Lexical surprisal as a general predictor of reading time

Irene Fernandez Monsalve, Stefan L. Frank and **Gabriella Vigliocco**
Division of Psychology and Language Sciences
University College London
{ucjtife, s.frank, g.vigliocco}@ucl.ac.uk

Abstract

Probabilistic accounts of language processing can be psychologically tested by comparing word-reading times (RT) to the conditional word probabilities estimated by language models. Using surprisal as a linking function, a significant correlation between unlexicalized surprisal and RT has been reported (e.g., Demberg and Keller, 2008), but success using lexicalized models has been limited. In this study, phrase structure grammars and recurrent neural networks estimated both lexicalized and unlexicalized surprisal for words of independent sentences from narrative sources. These same sentences were used as stimuli in a self-paced reading experiment to obtain RTs. The results show that lexicalized surprisal according to both models is a significant predictor of RT, outperforming its unlexicalized counterparts.

1 Introduction

Context-sensitive, prediction-based processing has been proposed as a fundamental mechanism of cognition (Bar, 2007): Faced with the problem of responding in real-time to complex stimuli, the human brain would use basic information from the environment, in conjunction with previous experience, in order to extract meaning and anticipate the immediate future. Such a cognitive style is a well-established finding in low level sensory processing (e.g., Kveraga et al., 2007), but has also been proposed as a relevant mechanism in higher order processes, such as language. Indeed, there is ample evidence to show that human language comprehension is both incremental and predictive. For example, on-line detection of semantic or syntactic anomalies can be observed in the brain's EEG signal (Hagoort et al., 2004) and eye gaze is directed in anticipation at depictions of plausible sentence completions (Kamide et al., 2003). Moreover, probabilistic accounts of language processing have identified *unpredictability* as a major cause of processing difficulty in language comprehension. In such incremental processing, parsing would entail a pre-allocation of resources to expected interpretations, so that effort would be related to the suitability of such an allocation to the actually encountered stimulus (Levy, 2008).

Possible sentence interpretations can be constrained by both linguistic and extra-linguistic context, but while the latter is difficult to evaluate, the former can be easily modeled: The predictability of a word for the human parser can be expressed as the conditional probability of a word given the sentence so far, which can in turn be estimated by language models trained on text corpora. These probabilistic accounts of language processing difficulty can then be validated against empirical data, by taking reading time (RT) on a word as a measure of the effort involved in its processing.

Recently, several studies have followed this approach, using "surprisal" (see Section 1.1) as the linking function between effort and predictability. These can be computed for each word in a text, or alternatively for the words' parts of speech (POS). In the latter case, the obtained estimates can give an indication of the importance of syntactic structure in developing upcoming-word expectations, but ignore the rich lexical information that is doubtlessly employed by the human parser

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 398–408,
Avignon, France, April 23 - 27 2012. ©2012 Association for Computational Linguistics

to constrain predictions. However, whereas such an *unlexicalized* (i.e., POS-based) surprisal has been shown to significantly predict RTs, success with *lexical* (i.e., word-based) surprisal has been limited. This can be attributed to data sparsity (larger training corpora might be needed to provide accurate lexical surprisal than for the unlexicalized counterpart), or to the noise introduced by participant's world knowledge, inaccessible to the models. The present study thus sets out to find such a lexical surprisal effect, trying to overcome possible limitations of previous research.

1.1 Surprisal theory

The concept of surprisal originated in the field of information theory, as a measure of the amount of information conveyed by a particular event. Improbable ('surprising') events carry more information than expected ones, so that surprisal is inversely related to probability, through a logarithmic function. In the context of sentence processing, if $w_1, ..., w_{t-1}$ denotes the sentence so far, then the cognitive effort required for processing the next word, w_t, is assumed to be proportional to its surprisal:

$$\begin{aligned} \text{effort}(t) &\propto \text{surprisal}(w_t) \\ &= -\log(P(w_t|w_1, ..., w_{t-1})) \end{aligned} \quad (1)$$

Different theoretical groundings for this relationship have been proposed (Hale, 2001; Levy 2008; Smith and Levy, 2008). Smith and Levy derive it by taking a scale free assumption: Any linguistic unit can be subdivided into smaller entities (e.g., a sentence is comprised of words, a word of phonemes), so that time to process the whole will equal the sum of processing times for each part. Since the probability of the whole can be expressed as the product of the probabilities of the subunits, the function relating probability and effort must be logarithmic. Levy (2008), on the other hand, grounds surprisal in its information-theoretical context, describing difficulty encountered in on-line sentence processing as a result of the need to update a probability distribution over possible parses, being directly proportional to the difference between the previous and updated distributions. By expressing the difference between these in terms of relative entropy, Levy shows that difficulty at each newly encountered word should be equal to its surprisal.

1.2 Empirical evidence for surprisal

The simplest statistical language models that can be used to estimate surprisal values are n-gram models or Markov chains, which condition the probability of a given word only on its $n-1$ preceding ones. Although Markov models theoretically limit the amount of prior information that is relevant for prediction of the next step, they are often used in linguistic context as an approximation to the full conditional probability. The effect of bigram probability (or forward transitional probability) has been repeatedly observed (e.g. McDonald and Shillcock, 2003), and Smith and Levy (2008) report an effect of lexical surprisal as estimated by a trigram model on RTs for the Dundee corpus (a collection of newspaper texts with eye-tracking data from ten participants; Kennedy and Pynte, 2005).

Phrase structure grammars (PSGs) have also been amply used as language models (Boston et al., 2008; Brouwer et al., 2010; Demberg and Keller, 2008; Hale, 2001; Levy, 2008). PSGs can combine statistical exposure effects with explicit syntactic rules, by annotating norms with their respective probabilities, which can be estimated from occurrence counts in text corpora. Information about hierarchical sentence structure can thus be included in the models. In this way, Brouwer et al. trained a probabilistic context-free grammar (PCFG) on 204,000 sentences extracted from Dutch newspapers to estimate lexical surprisal (using an Earley-Stolcke parser; Stolcke, 1995), showing that it could account for the noun phrase coordination bias previously described and explained by Frazier (1987) in terms of a minimal-attachment preference of the human parser. In contrast, Demberg and Keller used texts from a naturalistic source (the Dundee corpus) as the experimental stimuli, thus evaluating surprisal as a wide-coverage account of processing difficulty. They also employed a PSG, trained on a one-million-word language sample from the Wall Street Journal (part of the Penn Treebank II, Marcus et al., 1993). Using Roark's (2001) incremental parser, they found significant effects of unlexicalized surprisal on RTs (see also Boston et al. for a similar approach and results for German texts). However, they failed to find an effect for lexicalized surprisal, over and above forward transitional probability. Roark et al. (2009) also looked at the effects of *syntactic* and *lexical* surprisal, using RT data for short narrative texts. However, their estimates of these two surprisal values differ from those described above: In order to tease apart semantic and syntactic effects, they used Demberg and Keller's lexicalized surprisal as a total surprisal measure, which they decompose into syntactic and lexical components. Their results show significant effects of both syntactic and lexical surprisal, although the latter was found to hold only for closed class words. Lack of a wider effect was attributed to data sparsity: The models were trained on the relatively small Brown corpus (over one million words from 500 samples of American English text), so that surprisal estimates for the less frequent content words would not have been accurate enough.

Using the same training and experimental language samples as Demberg and Keller (2008), and only unlexicalized surprisal estimates, Frank (2009) and Frank and Bod (2011) focused on comparing different language models, including various *n*-gram models, PSGs and recurrent networks (RNN). The latter were found to be the better predictors of RTs, and PSGs could not explain any variance in RT over and above the RNNs, suggesting that human processing relies on linear rather than hierarchical representations.

Summing up, the only models taking into account actual words that have been consistently shown to simulate human behaviour with naturalistic text samples are bigram models.[1] A possible limitation in previous studies can be found in the stimuli employed. In reading real newspaper texts, prior knowledge of current affairs is likely to highly influence RTs, however, this source of variability cannot be accounted for by the models. In addition, whereas the models treat each sentence as an independent unit, in the text corpora employed they make up coherent texts, and are therefore clearly dependent. Thirdly, the stimuli used by Demberg and Keller (2008) comprise a very particular linguistic style: journalistic editorials, reducing the ability to generalize conclusions to language in general. Finally, failure to find lexical surprisal effects can also be attributed to the training texts. Larger corpora are likely to be needed for training language models on actual words than on POS (both the Brown corpus and the WSJ are relatively small), and in addition, the particular journalistic style of the WSJ might not be the best alternative for modeling human behaviour. Although similarity between the training and experimental data sets (both from newspaper sources) can improve the linguistic performance of the models, their ability to simulate human behaviour might be limited: Newspaper texts probably form just a small fraction of a person's linguistic experience. This study thus aims to tackle some of the identified limitations: Rather than cohesive texts, independent sentences, from a narrative style are used as experimental stimuli for which word-reading times are collected (as explained in Section 3). In addition, as discussed in the following section, language models are trained on a larger corpus, from a more representative language sample. Following Frank (2009) and Frank and Bod (2011), two contrasting types of models are employed: hierarchical PSGs and linear RNNs.

2 Models

2.1 Training data

The training texts were extracted from the written section of the British National Corpus (BNC), a collection of language samples from a variety of sources, designed to provide a comprehensive representation of current British English. A total of 702,412 sentences, containing only the 7,754 most frequent words (the open-class words used by Andrews et al., 2009, plus the 200 most frequent words in English) were selected, making up a 7.6-million-word training corpus. In addition to providing a larger amount of data than the WSJ, this training set thus provides a more representative language sample.

2.2 Experimental sentences

Three hundred and sixty-one sentences, all comprehensible out of context and containing only words included in the subset of the BNC used to train the models, were randomly selected from three freely accessible on-line novels[2] (for additional details, see Frank, 2012). The fictional narrative provides a good contrast to the pre-

[1] Although Smith and Levy (2008) report an effect of trigrams, they did not check if it exceeded that of simpler bigrams.

[2] Obtained from `www.free-online-novels.com`. Having not been published elsewhere, it is unlikely participants had read the novels previously.

viously examined newspaper editorials from the Dundee corpus, since participants did not need prior knowledge regarding the details of the stories, and a less specialised language and style were employed. In addition, the randomly selected sentences did not make up coherent texts (in contrast, Roark et al., 2009, employed short stories), so that they were independent from each other, both for the models and the readers.

2.3 Part-of-speech tagging

In order to produce POS-based surprisal estimates, versions of both the training and experimental texts with their words replaced by POS were developed: The BNC sentences were parsed by the Stanford Parser, version 1.6.7 (Klein and Manning, 2003), whilst the experimental texts were tagged by an automatic tagger (Tsuruoka and Tsujii, 2005), with posterior review and correction by hand following the Penn Treebank Project Guidelines (Santorini, 1991). By training language models and subsequently running them on the POS versions of the texts, unlexicalized surprisal values were estimated.

2.4 Phrase-structure grammars

The Treebank formed by the parsed BNC sentences served as training data for Roark's (2001) incremental parser. Following Frank and Bod (2011), a range of grammars was induced, differing in the features of the tree structure upon which rule probabilities were conditioned. In four grammars, probabilities depended on the left-hand side's ancestors, from one up to four levels up in the parse tree (these grammars will be denoted a_1 to a_4). In four other grammars (s_1 to s_4), the ancestors' left siblings were also taken into account. In addition, probabilities were conditioned on the current head node in all grammars. Subsequently, Roark's (2001) incremental parser parsed the experimental sentences under each of the eight grammars, obtaining eight surprisal values for each word. Since earlier research (Frank, 2009) showed that decreasing the parser's base beam width parameter improves performance, it was set to 10^{-18} (the default being 10^{-12}).

2.5 Recurrent neural network

The RNN (see Figure 1) was trained in three stages, each taking the selected (unparsed) BNC sentences as training data.

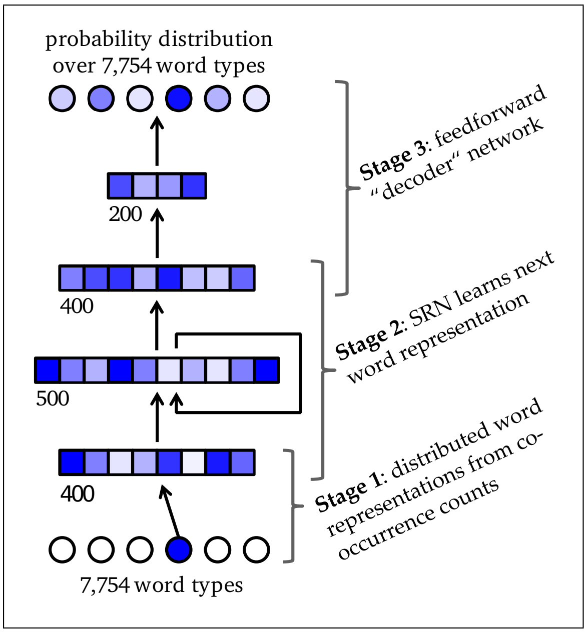
probability distribution over 7,754 word types
200
400
500
400
7,754 word types
Stage 3: feedforward "decoder" network
Stage 2: SRN learns next word representation
Stage 1: distributed word representations from co-occurrence counts

Figure 1: Architecture of neural network language model, and its three learning stages. Numbers indicate the number of units in each network layer.

Stage 1: Developing word representations

Neural network language models can benefit from using distributed word representations: Each word is assigned a vector in a continuous, high-dimensional space, such that words that are paradigmatically more similar are closer together (e.g., Bengio et al., 2003; Mnih and Hinton, 2007). Usually, these representations are learned together with the rest of the model, but here we used a more efficient approach in which word representations are learned in an unsupervised manner from simple co-occurrences in the training data. First, vectors of word co-occurrence frequencies were developed using Good-Turing (Gale and Sampson, 1995) smoothed frequency counts from the training corpus. Values in the vector corresponded to the smoothed frequencies with which each word directly preceded or followed the represented word. Thus, each word w was assigned a vector $(f_{w,1}, ..., f_{w,15508})$, such that $f_{w,v}$ is the number of times word v directly precedes (for $v \leq 7754$) or follows (for $v > 7754$) word w. Next, the frequency counts were transformed into Pointwise Mutual Information (PMI) values (see Equation 2), following Bullinaria and Levy's (2007) findings that PMI produced more psychologically accurate predictions than other measures:

$$\mathrm{PMI}(w,v) = \log\left(\frac{f_{w,v}\sum_{i,j} f_{i,j}}{\sum_i f_{i,v}\sum_j f_{w,j}}\right) \qquad (2)$$

Finally, the 400 columns with the highest variance were selected from the 7754×15508-matrix of row vectors, making them more computationally manageable, but not significantly less informative.

Stage 2: Learning temporal structure

Using the standard backpropagation algorithm, a simple recurrent network (SRN) learned to predict, at each point in the training corpus, the next word's vector given the sequence of word vectors corresponding to the sentence so far. The total corpus was presented five times, each time with the sentences in a different random order.

Stage 3: Decoding predicted word representations

The distributed output of the trained SRN served as training input to the feedforward "decoder" network, that learned to map the distributed representations back to localist ones. This network, too, used standard backpropagation. Its output units had softmax activation functions, so that the output vector constitutes a probability distribution over word types. These translate directly into surprisal values, which were collected over the experimental sentences 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). These will be denoted by RNN-1 to RNN-10.

A much simpler RNN model suffices for obtaining unlexicalized surprisal. Here, we used the same models as described by Frank and Bod (2011), albeit trained on the POS tags of our BNC training corpus. These models employed so-called Echo State Networks (ESN; Jaeger and Haas, 2004), which are RNNs that do not develop internal representations because weights of input and recurrent connections remain fixed at random values (only the output connection weights are trained). Networks of six different sizes were used. Of each size, three networks were trained, using different random weights. The best and worst model of each size were discarded to reduce the effect of the random weights.

3 Experiment

3.1 Procedure

Text display followed a self-paced reading paradigm: Sentences were presented on a computer screen one word at a time, with onset of the next word being controlled by the subject through a key press. The time between word onset and subsequent key press was recorded as the RT (measured in milliseconds) on that word by that subject.[3] Words were presented centrally aligned in the screen, and punctuation marks appeared with the word that preceded them. A fixed-width font type (Courier New) was used, so that physical size of a word equalled number of characters. Order of presentation was randomized for each subject. The experiment was time-bounded to 40 minutes, and the number of sentences read by each participant varied between 120 and 349, with an average of 224. Yes-no comprehension questions followed 46% of the sentences.

3.2 Participants

A total of 117 first year psychology students took part in the experiment. Subjects unable to answer correctly to more than 20% of the questions and 47 participants who were non-native English speakers were excluded from the analysis, leaving a total of 54 subjects.

3.3 Design

The obtained RTs served as the dependent variable against which a mixed-effects multiple regression analysis with crossed random effects for subjects and items (Baayen et al., 2008) was performed. In order to control for low-level lexical factors that are known to influence RTs, such as word length or frequency, a baseline regression model taking them into account was built. Subsequently, the decrease in the model's deviance, after the inclusion of surprisal as a fixed factor to the baseline, was assessed using likelihood tests. The resulting χ^2 statistic indicates the extent to which each surprisal estimate accounts for RT, and can thus serve as a measure of the psychological accuracy of each model.

However, this kind of analysis assumes that RT for a word reflects processing of only that word,

[3] The collected RT data are available for download at `www.stefanfrank.info/EACL2012`.

but spill-over effects (in which processing difficulty at word w_t shows up in the RT on w_{t+1}) have been found in self-paced and natural reading (Just et al., 1982; Rayner, 1998; Rayner and Pollatsek, 1987). To evaluate these effects, the decrease in deviance after adding surprisal of the *previous* item to the baseline was also assessed.

The following control predictors were included in the baseline regression model:

Lexical factors:

- *Number of characters:* Both physical size and number of characters have been found to affect RTs for a word (Rayner and Pollatsek, 1987), but the fixed-width font used in the experiment assured number of characters also encoded physical word length.
- *Frequency and forward transitional probability:* The effects of these two factors have been repeatedly reported (e.g. Juhasz and Rayner, 2003; Rayner, 1998). Given the high correlations between surprisal and these two measures, their inclusion in the baseline assures that the results can be attributed to predictability in context, over and above frequency and bigram probability. Frequency was estimated from occurrence counts of each word in the full BNC corpus (written section). The same transformation (negative logarithm) was applied as for computing surprisal, thus obtaining “unconditional” and bigram surprisal values.
- *Previous word lexical factors:* Lexical factors for the previous word were included in the analysis to control for spill-over effects.

Temporal factors and autocorrelation:

RT data over naturalistic texts violate the regression assumption of independence of observations in several ways, and important word-by-word sequential correlations exist. In order to ensure validity of the statistical analysis, as well as providing a better model fit, the following factors were also included:

- *Sentence position:* Fatigue and practice effects can influence RTs. Sentence position in the experiment was included both as linear and quadratic factor, allowing for the modeling of initial speed-up due to practice, followed by a slowing down due to fatigue.
- *Word position:* Low-level effects of word order, not related to predictability itself, were modeled by including word position in the sentence, both as a linear and quadratic factor (some of the sentences were quite long, so that the effect of word position is unlikely to be linear).
- *Reading time for previous word:* As suggested by Baayen and Milin (2010), including RT on the previous word can control for several autocorrelation effects.

4 Results

Data were analysed using the free statistical software package R (R Development Core Team, 2009) and the lme4 library (Bates et al., 2011). Two analyses were performed for each language model, using surprisal for either current or previous word as the dependent variable. Unlikely reading times (lower than 50ms or over 3000ms) were removed from the analysis, as were clitics, words followed by punctuation, words following punctuation or clitics (since factors for previous word were included in the analysis), and sentence-initial words, leaving a total of 132,298 data points (between 1,335 and 3,829 per subject).

4.1 Baseline model

Theoretical considerations guided the selection of the initial predictors presented above, but an empirical approach led actual regression model building. Initial models with the original set of fixed effects, all two-way interactions, plus random intercepts for subjects and items were evaluated, and least significant factors were removed one at a time, until only significant predictors were left ($|t| > 2$). A different strategy was used to assess which by-subject and by item random slopes to include in the model. Given the large number of predictors, starting from the saturated model with all random slopes generated non-convergence problems and excessively long running times. By-subject and by-item random slopes for each fixed effect were therefore assessed individually, using likelihood tests. The final baseline model included by-subject random intercepts, by-subject random slopes for sentence position and word position, and by-item slopes for previous RT. All factors (random slopes and fixed effects) were centred and standardized to avoid

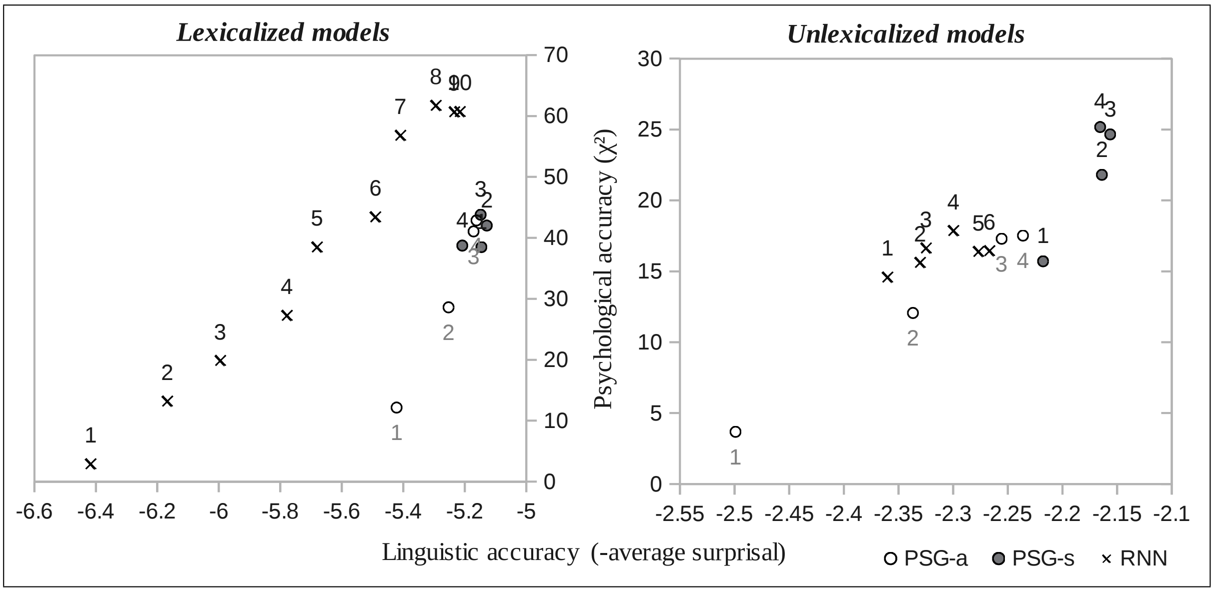

Figure 2: Psychological accuracy (combined effect of current and previous surprisal) against linguistic accuracy of the different models. Numbered labels denote the maximum number of levels up in the tree from which conditional information is used (PSG); point in training when estimates were collected (word-based RNN); or network size (POS-based RNN).

multicollinearity-related problems.

4.2 Surprisal effects

All model categories (PSGs and RNNs) produced lexicalized surprisal estimates that led to a significant ($p < 0.05$) decrease in deviance when included as a fixed factor in the baseline, with positive coefficients: Higher surprisal led to longer RTs. Significant effects were also found for their unlexicalized counterparts, albeit with considerably smaller χ^2-values.

Both for the lexicalized and unlexicalized versions, these effects persisted whether surprisal for the previous or current word was taken as the independent variable. However, the effect size was much larger for previous surprisal, indicating the presence of strong spill-over effects (e.g. lexicalized PSG-s$_3$: current surprisal: $\chi^2(1) = 7.29$, $p = 0.007$; previous surprisal: $\chi^2(1) = 36.73$, $p \ll 0.001$).

From hereon, only results for the combined effect of both (inclusion of previous and current surprisal as fixed factors in the baseline) are reported. Figure 2 shows the psychological accuracy of each model ($\chi^2(2)$ values) plotted against its linguistic accuracy (i.e., its quality as a language model, measured by the negative average surprisal on the experimental sentences: the higher this value, the "less surprised" the model is by the test corpus). For the lexicalized models, RNNs clearly outperform PSGs. Moreover, the RNN's accuracy increases as training progresses (the highest psychological accuracy is achieved at point 8, when 350K training sentences were presented). The PSGs taking into account sibling nodes are slightly better than their ancestor-only counterparts (the best psychological model is PSG-s$_3$). Contrary to the trend reported by Frank and Bod (2011), the unlexicalized PSGs and RNNs reach similar levels of psychological accuracy, with the PSG-s$_4$ achieving the highest χ^2-value.

Model comparison	$\chi^2(2)$	p**-value**
PSG over RNN	12.45	0.002
RNN over PSG	30.46	$\ll$0.001

Table 1: Model comparison between best performing word-based PSG and RNN.

Although RNNs outperform PSGs in the lexicalized estimates, comparisons between the best performing model (i.e. highest χ^2) in each category showed both were able to explain variance over and above each other (see Table 1). It is worth noting, however, that if comparisons are made amongst models including surprisal for current, but not previous word, the PSG is unable to explain a significant amount of variance over and above the RNN ($\chi^2(1) = 2.28; p = 0.13$).[4] Lexicalized models achieved greater psychological accuracy than their unlexicalized counterparts, but the latter could still explain a small amount of variance over and above the former (see Table 2).[5]

Model comparison	$\chi^2(2)$	*p*-value
Best models overall:		
POS- over word-based	10.40	0.006
word- over POS-based	47.02	≪0.001
PSGs:		
POS- over word-based	6.89	0.032
word- over POS-based	25.50	≪0.001
RNNs:		
POS- over word-based	5.80	0.055
word- over POS-based	49.74	≪0.001

Table 2: Word- vs. POS-based models: comparisons between best models overall, and best models within each category.

4.3 Differences across word classes

In order to make sure that the lexicalized surprisal effects found were not limited to closed-class words (as Roark et al., 2009, report), a further model comparison was performed by adding by-POS random slopes of surprisal to the models containing the baseline plus surprisal. If particular syntactic categories were contributing to the overall effect of surprisal more than others, including such random slopes would lead to additional variance being explained. However, this was not the case: inclusion of by-POS random slopes of surprisal did not lead to a significant improvement in model fit (PSG: $\chi^2(1) = 0.86, p = 0.35$; RNN: $\chi^2(1) = 3.20, p = 0.07$).[6]

5 Discussion

The present study aimed to find further evidence for surprisal as a wide-coverage account of language processing difficulty, and indeed, the results show the ability of lexicalized surprisal to explain a significant amount of variance in RT data for naturalistic texts, over and above that accounted for by other low-level lexical factors, such as frequency, length, and forward transitional probability. Although previous studies had presented results supporting such a probabilistic language processing account, evidence for word-based surprisal was limited: Brouwer et al. (2010) only examined a specific psycholinguistic phenomenon, rather than a random language sample; Demberg and Keller (2008) reported effects that were only significant for POS but not word-based surprisal; and Smith and Levy (2008) found an effect of lexicalized surprisal (according to a trigram model), but did not assess whether simpler predictability estimates (i.e., by a bigram model) could have accounted for those effects.

Demberg and Keller's (2008) failure to find lexicalized surprisal effects can be attributed both to the language corpus used to train the language models, as well as to the experimental texts used. Both were sourced from newspaper texts: As training corpora these are unrepresentative of a person's linguistic experience, and as experimental texts they are heavily dependent on participant's world knowledge. Roark et al. (2009), in contrast, used a more representative, albeit relatively small, training corpus, as well as narrative-style stimuli, thus obtaining RTs less dependent on participant's prior knowledge. With such an experimental set-up, they were able to demonstrate the effects of lexical surprisal for RT of closed-class, but not open-class, words, which they attributed to their differential frequency and to training-data sparsity: The limited Brown corpus would have been enough to produce accurate estimates of surprisal for function words, but not for the less frequent content words. A larger training corpus, constituting a broad language sample, was used in our study, and the detected surprisal effects were shown to hold across syntactic category (modeling slopes for POS separately did not improve model fit). However, direct comparison with Roark et al.'s results is not possible: They employed alternative definitions of structural and lexical surprisal, which they derived by decomposing the total surprisal as obtained with a fully lexicalized PSG model.

In the current study, a similar approach to that taken by Demberg and Keller (2008) was used to

[4] Best models in this case were PSG-a_3 and RNN-7.

[5] Since best performing lexicalized and unlexicalized models belonged to different groups: RNN and PSG, respectively, Table 2 also shows comparisons within model type.

[6] Comparison was made on the basis of previous word surprisal (best models in this case were PSG-s_3 and RNN-9).

define structural (or unlexicalized), and lexicalized surprisal, but the results are strikingly different: Whereas Demberg and Keller report a significant effect for POS-based estimates, but not for word-based surprisal, our results show that lexicalized surprisal is a far better predictor of RTs than its unlexicalized counterpart. This is not surprising, given that while the unlexicalized models only have access to syntactic sources of information, the lexicalized models, like the human parser, can also take into account lexical co-occurrence trends. However, when a training corpus is not large enough to accurately capture the latter, it might still be able to model the former, given the higher frequency of occurrence of each possible item (POS vs. word) in the training data. Roark et al. (2009) also included in their analysis a POS-based surprisal estimate, which lost significance when the two components of the lexicalized surprisal were present, suggesting that such unlexicalized estimates can be interpreted only as a coarse version of the fully lexicalized surprisal, incorporating both syntactic and lexical sources of information at the same time. The results presented here do not replicate this finding: The best unlexicalized estimates were able to explain additional variance over and above the best word-based estimates. However, this comparison contrasted two different model types: a word-based RNN and a POS-based PSG, so that the observed effects could be attributed to the model representations (hierarchical vs. linear) rather than to the item of analysis (POS vs. words). Within-model comparisons showed that unlexicalized estimates were still able to account for additional variance, although only reaching significance at the 0.05 level for the PSGs.

Previous results reported by Frank (2009) and Frank and Bod (2011) regarding the higher psychological accuracy of RNNs and the inability of the PSGs to explain any additional variance in RT, were not replicated. Although for the word-based estimates RNNs outperform the PSGs, we found both to have independent effects. Furthermore, in the POS-based analysis, performance of PSGs and RNNs reaches similarly high levels of psychological accuracy, with the best-performing PSG producing slightly better results than the best-performing RNN. This discrepancy in the results could reflect contrasting reading styles in the two studies: natural reading of newspaper texts, or self-paced reading of independent, narrative sentences. The absence of global context, or the unnatural reading methodology employed in the current experiment, could have led to an increased reliance on hierarchical structure for sentence comprehension. The sources and structures relied upon by the human parser to elaborate upcoming-word expectations could therefore be task-dependent. On the other hand, our results show that the independent effects of word-based PSG estimates only become apparent when investigating the effect of surprisal of the previous word. That is, considering only the current word's surprisal, as in Frank and Bod's analysis, did not reveal a significant contribution of PSGs over and above RNNs. Thus, additional effects of PSG surprisal might only be apparent when spill-over effects are investigated by taking previous word surprisal as a predictor of RT.

6 Conclusion

The results here presented show that lexicalized surprisal can indeed model RT over naturalistic texts, thus providing a wide-coverage account of language processing difficulty. Failure of previous studies to find such an effect could be attributed to the size or nature of the training corpus, suggesting that larger and more general corpora are needed to model successfully both the structural and lexical regularities used by the human parser to generate predictions. Another crucial finding presented here is the importance of spill-over effects: Surprisal of a word had a much larger influence on RT of the following item than of the word itself. Previous studies where lexicalized surprisal was only analysed in relation to current RT could have missed a significant effect only manifested on the following item. Whether spill-over effects are as important for different RT collection paradigms (e.g., eye-tracking) remains to be tested.

Acknowledgments

The research presented here was funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant number 253803. The authors acknowledge the use of the UCL *Legion* High Performance Computing Facility, and associated support services, in the completion of this work.

References

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

Mark Andrews, Gabriella Vigliocco, and David P. Vinson. 2009. Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, 116:463–498.

R. Harald Baayen and Petar Milin. 2010. Analyzing reaction times. *International Journal of Psychological Research*, 3:12–28.

R. Harald Baayen, Doug J. Davidson, and Douglas M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59:390–412.

Moshe Bar. 2007. The proactive brain: using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, 11:280–289.

Douglas Bates, Martin Maechler, and Ben Bolker, 2011. *lme4: Linear mixed-effects models using S4 classes.* Available from: http://CRAN.R-project.org/package=lme4 (R package version 0.999375-39).

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Marisa Ferrara Boston, 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,*, 2:1–12.

Harm Brouwer, Hartmut Fitz, and John C. J. Hoeks. 2010. Modeling the noun phrase versus sentence coordination ambiguity in Dutch: evidence from surprisal theory. In *Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics*, pages 72–80, Stroudsburg, PA, USA.

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

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109:193–210.

Stefan L. Frank and Rens Bod. 2011. Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, 22:829–834.

Stefan L. Frank. 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*, pages 1139–1144, Austin, TX.

Stefan L. Frank. 2012. Uncertainty reduction as a measure of cognitive processing load in sentence comprehension. *Manuscript submitted for publication.*

Peter Hagoort, Lea Hald, Marcel Bastiaansen, and Karl Magnus Petersson. 2004. Integration of word meaning and world knowledge in language comprehension. *Science*, 304:438–441.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8, Stroudsburg, PA.

Herbert Jaeger and Harald Haas. 2004. Harnessing nonlinearity: predicting chaotic systems and saving energy in wireless communication. *Science*, pages 78–80.

Barbara J. Juhasz and Keith Rayner. 2003. Investigating the effects of a set of intercorrelated variables on eye fixation durations in reading. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 29:1312–1318.

Marcel A. Just, Patricia A. Carpenter, and Jacqueline D. Woolley. 1982. Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, 111:228–238.

Yuki Kamide, Christoph Scheepers, and Gerry T. M. Altmann. 2003. Integration of syntactic and semantic information in predictive processing: cross-linguistic evidence from German and English. *Journal of Psycholinguistic Research*, 32:37–55.

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

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics,*, pages 423–430.

Kestutis Kveraga, Avniel S. Ghuman, and Moshe Bar. 2007. Top-down predictions in the cognitive brain. *Brain and Cognition*, 65:145–168.

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

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19:313–330.

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

Andriy Mnih and Geoffrey Hinton. 2007. Three new graphical models for statistical language modelling. *Proceedings of the 25th International Conference of Machine Learning*, pages 641–648.

Keith Rayner and Alexander Pollatsek. 1987. Eye movements in reading: A tutorial review. In

M. Coltheart, editor, *Attention and performance XII: the psychology of reading.*, pages 327–362. Lawrence Erlbaum Associates, London, UK.

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

Brian Roark, 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: Volume 1 - Volume 1*, pages 324–333, Stroudsburg, PA.

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

Beatrice Santorini. 1991. Part-of-speech tagging guidelines for the Penn Treebank Project. Technical report, Philadelphia, PA.

Nathaniel J. Smith and Roger Levy. 2008. Optimal processing times in reading: a formal model and empirical investigation. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, pages 595–600, Austin,TX.

Andreas Stolcke. 1995. An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Computational linguistics*, 21:165–201.

Yoshimasa Tsuruoka and Jun'ichi Tsujii. 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*, pages 467–474, Stroudsburg, PA.