Word surprisal predicts N400 amplitude during reading

Stefan L. Frank[1,2] **Leun J. Otten**[3] **Giulia Galli**[3] **Gabriella Vigliocco**[2]
{s.frank, l.otten, g.galli, g.vigliocco}@ucl.ac.uk
[1]Centre for Language Studies, Radboud University Nijmegen
[2]Department of Cognitive, Perceptual and Brain Sciences, University College London
[3]Institute of Cognitive Neuroscience, University College London

Abstract

We investigated the effect of word surprisal on the EEG signal during sentence reading. On each word of 205 experimental sentences, surprisal was estimated by three types of language model: Markov models, probabilistic phrase-structure grammars, and recurrent neural networks. Four event-related potential components were extracted from the EEG of 24 readers of the same sentences. Surprisal estimates under each model type formed a significant predictor of the amplitude of the N400 component only, with more surprising words resulting in more negative N400s. This effect was mostly due to content words. These findings provide support for surprisal as a generally applicable measure of processing difficulty during language comprehension.

1 Introduction

Many studies of human language comprehension measure the brain's electrical activity during reading. Such electroencephalography (EEG) experiments have revealed that the EEG signal displays systematic variation in response to the appearance of each word. The different components that can be observed in this signal are known as event-related potentials (ERPs). Probably the most reliably observed (and most studied) of these components is a negative-going deflection at centroparietal electrodes that peaks at around 400 ms after word onset and is therefore referred to as the N400 component.

It is well known that the N400 increases in amplitude (i.e., becomes more negative) when the word leads to comprehension difficulty. To study the general relation between word predictability and the N400, Dambacher et al. (2006) obtained subjective word-probability estimates (so-called *cloze* probabilities) by asking participants to predict the upcoming word at each point in a large number of sentences. A different group of subjects read these same sentences while their EEG signal was recorded. Results showed a correlation between N400 amplitude and cloze probability: Less predictable words yielded stronger N400s.

We investigated whether similar results can be obtained using more objective, model-based word probabilities. For each word in a collection of English sentences, estimates of its *surprisal* (i.e., its negative log-transformed conditional probability: $-\log P(w_t|w_1,\ldots,w_{t-1})$) were generated by three types of language model: Markov (i.e., n-gram) models, phrase-structure grammars (PSGs), and recurrent neural networks (RNNs). Next, EEG signals of participants reading the same sentences were recorded. A comparison of word surprisal to different ERP components revealed that, indeed, N400 amplitude was predicted by surprisal values: More surprising words resulted in more negative N400s, at least for content words.

2 Language models

A range of models of each type was trained, allowing to investigate whether models that capture the language statistics more accurately also yield better predictions of ERP size. Such a relation is generally found in studies that use word-reading time as the dependent variable (Fernandez Monsalve et al., 2012; Frank and Bod, 2011; Frank and Thompson, 2012), providing additional support that these psychological data are indeed explained by the surprisal values and not by some confounding variable.

2.1 Corpus data

All models were trained on sentences from the written texts in the British National Corpus (BNC). First, the 10,000 word types with highest

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 878–883,
Sofia, Bulgaria, August 4-9 2013.

frequency were selected from the BNC. Next, all sentences were extracted that contained only those words. This resulted in a training corpus of 1.06 million sentences (12.6 million word tokens).

Each trained model estimated a surprisal value for each word of the 205 sentences (1931 word tokens) for which eye-tracking data are available in the UCL corpus of reading times (Frank et al., in press). These sentences, which were selected from three unpublished novels, only contained words from the 10,000 high-frequency word list.

2.2 Markov models

Markov models were trained with modified Kneser-Ney smoothing (Chen and Goodman, 1999) as implemented in SRILM (Stolcke, 2002). Model order was varied: $n = 2, 3, 4$. No unigram model was computed because word frequency was factored out during data analysis (see Section 4.2).

2.3 Recurrent neural networks

The RNN model architecture has been thoroughly described elsewhere (Fernandez Monsalve et al., 2012; Frank, in press) so it is not discussed here. The only difference with previous versions was that the current RNN was trained on a substantially larger data set with more word types. A range of RNN models was obtained by training on nine increasingly large subsets of the BNC data, comprising 2K, 5K, 10K, 20K, 50K, 100K, 200K, 400K, and all 1.06M sentences. In addition, the network was trained on the full set twice, making a total of ten instantiations of the RNN model.

2.4 Phrase-structure grammars

To prepare data for PSG training, the selected BNC sentences were parsed by the Stanford parser (Klein and Manning, 2003). The resulting treebank was divided into nine increasingly large subsets, equal to those used for RNN training.[1] Grammars were induced from these subsets using the algorithm by Roark (2001) with its standard settings. Next, surprisal values on the experimental sentences were generated by Roark's incremental parser. Since increasing the parser's beam width has been shown to improve both word-probability estimates and the fit to word-reading times (Frank, 2009), the parser's 'base beam threshold' parameter was reduced to 10^{-20}.

[1]Because not all experimental sentences could be parsed when the treebank comprised only 2K sentences, 1K sentences were added to the smallest subset.

3 EEG data collection

Twenty-four healthy, adult volunteers from the UCL Psychology subject pool took part in the reading study. Their EEG was recorded continuously from 32 channels during the presentation of 5 practice sentences and the 205 experimental items. Participants were asked to minimise blinks, eye movements, and head movements during sentence presentation.

Each sentence was preceded by a centrally presented fixation cross. As soon as the participant pressed a key, the cross was replaced by the sentence's first word, which was then automatically replaced by each subsequent word. Word presentation duration (in milliseconds) equalled $190 + 20k$, where k is the number of characters in the word (including any attached punctuation). After the word disappeared, there was a 390 ms interval before the next word appeared.

The sentences were presented in random order, one word at a time, always centrally located on the monitor. One-hundred and ten of the experimental sentences were followed by a yes/no-comprehension question, to ensure that participants tried to understand the sentences. All participants answered at least 80% of the comprehension questions correctly.

4 Data analysis

4.1 ERP components

Four ERP components of interest were identified from the literature on EEG and sentence reading: Early Left Anterior Negativity (ELAN), P200, N400, and a post-N400 positivity (PNP). Table 1 lists the corresponding time windows and approximate electrode sites.[2] For each component, the average electrode potential over the corresponding time window and electrodes was computed. These average ERP amplitudes served as the four dependent variables for data analysis.

The ELAN component is generally thought of as indicative of difficulty with constructing syntactic phrase structure (Friederici et al., 1999; Gunter et al., 1999; Neville et al., 1991). Hence, if any of the model types predicts ELAN size, we would expect this to be the PSG.

Dambacher et al. (2006) found effects of word frequency or length (which are strongly correlated

[2]The P600 component (Osterhout and Holcomb, 1992) was not included because the shortest interval between consecutive word onsets was only 600 ms.

Component	Time window	Location
ELAN	125–175 ms	left anterior
P200	140–200 ms	frontocentral
N400	300–500 ms	centroparietal
PNP	400–600 ms	frontopolar

Table 1: Investigated ERP components, their time windows, and approximate scalp locations.

and therefore difficult to tease apart) on the P200 amplitude. Since we factor out these two lexical factors in the analysis, we expect no additional effect of surprisal on P200.

If any of the components is sensitive to word surprisal, this is most likely to be the N400 as many studies have already shown that N400 amplitude depends on subjective word predictability (Dambacher et al., 2006; Kutas and Hillyard, 1984; Moreno et al., 2002). Whether an effect will appear on the PNP is more doubtful. Van Petten and Luka (2012) argue that word expectations that are confirmed result in reduced N400 size, whereas expectations that are *dis*confirmed increase the PNP. However, in a probabilistic setting, expectations are not all-or-nothing so there is no strict distinction between confirmation and disconfirmation. Nevertheless, surprisal effects on PNP may occur. Since the PNP has received relatively little attention, the component may not be such a reliable index of comprehension difficulty as the N400 has proven to be.

4.2 Regression analysis

Data were discarded on words attached to a comma, clitics, sentence-initial, and sentence-final words. Moreover, artifacts in the EEG data (mostly due to eye blinks) were identified and removed, leaving 32,010 analysed data points per investigated ERP component. For each data point and ERP component, a baseline potential was determined by averaging over the component's electrodes in the 100 ms leading up to word onset.

In order to quantify the fit of surprisal to ERP size, a linear mixed-effects regression model was fitted to each of the four ERPs, using the predictors: baseline potential, log-transformed word frequency, word length (number of characters), word position in the sentence, and sentence position in the experiment.[3] Also, all significant two-way interactions were included (main effects were removed if they were not significant and did not appear in any interaction). In addition, there were by-subject and by-item random intervals, as well as significant by-subject and by-item random slopes. Parameters for the correlation between random intercept and slope where also estimated, if they significantly contributed to model fit.

When the surprisal estimates by a particular language model are included in the analysis, the regression model's deviance decreases. The size of this decrease is the χ^2-statistic of a likelihood-ratio test for significance of the surprisal effect, and was taken as the measure of the surprisal values' fit to the ERP data.[4] Negative values will be used to indicate effects in the negative direction, that is, when higher surprisal results in more negative (or less positive) going ERP deflections.

5 Results

5.1 Surprisal effects

Figure 1 plots the fit of each model's surprisal estimates to ERP amplitude as a function of the average natural $\log P(w_t|w_1, \ldots, w_{t-1})$, which quantifies to what extent the model has acquired accurate language statistics.[5] For the ELAN, P200 and PNP components, there were no significant effects after correcting for multiple comparisons. In contrast, effects on the N400 were highly significant.

5.2 Model comparison

Table 2 shows results of pairwise comparisons between the best models of each type (i.e., those whose surprisal estimates fit the N400 data best). Clearly, RNN-based surprisal explains variance over and above each of the other two models whereas neither the n-gram nor the PSG model outperforms the RNN. Moreover, the RNN's surprisals explain a marginally significant ($\chi^2 = 3.47; p < .07$) amount of variance over and above the *combined* PSG and n-gram surprisals.

[3] For word and sentence position, both linear and squared factors were included in order to capture possible non-linear changes over the course of the sentence or experiment.

[4] This definition equals what Frank and Bod (2011) call 'psychological accuracy' in an analysis of reading times.

[5] This measure, which Frank and Bod (2011) call 'linguistic accuracy', equals the negative logarithm of the model's perplexity. Increasing the amount of training data (or the value of n) resulted in higher linguistic accuracy, except for the three PSG models trained on the smallest amounts of data. This shows that the models did not suffer from overfitting.

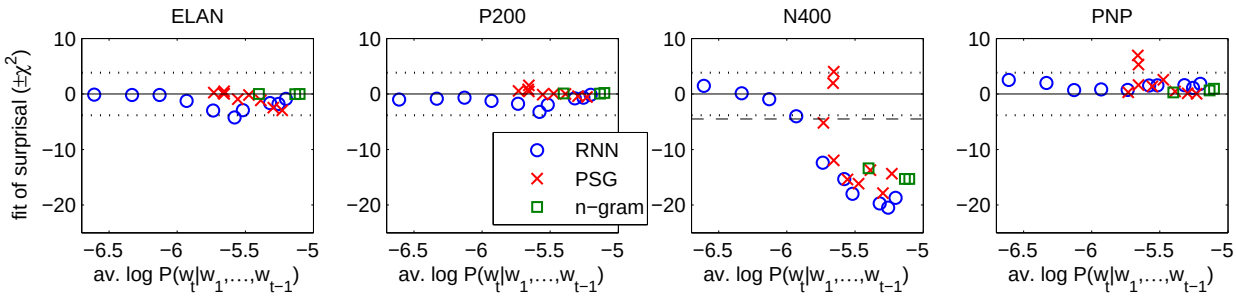

Figure 1: Fit to surprisal of ERP amplitude (for ELAN, P200, N400, and PNP components) as a function of average $\log P(w_t|w_1, \ldots, w_{t-1})$. Each plotted point corresponds to predictions by one of the trained models. Dotted lines indicate $\chi^2 = \pm 3.84$, beyond which effects are statistically significant ($p < .05$) if no correction for multiple comparisons is applied. The dashed line indicates the level below which effects are significant after applying the correction proposed by Benjamini and Hochberg (1995), on each ERP component separately because of our prior expectation that effects would occur mostly (if not exclusively) on the N400 component.

Model	n-**gram**	**RNN**	**PSG**
n-**gram**		$\chi^2 = 1.34$ $p > .2$	$\chi^2 = 1.66$ $p > .1$
RNN	$\chi^2 = 6.52$ $p < .02$		$\chi^2 = 4.78$ $p < .05$
PSG	$\chi^2 = 4.20$ $p < .05$	$\chi^2 = 2.14$ $p > .1$	

Table 2: Pairwise comparisons between surprisal estimates by the best models of each type. Shown are the results of likelihood-ratio tests for the effect of one set of surprisal estimates (rows) over and above the other (columns).

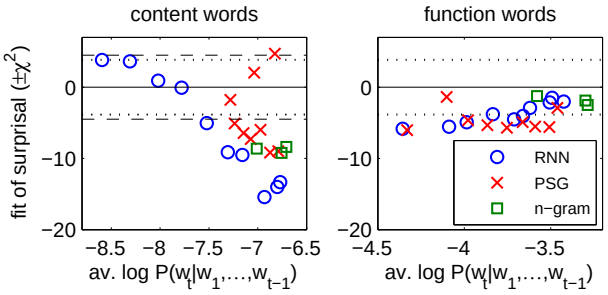

Figure 2: Fit to surprisal of N400 amplitude, for content words (left) and function words (right). Dotted lines indicate $\chi^2 = \pm 3.84$, beyond which effects are statistically significant ($p < .05$) without correcting for multiple comparisons. Dashed lines indicates the levels beyond which effects are significant after multiple-comparison correction (Benjamini and Hochberg, 1995).

5.3 Comparing word classes

N400 effects are nearly exclusively investigated on content (i.e., open-class) words. Dambacher et al. (2006), too, investigated the relation between ERP amplitudes and cloze probabilities on content words only. When running separate analyses on content and function words (constituting 53.2% and 46.8% of the data, respectively), we found that the N400 effect of Figure 1 is nearly fully driven by content words (see Figure 2). None of the models' surprisal estimates formed a significant predictor of N400 amplitude on function words, after correction for multiple comparisons.

6 Discussion

We demonstrated a clear effect of word surprisal, as estimated by different language models, on the EEG signal: The larger a (content) word's surprisal value, the more negative the resulting N400.

The N400 component is generally viewed as indicative of lexical rather than syntactic processing (Kaan, 2007), which may explain why surprisal under the PSG model did not have any significant explanatory value over and above RNN-based surprisal. The relatively weak performance of our Markov models is most likely due to their strict (and cognitively unrealistic) limit on the size of the prior context upon which word-probability estimates are conditioned.

Unlike the ELAN, P200, and PNP components, the N400 is known to be sensitive to the cloze probability of content words. The fact that surprisal effects were found on the N400 only, therefore suggests that subjective predictability scores and model-based surprisal estimates form opera-

tionalisations of one and the same underlying cognitive factor. Needless to say, our statistical models fail to capture many information sources, such as semantics and discourse, that do affect cloze probabilities. However, it is possible in principle to integrate these into probabilistic language models (Dubey et al., 2011; Mitchell et al., 2010).

To the best of our knowledge, only one other published study relates language model predictions to the N400: Parviz et al. (2011) found that surprisal estimates (corrected for word frequency) from an $n = 4$ Markov model predicted N400 size as measured by magnetoencephalography (rather than EEG). Although their PSG-based surprisals did not correlate with N400 size, a related measure derived from the PSG –lexical entropy– did. However, Parviz et al. (2011) only looked at effects on the sentence-final content word of items constructed for a speech perception experiment (Kalikow et al., 1977), rather than investigating surprisal's general predictive value across words of naturally occurring sentences, as we did here.

Our experimental design was parametric rather than factorial, which allowed us to study the effect of surprisal over a sample of English sentences rather than carefully manipulating surprisal while holding other factors constant. This has the advantage that our findings are likely to generalise to other sentence stimuli, but it can also raise a possible concern: The N400 effect may not be due to surprisal itself, but to an unknown confounding variable that was not included in the regression analysis. However, this seems unlikely because of two additional findings that only follow naturally if surprisal is indeed the relevant predictor: Significant results only appeared where they were most expected a priori (i.e., on N400 but not on other components) and there was a nearly monotonic relation between the models' word-prediction accuracy and their ability to account for N400 size.

7 Conclusion

Although word surprisal has often been shown to be predictive of word-reading time (Fernandez Monsalve et al., 2012; Frank and Thompson, 2012; Smith and Levy, in press), a general effect on the EEG signal has not before been demonstrated. Hence, these results provide additional evidence in support of surprisal as a reliable measure of cognitive processing difficulty during sentence comprehension (Hale, 2001; Levy, 2008).

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

Y. Benjamini and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57:289–300.

S. F. Chen and J. Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, 13:359–394.

M. Dambacher, R. Kliegl, M. Hofmann, and A. M. Jacobs. 2006. Frequency and predictability effect on event-related potentials during reading. *Brain Research*, 1084:89–103.

A. Dubey, F. Keller, and P. Sturt. 2011. A model of discourse predictions in human sentence processing. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 304–312. Edinburgh, UK: Association for Computational Linguistics.

I. Fernandez Monsalve, S. L. Frank, and G. Vigliocco. 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*, pages 398–408. Avignon, France: Association for Computational Linguistics.

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

S. L. Frank and R. L. Thompson. 2012. Early effects of word surprisal on pupil size during reading. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, pages 1554–1559. Austin, TX: Cognitive Science Society.

S. L. Frank, I. Fernandez Monsalve, R. L. Thompson, and G. Vigliocco. in press. Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods*.

S. L. Frank. 2009. Surprisal-based comparison between a symbolic and a connectionist model of sentence processing. In N. A. Taatgen and H. van Rijn, editors, *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, pages 1139–1144. Austin, TX: Cognitive Science Society.

S. L. Frank. in press. Uncertainty reduction as a measure of cognitive processing load in sentence comprehension. *Topics in Cognitive Science*.

A. D. Friederici, K. Steinhauer, and S. Frisch. 1999. Lexical integration: sequential effects of syntactic and semantic information. *Memory & Cognition*, 27:438–453.

T. C. Gunter, A. D. Friederici, and A. Hahne. 1999. Brain responses during sentence reading: Visual input affects central processes. *NeuroReport*, 10:3175–3178.

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

E. Kaan. 2007. Event-related potentials and language processing: a brief overview. *Language and Linguistics Compass*, 1:571–591.

D. N. Kalikow, K. N. Stevens, and L. L. Elliott. 1977. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *Journal of the Acoustical Society of America*, 61:1337–1351.

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

M. Kutas and S. A. Hillyard. 1984. Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307:161–163.

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

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

E. M. Moreno, K. D. Federmeier, and M. Kutas. 2002. Switching languages, switching *palabras* (words): an electrophysiological study of code switching. *Brain and Language*, 80:188–207.

H. Neville, J. L. Nicol, A. Barss, K. I. Forster, and M. F. Garrett. 1991. Syntactically based sentence processing classes: evidence from event-related brain potentials. *Journal of Cognitive Neuroscience*, 3:151–165.

L. Osterhout and P. J. Holcomb. 1992. Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, 31:785–806.

M. Parviz, M. Johnson, B. Johnson, and J. Brock. 2011. Using language models and Latent Semantic Analysis to characterise the N400m neural response. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 38–46. Canberra, Australia.

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

N. J. Smith and R. Levy. in press. The effect of word predictability on reading time is logarithmic. *Cognition*.

A. Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, pages 901–904. Denver, Colorado.

C. Van Petten and B. J. Luka. 2012. Prediction during language comprehension: benefits, costs, and ERP components. *International Journal of Psychophysiology*, 83:176–190.