gabriella Vigliocco for their help, as well as Daniel Müller-Feldmeth, Marshall Mayberry, John Hale, and two anonymous reviewers for their comments. The research presented here was funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant number 253803.

References

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, *116*, 463–498.

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*, 390–412.

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*, 12–28.

Baroni, M., & Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, *36*, 673–721.

Bartek, B., Lewis, R. L., Vasishth, S., & Smith, M. R. (2011). In search of on-line locality effects in sentence comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 1178–1198.

Blache, P., & Rauzy, S. (2011). Predicting linguistic difficulty by means of a morpho-syntactic probabilistic model. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation* (pp. 160–167).

Blachman, N. M. (1968). The amount of information that y gives about X. *IEEE transactions on information theory*, *14*, 27–31.

Boston, M. F., Hale, J., Patil, U., Kliegl, R., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, *2*, 1–12.

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

Brakel, P. (2009). *Testing surprisal theory using connectionist networks and dimensionality reduction methods*. Unpublished master's thesis, University of Amsterdam, The Netherlands.

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: a computational study. *Behavior Research Methods*, *39*, 510–526.

Burgess, C., Livesay, K., & Lund, K. (1998). Explorations in context space: words, sentences, discourse. *Discourse Processes*, *25*, 211–257.

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

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

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

Fossum, V., & Levy, R. (2012). Sequential vs. hierarchical syntactic models of human incremental sentence processing. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics* (pp. 61–69). Montréal, Canada: Association for Computational Linguistics.

Frank, S. L. (2010). Uncertainty reduction as a measure of cognitive processing effort. In *Proceedings of the 2010 workshop on cognitive modeling and computational linguistics* (pp. 81–89). Uppsala, Sweden: Association for Computational Linguistics.

Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, *22*, 829–834.

Frank, S. L., & Čerňanský, M. (2008). Generalization and systematicity in echo state networks. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 733–738). Austin, TX: Cognitive Science Society.

Frank, S. L., & Vigliocco, G. (2011). Sentence comprehension as mental simulation: an information-theoretic perspective. *Information*, *2*, 672–696.

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

Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentenial complexity. *Cognitive Science*, *29*, 261–290.

Hale, J. T. (2001). A probabilistic Early parser as a psycholinguistic model. In *Proceedings of the second conference of the North American chapter of the Association for Computational Linguistics* (Vol. 2, pp. 159–166). Pittsburgh, PA: Association for Computational Linguistics.

Hale, J. T. (2003). The information conveyed by words. *Journal of Psycholinguistic Research*, *32*, 101–123.

Hale, J. T. (2006). Uncertainty about the rest of the sentence. *Cognitive Science*, *30*, 643–672.

Hale, J. T. (2011). What a rational parser would do. *Cognitive Science*, *35*, 399–443.

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

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, *111*, 228–238.

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

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

Marr, D. (1982). *Vision*. San Francisco: W.H. Freeman and Company.

Mitchell, J., Lapata, M., Demberg, V., & Keller, F. (2010). Syntactic and semantic factors in processing difficulty: an integrated measure. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (p. 196-206). Uppsala, Sweden: Association for Computational Linguistics.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*, 280–291.

Piccinini, G., & Scarantino, A. (2011). Information processing, computation, and cognition. *Journal of Biological Physics*, *37*, 1–38.

Roark, B., Bachrach, A., Cardenas, C., & Pallier, C. (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* (pp. 324–333). Association for Computational Linguistics.

Santorini, B. (1991). *Part-of-speech tagging guidelines for the Penn Treebank project* (Tech. Rep.). Philadelphia, PA: University of Pennsylvania.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423, 623–656.

Smith, N. J., & Levy, R. (2008). Optimal processing times in reading: a formal model and empirical investigation. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 595–600). Austin, TX: Cognitive Science Society.

Tsuruoka, Y., & Tsujii, J. (2005). Bidirectional inference with the easiest-first strategy for tagging sequence data. In *Proceedings of the conference on human language technology and empirical methods in natural language processing* (pp. 467–474). Association for Computational Linguistics, Morristown, NJ.

Wu, S., Bachrach, A., Cardenas, C., & Schuler, W. (2010). Complexity metrics in an incremental right-corner parser. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 1189–1198). Association for Computational Linguistics.

A Language model training details

Stage 1 To obtain vector representations of words, a matrix of co-occurrence frequencies was constructed in which each value at (w, v) equals the number of times word type v directly precedes (for $v \leq 7754$) or follows (for $v > 7754$) word type w. These frequencies were turned into probabilities $p_{w,v}$ using Simple Good-Turing smoothing (Gale & Sampson, 1995). Next, the probabilities were transformed into pointwise mutual information values: $f_{w,v} = \log(p_{w,v}) - \log(p_w p_v)$, which Bullinaria and Levy (2007) found to result in vectors that perform well on a variety of tasks. Finally, the 400 columns with the highest variance were selected from the 7754×15508-matrix formed by the values $f_{w,v}$, resulting in a 400-dimensional word-representation space.

Stage 2 The SRN was trained using basic backpropagation (i.e., without momentum or backpropagation-through-time) with MSE as the minimized error measure. Hidden units had logistic activation function but output units were linear. The training data was presented five times (each time with a different random ordering of sentences) at which point the MSE over the experimental sentences no longer decreases. The learning rate was set to 10^{-6} during the first four presentations and was lowered to 2×10^{-7} for the fifth.

Stage 3 The decoder network was trained using backpropagation, minimizing cross-entropy. Hidden units were logistic, but output units had softmax activation functions. The training data (i.e., the trained Stage-2 outputs resulting from the training sentences) were presented twice, the first time with a learning rate of 0.001 and the second time with 0.0002. Connection-weight decay was applied in order to prevent overfitting: After each weight update, the weights were reduced by a small fraction (0.001 times the learning rate) of themselves.

B Regression model

To construct the baseline regression model, a number of predictors were chosen that are likely to account for reading times. These factors, all of which were standardized, are listed in Table 3. In addition, quadratic factors for SentPos and WordPos were present, as well as by-item and by-subject random intervals. Next, all two-way interactions were included and the model was fitted to the log-transformed reading times. Insignificant interactions were removed one at a time (starting with the least significant) until all remaining interactions were significant (i.e., $|t| > 2$). Following this, the three random slopes with the strongest effects were included. These were by-subject random slopes of PrevRT, SentPos and $(\text{SentPos})^2$. Finally, one more interaction was removed because it lost significance as a consequence of including the random slopes. Table 4 shows the factors and coefficients of the resulting baseline model.

Table 3: Predictors of word-reading time.

Abbreviation	Description
SentPos	position of sentence in presentation order
WordPos	position of word in sentence
Length	number of letters in word
PrevLength	number of letters in previous word
Freq	log of relative word frequency in written-text part of full BNC
PrevFreq	log of relative frequency of previous word
ForwProb	log of forward transitional probability ($\log P(w_{t+1}\|w_t)$)
	based on frequencies in written-text part of full BNC
PrevForwProb	previous word's forward transitional probability ($\log P(w_t\|w_{t-1})$)
PrevRT	reading time on previous word

Table 4: Fixed effects in the fitted baseline regression model (left) and in the model including surprisal and ΔH_4 (right).

Factor	Coeff ($\times 10^3$)	t	Coeff ($\times 10^3$)	t
Intercept	5542.48	252.88	5543.05	252.74
Surprisal			12.59	6.25
ΔH_4			3.52	4.67
SentPos	−85.04	−10.23	−85.10	−10.25
$(\text{SentPos})^2$	24.15	3.10	24.19	3.11
WordPos	3.65	3.43	3.54	3.31
$(\text{WordPos})^2$	−1.97	−3.87	−1.91	−3.73
Length	−0.24	−0.19	−0.63	−0.49
PrevLength	0.69	0.61	0.88	0.78
Freq	1.69	1.11	0.37	0.24
PrevFreq	−5.06	−3.20	0.84	0.47
ForwProb	−6.32	−4.69	−4.64	−3.41
PrevForwProb	−9.41	−6.48	−2.53	−1.32
PrevRT	109.27	20.74	109.16	20.71
SentPos × $(\text{SentPos})^2$	−17.57	−21.73	−17.58	−21.74
SentPos × Length	−5.44	−7.42	−5.43	−7.41
SentPos × PrevForwProb	4.85	6.06	4.87	6.10
SentPos × PrevRT	−11.04	−10.04	−10.99	−10.00
$(\text{SentPos})^2\times$ WordPos	1.73	3.36	1.73	3.36
$(\text{SentPos})^2\times$ Length	2.69	4.85	2.68	4.82
$(\text{SentPos})^2\times$ PrevForwProb	1.63	2.92	1.62	2.90
WordPos × ForwProb	−1.81	−2.59	−2.04	−2.94
WordPos × PrevRT	−4.58	−6.81	−4.59	−6.82
Length × PrevLength	−1.67	−2.31	−1.62	−2.25
Length × Freq	−5.85	−4.13	−5.60	−3.98
Length × ForwProb	2.76	2.11	2.30	1.76
PrevLength × PrevFreq	−3.40	−3.79	−2.37	−2.62
PrevLength × PrevRT	−2.00	−2.22	−1.96	−2.18
PrevFreq × PrevRT	−4.82	−3.55	−4.65	−3.42
ForwProb × PrevRT	−5.58	−7.94	−5.58	−7.95
PrevForwProb × PrevRT	3.14	2.52	3.05	2.45