Modeling the N400 brain potential as Semantic Bayesian Surprise

Lea Musiolek (leamusiolek@gmail.com), **Felix Blankenburg (felix.blankenburg@fu-berlin.de)**,
Dirk Ostwald* (dirk.ostwald@fu-berlin.de), **Milena Rabovsky* (milena.rabovsky@gmail.com)**
Freie Universität Berlin, Habelschwerdter Allee 45, 14195 Berlin, Germany (* equal contribution)

Abstract

In research on human language comprehension, the N400 component of the event-related brain potential (ERP) has attracted attention as an electrophysiological indicator of meaning processing in the brain. However, despite much research, the specific functional basis of the N400 remains widely debated. Recent neural network modeling work suggests that N400 amplitudes can be simulated as the stimulus-induced change in internally represented probabilities of aspects of meaning (Rabovsky, Hansen, & McClelland, 2018). Here, we assess this idea based on single-trial N400 amplitudes measured in an oddball-like roving paradigm with written words from different semantic categories varying in semantic feature overlap. We model the N400 as Semantic Surprise, the change in the probability distribution of a stimulus's semantic features for each trial. Simple condition-based analyses produced a significant effect of category switch on N400 amplitude, and the trial-by-trial modeling similarly revealed negative effects of Semantic Surprise on N400 amplitude. From fitting a forgetting parameter for each participant, we also gleaned insights into the rates of forgetting of past input to the semantic system. Thus, we provide a computationally explicit account of N400 amplitudes, which links the N400 and thus the neurocognitive processes involved in human language comprehension to the Bayesian brain hypothesis.

Keywords: N400; ERP; Bayesian surprise; semantic features; prediction

Introduction

Since its discovery in 1980, the N400 has received much attention due to its promise to uncover the brain basis of meaning processing. The first studies showed that verbal stimuli that were semantically incongruous or less expected in the preceding context reliably produced increased centro-parietal ERP negativities around 400ms after stimulus onset, which were insensitive to grammatical or visual violations of expectation. Subsequent experiments found that N400s are modulated not only by sentence context but also by a large variety of other lexical and semantic variables including the lexical frequency of single words, word repetition, and the semantic relatedness between word pairs, to name just a few examples. Overall, more than a thousand empirical studies have used the N400 as a dependent variable, but despite these large amounts of data, the specific functional basis of N400s is still unclear, as reviewed by Kutas and Federmeier (2011). To address this issue and systematically investigate the functional basis of the N400, in recent years there has been a growing interest in linking the N400 to explicit computational models. Most relevant for the current purpose, Rabovsky and McRae (2014) simulated typical word level N400 effects using a neural network model of word meaning and found that the semantic feature layer's error was consistently affected by a variety of experimental manipulations in the same way that N400 is. Because the network error in neural network models is often conceptualized as an implicit prediction error, these simulations were taken to suggest that N400 amplitudes reflect an implicit semantic prediction error or Bayesian surprise at the level of meaning (Rabovsky & McRae, 2014). Rabovsky et al. (2018) extended this approach to sentence meaning using a neural network model of sentence comprehension, the Sentence Gestalt model (St. John & McClelland, 1990). They found that the change each incoming word produced in the activation state of the model's hidden Sentence Gestalt layer, corresponding to the model's implicit prediction of all the semantic features involved in the event described by the sentence, patterned with the N400 in 16 distinct experimental paradigms. This activation change can be formally related to a change in probability distributions produced by a new piece of sequential input, i.e. the concept of Bayesian Surprise (Itti & Baldi, 2009), for different features/aspects of the meaning representation (Delaney-Busch, Morgan, Lau, & Kuperberg, 2017). We describe these distributions in more detail in the Methods section. In the current work, we explicitly model single trial N400 amplitudes as the sum of the Bayesian Surprise produced by different semantic features of German words in an oddball-like roving paradigm with words from different semantic categories (e.g., birds, land animals, kitchen utensils, etc.). We refer to this measure as "Semantic Surprise". The more semantic features a target stimulus shares with the preceding context, the smaller the Semantic Surprise should be. Modeling the N400 as Bayesian surprise at the level of meaning sets it in relation to other earlier ERP effects in oddball paradigms in other domains such as perceptual (i.e. auditory, visual, and tactile) mismatch negativities, which have featured prominently as indicators of Bayesian surprise in Bayesian accounts of brain function and predictive coding theories (Garrido, Kilner, Stephan, & Friston, 2009; Ostwald et al., 2012). From this perspective, the same fundamental mechanisms of brain function apply to processes across levels of representation and domain, in line with the Bayesian brain hypothesis.

Methods

Paradigm

We employed the "roving" paradigm developed by Baldeweg, Klugman, Gruzelier, and Hirsch (2004) and later used by

Ostwald et al. (2012) for modeling somatosensory mismatch negativities as Bayesian surprise. In this oddball-like stimulation protocol, rather than occasionally interrupting a train of "standard" stimuli with a single "deviant" stimulus, two categories of stimuli can each take on the role of standard and deviant simply by switching categories after every 4-8 trials. We modified this protocol to accommodate ten different stimulus categories. By presenting 100 different stimulus words (German nouns) from ten semantic categories in an ongoing sequence (3000 trials) made up of short sequences from each category, it became possible to model trial-by-trial amplitudes, but also to perform more simple condition-based analyses. Our categories were the following: tree species, vegetable species, land animals, birds, geographical formations, pieces of furniture, means of transport, tools, kitchen utensils, and items of clothing.

Semantic features

As features, we decided to use the hypernyms (umbrella terms) stored in the GermaNet lexical-semantic net (Hamp & Feldweg, 1997) for each of our stimulus words. Because of the hierarchical structure of the word net, this ensured varying degrees of feature overlap between words depending on their semantic similarity. Features were excluded if they occurred for only one stimulus word or if they had an absolute type frequency below 30 in the dlexDB corpus (Heister et al., 2011) and could therefore be assumed to be little-known. A word-feature table was then created with the words as rows and the hypernyms/features as columns and filled with values of 0 or 1 depending on whether a hypernym belonged to a word or not, as a basis for later trial-by-trial Bayesian Updating.

Participants and Procedure

40 right-handed German native speakers (8 of them men) between the ages of 19 and 34 participated. In order to give participants a task that would interfere as little as possible with their semantic representations but ensure they would actually process the stimuli, 200 non-words were interspersed between the stimulus words, and participants were instructed to push a certain key whenever they read a non-word. Inter-stimulus intervals were jittered around 800ms.

Analyses

Our dependent variable was the mean amplitude 300 to 500ms after stimulus onset, averaged across the electrode channels in an anterior region of interest (including the five middle electrodes of the F, FC and C rows, respectively).

Condition-based analysis Semantic categories are naturally characterized by high within-category and low between-category measures of overlap on semantic features. The last word in a sequence of words from the same category (a standard) is therefore expected to produce a significantly weaker N400 than a word immediately following a sequence of words from a different category (a deviant). Our conditions of interest were the standard (the last stimulus in each sequence of words from the same category, 475 trials per participant before artefact rejection), and the deviant (the first stimulus in a new sequence of words from the same category, also 475 trials per participant). We averaged N400 ROI mean amplitudes by participant and condition in order to test for differences between the conditions via a paired-samples t test.

Trial-by-trial Bayesian modeling On a trial-by-trial basis, we expect N400 amplitude to be influenced by the respective trial's Semantic Surprise, based on the current and preceding words' semantic features. Our Semantic Surprise measure is essentially the sum of the Bayesian Surprise elicited by all semantic features. For each semantic feature, we implemented a Bayesian sequential updating scheme which uses past occurrences and non-occurrences of the respective feature to compute a beta probability distribution for that feature's occurrence probability $\mu \in [0,1]$ on the next trial. The occurrence or non-occurrence of each semantic feature $i = 1,...,k$ at a given trial is modeled as the outcome of an independent Bernoulli process based on the parameter μ_i. On each trial $t = 1,...,u$, the stimulus word carries a subset of these k semantic features, corresponding to a trial-feature matrix $Y \in B^{u \times k}$ with $B \in \{0,1\}$, i. e. containing zeros and ones to mark the presence or absence of the different features. To model the greater importance of more recent semantic input compared to input further in the past, we conceive the system underlying the N400 as one that down-weights past trials with an exponential forgetting mechanism determined by a parameter $\tau \geq 0$ (Ostwald et al., 2012). Intuitively, the lower τ, the steeper the down-weighting and thus the higher the rate of forgetting past input. The α and β parameters of a beta probability distribution can be conceived as counters of past successes and failures in a Bernoulli process (occurrences and non-occurrences of features). To reflect our assumption of a uniform prior, our initial value for α and β before the first trial is 1. Thus, at a given trial t and for a given semantic feature i, α_{t_i} equals the sum of the vector of the feature's occurrences and β_{t_i} equals the sum of the vector of the feature's non-occurrences, each supplemented by an initial 1. Our forgetting mechanism can be implemented by multiplying each vector element-wise with a weighting vector d before computing the sum. The weighting vector d is obtained by computing the down-weighting function of all integers from 1 through $u+1$, with the down-weighting function being

$$f(x) = \exp\left(-\frac{1}{\tau}(u+1-x)\right) \tag{1}$$

This ensures that the highest weight is always 1. At each trial t, the past feature occurrences and initial 1 are multiplied element-wise with the last $t+1$ elements of d, such that the current trial always has a weight of 1:

$$\alpha_{t_i} = d_{u+1-t} + \sum_{s=1}^{t} y_{s_i} \cdot d_{u+1-t+s} \tag{2}$$

$$\beta_{t_i} = d_{u+1-t} + \sum_{s=1}^{t} (1 - y_{s_i}) \cdot d_{u+1-t+s} \tag{3}$$

The change in the beta distribution for feature i from trial $t-1$ to trial t, or rather the inefficiency of assuming that the distribution is $p_{i_{t-1}}$ (prior distribution) when it is really p_{i_t} (posterior distribution) may be computed using the Kullback-Leibler divergence (Kullback & Leibler, 1951; Itti & Baldi, 2009):

$$\text{BS}_{i_t} = KL(p_t(\mu_i|y_{i_t})||p_{t-1}(\mu_i|y_{i_t-1})) \quad (4)$$

We defined Semantic Surprise as the sum of this divergence across features at each trial. The Semantic Surprise for all trials of each participant was then re-scaled by its own range and used as a regressor for N400 amplitude in a simple linear regression model. This was done for each participant individually to allow for variability between participants. As a consequence, the parameters to be fitted to each participant's data were τ as well as the intercept, slope and error variance parameters of the linear regression. For a given value of τ, the three parameters of the linear model can be analytically fitted using maximum-likelihood estimation. However, formulating an analytical function mapping τ onto a simple linear regression likelihood is complex. Therefore, τ was fitted iteratively using the SciPy implementation of the Brent-Dekker method for unimodal minimization (Jones, Oliphant, Peterson, et al., 2001). At each iteration, a linear model using the Semantic Surprise with the current τ value was fitted and its negative log likelihood used as cost function for the minimization. As the down-weighting function exceeded computational capacities for $\tau < 5$, the lower bound for τ was set to 5. The upper bound was set to 1000 to reflect the fact that with increasing τ, the change in the Semantic Surprise regressor decreases (please see Figure 1).

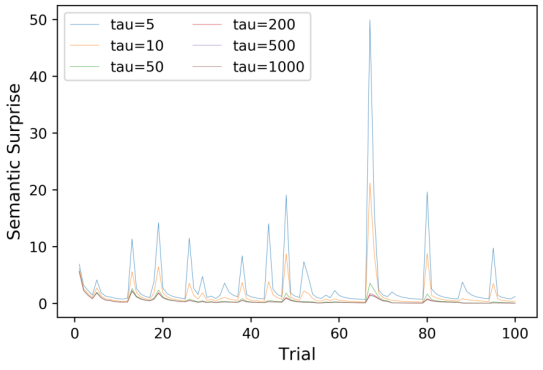

Figure 1: Semantic Surprise on 200 example trials for different values of τ.

Results

Condition-based results

There was a clear N400 effect of category switch between words. Figure 2 shows grand averages for standard and deviant stimuli at FCz. The mean difference of N400 amplitude in our ROI between the deviant and standard conditions across participants was -0.53 (SD=0.61). The difference was significant at $t(39) = -5.51, p < 0.0001$.

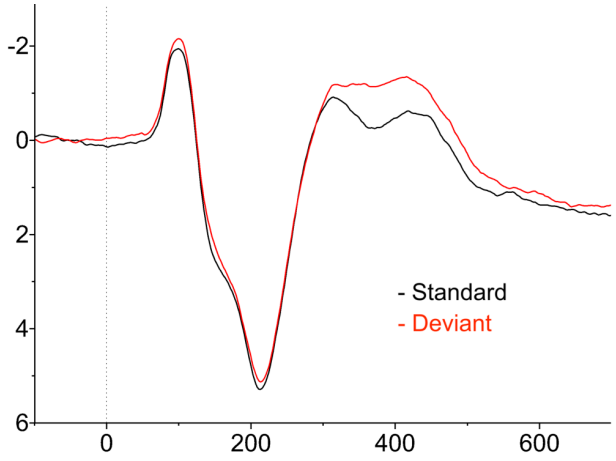

Figure 2: Grand averages at FCz averaged over all participants for the standard and deviant conditions.

Results of Semantic Surprise Modeling

At the level of individual participants, optimal tau values showed a bipolar distribution (see Figure 3). The slope of Semantic Surprise's effect on N400 mean amplitude also showed some variability (please see Figure 4). As the Semantic Surprise regressor for each participant was re-scaled by its own range, giving it a maximum of 1 and a minimum of 0, the slope may be interpreted as the amount by which Semantic Surprise at its maximum changes N400 mean amplitude compared to its minimum, and can be compared across participants.

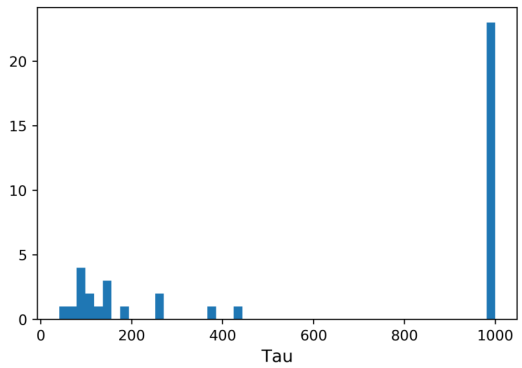

Figure 3: Histogram of optimal tau values by participant with a lower bound of 5 and an upper bound of 1000.

Discussion and Outlook

Our condition-based results confirm the basic idea that a high overlap of semantic features from one stimulus to the next in-

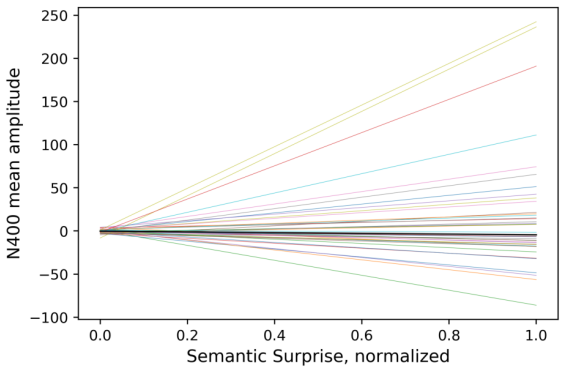

Figure 4: Regression lines for Semantic Surprise's effect on N400 mean amplitude by participant, with a black line representing the median slope and intercept.

creases N400 amplitudes, as semantic feature overlap is what characterizes words within each of our ten categories. The results of our Semantic Surprise Modeling, while showing the variability of Semantic Surprise's effect on the N400 across participants, mostly produced negative effects as expected. In addition, we found that the τ forgetting parameter varied widely, suggesting that it may be useful to take participants' individual rates of forgetting past semantic stimuli into account when using priming-related paradigms to examine the N400. The very high τ values for some participants may be interpreted to mean either that these participants had extremely low rates of forgetting, or that the massive repetition of stimuli (30 times per word) essentially prevented forgetting of past semantic input towards the end of an experimental session. This should be further explored, for example by examining the evolution of the forgetting parameter over the course of a participant's session. We demonstrate the feasibility of modeling trial-by-trial N400 amplitudes explicitly as an aspect of Bayesian semantic processing, in line with Rabovsky et al. (2018). In future analyses, we will make inferences on population parameters, and evaluate the relative model plausibility of other agent models and cognitive null models.

References

Baldeweg, T., Klugman, A., Gruzelier, J., & Hirsch, S. R. (2004). Mismatch negativity potentials and cognitive impairment in schizophrenia. *Schizophrenia Research*, *69*(2), 203–217.

Delaney-Busch, N., Morgan, E., Lau, E., & Kuperberg, G. (2017). Comprehenders Rationally Adapt Semantic Predictions to the Statistics of the Local Environment: a Bayesian Model of Trial-by-Trial N400 Amplitudes. In *Proceedings of the 39th annual conference of the cognitive science society.* London, UK.

Garrido, M. I., Kilner, J. M., Stephan, K. E., & Friston, K. J. (2009). The mismatch negativity: A review of underlying mechanisms. *Clinical Neurophysiology*, *120*(3), 453–463.

Hamp, B., & Feldweg, H. (1997). GermaNet - a Lexical-Semantic Net for German. In *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications.* Madrid.

Heister, J., Wrzner, K.-M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A., & Kliegl, R. (2011). dlexDB eine lexikalische Datenbank fr die psychologische und linguistische Forschung. *Psychologische Rundschau*, *62*(1), 10–20.

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, *49*(10), 1295–1306.

Jones, E., Oliphant, T., Peterson, P., et al. (2001). *SciPy: Open source scientific tools for Python.* ([Online; accessed May 2019])

Kullback, S., & Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86.

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*(1), 621–647.

Ostwald, D., Spitzer, B., Guggenmos, M., Schmidt, T. T., Kiebel, S. J., & Blankenburg, F. (2012). Evidence for neural encoding of Bayesian surprise in human somatosensation. *NeuroImage*, *62*(1), 177–188.

Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, *2*, 693–705.

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.

St. John, M. F., & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence*, *46*(1), 217–257.