UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Comprehenders Rationally Adapt Semantic Predictions to the Statistics of the Local Environment: a Bayesian Model of Trial-by-Trial N400 Amplitudes

Permalink

https://escholarship.org/uc/item/7wj6w9tm

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 39(0)

Authors

Delaney-Busch, Nathaniel
Morgan, Emily
Lau, Ellen
et al.

Publication Date

2017

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Comprehenders Rationally Adapt Semantic Predictions to the Statistics of the Local Environment: a Bayesian Model of Trial-by-Trial N400 Amplitudes

Nathaniel Delaney-Busch (ndelan02@tufts.edu)[1], **Emily Morgan** (emorga08@tufts.edu)[1], **Ellen Lau** (ellenlau@umd.edu)[2], **Gina Kuperberg** (kuperber@nmr.mgh.harvard.edu)[1,3]

[1] Department of Psychology, Tufts University

[2] Department of Linguistics, University of Maryland

[3] Department of Psychiatry and the Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital

Abstract

When semantic information is activated by a context prior to new bottom-up input (i.e. when a word is predicted), semantic processing of that incoming word is typically facilitated, attenuating the amplitude of the N400 event related potential (ERP) – a direct neural measure of semantic processing. This N400 modulation is observed even when the context is just a single semantically related "prime" word. This so-called "N400 semantic priming effect" is sensitive to the probability of seeing a related prime-target pair within experimental blocks, suggesting that participants may be adapting the strength of their predictions to the predictive validity of their broader experimental environment. We formalize this adaptation using an optimal Bayesian learner, and link this model to N400 amplitudes using an information-theoretic measure, surprisal. We found that this model could account for the N400 amplitudes evoked by words (whether related or unrelated) as adaptation unfolds across individual trials. These findings suggest that comprehenders may rationally adapt their semantic predictions to the statistical structure of their broader environment, with implications for the functional significance of the N400 component and the predictive nature of language processing.

Keywords: language; prediction; rational adaptation; semantic priming; EEG/ERP; word processing; information theory; Bayesian modeling; surprisal

Introduction

How a word is processed fundamentally depends on the context. Predictable words are processed more quickly than unpredictable words (Fischler & Bloom, 1979), with shorter fixations (and more frequent skips) during reading (see Staub, 2015 for review). A similar facilitation pattern is found on the N400 component (Kutas & Hillyard, 1984), an event-related potential (ERP) that reflects semantic processing (Kutas & Federmeier, 2011). The degree to which any particular word is facilitated is proportional to the probability of encountering that word given the context (DeLong, Urbach, & Kutas, 2005; Smith & Levy, 2013).

Semantic facilitation is observed even when the preceding context is only a single word. For example, the processing of a "target" word is facilitated when it is preceded by a semantically related (versus an unrelated) "prime" word: the so-called "semantic priming effect" (Neely, 1991). Semantic priming is also apparent on the N400 component, with more predictable words eliciting a smaller (less negative) N400 amplitude than less predictable words (Bentin, McCarthy, & Wood, 1985).

Importantly, the strength of the behavioral semantic priming effect is sensitive not only to the degree to which the prime and target are semantically related, but to the probability of receiving a related trial in the first place (Brown, Hagoort, & Chwilla, 2000; Grossi, 2006). This has also been found with ERPs: experimental blocks with a higher proportion of related trials elicit a larger N400 semantic priming effect (Lau, Gramfort, Hamalainen, & Kuperberg, 2013; Lau, Holcomb, & Kuperberg, 2013). The semantic priming effect is likely sensitive to predictive validity because participants implicitly track and adapt to changes in the statistical contingencies over time.

Here we utilized data from Lau and colleagues (Lau, Holcomb, et al., 2013) to build and test a quantitative hypothesis of what drives this effect of predictive validity on the N400 semantic priming effect. Specifically, we asked whether the larger N400 priming effect in the high predictive-validity block could have been achieved through a "rational" probabilistic model of trial by trial adaptation. Although there are an infinite number of ways to build such a model, there are some particular theoretical constraints we can start from. Current evidence suggests that (a) prediction in language processing is probabilistic in nature, (b) predictions incrementally adapt to new information (where adaptation should be rapid when the environment changes), and (c) the brain calculates something like prediction error.

Probabilistic Prediction in Language Processing and Semantic Priming

The role of prediction in language has long been debated, with differing definitions of what a "prediction" actually entails (see Kuperberg & Jaeger, 2016 for discussion). Here, we view *probabilistic* prediction as a central feature of language comprehension (DeLong et al., 2005; Federmeier, 2007; Smith & Levy, 2013), which does not necessarily need to be strategic or even conscious in nature. For any given context, there exists a probability distribution over the words that could be encountered next. A "prediction" is simply the presence of this probability distribution. While there is evidence for probabilistic prediction at multiple different levels of representation (Kuperberg & Jaeger, 2016), here we focus only on prediction at the lexical level.

Note that, defined in this way, prediction exists even in the absence of a local context. Consider an experiment where words are being serially presented to participants at random. The "prediction" that participants make in such an experiment could be expressed as a probability distribution over words given an *average* context. This is functionally identical to word frequency, where high frequency words are more probable given a random/average context and low frequency words are less probable given a random/average context (Norris, 2006).

Given these assumptions about the nature of linguistic predictions, we can view the semantic priming effect as a type of probabilistic prediction. If a participant knows that a prime informs the target, they will implicitly generate a probability distribution over possible targets given that prime. Target processing is facilitated proportional to its probability. Though these prime-target transition probabilities are not easy to estimate from corpus studies (people don't often write in prime-target pairs), it can be estimated using production tasks like word association.

A Rational Model of Adaptation

Probabilistic prediction is only beneficial if it actually approximates the statistical structure of the environment. Bad predictions aren't helping anybody. A number of recent language studies suggest that people rapidly *adapt* local models based on changes in their environment (Kleinschmidt & Jaeger, 2015). For example, an environment with a high proportion of typically dis-preferred syntactic parses can attenuate or even reverse the so-called "garden-path" effect in ambiguous sentences (Fine, Jaeger, Farmer, & Qian, 2013). At the phonemic level, participants change their perception of ambiguous phonemes if one of the two competing options was locally repeated (Kleinschmidt & Jaeger, 2016). And across longer periods, people show signs of adaptation to foreign accents that can generalize between different speakers with that accent (Bradlow & Bent, 2008).

We can view the predictive validity manipulation in Lau et al (2013) through a similar lens. In experimental contexts with a higher proportion of semantically related word-pairs, participants will rely relatively more on the prime (versus relying more on a random/average context) to inform their predictions of the target. This means that the probability mass assigned to any *particular* target word depends not only on the associative strength of its prime, but also the likelihood that the prime provides information about the target in the first place (i.e. the predictive validity effect). When the proportion of semantically related and unrelated trials within a block changes, people adapt.

Although there are many ways to quantify adaptive learning, one attractive theoretically motivated implementation is Bayesian updating. This assumes that adaption is "rational" in nature (Anderson, 1990). Here, initial belief about the probability of obtaining a particular type of trial (i.e. a related versus unrelated prime-target pair) is denoted by the prior probability p(h). Upon receiving a trial, this prior belief is updated using Bayes Law. This "posterior" is then used as the prior belief for the next trial.

Applying this rational, Bayesian framework to the predictive validity effect (on semantic priming) has a number of advantages. It would allow for prior beliefs from an initial lower predictive validity block to influence expectations for a subsequent higher predictive validity block in a principled way (i.e. the prior). It would adapt incrementally across trials. It would adapt more quickly near the change point, when evidence is low. And beliefs would slowly asymptote to the known true probability of receiving a related trial.

In the present investigation, we will use this type of Bayesian framework as the starting point for explaining how the brain adapts to changes in the statistical contingencies of incoming language input. We will refer to this as a Rational Adapter model.

The N400 Measures Information Content

So our model should be probabilistic, and it should adapt, but a third component of this model is required before it can be tested: a linking function to actual brain activity. It is not necessarily the case that N400 amplitudes — the brain activity we are attempting to model — need be linearly dependent on the probability of a word given a context. Here, we argue that the N400 is best thought of not as a measure of probability, but as a measure of *information* (see Rabovsky & McRae, 2014 for discussion).

In information theory (Shannon & Weaver, 1949), the amount of information conveyed by an event is "whatever was not known ahead of time". An input that was perfectly predicted does not convey any new information. In contrast, messages that are not very predictable convey a lot of information. The amount of information (in units of "bits") that was not predicted ahead of time is called "surprisal", quantified as $-\log_2[p(\text{word}|\text{context})]$. One bit is the amount of information provided by flipping a fair coin once. A halving of the probability (e.g. conveying a sequence of two coin flips instead of one) corresponds with a 1 bit increase in surprisal.

This simple transformation has proved tremendously powerful in explaining language processing data (Hale, 2001; Levy, 2008). Smith and Levy (2013) provide empirical evidence that reading times relate to word probability logarithmically (e.g. as with the surprisal transformation) across six orders of magnitude. More recently, Frank and colleagues (Frank, Otten, Galli, & Vigliocco, 2015) discuss ERP evidence that the N400 component is sensitive to word surprisal. Given this evidence, for the present investigation, information-theoretic surprisal will be our linking function between the Rational Adapter model and N400 amplitude, rather than probability.

In the present investigation, we aim to build and test a Rational Adapter model of semantic priming against neural N400 component data. Specifically, we use data from Lau and colleagues (2013). Here, the N400 semantic priming effect was measured first in block 1 where 10% of the trials were related, and then in block 2 where 50% of the trials were related. We use the block 1 only to inform the prior, which is then fed into the Rational Adapter model to predict N400 amplitudes in block 2, as participants' beliefs about the predictive validity of the prime adapt.

We hypothesize that this Rational Adapter surprisal model will better account for the N400 data than a non-learning model of N400 data. Specifically, we hypothesize that the size of the N400 effect (the difference in amplitude of the N400 evoked by related and unrelated target words) will increase rapidly near the beginning of block 2, as the Rational Adapter shifts from a "10% related" prior towards a "50% related" asymptote.

Methods, Modeling and Results

ERP Data collection

We used data from Lau and colleagues (Lau, Holcomb, et al., 2013). Briefly, 32 right-handed participants (13 men) between age 19-24 were shown sequential prime-target pairs as event-related potentials (ERPs) were recorded and time-locked to the onset of the target. Participants were asked to perform a semantic monitoring task that was not directly related to the experimental manipulation. All participants saw an initial 400 trials (block 1) where 10% of the stimuli were related (e.g. "ladder… climb"), followed by 400 trials where 50% of the stimuli were related (block 2). The blocks were separated by a short break, but participants were not explicitly told about any changes in the experiment.

80 of the trials in block 2 (40 related, 40 unrelated) were critical prime-target pairs that were matched and counterbalanced across participants, alongside 320 fillers. Primes were presented with an SOA of 600ms and targets had a duration of 900ms. The N400 component was averaged across a time window of 300-500ms over the average of three centro-parietal channels. Extreme outliers in N400 measures were removed (4 standard deviations or more from the mean).

Visualization Lau et al originally reported that the N400 semantic priming effect was larger in block 2 than block 1. For the present investigation, we plotted how the N400 amplitudes of these critical trials changed over the course of block 2, as shown in Figure 1. This was estimated using a loess local regression over N400 amplitudes for related and unrelated words across the ordinal position of critical items in the experiment (this local regression was necessary because not every participant saw critical targets in the same place). As can be seen here, N400 amplitudes for related and unrelated words are initially similar, but then diverge as participants are exposed to more and more of the block.

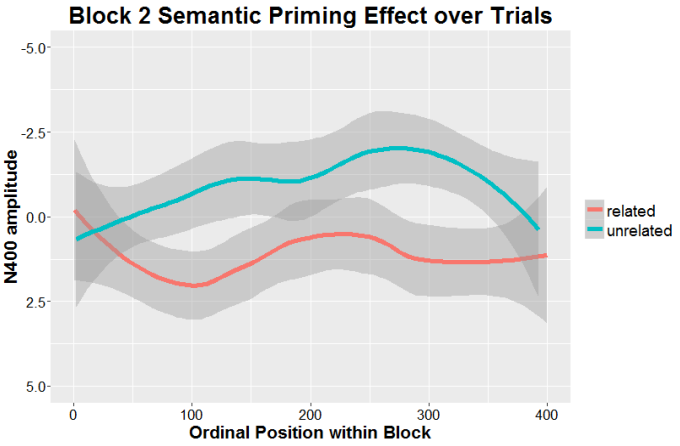

Figure 1: Block 2 N400 amplitudes over trials.

The Rational Adapter Word Surprisal Model

Our Rational Adapter model consists of three primary components: a) a Bayesian belief about the probability of receiving a related vs. an unrelated prime-target pair at any given point, b) a mixture of p(word|prime) and p(word|average context) given these beliefs about the trial types, and c) a conversion from these probabilistic predictions to word surprisal (as a linking function to the N400 component). The whole model takes the form:

$$\text{Word surprisal} = -\log_2[\lambda * p(\text{word}|\text{prime}) + (1-\lambda) * p(\text{word}|\text{average context})]$$

where λ is a point estimate of the probability with which a rational adapter expects a related trial at that point in time.

We use a beta-binomial model to estimate a participant's belief about the probability of seeing related versus unrelated trials. To set a prior on the beta distribution, we assume that participants enter block 2 assuming the proportion will be the same as block 1: a 10% chance of receiving a related trial. This prior is expressed using a 1:9 ratio of related:unrelated pseudocounts. Though participants see 400 trials in block 1, participants will likely discount their previous experience somewhat (reflecting some uncertainty). As a best-guess approximation, we set the prior going into block 2 at Beta(5, 45), i.e. pseudocounts equivalent to having seen 5 related and 45 unrelated trials. In other words, participants were assumed to put more weight on their experience with the new block (vs their prior given the old block) after about 50 trials.

After each new trial, this Beta distribution is updated by adding the observed trial counts to the prior pseudocounts. For example, after 5 related and 5 unrelated trials in block 2, a participant's beliefs would be modeled as Beta(10, 50). We took the mean of this beta distribution just before each critical trial to reflect the point probability λ with which that participant expects a related trial for that event.

This probability, λ, then provides a weighting term for a mixture model between the two ways that participants might generate more specific predictions about the upcoming target word at any given trial. Given a related trial, we

model these within-trial predictions as p(word|prime), estimated using "forward association strength" (FAS) from the Florida Word Association Norms (Nelson, McEvoy, & Schreiber, 2004), and then we weight this probability by λ. Given an unrelated trial, we model these within-trial predictions as p(word|average context), as estimated by word frequency from the SUBTLEX corpus (Brysbaert & New, 2009), and then we weight this probability by 1-λ.

The mixture of these two terms yields a "word probability", given the prime and beliefs about whether or not it will inform the target at any point in the experiment.

Finally, this word probability is transformed into "word surprisal", or the amount of information that was not predicted ahead of time (in bits), given by $-\log_2$[p(word probability)].

Word Surprisal and N400 Amplitudes

To numerically test whether our estimate of surprisal explains variance in the N400 amplitudes evoked by each target word, we conducted a linear mixed-effects regression using the lme4 package in R, with word surprisal as a predictor and centro-parietal N400 amplitude for each trial in block 2 as an outcome. Word surprisal was standardized. The maximal random effects structure across (crossed) subjects and items was used (Barr, Levy, Scheepers, & Tily, 2013).

Results We found that word surprisal significantly accounted for variance in N400 amplitudes ($\beta = -1.14$, $t = -5.24$, $p < 0.001$). As word surprisal increased, N400 amplitudes tended to be more negative (i.e. larger).

Word Surprisal Explains Trial-by-Trial Variance

There is an important caveat to this "rational adapter" word surprisal effect, however: by definition, unrelated words tend to have high surprisal, while related words tend to have low surprisal. As such, the word surprisal effect in block 2 could potentially be attributable to the categorical "Relatedness" effect already reported in the initial study.

To address this possibility, we ran a second linear mixed-effects regression that included *both* categorical Relatedness and word surprisal as predictors. This would show whether our rational adapter estimate of word surprisal could account for variance in N400 amplitudes above and beyond what could already be explained by the main effect of related vs unrelated trials. Again, the maximal random effects structure for word surprisal was used.

Results We found that word surprisal significantly accounted for variance in N400 amplitudes ($\beta = -2.21$, $t = -2.76$, $p = 0.006$) above and beyond the main effect of Relatedness. This indicates that the surprisal difference between related and unrelated words was not sufficient to account for the way that word surprisal related to the N400 in the first model.

We caution that given the multicollinearity between word surprisal and the relatedness effect (the primary motivation for running this test in the first place), this β estimate is likely inflated. We limit our conclusions to the explanatory power here, not the regression coefficient.

The "Rational Adapter" Word Surprisal Model Outperforms its Constituent Elements Alone

Another potential concern is that the model we used to estimate word surprisal simply includes more information about the trials. Namely, the word frequency and FAS of each trial are inputs to the Rational Adapter model calculations. These could have explained items-level variance in N400 amplitudes without resorting to adaptation, given that the N400 component is already known to be sensitive to both these factors. In short, perhaps the explanatory power of our rational adapter model is primarily due to the inclusion of trial-specific frequency and FAS information, rather than prediction and adaptation.

To address this possibility, we ran a third linear mixed-effects regression that includes not only word surprisal as a predictor, but also Frequency, FAS, and Relatedness predictors for each trial. This tests whether the particular *arrangement* of inputs into the "rational adapter" word surprisal model explains variance in N400 amplitudes marginal to the stationary main effect of Relatedness and to its constituent items-level elements. Again, the maximal random effects structure for word surprisal was used (across both items and subjects), and all continuous predictors were standardized.

Results We found that word surprisal significantly accounted for variance in N400 amplitudes ($\beta = -2.30$, $t = -2.11$, $p = 0.036$) above and beyond Frequency, FAS, and Relatedness. This indicates that the particular way items-level features were combined into the our model is an important source of explanatory power, and that the increased fit is not simply due to the fact that our model included additional information about items-level features.

Finding the Optimal Prior

Our model assumed that the rational adapter should approach $\lambda = 0.1$ (the actual block 1 proportion) as they go through the first 400 trials of block 1, regardless of what their expectations were coming into the experiment. Given that our model is explicitly a rational one, we kept constant this 1:9 related:unrelated ratio for the prior for block 2. However, that still leaves the prior strength (i.e. number of pseudocounts) as an assumption that can be explored. For hypothesis testing above, we assumed that participants entered block 2 with a Beta(4, 45) prior, i.e. that participants believed it would have the same 10% relatedness proportion as block 1 with a weight of 50 pseudotrials. This 50 pseudocount prior weighting, however, was essentially guesswork (we didn't want to bias our hypothesis tests by interrogating many models and selecting the best one). Here, we sought to ensure that our results were not idiosyncratically dependent on having made a "lucky" guess.

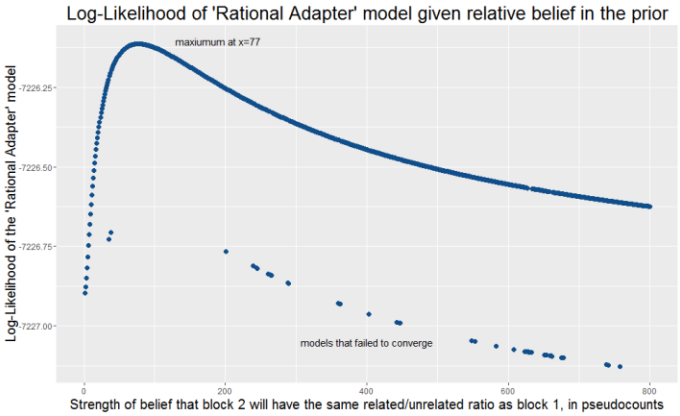

Figure 2: Deriving the optimal prior strength.

A low pseudocount prior like (1, 9) would cause rational adaptation to proceed very quickly, while a high prior pseudocount like (40, 360) would cause rational adaptation to proceed much more slowly, as participants enter block 2 with much more certainty about the environmental statistics. The pseudocount (and thus speed of adaptation) that best explains variance in N400 amplitudes is thus an empirical question: what is the optimal prior strength?

To find the optimal prior, we calculated the word surprisal (for all trials for all participants, as above) for every integer "prior strength" from 1 to 800 pseudocounts. We then ran a separate linear mixed-effects regression model for each prior strength with word surprisal and Relatedness as predictors and a maximal random effects structure. After fitting these 800 regression models, we extracted the log-likelihood of each.

Results These data are shown in Figure 2. The single maximum log-likelihood was obtained with a Beta(7.7, 69.3) prior, a "prior strength" of 77 pseudocounts. However, all pseudocounts between 70 and 85 yielded similar model fits, and performance degrades smoothly on either side.

This indicates that, on average, participants began giving more weight to the new block's data than the previous block's data 70-85 trials into block 2. A rational adapter with a very weak prior (below ~50 pseudocounts) does not account for N400 data well because it adapts too quickly. Similarly, a rational adapter entering block 2 with a very strong prior (above ~200 pseudocounts) also does not account for the N400 data well because it adapts too slowly.

We note that some models had poor fit because they did not converge. None were within the 70-85 range capturing the maximum.

Discussion

Previously, Lau and colleagues (2013) found that the N400 semantic priming effect shows evidence that adaptation occurred when the predictive validity of the local context changed. In the present investigation, we explored the nature of the adaptation process as it unfolded. Figure 1 shows the trial-by-trial nature of this adaptation over ordinal position in block 2. Our Rational Adapter model provides a theoretically-grounded quantitative account of how that adaptation may have occurred on an incremental trial-by-trial basis. It was built using three foundational considerations: that contexts can probabilistically inform lexico-semantic expectations for upcoming stimuli, that these expectations adapt rationally (in an optimal Bayesian manner), and that the N400 component is sensitive to units of information rather than units of probability (after the present analyses, we tested this assumption and found that word *surprisal* significantly accounted for variance in N400 amplitudes [t = -2.57, p = 0.011] above and beyond word *probability* and the categorical Relatedness effect.).

In a re-analysis of the original study, we provide empirical evidence that this model is consistent with how brain activity evoked by target words changed over the course of block 2. We showed that it accounted for variance in N400 amplitudes above and beyond the stationary effect of related versus unrelated trials, suggesting that it was capturing trial-by-trial differences within block 2. Further, we showed that this particular formulation of the rational adapter model accounted for significant variance in N400 amplitudes above and beyond even its own constituent elements, suggesting that the additional explanatory power was not simply due to the inclusion of items-level information (our single trial approach to ERP analysis). These findings extend previous work on rational adaptation to demonstrate that it can account for changes in predictions during lexico-semantic processing.

In addition, we used the rational adapter model to derive the rate of adaptation that best accounted for the ERP data. Even though participants saw 400 trials in Block 1, we estimate that participants adapted as if they had only seen 70-85 trials of bock 1 by the time they entered block 2. Although participants were not informed of the changing environmental statistics (and the manipulation was not overtly task-relevant), we speculate that the conspicuous block boundary may have prompted participants to adapt at a faster rate. Additionally, there may be a decay or filtering that occurs for distant exposures, which dynamical models of prediction and adaptation may be able to account for.

While the present study included data from a semantic priming paradigm, we suggest that a similar pattern may hold in comparable experiments with more expansive contexts, like sentences or discourses, as the theoretical underpinnings are functionally the same. For example, in experimental contexts with a high proportion of highly constraining sentences, we might expect participants to learn to predict more strongly. Finally, these data have implications for the functional significance of the N400 component. The N400 is often discussed as being sensitive to probabilities. We suggest that its sensitivity to probabilistic measures like cloze probability, forward association, and even frequency may be best conceptualized it as reflecting units of information rather than probability alone (see also Frank et al., 2015; Rabovsky & McRae, 2014; Smith & Levy, 2013).

Acknowledgments

This project was funded by NIMH-R01-MH071635 to GRK and the Sidney J. Baer Trust to GRK. Thank you to Florian Jaeger, Heather Urry, Eddie Wlotko, and Meredith Brown for input.

References

Anderson, J. (1990). *The Adaptive Character of Thought (Studies in Cognition)*: Psychology Press.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J Mem Lang, 68*(3).

Bentin, S., McCarthy, G., & Wood, C. C. (1985). Event-related potentials, lexical decision and semantic priming. *Electroencephalography and clinical neurophysiology, 60*(4), 343-355.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition, 106*(2), 707-729.

Brown, C. M., Hagoort, P., & Chwilla, D. J. (2000). An event-related brain potential analysis of visual word priming effects. *Brain and language, 72*(2), 158-190.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera 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.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature neuroscience, 8*(8), 1117-1121.

Federmeier, K. D. (2007). Thinking ahead: the role and roots of prediction in language comprehension. *Psychophysiology, 44*(4), 491-505.

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid Expectation Adaptation during Syntactic Comprehension. *PLoS ONE, 8*(10).

Fischler, I., & Bloom, P. A. (1979). Automatic and Attentional Processes in the Effects of Sentence Contexts on Word Recognition. *Journal of verbal learning and verbal behavior, 18*(1), 1-20.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and language, 140*, 1-11.

Grossi, G. (2006). Relatedness proportion effects on masked associative priming: an ERP study. *Psychophysiology, 43*(1), 21-30.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. *2nd Meeting of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, 159-166.

Kleinschmidt, D. F., & Jaeger, F. T. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological review, 122*(2), 148.

Kleinschmidt, D. F., & Jaeger, F. T. (2016). Re-examining selective adaptation: Fatiguing feature detectors, or distributional learning? *Psychonomic bulletin & review, 23*(3), 678-691.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Lang Cogn Neurosci, 31*(1), 32-59.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology, 62*, 621-647.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature, 307*(5947), 161-163.

Lau, E. F., Gramfort, A., Hamalainen, M. S., & Kuperberg, G. R. (2013). Automatic semantic facilitation in anterior temporal cortex revealed through multimodal neuroimaging. *J Neurosci, 33*(43), 17174-17181.

Lau, E. F., Holcomb, P. J., & Kuperberg, G. R. (2013). Dissociating N400 effects of prediction from association in single-word contexts. *J Cogn Neurosci, 25*(3), 484-502.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*(3), 1126-1177.

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. *Basic processes in reading: Visual word recognition, 11*, 264-336.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior research methods, instruments, & computers : a journal of the Psychonomic Society, Inc, 36*(3), 402-407.

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological review, 113*(2), 327-357.

Rabovsky, M., & McRae, K. (2014). Simulating the N400 ERP component as semantic network error: Insights from a feature-based connectionist attractor model of word meaning. *Cognition, 132*(1), 68-89.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana: University of Illinois Press.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition, 128*(3), 302-319.

Staub, A. (2015). The Effect of Lexical Predictability on Eye Movements in Reading: Critical Review and Theoretical Interpretation. *Language and linguistics compass, 9*(8), 311-327.