Prediction during natural language comprehension

Roel M. Willems, Stefan L. Frank, Annabel D. Nijhof, Peter Hagoort, Antal van den Bosch

Supplementary Materials

Supplementary Methods

Five probabilistic language models were trained on increasingly large subsets of the Dutch Corpus of Web (Schaefer and Bildhauer 2012), ranging in size from 1,000 to 10 million sentences. As shown in Supplementary Table S1, the models are identified by a label corresponding to the number of sentences in the training set: 1K, 10K, 100K, 1M, and 10M, respectively. After training, each model estimated one surprisal value, and one entropy value for each word, as explained in the main text. Note that surprisal and entropy were also estimated for punctuation marks because the models treat these as any other word. However, these values were not included in the analysis. A model that is trained on more data will more accurately capture the language statistics, hence, the stepwise increase in training corpus size yields surprisal and entropy values from increasingly accurate language models. If surprisal and entropy are indeed predictive of the Blood Oxygenation Level Dependent (BOLD) signal (Ogawa et al. 1990), we should find that the more accurate estimates explain more variance in the fMRI data.

To test this hypothesis, we looked at the correlation of brain data (BOLD response) and the estimates of entropy and surprisal from models trained on different corpus sizes in brain regions that were found to be related to surprisal / entropy in the whole-brain analysis (results reported in the main text). Because the regions are obtained from the whole-brain analysis, we refrained from doing statistical testing on the data from the different models; they are presented for illustration purposes only.

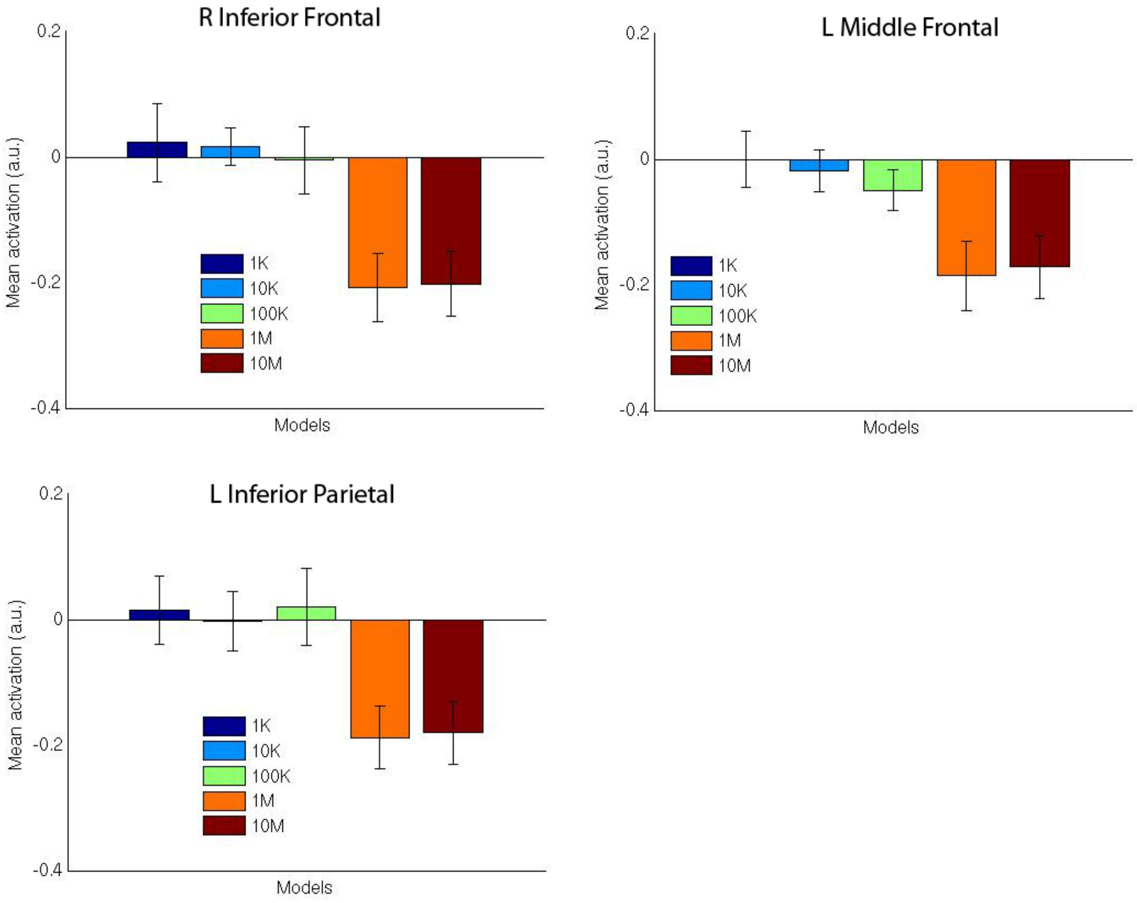
Model	# sentences	# word tokens	# word types
1K	1,000	19,069	5,642
10K	10,000	195,336	30,121
100K	100,000	1,959,298	132,027
1M	1,000,000	19,656,612	527,827
10M	10,000,000	196,521,450	2,164,522

Supplementary Table S1. Word and sentence count in the training corpora. Words, punctuations marks, and numbers are counted as individual tokens. Each smaller training data set formed a subset of all larger sets.

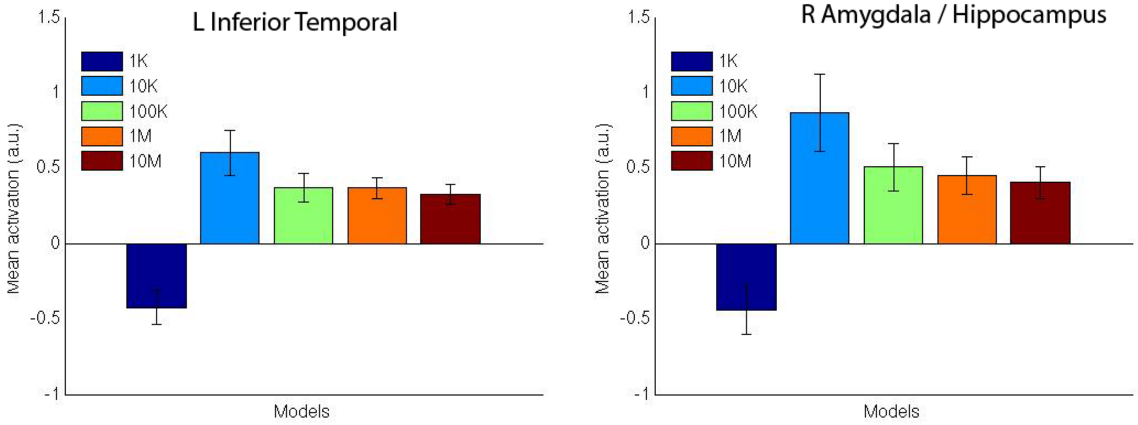
A separate regression model was fit for each of the five trigram models, for entropy and surprisal, as described in the text of the main paper. This was done in order to investigate whether more accurate entropy or surprisal estimates (i.e. based on a larger training corpus) more accurately predicted the experimental results, as has been found before for surprisal and word reading time and N400 amplitude, in studies employing single sentence reading (Monsalve et al. 2012; Frank 2013;Frank et al. 2015). All whole-brain analyses (reported in the main paper) were performed only on estimates from the model trained on the largest corpus (10 million words), given that this is – at least theoretically – the most accurate model.

Supplementary Results

Supplementary Figures S1 and S2 illustrate the effect of model quality (i.e., training corpus size) on the response in areas sensitive to entropy (Fig. S1) and surprisal (Fig. S2) in the whole-brain analysis. Fig. S1 shows that three areas that were activated to entropy in the whole-brain analysis are sensitive to model quality, with effects present for estimates by the 1M and 10M models only. Fig. S2 shows that for surprisal an effect is present for all models except for the 1K model (the model trained on the smallest corpus size), and that the lowest variance (standard error) is present in the 10M model.



Supplementary Fig. S1. Mean contrast values to the entropy regressor in three regions that were activated in the whole-brain analysis (Fig. 1 of the main text). In the whole-brain analysis, statistical analysis was done by taking the entropy estimates from the computational model which was trained on the largest corpus (10 Million words). Here we show for activated regions the responses to entropy estimates from models which were trained on smaller sets of words (Table S1). For each region, results from the five language models are displayed. It can be seen that the fit of the regressor improves with larger corpus size, and stabilizes around 1 Million words. Error bars show standard error of the mean (s.e.m.).



Supplementary Fig. S2. Mean contrast values to the surprisal regressor in two regions that were activated in the whole-brain analysis (Fig. 1 of the main text). In the whole-brain analysis, statistical analysis was done by taking the entropy estimates from the computational model which was trained on the largest corpus (10 Million words). Here we show for activated regions the responses to word surprisal estimates from models which were trained on smaller sets of words (Table S1). For each region, results from the five language models are displayed. It can be seen that all models differ from the 1K model, and that the variance decreases as corpus size increases (smaller standard error). Error bars show standard error of the mean (s.e.m.).

References

Frank SL. 2013. Uncertainty reduction as a measure of cognitive load in sentence comprehension. Top Cogn Sci. 5:475–494.

Frank SL, Otten LJ, Galli G, Vigliocco G. 2015. The ERP response to the amount of information conveyed by words in sentences. Brain Lang. 140:1–11.

Monsalve IF, Frank SL, 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. Presented at the Association for Computational Linguistics. Avignon, France. p. 398–408.

Ogawa S, Lee TM, Kay AR, Tank DW. 1990. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. Proc Natl Acad Sci U S A. 87:9868–9872.

Schaefer R, Bildhauer F. 2012. Building large corpora from the web using a new efficient tool chain. In: Proceedings of the 8th international conference on Language Resources and Evaluation. European Language Resources Association.