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

Peter Hendrix
University of Tübingen, Germany

Patrick Bolger
University of Alberta, Canada

Harald Baayen
University of Tübingen, Germany

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, persistent negativities were present for both high and low frequency phrases. Furthermore, independent effects with different temporal and topographical signatures characterized phrasal prototypicality. These results pose a challenge to exemplar-based models and fit more readily with a discrimination learning approach to language processing. In a simulation study we demonstrate that naive discrimination learning (Baayen et al., 2011) offers a competitive account of the ERP signal following picture onset as compared to standard lexical predictors that offers further insight into the nature of n-gram frequency effects.

Keywords: ERP , picture naming, prepositional paradigm, phrase frequency, relative entropy, discrimination learning

Introduction

Few effects in the psycholinguistic literature are better documented than the word frequency effect: the more often a word occurs in the language, the faster and more accurate people respond to that word in a wide range of linguistic tasks, including lexical decision (see, e.g.; (Scarborough et al., 1977; Balota et al., 2004) and word naming (Forster & Chambers, 1973; Balota & Chumbley, 1985; Jared, 2002). Recently, a number of studies have shown that word frequency effects are also present in electroencephalograms (EEGs) following the onset of a (linguistic) stimulus, which are commonly referred to as event-related potentials (ERPs).

Typically, the effects of word frequency on ERPs arise rapidly after the onset of the stimulus. Hauk et al. (2006), for instance, found an effect of word frequency in a visual lexical decision task as early 110 ms after stimulus onset. This early effect of word frequency was most prominent in left-lateralized temporal and parietal areas. Similarly, Sereno et al. (1998) found a word frequency effect in a visual lexical decision task that first reached significance at 132 ms after stimulus onset, whereas Penolazzi et al. (2007) observed an effect of word frequency on the ERP signal in a sentence-reading task that started at 120 ms after written word onset. The topographically widespread effect of word frequency in the picture naming task used by Strijkers et al. (2010) arose somewhat later, with more positive mean amplitudes for high frequency words than for low frequency word from 150 ms until voice onset.

The effect of frequency, however, is not limited to the word level. Arnon and Snider (2010) showed that phrasal decision latencies for high frequency phrases such as "all over the plac" 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 n-gram frequency effect has been replicated in a number of recent studies, showing n-gram frequency effects in sentence repetition (Bannard & Matthews, 2008), sentence reading (Siyanova-Chanturia et al., 2011), sentence recall (Tremblay et al., 2011) and frequency rating (Shaoul et al., 2013) tasks. Tremblay and Baayen (2010) added to these findings by observing an n-gram frequency effect in a free recall ERP study. The temporal onset of this effect was similar to that of the effects of word frequency described above, with n-gram probability first being significant around 110 ms after stimulus onset.

The n-gram frequency effect is theoretically interesting. At the very least, it "add[s] multi-word phrases to the units that influence processing in adults" (Arnon & Snider, 2010, p.76), which suggests that language users "seem to have [...] some experience-derived knowledge of specific four-word sequences" (Bannard & Matthews, 2008, p.246). Much, however, remains unclear about how this knowledge is implemented, and, therefore, about the implications of n-gram frequency effects for different models of language processing.

One interpretation of n-gram frequency effects is to consider these effects as evidence for whole-phrase representations. As noted by Baayen et al. (2013), such an interpretation 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 these stored sequences. In these exemplar-based approaches n-gram frequency effects are directly related to the n-gram representations that are stored in memory.

Baayen et al. (2013), however, noted that storing each multiword sequence and its associated frequency in memory 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. Baayen et al. (2013) continue their argument by stating that even if the storage of gigantic numbers of word n-grams were neuro-biologically possible, on-line processing over an instance space of this size would be very time-consuming. To side-step this problem, the memory-based learning system implemented in TiMBL (Daelemans et al., 2010) uses information gain trees (Daelemans et al., 1997) as a compression algorithm to reduce the computational demands of on-line searches.

An additional problem with n-gram representations described by Baayen et al. (2013) 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 and Rumelhart (1981) and D. Norris and McQueen (2008) for models that integrate word unigram frequencies as a priori-probabilities). The application of shortlists in interactive activation models (D. G. 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 (Ramscar et al., 2010; Baayen et al., 2013). Ramscar et al. (2010) argued that holistic linguistic representations may be beneficial at the earliest stages of learning (Dabrowska, 2000; Tomasello, 2003), but that additional experience will reduce the association strength 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) and Arnon and Ramscar (2012) for simulations that confirm this prediction).

Baayen et al. (2013) provided computational support for such an interpretation of the n-gram frequency effect by successfully simulating the findings of Arnon and 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, Filipović Durđević, & Moscoso del Prado Martín, 2009; Milin, Kuperman, et al., 2009; 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

$$\text{Relative Entropy} = \sum_{i=1}^{n}(p_i * \log_2 (p_i/q_i)) \tag{1}$$

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-a-vis the distribution of prepositional frequencies in the language as a whole. As such, the prepositional relative entropy effect poses a further 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 naive discrimination learning models 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

Experiment

In what follows we present the results of a primed picture naming experiment that gauges the effects of word frequency, phrase frequency and phrase prototypicality using event-related potentials (ERPs). The current work seeks to extend previous findings in two ways. First, while previous studies have investigated the effects of word frequency on ERPs in a variety of tasks, 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 a picture naming task. 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 measures derived from a naive discrimination learning model provide further insight into the temporal and spatial dynamics of the ERP signal in a primed picture naming task. The discriminative learning approach adopted by the ERP model has been shown to successfully simulate a variety of behavioral measures, including lexical decision latencies Baayen et al. (2011), word naming latencies (Hendrix, Ramscar, & Baayen, 2015) and eye movements during full text reading (Hendrix, Nick, & Baayen, 2015). 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 discrimination learning approach 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 e ntropy 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 of the pudding is in the eating 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 statistical methods used to analyze the data and the results of the experiment. In the second part, we will present a simulation study in which we explore to what extent the discriminative learning framework can provide further insight into the temporal and spatial dynamics of the ERP signal following picture onset.

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 20.43 (sd: 4.67). 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 1*T* *n*-gram data (Brants & Franz, 2006). Trigram frequencies for all prepositional phrases consisting of a preposition, an article ("a" or "the") and one of the 68 concrete nouns were extracted. For a given noun, the phrases at 25%, 50%, 75% and 100% of the summed phrase frequency distributions ("[preposition] a [noun]" + "[preposition] the [noun]") were included as stimuli. For the noun "saw", for instance, this procedure generated the experimental items "into the saw" (summed frequeny: 2061; frequency: 2061), "from the saw" (summed frequency: 5358; frequency: 4525), "to the saw" (summed frequency: 9781, frequency: 8436) and "with the saw" (summed frequency: 20464; frequency: 8691). 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 (logged) 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 *Picture Complexity*, *Preposition Length*, *Word Length*, *Preposition Frequency*, *Word Frequency*, *Phrase Frequency* and *Relative Entropy*.

Picture Complexity is the size of the picture file in bytes. *Preposition Length* and *Word Length* are the length of the preposition and the target noun in letters. *Preposition Frequency*, *Word Frequency* and *Phrase Frequency* are the frequency of the preposition (e.g., "with"), target noun (e.g., "saw") and phrase (e.g., "with the saw") in the Google *n*-gram data. *Picture Complexity*, *Preposition Length*, *Word Length*, *Preposition Frequency*, *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 prepositional phrases with definite article 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 0.001.544.62% of all predictor values for *Word Frequency*, 5.77% of all predictor values for *Phrase Frequency* and 4.62% 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,

Table 1: Summary of the independent variables (*log*) *Picture Complexity*, (*log*) *Preposition Length*, (*log*) *Word Length*, (*log*) *Preposition Frequency*, (*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.

predictor	range	adjusted range	mean	median	sd
Picture Complexity	8.53 - 11.13	8.69 - 10.83	9.88	9.91	0.50
Preposition Length	0.69 - 1.95	0.69 - 1.95	1.15	1.38	0.45
Word Length	1.10 - 2.30	1.10 - 2.08	1.58	1.61	0.26
Preposition Frequency	15.65 - 23.17	17.63 - 23.17	21.09	21.81	1.61
Word Frequency	12.90 - 18.96	13.60 - 18.37	15.74	15.50	1.25
Phrase Frequency	0.00 - 14.69	6.77 - 12.65	8.73	8.57	1.23
Relative Entropy	0.10 - 2.34	0.10 - 1.39	0.54	0.55	0.28

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.

The resulting data set is characterized by a considerable amount of collinearity ($\kappa = 123.16$). *Word Frequency*, for instance, correlates positively with *Phrase Frequency* ($r = 0.42$) and negatively with *Preposition Frequency* ($r = -0.40$), *Relative Entropy* ($r = -0.40$) and *Word Length* ($r = -0.51$). Similarly, *Preposition Frequency* correlates not only with *Word Frequency*, but also shows a strong negative correlation with *Preposition Length* ($r = -0.76$).[1]

One approach for dealing with collinearity is predictor residualization. In this approach, rather than entering the raw predictors into a regression model, one or more of the predictors are residualized prior to analysis by running a preliminary regression analysis with the predictor that is to be residualized as the dependent variable and one or more other predictor as the independent variable. For the current data, for instance, it would be an option to residualize *Phrase Frequency* from *Word Length*, *Word Frequency*, *Preposition Frequency* and *Relative Entropy*. The resulting *Phrase Frequency* measure would then no longer correlate with these other predictors.

Recently, however, Wurm and Fisicaro (2014) argued that residualizing is not a useful remedy for collinearity. Contrary to popular believe, they state, residualization "does not change the results for the predictor that was residualized [... and ...] does not create an improved, purified, or corrected version of the original predictor" (Wurm & Fisicaro, 2014, p.45). What residualization does do, the authors continue, is introduce an additional statistical problem: depending on the correlation between predictor X_1 and predictor X_2 and the correlations between the dependent variable Y and predictors X_1 and X_2, residualization of X_1 results in either underestimating or overestimating the statistical importance of the non-residualized predictor X_2. Given these consideration, they therefore conclude that, in the context of collinearity issues, "residualization of predictor variables is not the hoped-for panacea" (Wurm & Fisicaro, 2014, p.47).

[1] We explicitly mention correlations with an absolute value greater than 0.30 only here.

Not all is bad, however. While suppression is a serious problem when it occurs, it may not be as common as previously thought. Darlington (1990, p.155) (as cited in Wurm & Fisicaro, 2014), for instance, states that "suppression rarely occurs in real data", and Cohen et al. (2003) (as cited in Wurm & Fisicaro, 2014) state that "it is more likely to be seen in fields like economics, where variables or actions often have simultaneous equilibrium-promoting effects". While the correlation threshold for potential suppression depends on the correlation of the involved predictor with the dependent variable, suppression artifacts are highly uncommon for weak or moderate correlations.

For the current data set, these statements suggests that while suppression is not outside the realm of possibilities for the effects of *Preposition Length* and *Preposition Frequency*, our analysis of the main predictors of interest (*Word Frequency*, *Phrase Frequency* and *Relative Entropy*) is unlikely to suffer from this problem. We therefore decided to use the raw, non-residualized measures of *Picture Complexity*, *Preposition Length*, *Word Length*, *Preposition Frequency*, *Word Frequency*, *Phrase Frequency* and *Relative Entropy* described above as predictors in our analysis.

Procedure

Data were recorded from 32 Ag/AgCl active electrodes (*Fp1*, *Fp*, *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 (4.41%). In addition, we removed incorrect naming responses from the data (2.68%). Trials for which the maximum absolute voltage after signal correction exceeded 100 $\hat{I}ijV$ at any channel were removed from the data for all channels (5.25%). Furthermore, 39 trials (0.48%) were removed due to technical failure. No averaging over participants or items was done prior to 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 $1.8-3$)). GAMs have recently been used in a number of ERP studies on language processing (Kryuchkova et al., 2011; Baayen et al., 2015).

Reaction time analysis

We fitted a GAM with by-participant factor smooths for trial, a random intercept for prepositional phrase (e.g.; "with the") and noun (e.g.; "saw") and a smooth function for the previous naming latency to the naming latencies to the naming latency data. Naming latencies and previous naming latencies further than 2 standard deviations from the mean were removed from the data. A log transformation was applied to the naming latencies and previous naming latencies to remove a rightward skew from the data. We modeled the predictor effects for *Picture Complexity*, *Preposition Frequency*, *Word Frequency*, *Phrase Frequency* and *Relative Entropy* using smooth functions. We modeled the effects of *Preposition Length* and *Word Length* with a parametric term, because of the limited number of unique values for these predictors.

ERP analysis

For each electrode, we fitted a GAM with by-participant factor smooths for trial and time, as well as random intercepts for prepositional phrase and noun to the ERP from 0 to 600 ms after picture onset. For each of the predictors *Picture Complexity*, *Preposition Frequency*, *Word Frequency*, *Phrase Frequency* and *Relative Entropy* we furthermore included a main effect smooth, as well as a tensor product interaction with time. We furthermore included main effect smooths for *Preposition Length* and *Word Length*. The main effect smooths for *Word Length* and *Preposition Length*, however, reached significance at 1 electrode only (*Word Length*: electrode *C4*, $p = 0.023$; *Preposition Length*: electrode $AF4$, $p = 0.020$). Given the number of comparisons, these results provide little evidence for a statistically robust effect of *Word Length* and *Preposition Length*. We therefore decided not to include the main effect smooths for *Preposition Length* and *Word Length* in the GAMs reported in this paper. Effects in the predictor dimension were limited to 5th order non-linearities ($k = 5$), whereas effects in the time dimension were to 20th order non-linearities ($k = 20$). To

control for AR1 autocorrelation processes, we included an autocorrelation parameter ρ in the GAMs, which was set to 0.75.

Figure 1 shows the predicted values of our GAM at electrode *C3* (black line). Predicted main trend values correlate highly with average observed voltages (red dots): $r = 0.999$. This indicates that our GAM successfully captures the general trend of the ERPs over time. GAM fits correlated highly with averaged observed voltages across all electrodes, with an average correlation of $r = 0.997$ between predicted values and average observed values.

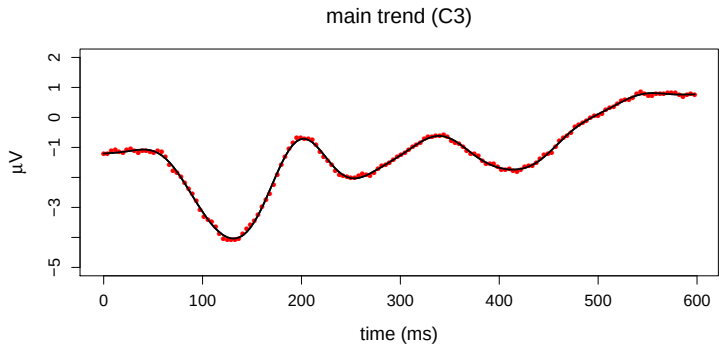

Figure 1. Main trend in the ERP signal at electrode *FC1* as predicted by the main trend GAM (black line) and as observed (red dots).

The average reaction time in the experiment was 854 ms (median: 800 ms). The earliest responses started coming in much earlier than that. As can been seen in the left panel of Figure 2 articulation has begun for a significant proportion of trials at the end of our 600 ms analysis window (13.6%). 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.

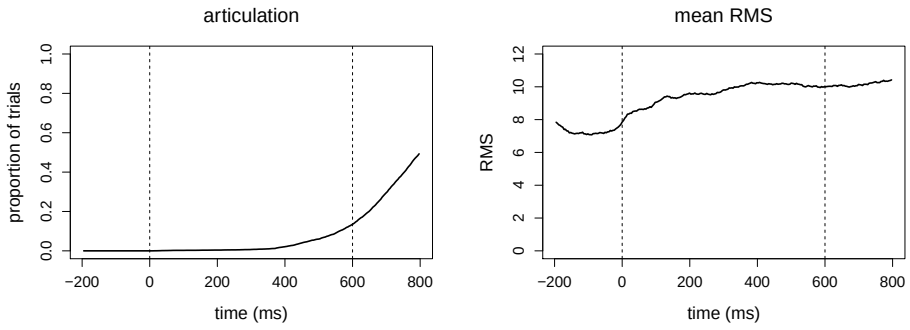

Figure 2. Left panel: percentage of data points after the onset of articulation as a function of time. Right panel: average root mean square (RMS) across all electrodes from -200 to 800 ms after picture onset.

There are two options for dealing with EMG activity in our data. First, we could remove all data points after the onset of articulation. As noted by Hillyard and Picton (1987), however, 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. Second, as noted above, articulation has started for 13.6% of all trials before the end of the 600 ms analysis window. Furthermore, the voice key did not register naming latencies for a non-trivial number of trials (for details, see the reaction time results section). Given that we are unsure about whether or not articulation started before the end of our analysis window, we would have to exclude these trials entirely avoid articulation artifacts altogether. Removing these data points and trials from the analysis would result in a substantial loss of statistical power.

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 (−200 to 0 ms), the average RMS across all electrodes and time points is 7.31, whereas in the post-stimulus interval (0 to 600 ms) it is 9.96. As predicted, the RMS does increase as a function of time. The increase, however, is fairly limited: the average RMS is 8.98 in the 0-200 ms interval, 9.83 in the 200-400 ms interval and 10.13 in the 400-600 ms interval. Furthermore, the increase in RMS primarily occurs in the first 400 ms after picture onset, but stabilizes in the 400-600 ms time window. Given that only 2.11% of the articulations began prior to the 400 ms mark, the early increase in RMS values is unlikely to be due to artifacts following the onset of articulation.

To further inspect the potential problem of a decreased SNR due to articulation artifacts we looked at the SNR across electrodes in the last 200 ms of our analysis window (i.e.; 400-600 ms after picture onset). 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 elevated at frontal locations. While the average RMS across all electrodes in the last epoch was 10.13, the average RMS values in 400-600 ms time window at frontal electrodes were 15.02 (*Fp1*), 14.01 (*Fp2*), 13.13 (*AF3*), 11.67 (*AF4*), 12.51 (*F7*), 11.66 (*F3*), 8.62 (*Fz*), 9.72 (*F4*), 12.10 (*F8*), 10.32 (*FC5*), 10.34 (*FC1*), 6.51 (*FC2*) and 9.50 (*FC6*). As such, the average RMS values at frontal electrodes show an increase in the last 200 ms. This increase, however, is limited to the most frontal electrodes only.

Despite the topographically limited and quantitatively moderate 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 excluded all trials with naming latencies shorter than 600 ms, as well as trials for which no naming latencies were available. As such, this analysis excluded all potential muscle artifacts following articulation onset. The results of this analysis were highly similar to the

results of the analysis on the full data set. We therefore decided to carry out our analysis on the full data set, including data points after articulation onset and trials for which no naming latencies were available.

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 ERP analysis to a traditional ERP analysis for simulated data, as well as for some of the key predictor effects described below in the Appendix.

Results

Reaction time results

During the experiment there were some technical difficulties regarding the sensitivity of the voice key. This resulted in response times not being registered for 2 participants. These participants therefore could not be included in the reaction time analysis. In addition, we removed all further trials for which the voice key did not register a response (7.82%) from the data prior to the reaction time analysis

The naming latencies showed a significant random intercept for the target noun ($F = 11.60$, $p < 0.001$), but not for the prepositional phrase ($F = 0.06$, $p = 0.267$). Furthermore, we found significant by-participant factor smooths for trial ($F = 8.30$, $p < 0.001$), as well as a significant smooth for (*log*) previous RT ($F = 13.21$, $p < 0.001$). Finally, we observed a significant effect of *Picture Complexity* ($F = 3.29$, $p = 0.034$). The effect of *Picture Complexity* is depicted in Figure 3. For ease of interpretation, normal linear naming latencies are plotted rather than the log transformed latencies used for modeling.

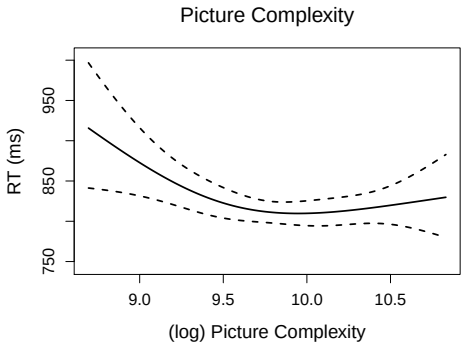

Figure 3. Effect for (log) Picture Complexity in the naming latencies.

As can be seen in Figure 3, the effect of *Picture Complexity* is quadratic in nature, with low *Picture Complexity* leading to longer naming latencies and the effect leveling off for high predictor values. This effect of *Picture Complexity* is perhaps most easily interpreted by taking into consideration that *Picture Complexity* is proportional to information: the more complex a picture, the more information it contains. The longer naming latencies for pictures with limited complexity, therefore, may be a result of the fact that less complicated pictures do not contain enough information for a rapid identification of the depicted object.

ERP results

In this section, we will discuss the results for the predictors *Picture Complexity*, *Preposition Frequency*, *Word Frequency*, *Phrase Frequency* and *Relative Entropy*. For each predictor, we visualize the time by predictor tensor product, as well as the main effect over time at a representative example electrode. Given the fact that GAMs tend to be somewhat unreliable near the edges, we selected representative example electrodes that did not display potentially unreliable behavior near the edges of the analysis window whenever possible.

Picture Complexity

Figure 4 shows the contour plot of the tensor surface for time by *Picture Complexity*. The x-axis represents time (in ms) at a representative example electrode. *Picture Complexity* 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 $\hat{I}ijV$. The p-value for the effect at the depicted electrode is presented in brackets in the figure title.

Figure 4 furthermore contains a picture inset. This picture inset shows the topography of the effect, with dark red indicating significance at an alpha level of 0.05 and bright red indicating significance at a Bonferroni-corrected alpha level of $(0.05/32 =)$ 0.0016. As can be seen in the inset in Figure 4, the tensor product between time and *Picture Complexity* is highly significant for a large number of electrodes across the scalp. A visual inspection of the results, however, reveals that the effect is most prominent in left and central parietal and occipital regions.

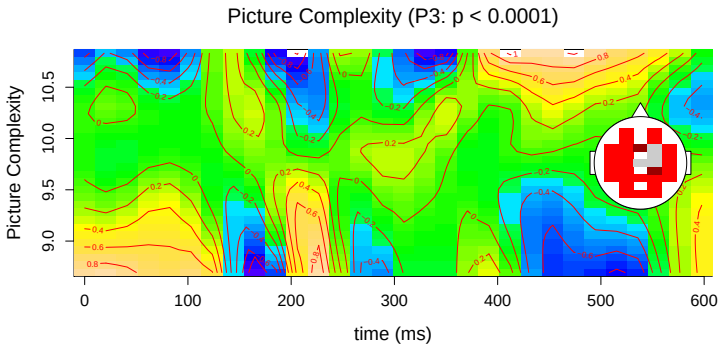

Figure 4. Effect for the tensor product interaction between time and (*log*) *Picture Complexity* at electrode *P3*. Color coding indicates voltages (in $\hat{I}ijV$), with warmer colors representing higher voltages. Picture insets show the topography of the effect, with bright red indicating significance at the Bonferroni-corrected alpha level ($p < 0.0016$) and dark red indicating significance at the non-corrected alpha level ($p < 0.05$).

For both high and low values of *Picture Complexity*, Figure 4 shows that voltages are negative, then positive, then negative, then positive, et cetera. In other words, oscillations tied to the complexity of the presented picture are present in the ERP following picture onset. These oscillations have the opposite phase for low and high values of *Picture Complexity*: when very complex pictures show negativities, show high voltages, less complex pictures

show low voltages and vice versa. To determine the frequency of the oscillations, we converted the time domain representation of the ERP signal seen in Figure 4 to the frequency domain. Although the frequency of the oscillations varies with time and predictor values, a peak in spectral intensity that corresponds to the early oscillations for highly complex pictures and the oscillations for pictures with low complexity in the middle of the analysis widow is reached at 7 *Hz*. As such, these oscillations tied to *Picture Complexity* are in the upper part of the theta range (3 to 7.5 *Hz*).

To gauge the temporal onset of time by predictor tensor products, 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 for high values of *Picture Complexity* is 46 ms after picture onset. The early positive voltages for low values of Picture Complexity, however, are already significant right after picture onset.[2]

On the one hand, finding ERP activity tied to the presentation of a visual stimulus at or even prior to picture onset is unsurprising. Given that the time between the presentation of the fixation mark and picture onset was fixed throughout the experiment, participants were able to accurately predict when the next picture onset would appear on the screen. On the other hand, however, finding ERP activity tied to the properties of a specific visual stimulus at picture onset is less expected.

There are at least two possible explanations for the extremely early effect of *Picture Complexity*. First, GAM estimates can be somewhat unreliable near the edges of the analysis window. It could be the case that uncertainty about the effect for low complexity pictures in the first 50 ms led to a temporal overestimation of a positivity that started somewhat later in time. An alternative explanation for the early onset of the *Picture Complexity* effect comes from the effect of the simple smooth term for *Picture Complexity*, which represents the main effect of *Picture Complexity* over time.

[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. In addition, as a result of phase shifts across the scalp these estimates are sensitive to the choice of the example electrode.

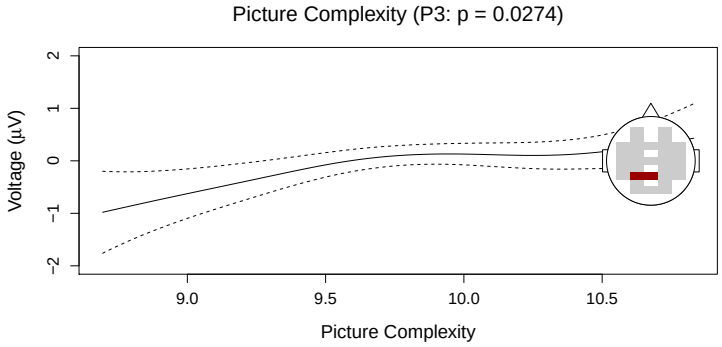

Figure 5. Effect for the main effect smooth of (*log*) *Picture Complexity* over time at electrode *P3*. Picture insets show the topography of the effect, with bright red indicating significance at the Bonferroni-corrected alpha level ($p < 0.0016$) and dark red indicating significance at the non-corrected alpha level ($p < 0.05$).

The main effect of *Picture Complexity* is presented in Figure 5. In contrast to the widespread effect of the time by *Picture Complexity* tensor product interaction, the main effect of *Picture Complexity* showed a topographically limited effect at a non-corrected alpha level only. Nonetheless, voltages seem to be somewhat increased for pictures with a higher visual complexity as compared to pictures with a lower visual complexity. Although the statistical evidence for this main effect of *Picture Complexity* is limited, this suggests that the early positivity for low values of *Picture Complexity* may indicate the absence of any main effect of *Picture Complexity* in the first 100 ms after picture onset. In other words: the early significance of the time by *Picture Complexity* tensor product may be a significant adjustment to the non-significant main effect smooth for *Picture Complexity* rather than a significant effect of *Picture Complexity* as such.

Preposition Frequency

Figure 6 presents the tensor product interaction of time by *Preposition Frequency*. The effect of *Preposition Frequency* is most prominent for low predictor values, with higher voltages for low frequency prepositions as compared to higher frequency preposition in the first 200 ms after picture onset. The fact that we see a significant effect of *Preposition Frequency* right after picture onset is unsurprising, given the fact that prepositions temporally preceded pictures in the experimental paradigm adopted here.

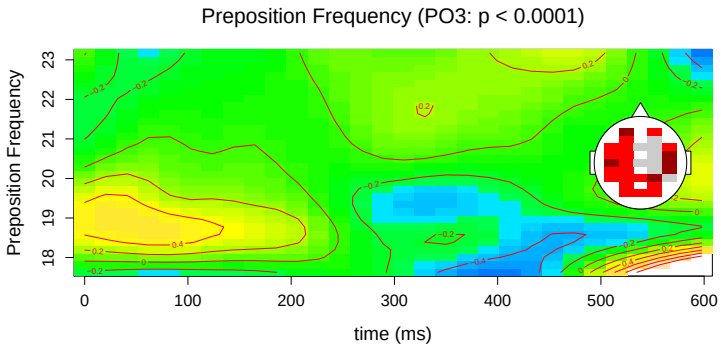

Figure 6. Effect for the tensor product interaction between time and (*log*) *Preposition Frequency* at electrode textitPO3.

After about 300 ms, the effect of *Preposition Frequency* reverses, with lower voltages for low frequency prepositions as compared to high frequency prepositions starting from 300 ms after picture onset. The effect of *Preposition Frequency* is topographically widespread, but more prominent in the left hemisphere than in the right hemisphere. The greatest effect sizes, however, were observed at left-lateralized parietal electrodes and bilateral occipital electrodes.

As for *Picture Complexity*, the results for the main effect smooth of *Preposition Frequency* showed little evidence for a *Preposition Frequency* effect over time. As can be seen in Figure 7, we found an effect at 2 electrodes at a non-corrected alpha level only, with slightly higher voltages for high frequency prepositions than for low frequency prepositions. As such, the effect of *Preposition Frequency* is much better described by a time by predictor interaction than by a main effect smooth.

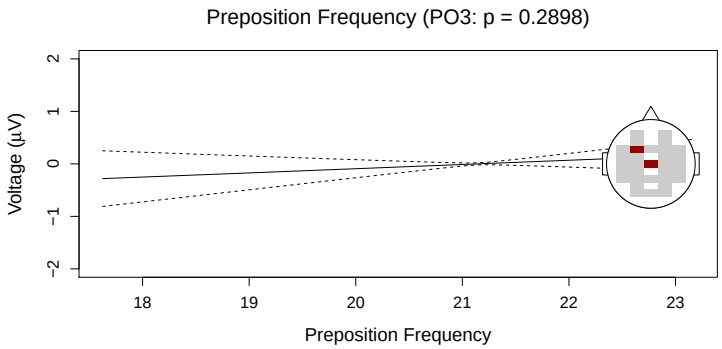

Figure 7. Effect for the main effect smooth of (*log*) *Preposition Frequency* over time at electrode *PO3*.

Word Frequency

Figure 8 shows the results for the time by *Word Frequency* tensor product interaction. The effect is characterized by oscillations for both high and low frequency words that are in opposite phrase and that reach maximum spectral intensity at 3 *Hz*. As such, these oscillations can be characterized as oscillations near the lower edge of the theta range. 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, 2008), syntactic processing (Bastiaansen et al., 2002) and translation (Grabner et al., 2007). In a regression study using GAMs, (Kryuchkova et al., 2011) 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).

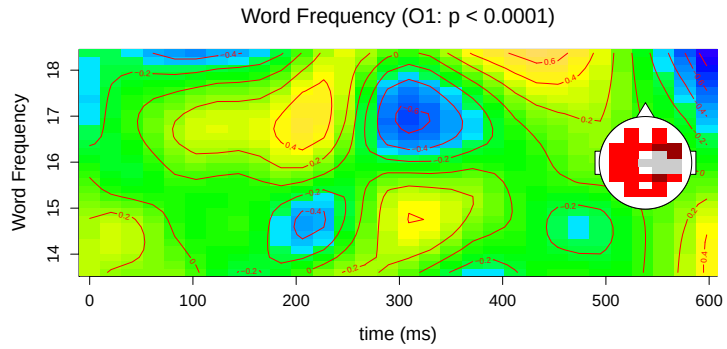

Figure 8. Effect for the tensor product interaction between time and (*log*) *Word Frequency* at electrode *O1*.

The effect of *Word Frequency* arises early. It is first significant at 95 ms after picture onset for medium to high predictor values. The early onset of the frequency effect for high frequency words is in line with previous findings (Hauk et al., 2006; Penolazzi et al., 2007; Sereno et al., 1998), reporting effects of lexical frequency in visual word recognition starting between 110 and 132 ms after word onset. The oscillations for low frequency words are somewhat more subtle in nature than those for high frequency words, with smaller amplitudes and a later onset (these oscillations first reach significance at 183 ms after picture onset).

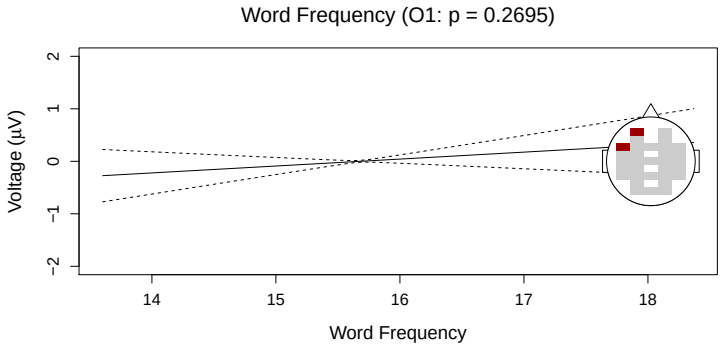

Figure 9. Effect for the main effect smooth of (*log*) *Word Frequency* over time at electrode *O1*.

The time by *Word Frequency* tensor product is significant at a large number of electrodes, with robust effects across frontal-to-occipital electrodes in the left hemisphere. By contrast, we found little to no evidence for a main effect of *Word Frequency* over time. Figure 9 shows that the main effect smooth for *Word Frequency* was significant at a non-corrected alpha level at 2 of the most frontal electrodes only. At these electrodes, we observed a small increase in voltages for higher values of *Word Frequency*, similar to the non-significant effect depicted in Figure 9 for electrode *O1*. As for the effect of *Preposition Frequency*, therefore, the effect of *Word Frequency* is much better described by a time by predictor interaction than by a main effect smooth.

Phrase Frequency

Figure 11 shows the tensor product interaction of time by *Phrase Frequency*. At first glance, it seems like there is a strong early positivity for high frequency phrases and a less pronounced early negativity for low frequency phrases, followed by a reversal of this patterns, with later negative voltages for high frequency phrases and positive voltages for low frequency phrases.

The main effect smooth of *Phrase Frequency*, however, reveals further insight into the tensor product interaction of time by *Phrase Frequency*. This main effect is presented in Figure 11. In contrast to *Preposition Frequency* and *Word Frequency*, *Phrase Frequency* shows a statistically robust main effect over time, with lower voltages for high frequency phrases as compared to low frequency phrases. The effect is present at electrodes across the left hemisphere and is most prominent in left-lateralized parietal and occipital areas.

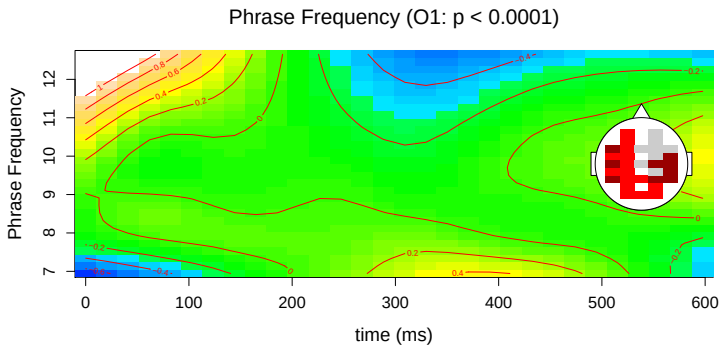

Figure 10. Effect for the tensor product interaction between time and (*log*) *Phrase Frequency* at electrode *O1*.

As can be seen in Figures 10 and 11, the pattern of results for the time by *Phrase Frequency* interaction at the start of the analysis window is opposite to the main effect of *Phrase Frequency* over time, such that the main effect of *Phrase Frequency* is initially cancelled out by the time by *Phrase Frequency* interaction. To illustrate this point, Figure 12 presents]the additive contour surface for the main effect of *Phrase Frequency* (Figure 11) and the tensor product interaction between time and *Phrase Frequency* (Figure 10).

Figure 12 shows that the effect of *Phrase Frequency* is best characterized as a near-linear main effect over time, with more positive voltages for low frequency phrases and more positive voltages for high frequency phrases. This effect arises somewhat earlier for low frequency phrases than for high frequency phrases and continues throughout the 600 ms analysis window. As such, the effect of *Phrase Frequency* seems to be qualitatively different from the effect of *Word Frequency*, which was characterized by theta range oscillations, rather than prolonged effects over time.

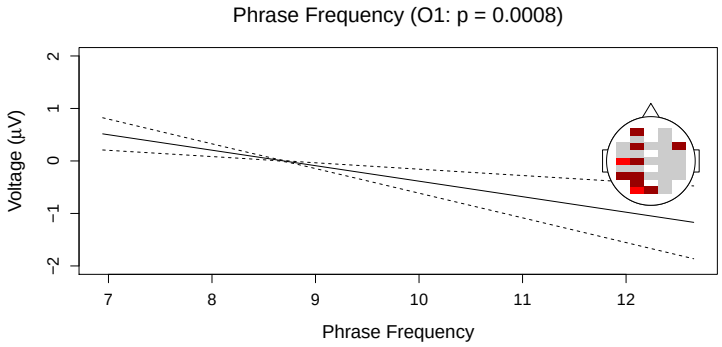

Figure 11. Effect for the main effect smooth of (*log*) *Phrase Frequency* over time at electrode *O1*.

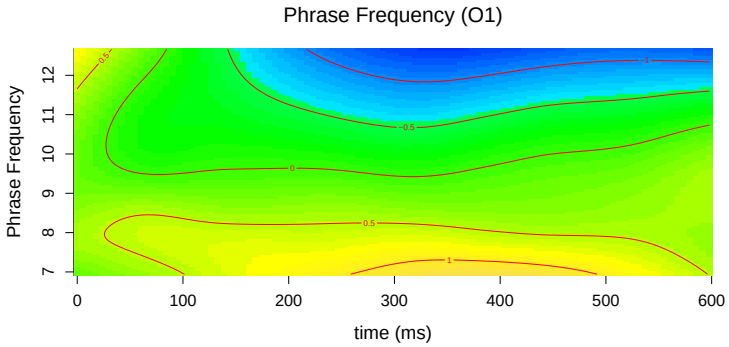

Figure 12. Additive contour surface for the tensor product interaction between time and (*log*) *Phrase Frequency* (Figure 10) and the main effect of (*log*) *Phrase Frequency* over time (Figure 11) at electrode *O1*.

Relative Entropy

Figure 13 presents the tensor product interaction of time by *Relative Entropy*. Similar to the effect of *Word Frequency*, the effect of *Relative Entropy* is characterized by theta range oscillations (4 *Hz*). These oscillations are most prominent high values of *Relative Entropy*, although opposite-phase oscillations with a lower amplitude are present for medium-to-low values of *Relative Entropy* as well.

The effect of the tensor product interaction of time by *Relative Entropy* is topographically widespread, with significant effects across the left - and to a lesser extent - the right hemisphere. The effect is most prominent at parietal and occipital electrodes. For high values of *Relative Entropy*, the effect is first significant at 95 ms after picture onset, whereas for medium-to-low values of *Relative Entropy* the effect first reaches significance at 104 ms after picture onset. As such, the temporal onset of the *Relative Entropy* effect is highly similar to that of the *Word Frequency* effect.

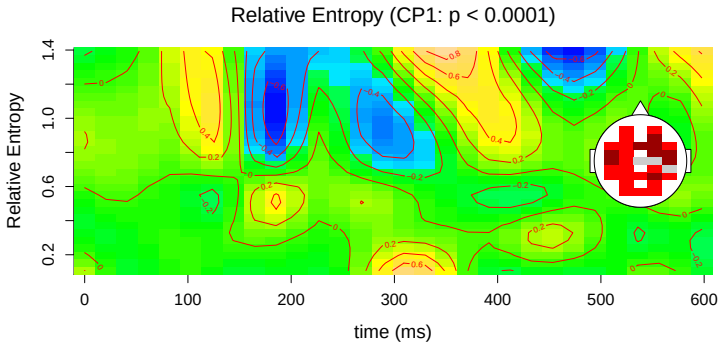

Figure 13. Effect for the tensor product interaction between time and *Relative Entropy* at electrode *CP1*.

Reaction time studies reported increased response latencies for words with high relative entropies (Milin, Filipović Durđević, & Moscoso del Prado Martín, 2009; Milin, Kuperman, et al., 2009; Kuperman et al., 2010; Baayen et al., 2011). The current pattern of results fits well with these findings if we interpret the increased amplitude of the oscillations for high values of *Relative Entropy* as evidence for increased processing costs. The current results then 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.

For completeness, we conclude with the main effect smooth of *Relative Entropy*. As can be seen in Figure 14, we found little evidence for an effect of *Relative Entropy* over time. An effect at a non-corrected alpha level was found at 2 electrodes only, with somewhat decreased voltages for higher values of *Relative Entropy*. As for the effects of *Preposition Frequency* and *Word Frequency*, however, it is clear that the effect of *Relative Entropy* is best described by a tensor product interaction of time by *Relative Entropy*.

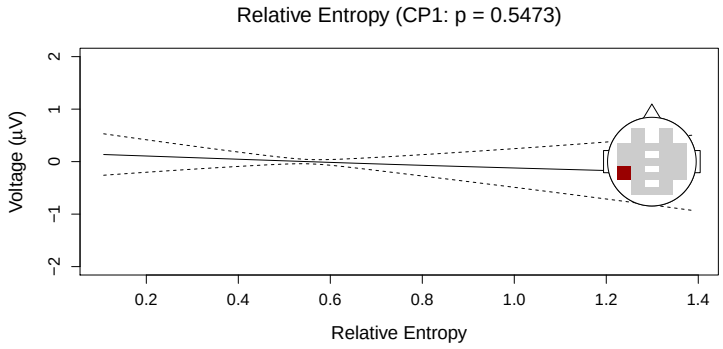

Figure 14. Effect for the main effect smooth of *Relative Entropy* over time at electrode *CP1*.

Discussion

In the current experiment, we observed effects of both word-level and phrase-level predictors in a primed picture naming paradigm. The effects of *Relative Entropy* and *Word Frequency* showed remarkable similarities. Both effects are characterized by oscillations in the lower end of the theta range. In addition, both effects showed similar topographical distributions and increased effect sizes in the left hemisphere as compared to the right hemisphere. Furthermore, the temporal onset of the effects was similar, with the onset of both effects being no more than 2 ms apart (*Word Frequency*: 97 ms after picture onset, *Relative Entropy*: 95 ms after picture onset). Neither *Word Frequency*, nor *Relative Entropy* showed a statistically robust main effect over time.

Similar to the effects of the word-level predictors *Word Frequency* and *Relative Entropy*, the effect for the phrase-level predictor *Phrase Frequency* was most prominent in the left hemisphere. In contrast to the effects of these word-level predictors, however, the effect for *Phrase Frequency* was not characterized by theta range oscillations. Instead, we observed a prolonged near-linear effect, with more negative voltages for high frequency phrases as

compared to low frequency phrases. 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 (see Baayen et al., 2013). 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 a discrimination learning model can provide further insight into the ERP signal in the current primed picture naming study as it evolves over time. First, we will introduce naive discrimination learning model in more detail. Next, we will describe a simulation study in which we used four measures derived from two discrimination learning networks to predict the ERP signal after picture onset. Finally, we will present the results of this simulation study for each of these four discrimination learning measures.

Naive Discrimination Learning

In this section we will describe Naive Discrimination Learning (NDL) as implemented in Baayen et al. (2011). The description below is a shortened version of the more detailed descriptions in Baayen et al. (2011) and Baayen et al. (2013). For more details we refer the interested reader to these papers. NDL networks learn associations between input cues and outcomes 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 $t+1$:

$$V_i^{t+1} = V_i^t + \Delta V_i^t. \tag{2}$$

with the change in association strength, ΔV_i^t, defined as:

$$\Delta V_i^t = \begin{cases} 0 & \text{if } \textsc{absent}(C_i, t) \\ \alpha_i\beta_1 \left(\lambda - \sum_{\textsc{present}(C_j,\, t)} V_j\right) & \text{if } \textsc{present}(C_j, t) \;\&\; \textsc{present}(O, t) \\ \alpha_i\beta_2 \left(0 - \sum_{\textsc{present}(C_j,\, t)} V_j\right) & \text{if } \textsc{present}(C_j, t) \;\&\; \textsc{absent}(O, t) \end{cases}. \tag{3}$$

The parameter settings in the NDR default to $\lambda = 1$, all α's equal, and $\beta 1 = \beta 2$. The association strength 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 NDL framework 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 $t+1$. In the Danks (2003) equilibrium equations the association strength (V_ik) between cue (C_i) and outcome (O_k) is defined as:

$$\Pr(O_k|C_i) - \sum_{j=0}^{n} \Pr(C_j|C_i) V_{jk} = 0, \tag{4}$$

where $\Pr(C_j|C_i)$ is the conditional probability of cue C_j given cue C_i, $\Pr(O_k|C_i)$ is the conditional probability of outcome O_k given cue C_i and $n+1$ is the number of different cues. As shown in Equation 4, 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.

At a given point in time, only a subset of all cues is present in the input. The extent to which these cues activate the target outcome is a measure of how hard it is to access that target in the context of the current input features. The NDR defines the activation of the target outcome O_k given the input cues C as:

$$a_k = \sum_{j \in C} V_{jk}. \tag{5}$$

where j ranges over the active cues and V_{jk} is the equilibrium association strength between cue C_j and outcome O_k.

The NDR network in Baayen et al. (2011) maps orthographic units onto lexemes. As such, this network provides a model of silent reading. The task in the current experiment, however, involves much more than silent reading. The orthographic presentation of the preposition and definite article is line with the nature of the orthography-to-lexeme network in Baayen et al. (2011). By contrast, the target noun 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 has successfully been implemented thus far. 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 over-simplification 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 orthography-to-lexeme network in Baayen et al. (2011) concerns the nature of the task. While the orthography-to-lexeme network provides a silent reading model, the task in the current experiment involves naming the target noun. Recently, Hendrix, Ramscar, and Baayen (2015) implemented the NDR$_a$ model, an extension of the original NDR model in Baayen et al. (2011) for reading aloud. The NDR$_a$ consists of two networks: a network mapping orthographic cues onto lexemes and a network mapping lexemes 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 a simple orthography-to-lexeme network in the current simulation for two reasons. First, the current task is a somewhat of a hybrid between production and comprehension. At the word level, the task very much resembles a reading aloud task, albeit with visual rather than orthographic input. At the phrase level, however, no overt response is 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 $2,524$ monosyllabic words used by Hendrix, Ramscar, and Baayen (2015) for instance, the (log and inverse transformed) activation of the target word meaning from the orthography is highly correlated with the (log and inverse transformed) activation of the target word meaning from the phonology ($r = 0.48$, $p < 0.001$). Before using a more complex model architecture, it is therefore useful to see how much explanatory power a simple orthography-to-lexeme network can provide for the current data.

While we decided not to train a network mapping acoustic features onto lexeme, we did expand the architecture of the original NDR model with a different type of additional network in the current simulations. To gauge contextual learning at the lexeme level, Hendrix, Nick, and Baayen (2015) recently used a discrimination learning network in which both the input cues and the outcomes are lexemes. As for the orthography-to-lexeme network, they trained this lexeme-to-lexeme network on word trigrams. In the lexeme-to-lexeme network, however,

[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.

the cues were words n-2 and n-1, whereas the outcome was word n. They found that the lexeme-to-lexeme network provided explanatory value over and above and orthography-to-lexeme network for the eye movement patterns on compounds in natural discourse reading. In the simulations reported below, we therefore use a set of predictors derived from both orthography-to-lexeme and lexeme-to-lexeme discrimination learning networks.

NDR simulation

In order to learn the associations between input cues and outcomes discrimination learning networks need to be trained on a representative language sample. Following Baayen et al. (2011) we trained both the orthography-to-lexeme and lexeme-to-lexeme networks 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. For the orthography-to-lexeme network the input cues were the letter trigrams and the outcomes were lexemes. For the lexeme-to-lexeme network, the input cues were lexemes n-2 and n-1 in a word trigram and the outcome was lexeme n.

We extracted three systemic measures of language processing from the orthography-to-lexeme network. These three measures are the activation of (1) the preposition, (2) the definite article and (3) the target noun given the presentation of the preposition, the definite article and the target noun. We obtained these simulated activations for all of the 272 phrases that were used in the experiment by summing the associations between all letter trigrams in the input phrase and the preposition, the definite article and the target noun (see Equation 5. For the example phrase "into the onion", for instance, we calculated the simulated activation of the target noun "onion" by summing the associations between the letter trigrams *#in*, *int*, *nto*, *to#*, *o#t*, *#th*, *the*, *he#*, *e#o*, *#on*, *oni*, *nio*, *ion* and *on#* (hash marks indicate word boundaries) and the lexeme *ONION*. Similarly, the simulated activations of the preposition "into" and the definite article "the" were defined as the summed association between these letter trigrams and the lexemes *INTO* and *THE*, respectively.

The simulation activations for the preposition, determiner and target noun will henceforth be referred to as *NDL Activation Preposition*, *NDL Activation Determiner* and *NDL Activation Word*. Following Baayen et al. (2011), we applied an inverse and logarithmic transformation to all activations prior to analysis to remove a rightward skew from the data. As such, the activation measures are proportional to the estimated time required for accessing a lexeme. Furthermore, we added a back off constant of 0.05 to all activations to prevent division by zero when applying the inverse transformation.

From the lexeme-to-lexeme network, we derived a more general systemic property of the target word lexeme. For each of the 68 target nouns that were used in the experiment, we extracted the association between the corresponding lexeme and all other lexemes in the training lexicon. We then calculated the median absolute deviation of the resulting vector of associations for each target word lexeme.

The median absolute deviation (henceforth MAD) is a measure of dispersion that is more robust to outliers than the standard deviation. Recently, Milin et al. (2015) and Hendrix, Nick, and Baayen (2015) successfully applied the MAD measure in the context of discrimination learning and described it as a measure of network connectivity: the greater the MAD of a lexeme, the greater its network connectivity and the easier it is to access that lexeme. As such, one could think of the MAD measure as a systemically motivated account

Table 2: Summary of the independent variables (*log*) *Picture Complexity*, (log and inverse transformed) *NDL Activation Preposition*, (log and inverse transformed) *NDL Activation Determiner*, (log and inverse transformed) *NDL Activation Word* and (*log*) *NDL MAD*. Range is the original range of the predictors. Adjusted range is the range after removing predictor outliers. Mean, median and sd are the means, medians and standard deviations after outlier removal.

predictor	range	adjusted range	mean	median	sd
Picture Complexity	8.53 - 11.13	8.69 - 10.83	9.88	9.91	0.50
NDL Activation Preposition	−0.65 - −1.70	−0.19 - 0.58	0.04	0.00	0.13
NDL Activation Determiner	−0.20 - 0.17	−0.12 - 0.04	−0.04	−0.04	0.02
NDL Activation Word	0.00 - 2.88	0.13 - 2.88	1.62	1.83	0.76
NDL MAD	−15.12 - −8.88	−14.57 - −9.56	−12.07	−12.15	1.26

of frequency effects. The greater the frequency of a lexeme, the better a discrimination learning network is able to learn which lexemes are positively or negatively associated with that lexeme (and therefore the greater the MAD). Indeed, MAD typically shows a correlation equal to or greater than 0.90 with word frequency. We will henceforth refer to the MAD measure as NDLMAD. We log-transformed *NDL MAD* prior to analysis to remove a rightward skew from the MAD distribution.

As for the lexical predictor analysis, we removed predictor outliers further than two standard deviations from the mean from the data prior to analysis. As such, we excluded 1.54% of predictor values for *NDL Activation Word*, 5.00% of all predictor values for *NDL Activation Determiner*, 6.92% of all predictor values for *NDL Activation Preposition* and 4.62% of all predictor values for *NDL MAD*. Table 2 shows the range, adjusted range, mean, median and standard deviation for all NDL predictors.

As for the lexical predictor data set, the NDL predictors are characterized by a considerable amount of collinearity ($\kappa = 59.97$). Most notably, there is a medium strength correlation between *NDL MAD* and *NDL Activation Word* ($r = 0.52$). Nonetheless, suppression is unlikely given the strength of this correlation. As for the lexical predictor analysis, we therefore decided not to decorrelate the NDL predictors.

Analogous to the analysis for the lexical predictors, we fitted a GAM with by-participant factor smooths for trial and time, as well as random intercepts for prepositional phrase and noun to the ERP signal at each electrode. In addition, we included a main effect smooth as well as a tensor product interaction between time and predictor for each of the predictors *Picture Complexity*, *NDL Activation Preposition*, *NDL Activation Determiner*, *NDL Activation Word* and *NDL MAD*. As before, non-linearities in the predictor dimension were limited to 5 knots, whereas non-linearities in the time dimension were limited to 20 knots. Again, we set the autocorrelation parameter ρ to 0.75 to control for AR1 autocorrelation.

Simulation Results

In this section, we will present the results for the predictors *NDL Activation Preposition*, *NDL Activation Determiner*, *NDL Activation Word* and *NDL MAD*. The effect of *Picture Complexity* was highly similar to that reported in the lexical predictor analysis and

is therefore not repeated below.

NDL Activation Preposition

Figure 15 shows the contour plot of the tensor surface for *NDL Activation Preposition* at electrode *P3*. The effect of *NDL Activation Preposition* is characterized by a positivity for prepositions with high (log and inverse transformed) activation values in the first 200 ms after picture onset, which is followed by a negativity for the same prepositions. This effect is highly significant across the left hemisphere, but shows peak amplitudes at left and central parietal and occipital electrodes.

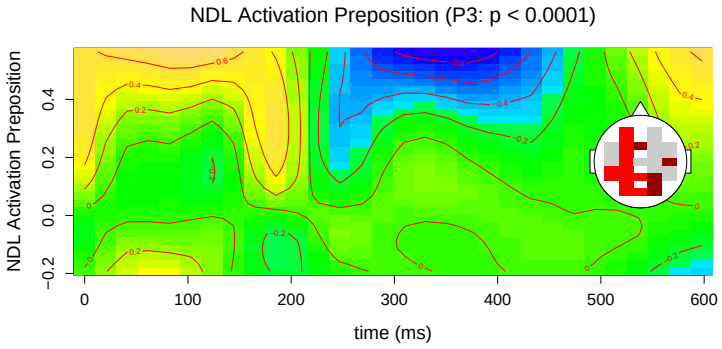

Figure 15. Effect for the tensor product interaction between time and (log and inverse transformed) *NDL Activation Preposition* at electrode *P3*.

Given that log and inverse transformed *NDL* activations are proportional to (simulated) naming latencies, whereas frequency measures are inversely proportional to naming latencies, the effect of *NDL Activation Preposition* is qualitatively and topographically similar to the effect of *Preposition Frequency* described for the lexical predictor analysis. Both predictors show positivities in the first 200 ms followed by negativities at later points in time for predictor values for which longer naming latencies are expected (i.e.; high activation, low frequency). In both cases, the effect is present across the left hemisphere, but is most prominent in left-central parietal-occipital areas. The similarity of the effects for *Preposition Frequency* and *NDL Activation Preposition* is unsurprising given the correlation between both predictors ($r = -0.60$).

Consistent with absence of a main effect of *Preposition Frequency*, we found little evidence for a significant main effect smooth for *NDL Activation Preposition* in the left hemisphere. In contrast to the main effect of *Preposition Frequency*, however, the main effect of *NDL Activation Preposition* did reach significance in frontal and frontal-central areas in the right hemisphere. Figure 16 shows the main effect of *NDL Activation Preposition* over time at electrode *F4*, with more positive voltages for predictor values that correspond expected processing difficulties (i.e.; long simulated naming latencies). The main effect of *NDL Activation Preposition* is theoretically interesting in the light of the *n*-gram frequency effect reported for the lexical predictor analysis. We will return to this issue in the Discussion of the NDL Simulation.

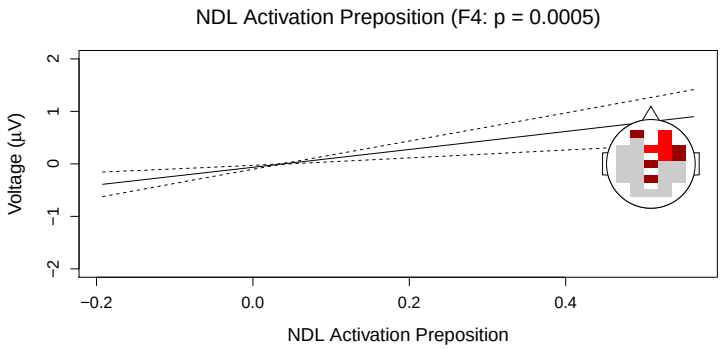

Figure 16. Effect for the main effect smooth of (log and inverse transformed) *NDL Activation Preposition* over time at electrode *P3*.

NDL Activation Determiner

Figure 17 presents the time by predictor tensor product interaction for *NDL Activation Determiner*. This effect is characterized by a complicated pattern of oscillatory activity in both the time and predictor dimensions. For a substantial number of time values, the effect seems to be mirrored with respect to the middle of the *NDL Activation Determiner* range. We see a concave effect in the predictor dimension that starts around 80 ms after picture onset and returns from 220 to 300 milliseconds. After that, the effect reverses, with a convex effect of *NDL Activation Determiner* from 320 ms onwards. This effect is most prominent in left and central parietal-occipital areas, but reaches significance across the left hemisphere.

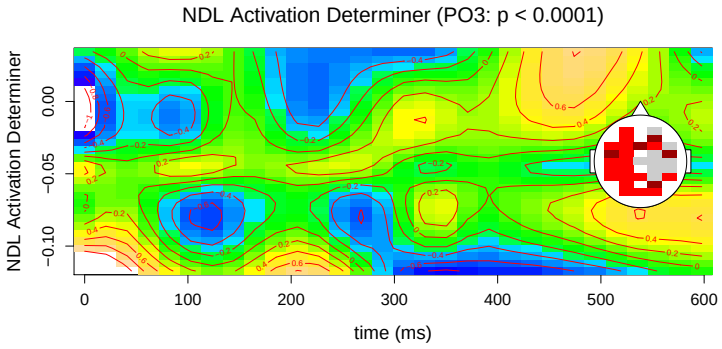

Figure 17. Effect for the tensor product interaction between time and (log and inverse transformed) *NDL Activation Determiner* at electrode *PO3*.

Given the fact that the determiner is identical in all stimuli, the presence of a statistically robust time by *NDL Activation Determiner* tensor product interaction with a relatively large effect size may seem surprising at first sight. Note, however, that *NDL Activation Determiner* was defined as the activation of the determiner lexeme given the orthographic cues in the preposition, the determiner and the target noun. As such, the current effect s uggests the context in which a determiner appears has considerable influence

on how that determiner is processed.

In addition to the time by predictor tensor product interaction, we also found evidence a main effect of *NDL Activation Determiner*. As can be seen in Figure 18, we found higher voltages for higher values of *NDL Activiation Determiner* at left and central parietal-occipital electrodes. Given that high values of (inverse-transformed) *NDL Activation Determiner* are expected to correspond to increased processing difficulty, this effect of *NDL Activation Determiner* over time is qualitatively similar to the effect of *Phrase Frequency* described in the lexical predictor analysis, which was characterized by a similar positivity over time for low frequency phrases.

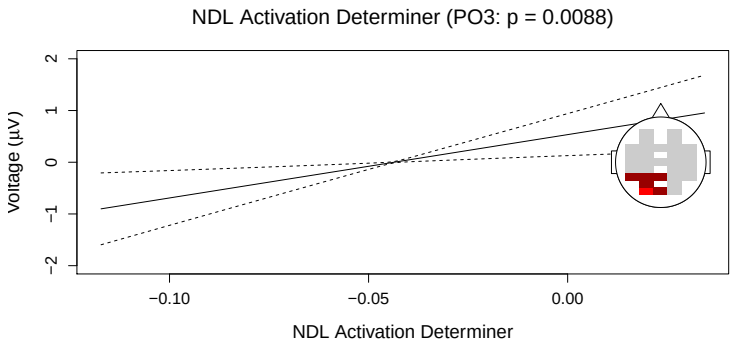

Figure 18. Effect for the main effect smooth of (log and inverse transformed) *NDL Activation Determiner* over time at electrode *PO3*.

NDL Activation Word

The time by predictor tensor product interaction for *NDL Activation Word* at example electrode *FC1* is presented in Figure 19. The effect is characterized by oscillations near the lower edge of the theta range (3 *Hz*) across the *NDL Activation Word range*. The oscillations are most prominent for high predictor values, but are also present for lower predictor values. The effect of *NDL Activation Word* is topographically widespread, with significant time by *NDL Activation Word* tensor product interactions across the scalp. Peak amplitudes, however, are reached in frontal and central areas in the left hemisphere and parietal and occipital areas in the right hemisphere. The effect of *NDL Activation Word* first reaches significance at 149 ms after picture onset for medium values of *NDL Activation Word*.

The time by *NDL Activation Word* interaction shows some similarities with the time by *Word Frequency* interaction described earlier. The effect is topographically widespread and characterized by oscillations in the lower part of the theta range. The onset of the effect, however, is later than that of the Word Frequency effect, which was first significant at 97 ms after picture onset. Furthermore, *NDL Activation Word* shows clear non-linearities in the predictor dimension. By contrast, the time by *Word Frequency* interaction was mostly characterized by simple linear effects in the predictor dimension with alternating positive and negatives slopes. The moderate similarities between the effect of *NDL Activation Word* and the effect of *Word Frequency* are in line with the moderate correlation between both predictors (r = -0.41).

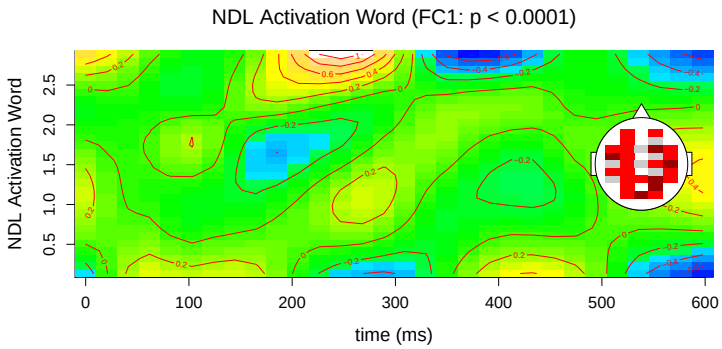

Figure 19. Effect for the tensor product interaction between time and (log and inverse transformed) *NDL Activation Word* at electrode *FC1*.

For *Word Frequency* we found little evidence for a main effect over time. Although the effect did not reach significance at the example electrode *FC1* discussed here (see Figure 20), we did find a subtle main effect of *NDL Activation Word* at a non-corrected alpha level at 6 electrodes located in bilateral frontal areas, with more positive voltages for high predictor values (i.e.; for words with longer expected naming latencies).

The main effect of *NDL Activation Word* did not reach significance at a Bonferroni-corrected alpha level. In addition, the electrodes at which we saw significant effects at a non-corrected alpha level were limited to frontal electrodes. Given the increased RMS values at these electrodes these effects need to be interpreted with care. As such, any strong conclusions regarding the main effect of *NDL Activation Word* would be unwarranted. Nonetheless, the current results provide somewhat more evidence for a main effect over time for *NDL Activation Word* as compared to *Word Frequency*. We will return to this issue shortly in the Discussion section of the NDL Simulation.

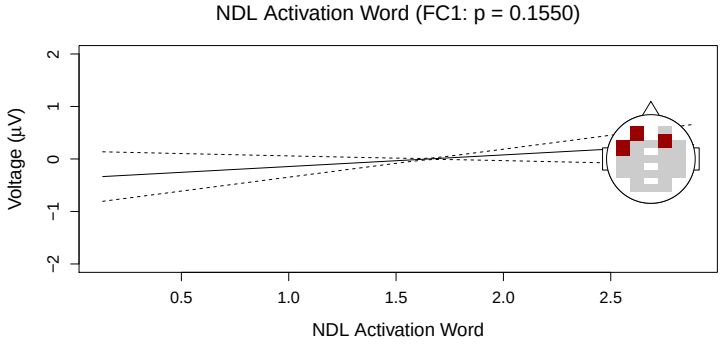

Figure 20. Effect for the main effect smooth of (log and inverse transformed) *NDL Activation Word* over time at electrode *FC1*.

NDL MAD

NDL MAD is a measure of the network connectivity of a word that is perhaps best perceived of as a systemic alternative to word frequency measures. Indeed, *NDL MAD* correlates much more strongly with *Word Frequency* ($r = 0.90$) than does *NDL Activation Word* ($r = -0.41$). As such, we would expected the effect of *NDL MAD* to be more similar to the effect of *Word Frequency* than the effect of *NDL Activation Word*. As can be seen in Figure 21, this prediction is borne out.

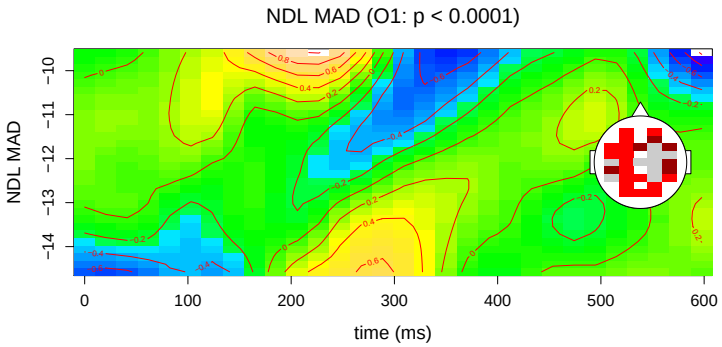

Figure 21. Effect for the tensor product interaction between time and (*log*) *NDL MAD* at electrode *O1*.

The effect of *NDL MAD* is characterized by 3 to 4 *Hz* oscillations for both high and low predictor values. For high values of *NDL MAD* the phase of these oscillations is highly similar to the phase of the oscillations observed for *Word Frequency*. For low predictor values, there is a phase mismatch with the oscillations in the first 250 ms. From 250 to 600 ms after picture onset, however, the phase of the oscillations is highly similar to that of the oscillations for *Word Frequency* once more.

The topographical distribution of the time by *NDL MAD* interaction is similar to that of the time by *Word Frequency* interaction as well, with a widespread effect that is significant across the left hemisphere, as well as in central and right parietal-occipital areas. Furthermore, the effect of *NDL MAD* at high predictor values is first significant at 100 ms after picture onset. As such, the temporal onset of the *NDL MAD* effect is highly similar to that of the *Word Frequency* effect, which was first significant at 97 ms after picture onset. In conclusion, therefore, the effect of *NDL MAD* and *Word Frequency* are qualitatively similar.

For low values of *NDL MAD*, we see an early negativity that is first significant at 20 ms after picture onset. Perhaps, this effect is an artifact due to the unreliability of GAMs near the edges of the analysis window. As such, a negativity for low values of *NDL MAD* around 100 ms after picture onset may incorrectly be present in the first 50 ms of the analysis window as well. Alternatively, the early negativity for low values of *NDL MAD* may be a modification of the main effect smooth for *NDL MAD*, indicating the absence of a main effect right after picture onset.

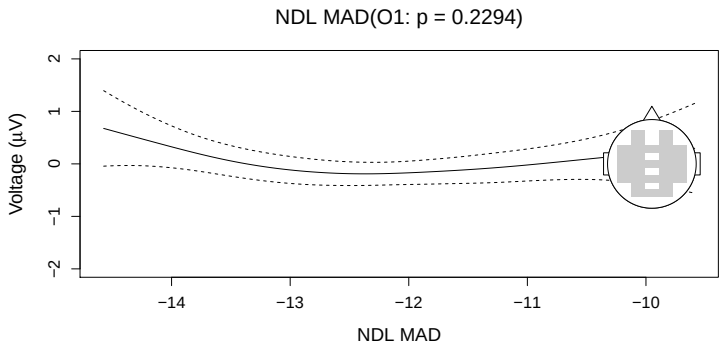

Figure 22. Effect for the main effect smooth of (*log*) *NDL MAD* over time at electrode *O1*.

Figure 22 presents the main effect of *NDL MAD* over time. We found no statistically significant evidence for a main effect of *NDL MAD*, neither at a corrected, nor at an uncorrected alpha level. Nonetheless, Figure 22 suggests that the spurious negativity for low values of *NDL MAD* at the start of the analysis window may indeed be a modification of the main effect estimate. Although the wide intervals for low values of *NDL MAD* suggest there is considerable uncertainty with regard to the main effect of *NDL MAD*, the predicted voltages for low values of *NDL MAD* are positive. As such, the negativity in the lower left of Figure 22 may indicate that to the extent that a positivity may be present for low values of *NDL MAD*, this positivity is not present right after picture onset.

Discussion

The NDL simulation described above demonstrated that the ERP signature of *Preposition Frequency* as it evolves over time can accurately be captured by *NDL Activation Preposition*, with a qualitatively and topographically similar effects for both predictors. The effects of *Word Frequency* and *NDL Activation Word* showed some similarities as well, but the NDL measure that most closely resembled the pattern of results for *Word Freuqency* was *NDL MAD*, a systemic measure of the out-of-context probability of a word. The time by predictor tensor product interactions for both *Word Frequency* and *NDL MAD* showed theta range oscillations with similar phases and similar topographical distributions.[4]

The lexical predictor analysis provided little evidence for main effects of *Preposition Frequency* or *Word Frequency*. By contrast, the effect of *Phrase Frequency* was primarily characterized by a main effect over time, with higher voltages for lower frequency phrases. Interestingly, the NDL activation of the preposition, the determiner and - to a lesser extent - the target noun showed similar main effects over time. Consistent with the higher voltages for low frequency phrases, *NDL Activation Preposition*, *NDL Activation Determiner* and *NDL Activation Word* all showed more positive voltages for higher (log and) inverse-

[4]Given the topographical and temporal overlap of the *Relative Entropy* effect with the effects of *Word Frequency* and *Preposition Frequency*, a direct comparison of *Relative Entropy* with NDL predictors was not possible. Previous findings, however, suggests that the NDL framework is fully compatible with relative entropy effects (see Baayen et al., 2011).

transformed activation values (i.e.; for predictor values that are expected to result in longer naming latencies).[5]

Baayen et al. (2011) successfully simulated the phrase frequency effect in lexical decision in the NDL framework through a simple additive integration of the activations of the component words given the orthographic features in a phrase. While there are some topographical differences between the widespread effect of phrase frequency and the spatially more restricted effects of *NDL Activation Preposition*, *NDL Activation Determiner* and *NDL Activation Word* in the current study, the fact that context-sensitive NDL activation measures of the component words show main effects over time that are qualitatively similar to the main effect of *Phrase Frequency* provides further support for the idea that *n*-gram representations may not be necessary to account for phrase frequency effects.

A final point of interest regarding the NDL simulation reported above is the quantitative performance of the NDL measures as compared to the lexical predictors *Preposition Frequency*, *Word Frequency*, *Relative Entropy* and *Phrase Frequency*. Given the different size of the data sets for both analyses after outlier removal, a direct comparison of the quantitative performance of both models through goodness-of-fit measures was not possible. For the lexical predictor models, we therefore constructed baseline models for the same data set as the original lexical predictor models that had an identical model structure, but that excluded the lexical predictors of interest (*Preposition Frequency*, *Word Frequency*, *Relative Entropy* and *Phrase Frequency*). Similarly, we constructed baseline models for the NDL models that excluded the NDL predictors of interest (*NDL Activation Preposition*, *NDL Activation Determiner*, *NDL Activation Word* and *NDL MAD*). We then looked at the difference in deviance explained between the lexical predictor model and the baseline lexical predictor model, as well as between the NDL model and the baseline NDL model.

Generally speaking, the contribution of both the lexical variables and the NDL measures to the deviance explained by the models was small, with improvements in the overall percentage of deviance explained (i.e.; deviance explained by full model minus deviance explained by baseline model) being substantially smaller than 1%. The average additional percentage of deviance explained across all electrodes was highly similar for the lexical predictor model (0.100%) and the NDL model (0.098%), with a paired t-test on the vectors of additional deviance explained for all electrodes in the lexical predictor and NDL model showing no significant difference ($p = 0.512$). As such, the quantitative performance of the NDL measures is highly competitive with that of standard lexical predictors. The competitive performance is of the NDL measures as compared to the lexical predictor measures is particularly impressive when taking into account the fact that the lexical predictors were derived from the Google *n*-gram corpus, whereas the NDL networks were trained on the much smaller British National Corpus (BNC).

[5]Note that some variation with respect to the reported main effects exists. At electrode *F8*, for instance, the main effect of *Phrase Frequency* shows the opposite pattern of results as compared to the effect reported for example electrode *O1*. For all main effects, we selected example electrodes that give a good impression of the overall nature of the effect. While the main effect of *Phrase Frequency* at electrode *F8* is qualitatively different from the main effect of *Phrase Frequency* at the reported example electrode *O1*, for instance, the other electrodes that show a significant main effect of Phrase Frequency over time (*Fp1*, *F3*, *T7*, *C3*, *P7*, *P3*, *PO3*, *Oz*) show an effect that is qualitatively similar to the reported effect at electrode *O1*.

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 significant time by predictor interactions for the frequency of the preposition and the target word, as well as for the prepositional relative entropy of the target word. For word frequency, we observed oscillations in the time dimension with a frequency near the lower edge of the theta range (3-7.5 Hz) across the left hemisphere, as well as in bilateral occipital-parietal areas. As mentioned above, 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 & Hagoort, 2003) and have previously been observed in a variety of language processing tasks (see, e.g.; Bastiaansen et al., 2005, 2008; Grabner et al., 2007). The effect of target word frequency was first significant at 97 ms after picture onset. This early onset of the word frequency effect is in line with previous studies that established the onset of word frequency effects (Hauk et al., 2006; Penolazzi et al., 2007; Sereno et al., 1998) soon after the 100 ms mark.

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, Filipović Durđević, & Moscoso del Prado Martín, 2009; Milin, Kuperman, et al., 2009; Kuperman et al., 2010; Baayen et al., 2011). The current study is the first to document a relative entropy effect in an ERP study, with oscillations near the lower edge of the theta range that were most prominent in parietal and occipital areas. These oscillations had greater amplitudes for high predictor values as compared to low predictor values. Similar to the reaction time studies mentioned above, therefore, the current results suggest that additional processing is necessary when a noun's use of prepositions is less prototypical. The effect of relative entropy emerged early, showing a significant effect as early as 95 ms after picture onset. The temporal onset of the relative entropy effect is therefore similar to that of word frequency (97 ms after picture onset).

The effect of relative entropy in the current study demonstrates that language users are sensitive to the extent to which the frequency distribution for a given noun's prepositional paradigm differs from the frequency distribution of prepositions 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. 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 effect of word frequency. While the word frequency effect was characterized by oscillations in the time domain, the phrase frequency effect is best described as

a near-linear effect over time with more positive voltages for low frequency phrases and more negative voltages for high frequency phrases. This effect was most prominent in left-lateralized parietal and occipital areas.

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., 2013; Tremblay et al., 2011; Siyanova-Chanturia et al., 2011). Recently, Tremblay and Baayen (2010) documented a phrase frequency effect in an ERP study 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 n-gram frequency effects in chronometric studies are evidence for "some experience-derived knowledge of specific four-word sequences" (Bannard & Matthews, 2008, p.246). These reaction time studies, however, provide little insight into the nature of this knowledge. One possibility is that phrase representations are stored holistically, much like word representations. Such a perspective on n-gram frequency effects fits well with the architecture of exemplar-based models of language processing. The current results, however, argue against 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 interpretation of phrase frequency effects. 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 effect of phrase frequency (Baayen et al., 2013). The second part of this paper presents a simulation study in which we demonstrate that the ERP signatures of context-sensitive word-level discrimination learning measures show remarkable similarities to the effect of phrase frequency.

In this simulation we constructed statistical models similar to those for the lexical predictor analysis. The lexical predictors, however, were replaced by 4 measures derived from discrimination learning networks: 1 measure regarding the network connectivity of the target word, and three measures gauging the amount of bottom-up support for the preposition, the determiner and the target noun given the presence of all three words in the visual input. As expected, the time by predictor interactions for the activation of the preposition and the target noun showed similarities to the time by predictor interactions for preposition frequency and word frequency. Furthermore, the results for the time by predictor interaction for the network connectivity of the target word were similar to those for the time by word frequency interaction in the lexical model.

The lexical predictor analysis showed little evidence for main effects over time tied to preposition frequency or word frequency. By contrast, however, we found statistically robust main effects over time for the activation of the preposition and the determiner, as well as somewhat less robust evidence for a main effect of the activation of the word. These main effects were qualitatively similar to the main effect of phrase frequency, with a linear effect showing higher voltages for predictor values that are typically associated with additional processing costs.

Although the effects of the NDL activation measures have more restricted topographical distributions than the phrase frequency effect, the qualitative similarity of the effects of the NDL activation measures and the effect of phrase frequency suggests that positing phrase level representations to explain the phrase frequency effect is unnecessary. In the NDL framework phrase frequency effects emerge as a result of both learning and presenting words in context. A high frequency phrase such as "all over the place" is read faster than a low frequency phrases such as "all over the city" (examples taken from Arnon & Snider, 2010), because the letters and letter combinations in "all over the place" have become more associated with the lexemes *ALL*, *OVER*, *THE* and *PLACE* than the letters and letter combinations in "all over the city" have become associated with the lexemes *ALL*, *OVER*, *THE* and *CITY* during the learning process (Baayen et al., 2013).

The fact that measures derived from a fully decompositional learning approach shows effects that are remarkably similar to the effect of phrase frequency 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.

The quantitative performance of the NDL measures was highly similar to that of the lexical predictors preposition frequency, word frequency, phrase frequency and relative entropy. As such, discrimination learning offers a highly competitive account of the ERP signal in the current primed picture naming paradigm that is entirely based on systemic estimates of the learnability of lexical items given the properties of the linguistic input space. It is important to note, however, that the NDL framework 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 NDL framework 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., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition*, *122*(3), 292–305.

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

Baayen, R. H., Hendrix, P., & Ramscar, M. (2013). Sidestepping the combinatorial explosion: an explanation of n-gram frequency effects based on naive discriminative learning. *Language and Speech*, *56*(3), 329–347.

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., Tremblay, A., & Hendrix, P. (2015). 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*.

Balota, D., & Chumbley, J. I. (1985). The locus of word frequency effects in the pronunciation task: Lexical access and/or production? *Journal of Memory and Language*, *24*, 89-106.

Balota, D., Cortese, M., Sergent-Marshall, S., Spieler, D., & Yap, M. (2004). Visual word recognition for single-syllable words. *Journal of Experimental Psychology:General*, *133*, 283–316.

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., 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.

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.

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.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral science (3rd ed.).* Mwahaw, NJ: Lawrence Erlbaum Associates.

Dabrowska, E. (2000). From formula to schema: the acquisition of english questions. *Cognitive Linguistics*, *11*, 83–102.

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

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., Zavrel, J., Sloot, K. Van der, & Bosch, A. Van den. (2010). *TiMBL: Tilburg Memory Based Learner Reference Guide. Version 6.3* (Technical Report No. ILK 10-01). Computational Linguistics Tilburg University.

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

Darlington, R. B. (1990). *Regression and linear models.* New York: McGraw-Hill Publishing Company.

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior*, *12*, 627-635.

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.

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

Hauk, O., Davis, M., Ford, M., Pulvermüller, 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., Nick, J., & Baayen, R. (2015). Compound reading in natural discourse contexts. *Manuscript*.

Hendrix, P., Ramscar, M., & Baayen, R. (2015). Ndra: a single route model of reading aloud based on discriminative learning. *Manuscript*.

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

Jared, D. (2002). Spelling-sound consistency and regularity effects in word naming. *Journal of Memory and Language*, *46*, 723–750.

Kryuchkova, T., Tucker, B. V., Wurm, L., & Baayen, R. H. (2011). Danger and usefulness in auditory lexical processing: evidence from electroencephalography. *Under revision for Brain and Language*.

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 effects in letter perception: Part i. an account of the basic findings. *Psychological Review*, *88*, 375-407.

Milin, P., Filipović Durđević, D., & Moscoso del Prado Martín, F. (2009). 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. (2009). 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.

Milin, P., Ramscar, M., Coch, K., Feldman, L., & Baayen, R. H. (2015). *Processing partially and exhaustively decomposable words: an amorphous approach based on discriminative learning.* (Manuscript)

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

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

Penolazzi, B., Hauk, O., & Pulverm§ller, F. (2007). Early lexical access and semantic context integration as revealed by event-related brain potentials. *Biological Psychology*, *74*(3), 374–388.

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*(7), in press.

Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance*, *3*, 1-17.

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. F., & Baayen, R. H. (2013). The subjective frequency of word n-grams. *Psihologija*, *46*, 497–537.

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.

Strijkers, K., A., C., & G., T. (2010). Tracking lexical access in speech production: electrophysiological correlates of word frequency and cognate effects. *Cerebral Cortex*, *20*(4), 912–928.

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

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 inflationist lexicon. *Proceedings of the Annual Meeting of the Linguistics Society of America*.

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.

Wurm, L. H., & Fisicaro, S. A. (2014). What residualizing predictors in regression analysis does (and what it does not do). *Journal of Memory and Language*, *72*, 37–48.

Appendix

We used generalized additive models (GAMs) to analyze the ERP data for the current experiment (Hastie & Tibshirani, 1986; Wood, 2006). 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 simulated data, as well as 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 23 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 23 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.

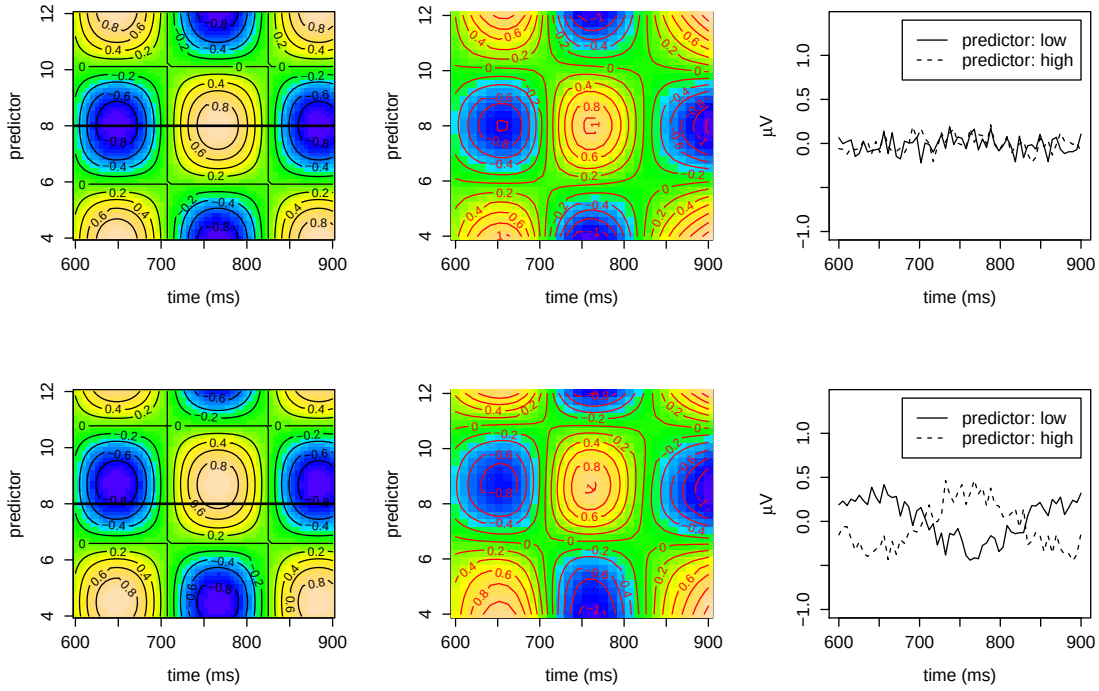

Figure 23. Simulated predictor effect with an oscillation in both the time and predictor dimension (left panels) and model fits for this effect in a GAM analysis (middle panels) and a traditional analysis using predictor dichotomization (right panels).

The top right panel of Figure 23 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.

The simulated data in the top left panel of Figure 23 are symmetrical with respect to the mid-point of the predictor range. For the bottom left panel of Figure 23 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 23 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 23 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.

More subtle examples of the problems associated with the dichotomization of 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 top panel of Figure 24 shows the effect of *Word Frequency* at electrode *O1* in the GAM analysis reported in this paper. The effect is characterized by 3 *Hz* oscillations for both high and low frequency words with opposite phases. The dashed line indicates the mean value of *Word Frequency*. The bottom panel of Figure 24 shows the results of a traditional analysis in which we dichotomized *Word Frequency* into high and low frequency words (split with respect to the mean value of *Word Frequency*). In this analysis we investigated the significance of the dichotomized Word Frequency predictor for each sample point in the time domain by running subject and item ANOVAs on a subset of the data that included all measurements for that sample point, as well as for the previous sample point and the next sample point. The significance of the *Word Frequency* effect in these subject and item analyses is indicated by the dark red ($\alpha = 0.05$) and bright red (Bonferroni-corrected alpha level; $\alpha = 0.0016$) squares at the bottom of the bottom panel of Figure 24.

The grand mean curves for *Word Frequency* show a similar pattern of results as compared to the GAM analysis. The difference between high and low frequency phrases first reaches significance at a non-corrected alpha level at 117 ms after picture onset, with higher voltages for high frequency phrases. As such, the temporal onset of the Word Frequency effect is somewhat later than the temporal onset of the Word Frequency effect in the GAM analysis (97 ms after picture onset), presumably due to a loss of statistical power as a result of the predictor dichotomization. Overall, the pattern of results in the GAM analysis and the dichotomization analysis show a highly similar pattern of results, with higher frequency

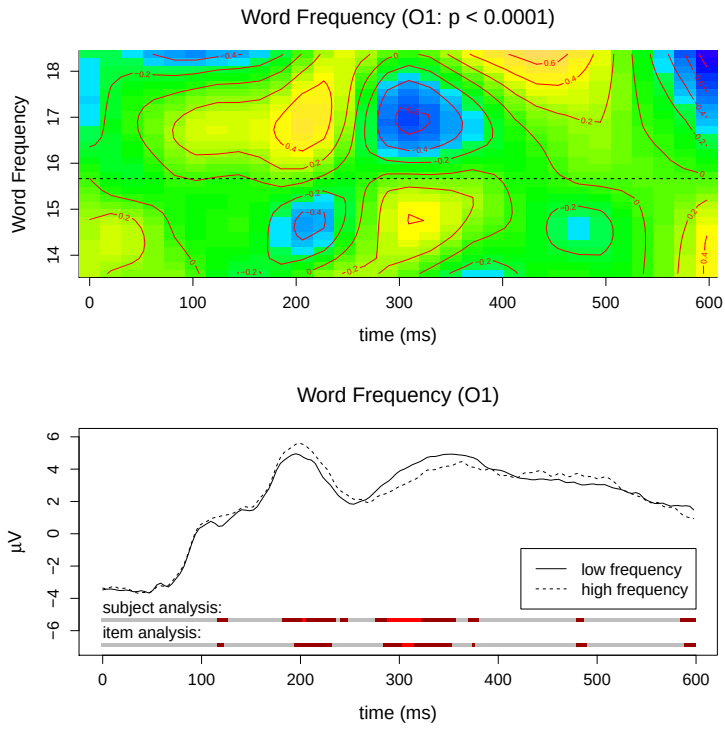

Figure 24. The effect of *Word Frequency* at electrode *O1* in a GAM analysis (top panel) and a traditional analysis in which *Word Frequency* is dichotomized (bottom panel). Color coding at the bottom of the second panel indicates significance of the *Word Frequency* effect in item and subject ANOVAs for each point in time.

words showing higher voltages from 180 to 260 ms after picture onset, lower voltages from 260 to 400 ms and higher voltages once more from 400 to about 530 ms (as compared to lower frequency words). Both in the GAM analysis and in the traditional analysis, the effect of *Word Frequency* is most pronounced from 180 to 400 ms after picture onset.

The comparison of the GAM analysis and the traditional analysis for the *Word Frequency* effect demonstrates that the oscillatory effect of *Word Frequency* 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 a decreased *P350* for high frequency words as compared to 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* represents a somewhat more complicated scenario. The top panel of Figure 25 shows the effect of *Relative Entropy* at electrode *CP1* in the *gam* analysis.

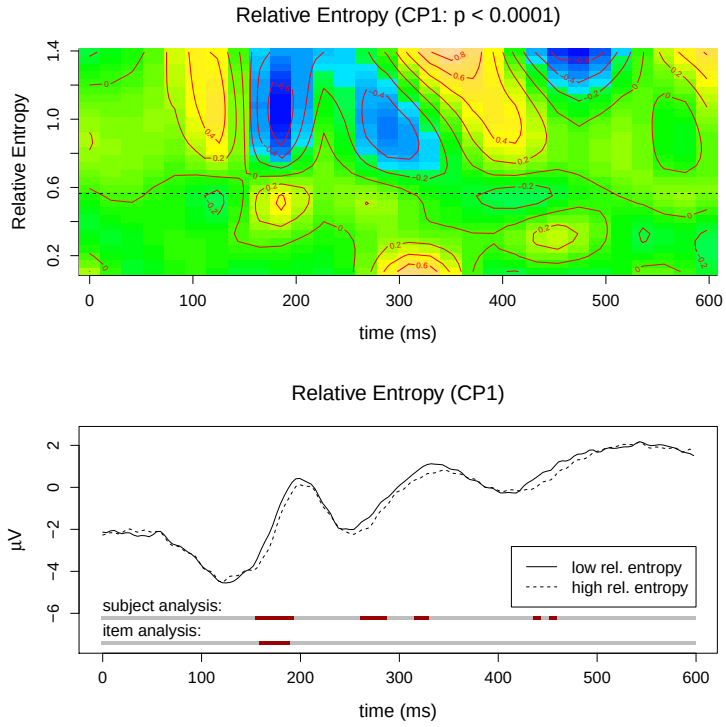

Figure 25. The effect of *Relative Entropy* at electrode *CP1* in a GAM analysis (top panel) and a traditional analysis in which *Relative Entropy* is dichotomized (bottom panel).

As can be seen in the top panel of Figure 25, the effect of *Relative Entropy* is characterized by oscillations in the time that arise around 100 ms after picture onset. The oscillations are most prominent for high predictor values, but lower amplitude oscillations are also present for medium-to-low and low predictor values. To complicate things further, the phase difference between the oscillations for high predictor values and the oscillations for low predictor values is not constant, due to small differences in the frequencies of these oscillations.

The bottom panel of Figure 25 shows the effect of *Relative Entropy* at electrode *CP1* in a traditional ERP analysis in which we dichotomized *Relative Entropy* into high and low relative entropy on the basis of a split at the mean (see black line in the top panel of Figure 25). The grand mean curves for high and low *Relative Entropy* correctly capture the fact that high values of *Relative Entropy* correspond to lower voltages from 150 to 220 ms, from 250 to 340 ms and from 420 to around 500 ms after picture onset (as compared to low values of *Relative Entropy*), although these effects reach significance at non-corrected alpha level only.

The traditional analysis fails to pick up on the more positive voltages for high values of *Relative Entropy* around 100 and 400 ms after picture onset. Potentially, this is due to the fact only *Relative Entropy* was entered into the traditional analysis, whereas the GAM analysis uses a multiple regression approach. As such, the effects of other predictors are

not taken into account in the traditional analysis. The main effect of phrase frequency, for instance, was marginally significant at electrode *CP1*, $p = 0.077$). Given the nature of the phrase frequency effect (i.e.; lower voltages for higher frequency phrases) and the negative correlation between *Relative Entropy* and *Phrase Frequency* ($r = -0.19$), the grand average curve for high values of relative entropy in Figure 25 may be somewhat lower than it would be if the effect of Phrase Frequency was properly accounted for.

Whereas the qualitative nature of the effect of *Word Frequency* was accurately captured by a traditional ERP analysis, a lot of detail is lost about the effect of *Relative Entropy* through dichotomization. While it might be possible to tell that the *Relative Entropy* effect is characterized by theta range oscillations from the bottom panel of Figure 25, for instance, it would be impossible to tell that these oscillations are most prominent for high predictor values. Furthermore, the nature of the effect across the predictor dimension is lost through dichotomization. The information that the effect of *Relative Entropy* is U-shaped in nature around 320 ms, for instance, cannot be retrieved from the bottom panel of Figure 25.

Theta range oscillations in the time dimension characterized he effects of *Word Frequency* and *Relative Entropy*. For *Phrase Frequency*, we found a near-linear effect that persisted over time. The top panel of Figure 26 shows the effect of *Phrase Frequency* at electrode *O1*, with a long-lasting positivity for low frequency words and a long-lasting negativity for high frequency words.

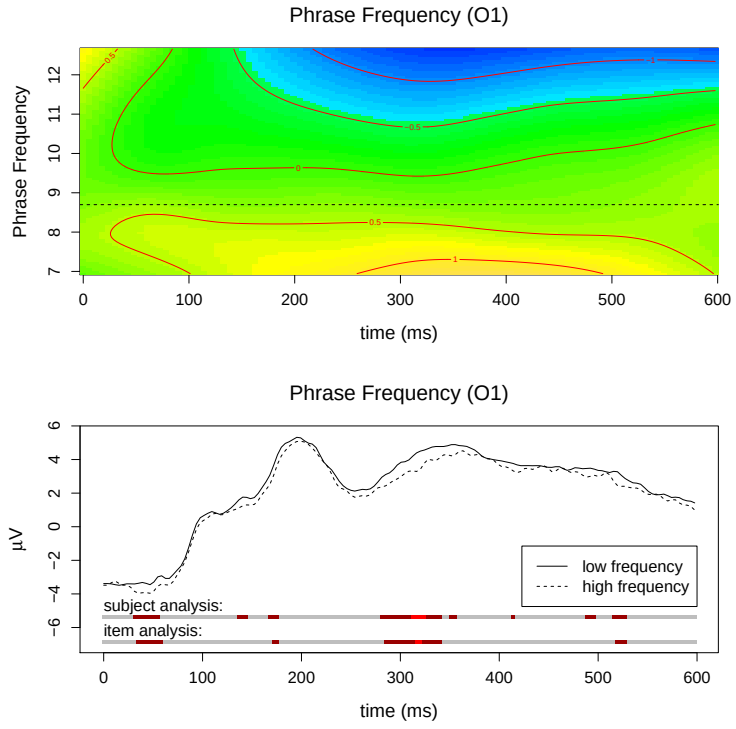

Figure 26. The effect of *Phrase Frequency* at electrode *O1* in a GAM analysis (top panel) and a traditional analysis in which *Phrase Freuqency* is dichotomized (bottom panel).

The bottom panel of Figure 26 shows the results of a traditional ERP analysis in which *Phrase Frequency* was dichotomized with respect to the mean predictor value (see black line in the top panel of Figure 26). The general nature of the *Phrase Frequency* effect is similar to that in the GAM analysis, with more positive voltages for low frequency words as compared to high frequency words over time. Consistent with the top panel of Figure 26, the difference between high and low predictor values is greatest around 300 ms after picture onset, with significant effects in both the item and the subject analysis.

At other points in time, the grand mean curve for high frequency phrases is below that for low frequency phrases as well, but this difference reaches significance for a limited number of sample points at a non-corrected alpha level only. The inability of the subject and item analyses to pick up on the phrase frequency effect throughout the analysis window may be the result of a loss of statistical power in the traditional analysis as compared to the GAM analysis. This loss in statistical power is a consequence of both the dichotomization of phrase frequency and the fact that other parts of the ERP are not properly controlled for in the traditional analysis (e.g.; trial-effects, random effects of subject, preposition, target noun and phrase).

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 typically converge as long as dichotomization of a predictor is relatively unproblematic given the nature of a predictor effect. When this is not the case, such as in our simulation example, the differences that arise between the results from a GAM analysis and a traditional analysis are easily explained given the nature of the predictor effect.

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, especially when other sources of variance in the ERP signal are not (properly) taken into account. 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.