# Distinct ERP signatures of word frequency, phrase frequency, and prototypicality in speech production.

Peter Hendrix[1]

University of Tuebingen

Patrick Bolger

University of Alberta

Harald Baayen

University of Tuebingen

[1]Correspondence to:

Telephone: +49 7071 29-76893

E-mail: peter.hendrix@gmail.com

**Abstract**

Recent studies have documented frequency effects for word n-grams, independently of word unigram frequency. Further studies have revealed constructional prototype effects, both at the word level as well as for phrases. The present speech production study investigates the time course of these effects for the production of prepositional phrases in English, using event related potentials (ERPs). For word frequency, oscillations in the theta range emerged. By contrast, phrase frequency showed a persistent effect over time. Furthermore, independent effects with different temporal and topographical signatures characterized phrasal prototypicality. In a simulation study we demonstrate that naive discrimination learning provides an alternative account of the data that is as least as powerful as a standard lexical predictor analysis. The implications of the current findings for models of language processing are discussed.

**Keywords:** ERP, picture naming, prepositional paradigm, phrase frequency, relative entropy, discrimination learning, Naive Discriminative Reader

## Introduction

N-gram frequency is the frequency of a phrase of length *n*. In recent work, Arnon & Snider (2010) showed that phrasal decision latencies for high frequency phrases such as "all over the place" are shorter than those for low frequency phrases, such as "all over the city". This effect did not reduce to frequency effects of single words or smaller n-grams. The word n-gram effect has been replicated in a number of recent studies (Bannard & Matthews, 2008; Shaoul et al., 2009, Tremblay et al., 2010; Tremblay & Baayen, 2011; Siyanova-Chanturia et al., 2011; Baayen et al., 2011).

Much, however, remains unclear about the nature of the n-gram frequency effect. Arnon & Snider (2010:76) indicate that: "At a minimum, the current findings add multi-word phrases to the units that influence processing in adults". Bannard & Matthews (2008) take the argument one step further and interpret their finding that young children process frequent phrases (e.g.; "a drink of milk") faster than infrequent phrases (e.g.; "a drink of tea") as "evidence for whole-form storage" and "representations at different levels of granularity".

An interpretation of n-gram frequency effects in terms of representations for n-grams fits well with theoretical approaches like data-oriented parsing (Bod, 2006) or memory-based learning (Daelemans & Bosch, 2005), in which large numbers of multiword sequences (or parse trees for these sequences) are stored in memory and optimal performance is ensured through on-line generalization over stored sequences. In these exemplar-based approaches n-gram frequency effects are directly related to the n-gram representations that are stored in memory.

Storing each multiword sequence and its associated frequency in memory, however, is problematic for a number of reasons. Given the Zipfian shape of frequency distributions, the number of unique n-grams is extremely large. The British National Corpus, for instance, contains 40 million unique word trigrams. Even if the storage of hundreds of millions of word n-grams were neuro-biologically possible, on-line processing over an instance space of this size would be very time-consuming. Memory-based learning as implemented in TiMBL (Daelemans, Zavrel, Sloot & Bosch, 2007), for instance, uses information gain trees (Daelemans, Bosch & Weijters, 1997) as a compression algorithm to reduce the computational demands of on-line searches.

An additional problem with n-gram representations is that it is not immediately clear what the function of such representations would be. Positing representations as a locus for a frequency "counter in the head" seems unconvincing (see, e.g.; McClelland & Rumelhart (1981) and Norris & McQueen (2008) for models that integrate word unigram frequencies as a priori-probabilities). The application of shortlists in interactive activation models (Norris, 1994) raises further questions about the necessity of n-gram representations. These models use shortlists of stored candidates as a computational shortcut that allows for simulations with realistic input sizes. The success of shortlists in these types of models indicates that at least some stored multiword sequences are not relevant for on-line processing.

These concerns have led researchers to propose alternative explanations for the effect of n-gram frequency. Tremblay et al. (2011) suggest that n-gram frequency effects may reflect past experience with (de)compositional processing. Such an interpretation fits well evidence from the learning literature demonstrating that learning is a dynamic discriminative process that is associative in nature (see Ramscar et al. (2010)). Holistic linguistic representations may be beneficial at the earliest stages of learning (Dabrowska, 2000; Tomasello, 2003), but additional experience will weaken the associations between the components of these holistic initial representations and lead to an increased importance of decomposed, lower-level representations. Learning theory therefore predicts that the adult language processing system is less likely to have separate representations for multiword units (see Dabrowska (2000), Arnon & Ramscar (2012) for a computational simulation of this process).

Baayen et al. (2013a) provided computational support for such an interpretation of the n-gram frequency effect by successfully simulating the findings of Arnon & Snider (2010) in a full decomposition model based on discrimination learning. The Naive Discriminative Reader (NDR) model used in their simulations has no representations beyond the simple word level. In the NDR model the n-gram frequency effect arises as a result of the associative learning process that maps orthographic input units (letters and letter combinations) to semantic outcomes (word meanings). A high frequency phrase such as "all over the place" is read faster than a low frequency phrase such as "all over the city", because the letters and letter combinations in "all over the place" are more associated with the meanings ALL, OVER, THE and PLACE than the letters and letter combinations in "all over the city" are associated with the meanings ALL, OVER, THE and CITY.

Thus far we discussed effects of the frequency of multi-word sequences. The prototypicality of phrases is likewise reflected in behavioral measures of language processing. Several studies have documented prototypicality effects at the word level, using relative entropy to gauge the similarity of an exemplar to its constructional prototype (Milin et al., 2009a; Milin et al., 2009b; Kuperman et al., 2010). Above the word level, relative entropy effects have been observed for English prepositional phrases (Baayen et al., 2011). Given estimated probabilities *p* (relative frequencies) of prepositional phrases for a given noun and estimated probabilities *q* (relative frequencies) of prepositions across all nouns, prepositional relative entropy is defined as

$$\text{Relative Entropy} = \sum_{i=1}^{n} \left( p_i * log_2 \left( \frac{p_i}{q_i} \right) \right) \quad (1)$$

where *n* is the number of prepositions taken into account.

The relative entropy measure compares how similar the distribution of prepositional phrase frequencies for a given noun is to the distribution of preposition frequencies in the language as a whole. Values for relative entropy are low when the prepositional phrase frequency distribution for a given noun (exemplar) is similar to the overall prepositional phrase frequency distribution (prototype) and high when the prepositional phrase frequency distribution for a given noun differs substantially from the overall prepositional phrase frequency distribution. Higher relative entropies are typically associated with greater processing costs. Nouns that use prepositions in an atypical way, for instance, take longer to process than nouns that use prepositions in a typical way (Baayen et al., 2011).

The effect of prepositional relative entropy implies that the language processing system is sensitive to the distributional properties of a noun's prepositional paradigm vis-à-vis the distribution of prepositional frequencies in the language as a whole. As such, the prepositional relative entropy effect poses a challenge to exemplar-based models. Accounting for the effect of prepositional relative entropy in such models involves three assumptions. First, in order for the distributional properties of a noun's prepositional paradigm to be available, prepositional phrases would need to be stored in the mental lexicon. We outlined the problems associated with the assumption of representations for multiword sequences above.

Second, the frequency distribution of the prototype (i.e., the frequency distribution of prepositions across all nouns) would need to be available. Storing the frequency distribution of the prototype would further increase the memory demands on the language processing system. In addition, it is unclear what function prototype representations would have beyond accounting for the effect of relative entropy. Perhaps the frequency distribution of prepositions in the language as a whole provides a reasonably accurate estimation of the frequency distribution of prepositions across all nouns that would obviate the need for the explicit storage of prototype frequency distributions.

Third, even if the language processing system contains information about exemplar and prototype frequency distributions for prepositional phrases, the distance between these distributions would need to be computed on-line. Given that Baayen et al. (2011) observed effects of prepositional relative entropy in isolated word reading, this on-line computation would need to be carried out not only when processing prepositional phrases, but any time a noun is encountered. Furthermore, if we assume that the distance between exemplars and their prototype is computed on-line for prepositional phrases, do we need to posit similar computations for other types of constructions by analogy?

Unlike exemplar-based models, discrimination learning does not need to posit any representations beyond the basic word level to account for relative entropy effects. Baayen et al. (2011) showed that the NDR model successfully captures the fact that nouns with high prepositional relative entropies (i.e.; nouns that use prepositions in an atypical way) take longer to process than nouns with low relative entropy. In the NDR model the effect of relative entropy arises as a straightforward consequence of way the distributional properties of English shape the associations between orthographic input cues and semantic outcomes across sequences of words.

## Experiment

In what follows we present the results of a primed picture naming experiment that gauges the effects of phrasal frequency and phrasal prototypicality for prepositional phrases using event-related potentials (ERPs). The current work seeks to extend previous findings in two ways. First, the experimental results for phrase frequency and relative entropy discussed thus far were mostly obtained in chronometric studies. While these studies demonstrated that both frequency and relative entropy influence how (prepositional) phrases are processed, they offer little information on the temporal details of these effects. The temporal resolution of ERPs will allow us to gauge the millisecond-by-millisecond temporal development of the phrase frequency and relative entropy effects. In addition, while the spatial resolution of ERPs is limited, the current work may provide us with a general idea about the topographical dynamics of these effects. The first goal of the current study, therefore, is to obtain a more detailed picture of the effects of phrase frequency and relative entropy that arise during prepositional phrase processing.

The second goal of the current work is to find out to what extent the temporal and spatial dynamics of the ERP signature of the phrase frequency and relative entropy effects can be replicated in the NDR model. The discriminative learning approach adopted by the NDR model has been shown to capture a wide range of effects documented in the chronometric experimental literature, including the effects of phrase frequency and phrase prototypicality. Predicting the ERP signal following the presentation of a prepositional phrase stimulus, however, involves predicting a signal as it evolves over both time and space. This stringent test of the NDR model will help gain more insight into the strengths and shortcomings of the discriminative learning approach to language processing.

The setup of the current experiment closely resembles the simulations by Baayen et al. (2011). Participants are presented with a preposition plus definite article prime, followed by a picture of a concrete noun that they have to name as fast and accurately as possible. The use of a primed picture naming paradigm might seem at odds with our interest in phrase frequency and prototypicality effects. Technically, there is no need for participants to read the preposition plus definite article primes and therefore to process the stimuli at the phrase level.

We decided to nonetheless use a picture naming paradigm for a number of reasons. First, while prepositional relative entropy is a measure of constructional prototypicality, it describes how prototypical a given noun's use of prepositions is. The effect of relative entropy is therefore best measured at the noun. In the current picture naming paradigm the earliest possible point in time where noun processing can take place is precisely defined as the moment the target noun picture appears on the screen. If we were to present the prepositional phrases as a whole it would be much harder to identify the temporal onset of target noun processing.

A related reason for using a primed picture naming paradigm is that it reduces the temporal overlap between processes related to the preposition and definite article and processes related to the noun. Experienced readers are able to read prepositional phrases in a few hundred milliseconds. Nonetheless, as will become apparent soon, ERP effects related to the lexical properties of a given word can last many hundreds of milliseconds (see, e.g.; Kryuchkova et al. (2011)). This implies that there is a temporal overlap between processes related to the different words in the prepositional phrase. In the current setup, the temporal distance between the onset of the prime and the onset of the target is 2000 ms. This allows a substantial part of the initial processing of the preposition and definite article to complete prior to the presentation of the target noun.

A third reason for using the current experimental setup is that the proof is in the pudding as far as phrase frequency effects are concerned. As noted above, the current paradigm does not guarantee that the information in the preposition plus definite article primes and that the target noun picture is integrated to obtain a phrase-level understanding of the stimulus. It is therefore possible that the current setup does not allow us to replicate the phrase frequency effect. If we do observe an effect of phrase frequency, however, this unequivocally entails that the stimuli were processed at the phrase level.

The first part of what follows describes in more detail the experiment outlined above, the analysis techniques used to analyze the data and the results of the experiment. In the second part, we will present a simulation study in which we try to replicate the temporal and spatial dynamics of the observed effects in the discriminative learning framework of the NDR model.

## Methods

*Participants*

Thirty participants took part in the experiment. All participants were students of the University of Alberta in Edmonton and native speakers of English. Their mean age was 20.4 (sd: 4.7). Nineteen participants were female, eleven were male. All participants were right-handed, had normal or corrected to normal vision and did not have a history of neurological illness. Participants received partial course credits for their participation.

*Materials*

Sixty-eight concrete nouns were paired with photographs, depicting the referent of these nouns on a beige background. For each of the nouns, four three-word prepositional phrases were constructed, consisting of a preposition, the definite article "the" and the noun itself (e.g., "with the saw", "against the strawberry").

Phrases were selected on the basis of trigram frequencies as available in the Google 1T n-gram data (Brants & Franz, 2006). Trigram frequencies for all prepositional phrases consisting of a preposition, the definite article "the" and one of the 68 concrete nouns were extracted. For a given noun, the phrases at 25%, 50%, 75% and 100% of the phrase frequency distributions were included as stimuli. For the noun *saw*, for instance, this procedure generated the experimental items "into the saw" (frequency: 2061), "from the saw" (5358), "to the saw" (9781) and "with the saw" (20464). The total number of stimuli was 272.

Only prepositions from a pre-compiled list of 35 prepositions were included in the trigram frequency list. Selecting the phrases at the quantiles of the phrase frequency distribution led to 29 of these prepositions being used in the experiment. As a result of this selection procedure, there was a significant correlation between preposition frequency and number of times a preposition was used in the experiment ($r = 0.85, p < 0.001$), with frequent prepositions such as "in" (44 times) or "on" (23 times) being included more often than infrequent prepositions such as "under" (6 times) or "against" (5 times). The experience with prepositions in the context of the current experiment therefore matches the experience with prepositions in the language as a whole.

*Design*

The experiment consisted of 272 picture naming trials. Prior to the experiment, a practice phase was included, consisting of 10 items. The order in which the stimuli were presented was randomized between participants. The dependent variable was the ERP signal measured at 32 locations on the scalp. The independent variables were *Word Length*, *Word Frequency*, *Phrase Frequency* and *Relative Entropy*.

*Word Length* is the length of the target noun in letters. *Word Frequency* and *Phrase Frequency* are the frequency of the target noun (e.g., "saw") and phrase (e.g., "with the saw") in the Google n-gram data. *Word Length, Word Frequency* and *Phrase Frequency* were log-transformed to remove a rightward skew from the predictor value distribution. *Relative Entropy* was calculated on the basis of the Google n-gram phrase frequencies for all 272 nouns used in the experiment and all 35 prepositions in the precompiled list of prepositions. Prepositional phrase frequencies were converted to relative frequencies (i.e.; estimated probabilities) for each noun and across all nouns to obtain estimated probability distributions $p$ (for a given noun) and $q$ (across all nouns). *Relative Entropy* was then calculated as the Kullback-Leibler divergence between $p$ and $q$ (see Equation 1).

Prior to analysis, we removed predictor outliers (i.e.; predictor values further than two standard deviations from the mean) from the data. This resulted in the exclusion of 1.53% of predictor values for *Word Length*, 4.61% of all predictor values for *Word Frequency*, 5.76% of all predictor values for *Phrase Frequency* and 4.61% of all predictor values for *Relative Entropy*. Outliers for *Phrase Frequency* included the 2.76% of all phrases that did not occur in the Google n-gram data, such as "up the sock" or "into the pencil". Table 1 shows the range and adjusted range for all independent variables. In addition, it presents the mean, median and standard deviation of the predictor distributions after outlier removal.

*[INSERT TABLE 1 AROUND HERE]*

*Phrase Frequency* was significantly correlated with *Word Frequency* (*r = 0.42*) and *Preposition Frequency* (the Google n-gram frequency of the preposition, *r = 0.27*). To ensure that any effect of *Phrase Frequency* was not an artifact of unigram frequency effects we therefore decorrelated *Phrase Frequency* by taking the residuals from a linear model regressing *Phrase Frequency* on *Word Frequency* and *Preposition Frequency*. The correlation of the original *Phrase Frequency* measure and the residualized *Phrase Frequency* measure was 0.77.

Given the nature of the bigrams in the current phrases the correlation of bigram frequencies with unigram frequencies was extremely high. Preposition plus definite article (e.g. "with the") bigram frequencies correlated 0.96 with *Preposition Frequency*, whereas the correlation of definite article plus noun (e.g.; "the saw") bigram frequencies with *Word Frequency* was 0.89. Both bigram frequencies, however, were not significantly correlated with residualized *Phrase Frequency* (*r = 0.07, r = 0.11*). We therefore did not decorrelate *Phrase Frequency* from the component bigram frequencies.[1]

*Relative Entropy* was significantly correlated with *Word Frequency* (*r = -0.40*) and *Phrase Frequency* (*r = -0.22*). We therefore decorrelated *Relative Entropy* from *Word Frequency* and *Phrase Frequency* by taking the residuals of a linear model regressing *Relative Entropy* on *Word Frequency* and *Phrase Frequency*. The correlation between the original and residualized *Relative Entropy* measures was 0.92.

We end this section with a methodological note on decorrelation. In the design described here, we decorrelated predictor A from predictor B by taking the residuals of a linear model regressing predictor A on predictor B. If we consider the example of decorrelating *Phrase Frequency* from *Word Frequency*, this procedure removes the variance shared by both frequency measures from *Phrase Frequency*. It does, however, not remove the same variance from *Word Frequency*. Residualizing *Phrase Frequency* from *Word Frequency* as done here allows us to determine whether there is an effect of *Phrase Frequency* over and above an effect of *Word Frequency*. It does, however, not allow us to conclude whether or not the effects of *Word Frequency* and *Phrase Frequency* are different in nature. To allow for such a conclusion, a complete decoupling of both predictors is necessary.

To completely decouple *Word Frequency* and *Phrase Frequency* we carried out a post-hoc analysis. In this analysis we used a reverse decorrelation procedure in which we decorrelated *Word Frequency* from *Phrase Frequency*. Residualized *Word Frequency* correlated 0.91 with the original *Word Frequency* measure. We then re-ran all reported models using the raw non-residualized *Phrase Frequency* measure and the decorrelated *Word Frequency* measure. The results for residualized *Word Frequency* and raw *Phrase Frequency* in this post-hoc analysis are reported after discussing the effects of the original raw *Word Frequency* and residualized *Phrase Frequency* measures introduced above and demonstrate that the effects reported here reflect true qualitative differences between the effects of *Word Frequency* and *Phrase Frequency*.

*Procedure*

Data were recorded from 32 Ag/AgCl active electrodes (Fp1, Fp2, AF3, AF4, F7, F3, Fz, F4, F8, FC5, FC1, FC2, FC6, T7, C3, Cz, C4, T8, CP5, CP1, CP2, CP6, P7, P3, Pz, P4, P8, PO3, PO4, O1, Oz, O2), which were mounted on an electrode cap (BioSemi, international 10/20 system). Reference electrodes were placed at the left and right mastoids. The EOG was recorded using electrodes below and above the left eye and at the outer canthi of both eyes. Electrode cap sizes varied from 54 to 60 cm between participants to allow for an optimal fit.

Data were sampled at 8,102 Hz using a BioSemi Active II amplification system. Prior to analysis, the signal was downsampled to 256 Hz, band-pass filtered from 0.5 to 50 Hz, baseline corrected (-200 to 0 ms interval) and re-referenced to the average of the left and right mastoids using Brain Vision Analyzer (version 1.05). In addition, the signal was corrected for eye-movements and eye blinks using the *icaOcularCorrection* package for R (Tremblay, 2010).

Verbal responses were recorded using a microphone (Sennheiser) and response box including a voice key (Serial Response Box) for the E-Prime experimental software package (version 2.0.1). The same package was used to present the stimuli on a 17 inch CRT monitor using a 1024 by 768 resolution.

A fixation mark was shown for 1000 ms prior to each trial. Next, participants were presented with a preposition plus definite article prime (e.g., “in the”) for 1000 ms. This screen was followed by another 1000 ms fixation mark screen. We then presented the photograph depicting the target noun (512 by 384 pixels) for 3000 ms. Participants were instructed to name the target noun, as depicted by the photograph. They were instructed to

respond as fast a possible, while retaining accuracy. In addition, participants were instructed to limit eye blinking and body movements to a minimum.

All fixation marks and texts were presented in white Courier New 24 point font. All fixation marks, texts and photographs were presented in the center of the screen against a black background. Each photograph was followed by a 2000 ms pause prior to the next stimulus, to allow the EEG signal to return to baseline. The experiment had a duration of about 40 minutes, excluding a preparation phase of about 30 minutes. Halfway through the experiment, participants were given a break to prevent fatigue.

## Analysis

Prior to analysis we removed 12 items corresponding to 3 problematic photographs from the data, as error rates were high for these photographs across participants. In addition, we removed incorrect naming responses from the data (7.61%). No averaging over participants or items was done prior to analysis. No channels were excluded from the analysis.

### *Generalized Additive Models (GAMs)*

This experiment examines the effect of numerical predictors over time. These effects are potentially non-linear in both the predictor dimension (at a given point in time) and the time dimension (for a given predictor value). To allow for non-linearities in multiple dimensions, we used Generalized Additive Models (GAMs) to analyze our data (Hastie & Tibshirani (1986); Wood (2006), R package *MGCV* (version 1.6-2)). GAMs have recently been used in a number of ERP studies on language processing (Baayen et al., 2013b; Kryuchkova et al., 2011).

GAMs are regression models of the form

$$y = X\beta + f(x_1, x_2, \dots) + \dots + \varepsilon \qquad (2)$$

where $y$ is the response variable, $X\beta$ is a linear predictor and $f_i$ are smooth functions of the covariates $x_k$. The parametric part of this equation ($X\beta$) is identical to that in standard regression models. The non-parametric part ($f(x_1, x_2, \ldots) + \ldots$) consists of a number of smooth functions $f_i$ and is unique to GAMs. This part of the model allows GAMs to model non-linearities in multiple dimensions.

*Reaction time analysis*

We fitted a GAM with by-participant smooth functions for trial, a random intercept for item and a smooth function for the previous reaction time to the naming latencies. Naming latencies further than 2.5 standard deviations from the mean were removed from the data. A square root transformation was applied to the naming latencies to remove a rightward skew from the data.

We modeled the predictor effects for *Word Frequency, Phrase Frequency* and *Relative Entropy* using smooth functions. We modeled the effect of *Word Length* with a parametric term, because of the limited number of unique values for *Word Length*. The reported effect for *Word Length*, however, is identical when modeled with a smooth function.

*ERP analysis*

We fitted a two stratum hierarchical GAM to the ERP data. In the first stratum we removed participant- and item-related variability, as well as task effects and the grand average over time from the data. In the second stratum we looked at the predictive power of our linguistic covariates using tensor products smooths for time by predictor. Alternatively, we could have opted for a single stratum modeling strategy, in which participant- and item-related variance, task effects, the grand average over time and predictor effects are entered into the model simultaneously. The results of such a holistic modeling strategy for the key effects reported below are presented in Appendix B.

*Stratum one*

We used a restricted cubic spline for the main trend over *Time* and the effect of *Trial Count*, which gauges effects of fatigue or habituation. The effect of *Participant* was modeled with participant-specific

quadratic polynomials. To model item-related variance, we included quadratic polynomials for both prepositional phrase (e.g., “with the”) and noun (e.g., “saw”).

Due to the computational demands of GAMs and the size of our data set, we decided to analyze the ERP signal in four epochs: 0 to 300 ms, 200 to 500 ms, 400 to 700 ms and 600 to 900 ms after picture onset. We use a 100 ms overlap between epochs to verify consistency of results for subsequent epochs. We refer the interested reader to Appendix B for an analysis of the key predictor effects reported below using larger time windows and a more detailed discussion of why a 100ms overlap between subsequent epochs is necessary.

Figure 1 shows the main trend over time at electrode Cz as predicted by our main trends GAM (solid black lines). Predicted main trend values correlate highly with average observed voltages (red dots), $r = 0.87$. This indicates that our main trends model successfully captures the general trend of the ERPs over time. Main trend model fits correlated highly with averaged observed voltages across all electrodes, with an average correlation of $r = 0.97$ between predicted main trend values and average observed values.

*[INSERT FIGURE 1 AROUND HERE]*

The average reaction time in the experiment was 823 ms (median: 794 ms). The earliest responses started coming in much earlier than that, with a minimum reaction time of 238 ms. As a consequence, electromyographic (EMG) potentials arising from the facial, jaw and tongue muscles are present in a substantial subset of our data. These EMG potentials could therefore impoverish the signal-to-noise ratio (SNR) for this subset of the data.

There are two options for dealing with EMG activity in our data. First, we could remove all data points after the onset of articulation. There are a number of problems with this approach. As noted by Hillyard & Picton (1987) muscle artifacts may well be present long before speech onset. Even if we were to remove all data points following the onset of articulation, EMG artifacts would therefore remain in the data. In addition, the number of data points would substantially differ between epochs. The left panel of Figure 2 illustrates this problem. From 0 to 400 ms very few data points are potentially affected by articulation artifacts: at 400 ms after stimulus onset, articulation has started for only 1.8% of all trials. In the third and fourth epoch, however, articulation has begun

for a significant portion of data points: 5.2% at 500 ms, 11.3% at 600 ms, 25.0% at 700 ms and 42.1% at 800 ms. At 900 ms after stimulus onset, articulation has already started for 57.1% of trials. Removing all these data points would lead to a drastic reduction of statistical power in the third and fourth epoch.

The second option for dealing with EMG activity is to include all data points, even those for which articulation artifacts might be present. While this approach ensures an equal amount of data for each point in time, it does not necessarily solve the problem of reduced statistical power in the later epochs. If EMG artifacts have a negative effect on the SNR in the last two epochs it becomes harder for statistical models to identify predictor effects in these epochs. To gauge the severity of this problem, we calculated the root mean square (RMS) for all electrodes. The right panel of Figure 2 shows the average RMS across all electrodes as a function of time. In the pre-stimulus interval (-100 to 0 ms), the average RMS across all electrodes and time points is 6.47, whereas in the post-stimulus interval (0 to 900 ms) it is 7.94. As predicted, the RMS does increase as a function of time. Fortunately, however, the increase is limited: the average RMS is 7.67 in the 0-300 ms interval, 8.10 in the 300-600 ms interval and 8.05 in the 600-900 ms interval.

To further inspect the potential problem of a decreased SNR due to articulation artifacts we looked at the RMS across electrodes in the last epoch (600 to 900 ms). If articulation introduces noise in the signal, we would expect this noise to be most prominent at frontal electrodes, which are closest to the facial and tongue muscles. RMS averages in the last epoch were indeed somewhat elevated at frontal locations. While the average RMS across all electrodes in the last epoch was 8.05, the average RMS values in the last epoch at frontal electrodes were 9.39 (Fp1), 9.27 (Fp2), 9.19 (AF3), 8.61 (AF4), 8.92 (F7), 8.81 (F3), 7.22 (Fz), 7.75 (F4), 8.80 (F8). Even though the average RMS values at frontal electrodes in the last epoch are somewhat elevated, these values suggest that the amount of noise introduced by EMG activity is fairly limited.

Despite the limited increase in RMS values over time, articulation artifacts could nonetheless be problematic if they vary systematically with our predictors of interest. To rule out this possibility, we compared the results of an analysis on the full data set to the results of an analysis on a subset of the data that only included data points before articulation onset. Most of the predictor effects that were significant in the full data set remained significant for the subset of the data. Furthermore, these predictor effects were qualitatively highly similar for the full data set and the pre-articulation subset of the data. We therefore decided to carry out our

analysis on the full data set, including data points after articulation onset. Whenever an effect was not significant in the pre-articulation subset of the data we explicitly mention this when discussing this effect.

*Stratum two: predictor tensor products*

We fitted GAMs with tensor product smooths for time by predictor on the residuals of the stratum one models. We observed similar, but less conservative results when fitting the predictor GAMs to the original ERP signal. For computational efficiency, we fitted a separate model for each of our predictors. Similar results were obtained when using a hierarchical approach, in which each predictor GAM was fitted on the residuals of the previous predictor GAM and when using a multiple regression approach in which all predictor smooths were entered simultaneously. We used tensor product smooths with restricted cubic spline basis functions. To address the problem of multiple comparisons (32 electrodes, 4 epochs), we adopted a Bonferroni-corrected significance level of 0.0004 for all predictor GAMs.

The use of regression models has become commonplace in experimental studies investigating predictor effects on unidimensional dependent variables, such as reaction time studies. The application of regression type models in ERP studies, however, is much less widespread. To allow for a better understanding of the analysis technique used here and the advantages GAMs offer in comparison to a traditional ERP analysis we compare the current GAM analysis to a traditional ERP analysis for simulated data, as well as for some of the key predictor effects described below in Appendix A.

## Results

*Reaction time results*

The naming latencies showed a marginally significant effect of *Word Length* ($F = 3.367$, $p = 0.067$). This marginally significant effect of *Word Length* was linear in nature, with longer naming latencies for longer words. The effect of *Word Length* is depicted in Figure 3. For ease of interpretation, normal linear naming latencies are plotted rather than the square root transformed latencies used for modeling. No significant effects were observed

for *Word Frequency, Phrase Frequency* and *Relative Entropy*.

*ERP results*

In this section, we will discuss the results for the predictors *Preposition Frequency*, *Word Frequency*, *Phrase Frequency* and *Relative Entropy*. For each predictor, we visualize the observed effect over time at a representative example electrode.

*Word Length*

Figure 4 shows the contour plot of the tensor surface for *Word Length*. The x-axis represents time (in ms), with the four panels showing the development of the effect over the four epochs (0 to 300 ms, 200 to 500 ms, 400 to 700 ms and 600 to 900 ms) at a representative example electrode. *Word Length* is on the y-axis. The contour plot represents voltages at the depicted electrode, with warmer colors representing higher voltages. Contour lines are shown at intervals of 0.2 $\mu V$. Above each panel, the *p* value for the effect at the depicted electrode is given. As recommended by Wood (Wood, 2006) we used Bayesian *p*-values rather than the standard frequentist *p*-values, as Bayesian *p*-values have improved frequentist performance over the strictly frequentist approximation. Significant *p*-values at the Bonferroni-corrected alpha level are displayed in red.

*[INSERT FIGURE 4 AROUND HERE]*

The first panel of Figure 4 shows an effect of *Word Length*. For long words, voltages are negative, then positive, then negative again and then positive again. In other words, we see oscillations for long words. To determine the frequency of these oscillations, we converted the time domain representation of the ERP signal seen in the first panel of Figure 4 to the frequency domain. Maximum spectral intensity for the oscillations is reached at 7 Hz. These oscillations are therefore theta range oscillations (3-7 Hz).

Previously, theta range activity has been observed in a number of language processing studies and has been demonstrated to be related to, for instance, lexical-semantic retrieval (Bastiaansen et al., 2005, Bastiaansen et al., 2008), syntactic processing (Bastiaansen et al., 2002) and translation (Grabner et al., 2007). In a regression

study using GAMs, Kryuchkova et al. (2012) recently reported theta range oscillations in auditory comprehension tied to word frequency, phonological neighborhood density and morphological family size. Theta range oscillations are thought to reflect (working) memory demands in language processing that arise from the synchronous firing of neurons in hippocampal areas (see Bastiaansen and Hagoort (2003) for a comprehensive discussion of theta range oscillations).

Each panel of Figure 4 contains a picture inset. Picture insets show the topography of the effect in each epoch, with bright red indicating significance at the Bonferroni-corrected alpha level ($p < 0.0004$) and dark red indicating significance at the non-corrected alpha level ($p < 0.05$). As can be seen in the inset in the left panel of Figure 4, the early oscillatory effect of *Word Length* is topographically widespread, with peak amplitudes at central-parietal electrodes in the left hemisphere.

The oscillatory effect of *Word Length* continues into the second epoch, where, again, we observed oscillations for long words. These oscillations now have a somewhat reduced frequency (5 Hz) and are topographically less widespread, with peak amplitudes in left-lateralized parietal-occipital regions. In the third epoch the oscillations for long words fade out, with low amplitude oscillations at the start of the third epoch in left-central areas only. In the fourth epoch, no topographically consistent effect of *Word Length* was observed at the Bonferroni-corrected alpha level.

To gauge the temporal development of the oscillatory effect for *Word Length*, we calculated three sigma (99.7%) confidence intervals around the contour surfaces. The first point in time at which 0 is not within this three sigma confidence interval is 131 ms after picture onset.[2] The early onset of the Word Length effect is in line with previous work by Hauk et al. (2006), who found an effect of word length in visual word recognition starting at 90 to 100 ms after stimulus onset. The last point in time at which the effect of *Word Length* is statistically significant is 446 ms after picture onset.

*Word Frequency*

Figure 5 shows the effect of *Word Frequency*. As for *Word Length,* oscillations tied to *Word Frequency* arise in the first epoch, for high frequency and - to a lesser extent - for low frequency words. These theta range oscillations have maximum spectral intensity at 5 Hz. The effect of *Word Frequency* in the first epoch is

topographically widespread, but the oscillations reach peak amplitudes in left-frontal areas. As for *Word Length*, the effect of *Word Frequency* arises early: it is first significant at 96 ms post stimulus onset. The early onset of the frequency effect is in line with previous findings (Hauk et al., 2006; Sereno et al., 1998), reporting effects of lexical frequency in visual word recognition starting at 110 and 132 ms.

The early oscillatory activity for *Word Frequency* fades out in the second epoch. Oscillations for high frequency words are still present, but decrease in amplitude throughout the second epoch. They are last significant at 379 ms after picture onset. Topographically, the effect remains widespread and still peaks at left-frontal electrodes.

*[INSERT FIGURE 5 AROUND HERE]*

While we primarily see an effect for high frequency words in the first two epochs, an effect for low frequency words emerges in the last two epochs. This late re-emergence of the *Word Frequency* effect is present in the full data set that includes data points post articulation onset, but not in the subset of the data that includes data points before the onset of articulation only. Subtle oscillations at the bottom of the theta range (peak spectral intensity: 4 Hz) arise in the third epoch at left-lateralized frontal, central and parietal electrodes and first reach significance at 584 ms after picture onset. In the fourth epoch, these oscillations become more pronounced: the amplitude increases and the topographical distribution becomes wider. The effect remains significant until the end of the fourth epoch, 900 ms after picture onset.

The earlier onset of the oscillations for high frequency words as compared to low frequency words fits well with the classic frequency effect in reaction time studies on speech production: high frequency words are produced faster than their low frequency counterparts (Oldfield & Wingfield, 1965; Bates et al., 2003; Jescheniak & Levelt, 1994). Recently, a very similar effect of word frequency has been observed in an auditory comprehension ERP study. Consistent with the current findings, Kryuchkova et al. (2012) reported early theta range oscillation for both high and low frequency words, with maximum amplitudes for high frequency words, as well as late oscillations that were exclusive to low frequency words.

In a post-hoc analysis we residualized *Word Frequency* from *Phrase Frequency* to completely decouple

*Word Frequency* and *Phrase Frequency*. As for raw *Word Frequency* we observed theta range oscillations for both high and low frequency words in the first and second epoch. These oscillations had similar phases to the oscillations reported above. In addition to these theta range oscillations we observed later prolonged negativities for high frequency words and prolonged positivities for low frequency words. When the 2.00% most extreme predictor values at both ends of the residualized *Word Frequency* distribution were removed the data, however, this positivity and negativity were no longer present in the data. Instead, we found 4 Hz oscillations in the fourth epoch that are most pronounced for low frequency words (but also present for medium and high frequency words near the end of the fourth epoch) and that are similar in phase to the late 4 Hz oscillations reported for *Word Frequency* below. This deviation from the results reported above therefore seems to be an outlier effect with limited statistical robustness. As such, we conclude that the results for residualized *Word Frequency* are similar to those reported for raw *Word Frequency* above.

*Phrase Frequency*

Figure 6 shows the effect of *Phrase Frequency*. In contrast to the effects of *Word Length* and *Word Frequency*, the effect of *Phrase Frequency* is not oscillatory in nature. Instead, we see persistent negativities for both high and low frequency phrases at left frontal, central and parietal electrodes.[3] These negativities arise in the first epoch and are first significant at 70 ms after stimulus onset for high frequency phrases. For low frequency phrases, the onset of the effect is at 172 ms after stimulus onset. The negativities for both low and high frequency words continue in the second epoch, with a widespread topographical distribution and peaking at left frontal, central and parietal electrodes. In addition, we observed a more transient positivity for medium to high frequency words that arises in the second half of the second epoch. This positivity continues until halfway through the third epoch.

*[INSERT FIGURE 6 AROUND HERE]*

In the third epoch, negativities for both high and low frequency words that peak at left frontal, central and parietal electrodes remain. While the negativities for low frequency words continue throughout the epoch,

the negativities for high frequency words are significant only until 591 ms after picture onset. Persistent negativities for low frequency phrases remain throughout the fourth epoch, but only reach significance until 768 ms after stimulus onset. In addition, positivities arise for medium to low frequency words. As for the earlier effects of *Phrase Frequency*, this late effect is widespread, but most prominent at left-lateralized frontal and central electrodes.

The effect for high frequency phrases starts and ends earlier than that for low frequency phrases. This finding is in line with the temporal development of the *Word Frequency* effect, which showed earlier oscillations for high frequency words than for low frequency words. As for the effect of word frequency, the effect of *Phrase Frequency* therefore fits well with reaction time studies that found faster responses to high frequency phrases than to low frequency phrases (Arnon & Snider, 2010, Bannard & Matthews, 2008; Shaoul et al., 2009, Tremblay et al., 2009; Tremblay & Baayen, 2010; Siyanova-Chanturia et al., 2011).

The effect of *Phrase Frequency* is primarily characterized by prolonged effects that continue over substantial periods of time. In addition to the effects of *Phrase Frequency* reported above we found some evidence for oscillatory activity tied to *Phrase Frequency*, with theta range oscillations for both high and low frequency words that were most prominent at left-central frontal to parietal electrodes in the 0-300 ms time window and later theta range oscillations at right frontal electrodes in the third epoch, at right-lateralized central-parietal electrodes in the fourth epoch and at left-lateralized and central central-to-occipital electrodes in the third and fourth epoch. It is therefore possible that the prolonged effects of *Phrase Frequency* reported above exist in addition to theta range oscillations for *Phrase Frequency*.

The statistical robustness of the oscillatory activity tied to *Phrase Frequency*, however, is questionable. Even for the overlapping segment of 200-300 ms, the analysis for the 200-500 ms time window does not provide any evidence for the early theta range oscillations in the 0-300 ms time window. Similarly, the oscillations at right frontal electrodes in the third epoch are not supported for the overlapping segments in the second and in the fourth epoch. In addition, the oscillations at left-lateralized and central central-to-occipital in the third and fourth epoch are attenuated or replaced by prolonged negativities or positivities when looking at the pre-articulation subset of the data only. While it is possible that oscillatory effects of *Phrase Frequency* exist in addition to the prolonged effects of *Phrase Frequency* reported above, the evidence for such oscillations is therefore too limited

to consider these effects statistically robust. As such, we decided not to discuss these oscillations in more detail here.

To allow for a complete decoupling of the *Word Frequency* and *Phrase Frequency* effects we carried out a post-hoc analysis using raw *Phrase Frequency* as a predictor, rather than the *Phrase Frequency* measure residualized on *Word Frequency* that we used in the analyses reported above. Overall, we found a similar pattern of results for raw *Phrase Frequency* and residualized *Phrase Frequency*. Although the temporal onset of these effects was somewhat delayed as compared to the effect of *Phrase Frequency* reported above (see also the simulated effect of *Phrase Frequency* in the NDR model reported below), we found persistent negativities for both high and low values of raw *Phrase Frequency*. As for residualized *Phrase Frequency*, the negativity for high values of raw *Phrase Frequency* faded around 600 ms after picture onset, whereas the negativity for low values of *Phrase Frequency* continued into and throughout the fourth epoch. Furthermore, we again observed more transient positivities for medium-to-high values of raw *Phrase Frequency* at the end of the second epoch and the start of the third epoch, as well as for low-to-medium values of raw *Phrase Frequency* at the end of the third and in the middle of the fourth epoch.

As for the post-hoc analysis for *Word Frequency*, we found some deviations from the results reported above for the most extreme values of raw *Phrase Frequency*. First, we saw a transient positivity for high values of *Phrase Frequency* in the fourth epoch (600-875 ms) that was not present for residualized *Word Frequency*. Second, below the negativity for low values of *Phrase Frequency*, we saw a positivity for extremely low values of *Phrase Frequency* throughout all 4 epochs, that started at 0 ms after picture onset and continued until 900ms after picture onset. Potentially, these effects may be related to properties of *Word Frequency* that are present in the raw, but not in the residualized *Phrase Frequency* measure. Given that we found no corresponding effects for *Word Frequency*, however, it is unclear how likely such an interpretation of this late transient positivity for raw *Phrase Frequency* is. Also, the onset of the persistent positivity for the lowest values of *Phrase Frequency* at 0 ms post picture onset poses some questions regarding the robustness of this effect. Independent of the status of these additional effects observed for raw *Phrase Frequency*, however, it is clear that the effect of residualized *Phrase Frequency* reported above reflects lexical properties that are present in the raw *Phrase Frequency* measure as well. Given the results of the post-hoc analyses for raw *Word Frequency* and raw *Phrase Frequency*

we conclude that the results reported here reflect true qualitative differences between the effects of *Word Frequency* and *Phrase Frequency*.

The temporal onset of the effects of both *Word Frequency* and *Phrase Frequency* is similar, with early effects for both predictors. In addition, the effects for both predictors peak in left-frontal and left-central areas throughout all four epochs. Nonetheless, the qualitative nature of the *Word Frequency* and *Phrase Frequency* effects is different, with theta range oscillations characterizing the effect of *Word Frequency* and persistent negativities being the most prominent feature of the *Phrase Frequency* effect. The qualitative differences between the effects for *Word Frequency* and *Phrase Frequency* are reflected in the absence of a correlation between the contour surfaces for *Word Frequency* and *Phrase Frequency* ($r < 0.01$, $p = 0.897$). We will return to the dissociation between the effects of *Word Frequency* and *Phrase Frequency* shortly.

*Relative Entropy*

Figure 7 presents the effect of *Relative Entropy*. In the first epoch, we observed 7 Hz oscillations throughout the predictor range at parietal-occipital electrodes. This effect first reaches significance at 111 ms after picture onset. In the second, epoch, 7 Hz theta oscillations continue throughout the predictor range at parietal and occipital electrodes across both hemispheres. The oscillations are last significant at 381 ms after stimulus onset.

*[INSERT FIGURE 7 AROUND HERE]*

In the third and fourth epoch the effect of *Relative Entropy* re-emerges, with low-frequency 4 Hz oscillations for high predictor values that first reach significance at 564 ms and that are last significant until 831 ms after stimulus onset. The effect in the third epoch is present for both the full set of the data and the subset of the data that includes data points before the onset of articulation only. The effect of *Relative Entropy* in the fourth epoch, however, is absent for the pre-articulation subset of the data. In contrast to the early effect of *Relative Entropy*, this later effect peaks at left-lateralized central and frontal electrodes. While we also see more subtle oscillations for words with low *Relative Entropy* in the bottom left of the fourth panel of Figure 5, these

oscillations do not reach significance.

Reaction time studies reported increased response latencies for words with high relative entropies (Milin et al., 2009a; Milin et al., 2009b; Kuperman et al., 2010; Baayen et al., 2011). The current pattern of results fits well with these findings. The early 7 Hz oscillations reach maximum amplitude for high values of *Relative Entropy* and the late 4 Hz oscillations are exclusive to words with high *Relative Entropy*. The current results therefore indicate that additional processing is required for nouns with atypical prepositional phrase frequency distributions as compared to nouns that use prepositions in a more typical way.

The effects of *Relative Entropy* and *Word Frequency* show remarkable similarities. While the topographies of the early effects are different, both predictors give rise to 7 Hz oscillations for both high and low predictor values that arise around 100 ms after picture onset. For both predictors, these oscillations fade out around 400 ms after stimulus onset. In addition, we see a late effect of both *Relative Entropy* and *Word Frequency* for those predictor values that have been demonstrated to lead to increased reaction times in chronometric studies: both low values of *Word Frequency* and high values of *Relative Entropy* give rise to late 4 Hz oscillations in left-lateralized frontal areas.

The similarity of the effect of *Relative Entropy* and *Word Frequency* is complimented by the nature of the *Word Length* effect. While there was no late manifestation of *Word Length* in the ERP signal, the effect of this third lexical predictor was characterized by left-lateralized 7 Hz oscillations that arise around 100 ms after picture onset and fade out around 400 ms. The effect of *Word Length* is therefore similar to the early effects of *Word Frequency* and *Relative Entropy*. The qualitative, temporal and topographical overlap between the effects of the three word level predictors is food for thought. Given the relevance of the simulation study that follows for the implications of these similarities, we will return to this issue in the General Discussion section of this paper.

*Controls*

In addition to the effects reported above, we observed theta range (7 Hz) oscillations related to picture complexity (Jpg size in bytes) throughout all epochs and for all predictor values. These oscillations peaked in left-central parietal-occipital areas and had higher amplitudes for more complicated pictures than for less

complicated pictures. Furthermore, we observed effects of both preposition length and preposition frequency. For preposition length we observed early 7 Hz oscillations for long prepositions at left-frontal electrodes, as well as late 5 Hz oscillations at left-frontal and left-parietal electrodes. For preposition frequency, we observed 7 Hz oscillations for low frequency prepositions throughout the second, third and fourth epoch. These oscillations peaked at left-lateralized parietal locations, but were also present in central and occipital areas. Controlling for the effects of picture complexity, preposition length and preposition frequency did not significantly affect the results reported for the predictors of interest above.

## Discussion

In the current experiment, we observed effects of both word-level and phrase-level predictors in a primed picture naming paradigm. At the word level, theta range oscillations characterized the ERP signatures of *Word Length*, *Word Frequency* and *Relative Entropy*. All three word level effects arose early: they were first significant at 110 ms, 96 ms and 108 ms after picture onset. In addition, while some differences in topographies were observed, all word level effects were significant across a wide range of electrodes in the left hemisphere. The ERP signature of *Phrase Frequency* was qualitatively different and characterized by persistent negativities for both extreme predictor values and more transient positivities for phrases with intermediate frequencies. The timing and topography of the *Phrase Frequency* effect, however, were more similar to the timing and topography of the word level effects: the effect for *Phrase Frequency* arose early and was most prominent in the left hemisphere. How should we interpret this pattern of results?

In exemplar-based approaches such as data-oriented parsing (Bod, 2006) or memory-based learning (Daelemans & Bosch, 2005) phrase frequency effects are explained through the existence of phrase representations. The frequency count associated with a phrase representation determines how quickly that phrase representation can be accessed, just like the frequency count associated with a word representation determines how quickly that word can be accessed. While exemplar-based models correctly predict that there should be

temporal and spatial overlap between the effects of word frequency and phrase frequency, it is unclear how such models would account for the qualitatively different pattern of results observed for *Word Frequency* and *Phrase Frequency* in the current experiment.

Perhaps the apparent incompatibility of exemplar-based models with the current findings results from the fact that exemplar-based models are implemented at a certain level of abstraction. Exemplar-based models represent words and phrases as discrete units or sets of finer-grained discrete feature-value pairs. This discretization is an obvious oversimplification of the neuro-biological processes that the ERP signal taps into. In these processes word or phrase representations are more likely to consist of firing patterns of assemblies of neurons. Given our limited understanding of the neuro-biological reality of language processing it is possible that conceptually similar representations for words and phrases correspond to qualitatively different neural firing patterns with qualitatively different manifestations in the ERP signal.

Nonetheless, it is clear that at this point in time exemplar-based models do not straightforwardly account for the differences between the observed word and phrase frequency effects. Furthermore, accounting for relative entropy effects in exemplar-based models would involve the conceptually and computationally unappealing assumption that online computation over stored frequency distributions for both exemplars and prototypes takes place. The current pattern of results therefore poses a challenge to exemplar-based models.

Discrimination learning provides an alternative account for the effects of word frequency, phrase frequency and relative entropy. Baayen et al. (2011) successfully replicated chronometric effects of prepositional relative entropy and phrase frequency in the Naive Discriminative Reader (NDR) model. In what follows, we will explore to what extent the NDR model is able to capture the complex ERP signatures for *Word Length*, *Word Frequency*, *Phrase Frequency* and *Relative Entropy* observed here. First, we will introduce the NDR model in more detail. Next, we will describe a simulation study in which we sought to replicate the current experimental results in the NDR model. Finally, we will present the results of this simulation for each of our four predictors of interest.

**Naive Discriminative Reader**

The Naive Discriminative Reader (Baayen et al., 2011) is a model of language processing that learns associations between letters and letters combinations on the one hand and basic word meanings on the other hand. The associations are learned through the Rescorla-Wagner equations (Wagner & Rescorla, 1972), which are mathematically equivalent to the delta rule (Sutton & Barto, 1981). Given the association strength $V_i^{t+1}$ between outcome *O* and cue $C_i$ at time *t*, the Rescorla-Wagner equations provide the association strength at time $t + 1$:

$$V_i^{t+1} = V_i^t + \Delta V_i^t \tag{3}$$

where the change in association strength, $\Delta V_i^t$, is defined as:

$$\Delta V_i^t = \begin{cases} 0 & \text{if } \mathrm{ABSENT}(C_i, t) \\ \alpha_i \beta_1 (\lambda - \sum_{\mathrm{PRESENT}(C_j, t)} V_j) & \text{if } \mathrm{PRESENT}(C_j, t) \ \& \ \mathrm{PRESENT}(O, t) \\ \alpha_i \beta_2 (0 - \sum_{\mathrm{PRESENT}(C_j, t)} V_j) & \text{if } \mathrm{PRESENT}(C_j, t) \ \& \ \mathrm{ABSENT}(O, t) \end{cases} \tag{4}$$

The NDR uses the default settings for all parameters: $\lambda = 1$, all $\alpha$'s equal, and $\beta_1 = \beta_2$. As can be seen in Equation 4, the association between a cue and an outcome increases if the outcome occurs when the cue is present and decreases if the outcome does not occur when the cue is present.

The Rescorla-Wagner equations have a temporal dimension: they describe the development of the association strengths over time. The NDR model uses the Danks equations (Danks, 2003) as a mathematical shortcut to the association strength for the equilibrium state of the model – i.e.; the state of the model in which the association strengths do not change from time *t* to time *t + 1*. These equilibrium equations define the association strength ($V_{ik}$) between cue ($C_i$) and outcome ($O_k$) as:

$$\Pr(O_k|C_i) - \sum_{j=0}^{n} \Pr(C_j|C_i)V_{jk} = 0 \quad (5)$$

with $\Pr(C_j|C_i)$ the conditional probability of cue $C_j$ given cue $C_i$, $\Pr(O_k|C_i)$ the conditional probability of outcome $O_k$ given cue $C_i$ and $n + 1$ the number of unique cues. As shown in equation 5, the association strengths are calculated independently for each outcome. This is simplification is similar to that in Naive Bayesian Classifiers and inspired Baayen et al. (2011) to refer to their model as an instantiation of *naive* discrimination learning.

When a specific word or phrase is presented as an input, only the subset of letter combination cues present in that word or phrase will become active. The extent to which these cues activate the target meaning outcome (in case a single word is presented) or outcomes (in case multiple words are presented) is a measure of how hard it is to access the meaning of a word or phrase. The activation of the set of target meanings $O$ given the set of active input cues $C$ is defined in the NDR as:

$$a_i = \sum_{k \in O} \sum_{j \in C} V_{jk} \quad (6)$$

where $j$ ranges over the active cues, $k$ ranges over the active outcomes and $V_{jk}$ is the equilibrium association strength for cue $C_j$ and outcome $O_k$.

## NDR Simulation

The NDR model is a model of reading: the input cues are orthographic in nature, while the outcomes are word meanings. The task in the current experiment, however, involves much more than simple reading. The orthographic presentation of the preposition and definite article is line with the nature of the NDR model. The target noun, however, is depicted in a photograph. Ideally, therefore, a simulation of the current data would involve an additional discrimination network mapping visual features of the photograph onto the word meaning

of the target noun. While we are exploring how to implement a visual discriminative learning network in ongoing research, no such network is implemented in the current version of the NDR model. We therefore decided to use orthographic input cues not only for the preposition and the definite article, but also for the target noun. While orthographic cues are an obvious oversimplification of the rich visual input provided by the photographs, the simulation results reported below indicate that the orthography to meaning mappings are a satisfactory proxy for the mappings from visual features to meanings.

A second discrepancy between the experimental setup and the current implementation of the NDR model concerns the nature of the task. While the NDR model is a reading model, the task in the current experiment involves naming the target noun. Recently, Hendrix et al. (2013) implemented the $NDR_a$ model, an extension of the NDR model for reading aloud. The $NDR_a$ consists of two networks: a network mapping orthographic cues onto meanings outcomes and a network mapping meanings onto acoustic features (diphones). The $NDR_a$ replicates the successful simulation of a large number of predictor effects in the NDR model – including the effects of word frequency, word length and relative entropy[3]. In addition it captures a number of findings that are specific to the reading aloud literature, such as effects of the consistency of orthography to phonology mappings and a pseudohomophone advantage for nonwords.

Nonetheless, we decided to use the original NDR reading model for the current simulation. A first reason for using the original NDR model is that the current task is a somewhat of a hybrid between production and comprehension. At the word level, the task very much resembles a naming aloud task, albeit with visual rather than orthographic input. At the phrase level, however, no overt responses are required. The effect of *Phrase Frequency* is an effect of implicit phrase-level comprehension, not of phrase-level production. While ideal for word-level simulations, therefore, the architecture of the $NDR_a$ is less than optimal for phrase-level simulations.

Second, despite the fact that the orthography to phonology mapping in English is inconsistent at times, there is considerable isomorphism between the orthographic and the phonological representations of words. As a result, there is a fair amount of overlap between the information learned by a discriminative learning network from orthography to semantics and the information learned by a discriminative learning network from phonology to semantics. For the set of 2,416 monosyllabic words used by Hendrix et al. (2013), for instance, the activation of the target word meaning from the orthography is highly correlated with the activation of the target word

meaning from the phonology ($r = 0.41$, $p < 0.001$).

Third, the original NDR model captures the chronometric effects of the predictors of interest in this study. If the same holds for the ERP effects reported here, the most parsimonious account for the data is the original NDR model, which consists of one, rather than two discriminative learning networks. Before attempting a simulation in the more complex $NDR_a$ model, it is therefore worth seeing how well the original NDR model captures the current pattern of results.

In order to learn the associations between input cues and word meanings the NDR model needs to trained on a representative language sample. Following Baayen et al. (2011) we trained the NDR model on the British National Corpus (henceforth BNC; Burnard (1995)). The training data for the current simulation consisted of 100 million word trigrams from the BNC, using letter trigrams as input cues and word meanings as outcomes. This training regime resulted in a set of 9238 unique orthographic input cues and a set of 71067 unique meaning outcomes.

In the current primed picture naming paradigm, participants are presented with a preposition plus definite article prime prior to seeing the target picture. Their task is to name the target picture as fast and accurately as possible given the presentation of the prime and target. For the simulation of the word level effects of *Word Length, Word Frequency* and *Relative Entropy,* we are therefore interested in the activation of the meaning of the target noun given the presentation of the preposition, the definite article and the noun. We obtained this simulated target noun activation by summing the associations between all letter trigrams in the input phrase and the target noun meaning (see Equation 6). For the example phrase "into the onion", for instance, we summed the associations between the letter trigrams *#in, int, nto, to#, o#t, #th, the, he#, e#o, #on, oni, nio, ion* and *on#* (word boundaries are represented by hash marks in the NDR input encoding) and the meaning ONION.

For the simulation of the *Phrase Frequency* effect, the activation of the meaning of the full phrase is of interest. In the full-decomposition NDR model simulated phrase activations are defined as the summed activations of the meanings of the component words. For the example phrase "into the onion", we therefore summed the association between the component letter trigrams (*#in, int, nto, to#, o#t, #th, the, he#, e#o, #on, oni, nio, ion* and *on#*) and the meaning INTO, between the letter trigrams and the meaning THE and between the letter trigrams and the meaning ONION. We then summed these activations to obtain the activation of the phrase

meaning INTO THE ONION.

We calculated simulated *Word Activation* values for all 68 target nouns and simulated *Phrase Activation* values for all 272 phrases used in the experiment. Following Baayen et al. (2011) we applied an inverse and logarithmic transformation to all activations to remove a rightward skew from the data. Consistent with previous NDR simulations, we added a back off constant of 0.10 to all activations to prevent division by zero when applying the inverse transformation.

Prior to the analysis of the experimental data, we decorrelated *Phrase Frequency* from *Word Frequency* and *Preposition Frequency*. To allow for a direct comparison of the effect for *Phrase Activation* to the effect for *Phrase Frequency* we applied the same decorrelation procedure to *Phrase Activation* (i.e.; we use the residuals of a linear model predicting *Phrase Activation* from *Word Frequency* ($r = -0.09$) and *Preposition Frequency* ($r = -0.30$)). The residualized *Phrase Activation* measure correlated 0.91 with the non-residualized, original *Phrase Activation* measure. *Phrase Activation* did not significantly correlate with *Phrase Frequency*, neither before ($r = 0.07$, $p = 0.264$) nor after residualization ($r = 0.07$, $p = 0.293$). By contrast, *Word Activation* correlated significantly with the word level predictors *Word Length* ($r = 0.29$, $p < 0.001$), *Word Frequency* ($r = -0.39$, $p < 0.001$) and *Relative Entropy* ($r = 0.15$, $p = 0.025$).

The rationale behind this simulation is to compare the effects of *Word Activation* and *Phrase Activation* to the observed effects for *Word Length, Word Frequency, Phrase Frequency* and *Relative Entropy*. Analogous to our analyses for the lexical predictors, we therefore fitted separate GAMs with tensor product smooths (with restricted cubic spline basis functions) for time by *Word Activation* and time by *Phrase Activation* on the residuals of the stratum one GAMs described in the Analysis section of this paper. No activations were removed prior to analysis. As before, we adopted a Bonferroni-corrected significance level of 0.0004 for all activation predictor GAMs.

## Simulation Results

In this section, we will present the simulation results for the observed effects of *Word Length, Word Frequency, Phrase Freqency* and *Relative Entropy* at representative example electrodes. Simulation results for the word level predictors *Word Length, Word Frequency* and *Relative Entropy* are the outcome of the tensor product smooths for time by *Word Activation*, whereas the simulation results for *Phrase Frequency* are the outcome of the tensor product smooths for time by *Phrase Activation*.

*Word Length*

Figure 8 shows the contour plot of the tensor surface for *Word Activation*. Selected example electrodes are electrodes at a maximum horizontal, vertical or diagonal distance of two electrodes from the example electrodes for the observed effect of *Word Length* in Figure 4. As before, the x-axis shows time in milliseconds after picture onset, with the four panels showing the development of the effect over the four epochs (0 to 300 ms, 200 to 500 ms, 400 to 700 ms and 600 to 900 ms). *Word Activation* is on the y-axis. Following the description of the observed effects, the z-axis shows voltages, with warmer colors representing higher voltages. Above each panel, the Bayesian *p* value for the effect of *Word Activation* at the depicted electrode is given, with significant *p*-values at the Bonferroni-corrected alpha level displayed in red.

If the NDR model correctly captures the observed effects, the ERP signature effect of *Word Activation* should be a superposition of the ERP signatures for *Word Length*, *Word Frequency* and *Relative Entropy*. The effect of *Word Activation* should therefore be significant whenever the combination of the effects of *Word Length*, *Word Frequency* or *Relative Entropy* is significant. As a result, the topographical distribution of the *Word Activation* effect is not directly comparable to that of the individual lexical predictors. We therefore omitted the picture inset showing the topographical distribution of the *Word Activation* effect from Figure 6.

*[INSERT FIGURE 8 AROUND HERE]*

The first panel of Figure 8 shows an oscillatory effect for high values of *Word Activation* in the first epoch. This effect is most prominent in parietal-occipital areas. Given the inverse transform that we applied to the activations, *Word Activation* correlates positively with *Word Length* ($r = 0.29$, $p < 0.001$). The oscillations

for high values of *Word Activation* therefore correspond to the oscillations for high values of *Word Length* in Figure 2. With peak spectral intensity at 6 Hz, the frequency of the oscillations observed here is slightly lower than that of the observed 7 Hz oscillations for *Word Length*.

The effect of *Word Activation* is first significant at 93 ms after stimulus onset. This is 17 ms earlier than the observed effect of *Word Length*, which was first significant at 110 ms. This difference likely arises as a result of the fact that we are determining the temporal onset of oscillatory effects. The slightly reduced frequency of the oscillations for *Word Activation* as compared to *Word Length* in the first 150 ms after picture onset leads to a phase shift, with maximum positive amplitudes for *Word Activation* preceding maximum positive amplitudes for *Word Length*. As a result, the effect of *Word Activation* reaches significance a bit earlier than the effect of *Word Length*.

As for the observed effect of *Word Length,* the oscillations for high values of *Word Activation* continue in the second epoch in left-lateralized central-parietal areas. Whereas the observed effect of *Word Length* was significant until 446 ms after picture onset, the oscillations for *Word Activation* are last significant at 383 ms. Consistent with the results for *Word Length*, the effect of *Word Activation* is no longer significant at left-lateralized frontal-central electrodes in the third epoch and at parietal-occipital electrodes in the fourth epoch.

Overall, the word level NDR activations successfully capture the observed effect of *Word Length*. The contour plots in Figure 4 and Figure 8 show a similar qualitative pattern, with 5-7 Hz oscillations for high predictor values in the first and second epoch and no effects in the third and fourth epoch. To quantify the similarity of the observed and simulated effects of *Word Length*, we calculated the correlation between the tensor surfaces displayed in Figure 2 and Figure 6. At $r = 0.29$ this correlation was highly significant ($p < 0.001$).

*Word Frequency*

Figure 9 shows the simulation of the *Word Frequency* effect, as depicted at representative example electrodes for *Word Activation*. Given the inverse transform that we applied to the activations, *Word Activation* correlates negatively with *Word Frequency* ($r = 0.39$, $p < 0.001$). To simplify the comparison of the simulated effects with the observed effects, we therefore flipped the y-axis in Figure 9, which now has high values for *Word Activation* at the bottom and low values for *Word Activation* at the top.

*[INSERT FIGURE 9 AROUND HERE]*

In the first epoch, oscillations arise for both high and low frequency words. These oscillations reach peak spectral intensity at 7 Hz. As such, these oscillations have a somewhat higher frequency than the 5 Hz oscillations in the first epoch for *Word Frequency* in Figure 5. In addition, the amplitude of the oscillations is greater for low frequency words than for high frequency words, while the observed effect showed maximum amplitudes for high frequency words. Furthermore, the effect reaches significance a bit later than the observed effect of *Word Frequency*: whereas the oscillatory effect of *Word Activation* is first significant at 107 ms after picture onset[4], the observed effect of *Word Frequency* was first significant 11 ms earlier, at 96 ms after picture onset. Despite these differences, the overall ERP signature of the simulated effect in the first epoch is highly similar to that of the early observed effect of *Word Frequency*: theta range oscillations for both high and low frequency words at left-lateralized frontal-central electrodes, with identical phases for the observed and simulated oscillations.

The oscillations for both high and low frequency words continue into the second epoch, where they remain more pronounced (i.e.; characterized by higher amplitudes) for low frequency words. At 7 Hz, the frequency of these oscillations now matches the frequency of the 7 Hz oscillations in the second epoch for the observed effect of *Word Frequency*. The effect fades out in the second half of the epoch and is last significant at 383 ms after picture onset. As such, the temporal offset of the simulated *Word Frequency* effect closely resembles that of the observed *Word Frequency* effect, which was last significant at 379 ms.

The observed effect of *Word Frequency* was characterized by a late re-emergence of oscillatory activity, with low-frequency oscillations for low frequency words at the end of the third and throughout the fourth epoch. This effect was only marginally significant in the third epoch, but highly significant in the fourth. Here, we see a similar pattern of results. While no effect is present in the third epoch, highly significant 4 Hz oscillations characterize the bottom half of the fourth panel of Figure 9. These oscillations have the same phase and frequency as the oscillations for low-frequency words in the fourth panel of Figure 5. The late oscillations for *Word Activation* are first significant at 604 ms after picture onset and remain significant throughout the fourth

epoch. As such, their time-course is comparable to the late observed effect of *Word Frequency,* which started at 584 ms after picture onset and remained significant until 900 ms after picture onset.

In summary, the NDR activations successfully capture the complicated pattern of results observed for *Word Frequency*. The model correctly simulates the early 5-7 Hz oscillations for high and low frequency words at left-lateralized frontal-central electrodes, as well as the late 4 Hz oscillations for low frequency words in the same areas. The phase of the simulated oscillations was highly similar to the phase of the observed oscillations for both the early and late effect of *Word Frequency*. The successful simulation of the *Word Frequency* effect is confirmed by a significant correlation ($r = 0.208, p < 0.001$) between the tensor surfaces for the observed effect of *Word Frequency* and the effect of *Word Activation*.

*Phrase Frequency*

Figure 10 presents the effect of *Phrase Frequency* as simulated by the *Phrase Activation* measure from the NDR simulation. As for *Word Frequency*, we flipped the y-axis in Figure 10 to allow for an easy comparison of the simulated effect with the observed effect for *Phrase Frequency* in Figure 6.

*[INSERT FIGURE 10 AROUND HERE]*

In the first panel of Figure 10, we see a deviation from the observed effect of *Phrase Frequency*, which was characterized by early negativities for both high and low frequency phrases. For *Phrase Activation*, no such negativities are present. Instead, see a transient early negativity for low predictor values in the top left of panel 1 of Figure 10. This effect arises earlier than any effect we have seen so far, either in the observed data or in the NDR simulation. In addition, the effect is topographically inconsistent: a similar transient early negativity is present only at one other electrode (P7). This suggests that the early negativity for low predictor values seen here may be a statistical fluke. We will therefore not discuss this effect in further detail. In addition to the early transient negativity for low predictor values, we also see a hint of an early negativity for high predictor values (bottom right of panel 1). This negativity, however, does not reach significance in the first epoch.

In the second epoch the simulation results look highly similar to the observed effect of *Phrase*

*Frequency*, with persistent left-lateralized negativities at frontal, central and parietal electrodes for both high and low predictor values. The negativity for low predictor values (top of panel 2 of Figure 10) reaches significance at 267 ms after picture onset, whereas the negativity for high predictor values (bottom of panel 2) is first significant at 302 ms. The effect size of the negativity for high values of *Phrase Activation* is somewhat greater than the effect size for the observed effect of *Phrase Frequency.* Given that there is only one item for which *Phrase Activation* is greater than 0.30, however, the effect of *Phrase Activation* for high predictor values might be somewhat overrepresented in Figure 10. In addition, a more transient positivity for low to medium values of *Phrase Activation* emerges in the second half of the second epoch. This positivity is similar to the positivity observed for medium to high frequency phrases in the observed effect for *Phrase Frequency*. As for the observed effect, this transient positivity continues in the first part of the third epoch.

Consistent with the observed effect of *Phrase Frequency*, the negativities for extreme predictor values continue in the third epoch at left frontal, central and parietal electrodes. For high values of *Phrase Activation* these negativities continue throughout the epoch (bottom half of panel 3 of Figure 10), but for low values of *Phrase Activation* they are last significant at 611 ms after picture onset. This closely resembles the pattern of results for the observed effect of *Phrase Frequency*, where negativities for low frequency phrases continued throughout the third epoch, but negativities for high frequency phrases were last significant at 591 ms after picture onset.

The negativities for high values of *Phrase Frequency* at left frontal, central and parietal electrodes remain significant throughout the fourth epoch. As such, the simulated effect for low frequency phrases remains significant substantially longer than the observed effect, which was last significant at 768 ms. Furthermore, we see a continuation of the positivity for low to medium values of *Phrase Activation*. This positivity was not present in the fourth epoch for the observed effect of *Phrase Frequency*. The positivity for low to medium values of *Phrase Frequency* in the second half of the fourth epoch is reflected in the subtle positivities for medium to high values of *Phrase Activation* in the bottom half of the fourth panel of Figure 8. Although the effect size of these positivities is limited, this effect reaches significance from 600 to 828 ms.

The observed effect of *Phrase Frequency* is quite complicated, with widespread left-lateralized persistent negativities for high and low frequency phrases and more transient positivities for intermediate values

of phrase frequency. Although the NDR model does not pick up the negativities for extreme predictor values in the first epoch, the overall nature of the *Phrase Frequency* effect is successfully captured by the NDR phrase activations. The successful replication of the *Phrase Frequency* effect in the NDR simulation is confirmed by a high correlation between the tensor surfaces for the observed and simulated effects ($r = 0.539$, $p < 0.001$).

*Relative Entropy*

Figure 11 presents the simulation of the observed effect for *Relative Entropy* using the NDR word activations. *Word Activation* correlates positively with *Relative Entropy* ($r = 0.15$, $p = 0.025$). High values for *Word Activation* therefore correspond to high values for *Relative Entropy* and, as such, represent words with atypical prepositional phrase frequency distributions.

*[INSERT FIGURE 11 AROUND HERE]*

The first panel of Figure 11 shows 7 Hz oscillations for both high and low values of *Word Activation*. These oscillations are most prominent at central and right-lateralized parietal-occipital electrodes. The oscillations for high values of *Word Activation* are similar to those for high values of *Relative Entropy*, with a similar frequency and phase, but a somewhat increased amplitude. The temporal onset of both effects is similar as well, with the oscillations for high predictor values first being significant at 107 ms for *Word Activation* and at 108 ms for *Relative Entropy*. At the bottom end of the predictor range, we see oscillations for low and medium values of *Word Activation* that correspond in frequency and phase to the oscillations observed for medium values of *Relative Entropy*. The extra row of oscillations for words with low *Relative Entropy*, however, is not captured by the NDR activations.

In the second epoch the 7 Hz oscillations for high values of *Word Activation* continue at right-lateralized parietal-occipital electrodes. These correlations correspond temporally and topographically to the oscillations seen for high values of *Relative Entropy*. In addition, the NDR simulation captures the oscillations seen for words with both medium and low values of *Relative Entropy*. While the effect of *Relative Entropy* was last significant at 381 ms after picture onset, the significant effect of *Word Activation* lasts 38 ms longer: it is last

significant at 419 ms.

In the third and fourth epoch, the effect for high values of *Word Activation* re-emerges. This late effect of *Word Activation* corresponds to the late effect for words with high *Relative Entropy*. Like the observed effect of *Relative Entropy*, the late effect of *Word Activation* is characterized by low-frequency 4 Hz oscillations for high predictor values at left-lateralized central and frontal electrodes. While the onset of the observed *Relative Entropy* effect was 584 ms, the onset of the late effect in the NDR simulation is 608 ms. The offset of the observed effect of *Relative Entropy* is somewhat later than that of the simulated effect as well: 831 ms versus 900 ms.

There is one striking difference between the late effect observed for *Relative Entropy* and the late effect for *Word Activation*. The observed effect is characterized by a single row of oscillations for all high values of *Relative Entropy*. The simulated effect, by contrast, shows two rows of oscillations. The oscillations for the highest values of *Word Activation* correspond in phase and frequency to those observed for *Relative Entropy*. The second row of oscillations has the same frequency, but is opposite in phase. This discrepancy between the observed and simulated effects is a result of the fact that the NDR model simultaneously captures the observed effects of *Word Length, Word Frequency* and *Relative Entropy*. As a result, simulation contour plots are a superposition of the simulated effects of *Word Length, Word Frequency* and *Relative Entropy* (as well as any other effects that might arise in the NDR model, but are not of interest in the current simulation).

For all simulated effects reported so far, we were able to select example electrodes for which the simulation results (almost) exclusively correspond to the results of a single predictor. The late effects of *Word Frequency* and *Relative Entropy*, however, coincide both temporally and topographically. The fourth panel of Figure 11, therefore, is a superposition of the late 4 Hz oscillations for low values *Word Frequency* (the vertical mirror image of panel 4 of Figure 5) and high values of *Relative Entropy* (panel 4 of Figure 7). As such, the extra row of oscillations seen in panel 4 of Figure 11 is not an incorrect simulation of the *Relative Entropy* effect, but the correct simultaneous simulation of the late *Relative Entropy* and *Word Frequency* effects. In addition, the superposition of the phase-synchronized 4 Hz oscillations tied to low values of *Word Frequency* and high values of *Relative Entropy* effects explains the increased amplitude of the oscillations for the highest values of *Word Activation*.

Finally, more subtle oscillations for low values of *Word Activation* arise in the fourth epoch. These oscillations correspond to the subtle oscillations seen for low values of *Relative Entropy*, but reach peak amplitudes somewhat later in time. While the subtle oscillations for *Relative Entropy* failed to reach significance, the oscillations for low values of *Word Activation* observed here briefly reach significance in the last 7 ms of the fourth epoch. Given the limited reliability of GAMs near the edges of the analysis windows, however, it is unclear how robust this effect of *Word Activation* is.

In summary, the NDR model successfully replicates the pattern of results for *Relative Entropy*, with 7 Hz oscillations across the predictor range at parietal-occipital locations in the first two epochs and a late re-emergence of the effect for words with high *Relative Entropy* in the form of slower 4 Hz oscillations in left-lateralized frontal-central areas. This is confirmed by a highly significant correlation of the tensor surfaces for the observed and simulated effects ($r = 0.421$, $p < 0.001$)

## Discussion

The simulation results demonstrate that the predictive power of the discriminative learning approach as implemented in the NDR model extends beyond the realm of chronometric studies. The effects of *Word Length, Word Frequency* and *Relative Entropy* were characterized by theta range oscillations with different temporal dynamics, amplitudes and phases across the predictor dimension. By contrast, we observed persistent negativities for extreme values of *Phrase Frequency*. While all predictor effects were most prominent in the left hemisphere, each predictor effect showed a unique topographical development over time. The NDR model successfully captures the observed non-linear predictor effects and their temporal and spatial dynamics.

Three aspects of these simulation results are of particular theoretical interest. First, the NDR model successfully replicates the qualitative differences between the effects for *Word Frequency* and *Phrase Frequency*. Previously, Baayen et al. (2013a) demonstrated that the NDR model correctly captures phrase frequency effects in chronometric studies. The current results extend the findings by Baayen et al. (2013a) by showing that the NDR model not only simulates the existence of a phrase frequency effect, but also correctly predicts how the

non-linear temporal and spatial dynamics of this effect differ from those for the effect of *Word Frequency*.

*Word Frequency* effects in the NDR model are directly reflected in the word level activations, which are a measure of the bottom-up support for the meaning of the target noun given the prepositional phrase input. These same word level activations enter the phrase level activation measure, which leads to a strong correlation between *Word Activation* and *Phrase Activation* ($r = 0.72$, $p < 0.001$, before residualization of *Phrase Activation*: $r = 0.65$, $p < 0.001$). Crucially, however, it is the integration of the activations of the target noun with the activations of the preposition and the definite article that enables the NDR model to successfully replicate the qualitative differences between the *Word Frequency* and *Phrase Frequency* effects.

In the current NDL simulation the activations of the preposition, definite article and noun equally contribute to the simulated phrase activations. While this approach allowed us to simulate the observed effects in a parameter-free model, it is possible that a weight parameter for the contribution of each word to the phrase activations would help further improve the performance of the model. Given the primed picture naming paradigm used in the current study, preposition and definite article activations may contribute to phrase activations to a different degree than target noun activations. In addition, the relative contribution of preposition and definite article activations and target noun activations may vary over time as a consequence of the 2000 ms time lag between the onset of the presentation of the prime and the onset of the presentation of the target. A time-sensitive weight parameter might therefore provide the NDR model with an opportunity to capture the negativities for *Phrase Frequency* in the first epoch.

The second aspect of this simulation that is of particular theoretical interest concerns the *Relative Entropy* measure. While *Relative Entropy* effects pose a challenge to exemplar-based models, they fit well with the architecture of the NDR model. The computational engine of the NDR is a discrimination learning algorithm that learns to associate orthographic input units with semantic outcomes on the basis of the distributional properties of the linguistic input space. Prepositional relative entropy is a constructional measure that taps into a subset of these distributional properties. If discrimination learning is an adequate description of how we become sensitive to the distributional properties of a language, we would therefore expect the NDR model to replicate the observed effects of *Relative Entropy*. Baayen et al. (2011) showed that the NDR model correctly predicts the chronometric effect of prepositional relative entropy. The current simulation results demonstrate that the NDR

model also captures the complex non-linearities in the time, predictor and topographical dimensions that characterize the ERP signature of the *Relative Entropy* effect in the current primed picture naming study.

A final interesting aspect of the current simulations is a comparison between the model fits for the lexical predictors and the model fits for the NDR activations as alternative predictors. In a post-hoc analysis, we therefore compared the performance of multiple regression GAMs with tensor product smooths of time by *Word Length,* time by *Word Frequency,* time by *Phrase Frequency* and time by *Relative Entropy* to the performance of multiple regression GAMs with tensor product smooths of time by *Word Activation* and time by *Phrase Activation* for all 4 epochs at all 32 electrodes. All models were fitted on the residuals of the Stratum 1 models described in the Analysis section of this paper.

Given the fact that the contribution of linguistic predictors to the ERP signal is limited, r-squared values were very small for both the lexical predictor GAMs and the NDR activation GAMs. On average, the lexical predictor GAMs (average r-squared: 0.00027) had somewhat higher r-squared values than the lexical predictor GAMs (average r-squared: 0.00014). The AIC scores of the NDR models (average AIC score: 4263390), however, were significantly lower ($t = 10.23, p < 0.001$) than the AIC scores of the predictor models (average AIC score: 4522202). For all 128 epoch-electrode combinations the AIC score of the NDR model was lower than that of the corresponding lexical predictor model. The lower AIC scores for the NDR models suggest that the NDR activations provide a better account of the ERP signal than the lexical predictors. We will return to this issue shortly.

## General Discussion

The first half of this paper presents the results of a primed picture naming study on prepositional phrase processing. In this experiment participants were presented with preposition plus definite article primes (e.g.; “on the”) followed by target photographs depicting concrete nouns (e.g.; “strawberry”). Participants were asked to name the target noun as fast and accurately as possible. We measured the ERP signal after picture onset and

analyzed the correlates of four linguistic predictors in this signal using generalized additive models.

At the word level we observed theta range (4-7 Hz) oscillations in the left hemisphere tied to the length, frequency and prepositional relative entropy of the target word. Theta range oscillations are thought to reflect (working) memory demands in language processing and have previously been observed in a variety of language processing tasks, including lexico-semantic retrieval, syntactic processing and translation (see, e.g.; Bastiaansen et al. (2005); Bastiaansen et al. (2008); Grabner et al. (2007)). The oscillatory activity for all word level predictors arose around 100 ms after picture onset. The early onset of the effects for the word level predictors is in line with previous studies that established the onset of word length (Hauk et al., 2006) and word frequency effects (Hauk et al., 2006; Sereno et al., 1998) around the 100 ms mark. The qualitative, temporal and topographical similarity between the effects of the word level predictors is an interesting issue that we will return to shortly.

Of the word level effects, the effect of relative entropy is of particular theoretical interest. Previously, relative entropy effects had only been observed in reaction time studies (see e.g.; Milin et al., 2009a, Milin et al., 2009b; Kuperman et al., 2010, Baayen et al, 2011). The current study is the first to document a relative entropy effect in an ERP study. The effect of relative entropy suggests that the language processing system is sensitive to the distributional properties of a noun's prepositional paradigm as compared to the prepositional frequency distribution in the language as a whole. As such, the effect of relative entropy observed here poses a challenge to exemplar-based approaches to language processing, such as data-oriented parsing (Bod, 2006) or memory-based learning (Daelemans & Bosch, 2005). To account for relative entropy effects exemplar-based models would have to assume that frequency information about prepositional phrases and the prepositional phrase prototype is available during processing and that the distance between a noun's prepositional phrase frequency distribution and the prototypical prepositional phrase frequency distribution is computed online.

At the phrase level, we observed an effect of phrase frequency that was qualitatively different from the effects of the word level predictors. Persistent negativities arose for prepositional phrases with atypical frequencies in the first 300 ms epoch and continued throughout our 900 ms analysis window. More transient positivities were present for phrases with more typical frequencies. As for the effect of relative entropy, the effect of phrase frequency is well-documented in chronometric studies (see e.g.; Arnon & Snider (2010); Bannard &

Matthews, 2008; Shaoul et al., 2009, Tremblay et al., 2009; Tremblay & Baayen, 2010; Siyanova-Chanturia et al., 2011; Baayen et al., 2011). Recently, Tremblay & Baayen (2010) documented a phrase frequency effect for 4-word sequences in a free recall task. The current study adds to these findings by showing a phrase frequency effect in a primed picture naming paradigm. The effect of relative entropy was qualitatively similar to that of the other word level predictors. By contrast, the effect of phrase frequency differs substantially from the other observed effects, including the effect of word frequency.

The phrase frequency effect in chronometric studies has been interpreted as evidence for the existence of phrasal representations. Bannard & Matthews (2008), for instance, suggest that their finding that young children process frequency phrases faster than infrequent phrases indicates the existence of representations at different levels of granularity. This fits well with exemplar-based models of language processing, in which phrase representations can be stored in the same way word representations are stored. The current pattern of results, however, does not straightforwardly support an interpretation of phrase frequency effects in terms of phrasal representations: if word representations and phrase representations are stored and accessed in the same way we would expect the effects of word frequency and phrase frequency to be highly similar.

Discrimination learning offers an alternative to exemplar-based approaches to language processing. In the Naive Discriminative Reader (NDR) model (Baayen et al., 2011) no representations beyond the simple word level exist. Nonetheless, the NDR successfully replicates the chronometric effects of relative entropy (Baayen et al., 2011) and phrase frequency (Baayen et al., 2013a). The second part of this paper presents a simulation study in which we demonstrate that the NDR model also captures the complex non-linearities that characterize the qualitative, temporal and topographical dynamics of the effects of word length, word frequency, phrase frequency and relative entropy in the ERP data.

The discriminative learning algorithm that forms the computational core of the NDR model learns associations between orthographic input units and semantic outcomes on the basis of the distributional properties of the language input space. The relative entropy measure gauges a subset of these distributional properties by comparing the prepositional phrase frequency distribution for a given noun against the constructional prototype. The similarities between the observed effect of relative entropy and the effect of relative entropy as simulated in the NDR model suggest that the human language processing system is at the very least sensitive to those

distributional properties of the linguistic input space that are captured by the relative entropy measure.

Importantly, the effect of relative entropy in the NDR model is a side-effect of the basic process of learning a language. No representations beyond the simple word level or frequency counters in the head have to be assumed to account for the relative entropy effect. The NDR model therefore offers a parsimonious account of relative entropy effects that is grounded in well-established principles of human learning (Wagner & Rescorla, 1972; Miller et al., 1995; Siegel & Allan, 1996; Chater et al., 2006) that have recently proved insightful for child language acquisition (Ramscar et al., 2010; Hsu et al., 2010) and second language learning (Ellis, 2006).

As noted above, the effects of the word level predictors word length, word frequency and relative entropy are qualitatively, temporally and topographically remarkably similar. The NDR model replicates these similarities by correctly predicting left-lateralized theta range oscillations that arise around 100 ms after picture onset for all three predictors. The similarity of the word length, word frequency and relative entropy effects suggests that a single processing mechanism underlies all three effects. From a discriminative learning perspective this makes sense. The NDR model has no explicit representations for the distributional properties of a word's prepositional paradigm, nor does it explicitly encode a word's length or frequency. Instead, the effects of these three predictors arise as a straightforward consequence of linguistic discrimination learning.

An interpretation of the current results in terms of discriminative learning is supported by the lower AIC scores for the NDL multiple regression models as compared to the lexical predictor multiple regression models, which suggest that discriminative learning offers a superior explanation of the ERP data as compared to standard lexical predictors. Although the similarity of the word level predictors is unsurprising from a discriminative learning point of view, the lower AIC scores for the NDL multiple regression models remind us of an important fact about psycholinguistic research: lexical predictors are descriptive level abstractions from the underlying language processing system. While lexical predictors describe the behavioral correlates of (properties of) the language processing system, they do not necessarily provide insight into the processing system itself. One of the consequences of this is that the presence of a behavioral effect for a lexical predictor therefore does not imply the existence of corresponding representations. Bearing Ockham's razor in mind, quite the opposite is true: if a model is able to account for the effect of a lexical predictor *without* assuming dedicated representations tied to that predictor, this model should be preferred above a model that requires additional representations to explain

an effect.

Ockham's lessons resonate in the correct simulation of the differences between the word frequency and phrase frequency effects in the NDR model. As mentioned earlier, phrase frequency effects have been interpreted as evidence for the existence of phrase level representations, be they discrete or distributed. In the NDR model phrase frequency effects emerge from a simple integration over word level activations. A high frequency phrase such as "all over the place" is read faster than a low frequency phrases such as "all over the city", because the letters and letter combinations in "all over the place" have become more associated with the meanings ALL, OVER, THE and PLACE than the letters and letter combinations in "all over the city" have become associated with the meanings ALL, OVER, THE and CITY during the learning process. The simulation of the phrase frequency effect in the NDR model therefore demonstrates that positing phrase level representations to explain the phrase frequency effect is unnecessary.

The NDR activations provide a systematic estimate of the language processing system that gauges the learnability of lexical items given the properties of the linguistic input space. The simulation results suggest that the learnability of target words and phrases has improved predictive power for the ERP signal in a primed picture naming paradigm over a set of standard lexical predictors, while avoiding the multi-collinearity issues associated with these predictors. It is important to note, however, that the NDR model itself is an abstractive level description that tells us little about the neuro-biological implementation of the discriminative learning mechanism it posits. The discrete representations in the NDR model do not do justice to the complex architectural and topographical neuro-biological reality of neural networks. Nonetheless, the current simulations demonstrate that discrimination learning can help us provide more insight into the behavioral effects of lexical predictors and further our understanding of the language processing system. When trying to understand the complex dynamic system that language is, there is no harm in starting small.

## References


Arnon, I., & Snider, N. (2010). Syntactic probabilities affect pronunciation variation in spontaneous speech. *Journal of Memory and Language, 62*, 67-82.

Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. Proceedings of the 31st annual meeting of the cognitive science society, 2112-2117.

Baayen, R. H., Dijkstra, T., & Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual route model. *Journal of Memory and Language, 36*, 94-117.

Baayen, R. H. (2010). Demythologizing the word frequency effect: a discriminative learning perspective. Mental Lexicon 5, 436-461.

Baayen, R. H., Milin, P., Filipović Durđević, D., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review, 118 (3)*, 438-482.

Baayen, R. H., Hendrix, P., & Ramscar, M. (2013a). Sidestepping the combinatorial explosion: An Explanation of n-gram Frequency Effects based on Naive Discriminative Learning. Language and Speech, *56 (3)*.

Baayen, R. H., Tremblay, A., & Hendrix, P. (2013b). An introduction to analyzing the ERP signal with generalized additive modeling using the gam-erp package, vignette for the GAM-eRp package for R. *Vignette and package in preparation*.

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning: the effect of familiarity on children's repetition of four-word combinations. *Psychological Science, 19*, 241-248.

Bastiaansen, M., Berkum, J. v., & Hagoort, P. (2002). Syntactic processing modulates the theta rhythm of the human eeg. *NeuroImage, 17 (3)*, 1479-1492.

Bastiaansen, M., & Hagoort, P. (2003). Event-induced theta-responses as a window on the dynamics of memory. *Cortex , 39 (4-5)*, 967-992.

Bastiaansen, M., Van Der Linden, M., Ter Keurs, M., Dijkstra, T., & Hagoort, P. (2005). Theta responses are involved in lexical-semantic retrieval during language processing. *Journal of Cognitive Neuroscience, 17 (9)*, 530-541.

Bastiaansen, M., Oostenveld, R., Jensen, O., & Hagoort, P. (2008). I see what you mean: theta power increases are involved in the retrieval of lexical semantic information. *Brain and language, 106 (1),* 15-28.

Bates, E., D'Amico, S., Jacobsen, T., Szekely, A., Andonova, E., Devescovi, A., et al. (2003). Timed picture naming in seven languages. *Psychonomic Bulletin and Review*, 344-380.

Bertram, R., Schreuder, R., & Baayen, R. H. (2000). The balance of storage and computation in morphological processing: the role of word formation type, affixal homonymy, and productivity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 419-511.

Bod, R. (2006). Exemplar-based syntax: How to get productivity from examples. The Linguistic Review, 23, 291-320.

Brants, T., & Franz, A. (2006). *Web 1t 5-gram version 1*. Philadelphia: Linguistic Data Consortium.

Burnard, L. (1995). *Users guide for the British National Corpus.* Oxford university computing service: British National Corpus consortium.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. Trends in Cognitive Science, 10 (7), 287-291.

Dabrowska, E. (2000). From formula to schema: the acquisition of English questions. Cognitive Linguistics, 11 (1/2), 83-102.

Daelemans, W., Bosch, A. Van den, & Weijters, A. (1997). IGTree: Using trees for compression and classification in lazy learning algorithms. Artificial Intelligence Review, 11, 407-423.

Daelemans, W., & Bosch, A. Van den. (2005). Memory-based language processing. Cambridge: Cambridge University Press.

Dealemans, W., Zavrel, J., Sloot, K. & Bosch, A. Van den. (2007). TiMBL: Tilburg Memory Based Learner Reference Guide. Version 6.1 (Technical Report No. ILK 07-07). Computational Linguistics Tilburg University.

Danks, D. (2003). Equilibria of the Rescorla-Wagner model. Journal of Mathematical Psychology, 47 (2), 109-121.

Ellis, N. C. (2006). Language acquisition as rational contingency learning. Applied Linguistic, 27 (1), 1-24.

Grabner, R. H., Brunner, C., Leeb, R., Neuper, C., & Pfurtscheller, G. (2007). Event-related eeg theta and alpha band oscillatory responses during language translation. *Brain research bulletin, 72 (1)*, 57-65.

Haskell, T. R., Thornton, R., & MacDonald, M. C. (2010). Experience and grammatical agreement: Statistical learning shapes number agreement production. *Cognition, 114 (2)*, 151-164.

Hastie, T., & Tibshirani, R. (1986). Generalized additive models (with discussion). *Statistical Science, 1(3)*, 297-318.

Hauk, O., Davis, M., Ford, M., Pulverműller, F., & Marslen-Wilson, W. (2006). The time course of visual word recognition as revealed by linear regression analysis of ERP data. *NeuroImage, 30*, 1383-1400.

Hendrix, P., & Ramscar, M. & Baayen, R. H. (2013). NDRa: a single route model of reading aloud based on discriminative learning. *Manuscript submitted for publication.*

Hillyard, S., & Picton, T. (1987). Electrophysiology of cognition. *Handbook of physiology, 5*, 519-584.

Hsu, A. S., Chater, N., & Vitanyi, P. (2011). The probabilistic analysis of language acquisition: Theoretical, computational, and experimental analysis. Cognition, 120 (3), 380-390.

Jescheniak, J. D., & Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory and Cognition, 20 (4)*, 824-843.

Kryuchkova, T., Tucker, B. V., Wurm, L., & Baayen, R. H. (2011). Danger and usefulness in auditory lexical processing: evidence from electroencephalography. *Brain and Language*, 122 (2), 81-91.

Kuperman, V., Bertram, R., & Baayen, R. H. (2008). Morphological dynamics in compound processing. *Language and Cognitive Processes, 23*, 1089-1132.

Kuperman, V., Bertram, R., & Baayen, R. H. (2010). Processing trade-offs in the reading of Dutch derived words. *Journal of Memory and Language, 62*, 83-97.

McClelland, J. L. & Rumelhart, D. E. (1981). An interactive activation model of context effect in letter perception: Part i: an account of the basic findings. Psychological Review, 88, 375-407.

Milin, P., Filipović Durđević, D., & Moscoso del Prado Martin, F. (2009a). The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from serbian. *Journal of Memory and Language*, 50-64.

Milin, P., Kuperman, V., Kostic, A., & Baayen, R. H. (2009b). Paradigms bit by bit: an information-theoretic approach to the processing of paradigmatic structure in inflection and derivation. In J. P. Blevins & J. Blevins (Eds.), *Analogy in grammar: form and acquisition* (pp. 214-252). Oxford: Oxford University Press.

Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the rescorla-wagner model. Psychological Bulletin, 117 (3), 363-386.

Oldfield, R. C., & Wingfield, A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology, 17*, 273-281.

Norris, D. G. (1994). Shortlist: A connectionist model of continuous speech recognition. Cognition, 52, 189-234.

Norris, D. G. & McQueen, J. (2008). Shortlist B: A Bayesian model of continuous speech recognition. Psychological Review, 115 (2), 357-395.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. Cognitive Science, 34 (6), 909-957.

Sereno, S. C., Rayner, K., & Posner, M. (1998). Establishing a time-line of word recognition: evidence from eye movements and event-related potentials. *Neuroreport, 9 (10)*, 2195-2200.

Shaoul, C., Westbury, C., & Baayen, R. H. (2009). *Agreeing with Google: We are sensitive to the relative orthographic frequency of phrases.* Poster presented at Psychonomics 2009.

Siegel, S., & Allan, L. G. (1996). The widespread influence of the rescorla-wagner model. Psychonomic Bulletin & Review, 3 (3), 314-321.

Siyanova-Chanturia, A., Conklin, K. & Van Heuven, W. (2011). Seeing a phrase 'time and again' matters: The role of phrasal frequency in the processing of multi-word sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37 (3),* 776-784.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. Psychological Review, 88, 135-170.

Taft, M. (1979). Recognition of affixed words and the word frequency effect. *Memory and Cognition, 7*, 263-272.

Taft, M. (1994). Interactive-activation as a framework for understanding morphological processing. *Language and Cognitive Processes, 9 (3)*, 271-294.

Taft, M., & Forster, K. I. (1976). Lexical storage and retrieval of polymorphemic and polysyllabic words. *Journal of Verbal Learning and Verbal Behavior, 15*, 607-620.

Tomasello, M. (2003). Constructing a language: A usage-based theory of language acquisition. Cambridge, Mass.: Harvard University Press.

Tremblay, A. (2010). Independent components analysis (ica) based eye-movement correction [Computer software manual]. (R package version 1.2)

Tremblay, A., & Baayen, R. H. (2010). Holistic processing of regular four-word sequences: A behavioral and erp study of the effects of structure, frequency, and probability on immediate free recall. In D. Wood (Ed.), *Perspectives on formulaic language: Acquisition and communication* (pp. 151-173). London: The Continuum International Publishing Group.

Tremblay, A., Baayen, R. H., Derwing, B., Libben, G., Tucker, B., & Westbury, C. (2011). *Empirical evidence for an inationist lexicon.* Poster presented at LSA Annual Meeting 2011.

Van Gompel, R. P., & Pickering, M. J. (2007). The oxford handbook of psycholinguistics. In M. G. Gaskell (Ed.), (p. 289-307). Oxford: Oxford University Press.

Wagner, A., & Rescorla, R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning ii* (pp. 64-99). New York: Appleton-Century-Crofts.

Wood, S. N. (2006). *Generalized additive models*. New York: Chapman & Hall/CRC.

## Appendix A

We used generalized additive models (GAMs) to analyze the ERP data for the current experiment (Hastie & Tibshirani (1986); Wood (2006), R package *MGCV* (version 1.6-2)). Unlike traditional ERP analysis techniques, GAMs allowed us to investigate the non-linear effects of numerical predictors as they evolve over time in the ERP signal. By contrast, traditional ERP analysis typically operate on the basis of dichotomized versions of numerical predictors such as word frequency, phrase frequency or relative entropy. The average curves for the dichotomized predictors values are then compared in by-item or by-subject analyses (i.e.; low frequency versus high frequency). In this appendix we will compare the performance of GAMs to the performance of a traditional analysis method for both simulated data and for some of the key predictor effects documented in this paper. We will demonstrate that the patterns of results for both types of analyses converge in some cases, but that a traditional analysis results in a loss of information or dichotomization artifacts in other cases.

First, consider the simulated predictor effect in the top left panel of Figure 12. The effect is characterized by a two-dimensional sinusoid, with oscillations in both the time and the predictor dimension. White noise with a mean of 0 and a standard deviation of 0.5 was added to each simulated data point. The middle panel of the top row of Figure 12 shows the results of a GAM analysis on this simulated predictor effect. The two-dimensional sinusoid in the simulated data is replicated in the GAM analysis. The frequencies of the oscillations in both directions and the effect sizes match those in the simulated data. The top right panel of Figure 12 shows the results of a dichotomization of the predictor into low and high predictor values based on a split halfway the predictor range. No sinusoidal activity is seen for either high or low frequency words and no difference is observed between high and low frequency words at any point in time. Dichotomization of the predictor therefore entirely masks the two-dimensional oscillatory activity that is present in the simulated data.

*[INSERT FIGURE 12 AROUND HERE]*

The simulated data in the top left panel of Figure 12 were symmetrical with respect to the mid-point of the predictor range. For the bottom left panel of Figure 12 we shifted the effect upwards on the y-axis, such that the simulated predictor effect is no longer symmetrical with respect to the mid-point of the predictor range. The middle panel of the bottom row of Figure 12 demonstrates that this does not constitute a problem for GAMs. As before the two-dimensional sinusoid is replicated with the correct frequency in both dimensions and the correct effect size. The bottom right panel of Figure 12 shows what happens if the predictor is dichotomized into high and low predictor values with a split at the mid-point of the predictor range. Due to the vertical shift of the oscillations a traditional analysis now reflects some of the oscillatory activity in the simulated data. The observed differences between high and low predictor values, however, reflect the differences between medium and low predictor values in the simulated data. All information about the fact that high predictor values and low predictor values show a highly similar pattern of results is lost.

The problems of dichotomizing numerical predictors outlined above arise in the ERP data reported in this paper as well. In what follows, we will examine the performance of a traditional ERP analysis for the most typical effects of word frequency, phrase frequency and relative entropy effects in the current data. For each of these three predictors, we will compare the GAM analyses in this paper to a traditional analysis of the data for the same epoch at the same electrode.

The left panel of Figure 13 shows the effect of Word Frequency at electrode F3 in the 0 to 300 ms time window in the GAM analysis reported here. The effect is characterized by theta range oscillations for both high and low frequency words with opposite phases. The oscillations arise around 100 ms after picture onset, but are most prominent in the second half of the 0 to 300 ms time window.

*[INSERT FIGURE 13 AROUND HERE]*

The right panel of Figure 13 shows the results of a traditional analysis in which we dichotomized *Word Frequency* into high and low frequency words (split halfway the *Word Frequency* range). The grand mean curves for *Word Frequency* show a similar pattern of results as compared to the GAM analysis, with higher voltages for low frequency words from 180 to 260 ms after picture onset and higher voltages for high frequency words from

260 to 300 ms after picture onset. Both of these effects reach significance in the item-analysis, as indicated by the dark red ($\alpha = 0.05$) and bright red (Bonferroni-corrected alpha level; $\alpha = 0.0004$) squares at the bottom of the right panel of Figure 13.

The comparison of the GAM analysis and the traditional analysis for the *Word Frequency* effect demonstrates that the oscillatory effect of *Word Frequency* in the 0 to 300 ms time window is reflected in the grand means curves for high and low frequency words. Rather than being interpreted as theta range oscillations, however, this effect would likely be described in terms of ERP components in a traditional analysis – with an increased P200 and N300 for low frequency words.

The effect of *Word Frequency* in the GAM analysis is relatively simple in nature, with oscillations for high and low frequency words that are nicely separated with respect to the middle of the *Word Frequency* range and that have opposite phases. This is close to an ideal scenario for a traditional ERP analysis. The effect of *Relative Entropy,* however, is much more complicated in nature. The left panel of Figure 14 shows the effect of *Relative Entropy* at electrode PO3 in the 0 to 300 ms time window in the GAM analysis. We slightly adjusted the z-range of Figure 14 in comparison to the left panel of Figure 7 to reveal in more detail the complicated nature of the early effect of *Relative Entropy*.

*[INSERT FIGURE 14 AROUND HERE]*

As can be seen in the left panel of Figure 14, the early effect of *Relative Entropy* is characterized by oscillations in both the time and predictor dimension that arise at around 100 ms after picture onset. The oscillations are most prominent for extreme predictor values. In contrast to the effect of *Word Frequency* that we saw above, the oscillations for these extreme predictor values are in phase. In between the oscillations for extreme values of *Relative Entropy* are more subtle oscillations for low to medium values of *Relative Entropy* that are opposite in phase to the oscillations for the extreme predictor values.

The right panel of Figure 14 shows the effect of *Relative Entropy* at electrode PO3 in a traditional ERP analysis in which we dichotomized *Relative Entropy* into high and low relative entropy on the basis of a split halfway the *Relative Entropy* range (see black line in the left panel of Figure 14). In the grand mean curves for

high and low *Relative Entropy* we see an early positivity from 100 to 170 ms after picture onset for high values of *Relative Entropy*, followed by a negativity for words with high *Relative Entropy* in the 170 to 275 ms range. Both of these effects reach significance in both by-item and by-subject ANOVAs, although not always at Bonferroni-corrected alpha levels.

Interestingly, the positivities at the right edge of the left panel of Figure 14 are not reflected in the grand mean curves for high and low values of *Relative Entropy*. Although this might be related from the fact that item-, subject- and trial-related variance were not properly accounted for in the traditional analysis, an alternative explanation for this discrepancy is that GAMs tend to be somewhat less reliable near the edges. This is the reason we used a 100 ms overlap between subsequent time windows in the GAM analysis reported here to establish the consistency of the results for subsequent time windows.

The overall pattern of results in the right panel of Figure 14 is consistent with the results of the GAM analysis, with the difference between high and low values of *Relative Entropy* reflecting the facts that 1) the oscillations for high values of *Relative Entropy* cover a larger part of the predictor range than do the oscillations for low values of *Relative Entropy* and 2) the opposite-phase oscillations for low to medium values of *Relative Entropy* partly cancel out the oscillations for low values of *Relative Entropy* in the dichotomized *Relative Entropy* measure. As for the effect of *Word Frequency*, this demonstrates that the effects observed in the GAM analysis reflect properties of the ERP signal that are visible in the grand mean curves.

Whereas the qualitative nature of the effect of *Word Frequency* was accurately captured by a traditional ERP analysis, however, a lot of detail is lost about the effect of *Relative Entropy* through dichotomization. From the right panel of Figure 14, for instance, it would be impossible to tell that the effects for high and low values of *Relative Entropy* are in fact highly similar and that the differences in the grand mean curves for the dichotomized predictor are driven primarily by the opposite-phase theta range oscillations for low to medium values of *Relative Entropy*.

Theta range oscillation in the time dimension characterized he effects of *Word Frequency* and *Relative Entropy*. For *Phrase Frequency,* we observed effects that persisted over time. The left panel of Figure 15 shows the effect of *Phrase Frequency* at electrode CP5 in the 400 to 700 ms time window. For both high and low frequency words we see long-lasting negativities. For high frequency phrases these negativities fade out near the

end of the time window, whereas for low frequency phrases the negativities persist through the epoch. By contrast, the *Phrase Frequency* effects for phrases with intermediate frequencies are characterized by more transient positivities.

*[INSERT FIGURE 15 AROUND HERE]*

The right panel of Figure 15 shows the results of a traditional ERP analysis in which *Phrase Frequency* was dichotomized halfway the phrase frequency range (see black line in the left panel of Figure 15). This analysis shows a difference between high and low frequency words in the first half of the 400 to 700 ms time window. As we will argue below, this effect is likely to be an artifact of dichotomizing *Phrase Frequency*.

The artifactual nature of the *Phrase Frequency* effect in the traditional analysis does not immediately become clear from a visual inspection of the left panel of Figure 15. At first glance, it seems that the grand mean curve for high frequency phrases should show an early null effect or even a small positivity as compared to low frequency phrases. The probability distribution of predictor values, however, is not uniform in nature, nor is it normally distributed with a mean halfway the *Phrase Frequency* predictor range. The yellow bulge for medium to high frequency phrases in the left panel of Figure 15 represents relatively few data points (24% of the data points are between the mid-point of the *Phrase Frequency* range (0.40) and a predictor value of 2.00). By contrast, the green area for medium to low frequency words represents a large number of data points (65% of the data points are between -1.2 and 0.4). As a result the negativity for the highest frequency phrases is cancelled out to a much lesser extent by small positive voltages for medium to high frequency phrases than the negativity for the lowest frequency phrases is cancelled out by average voltages for medium to lower frequency phrases.

Due to the nature of the probability distribution of the phrase frequencies, therefore, a dichotomization with a split halfway the *Phrase Frequency* range leads to the incorrect conclusion that voltages for high frequency phrases are lower than those for low frequency phrases. This conclusion would be supported by the fact that this difference is significant at a substantial number of data points in the time dimension in both by-item and by-subject analyses, albeit rarely at a Bonferroni-corrected alpha level. As such a dichotomization of *Phrase Frequency* could lead to incorrect conclusions about the nature of the *Phrase Frequency* effect observed here.

A further problem with a traditional analysis of the *Phrase Frequency* effect is that dichotomization of the *Phrase Frequency* predictor results in a loss of information with respect to the U-shaped nature of the phrase frequency effect along the y-axis. The right panel of Figure 15 does not provide any information about the fact that intermediate values of *Phrase Frequency* are characterized by higher voltages than low or high values of *Phrase Frequency* – let alone about how these positivities for medium values of *Phrase Frequency* evolve over time within the 400 to 700 ms window.

To end our conclusion of the traditional analysis of the *Phrase Frequency* effect on a positive note, we should mention that the traditional analysis does pick up on the fact that negativities for high frequency phrases fade out over time: as a result of the relative increase of the voltages for the highest frequency phrases as compared to the lowest frequency phrases the difference between the grand mean curves for high and low frequency phrases decreases as a function of time in the right panel of Figure 15.

In this appendix we compared the GAM analyses reported in this paper to traditional ERP analyses using predictor dichotomization for simulated data, as well as for some of the key effects reported in this paper. Generally speaking, two conclusions can be drawn from this comparison. First, the GAM analyses reported here seem to provide estimates of predictor effects that are compatible with the grand mean curves. The results of a GAM analysis and a traditional analysis converge when dichotomization of a predictor is relatively unproblematic given the nature of a predictor effect. When this is not the case, the differences that arise between the results from a GAM analysis and a traditional analysis are easily explained given the information about the nature of the predictor effect provided by the GAMs.

Second, a GAM analysis provides much more information than does a traditional analysis in which predictors are dichotomized. In a dichotomization analysis predictor values with very different patterns of results are grouped together, which can result in a loss of statistical power. In addition, the nature of tri- or multipartite predictor effects is – by definition – lost when a predictor is dichotomized. This can lead to a loss of information or misguided conclusions about the nature of an effect. By contrast, as seen in the analysis of the simulated data GAM analyses accurately capture non-linear predictor effects as they evolve over time.

Some of the problems associated with a traditional dichotomization analysis can be overcome by choosing an experimental design that investigates the effect of a single categorical predictor with carefully

selected predictor values that fall into two or more discrete categories. Many of the questions in psycholinguistic research, however, are easier to answer in multiple regression designs that allow for the simultaneous investigation of the effect of multiple numerical predictors with continuous distributions. The experimental design and analysis techniques presented here provide an example of how the multiple regression techniques that have become commonplace in reaction time studies can be applied in ERP studies through the use of GAMs. As demonstrated in this appendix, the results from such a GAM analysis provide precise information about the linear and non-linear nature of the effects of multiple numerical predictors as they evolve over time.

## Appendix B

For the GAM analyses reported in this paper we used a hierarchical modeling strategy with analysis windows of 300 milliseconds and a 100 ms overlap between time windows to verify the consistency of results between subsequent time windows. This modeling approach ensures that participant- and item-related variance, as well as task effects and the grand average over time are removed from the data prior to the estimation of the predictor effects as they evolve over time.

An alternative modeling strategy would involve fitting non-hierarchical GAMs on larger time windows. In this appendix, we present the results of such a non-hierarchical modeling strategy in which we included the main trend over time, by-participant smooths over time (restricted to 20 knots), by-participant trial smooths, random intercepts for prepositional phrase (e.g., "with the") and noun (e.g., "saw"), an autocorrelation correction parameter ($\rho = 0.75$), as well as predictor main effect smooths and time by predictor tensor product interactions (restricted to 10 knots in the time dimension and 5 knots in the predictor dimension) in models fitting the ERP signal from 0 to 600 ms after picture onset.

The results of this analysis for the key effects of *Word Frequency*, *Phrase Frequency*, *Relative Entropy* and *NDL Activation* reported in this paper are shown in Figure 16. The top row of each of the four panels of Figure 16 shows the results for each predictor in the 0-300 and 200-500 ms time windows as reported in this paper. The bottom row of each panel presents the results from non-hierarchical analyses for *Word Frequency*, *Phrase Frequency*, *Relative Entropy* and *NDL Activation* on a larger 0-600 ms time window.

*[INSERT FIGURE 16 AROUND HERE]*

Overall, the qualitative nature of the effects for *Word Frequency*, *Phrase Frequency*, *Relative Entropy* and *NDL Activation* in the non-hierarchical analysis on a 600 ms time window is highly similar the qualitative nature of these effects in hierarchical GAMs on 300 ms time windows. For *Word Frequency* and *NDL Activation* we continue to see theta range oscillations for extreme predictor values. In addition, theta range oscillations throughout the predictor dimension characterize the effect of *Relative Entropy* in both analyses.

For *Phrase Frequency*, there is limited evidence for a tensor product interaction with *Time* (*p = 0.07)* in the non-hierarchical GAM. Instead, we observed a main effect of *Phrase Frequency* over time that closely resembles the nature of the *Phrase Frequency* effect in the second half of the 0 to 300 ms time window and throughout the 200 to 500 ms time window, with negativities for low frequency phrases, more positive voltages for phrases with intermediate frequencies and negativities for high frequency phrases. The non-hierarchical GAM on the full 0-600 ms time window shows a further positivity for phrases with extremely high frequencies. Given the limited number of predictor values greater than 2 (2.98% of the data), however, it is unclear how robust this effect is. The fact that a tensor product interaction with *Time* characterizes the *Word Frequency* effect in the non-hierarchical GAM on the 0-600 ms time window, whereas the effect of *Phrase Frequency* is best described by a main effect of *Phrase Frequency* confirms once more that the effects of *Word Frequency* and *Phrase Frequency* are qualitatively different.

Despite the overall similarity of the results for the hierarchical GAMs fit to 300 ms time windows and the results for the non-hierarchical GAMs on 600 ms time windows, there are two differences between the results presented in the top rows and in the bottom rows of the four panels of Figure 16. The first difference concerns the reliability of GAMs near the edges of the analysis window. In the analysis using 300 millisecond time windows, we see effects in the last 50 ms of the 0-300 ms time window and in the first 50 ms of the 200-500 ms time window for *Word Frequency* (positivities for high predictor values), *Relative Entropy* (positivities for high predictor values) and *NDL Activation* (positivities for high predictor values, negativities for low predictor values) that are not reflected in the corresponding analyses for the larger 600 millisecond time windows. This demonstrates that effects in the first 50 ms and the last 50 ms of our 300 ms time windows have to be interpreted with caution. The same caution is required with respect to the interpretation of the first and last 50 milliseconds in analyses for the 600 ms time intervals.

The second difference between the analyses in the top and bottom rows of the four panels of Figure 16 is that the analysis using separate 0 to 300 and 200 to 500 ms time windows results in more conservative estimates of the effect sizes of the *Word Frequency*, *Phrase Frequency*, *Relative Entropy* and *NDL Activation* effects as compared to the analysis for the full 0 to 600 ms time window. It is likely that nearly doubling the number of data points both increases the power of the GAMs and allows it to estimate amplitudes with greater precision.

This comes at the price of having to increase the number of basis functions for the time dimension, in order to allow the model to pick up changes in amplitude over time with the same granularity as in the analyses using restricted 300 ms time windows.

For this study, we decided to follow a conservative modeling strategy, using overlapping time windows of 300 ms and the default settings for the number of basis functions used by the mgcv package for R. This choice also comes with a practical advantage, namely that the time required for fitting a model to hundreds of thousands if not millions of data points is substantially reduced.

## Footnotes

[1] One of our reviewers pointed out that due to the high correlation between word frequency and determiner plus noun bigram frequency any effect of word frequency may instead be an effect of bigram frequency. We would like to make explicit here that we are in no way opposed to such an interpretation of any word frequency effects documented here. By contrast, we believe that word frequency effects for a large part reflect local syntactic co-occurences (see e.g.; Baayen (2010)).

[2] Note that for oscillatory effects the phase of an oscillation co-determines the significance of an effect at a given point in time. Converting the signal to the frequency domain does not help solve this problem. Potential oscillations in the predictor dimension further complicate the process of determining the exact onset of an effect. As a result, the numbers reported for oscillatory effects here are conservative estimates for the temporal onset of these effects.

[3] Phrase frequency effects have not been documented in the naming aloud literature. We therefore have not yet attempted to simulate phrase frequency effects in the $NDR_a$ model.

[4] The effect of *Word Activation* also reached significance from 57 to 81 ms after picture onset for predictor values between 0.59 and 0.89. Given the limited effect size of this extremely early effect of *Word Activation* (see Figure 9), however, we will not discuss this effect in more detail.

## Table Captions

Table 1. Summary of the independent variables (log) *Word Length*, (log) *Word Frequency*, (log) *Phrase Frequency* and *Relative Entropy*. Range is the original range of the predictor. Adjusted range is the range after removing predictor outliers. Mean, median and sd are the means, medians and standard deviations after outlier removal, but prior to residualization.

| predictor | range | adjusted range | mean | median | sd |
|---|---|---|---|---|---|
| *WordLength* | 1.10 - 2.30 | 1.10 – 2.08 | 1.58 | 1.61 | 0.26 |
| *WordFrequency* | 12.90 – 18.96 | 13.60 – 18.37 | 15.74 | 15.50 | 1.25 |
| *PhraseFrequency* | 0 – 14.69 | 6.77 – 12.65 | 8.73 | 8.57 | 1.23 |
| *RelativeEntropy* | 0.10 – 2.34 | 0.10 – 1.39 | 0.54 | 0.55 | 0.28 |

**Figure Captions**

*Figure 1*. Main trend in the ERP signal at electrode Cz as predicted by the main trend GAM (black lines) and as observed (red dots).

*Figure 2.* Left panel: percentage of data points after the onset of articulation as a function of time. Right panel: average root mean square (RMS) across all electrodes from -100 to 900 ms after picture onset.

*Figure 3.* Effect for (log) *Word Length* in the naming latencies.

*Figure 4*. Effect for (log) *Word Length* over time at representative example electrodes. Color coding indicates voltages (in $\mu V$), with warmer colors representing higher voltages. Picture insets show the topography of the effect, with bright red indicating significance at the Bonferroni-corrected alpha level ($p < 0.0004$) and dark red indicating significance at the non-corrected alpha level ($p < 0.05$).

*Figure 5*. Effect for (log) *Word Frequency* over time at representative example electrodes.

*Figure 6*. Effect for (residualized log) *Phrase Frequency* over time at representative example electrodes.

*Figure 7*. Effect for (residualized) *Relative Entropy* over time at representative example electrodes.

*Figure 8.* Effect for *Word Length* as simulated by the Naive Discriminative Reader. The reported correlation is the correlation between the observed and simulated contour surfaces at the presented example electrodes.

*Figure 9.* Effect for *Word Frequency* as simulated by the Naive Discriminative Reader. The y-axis is flipped vertically to allow for an easy comparison with the observed effect of *Word Frequency*.

*Figure 10.* Effect for *Phrase Frequency* as simulated by the Naive Discriminative Reader. The y-axis is flipped vertically to allow for an easy comparison with the observed effect of *Phrase Frequency*.

*Figure 11*. Effect for *Relative Entropy* as simulated by the Naive Discriminative Reader.

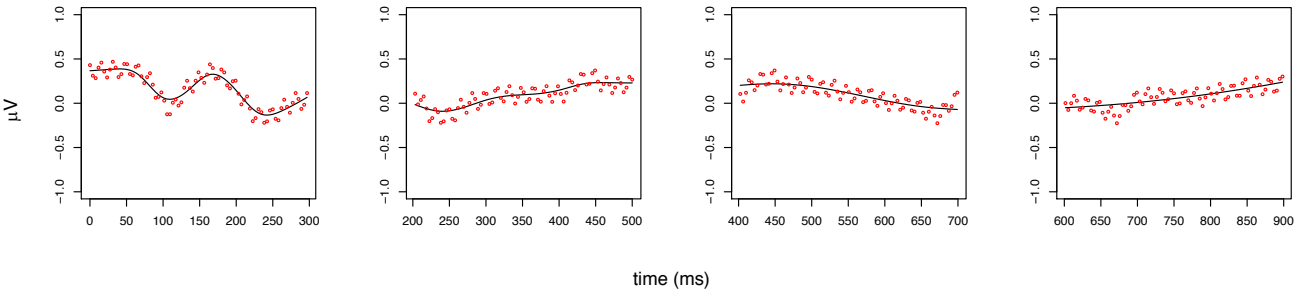
*Figure 12.* Simulated predictor effect with an oscillation in both the time and predictor dimension (left panels) and model fits for this effect in a GAM analysis (middle panel) and a traditional analysis using predictor dichotomization (right panels).

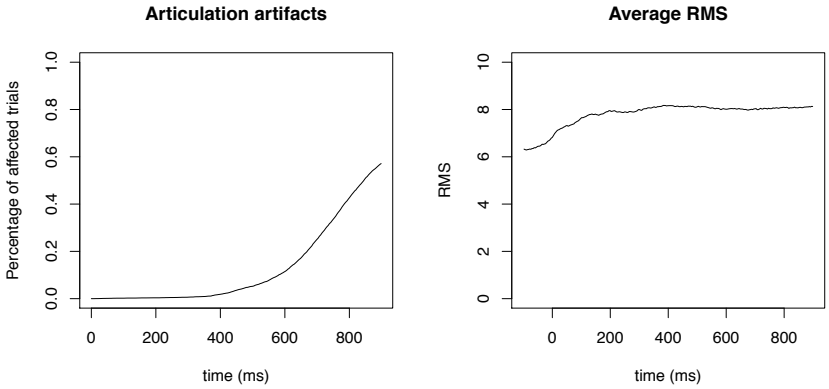
*Figure 13*. The effect of *Word Frequency* at electrode F3 in the 0 to 300 ms time window in a GAM analysis (left panel) and a traditional analysis in which *Word Frequency* is dichotomized (right panel). Color coding at the bottom of the right panel indicates significance of the *Word Frequency* effect in by-item and by-subject ANOVAs for each point in time.

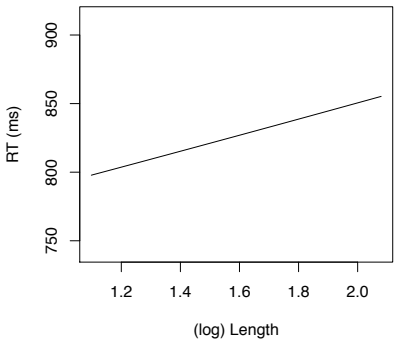
*Figure 14*. The effect of *Relative Entropy* at electrode PO3 in the 0 to 300 ms time window in a GAM analysis (left panel) and a traditional analysis in which *Relative Entropy* is dichotomized (right panel).

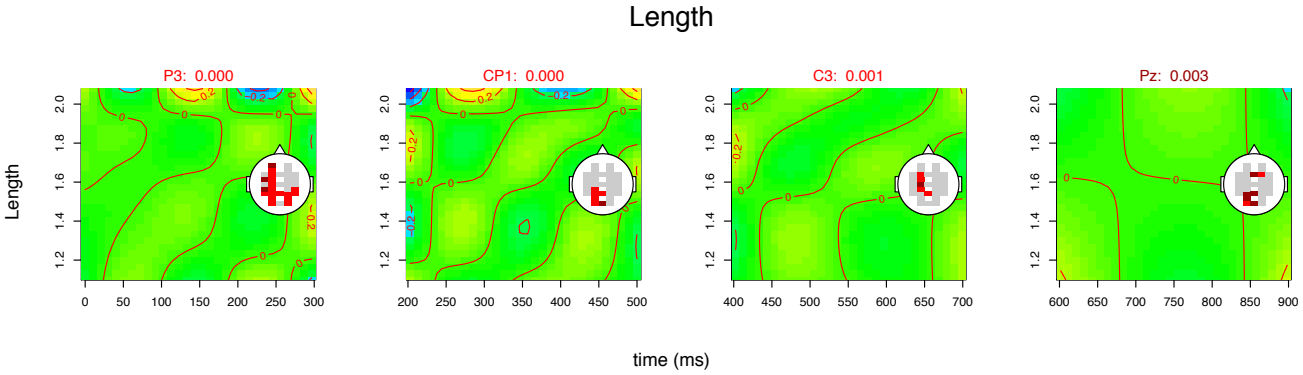
*Figure 15*. The effect of *Phrase Frequency* at electrode CP5 in the 400 to 700 ms time window in a GAM analysis (left panel) and a traditional analysis in which *Phrase Frequency* is dichotomized (right panel).

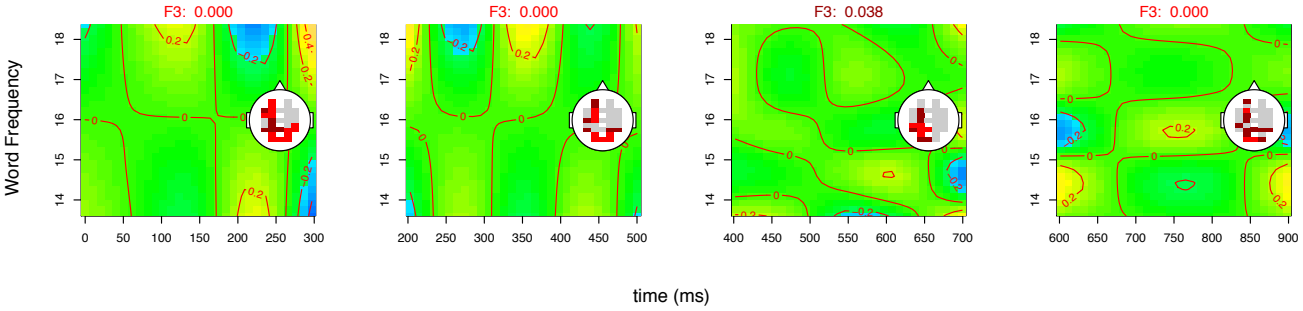
*Figure 16*. The effects of *Word Frequency* (top left panel)*, Phrase Frequency* (top right panel), *Relative Entropy* bottom left panel) and *Activation Word* (bottom right panel) at electrode relevant example electrodes in a hierarchical GAM analysis using 300 ms time windows with a 100 ms overlap between subsequent time windows (top row of each panel) and a non-hierarchical GAM analysis using 600 ms time windows (bottom row of each panel).

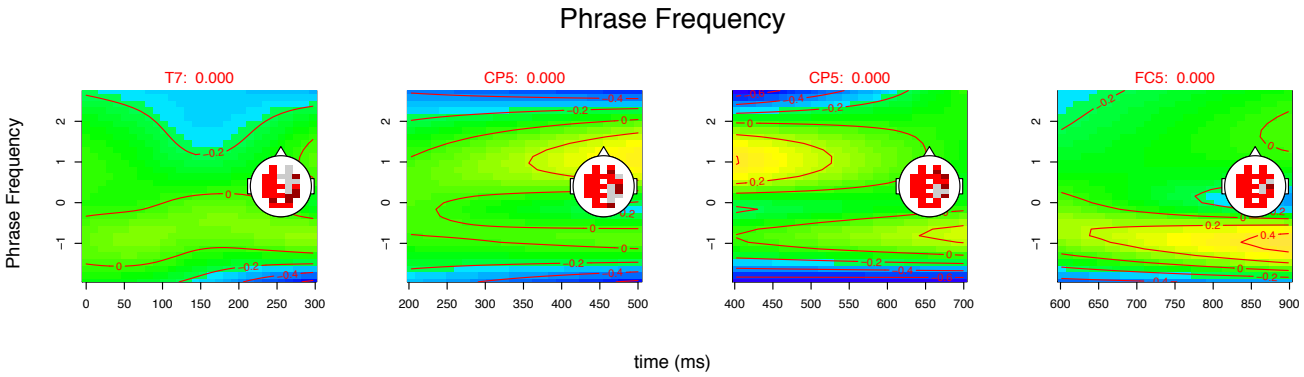

main trend
μV
1.0
0.5
0.0
−0.5
−1.0
0
50
100
150
200
250
300
200
250
300
350
400
450
500
400
450
500
550
600
650
700
600
650
700
750
800
850
900
time (ms)

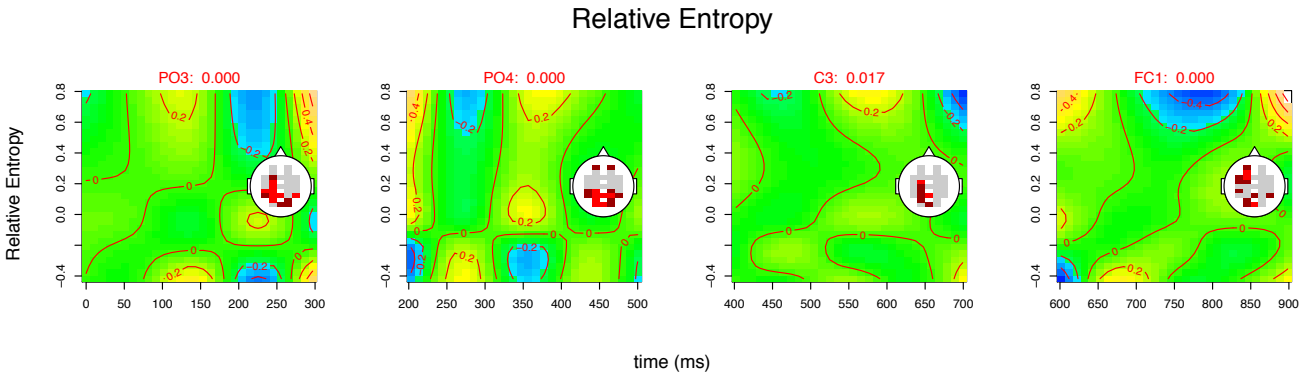

Articulation artifacts
Percentage of affected trials
0.0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
time (ms)
Average RMS
RMS
0
2
4
6
8
10
0
200
400
600
800
time (ms)

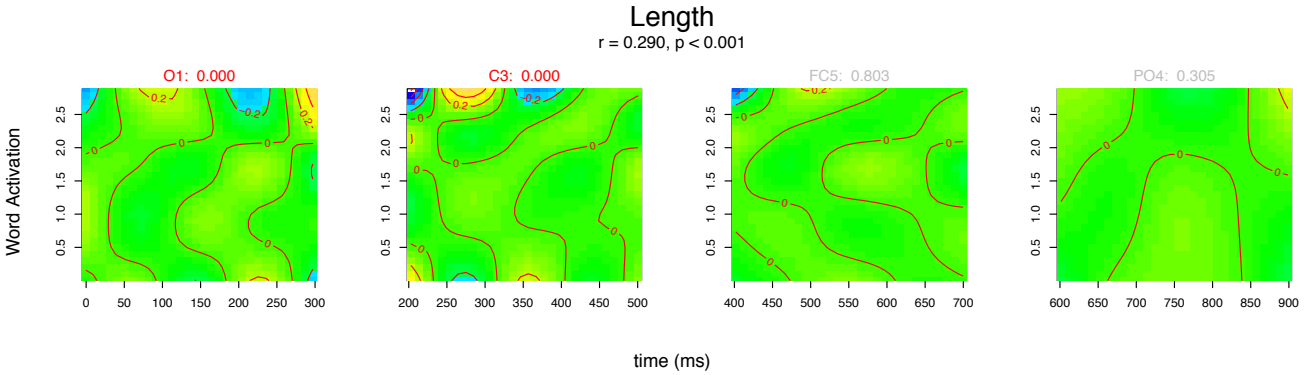

RT (ms)
750
800
850
900
1.2
1.4
1.6
1.8
2.0
(log) Length

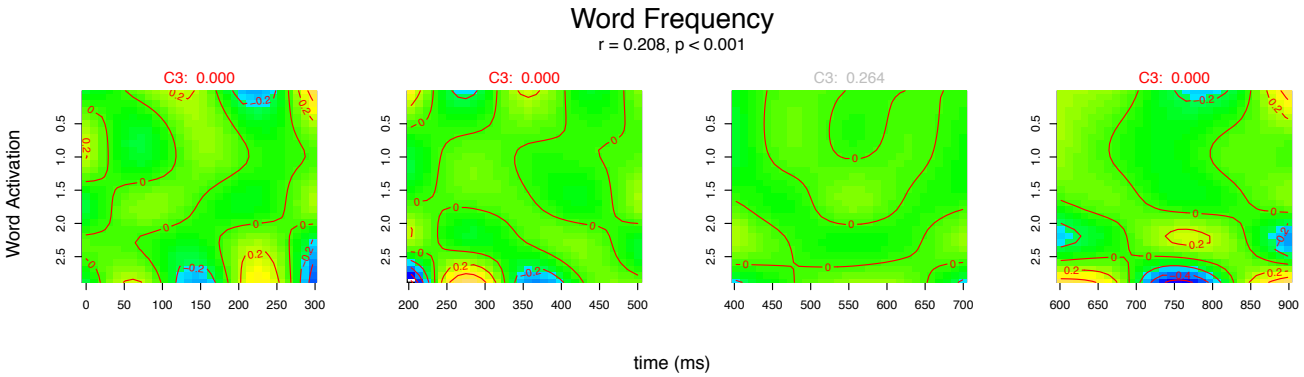

Length
P3: 0.000
CP1: 0.000
C3: 0.001
Pz: 0.003
Length
time (ms)

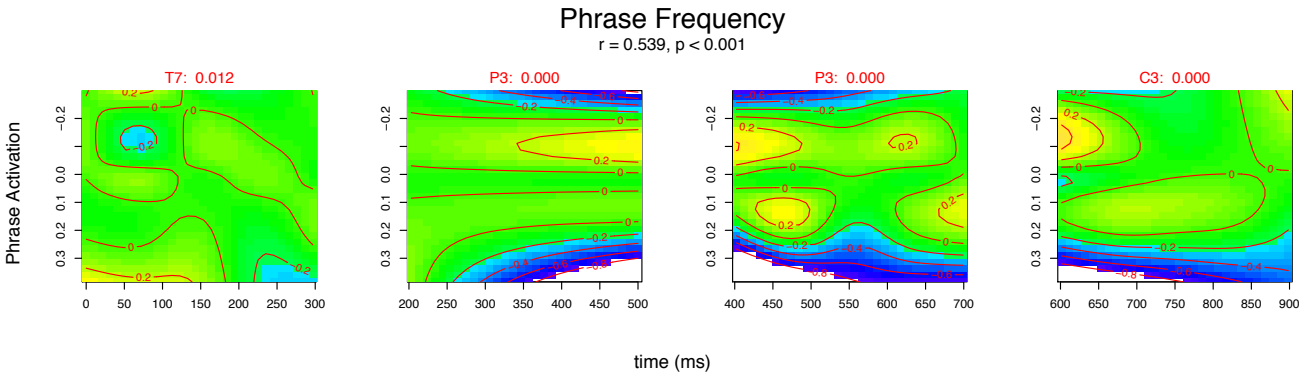

Word Frequency
F3: 0.000
F3: 0.000
F3: 0.038
F3: 0.000
Word Frequency
time (ms)

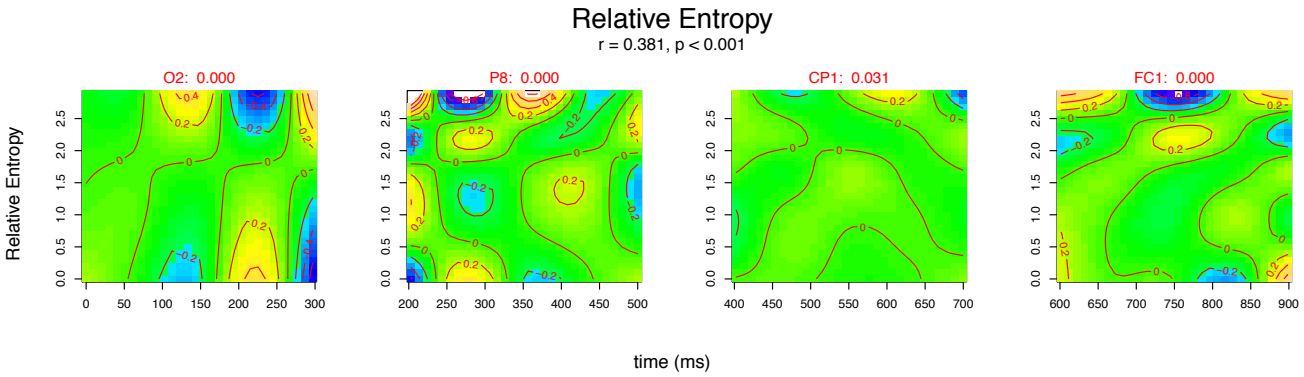

Phrase Frequency
T7: 0.000
CP5: 0.000
CP5: 0.000
FC5: 0.000
Phrase Frequency
time (ms)

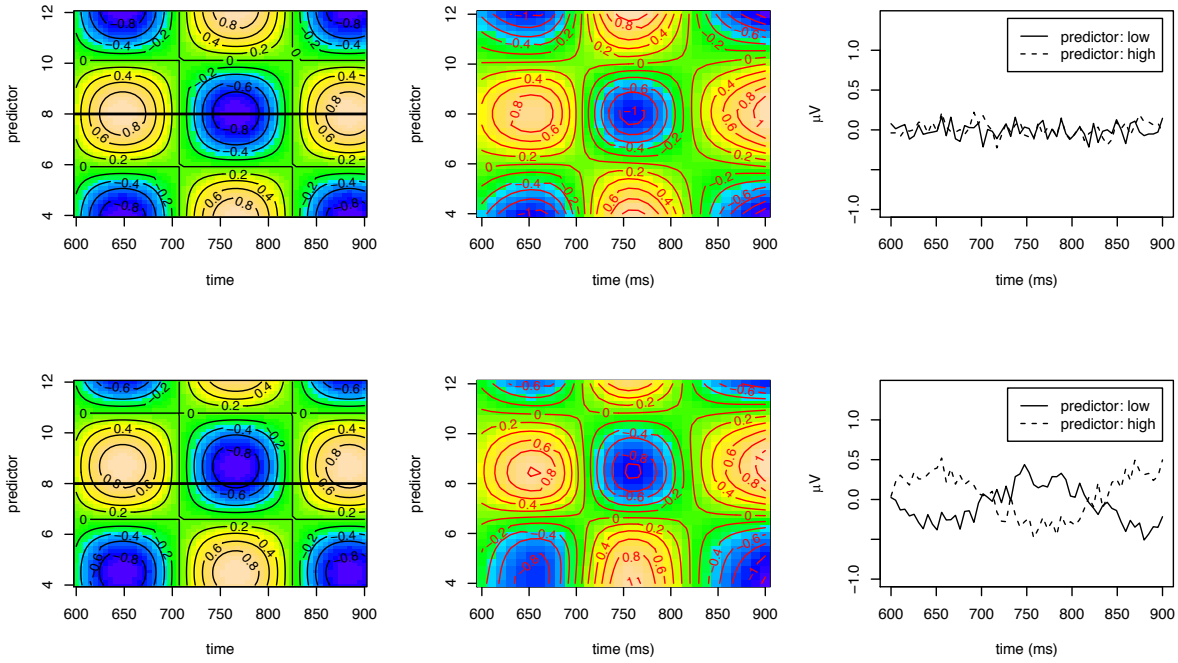

Relative Entropy
PO3: 0.000
PO4: 0.000
C3: 0.017
FC1: 0.000
Relative Entropy
time (ms)

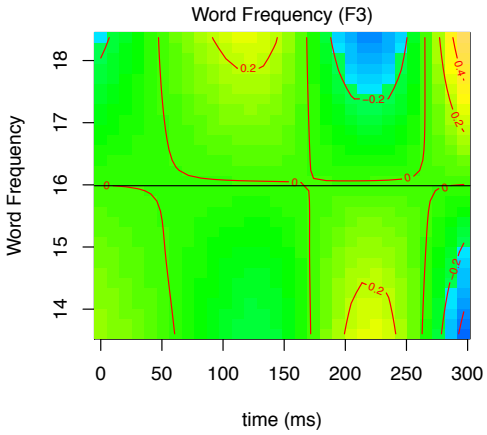

Length
r = 0.290, p < 0.001
O1: 0.000
C3: 0.000
FC5: 0.803
PO4: 0.305
Word Activation
time (ms)

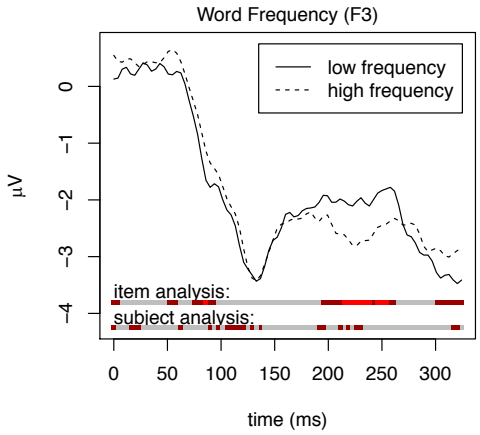

Word Frequency
r = 0.208, p < 0.001
C3: 0.000
C3: 0.000
C3: 0.264
C3: 0.000
Word Activation
time (ms)

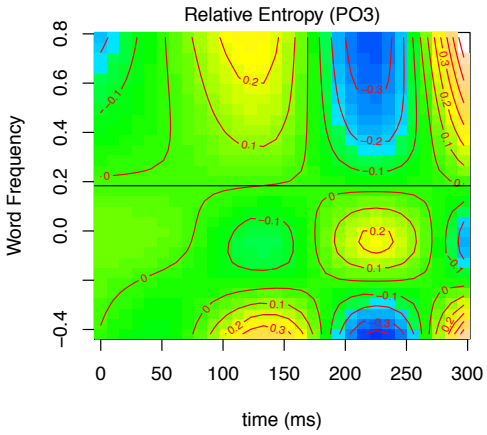

Phrase Frequency
r = 0.539, p < 0.001
T7: 0.012
P3: 0.000
P3: 0.000
C3: 0.000
Phrase Activation
time (ms)

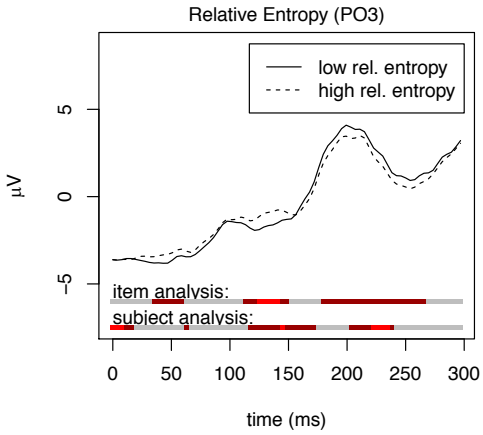

Relative Entropy
r = 0.381, p < 0.001
O2: 0.000
P8: 0.000
CP1: 0.031
FC1: 0.000
Relative Entropy
time (ms)

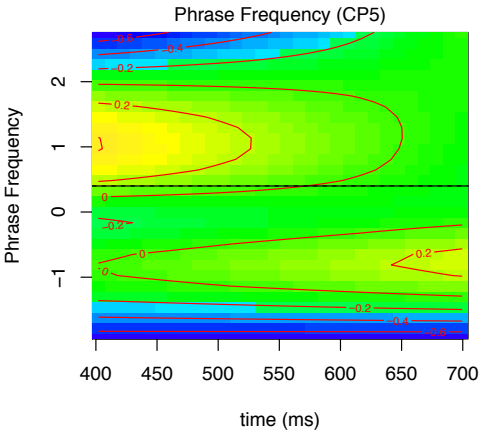

predictor
time
time (ms)
μV
predictor: low
predictor: high

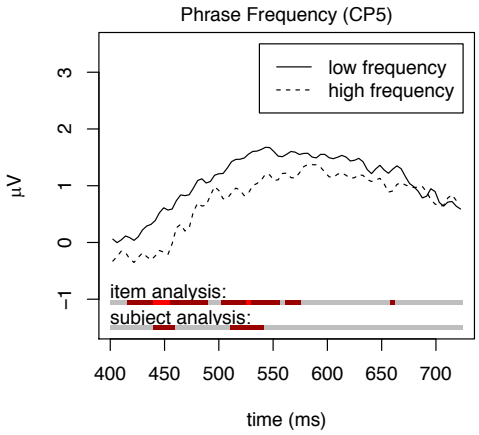

Word Frequency (F3)
Word Frequency
18
17
16
15
14
0
50
100
150
200
250
300
time (ms)

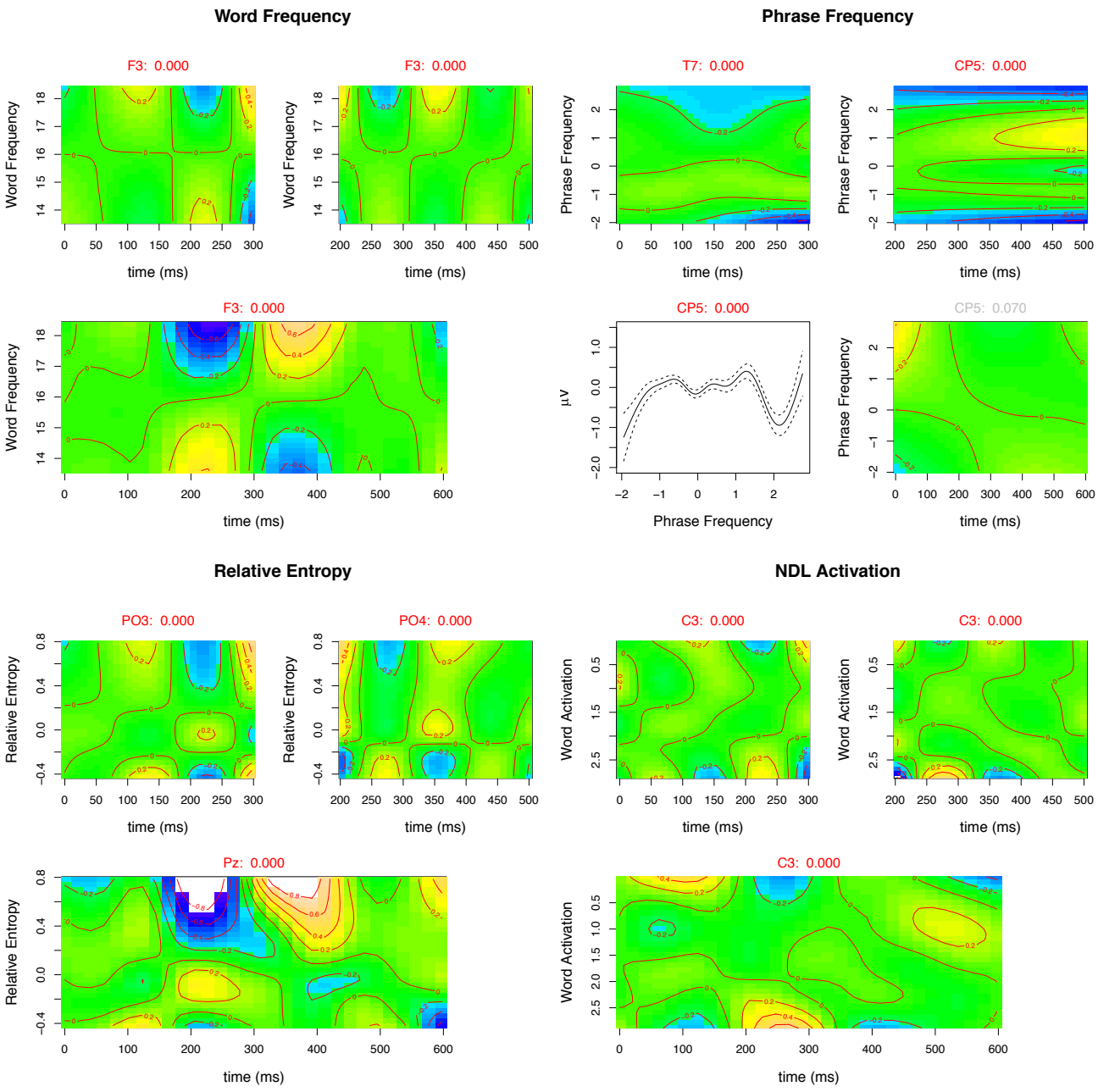

Word Frequency (F3)
low frequency
high frequency
μV
0
−1
−2
−3
−4
item analysis:
subject analysis:
0
50
100
150
200
250
300
time (ms)


Relative Entropy (PO3)
Word Frequency
0.8
0.6
0.4
0.2
0.0
−0.4
0
50
100
150
200
250
300
time (ms)


Relative Entropy (PO3)
low rel. entropy
high rel. entropy
μV
5
0
−5
item analysis:
subject analysis:
0
50
100
150
200
250
300
time (ms)


Phrase Frequency (CP5)
Phrase Frequency
time (ms)
400
450
500
550
600
650
700
−1
0
1
2


Phrase Frequency (CP5)
low frequency
high frequency
item analysis:
subject analysis:
μV
time (ms)
400
450
500
550
600
650
700
−1
0
1
2
3


Word Frequency
Phrase Frequency
F3: 0.000
F3: 0.000
T7: 0.000
CP5: 0.000
F3: 0.000
CP5: 0.000
CP5: 0.070
Relative Entropy
NDL Activation
PO3: 0.000
PO4: 0.000
C3: 0.000
C3: 0.000
Pz: 0.000
C3: 0.000
Word Frequency
Phrase Frequency
Relative Entropy
Word Activation
μV
time (ms)