The neural dynamics of auditory word recognition and integration

Jon Gauthier and **Roger Levy**
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
jon@gauthiers.net, rplevy@mit.edu

Abstract

Listeners recognize and integrate words in everyday speech by combining expectations about upcoming content with incremental sensory evidence. We present a computational model of word recognition and its downstream neural correlates, and fit this model to explain EEG signals recorded as subjects listened to a fictional story. The model reveals distinct neural processing of words depending on whether or not they can be quickly recognized. While all words trigger a neural response characteristic of probabilistic integration — voltage modulations predicted by a word's surprisal in context — these modulations are amplified for words which require more than roughly 100 ms of input to be recognized. We observe no difference in the latency of these neural responses according to words' recognition times. Our results support a two-part model of speech comprehension, combining an eager and rapid process of word recognition with a temporally independent process of word integration.

Keywords: word recognition; speech comprehension

The N400 ERP is a centro-parietally distributed negative voltage modulation measured at the scalp by electroencephalogram (EEG) which peaks around 400 ms after the onset of a word (Kutas & Hillyard, 1984). Studies of the N400 in naturalistic reading and listening suggest that the amplitude of this response measures the difficulty of integrating a recognized word with a model of the broader linguistic context, and that this amplitude is well predicted by objective estimates of a word's contextual probability (Frank, Otten, Galli, & Vigliocco, 2015; Heilbron, Armeni, Schoffelen, Hagoort, & De Lange, 2022).

This paper investigates the relationship between the N400 response and upstream cognitive mechanisms of word recognition in naturalistic speech comprehension. First, we ask whether there is a regular *temporal* relationship between the process of word recognition and the integration processes observable in neural data. Second, we ask whether this neural response differs according to words' recognition times in ways not captured by mere latency differences.

Model

We first design a cognitive model of the dynamics of word recognition, capturing how a listener forms incremental beliefs about the word they are hearing as a function of the linguistic context C and some partial acoustic evidence $I_{\leq k}$. We formalize the listener's belief in the intended word w_i as a Bayesian posterior (Norris & McQueen, 2008):

$$P(w_i \mid C, I_{\leq k}) \propto P(w_i \mid C)\, P(I_{\leq k} \mid w_i) \tag{1}$$

which factorizes into a prior expectation of the word w_i in context (first term) and a likelihood of the partial evidence of k phonemes $I_{\leq k}$ (second term). We use a neural network language model (GPT Neo 2.7B; Black, Gao, Wang, Leahy, & Biderman, 2021) for the prior. The likelihood is a noisy-channel phoneme recognition model:

$$P(I_{\leq k} \mid w_i) = \prod_{1 \leq j \leq k} P(I_j, w_{ij})^{\frac{1}{\lambda}} \tag{2}$$

where per-phoneme confusion probabilities are drawn from prior phoneme recognition studies (Weber & Smits, 2003) and reweighted by a temperature parameter λ.

We evaluate this posterior for every word with each incremental phoneme, from $k = 0$ (no input) to $k = |w_i|$ (conditioning on all of the word's phonemes). We say a word is *recognized* at a phoneme $0 \leq k_i^* \leq |w_i|$ when this posterior exceeds a confidence threshold parameter γ.

We take a word's recognition time τ_i to be some fraction α of the way through the span of the k_i^*-th phoneme; in the special case where $k_i^* = 0$ and the word is confidently identified prior to acoustic input, we take τ_i to be a fraction α_p of its first phoneme's duration (where α, α_p are free parameters fitted jointly with the rest of the model).

We next define a set of candidate linking models which describe how word recognition times τ_i affect neural responses. These models are variants of a temporal receptive field model (TRF; Crosse, Di Liberto, Bednar, & Lalor, 2016), which predicts multivariate scalp EEG data as a convolved linear response to lagged features of the stimulus. We define two time series: X_t, control features of the auditory stimulus, and X_v, features of words in the stimulus. We assume that X_t causes a neural response independent of recognition times τ_i, while the neural response to features X_v may vary as a function of τ_i.

We consider three distinct TRF models linking the cognitive dynamics of word recognition to neural responses (fig. 1): 1) a unitary response, aligned to τ_i (**shift model**); 2) a variable response by τ_i, aligned to word onset (**variable model**); 3) a unitary response aligned to word onset (**baseline model**). In the variable model, we estimate independent TRFs for words as a function of their recognition times. For a tercile split of words based on recognition time τ_i into "early," "intermediate," and "late" bins, we learn distinct TRF parameters mapping word features X_v in each quantile to the neural response.

We analyze EEG data recorded as 19 subjects listened to one hour of a fictional story, published in Heilbron et al. (2022). We follow the authors' preprocessing methods and control predictors X_t. Our word-level feature vectors $X_v \in \mathbb{R}^{n_w \times 2}$ consist of 1) word surprisal in context, computed using GPT Neo 2.7B (Black et al., 2021), and 2) word log-frequency

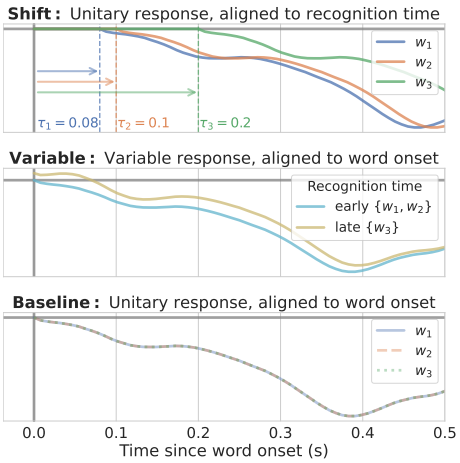

Figure 1: Possible neural models linking word recognition times τ_i to neural modulations by word-level features X_v.

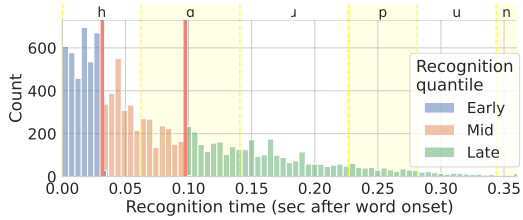

Figure 2: Inferred distribution of word recognition times. Salmon lines mark a tertile split by recognition time; yellow regions mark median phoneme durations. An example word from the data, *harpoon*, is aligned above the graph.

(Brysbaert & New, 2009).

We infer the parameters of the cognitive model jointly with within-subject neural TRF parameters in order to minimize EEG prediction loss, and evaluate models on held-out EEG data.

Results

The baseline model exceeds the performance of an ablated model without word-level features X_v ($t = 4.63, p < 0.001$), and recovers a naturalistic N400 response. The variable model in turn significantly exceeds this baseline ($t = 6.57, p < 10^{-5}$), while the shift model does not ($t = 0.515, p > 0.6$).

We next examine the fitted variable model's estimates of the cognitive dynamics of word recognition and the neural correlates of word integration. The variable model fit predicts (Figure 2) that a lower third of "early" words are recognized prior to 32 ms, and an upper third of "late" words are recognized after 97 ms post word onset. This finding aligns with prior work suggesting that listeners frequently pre-activate features of lexical items far prior to their acoustic onset in the stimulus (Goldstein et al., 2022; Wang, Kuperberg, & Jensen, 2018). Figure 3 shows the variable model's parameters describing a neural response to word surprisal for each of three recognition time quantiles, time locked to word onset. We see that late-recognized words show an exaggerated negative modulation

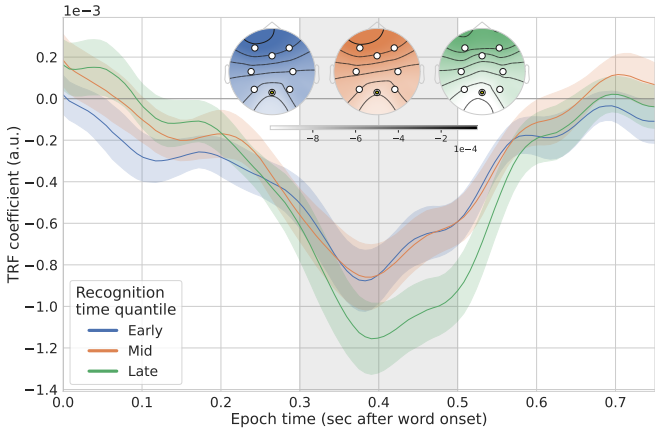

Figure 3: Centro-parietal signal modulations by word surprisal vary by recognition time. Average over subject coefficients; error regions denote s.e.m. ($n = 19$). Inset: spatial distribution of surprisal modulations averaged for each recognition time quantile within vertical gray regions.

due to word surprisal (green line peak minus blue line peak in the shaded region; within-subject paired $t = -5.14, p < 10^{-4}$). However, there is no significant difference in the latency of the N400 peak for words recognized early vs. late (green line peak time minus blue line peak time; within-subject paired $t = 1.391, p > 0.1$).

Discussion

Our analyses reveal two major findings about the link between word recognition and integration: 1) The onset of word integration effects does not vary as a function of word recognition times. 2) Neural integration responses show a different morphology (with exaggerated modulations by surprisal) for words recognized late after their acoustic onset.

These results are consistent with a two-part model of speech comprehension (van den Brink, Brown, & Hagoort, 2006): Listeners continuously update posterior beliefs about the word being heard, loading incremental interpretations into a memory buffer. The recognition time estimate τ_i predicts when this buffer will resolve into a clear lexical inference.

A second integration process reads the contents of this buffer and merges it with representations of the linguistic context. Our latency results show that this process happens independently of a listener's confidence in their lexical interpretations, and is instead time-locked to word onset. This integration process thus exhibits two modes depending on the buffer's contents: one *standard*, in which the buffer is clearly resolved, and one *exceptional*, in which the buffer contents are still ambiguous, and additional inferential or recovery processes must be deployed in order to proceed with integration. Future research should address what drives the regular and independent timing of integration processes (cf. Federmeier & Laszlo, 2009), and further characterize the mode of "exceptional" word integration.

References

Black, S., Gao, L., Wang, P., Leahy, C., & Biderman, S. (2021, March). *GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.5297715 doi: 10.5281/zenodo.5297715

Brysbaert, M., & New, B. (2009, November). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. doi: 10.3758/BRM.41.4.977

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The multivariate temporal response function (mtrf) toolbox: A matlab toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience*, *10*. Retrieved from https://www.frontiersin.org/articles/10.3389/fnhum.2016.00604 doi: 10.3389/fnhum.2016.00604

Federmeier, K. D., & Laszlo, S. (2009). Time for meaning: Electrophysiology provides insights into the dynamics of representation and processing in semantic memory. *Psychology of learning and motivation*, *51*, 1–44.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015, January). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11. doi: 10.1016/j.bandl.2014.10.006

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., ... Hasson, U. (2022, March). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), 369–380. doi: 10.1038/s41593-022-01026-4

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & De Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, *119*(32), e2201968119.

Kutas, M., & Hillyard, S. A. (1984, January). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161–163. doi: 10.1038/307161a0

Norris, D., & McQueen, J. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological review*. doi: 10.1037/0033-295X.115.2.357

van den Brink, D., Brown, C. M., & Hagoort, P. (2006). The cascaded nature of lexical selection and integration in auditory sentence processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(2), 364–372. doi: 10.1037/0278-7393.32.3.364

Wang, L., Kuperberg, G., & Jensen, O. (2018). Specific lexico-semantic predictions are associated with unique spatial and temporal patterns of neural activity. *Elife*, *7*, e39061.

Weber, A., & Smits, R. (2003). Consonant And Vowel Confusion Patterns By American English Listeners. , 4.