




(Early) context effects on event-related potentials over natural inputs

Shaorong Yan & T. Florian Jaeger

To cite this article: Shaorong Yan & T. Florian Jaeger (2019): (Early) context effects on event-related potentials over natural inputs, Language, Cognition and Neuroscience, DOI: [10.1080/23273798.2019.1597979](https://doi.org/10.1080/23273798.2019.1597979)

To link to this article: <https://doi.org/10.1080/23273798.2019.1597979>

 [View supplementary material](#) 

 Published online: 30 Mar 2019.

 [Submit your article to this journal](#) 

 [View Crossmark data](#) 



(Early) context effects on event-related potentials over natural inputs

Shaorong Yan^a and T. Florian Jaeger^b

^aDepartment of Brain and Cognitive Sciences, University of Rochester, Rochester, NY, USA; ^bDepartment of Brain and Cognitive Sciences, Department of Computer Science, University of Rochester, Rochester, NY, USA

ABSTRACT

Language understanding requires the integration of the input with preceding context. Event-related potentials (ERPs) have contributed significantly to our understanding of what contextual information is accessed and when. Much of this research has, however, been limited to experimenter-designed stimuli with highly atypical lexical and context statistics. This raises questions about the extent to which previous findings generalise to everyday language processing of natural stimuli with typical linguistic statistics. We ask whether context can affect ERPs over natural stimuli early before the N400 time window. We re-analyse a data set of ERPs over ~700 visually presented content words in sentences from English novels. To increase power, we employ trial-level ms-by-ms linear mixed-effects regression simultaneously modelling random variance by subject and by item. To reduce concerns about Type I error inflation common to time series analyses, we introduce a simple approach to model and discount auto-correlations at multiple, empirically determined, time lags. We compare this approach to Bonferroni correction. Planned follow-up analyses employ Generalized Additive Mixed Models to assess the linearity of contextual effects, including lexical surprisal, within the N400 time window. We found that contextual information affects ERPs in both early (~200 ms after word onset) and late (N400) time windows, in line with a cascading, interactive account of lexical access.

ARTICLE HISTORY

Received 18 February 2018
Accepted 21 February 2019

KEYWORDS

Event-related potentials; time course analysis; mixed-effects regression; generalized additive mixture models; auto-correlation


Introduction

During language understanding, comprehenders infer meaning and intentions from the language input guided by the preceding linguistic (and non-linguistic) context. The time course of this information integration has received substantial attention in psycho- and neuro-linguistic research. A large body of research has investigated when – or where in the brain – different types of contextual information are accessed with regard to the onset of an input (for reviews, see Kuperberg, 2016; Van Petten & Luka, 2012). Answers to the questions have informed theories about the architecture of the cognitive and neural systems underlying language comprehension, for example, with regard to the relative degree of information encapsulation between “stages” of language processing (McClelland & Elman, 1986; for discussion, see Norris, McQueen, & Cutler, 2015; Seidenberg & MacDonald, 1999; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995).

Here we employ a paradigm that has been influential in revealing the time course of information access during language understanding: event-related potentials (ERPs) over electroencephalograph recordings (Dambacher, Kliegl, Hofmann, & Jacobs, 2006; Hauk, Davis, Ford,

Pulvermüller, & Marslen-Wilson, 2006; Hauk, Pulvermüller, Ford, Marslen-Wilson, & Davis, 2009; for a recent review, see Laszlo & Federmeier, 2014). Some ERP studies suggest that contextual information affects later time window, e.g. the N400 (a negative-going ERP component that peaks around 400 ms after word onset), but not in earlier time windows (e.g. Dambacher et al., 2006). This would stand in contrast with context-independent lexical information, such as word frequency and form-related predictors, that affects ERPs as early as about 100 ms after word onset (Hauk et al., 2006, 2009; Laszlo & Federmeier, 2014). However, other studies have found that contextual information affects ERPs as early as ~200 ms after word onset (Federmeier, Mai, & Kutas, 2005; Frank & Willems, 2017). This time window is argued to reflect the processing of form-related information, e.g. the orthographic form of a word (Laszlo & Federmeier, 2014). Some studies have even found contextual effects during perceptual processing as early as 100 ms after word onset (Grainger & Holcomb, 2009), although these effects seem to be confined to highly constraining contexts and words that are short (Kim & Lai, 2012; Penolazzi, Hauk, & Pulvermüller, 2007) or frequent (Lee, Liu, & Tsai, 2012). If contextual information

CONTACT Shaorong Yan  syan13@ur.rochester.edu

 Supplemental data for this article can be accessed <http://dx.doi.org/10.1080/23273798.2019.1597979>

© 2019 Informa UK Limited, trading as Taylor & Francis Group

can affect ERPs as early as lexical information can, this can be taken to rule out architectures in which lexical and contextual information are processes at different, strictly serially organised, stages (for review, see e.g. Dell & O'Seaghdha, 1992; Laszlo & Federmeier, 2014; McClelland & Elman, 1986).

However, research on this topic has almost exclusively employed stimuli that are *non-randomly selected* or *intentionally designed* by the experimenter. Such stimuli tend to exhibit linguistic statistics that deviate *strongly* from those observed in natural sentences. This is particularly true for several design properties that are common in ERP research: (1) focusing on only one or two predictors of interest while aiming to hold constant other predictors known to affect ERPs; (2) dichotomising continuous predictors into “bins” of high and low values, e.g. to facilitate factorial designs; (3) using designs in which strong expectations are repeatedly violated; (4) using stimuli with nonce words (e.g. to assess effects of lexicality, Hauk et al., 2006; Laszlo & Federmeier, 2014) or incongruent sentences (e.g. to assess effects of congruency, Payne, Lee, & Federmeier, 2015). Each of these common practices is well motivated and typically serves a purpose. Yet, the almost exclusive reliance of ERP studies on experimenter-designed stimuli comes with potential risks. In particular, it raises questions about the extent to which findings from these works generalise to everyday language processing of natural stimuli with typical statistics (for relevant discussion, see Dambacher et al., 2006; Hauk et al., 2006; Smith & Kutas, 2015b).

One *a priori* reason for further research on this question is provided by studies on implicit adaptation or learning during sentence processing (e.g. Chang, Dell, & Bock, 2006; Fine, Jaeger, Farmer, & Qian, 2013; Kaschak & Glenberg, 2004). There is now mounting evidence that deviation from typical statistics can lead comprehenders to change their processing behaviour, sometimes even within the course of a single experimental session (e.g. Arai & Mazuka, 2014; Creel, Aslin, & Tanenhaus, 2008; Domahs, Klein, Huber, & Domahs, 2013; Fine & Jaeger, 2016; Fine et al., 2013; Kurumada, Brown, & Bibyk, 2014; for recent reviews, see Dell & Chang, 2014). This suggests that experiments with linguistic statistics that strongly deviate from natural text risk confounding processing effects with adaptation effects (for related discussion, see Fine et al., 2013; Jaeger, 2010, pp. 52–54). Indeed, there are examples where strong deviation from typical statistics masks effects that are visible if that deviation is avoided (see Farmer, Monaghan, Misyak, & Christiansen, 2011 in response to Staub, Grant, Clifton, & Rayner, 2009). Similarly, ERP studies have sometimes found changes in effect sizes between earlier and later parts of an

experiment (e.g. Elston-Güttler, Gunter, & Kotz, 2005; Macizo & Herrera, 2011). This can undo effects that are present at the beginning of the experiment (for an example, see Hanulíková, van Alphen, van Goch, & Weber, 2012). Findings like these point to a need for further research on ERPs over natural sentences.

This motivates the present study. Our primary goals are to (1) examine when different types of contextual information come to affect language processing during the reading of natural stimuli, and (2) contribute to the methodological advance of this type of study.

With regard to the first goal, we investigate whether context can affect ERPs as early as non-contextual information can (e.g. lexical frequency). We know of only one ERP study that speaks to the time course of contextual effects in reading over natural stimuli (Frank & Willems, 2017; but see, Broderick, Anderson, Di Liberto, Crosse, & Lalor, 2018 on contextual effects in speech processing). Frank and Willems focused effects of lexical surprisal and semantic association during the N400 time window. Participants read sentences from three novels (UCL corpus of reading times, Frank, Fernandez Monsalve, Thompson, & Vigliocco, 2013). Lexical surprisal and semantic association were estimated from computational language models. Regression analyses found independent effects of the two predictors during the N400 time window. Though not the target of their planned analyses, Frank and Willems also observe evidence for early effects around the P2 window (~200 ms after word onset).

Interestingly, the only other study that has used somewhat comparable stimuli has come to a conflicting conclusion about early effects of contextual predictors. Dambacher et al. (2006) investigated ERPs over a reading corpus with lexically and syntactically heterogeneous sentences (Kliegl, Grabner, Rolfs, & Engbert, 2004). The sentences from this corpus were not sampled from natural text, but rather were experimenter-designed with “the goal to represent a large variety of grammatical structures around a set of target words [...] for which length and frequency are uncorrelated across the sentences” (Kliegl et al., 2004, p. 267). This differs from the sentences employed by Frank and Willems (2017), which were extracted from natural text and thus exhibited strong correlations between word length and frequency (see Figure 1 below). Critically, Dambacher and colleagues did not find reliable contextual effects during the P2 time window, instead finding contextual effects only during the N400 time window.

The two studies differ in a number of other methodological aspects. Crucially, this includes differences in the statistical analyses that are likely to have affected the Type I error rates and power. Specifically, Dambacher and colleagues identified two time windows, the P2

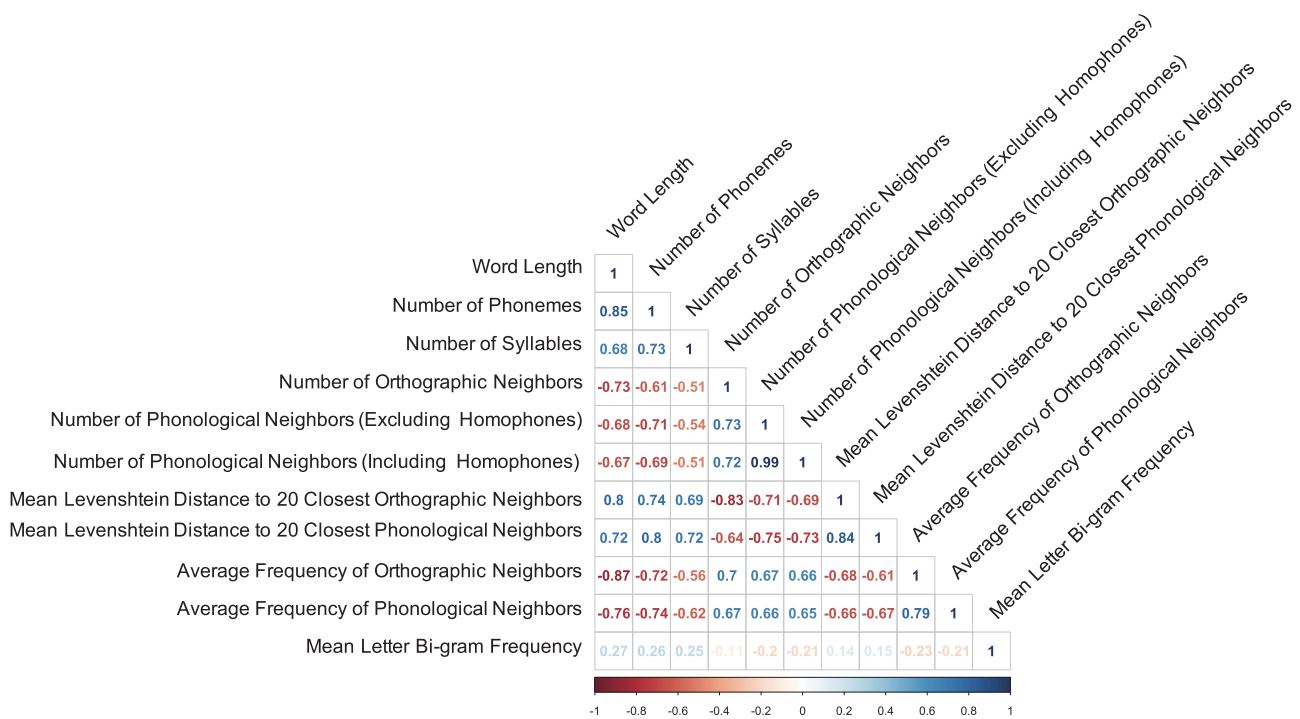


Figure 1. Correlation between word form properties of the critical word tokens that entered analyses.

window (140–200 ms from word onset) and the N400 window (300–500 ms). For each word within each subject, mean EEG amplitudes over each time window were calculated (collapsing across all ERP samples and selected electrodes for the components). The effect of contextual predictability and other factors was estimated separately within each subject, and significance was determined by analysing the distribution of effect sizes across subjects (following Lorch & Myers, 1990). This approach is now outdated, as it unnecessarily discards information (e.g. any uncertainty about the subject-specific effect size), which can reduce statistical power. In contrast, Frank and Willems (2017) employed linear mixed-effects regression with crossed random effects for subjects and items (words) to separately analyze the ERP signal from each time sample (every 4 ms). The use of mixed-effects regression has the advantage that it avoids discarding data, while still recognising the repeated measures structure of the data. However, analysing each ERP sample separately fails to account for auto-correlations between temporally adjacent ERP samples. Such correlations are a widely acknowledged problem (for discussion, see Guthrie & Buchwald, 1991; Piai, Dahlsätt, & Maris, 2015), and failure to correct for them can result in anti-conservativity and inflated Type I error rates.

In other words, Dambacher and colleagues did not find early context effects and had arguably lower power, whereas Frank and Willems found early context

effects but likely had inflated Type I error rates with regard to this test (we note that primary purpose of the study by Frank and Willems was not to assess time course and that the potential problem we focus on here is unlikely to affect the conclusions with regard to their question of interest). The present study seeks to address this issue by introducing a method for modelling and discounting auto-correlations between ERP samples at various, empirically-determined time-lags. By analysing auto-correlation-corrected ERPs, we can maintain all the advantages of the regression-based approach over time series data (see Hauk et al., 2006; Smith & Kutas, 2015a) while avoiding anti-conservativity. Specifically, we present separate mixed-effects analyses over uncorrected and auto-correlation-corrected ERPs from the time series data in Frank and Willems (2017). We compare the two analyses against each other, and against a more general, but potentially conservative (for discussion, see, e.g. Narum, 2006), Type I error correction (Bonferroni). We also compare this type of regression-based time series analysis against the more common approach of analysing aggregate ERPs over specific time windows. Specifically, we use the same two time windows as in Dambacher et al. (2006) and compare the results of this analysis to the time series analysis.

Further contributing to our first goal, we address another important caveat on the interpretation of previous studies on this topic: the lack of controls for word

Table 1. Summary statistics for the properties of the words that entered our analyses (aggregated over word types or token).

Predictor name	Range	by word token		by word type	
		Mean	SD	Mean	SD
Word frequency (relative, log-transformed)	(−12.85, −3.93)	−8.70	1.79	−9.34	1.58
Word length	(2, 10)	4.80	1.57	5.19	1.63
Number of phonemes	(2, 9)	3.86	1.23	4.15	1.35
Number of syllables	(1, 4)	1.30	0.54	1.42	0.61
Number of orthographic neighbours	(0, 34)	9.47	7.53	7.74	7.10
Number of phonological neighbours (excluding homophones)	(0, 58)	19.98	15.38	17.19	14.97
Number of phonological neighbours (including homophones)	(0, 64)	21.18	16.50	18.35	15.88
Mean Levenshtein distance to 20 closest orthographic neighbours	(1.00, 3.85)	1.64	0.48	1.75	0.53
Average frequency of orthographic neighbours	(3.87, 10.69)	7.98	0.93	7.78	0.93
Mean Levenshtein distance to 20 closest phonological neighbours	(1.00, 4.80)	1.47	0.52	1.56	0.58
Average frequency of phonological neighbours	(4.44, 11.89)	8.35	1.08	8.11	1.08
Mean letter bi-gram frequency	(1, 396)	187.20	113.75	197.80	114.63
Lexical surprisal	(0.33, 14.77)	6.67	2.98	–	–
Semantic distance	(−0.82, −0.064)	−0.28	0.11	–	–
Word position	(2, 14)	5.68	2.62	–	–

Note: Contextual properties, which constitute the focus of the present study, are highlighted by gray shading. Contextual predictors and word position are token-based predictors.

form properties, such as orthographic neighbourhood density or orthographic probability. This is problematic if one seeks to assess the existence of early context effects because form-related properties are known to affect early ERPs (Hauk et al., 2006; Laszlo & Federmeier, 2014), including during the P2 time window. Our analyses thus include control for form-related effects.

We make another contribution that focus on the “shape” of contextual effects. Some previous work has reported non-linear effects of context on ERPs (Dambacher et al., 2006; see Parviz, Johnson, Johnson, & Brock, 2011 for related findings using MEG). We thus assess the (non)linearity of all three contextual predictors we consider. This is relevant for methodological reasons: assuming a non-linear effect to be linear can increase the Type I error and decrease power (for discussion, see Baayen, Vasishth, Kliegl, & Bates, 2017). Understanding the “shape” of an effect can thus inform the design future studies. We assess linearity by means of generalized additive mixed models (GAMMs, Wood, 2006). This method has been previously used to assess assumptions about the linearity of, e.g. surprisal effects on reading times (Smith & Levy, 2013), and is now increasingly used within ERP analyses (e.g. Hendrix, Baayen, & Bolger, 2017; for introductions directed at ERP researchers, see Smith & Kutas, 2015b; Tremblay & Newman, 2015).

Methods

EEG data

We used the same dataset of EEG recordings as used in Frank and Willems (2017), shared by the first author of the study. In the study, 24 subjects read sentences drawn from a natural corpus (for details, see Frank,

Otten, Galli, & Vigliocco, 2015). Sentences were presented word by word using an RSVP paradigm, with stimulus onset asynchrony (SOA) manipulated as a function of word length. We took the ERPs of each word epoched between −100 and 700 ms time-locked to word onset (down sampled to 250 Hz, i.e. 200 time points per word). Since we are interested in effects of context, we follow Frank and Willems and excluded the first content word of each sentence. This left ERPs for 670 content word tokens per subject (399 different word types).

Predictors of lexical properties (control variables)

Table 1 summarises the statistics of all predictors we considered. Note all of these predictors were entered into the analysis. As we described below, we used principal component analysis to extract the three most important dimensions of the various word form properties in Table 1.

Word frequency

We used the same word frequency measurement used by Frank and Willems (2017). They used log-transformed (relative) word frequency in the COW14 corpus (Schäfer, 2015).

Word form

Table 1 lists the word-form-related predictors we considered. All predictors were obtained from the English Lexicon Project, Balota et al., 2007). Specifically, we considered variables that reflect word length (number of characters, number of syllables, number of phonemes), neighbourhood density (number of orthographic neighbours, mean Levenshtein distance from a word to its 20 closest orthographic neighbours, mean log-transformed

frequency of orthographic neighbours, and the phonological equivalent of all these measures) and orthographic probability (mean letter bi-gram frequency). We chose these measures because they all have been found in previous work using experimenter-designed stimuli to affect early time windows (Hauk et al., 2006, 2009; Laszlo & Fed-ermeier, 2014).

One of the challenges to be expected for stimuli with typical statistics are high correlations among predictors. This was the case for the different types of word-form-related predictors in the present study (Figure 1), with particularly high correlations between word length and neighbourhood density measures (we address correlations of form-related predictors with other types of predictors below). Here we are interested in controlling for potential confounds due to *any* type of form-related effect on the analysis of contextual effects, rather than to tease apart different types of form-related effects. There is thus little advantage to be gained from including all form-related predictors in our analysis. To reduce the number of parameters required to control for these effects, we performed a principal component analysis (PCA) over all form-related predictors (see also Hauk et al., 2006). PCA identifies orthogonal (uncorrelated) dimensions out of a cluster of correlated variables. This allows researchers to balance the complexity of their models (the number of predictors and degrees of freedom in the model) against the ability to capture effects.

All form-related predictors were centred and standardised before being entered into the PCA. The loadings of the first three factors on form-related predictors are shown in Table 2; they account for 83.2% of the variance

among form-related predictors. We include these top three components as predictors in our main analysis to control for potential form-related effects. Although the loadings of these components are not of primary interest to the present study, we note similarities with the PCA analysis of Hauk et al. (2006) for experimenter-selected words. For example, we find that the first principal component loads on word length and neighbourhood density, and the second component loads onto orthographic probability. We found no component that clearly distinguishes between word length and neighbourhood density. This replicates the findings of Hauk and colleagues for experimenter-selected words, and is thus not necessarily a limitation specific to natural stimuli.

Contextual predictors (predictors of primary interest)

Research on experimenter-designed stimuli often uses cloze norms to estimate the predictability of target words, or the constraint of the preceding context. This method can tap into the parts of language users' implicit knowledge that are explicitly retrievable. Cloze estimates thus take into account many (unknown) sources of information that might affect the subjective predictability of a word. Cloze norms have also sometimes been employed in studies over natural stimuli (Luke & Christianson, 2016), including in ERP studies over experimenter-designed stimuli with heterogeneous contextual properties (Dambacher et al., 2006).

Here, we took a different approach. Following Frank and colleagues, we used computational models to obtain multiple different measures of contextual information (Frank et al., 2015; Frank & Willems, 2017). This made it possible to test *what types* of contextual information affect ERPs and *when*. The two measures we employed were inherited from Frank and Willems (2017), and are part of their shared data set.¹

Lexical surprisal

Our first predictor is intended to capture the information contained in the sequential order of words in the preceding context. The surprisal values came from mixture of a 5-gram model and a "skip bi-gram" model, both trained by Frank and Willems (2017). Both models and the optimal mixture weights between them were fit to a corpus containing English text collected from the web (COW14; Schäfer, 2015). The best performing model had a weight of 0.98 for the 5-gram model and 0.02 for "skip bigram", so that the estimate of lexical probability in context primarily reflects the 5-gram model. Lexical

Table 2. Loadings of the top three factors of the principal component analysis over word-form-related measures.

Predictor name	Loadings		
	Factor 1	Factor 2	Factor 3
Word length	-0.333	-0.083	0.201
Number of phonemes	-0.327	-0.098	0.200
Number of syllables	-0.280	-0.160	0.482
Number of orthographic neighbours	0.305	-0.217	0.158
Number of phonological neighbours (excluding homophones)	0.318	-0.124	0.503
Number of phonological neighbours (including homophones)	0.314	-0.117	0.535
Mean Levenshtein distance to 20 closest orthographic neighbours	-0.326	0.147	0.141
Average frequency of orthographic neighbours	0.311	0.047	-0.052
Mean Levenshtein distance to 20 closest phonological neighbours	-0.320	0.094	0.108
Average frequency of phonological neighbours	0.309	0.031	-0.124
Mean letter bi-gram frequency	-0.099	-0.921	-0.269
Proportion of Variance Explained (for a total of 83.2%)	67.5%	9.1%	6.6%

surprisal is the log-transformed reciprocal of the lexical probability estimate.

Behavioural research on language processing now predominantly assesses contextual effects in terms of lexical surprisal (or, equivalent except for the sign, log-transformed contextual probability; for a recent review and discussion, see Kuperberg & Jaeger, 2016). This contrasts with most ERP studies, which typically employ untransformed cloze rate in the analyses (Nieuwland et al., 2018; but see Delaney-Busch, Lau, Morgan, & Kuperberg, 2019; Frank et al., 2015; Frank & Willems, 2017; Yan, Kuperberg, & Jaeger, 2017). This is an important difference, both for theoretical and for methodological reasons. For example, research on reading has found that lexical surprisal is a better linear predictor of reading times than raw untransformed contextually-conditioned lexical probability (Goodkind & Bicknell, 2018; Smith & Levy, 2013). This has been taken as evidence for specific models of incremental information gathering during visual word recognition (see Smith & Levy, 2013).

While computational models of ERP components that would be sufficiently specific to be subject to such arguments are largely still lacking (for notable exceptions, see Brouwer, Crocker, Venhuizen, & Hoeks, 2017; Delaney-Busch et al., 2019; Rabovsky, Hansen, & McClelland, 2018), questions about the functional relation between contextual probability and the amplitude of ERP components also has methodological consequences. Namely, if an ERP component's amplitude of interest is linear in surprisal – thus log-linear, rather than linear, in contextual probability – this unduly emphasises differences among relatively unpredictable words (the difference between $p=.5$ or 1.0 corresponds to 1 bit of surprisal; so does the difference between .0039 and .00195). But this stands in stark contrast to most ERP studies which have focused on the contrast between highly predictable and unpredictable words.

Planned follow-up analyses in the present study thus test whether lexical surprisal or probability is a better linear predictor of the N400 amplitude, while controlling for other factors known to affect N400 amplitude. This contributes to preliminary evidence that surprisal is a better linear predictors of N400 amplitude over experimenter-designed stimuli (Delaney-Busch et al., 2019; Yan et al., 2017).

Semantic association

To capture the effects of semantic association between a preceding context and a target word, Frank and Willems (2017) trained a skip-gram model (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) on COW14 corpora

(Schäfer, 2015). This creates a vectorised representation of word semantics based on word co-occurrence in both preceding and following contexts. The vector representation of the context is calculated by adding the word vectors of all content words in the context (see Figure 2). The semantic association between a word and its context is represented by the cosine distance between the two vectors, i.e. the negative cosine between the vector representations of each content word and its context (for further details, see Frank & Willems, 2017).

Assessing the feasibility of ERP analyses over natural stimuli: correlation among predictors

One potential limitation of ERP analyses over natural stimuli originates in correlations between the variables of interest. Such correlations can affect analyses in at least two ways. First, if a predictor is highly correlated with other predictors, caution is required when interpreting the effect of this predictor when the correlated predictors are not included in the model as control variables. Second, high correlation between predictors can potentially cause multi-collinearity when including these predictors in the same model, which reduces the power to detect effects for the collinear predictors, and can limit the interpretability of the estimated effect of the correlated predictors (Baayen, Feldman, & Schreuder, 2006). Hence, examining the correlation structure among predictors can inform us as to whether one can statistically tease their effects apart within this or similar data sets with typical statistics.

As shown in Figure 3, there are only moderate correlations between our predictors. Log-transformed word frequency correlates with lexical surprisal ($r=-0.58$) and the first form-related PCA factor. Both correlations are to be expected. More frequent words on average have higher n-gram probability, and hence lower surprisal.² Additionally, the first PCA factor loads strongly on to word length (Table 2), which is known to be negatively correlated with word frequency (Zipf, 1949). Second, word order negatively correlates with semantic distance ($r=-.39$). This makes intuitive sense: sentences with typical statistics (and thus without semantic anomalies) tend to become increasingly constraining as the sentence unfolds, with words towards the end of the sentence fitting more closely into the semantic context (Payne et al., 2015). None of these correlations are sufficiently high to cause concerns about collinearity (cf. Baayen et al., 2006). Follow-up analyses reported in Appendix C confirm that none of the results we report below is affected by collinearity.

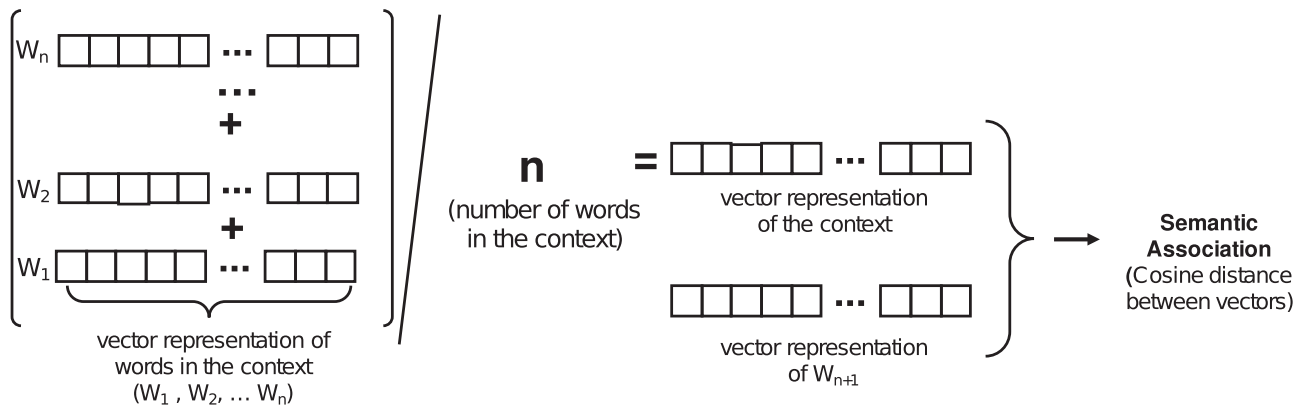


Figure 2. Measuring the semantic association between a word and its context (for details, see Frank & Willems, 2017).

Analysis approach

Below we present both millisecond-resolution time series analysis and analyses based on *a priori* defined time windows. This allows us to compare the effect of the different approaches on the same data set. This comparison also might provide an explanation for the seemingly conflicting results of Frank and Willems (2017) and Dambacher et al. (2006). Specifically, it is possible that the effects that are found in time series analyses are sometimes non-detectable in time window analyses. This provides a possible explanation. For the window-based analyses, we also entertain non-linear effects of context, as observed in some previous studies (e.g. Dambacher et al., 2006). In particular, we test whether lexical surprisal or lexical probability is a better linear predictor of the N400 amplitude, as this amplitude has been linked to the unexpected information associated with lexical processing (for reviews,

see Kuperberg, 2016; Kutas & Federmeier, 2011; Van Petten & Luka, 2012).

Before we present the results of these different analyses, we briefly discuss different approaches to time series analyses (over ERP or other data), and motivate the choices we made in this study. This leads us to introduce an approach to modelling and discounting auto-correlations that might hold promise for research on ERPs or similar types of signals, regardless of whether they are elicited over natural stimuli.

Pros and cons of different approaches to time course analyses

There are at least three broad classes of approaches for analysing ERP time series data (Hauk et al., 2009; Smith & Kutas, 2015a). The first and most common approach defines time-windows based on either *a priori* considerations (e.g. based on theory or previous findings) or after visual inspection of the data. ERPs are then typically averaged over that time window, and the averages are submitted for analysis (for discussion, see Burns, Bigdely-Shamlo, Smith, Kreutz-Delgado, & Makeig, 2013). One downside of this approach is that it makes assumptions about the relevant time windows, rather than detecting the relevant time course in exploratory analyses. On the upside, this approach reduces the risk of inflated Type I errors due to correlations between ERPs at different time points.

A second approach is to use methods for linear or non-linear time series modelling. While this approach has been under-explored in ERP analyses (but see Baayen, van Rij, de Cat, & Wood, 2016), it has been increasingly influential in, for example, eye-tracking analyses for visual world experiments (Mirman, Dixon, & Magnuson, 2008; Nixon, van Rij, Mok, Baayen, & Chen, 2016). In the initial stages of this project, we employed Generalized Additive Mixed Models (GAMMs, Wood,

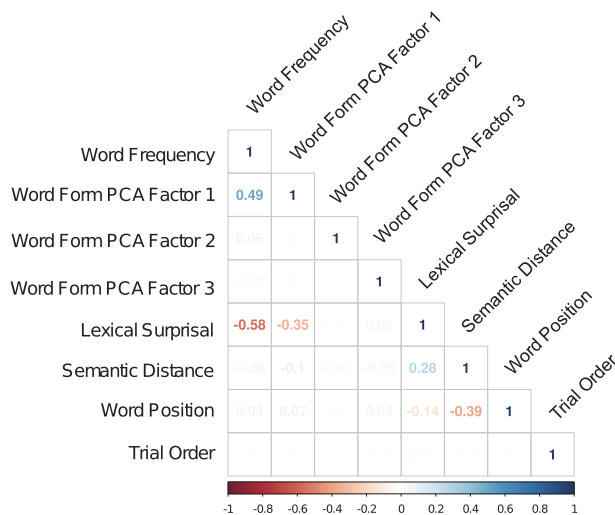


Figure 3. Correlations between predictors in main analysis. Blue colours indicate positive correlations and red colours indicate negative correlations.

2006) to analyze the time series data from each electrode. One advantage of this approach is that it provides an elegant and powerful way to model non-linear effects of predictors on ERPs over time. There are, however, also downside to this approach. GAMMs offer researchers a large number of degrees of freedoms (e.g. the choice of basis function, number of knots, knot locations; for an overview, see Baayen et al., 2017), and the consequences of those choices for ERP analyses are not yet known. Another potential downside of this approach is that it requires additional analysis steps (based on the initial GAMM results) if the researcher's primary interest is to determine *when* an effect emerges. The downside that eventually convinced us to explore another approach is that currently available implementations of GAMM are limited in their ability to model auto-correlations (but see additional auto-correlation models in, e.g. the brms library, Bürkner, 2017). A failure to account for such correlations will inflate Type I error rates. We note though that the patterns detected by both approaches were very similar for the present study. Our choice to switch methods was purely based on methodological considerations.

Another approach, and the one that we present here, involves separate analyses for the ERP signal *at each point in time* (e.g. Frank & Willems, 2017; Groppe, Urbach, & Kutas, 2011; Smith & Kutas, 2015a). This approach *detects*, rather than *assumes*, the emergence of ERP components related to different processes over time. However, the approach also has two interrelated critical downsides. First, it has the potential to inflate Type I error by running multiple comparisons. For example, the present study requires 200 separate linear mixed models (LMMs) for each electrode given the time window of interest and sampling rate. While this does not inflate the Type I error rate of each individual model, it makes it harder to interpret the presence of significances (as an average of 10 spuriously significant effects per electrode is expected by chance). The second weakness is that this approach assumes that each of the 200 analyses are independent. As discussed in the previous paragraph, this assumption that is wrong: it is well known that ERPs exhibit strong auto-correlations (Guthrie & Buchwald, 1991).

To address these issues, we performed two additional analyses aimed to correct the false positive rate of our results. The first of these employs Bonferroni correction. Given that we are conducting 200 regressions on each electrode, we used an alpha level of $0.00025 = 0.05/200$. Bonferroni correction can, however, be conservative (for discussion, see e.g. Narum, 2006). Here we use this correction as a lower bound of what effects are the most reliable and trust-worthy. The second additional

analysis explores a new approach to more directly estimate and discount the general autocorrelation (AC) structure on the EEG signal. This approach is described in the next section.

The three main analyses reported below include all the predictors described above: word frequency, the three PCA factors over form-related measures, and our two predictors of interest (lexical surprisal, semantic distance). Following Frank and Willems (2017), we also included three control predictors in each analysis, including the ERP baseline³ (the average amplitude between -100 and 0 ms before word onset), trial order (the order of presentation in the experiment), and word position (word position within a sentence). To facilitate the comparison of effect sizes across different predictors and with previous work (Frank & Willems, 2017), we z-scored all predictors. All models also included random intercepts by subject and by word token. We treated word tokens, rather than types, as random effects because this is conservative with regard to our question of interest: context effects vary *between* the random intercepts for word tokens. Because contextual effects mostly affect centro-parietal electrodes, we focused our analyses for each time point of 3 midline electrodes (Fz, Cz, Pz).

All analyses were performed using the *lmer* function (Bates, Mächler, Bolker, & Walker, 2015) for linear mixed-effects regression from the *lme4* package in the statistical software R core team, (2016).

Auto-correlation (AC)-corrected ERPs

An overview of our procedure to obtain AC-corrected ERPs is given in Figure 4. The goal of this procedure is to capture the *general* autocorrelation structure across all time points. We repeated the following process 1,000 times. For each electrode of each subject, we first randomly sampled 10,000 EEG samples across the whole recording session (excluding EEG during inter stimulus intervals, i.e. only including EEG that were recorded during stimulus presentation). We refer to this as the training data. We then extracted the preceding EEG at lag 1 (4 ms earlier) to lag 30 (120 ms earlier) for each of these 10,000 time points (EEG from the first 120 ms of a recording session were excluded from this procedure). This allows us to capture correlations due to oscillations between ~8 and 25 Hz. This matrix of 30 lag predictors over 10,000 data points was submitted to a principal component analysis (PCA). The top 5 PCA factors – capturing most of the auto-correlational structure of the 10,000 data points – were then used to predict the 10,000 ERP data points (R_{fit}^2 mean over by-subject means = 78%; SD = 7%).

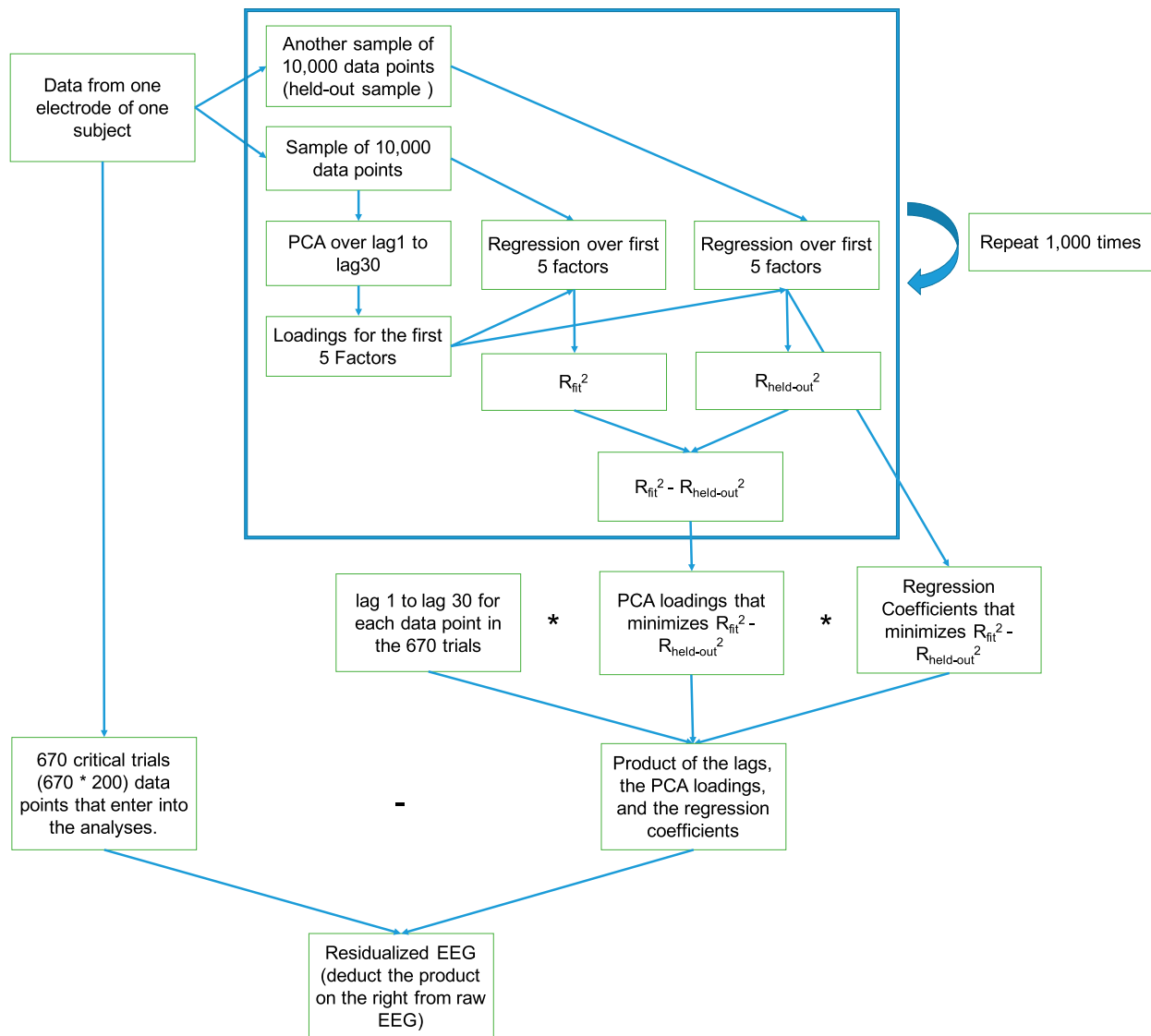


Figure 4. Illustration of the procedure employed to obtain autocorrelation-corrected ERPs.

Next, we assessed the reliability of the coefficients from this analysis. For each of the training data sets, we randomly sampled another 10,000 time points (held-out test data). We used the PCA loadings derived from the training data to and assessed how well they predicted the EEG activity from the test data ($R_{\text{held-out}}^2$ mean over by-subject means = 78%; SD = 7% [sic]) and calculated the difference between $R_{\text{fit}}^2 - R_{\text{held-out}}^2$ (mean = 0.004%; SD = .002%).

We selected the PCA loadings that minimised the difference between $R_{\text{held-out}}^2$ and R_{fit}^2 . We used these PCA loadings together with corresponding coefficients derived from the regression over the held-out sample to predict the EEGs analyzed in the main analyses. The residuals of this prediction constitute the AC-corrected ERPs.

The loading of the five factors for the AC-correction from the Pz electrode of Subject 1 are shown in [Figure](#)

5. Although nothing in our approach constrained loadings to be cyclic, the loadings reflect oscillations at various frequencies (e.g. 10 Hz for the top panel). [Figure 6](#) shows that AC-corrected ERPs exhibit substantially less variance, and less extreme amplitudes.

Next, we report the results from the analyses over the whole ERP time series. These constitute the core of our analyses. Then we report the results of the time window analyses. Finally, we turn to the question of whether lexical surprisal or untransformed lexical probability is a better linear predictor of N400 amplitude over natural stimuli.

Results of time course analysis

We first present the results for the two contextual predictors of interest (lexical surprisal and semantic association). Then we present the results for the control

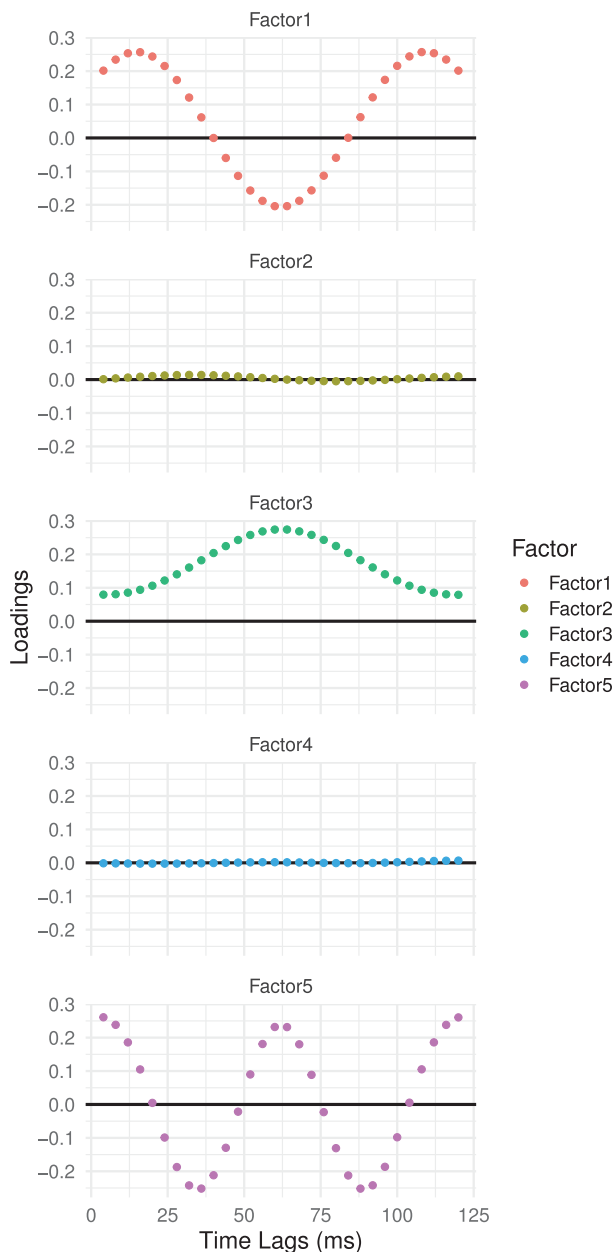


Figure 5. Loadings of the five factors that went into the auto-correlation correcting step from the Pz electrode of Subject 1.

predictor that are moderately correlated with these contextual predictors (word frequency and word position). The time course of frequency effects is of particular relevance, as it provides a baseline for the effects of the contextual predictors. Recall, for example, that Dambacher et al. (2006) argued against the existence of early contextual effects on ERPs because they found effects of frequency on P2 (140–200 ms after word onset), but no effects of cloze scores during the same time window. Word-form-related predictors and additional controls (word position and ERP baseline) are reported in the appendix. We note that we found early word form effects for both AC-corrected and uncorrected

ERPs around 100, 150, and 200 ms after word onset, as well as highly significant effects at later time points (e.g. ~350 ms). These results resemble previous findings from experimenter-designed stimuli (e.g. Hauk et al., 2006, 2009; Laszlo & Federmeier, 2014). Of note is that we found these effects on electrodes that were selected because of our focus on context effects, and they differ from the electrodes for which previous literature suggests the strongest form-related effects. This validates the need to account for form-related controls in research on the time course of context effects.

Lexical surprisal

Figure 7 shows the effect of lexical surprisal on ERP amplitudes as a function of time. In the left panel, the y-axis shows the coefficient of lexical surprisal – i.e. the change in ERP amplitudes with 1 unit change in the scaled lexical surprisal. For a more intuitive interpretation of the results, we also plotted the model predicted ERP amplitudes for high (90th quantile) and low (10th quantile) lexical surprisal (Figure 7, right panel). Following conventions in the field, we plotted ERP with negative values up. We will present the same type of plots for all the predictors we discuss.

Before we discuss the significant patterns, we make two observations about the relation between the main analysis and the two approaches to Type I error correction that we present. First, the corrections are more conservative, as intended. Second, the two types of correction approaches seem to differ in *how* they relate to the uncorrected analysis. The analysis of AC-corrected ERP tends to find significances that align with the *onset* of significances in the uncorrected ERPs. This suggests that later significances in the uncorrected ERPs are explained by autocorrelations with earlier ERPs. We take this to be a desirable property, in particular, for time course analyses. Bonferroni corrected analyses, on the other hand, tend to attribute the significances to the centre of the time range that exhibits significant effects in the uncorrected analysis (i.e. alignment to the peaks in ERP). These patterns hold across the other predictors presented below, and thus might be a general property of these two types of Type I error correction for time series analyses over ERP data.

Similar to Frank and Willems (2017), we found an early effect of lexical surprisal prior to the N400 time window, around 200 ms after word onset. We found this effect while controlling for word form predictors. Lexical surprisal has a positive effect on ERP amplitudes in this time window. However, this effect does not survive Bonferroni correction. The most robust effect of lexical surprisal is found in the (late) N400 time

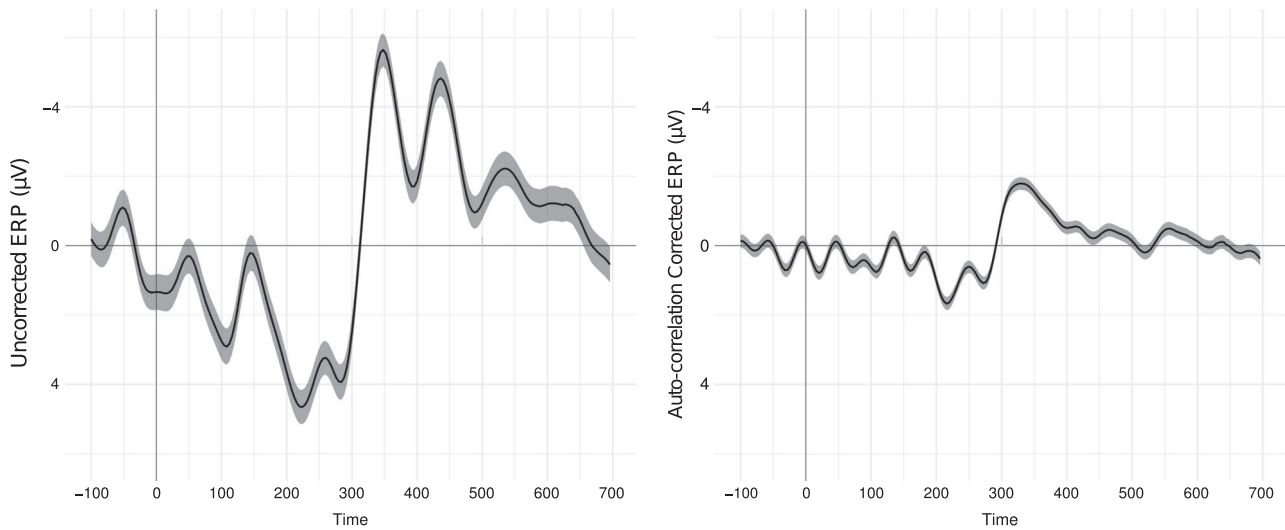


Figure 6. Grand average ERPs on the Pz electrode of subjects. Shaded areas represent ± 1 standard error over critical trial. Left panel: ERPs before auto-correlation correction. Right panel: ERPs after auto-correlation correction.

window. Lexical surprisal has a negative effect on ERP amplitudes in this time window, i.e. less predicted words elicit a larger N400. This effect is also significant after Bonferroni correction.

Semantic distance effect

Figure 8 shows the effect of semantic distance on ERP amplitude as a function of time. Similar to lexical surprisal, there is an early effect of semantic distance prior to

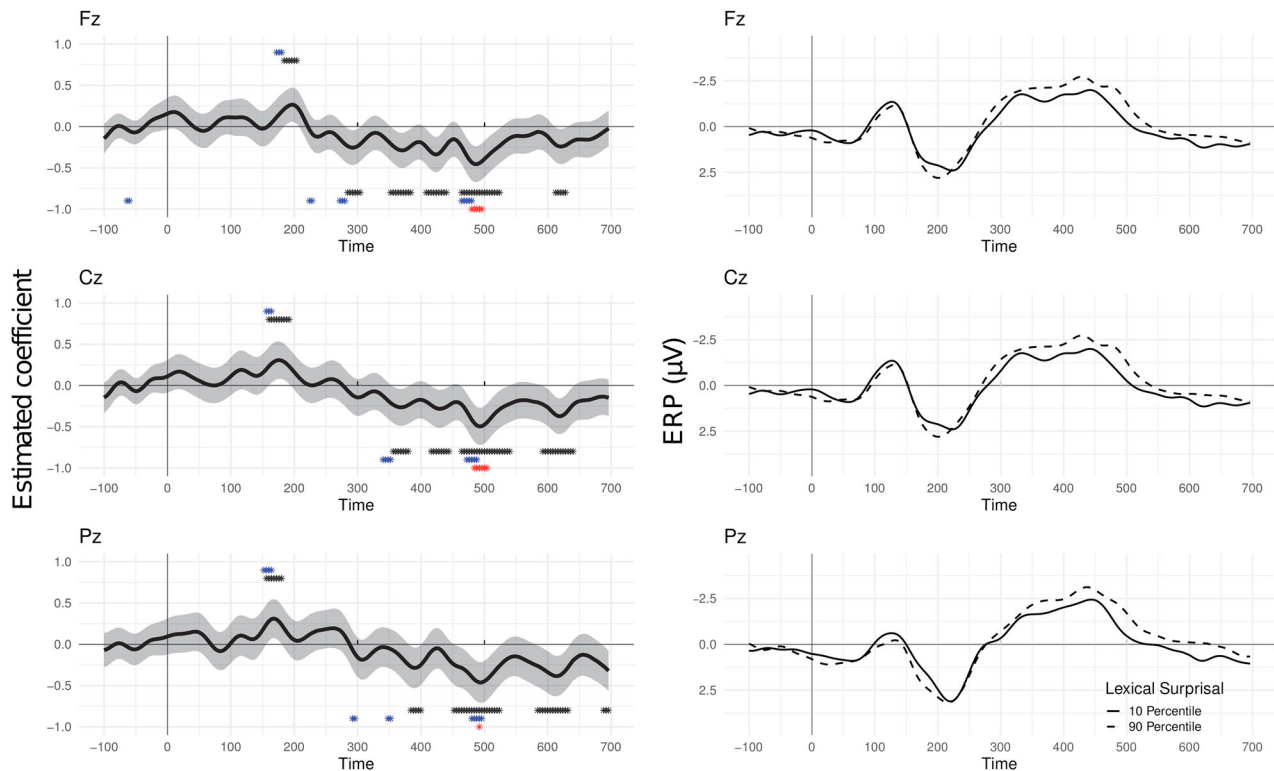


Figure 7. Lexical surprisal effect over time (–100 to 700 ms from word onset). Left panel: Coefficients from mixed-effects models for uncorrected ERPs. Shaded areas represent estimated 95% confidence intervals. Dots highlight time points where the effect of frequency is significant. Black dots: significant with an alpha level of .05. Blue dots: significant for AC-corrected ERPs. Red dots: significant for uncorrected ERPs after Bonferroni correction. Right panel: Model predicted ERP amplitudes at High (90th percentile, dotted line) and Low (10th percentile, solid line) lexical surprisal. Following conventions in the field, we plot ERPs with negative values up (the same is true for all the other ERP graphs we are plotting in this paper).

the N400 time window, around 250 ms after word onset. Semantic distance has a positive effect on ERP amplitudes in this time window, i.e. words that are semantically more distant from the context elicit a larger positivity. However, this effect is not significant after either of our Type I error corrections. The most robust effect of semantic distance is found in the (early) N400 time window. Semantic distance has a negative effect on ERP amplitudes in this time window, i.e. words that are semantically more distant from the context elicit a larger a larger N400. This effect is significant for AC-corrected ERPs but not significant after Bonferroni correction. There is also a late effect of semantic distance around 550 ms after word onset. This effect is significant for AC-corrected ERPs on electrode Fz and Cz but not after Bonferroni correction. To our knowledge, this effect is unexpected and not previously reported.

Frequency effect

Figure 9 shows the effect of word frequency on ERP amplitudes as a function of time. The earliest frequency effect that is significant for both corrected and uncorrected ERPs occurs around 100 ms after word onset on

Cz. Frequency positively correlates with ERP amplitudes, i.e. increases in frequency predict increases in ERP amplitudes, replicating the time course and direction of previous findings (e.g. Hauk et al., 2006; Sereno, Posner, & Rayner, 1998). However, it is not significant after Bonferroni correction. Between 200 and 300 ms from word onset, frequency has a negative effect, in the same direction as previous findings in this time window (Hauk et al., 2006) on Fz and Cz. It is not significant after Bonferroni correction. The most robust frequency effect is found around the N400 time window. Frequency has a positive effect on ERP amplitudes, reflecting a smaller N400 for more frequent words (Hauk et al., 2006; Van Petten & Kutas, 1990). This effect is also significant after Bonferroni correction.

Word position

Figure 10 shows the effect of word position on ERP amplitudes as a function of time. Overall, there is a sustained negativity effect, i.e. words appearing later in the sentence elicit a larger negativity. The analysis of AC-corrected ERPs seems to attribute much of this shift to autocorrelation, leaving just a few inflection points

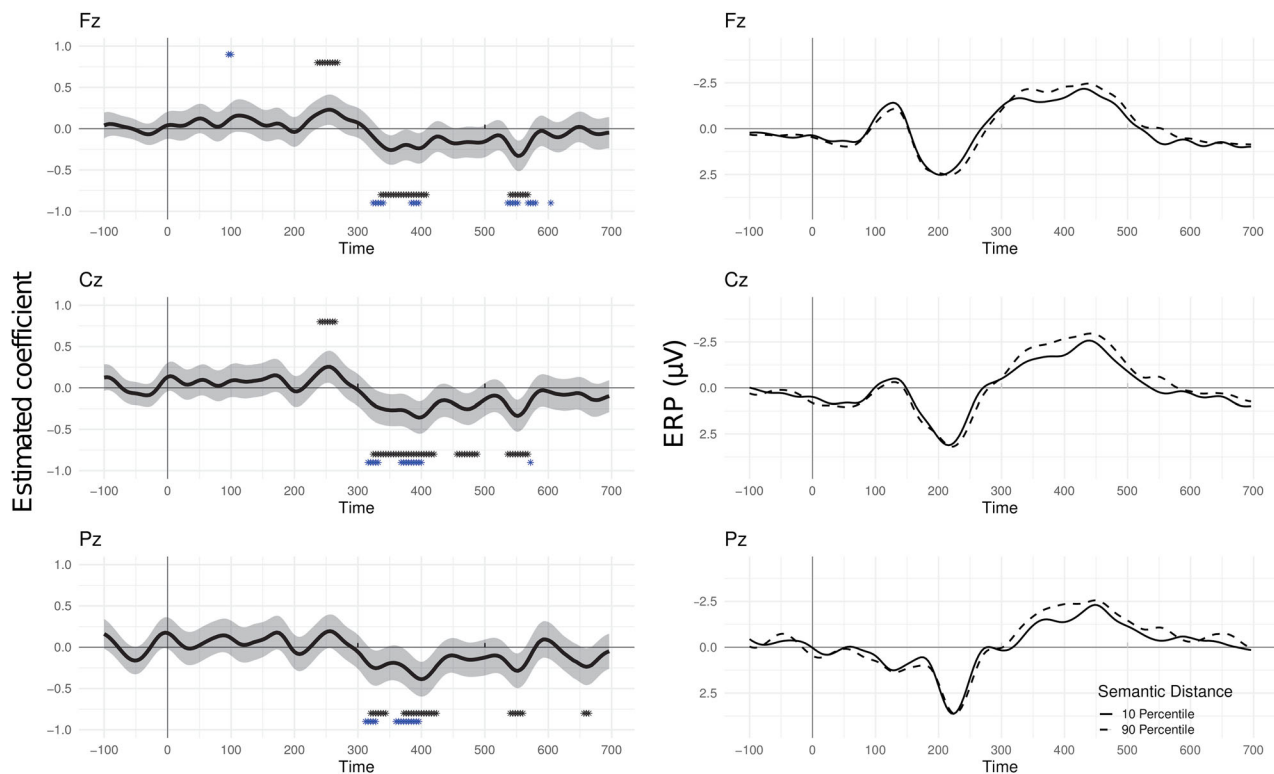


Figure 8. Semantic distance effect over time (–100 to 700 ms from word onset). Left panel: Coefficients from mixed-effects models for uncorrected ERPs. Shaded areas represent estimated 95% confidence intervals. Dots highlight time points where the effect of frequency is significant. Black dots: significant with an alpha level of .05. Blue dots: significant for AC-corrected ERPs. Red dots: significant for uncorrected ERPs after Bonferroni correction. Right panel: Model predicted ERP amplitudes at High (90th percentile, dotted line) and Low (10th percentile, solid line) semantic distance between a word and its context.

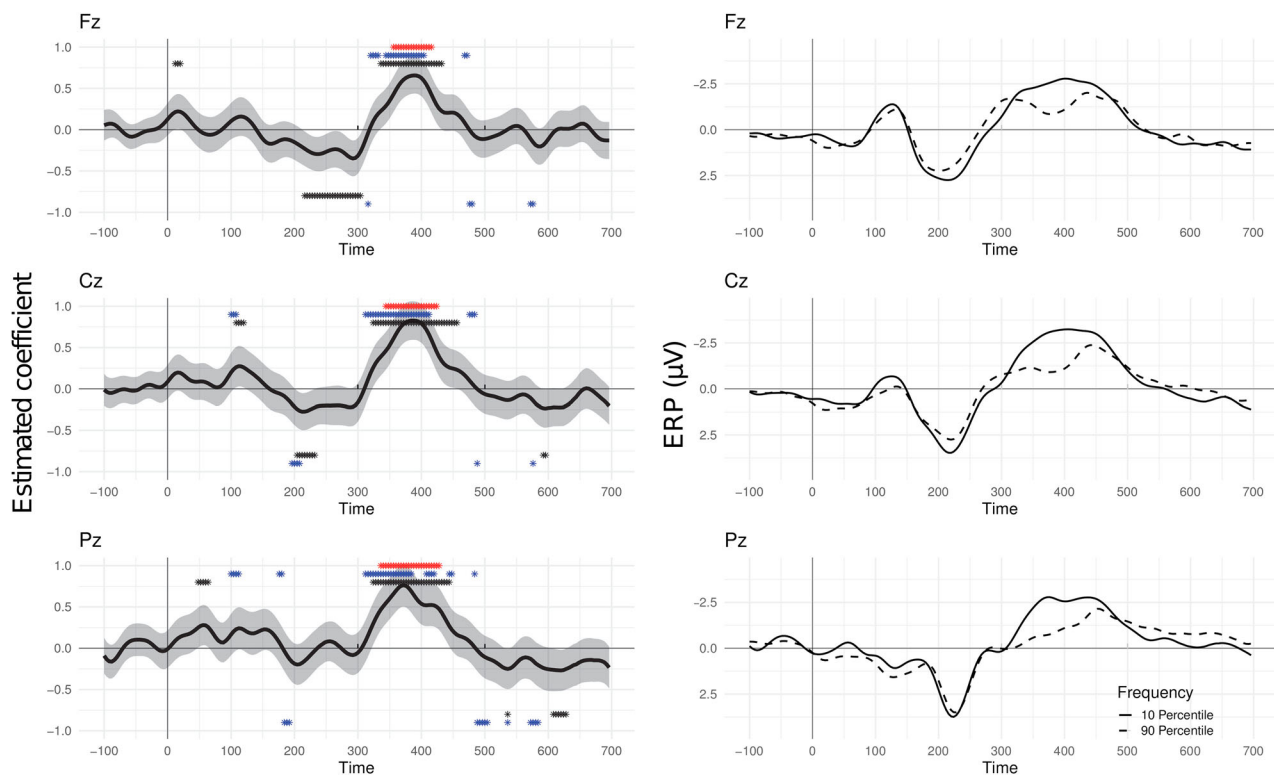


Figure 9. Frequency effect over time (–100 to 700 ms from word onset). Left panel: Coefficients from mixed-effects models for uncorrected ERPs. Shaded areas represent estimated 95% confidence intervals. Dots highlight time points where the effect of frequency is significant. Black dots: significant with an alpha level of .05. Blue dots: significant for AC-corrected ERPs. Red dots: significant for uncorrected ERPs after Bonferroni correction. Right panel: Model predicted ERP amplitudes at High (90th percentile, dotted line) and Low (10th percentile, solid line) frequency.

significant. Only three time windows reach significance after Bonferroni correction. Word position has a negative correlation with ERP amplitudes around 150 ms and around 400 ms after word onset on Fz, and around 700 ms on Cz.

Results of time window analyses

The results of the time series analysis broadly replicate the time course of contextual effects observed by Frank and Willems (2017). This includes evidence for early effects of lexical surprisal and semantic association that survives our AC-correction (but not Bonferroni correction). This raises the question of whether these effects would also be visible in a time window analysis. We address this question for the two time windows previously examined by Dambacher et al. (2006), i.e. P2 and N400. We followed the time window and electrodes used in their study. For P2, we averaged the ERP amplitudes between 140 and 200 ms after word onset and averaged across fronto-central electrodes (AFZ, FZ, F3, F4, FC3, FC4, FC5, FC6, CZ, C3, C4). For N400, we averaged the ERP amplitudes between 300 and 500 ms after word onset and averaged across centro-parietal electrodes (CZ, C3, C4, CP5, CP6, PZ, P3, P4, P7, P8, O1, O2).

Because some previous work suggests that context effects on aggregate ERP amplitudes can be non-linear (including Dambacher et al., 2006), we used GAMMs to examine the effects of the two context predictors (lexical surprisal and semantic distance). For each predictor, we first tested whether including the predictor as a linear predictor in the model significantly increased model fit against ERP amplitudes compared to when it was not included. We then examined whether including the predictor as a non-linear predictor in the model further increases fit compared to when only the linear predictor is included (along with all controls; for models see Appendix E). All model comparisons employed the *CompareML* function from the R package *itsadug* (van Rij, Wieling, Baayen, & van Rijn, 2017), and were based on maximum likelihood fits (as is recommended for comparisons of nested GAMMs that differ in fixed effect structure, van Rij, 2016). The GAMM analyses included the same controls and random effects employed in the time series analysis.

P2 time window

In line with the results of the time series analyses, lexical surprisal is a marginally significant linear predictor of P2

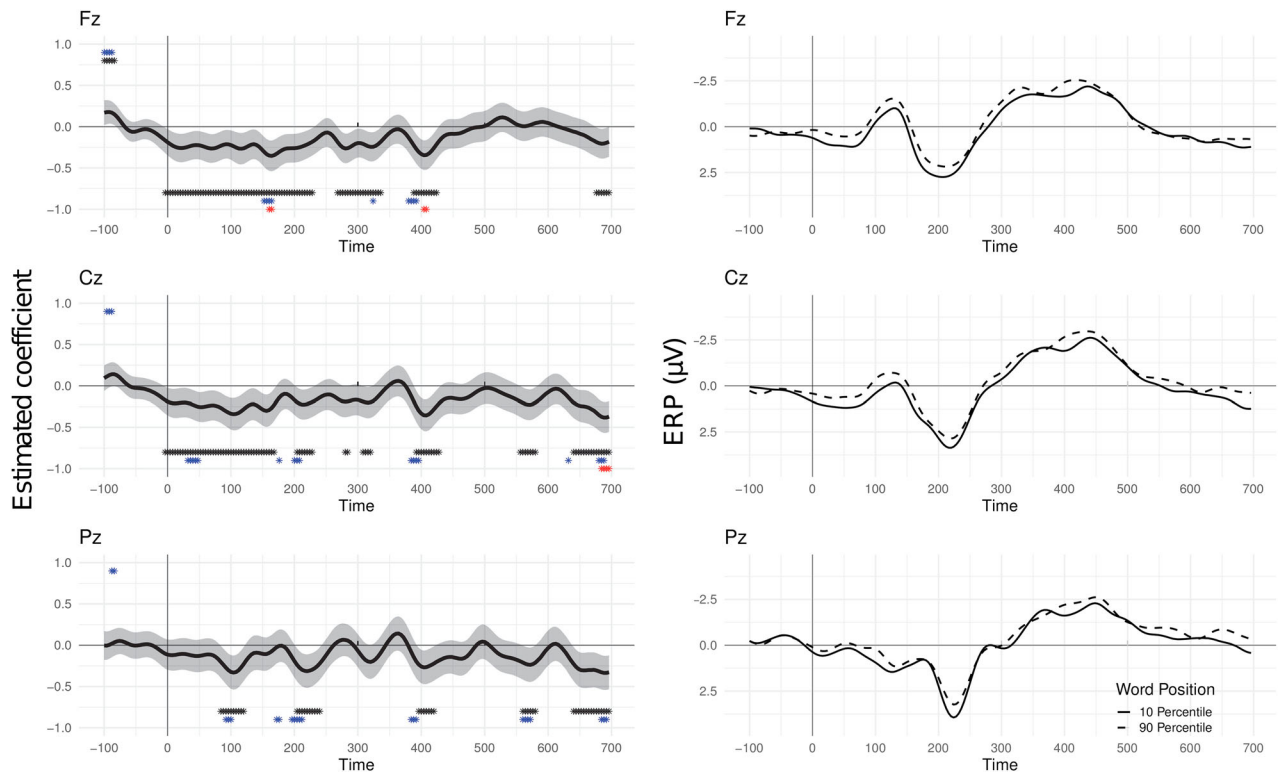


Figure 10. Word position effect over time (–100 to 700 ms from word onset). Left panel: Coefficients from mixed-effects models for uncorrected ERPs. Shaded areas represent estimated 95% confidence intervals. Dots highlight time points where the effect of frequency is significant. Black dots: significant with an alpha level of .05. Blue dots: significant for AC-corrected ERPs. Red dots: significant for uncorrected ERPs after Bonferroni correction. Right panel: Model predicted ERP amplitudes at High (90th percentile, dotted line) and Low (10th percentile, solid line) word position.

amplitude ($\chi^2 = 3.06$, $p = 0.08$). Semantic distance is not a significant linear predictor of P2 amplitude ($\chi^2 = 0.50$, n.s). Including any context predictors as non-linear predictors does not improve model fit beyond their inclusion as linear predictors ($ps > 0.9$).

N400 time window

In line with the results of the time series analyses, both lexical surprisal and semantic distance are significant linear predictors of N400 amplitude (lexical surprisal: $\chi^2 = 8.92$, $p < 0.005$; semantic distance: $\chi^2 = 10.06$, $p < 0.001$). Including lexical surprisal and semantic distance as a non-linear predictor does not improve model fit beyond their inclusion as linear predictors ($ps > 0.9$).

Summary

Thus, the window-based analyses are broadly in line with the time series analyses we conducted. Both types of analysis provide some support for early effects of context. However, the window-based analysis seems overall more conservative: even some of the patterns that survived AC-correction in the time series analyses

did not reach significance in the window-based analyses ($p = 0.08$). On the other hand, Bonferroni-correction of time series data seems to be even more conservative than the window-based analysis. (We note that our comparison leaves open which analysis is more adequate, as the ground truth – the time course of context effects – is the object of inquiry here, rather than being known.)

What predicts N400 amplitude? Lexical surprisal vs. lexical probability

Finally, we turn to our last question – whether the amplitude of the N400 is best described as a linear function of lexical surprisal (i.e. log-transformed contextually-conditioned lexical probability) or untransformed (contextually-conditioned) lexical probability. The absence of significant non-linearities in the effects of lexical surprisal, reported in the previous section, provide initial support for the hypothesis that the N400 is best described as a linear function of lexical surprisal, rather than lexical probability. However, it is also possible that the absence of significant non-linear effects for lexical is caused by lack of statistical power.

To assess this possibility, we first repeated the analysis of the N400 window reported in the previous section, but this time substituting lexical probability for lexical surprisal. Paralleling the results for lexical surprisal, we found that including lexical probability as a *non-linear* predictor does not improve model fit compared to when it is included a *linear* predictor of N400 amplitude ($\chi^2 = 0$, n.s.). That is, despite the fact that lexical probability is exponentially related to lexical surprisal, the effects of either predictor on N400 amplitude do not significantly deviate from linearity.

Since it is not possible that both lexical surprisal and lexical probability are linear predictors of the N400, we assessed which of the two null effects for the non-linearity test is more likely to reflect a true null effect (and thus evidence for a linear relation of that predictor with the N400). Figure 11 compares the predicted effects of lexical probability and lexical surprisal on the N400 amplitude. Panel A plots predictions aligned with lexical probability. This panel compares the predictions from the GAMM in which lexical probability is allowed to be *non-linear* (but, in fact, has a linear relation with the N400) with the predictions from the GAMM with lexical surprisal as a *linear* predictor transformed into probability space. Panel B plots the predictions aligned with lexical surprisal. Lexical probability values cluster close to zero, leaving little power to detect deviation from linearity where they are predicted. This contrasts with lexical surprisal values which are distributed rather uniformly across the entire interval over which we make predictions. It is thus likely that we had substantially more statistical power to detect deviation from linearity for lexical surprisal (under the hypothesis that the N400 is linear lexical probability), than we had to detect deviation from linearity for lexical probability (under the hypothesis that the N400 is linear in surprisal). This makes the absence of evidence for non-linearity of lexical surprisal more informative.

Finally, we examined whether including lexical surprisal as linear predictor in the model can improve model fit while lexical probability is already in the model and vice versa. Adding lexical surprisal as a linear predictor to a model with lexical probability does improve model fit ($\chi^2 = 5.00$, $p = 0.025$), but not vice versa ($\chi^2 = 0.036$, n.s.). This suggests that lexical surprisal is a better linear predictor of N400 amplitude than lexical probability.

Discussion

ERP has proven a powerful paradigm in advancing our understanding of the time course of information processing during language understanding. With very few exceptions, previous work has, however, almost

exclusively employed experimenter-designed stimuli. Such stimuli differ in many ways from naturally occurring language. ERP experiments on context effects often present participants with equal (50/50) proportions of “high” and “low predictability” target words (e.g. having a contextual probability of more or less than .5). ERPs are typically only analysed for those target words. This means that the words and contexts that are analysed in most ERP experiments make for a very odd subset of natural text. As an illustration, consider that less than 3% of the content words analysed in the present study had an n-gram probability above .5. This does not invalidate research on experimenter-designed stimuli. Psycho- and neurolinguistics paradigms that investigate the language system by presenting it with rare, taxing, and unexpected stimuli have a long history of critical contributions to the field. However, just as the field of modern neuroscience does not *exclusively* rely on findings from patients with brain damage, ERP studies on natural stimuli with typical statistics offer important validation of neurolinguistics theories (as well as the potential of unique insights).

In the present study, we examined when different contextual predictors affect ERPs in reading natural sentence stimuli. We found that both surprisal and semantic association are independent linear predictors of N400 amplitude (replicating Frank & Willems, 2017). We also found evidence that lexical surprisal affects early ERPs during the P2 time window, though this effect was only marginally significant in the time window analysis.

The early effects of some contextual predictors confirm the patterns found in Frank and Willems (2017) but contrasts with the findings of Dambacher et al. (2006). This suggests that contextual information can affect early lexical processing even in natural sentence reading where the context is not highly constraining. This would be in line with accounts that lexical processing is interactive and cascading (Dell & O’Seaghdha, 1992; McClelland & Elman, 1986). We note, however, that the contextual effect on the P2 is not as robust as the effect on the N400. In particular, the time window analyses only found a marginally significant effect of lexical surprisal in the P2 time window. This is not necessarily unexpected for two reasons. First, the cascading effects on form-based processing are expected to be small (Dell & O’Seaghdha, 1992; for discussion, see, e.g. Yan et al., 2017). Second, early ERP effects are likely smaller than later effects and harder to detect (see, e.g. Hauk et al., 2006). Further work on other data – including on ERPs over natural stimuli – are needed to cross-validate the effects.

Next, we discuss some further similarities and differences between our findings and previous work on

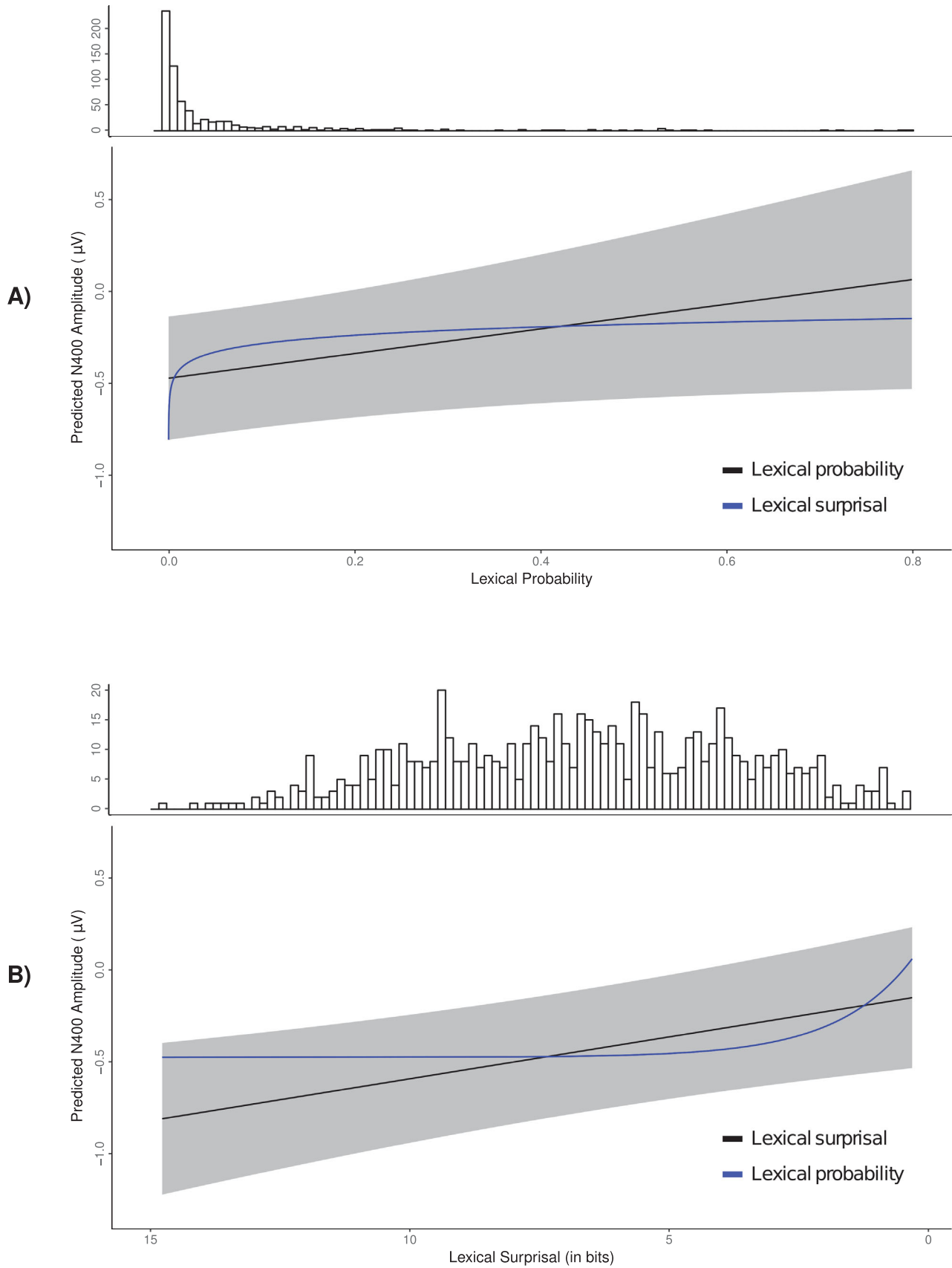


Figure 11. Comparing predictions of lexical probability and surprisal. Panel A: Predictions are shown on a probability scale. Panel B: Predictions are shown on a surprisal scale. Black line denotes predicted N400 amplitude from GAMMs with lexical probability (A) or lexical probability (B) as a predictor allowed to be non-linear. Shaded area represent estimated 95% confidence intervals. Blue line denotes predicted N400 amplitude from GAMMs with lexical surprisal (A) or lexical probability (B) as a linear predictor.

experimenter-designed stimuli. Then we discuss the implication of our finding that lexical surprisal is a better predictor of N400 amplitude. Lastly, we review some limitations of the methods employed in the present study, and how they might be resolved in future work.

ERP effects in reading natural text

Our time course analyses qualitatively replicated previously findings using experimenter-designed stimuli. Besides replicating the time course of context effects, we also found word-form-related effects with similar time course as in studies using experimenter-designed stimuli (Hauk et al., 2006; Laszlo & Federmeier, 2014), as described in Appendix D. One noticeable exception to this is that the PCA factor with a strong loading on orthographic probability was not found to have an early effect on ERPs (in contrast to, e.g. Hauk et al., 2006). This could indicate that orthographic probability of a word is not an important source of information in natural language reading compared to the meta-linguistic tasks that subjects were performing in the original study (lexical decision). It is, however, also possible that this null effect is due to our focus on centro-parietal sites, whereas the most robust effect of orthographic probability was found at peripheral sites. Future work is required to thoroughly compare the differences in accessing lexical information in reading natural and experimenter-designed stimuli by examining larger range of stimuli with different lexical properties and sample more electrodes that are sensitive to early lexical processing.

Another potential difference to previous work on experimenter-designed stimuli pertains to the effect of word position (Payne et al., 2015; Van Petten & Kutas, 1990). Unlike these works, we did not find that N400 amplitude decreases with the increase of word position. There are two possible reasons why we did not find such an effect. First, the sentences used in Payne et al. (2015) were longer (mean 15, range 5–27) compared to the dataset we analysed (mean 10, range 5–15). Second, the current dataset used naturalistic stimuli while Payne et al. (2015) and previous studies used experimenter designed stimuli. In experimenter-designed stimuli, each sentence is like a mini-story that has richer contextual information, e.g. “She kept checking the oven because the cake seemed to be taking an awfully long time to bake” (Payne et al., 2015). In our dataset, the naturalistic stimuli are excerpts from novels, information is likely more spread out across sentences but less constraining within a sentence. Therefore, later parts of the sentences are less constrained by

previous parts of the sentences in our stimuli, and word position is not as good a metric for sentential constraint.

Lexical surprisal vs. lexical probability

Another finding worth highlighting is that we found that (n-gram) lexical surprisal is a better linear predictor for N400 amplitude than (n-gram) lexical predictability. This complements findings from previous tests on ERP over experimenter-designed stimuli (Delaney-Busch et al., 2019; Yan et al., 2017). It also mirrors findings from reading time studies (Goodkind & Bicknell, 2018; Smith & Levy, 2013).

Although further work on this question is required to reach certainty about the functional relation between N400 amplitude and lexical probability, these findings call for caution with regard to a common practice in ERP design and analyses. Most research on the N400 has measured and manipulated predictability in terms of (non-log-transformed) cloze scores. However, if the N400 amplitude is linearly correlated with the surprisal, i.e. correlated with lexical probability log-linearly, then the difference in N400 amplitude between lexical items with a lexical probability of 9% and 90% is comparable to that between lexical items with a lexical probability of 0.9% and 0.09%. However, due to the lack of resolution in cloze tests, the lexical items with lexical probability of 0.9% and 0.09% will likely both have a cloze rate of 0. When untransformed cloze scores are used as predictors, this pools all the low lexical probability items together despite the fact that they potentially elicit very different N400 amplitude (for further discussion, see Yan et al., 2017). The same potential issue applies to factorial designs that dichotomise cloze scores into, for example, “low” vs. “high”. Besides loss of resolution for low probability words, this also potentially violates the homoscedasticity assumption of standard analysis approaches, if N400 amplitude associated with items in the “low” condition vary more or less than those in the “high” condition.

If the N400 amplitude is linearly correlated with the surprisal, this finding can also have potential theoretical implications. The fact that both the amplitude of the N400 and reading times are linear in lexical surprisal can be seen as in suggestive (albeit weak) support of the argument that reading times and N400 amplitude both reflect lexical processing difficulty (Dimigen, Sommer, Hohlfeld, Jacobs, & Kliegl, 2011). Whether a measure is a linear predictor can also have implications for the underlying processing mechanism reflected by this measure (Smith & Levy, 2013), although we note that computationally explicit models of ERP components

are still largely lacking (for a notable exception, see Rabovsky et al., 2018; for further discussion, see also Yan et al., 2017).

Auto-correlation correction

Our approach to AC-correction results in fewer effects reaching significance. This is expected if the procedure successfully corrects for auto-correlations. The present study cannot, however, assess whether our approach indeed successfully corrects the Type I error rate, or to what extent our approach is conservative (resulting in reduced power). The answer to this question would require Type I and II error simulations that are beyond the scope of the present study. Here, we note three properties of our approach that should be considered when applying it to other data.

First, when analysing AC-corrected ERPs, one will only find significant effects at time points where the effect size differs from what is predicted by the effect sizes in the preceding time points. This is, of course, the purpose of AC-correction, but it means that analyses over corrected ERPs will thus be most sensitive where there is a drastic change in the effect size (e.g. at the onset/offset of an effect). For example, for sustained effects that are relatively stable in size across time, the analyses on corrected ERPs might only find significant results at the onset of the effect but not at later time points. This is not per se a limitation, but rather a property that needs to be considered when interpreting results over AC-corrected ERPs.

Second, our approach to AC-correction at present does not take into account uncertainty about the EEG activity at different preceding time lags. As a consequence, the approach might “over-react” to fluctuations in the preceding EEG activity, and this in turn might exaggerate or mask true effects. This shortcoming could be addressed in analyses techniques that model auto-correlations and the effects of interest at the same time, considering uncertainty about one while evaluating the other.

Third and finally, our approach does not currently account for spatial correlations across electrodes. This is not necessarily a hard limitation of the approach: just as the present approach predicts EEG activity at different lags of the same electrode (see Figure 4), one could potentially include EEG from any other electrode in this prediction process.

Models of context

Our results need to be interpreted in the context of the modelling choices we make. We thus lay out the simplifying assumptions we made in the formalisation of the

context predictors we considered in our analyses, and discuss how these might be improved in future studies.

The estimate of lexical surprisal employed in the present study almost exclusively reflects local lexical (n-gram) information. It does not incorporate non-local information, for example, discourse context, that is known to affect N400 amplitude (Van Berkum, Hagoort, & Brown, 1999). One way in which future work can address this shortcoming would be through cloze norms based on large scale norming studies, following similar recent behavioural work on reading (Luke & Christianson, 2017). However, to at least approximate the resolution of an n-gram or other computational models, one would need to collect data from thousands of subjects for each word. This quickly becomes infeasible for word-by-word estimates over large sets of sentences.

A second caveat to our findings originates in the representational assumptions made in the models of context employed in the present study. We represented the semantics of the context by simply adding the vectors of all the content words in the context. While such approach has received empirical support from other studies (Broderick, Anderson, Di Liberto, Crosse, & Lalor, 2018; Ettinger, Feldman, Resnik, & Phillips, 2016; Frank & Willems, 2017), it does not include other factor that are shown to affect meaning building and affects N400 amplitude (e.g. thematic role assignment, Rabovsky et al., 2018). The current results confirm that even with simple “bag-of-words” representation of the context semantics can explain the variances of N400 amplitude. It remains to be tested whether the semantic measures will have better explanatory power when other information that can influence meaning building is also incorporated into the model. An alternative approach is to build language models that explicit incorporate other information that affects meaning construction, e.g. thematic information. There are now models of this type, but they have so far only been applied to smaller “toy” corpora and cannot yet be scaled to the type of broad-coverage natural corpora investigated in the present study (Brouwer et al., 2017; Rabovsky et al., 2018).

Future directions

The use of computational models makes it possible to test hypotheses about N400 and other neural processes on large scale, more naturalistic stimuli. This allows one to test language processing in more ecological task environments. The dataset we analysed contains sentences from corpora that captures sentences similar to those encountered in daily life. However, the sentences were presented in isolation and devoid of discourse context. With the development of more naturalistic

EEG paradigms, e.g. gaze-contingent EEG (Dimigen et al., 2011; Hauk et al., 2017; Plöchl, Ossandón, & König, 2012), it has become possible to study ERPs over entire coherent discourse, rather than isolated sentences. The statistical methods we have employed here can be adopted in future work on the N400 and other ERP components related to language processing in more naturalistic contexts.

Notes

1. We originally also examined a related alternative measures of semantic distance, based on an interpretation of context vectors as probability distributions over latent semantic spaces. This measure correlates neither with early ERP components nor with N400 amplitude. The results reported here do not change if this predictor is included in the analysis.
2. Additionally, the smoothing technique employed in the n-gram model (Kneser-Key smoothing, Chen & Goodman, 1999) uses back-off to smooth unreliable estimates of n-gram probabilities. While this is an effective smoothing technique and thus desirable, it can further increase the correlation between the estimated n-gram probabilities and word frequency.
3. Typically, the ERP baseline is simply deducted from the ERP wave to perform baseline correction. This assumes that the effect of baselines is constant across ERPs at different times. Here we include the ERP baseline as a regressor to keep our results comparable to Frank and Willems (2017). This decision does not affect our results. As reported in the appendix, the ERP baseline had different effect on different parts of the ERP signal.

Acknowledgments

We are grateful to Dr. Stefan Frank for sharing the data, including EEG data, sentence materials, predictors from language models, and for patiently answering all our questions about his original experiments. We would also like to thank Dr. Gina Kuperberg for providing insightful feedbacks on earlier drafts of the paper. We really appreciate the highly constructive feedback of Dr. Milena Rabovsky, another anonymous reviewer and the editor. We would also like to thank the members of the Human Language Processing Lab – in particular, Zachary Burchill, Wednesday Bushong, Dr. Linda Liu – for proof reading earlier versions of the paper. All remaining mistakes are our own.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was funded by the NICHD R01 grant [HD075797] to T. Florian Jaeger; Division of Human Development.

References

- Arai, M., & Mazuka, R. (2014). The development of Japanese passive syntax as indexed by structural priming in comprehension. *Quarterly Journal of Experimental Psychology*, 67(1), 60–78. doi:10.1080/17470218.2013.790454
- Baayen, R. H., Feldman, L. B., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, 55(2), 290–313. doi:10.1016/j.jml.2006.03.008
- Baayen, R. H., van Rij, J., de Cat, C., & Wood, S. N. (2016). Autocorrelated errors in experimental data in the language sciences: Some solutions offered by generalized additive mixed models. In D. Speelman, K. Heylen, & D. Geeraerts (Eds.), *Mixed-effects regression models in linguistics* (pp. 49–69). Springer International Publishing AG. doi:10.1002/pssb.201300062
- Baayen, R. H., Vasishth, S., Kliegl, R., & Bates, D. (2017). The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, 94, 206–234. doi:10.1016/j.jml.2016.11.006
- Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., ... Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, 39(3), 445–459. doi:10.3758/BF03193014
- Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. doi:10.18637/jss.v067.i0
- Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, 28(5), 803–809.e3. doi:10.1016/j.cub.2018.01.080
- Brouwer, H., Crocker, M. W., Venhuizen, N. J., & Hoeks, J. C. J. (2017). A neurocomputational model of the N400 and the P600 in language processing. *Cognitive Science*, 41, 1318–1352. doi:10.1111/cogs.12461
- Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1), 1–28. doi:10.18637/jss.v080.i01
- Burns, M. D., Bigdely-Shamlo, N., Smith, N. J., Kreutz-Delgado, K., & Makeig, S. (2013). Comparison of averaging and regression techniques for estimating Event Related Potentials. In *Conference proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference* (Vol. 2013, pp. 1680–1683). doi:10.1109/EMBC.2013.6609841
- Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, 113(2), 234–272. doi:10.1037/0033-295X.113.2.234
- Chen, S. F., & Goodman, J. T. J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, 13(13), 359–393.
- Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2008). Heeding the voice of experience: The role of talker variation in lexical access. *Cognition*, 106(2), 633–664. doi:10.1016/j.cognition.2007.03.013
- Dambacher, M., Kliegl, R., Hofmann, M., & Jacobs, A. M. (2006). Frequency and predictability effects on event-related potentials during reading. *Brain Research*, 1084(1), 89–103. doi:10.1016/j.brainres.2006.02.010

- Delaney-Busch, N., Lau, E. F., Morgan, E., & Kuperberg, G. R. (2019). Neural evidence for bayesian trial-by-trial adaptation on the N400 during semantic priming. *Cognition*, 187, 10–20. doi:10.1016/j.cognition.2019.01.001.
- Dell, G. S., & Chang, F. (2014). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1634), 20120394. doi:10.1098/rstb.2012.0394
- Dell, G. S., & O'Seaghdha, P. G. (1992). Stages of lexical access in language production. *Cognition*, 42(1–3), 287–314. doi:10.1016/0010-0277(92)90046-K
- Dimigen, O., Sommer, W., Hohlfeld, A., Jacobs, A. M., & Kliegl, R. (2011). Coregistration of eye movements and EEG in natural reading: Analyses and review. *Journal of Experimental Psychology: General*, 140(4), 552–572. doi:10.1037/a0023885
- Domahs, U., Klein, E., Huber, W., & Domahs, F. (2013). Good, bad and ugly word stress – fMRI evidence for foot structure driven processing of prosodic violations. *Brain and Language*, 125(3), 272–282. doi:10.1016/j.bandl.2013.02.012
- Elston-Güttler, K. E., Gunter, T. C., & Kotz, S. A. (2005). Zooming into L2: Global language context and adjustment affect processing of interlingual homographs in sentences. *Cognitive Brain Research*, 25(1), 57–70. doi:10.1016/j.cogbrainres.2005.04.007
- Ettinger, A., Feldman, N. H., Resnik, P., & Phillips, C. (2016). Modeling N400 amplitude using vector space models of word representation. In *Proceedings of the 38th annual conference of the Cognitive Science Society* (pp. 1445–1450).
- Farmer, T. A., Monaghan, P., Misyak, J. B., & Christiansen, M. H. (2011). Phonological typicality influences sentence processing in predictive contexts: Reply to Staub, Grant, Clifton, and Rayner (2009). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(5), 1318–1325. doi:10.1037/a0023063
- Federmeier, K. D., Mai, H., & Kutas, M. (2005). Both sides get the point: Hemispheric sensitivities to sentential constraint. *Memory & Cognition*, 33(5), 871–886. doi:10.3758/BF03193082
- Fine, A. B., & Jaeger, T. F. (2016). The role of verb repetition in cumulative structural priming in comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(9), 1362–1376. doi:10.1037/xlm0000236
- Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS ONE*, 8(10), e77661. doi:10.1371/journal.pone.0077661
- Frank, S. L., Fernandez Monsalve, I., Thompson, R. L., & Vigliocco, G. (2013). Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods*, 45(4), 1182–1190. doi:10.3758/s13428-012-0313-y
- Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, 140, 1–11. doi:10.1016/j.bandl.2014.10.006
- Frank, S. L., & Willems, R. M. (2017). Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Language, Cognition and Neuroscience*, 32(9), 1192–1203. doi:10.1080/23273798.2017.1323109
- Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics* (pp. 10–18).
- Grainger, J., & Holcomb, P. J. (2009). Watching the word go by: On the time-course of component processes in visual word recognition. *Language and Linguistics Compass*, 3(1), 128–156. doi:10.1111/j.1749-818X.2008.00121.x
- Groppe, D. M., Urbach, T. P., & Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. *Psychophysiology*, 48(12), 1711–1725. doi:10.1016/j.micinf.2011.07.011.Innate
- Guthrie, D., & Buchwald, J. S. (1991). Significance testing of difference potentials. *Psychophysiology*, 28(2), 240–244. doi:10.1111/j.1469-8986.1991.tb00417.x
- Hanulíková, A., van Alphen, P. M., van Goch, M. M., & Weber, A. (2012). When one person's mistake is another's standard usage: The effect of foreign accent on syntactic processing. *Journal of Cognitive Neuroscience*, 24(4), 878–887. doi:10.1162/jocn_a_00103
- Hauk, O., Davis, M. H., Ford, M., Pulvermüller, F., & Marslen-Wilson, W. D. (2006). The time course of visual word recognition as revealed by linear regression analysis of ERP data. *NeuroImage*, 30(4), 1383–1400. doi:10.1016/j.neuroimage.2005.11.048
- Hauk, O., Dreyer, F., van Casteren, M., Coutout, C., Fonteneau, E., & Weiss, B. (2017). Investigating brain mechanisms of natural reading by combining EEG, MEG and eye-tracking. In *Proceedings of the 9th Annual Meeting of the Society for the Neurobiology of Language* (pp. 222–223).
- Hauk, O., Pulvermüller, F., Ford, M., Marslen-Wilson, W. D., & Davis, M. H. (2009). Can I have a quick word? Early electrophysiological manifestations of psycholinguistic processes revealed by event-related regression analysis of the EEG. *Biological Psychology*, 80(1), 64–74. doi:10.1016/j.biopsycho.2008.04.015
- Hendrix, P., Baayen, R. H., & Bolger, P. (2017). Distinct ERP signatures of word frequency, phrase frequency, and prototypicality in speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(1), 128–149. doi:10.1037/a0040332
- Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1), 23–62. doi:10.1016/j.cogpsych.2010.02.002
- Kaschak, M. P., & Glenberg, A. M. (2004). This construction needs learned. *Journal of Experimental Psychology: General*, 133(3), 450–467. doi:10.1037/0096-3445.133.3.450
- Kim, A., & Lai, V. (2012). Rapid interactions between lexical semantic and word form analysis during word recognition in context: Evidence from ERPs. *Journal of Cognitive Neuroscience*, 24(5), 1104–1112. doi:10.1162/jocn_a_00148
- Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, 16(1/2), 262–284. doi:10.1080/09541440340000213
- Kuperberg, G. R. (2016). Separate streams or probabilistic inference? What the N400 can tell us about the comprehension of events. *Language, Cognition and Neuroscience*, 31(5), 602–616. doi:10.1080/23273798.2015.1130233
- Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, 31(1), 32–59. doi:10.1080/23273798.2015.1102299

- Kurumada, C., Brown, M., & Bibyk, S. (2014). Rapid adaptation in online pragmatic interpretation of contrastive prosody. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society* (pp. 791–796).
- Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event related brain potential (ERP). *Annual Review of Psychology*, 62, 621–647. doi:10.1146/annurev.psych.093008.131123
- Laszlo, S., & Federmeier, K. D. (2014). Never seem to find the time: Evaluating the physiological time course of visual word recognition with regression analysis of single-item event-related potentials. *Language, Cognition and Neuroscience*, 29(5), 642–661. doi:10.1080/01690965.2013.866259
- Lee, C.-Y., Liu, Y.-N., & Tsai, J.-L. (2012, August). The time course of contextual effects on visual word recognition. *Frontiers in Psychology*, 3, 1–13. doi:10.3389/fpsyg.2012.00285
- Lorch, R. F., & Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(1), 149–157. doi:10.1037/0278-7393.16.1.149
- Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology*, 88, 22–60. doi:10.1016/j.cogpsych.2016.06.002
- Luke, S. G., & Christianson, K. (2017). The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 1–8. doi:10.3758/s13428-017-0908-4
- Macizo, P., & Herrera, A. (2011). Cognitive control in number processing: Evidence from the unit-decade compatibility effect. *Acta Psychologica*, 136(1), 112–118. doi:10.1016/j.actpsy.2010.10.008
- McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18(1), 1–86. doi:10.1016/0010-0285(86)90015-0
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. doi:10.1162/jmlr.2003.3.4-5.951
- Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, 59(4), 475–494. doi:10.1016/j.jml.2007.11.006
- Narum, S. R. (2006). Beyond Bonferroni: Less conservative analyses for conservation genetics. *Conservation Genetics*, 7(5), 783–787. doi:10.1007/s10592-005-9056-y
- Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., ... Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife*, 7, e33468. doi:10.7554/eLife.33468
- Nixon, J. S., van Rij, J., Mok, P., Baayen, R. H., & Chen, Y. (2016). The temporal dynamics of perceptual uncertainty: Eye movement evidence from Cantonese segment and tone perception. *Journal of Memory and Language*, 90, 103–125. doi:10.1016/j.jml.2016.03.005
- Norris, D., McQueen, J. M., & Cutler, A. (2015, December). Prediction, Bayesian inference and feedback in speech recognition. *Language, Cognition and Neuroscience*, 3798, 1–15. doi:10.1080/23273798.2015.1081703
- Parviz, M., Johnson, M., Johnson, B., & Brock, J. (2011). Using language models and latent semantic analysis to characterise the N400m neural response. In *Proceedings of the Australasian Language Technology Association Workshop 2011* (pp. 38–46). Retrieved from <http://aclweb.org/anthology/U11-1007>
- Payne, B. R., Lee, C. L., & Federmeier, K. D. (2015). Revisiting the incremental effects of context on word processing: Evidence from single-word event-related brain potentials. *Psychophysiology*, 52(11), 1456–1469. doi:10.1111/psyp.12515
- Penolazzi, B., Hauk, O., & Pulvermüller, F. (2007). Early semantic context integration and lexical access as revealed by event-related brain potentials. *Biological Psychology*, 74(3), 374–388. doi:10.1016/j.biopsycho.2006.09.008
- Piai, V., Dahlslett, K., & Maris, E. (2015). Statistically comparing EEG/MEG waveforms through successive significant univariate tests: How bad can it be? *Psychophysiology*, 52(3), 440–443. doi:10.1111/psyp.12335
- Plöchl, M., Ossandón, J. P., & König, P. (2012). Combining EEG and eye tracking: identification, characterization, and correction of eye movement artifacts in electroencephalographic data. *Frontiers in Human Neuroscience*, 6, 278. doi:10.3389/fnhum.2012.00278
- Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, 2(9), 693–705. doi:10.1038/s41562-018-0406-4
- R: A language and environment for statistical computing. (2016). *R foundation for statistical computing*, Vienna, Austria. R Core Team. Retrieved from <https://www.r-project.org/about.html>
- Schäfer, R. (2015). Processing and querying large web corpora with the COW14 architecture. In *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3)* (pp. 28–34). Retrieved from <http://rolandschaefer.net/cow14-cmlc3-paper/>
- Seidenberg, M. S., & MacDonald, M. C. (1999). A probabilistic constraints approach to language acquisition and processing. *Cognitive Science*, 23(4), 569–588. doi:10.1207/s15516709cog2304
- Sereno, S. C., Posner, M. I., & Rayner, K. (1998). Establishing a timeline of word recognition: Evidence from eye movements and event-related potentials. *Neuroreport*, 9(10), 2195–2200.
- Smith, N. J., & Kutas, M. (2015a). Regression-based estimation of ERP waveforms: I. The rERP framework. *Psychophysiology*, 52(2), 157–168. doi:10.1111/psyp.12317
- Smith, N. J., & Kutas, M. (2015b). Regression-based estimation of ERP waveforms: II. Nonlinear effects, overlap correction, and practical considerations. *Psychophysiology*, 52(2), 169–181. doi:10.1111/psyp.12320
- Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3), 302–319. doi:10.1016/j.cognition.2013.02.013
- Staub, A., Grant, M., Clifton, C., & Rayner, K. (2009). Phonological typicality does not influence fixation durations in normal reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(3), 806–814. doi:10.1037/a0015123
- Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217), 1632–1634. doi:10.1126/science.7777863
- Tremblay, A., & Newman, A. J. (2015). Modeling nonlinear relationships in ERP data using mixed-effects regression with R examples. *Psychophysiology*, 52(1), 124–139. doi:10.1111/psyp.12299
- Van Berkum, J. J. A., Hagoort, P., & Brown, C. M. (1999). Semantic integration in sentences and discourse: Evidence from the

- N400. *Journal of Cognitive Neuroscience*, 11(6), 657–671. doi:10.1162/089892999563724
- Van Petten, C., & Kutas, M. (1990). Interactions between sentence context and word frequency in event-related brain potentials. *Memory & Cognition*, 18(4), 380–393. doi:10.3758/BF03197127
- Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, 83(2), 176–190. doi:10.1016/j.ijpsycho.2011.09.015
- van Rij, J. (2016). Testing for significance. Retrieved from <https://cran.r-project.org/web/packages/itsadug/vignettes/test.html>
- van Rij, J., Wieling, M., Baayen, R. H., & van Rijn, H. (2017). itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs. *R package*, version 2.2.
- Wood, S. N. (2006). *Generalized additive models*. doi:10.1002/0471667196.ess0297.pub2
- Yan, S., Kuperberg, G. R., & Jaeger, T. F. (2017). Prediction (or not) during language processing. A commentary on Nieuwland et al. (2017) and DeLong et al. (2005). *BioRxiv*. Retrieved from <http://biorxiv.org/content/early/2017/05/30/143750.abstract>
- Zipf, G. K. (1949). *Human behaviour and the principle of least-effort*. New York: Addison-Wesley.