Dissociable effects of prediction and integration during language comprehension: Evidence from a large-scale study using brain potentials

Mante S. Nieuwland[1,2], Dale J. Barr[3], Federica Bartolozzi[1,2], Simon Busch-Moreno[4], Emily Darley[5], David I. Donaldson[6], Heather J. Ferguson[7], Xiao Fu[4], Evelien Heyselaar[1,8], Falk Huettig[1], E. Matthew Husband[9], Aine Ito[2,9], Nina Kazanina[5], Vita Kogan[2], Zdenko Kohút[10], Eugenia Kulakova[11], Diane Mézière[2], Stephen Politzer-Ahles[9,12], Guillaume Rousselet[3], Shirley-Ann Rueschemeyer[10], Katrien Segaert[8], Jyrki Tuomainen[4], Sarah Von Grebmer Zu Wolfsthurn[5]

[1]Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands

[2]School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh, United Kingdom

[3]Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, United Kingdom

[4]Division of Psychology and Language Sciences, University College London, London, United Kingdom

[5]School of Experimental Psychology, University of Bristol, Bristol, United Kingdom

[6]Psychology, Faculty of Natural Sciences, University of Stirling, Stirling, United Kingdom

[7]School of Psychology, University of Kent, Canterbury, United Kingdom

[8]School of Psychology, University of Birmingham, Birmingham, United Kingdom

[9]Faculty of Linguistics, Philology & Phonetics; University of Oxford, Oxford, United Kingdom

[10]Department of Psychology, University of York, York, United Kingdom

[11]Institute of Cognitive Neuroscience, University College London, London, United Kingdom

[12]Department of Chinese and Bilingual Studies, the Hong Kong Polytechnic University, Kowloon, Hong Kong

Author note: In accordance with the Peer Review Openness Initiative (Morey et al., 2016), all materials, data and analysis scripts are available on the Open Science Framework for review purposes, and will be made publicly available upon publication of this manuscript.

Abstract

What makes predictable words easier to process than unpredictable words (e.g., ‘bicycle’ compared to ‘elephant’ in “You never forget how to ride a bicycle/an elephant once you’ve learned”)? Are predictable words genuinely predicted, or simply more plausible and therefore easier to integrate with sentence context? We addressed this persistent and fundamental question using data from a recent, large-scale (N = 334) replication study, by investigating the effects of word predictability and plausibility on the N400, the brain’s electrophysiological index of semantic processing. A spatiotemporally fine-grained, mixed-effects multiple regression analysis revealed overlapping effects of predictability and plausibility on the N400, albeit with distinct spatiotemporal profiles. Our results challenge the view that semantic facilitation of predictable words reflects the effects of *either* prediction *or* integration, and suggest that facilitation arises from a cascade of processes that access and integrate word meaning with context into a sentence-level meaning.

Introduction

Predictable words are easier to process than unpredictable words: 'bicycle' is easier to process than 'elephant' in "You never forget how to ride a bicycle/an elephant once you've learned". For example, predictable words are read and recognized faster than unpredictable words (e.g., Clifton, Staub & Rayner, 2007; Stanovich & West, 1981). However, it remains unclear whether such facilitation is driven by actual prediction (i.e., predictable words are activated before they appear), by integration (i.e., predictable words are semantically more plausible and therefore easier to integrate into sentence context than unpredictable words after they have appeared), or by both. We addressed this issue by exploring modulation of the N400 (Kutas & Hillyard, 1980), an event-related potential (ERP) component commonly considered the brain's index of semantic processing (Kutas & Federmeier, 2011). In a temporally fine-grained analysis of data from a large-scale (*N*=334) replication study (Nieuwland et al., 2017), we investigated whether prediction and integration have dissociable effects on N400 amplitude, and how these effects unfold over time.

Word predictability[1] is strongly correlated with an amplitude reduction of the N400 (Kutas & Hillyard, 1984), a negative deflection with a centroparietal scalp-distribution that occurs approximately 200-500 ms after word onset and peaks around 400 ms. Traditionally, there have been two opposing views on what this pattern means. According to the 'integration view' (e.g., Brown & Hagoort, 1993; Hagoort, Hald, Bastiaansen & Petersson, 2004; Van Berkum, Hagoort & Brown, 1999), the N400 directly reflects the integration of an activated word meaning with sentence context and general world knowledge. Integration is required to compose higher-level sentence meaning from individual word meanings, and is easier for predictable words because they yield a sentence meaning that is more plausible

[1] Defined as the likelihood of being used in an offline sentence completion test ('cloze probability'; Taylor, 1953).

with regard to world knowledge. Alternatively, according to the 'access view' (e.g., Brouwer, Fitz & Hoeks, 2012; Lau, Almeida, Hines & Poeppel, 2009; Kutas & Federmeier, 2000, 2011), the N400 reflects access to word meaning in long-term memory. People easily access the meaning of a predictable word because they pre-activate some (or all) of its meaning based on the context. In this view, integration of this word meaning with context does not happen until after the N400 component (Bornkessel-Schlesewsky & Schlesewsky, 2008; Brouwer et al., 2012; Kutas & Federmeier, 2000). These mutually exclusive views thus differ on whether or not people predict the meaning of an upcoming word, and on whether the N400 reflects a compositional process (combining word meanings into a higher-order representation) or a non-compositional process (accessing the meaning of a single word).

The integration-access debate has long engrossed the psychology and neuroscience of language (for reviews, see Kutas & Federmeier, 2011; Lau, Phillips & Poeppel, 2008; Van Berkum, 2009), but it has yet to reach a conclusion, and there is support for both views. Supporting the access-view (Kutas & Federmeier, 2011), numerous studies show that people can predict the meaning of upcoming words during sentence comprehension (for reviews, see Federmeier, 2007; Van Petten & Luka, 2012). In addition, some studies suggest that N400 amplitude is not a function of sentence plausibility[2] (e.g., Federmeier & Kutas, 1999; Ito, Corley, Pickering, Martin & Nieuwland, 2016), a pattern that is incompatible with the integration view. Supporting the integration view, however, several studies report N400 modulations by semantic or pragmatic plausibility that are not easily explained in terms of prediction alone (e.g., Calloway & Perfetti, 2017; Li, Hagoort & Yang, 2008; Rueschemeyer, Gardner & Stoner, 2015; see also Lau, Namyst, Fogel & Delgado, 2016; Steinhauer, Royle, Drury & Fromont, 2017).

[2] The plausibility of the described event with regard to real-world knowledge, as established in an offline plausibility-norming test.

This mixed evidence has led some researchers to question the viability of an access-only or integration-only view of the N400, and to propose a hybrid, '*multiple-process*' account (Baggio, 2012; Baggio & Hagoort, 2011; Lau et al., 2016; Newman, Forbes & Connolly, 2012). This account views N400 activity as reflecting cascading access- and integration-processes. Effects of prediction and of integration are therefore both visible in N400 activity, but effects of prediction would precede and be functionally distinct from those of integration. Consistent with this account, Brothers, Swaab and Traxler (2015) found that effects of prediction appeared earlier than effects of contextual integration, and on distinct ERP components (N250 and N400). However, because their participants were instructed to actively predict sentence-final words, the observed ERP patterns also reflected task-relevant decision-processes (Polich, 2007; Roehm, Bornkessel-Schlesewsky, Rösler & Schlesewsky, 2007), and may not generalize to situations where people do not strategically predict upcoming words.

Here, we tested the multi-process hypothesis using data from a direct replication attempt of a landmark study on prediction (DeLong, Urbach & Kutas, 2005). DeLong et al. capitalized on the phonotactic dependence of the English indefinite articles 'a/an' on whether the next word starts with a consonant or vowel. Participants read sentences containing an indefinite article (a/an) followed by a noun. The article-noun pairs were always morphologically consistent but differed in their predictability from sentence context (e.g., "You never forget how to ride <u>a bicycle</u>/<u>an elephant</u> once you've learned"). As expected, amplitude of the noun-elicited N400s gradually decreased with increased predictability (Kutas & Hillyard, 1984). Critically, however, DeLong et al. also observed this pattern of results at the preceding articles, which cannot arise from differences in the meaning of 'a/an' and therefore does not index integration costs. The article-effect was taken as strong evidence that participants predicted the nouns, including their phonological form (i.e., whether they

start with a consonant or vowel), and that the articles that disconfirmed this prediction resulted in processing difficulty (higher N400 amplitude at the article).

In a large-scale, direct replication attempt spanning 9 labs (Nieuwland et al., 2017), our pre-registered analyses failed to replicate the article-effect but successfully replicated the noun-effect. Crucially, without strong article-evidence for prediction, it remains uncertain whether the noun-effects reflect prediction, integration, or both. We therefore performed a further (non-pre-registered) analysis to dissociate the effects of prediction and integration. Like previous studies (Federmeier & Kutas, 1999; Ito et al., 2016), we investigated their effects by examining N400 activity as a function of word predictability and plausibility, which we established in offline norms. Improving on previously used methods, we simultaneously modelled variance associated with predictability and plausibility, while also controlling for semantic similarity (Landauer & Dumais, 1997), a measure of low-level semantic relatedness between word and context derived from distributional semantics. Using a spatiotemporally fine-grained analysis, modelling activity at each EEG channel and time-point within an extended time window (e.g., Hauk, Davis, Ford, Pulvermüller, & Marslen-Wilson, 2006), we examined the time-course and spatial distribution of the effect of predictability while appropriately controlling for plausibility and vice versa (Sassenhagen & Alday, 2016; see also Frank & Willems, 2017, using a similar approach to regress effects of corpus-based statistics).

If N400 amplitude reflects only the effects of prediction and not of integration (Brouwer et al., 2012; Kutas & Federmeier, 2000), plausibility would not impact processing until after the time window typically associated with the N400 component. In contrast, if the N400 reflects effects of prediction and integration (Baggio & Hagoort, 2011; Lau et al., 2016), plausibility would have an effect on N400 activity alongside the effect of

predictability, although any effects of plausibility would occur later than effects of predictability.

Methods

Our materials were the 80 sentences in two conditions (expected/unexpected article-noun combination), used by DeLong et al. (2005). Participants were native English speaking students from the University of Birmingham, Bristol, Edinburgh, Glasgow, Kent, Oxford, Stirling, York, or volunteers from the participant pool of University College London or Oxford University, who received cash or course credit for taking part in the ERP experiment. Each laboratory aimed to test 40 participants and tested at least 32 participants, which was the sample size of DeLong et al., (2005). Our data pre-processing was identical to Nieuwland et al., 2017, which used a pre-registered procedure (see https://osf.io/eyzaq) that led to the exclusion of 22 participants from the initial group of 356 participants, leaving a sample size of 334 participants.

Details about the stimulus materials, participants, EEG recording equipment and settings, and experimental procedure are described elsewhere (Nieuwland et al., 2017). This section only describes the changes and extensions to the methods from that study. A full list of the materials, including all norming results, is available as Supplementary materials.

Predictability and Plausibility pre-tests

Prior to collecting EEG data, we conducted predictability and plausibility pre-tests. For the predictability (cloze probability) pre-test, we truncated all sentences after the critical indefinite article and asked participants to complete each sentence with the first word or words that came to mind (for details, see Nieuwland et al., 2017). We presented two counterbalanced lists of 80 sentences to 30 participants each, such that no participant saw the same sentence context with the expected and the unexpected article. We computed the

predictability of each word as the percentage of participants who used the word to complete the sentence.

For the plausibility pre-test, we truncated all sentences after the critical nouns. We presented two counterbalanced lists of 80 sentences to 31 participants each, such that no participant saw the same sentence context with the expected and the unexpected noun. All participants were volunteers from the University of Edinburgh, who did not participate in the predictability pre-tests or the ERP experiment. They were asked to judge "the plausibility of the events described in the sentences", on a scale from 1 to 7 (from very implausible to very plausible, respectively; other values on the range were shown without a verbal label). We computed a plausibility score for each word as the average of the obtained plausibility ratings over participants. On average, relatively expected nouns were rated as plausible (M = 5.9, SD = 0.48) whereas relatively unexpected nouns were rated as neither plausible nor implausible (M = 3.8, SD = 1.31).

Mixed-effects multiple regression

After pre-processing, artefact-free segments were resampled to 250 Hz (i.e., one sample every 4 ms). Then, for each sample between -100 to 1000 ms relative to noun onset, and for each channel, we performed the following mixed-effects model analysis (Baayen, Davidson & Bates, 2008) using the 'lme4' package (Bates, Maechler, Bolker & Walker, 2014) as implemented in R (R CoreTeam, 2014):

eegdata ~ predictability + plausibility + similarity + laboratory + (predictability + plausibility + similarity || subject) + (predictability + plausibility + similarity || item)

Both predictability and plausibility were z-scored, and we removed random correlations to facilitate model convergence. We also included an additional fixed effect ('similarity') in order to control for semantic similarity between the critical word and the sentence context, which can influence N400 amplitude (Brothers, Swaab & Traxler, 2015;

Nieuwland et al., 2010; Van Petten, 2014). We measured z-transformed semantic similarity values obtained from pairwise Latent Semantic Analysis (LSA; Landauer & Dumais, 1997). LSA is a statistical technique for extracting and representing the similarity of meaning of words and text by analysis of large bodies of text. Our analysis used the General Reading – Up to First Year of College topic space (lsa.colorado.edu). In addition, the factor 'laboratory' was included as a deviation-coded, categorical fixed effect variable. Although this factor was not of theoretical interest, it was included because a previous pre-registered analysis (Nieuwland et al., 2017) showed that although the laboratories did not show significantly different N400 effects of noun-predictability, they did significantly differ in overall N400 amplitude (i.e., a main effect of 'laboratory'). Analysis without the factor 'laboratory' did not change the observed patterns of results and can be reproduced from our online data set.

Our measures of predictability and plausibility are more strongly correlated (r=0.72) than predictability and semantic similarity (r=0.19), and plausibility and semantic similarity (r=0.16). However, these correlation coefficients are not a principled obstacle to our approach. Variance Inflation Factors (VIF) for our continuous predictors were all below 2.1, which is well below the values deemed problematic due to high multicollinearity (e.g., Zuur, Ieno, Elphick, 2010).

For each analysis, we extracted a coefficient estimate with 95% confidence interval (CI), a t-value and p-value associated with each fixed effect except for 'laboratory'. We computed confidence intervals with the 'Wald' method, and p-values with the normal approximation.

Correction for multiple comparisons

We corrected for multiple comparisons using the Benjamini and Hochberg (1995) method to control the false discovery rate, the expected proportion of false discoveries amongst the rejected hypotheses. For predictability, plausibility and semantic similarity

separately, we applied this correction (implemented in R's p.adjust) to p-values associated with samples from all electrodes in three time windows of interest (1-200, 200-500, 500-1000 ms). Our main window of interest regarding N400 activity was the 200-500 ms time window, the pre-registered window of analysis in Nieuwland et al. (2017), which followed DeLong et al. (2005). We applied the correction separately to each window of interest because the false discovery rate procedure, when applied to the entire window, can be too lenient outside the 200-500 ms time window containing large N400 effects (see also DeLong, Quante & Kutas, 2014; Groppe, Urbach & Kutas, 2011).

Additional exploratory interaction analyses

We also explored potential interactions between our fixed effects, although we did not have an a priori theoretical basis to expect the effect size of plausibility or semantic similarity to change with predictability (or vice versa), nor to expect the effect size of plausibility to depend on semantic similarity (or vice versa). We repeated our analysis with the inclusion of all two-way interaction terms between the fixed effects predictability, plausibility and semantic similarity. To facilitate convergence and reduce computing time we did not include random slopes for the interaction terms. We applied the same correction for multiple comparisons to the resulting *p*-values as described previously.

Results

More predictable nouns elicited more positive amplitude (likely indicating a smaller N400 component) than unpredictable nouns within the N400 time window (200-500 ms) across all channels (see Figure 1). This effect was statistically significant as early as 200 ms after word onset at multiple channels, and peaked around 330 ms. Following the N400 component, the pattern of activation reversed, such that more predictable nouns elicited a more negative deflection than less predictable items. This post-N400 waveform was

statistically significant at frontal and central channels already within the 200-500 ms time window (see also DeLong et al., 2014), and appeared stronger and more extended at left- compared to right-hemisphere channels.

More plausible nouns elicited more positive (smaller) amplitude than implausible nouns within the N400 time window (200-500 ms) (see Figure 2). In contrast to the pattern observed for predictability, the effect of plausibility showed a less peaked, more extended time course that continued until about 650 ms after word onset. The effect of plausibility became statistically significant at about 350 ms after word onset, thus after the peak effect of predictability, and was most pronounced over right-posterior electrode sites.

We did not observe statistically significant effects of semantic similarity in the N400 time window, but semantically more similar words elicited more negative voltage than dissimilar words between 600-1000 ms at all channels (Figure 3).

To facilitate comparison of the effects of predictability, plausibility and semantic similarity, Figure 4 plots the scalp distribution of their respective effects in 50 ms time bins.

In our additional exploratory analysis of potential interactions, none of the interaction terms elicited significant effects after multiple comparison correction, although trends were visible in the N400 window: the effect of plausibility and of similarity became smaller with increasing predictability, whereas the effect of plausibility became greater with increasing similarity. The main effects of predictability and of and semantic similarity remained largely similar to the effects observed in absence of interaction terms. The main effect of plausibility was reduced to only a few statistically significant samples, although the direction of the plausibility effect remained clearly visible. Code and figures associated with this interaction analysis are available online along with the rest of our materials.

Figure 1. Effect of noun predictability (cloze probability) when controlling for noun plausibility. Because we z-transformed our measure of predictability, the voltage estimates (blue lines) and corresponding 95% confidence intervals (grey area) represent the change in voltage, for each time sample and EEG channel, associated with a 1 standard deviation increase in predictability. Dots underneath the voltage estimates indicate statistically significant samples after multiple comparisons correction based on the false discovery rate.

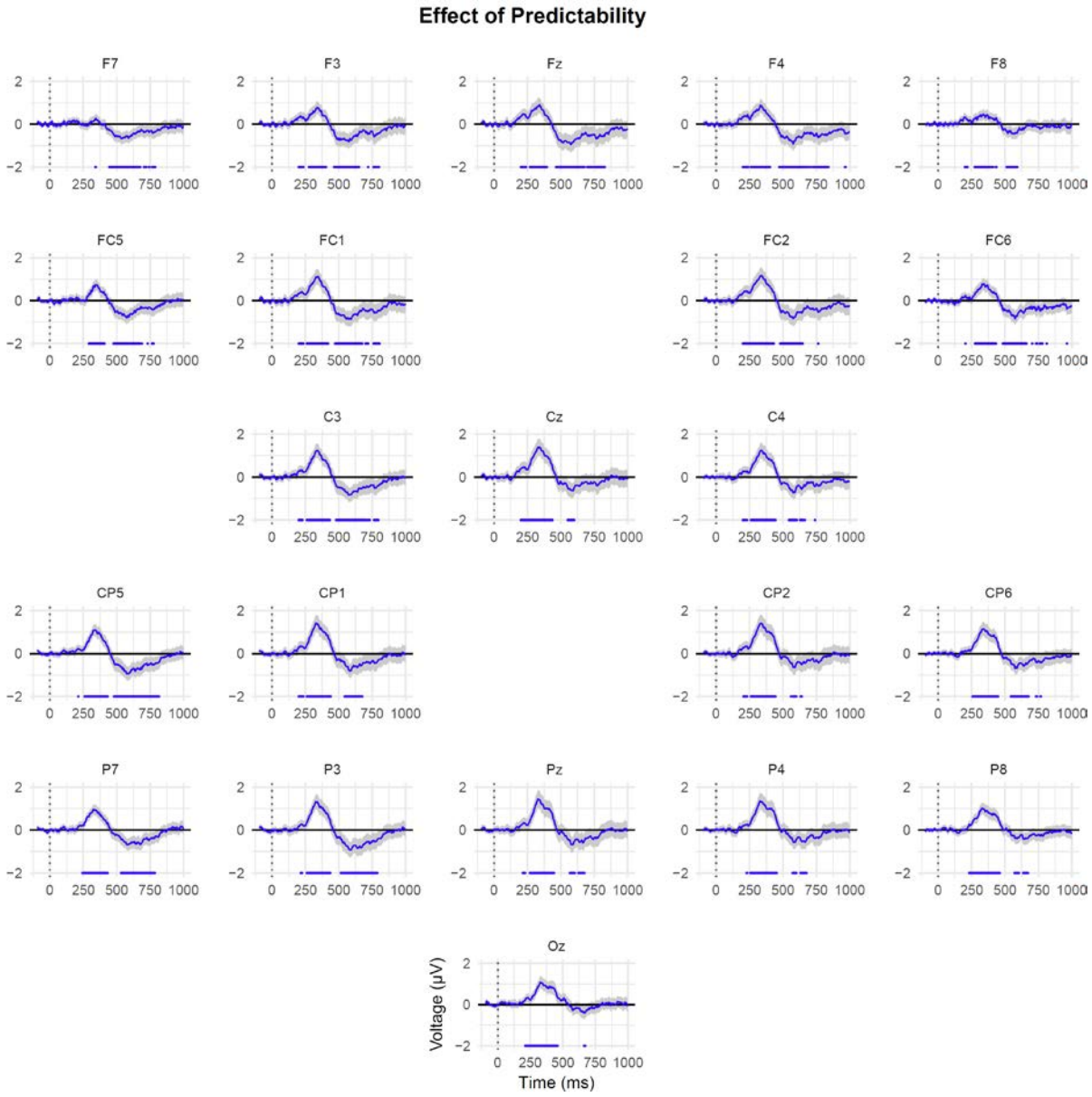


Figure 2. Effect of noun plausibility when controlling for noun predictability (cloze probability). Because we z-transformed our measure of plausibility, the voltage estimates (red lines) and corresponding 95% confidence intervals (grey area) represent the change in voltage, for each time sample and EEG channel, associated with a 1 standard deviation increase in plausibility. Dots underneath the voltage estimates indicate statistically significant samples after multiple comparisons correction based on the false discovery rate.

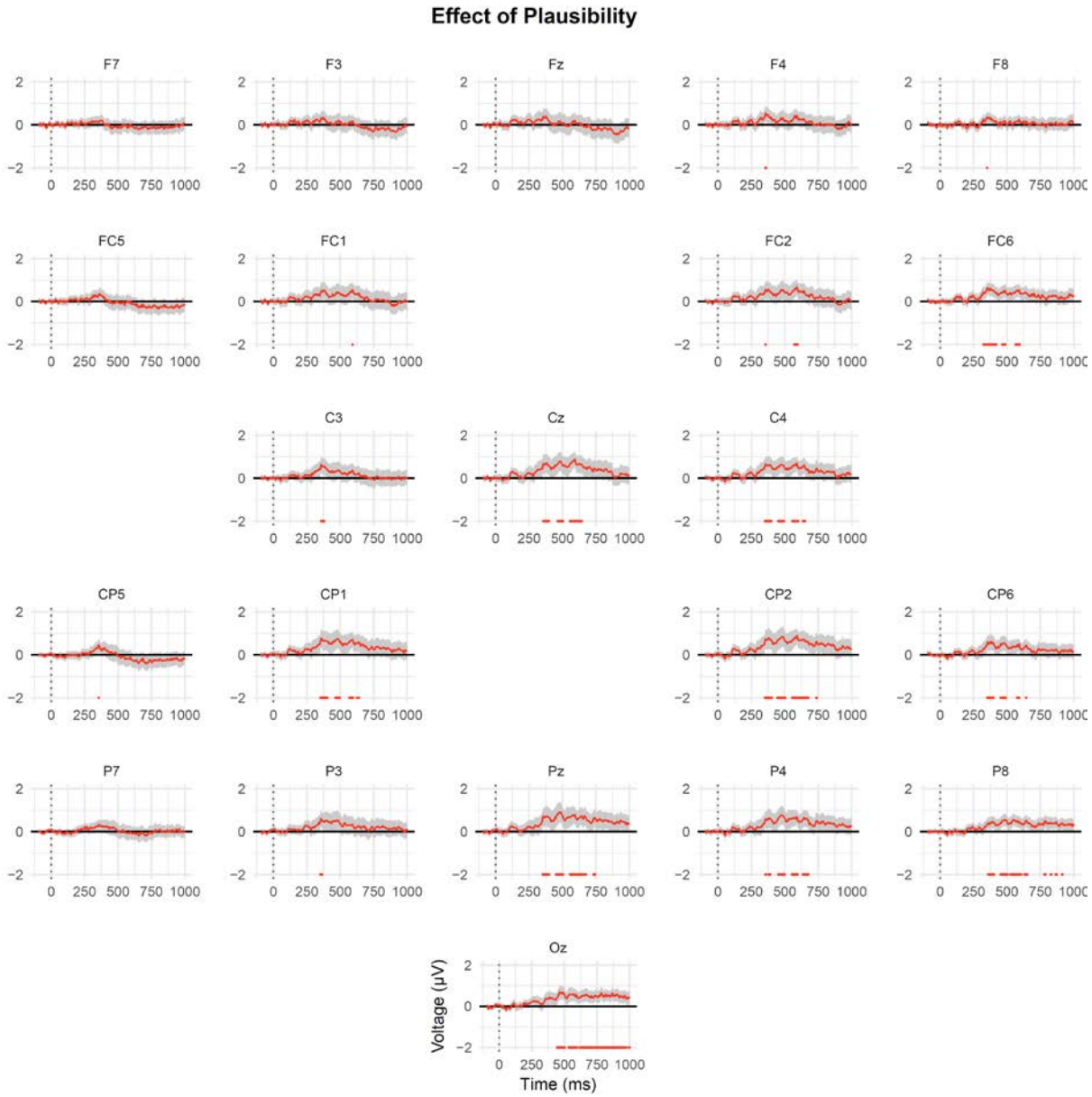


Figure 3. Effect of semantic similarity between the critical noun and the sentence context. Because we z-transformed our measure of semantic similarity, the voltage estimates (blue lines) and corresponding 95% confidence intervals (grey area) represent the change in voltage, for each time sample and EEG channel, associated with a 1 standard deviation increase in semantic similarity. Dots underneath the voltage estimates indicate statistically significant samples after multiple comparisons correction based on the false discovery rate.

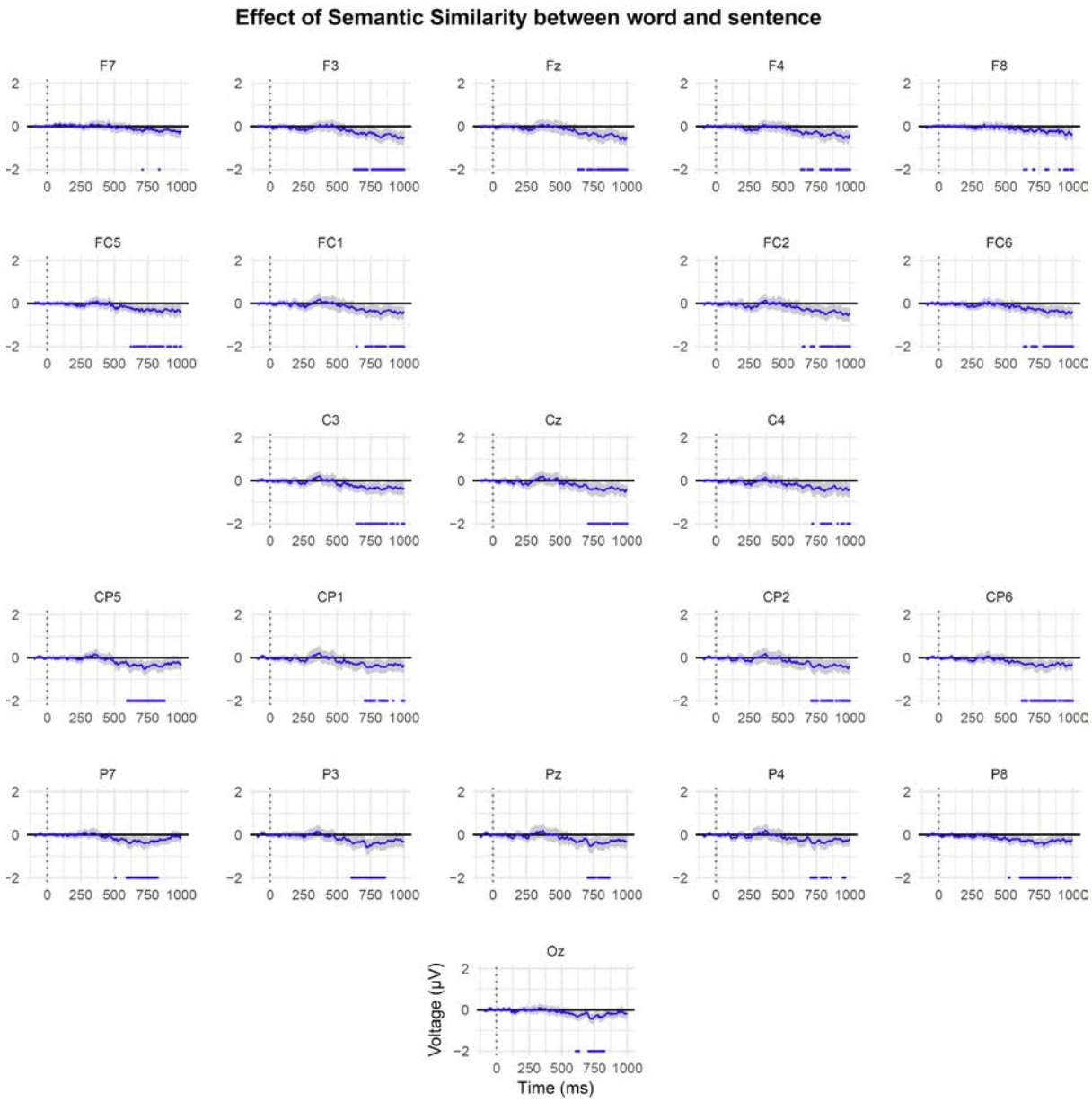
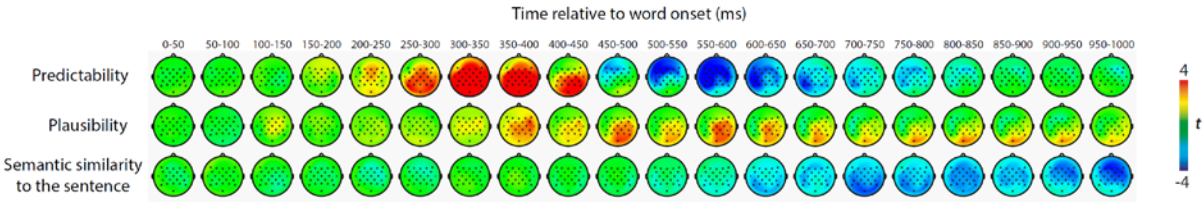


Figure 4. Scalp distribution associated with the effects of predictability, plausibility and semantic similarity between the critical noun and the sentence context. The colour scale indexes the average *t*-value within a 50 ms time window relative to word onset.



Discussion

We observed the combined effects of word predictability and sentence plausibility in activity of the N400, the brain's well-known index of semantic processing (Kutas & Federmeier, 2011). Our results suggest that, compared to unpredictable words, predictable words are relatively easy to process for two reasons: (1) their meaning is at least partly predicted (i.e., activated prior to presentation), and (2) they are easier to integrate with sentence context because they describe an event that is more plausible with regard to real-world knowledge.

Our results advance a resolution to the long-running access-integration debate in the psycholinguistic literature (for reviews, see Kutas & Federmeier, 2000; Lau et al., 2008). On one side of this debate, the 'access view' holds that the N400-component reflects the processes by which word meaning is accessed, which depends on whether its meaning is pre-activated, but not on whether this meaning renders the described event plausible with regard to real-world knowledge (Brouwer et al., 2012; Kutas & Federmeier, 2000). On the other side, the 'integration view' holds that the N400-component reflects processes that integrate word meaning with previous sentential context and world knowledge (Hagoort et al., 2004), which is easier for predictable words because they are more plausible.

The results from our large-scale study show there is merit in both views. We simultaneously modelled the effects of predictability and plausibility (along with a low-level semantic control variable), and we examined the time course of their effects. Predictability strongly predicted widespread N400 activity starting as early as 200 ms after word onset, and showed an effect reversal (high predictability eliciting more negative voltage) that started before the end of the N400 window and lasted several hundreds of milliseconds. In contrast, plausibility was associated with a smaller, right-lateralized effect that started only after the effect of predictability reached its peak (around 350 ms after word onset) and that continued

until well beyond the classical N400 window. Crucially, effects of predictability and plausibility both occurred in the N400 time window, but the former dominated the N400's rise (i.e., upward flank), while the latter set in at its fall (i.e., downward flank). In addition, our results offered some initial evidence that the N400 effect of plausibility partly, but not entirely, reflects the stronger effect of plausibility in relatively unexpected words.

Our results challenge accounts in which the predictability-dependent N400 (Kutas & Hillyard, 1984) reflects the effects of *only* prediction (DeLong et al., 2005) or *only* integration (Hagoort et al., 2004). Instead, they support a 'hybrid', multiple-process view (Baggio, 2012; Baggio & Hagoort, 2011; see also Lau et al., 2016; Newman et al., 2012), wherein N400 activity reflects a cascade of non-compositional and compositional processes that access and integrate word meaning within a sentence context. In a recent neurobiological account, Baggio and Hagoort (2011) propose that the onset and rising flank of the N400 reflect build-up of current in the temporal cortex when people access word meaning from long-term memory, followed by forward currents to the inferior prefrontal gyrus, where a context representation is generated and maintained. The peak and downward flank of the N400 reflect the moment when re-injection of currents back to the temporal cortex dominates activity as people integrate word meaning with a context representation held active in prefrontal cortex. Our results are broadly compatible with this proposal in terms of the observed time course of prediction and integration effects, although our results are inconclusive with regard to the assumed neural generators.

Our results obtained in the post-N400 time window (500-1000 ms) inform another current debate, namely on the processing consequences of words that disconfirm a strong prediction. Van Petten and Luka (2012) argue for a processing distinction between plausible unexpected words (prediction mismatch) and implausible unexpected words (plausibility violations). The former elicit a left-frontal positive ERP effect, whereas the latter elicit a parietal positive ERP

effect (see also Delong, Quante & Kutas, 2014). While our study also showed a left-frontal positive ERP effect of prediction mismatch, the effect was more short-lived than is typically reported (Van Petten & Luka, 2012), and we also observed a late positive ERP effect of semantic similarity, not of plausibility. Therefore, the post-N400 positive ERP effect of prediction mismatch, like the N400, does not reflect effects of prediction only.

Our results thus demonstrate a more general point, namely that perhaps any ERP component, and especially those extending over hundreds of milliseconds like the N400 or the post-N400 positivity, is likely to reflect the combined activity of multiple subcomponents that are associated with related yet distinct cognitive processes (e.g., Dien, Michelson & Franklin, 2010; Newman et al., 2012; Otten & Van Berkum, 2009; Pylkkänen & Marantz, 2003). This highlights the need for a detailed and computationally specific account of the transition between access and integration (see also, Baggio & Hagoort, 2011; Hauk, 2016).

Further research should establish the replicability and generalizability of our results. All our sentences elicited a strong expectation for a given noun (DeLong et al., 2005) and contained indefinite articles that were consistent or inconsistent with that noun (although people may not use inconsistent articles to revise their prediction, see Nieuwland et al., 2017). In sentences that generate weak or no predictions, integration processes may contribute more strongly to N400 activity (for discussion, see Lau et al., 2016). Furthermore, differences between the onset of prediction and integration effects may be exacerbated during spoken language comprehension, as listeners only need an initial phoneme to disconfirm a prediction (e.g., Van Petten et al., 1999), well before word meaning is available for contextual integration.

In sum, the results of our large-scale study challenge the view that semantic facilitation of predictable words reflects the effects of *either* prediction *or* integration, and suggest that

facilitation arises from a cascade of processes that access and integrate word meaning with context into a sentence-level meaning.

References

Baggio, G. (2012). Selective alignment of brain responses by task demands during semantic processing. *Neuropsychologia*, *50*(5), 655-665.

Baggio, G., & Hagoort, P. (2011). The balance between memory and unification in semantics: A dynamic account of the N400. *Language and Cognitive Processes, 26*(9), 1338-1367. doi: 10.1080/01690965.2010.542671

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2008). An alternative perspective on "semantic P600" effects in language comprehension. *Brain Research Reviews, 59*(1), 55-73. doi: 10.1016/j.brainresrev.2008.05.003

Brothers, T., Swaab, T. Y., & Traxler, M. J. (2015). Effects of prediction and contextual support on lexical processing: Prediction takes precedence. *Cognition, 136*, 135-149. doi: 10.1016/j.cognition.2014.10.017

Brouwer, H., Fitz, H., & Hoeks, J. (2012). Getting real about semantic illusions: rethinking the functional role of the P600 in language comprehension. *Brain Research, 1446*, 127-143. doi: 10.1016/j.brainres.2012.01.055

Brown, C., & Hagoort, P. (1993). The processing nature of the N400: Evidence from masked priming. *Journal of cognitive neuroscience*, *5*(1), 34-44.

Calloway, R. C., & Perfetti, C. A. (2017). Integrative and predictive processes in text reading: the N400 across a sentence boundary. *Language, Cognition and Neuroscience*, *32*(8), 1001-1016.

Clifton, C., Staub, A., & Rayner, K. (2007). Eye movements in reading words and sentences. In *Eye Movements* (pp. 341-371).

DeLong, K. A., Quante, L., & Kutas, M. (2014). Predictability, plausibility, and two late ERP positivities during written sentence comprehension. *Neuropsychologia, 61*, 150-162. doi: 10.1016/j.neuropsychologia.2014.06.016

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience, 8*(8), 1117-1121. doi: 10.1038/nn1504

Dien, J., Michelson, C. A., & Franklin, M. S. (2010). Separating the visual sentence N400 effect from the P400 sequential expectancy effect: cognitive and neuroanatomical implications. *Brain Research, 1355*, 126-140. doi: 10.1016/j.brainres.2010.07.099

Federmeier, K. D. (2007). Thinking ahead: the role and roots of prediction in language comprehension. *Psychophysiology, 44*(4), 491-505. doi: 10.1111/j.1469-8986.2007.00531.x

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language, 41*(4), 469-495. doi: DOI 10.1006/jmla.1999.2660

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*, 1-12.

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: 10.1111/j.1469-8986.2011.01273.x

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science, 304*(5669), 438-441. doi: 10.1126/science.1095455

Hauk, O. (2016). Only time will tell–why temporal information is essential for our neuroscientific understanding of semantics. *Psychonomic Bulletin & Review, 23*(4), 1072-1079.

Hauk, O., Davis, M. H., Ford, M., Pulvermuller, 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

Ito, A., Corley, M., Pickering, M. J., Martin, A. E., & Nieuwland, M. S. (2016). Predicting form and meaning: Evidence from brain potentials. *Journal of Memory and Language, 86*, 157-171. doi: 10.1016/j.jml.2015.10.007

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences, 4*(12), 463-470.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Reviews of Psychology, 62*, 621-647. doi: 10.1146/annurev.psych.093008.131123

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science, 207*(4427), 203-205.

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

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211-240. doi: Doi 10.1037//0033-295x.104.2.211

Lau, E., Almeida, D., Hines, P. C., & Poeppel, D. (2009). A lexical basis for N400 context effects: evidence from MEG. *Brain and Language, 111*(3), 161-172. doi: 10.1016/j.bandl.2009.08.007

Lau, E., Namyst, A., Fogel, A., & Delgado, T. (2016). A direct comparison of N400 effects of predictability and incongruity in adjective-noun combination. *Collabra: Psychology, 2*(1), 13.

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (de)constructing the N400. *Nature Reviews Neuroscience, 9*(12), 920-933. doi: 10.1038/nrn2532

Li, X., Hagoort, P., & Yang, Y. (2008). Event-related potential evidence on the influence of accentuation in spoken discourse comprehension in Chinese. *Journal of Cognitive Neuroscience, 20*(5), 906-915. doi: 10.1162/jocn.2008.20512

Newman, R. L., Forbes, K., & Connolly, J. F. (2012). Event-Related Potentials and Magnetic Fields Associated with Spoken Word Recognition. In: Spivey M, Joanisse M, McRae K, editors. *Cambridge Handbook of Psycholinguistics.* Cambridge University Press;

New York, NY.
Nieuwland, M., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., ... & Mézière, D. (2017). Limits on prediction in language comprehension: A multi-lab failure to replicate evidence for probabilistic pre-activation of phonology. *BioRxiv*, 111807.
Otten, M., & Van Berkum, J. J. (2007). What makes a discourse constraining? Comparing the effects of discourse message and scenario fit on the discourse-dependent N400 effect. *Brain Research, 1153*, 166-177. doi: 10.1016/j.brainres.2007.03.058
Polich, J. (2007). Updating p300: An integrative theory of P3a and P3b. *Clinical Neurophysiology, 118*(10), 2128-2148. doi: 10.1016/j.clinph.2007.04.019
Pylkkanen, L., & Marantz, A. (2003). Tracking the time course of word recognition with MEG. *Trends in Cognitive Sciences, 7*(5), 187-189.
Roehm, D., Bornkessel-Schlesewsky, I., Rosler, F., & Schlesewsky, M. (2007). To predict or not to predict: Influences of task and strategy on the processing of semantic relations. *Journal of Cognitive Neuroscience, 19*(8), 1259-1274.
Rueschemeyer, S. A., Gardner, T., & Stoner, C. (2015). The Social N400 effect: how the presence of other listeners affects language comprehension. *Psychonomic Bulletin and Review, 22*(1), 128-134. doi: 10.3758/s13423-014-0654-x
Sassenhagen, J., & Alday, P. M. (2016). A common misapplication of statistical inference: Nuisance control with null-hypothesis significance tests. *Brain and Language, 162*, 42-45. doi: 10.1016/j.bandl.2016.08.001
Stanovich, K. E., & West, R. F. (1981). The Effect of Sentence Context on Ongoing Word Recognition: Tests of a Two-Process Theory. *Journal of Experimental Psychology: Human Perception and Performance, 7*(3), 658-72.
Steinhauer, K., Royle, P., Drury, J. E., & Fromont, L. A. (2017). The priming of priming: Evidence that the N400 reflects context-dependent post-retrieval word integration in working memory. *Neuroscience Letters, 651*, 192-197. doi: 10.1016/j.neulet.2017.05.007
Van Berkum, J. J. (2009). The neuropragmatics of 'simple' utterance comprehension: An ERP review. In *Semantics and pragmatics: From experiment to theory* (pp. 276-316). Palgrave Macmillan.
Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory and Cognition, 31*(3), 443-467. doi: 10.1037/0278-7393.31.3.443
Berkum, J. J. V., Hagoort, P., & Brown, C. M. (1999). Semantic integration in sentences and discourse: Evidence from the N400. *Journal of cognitive neuroscience, 11*(6), 657-671.
Van Petten, C. (2014). Examining the N400 semantic context effect item-by-item: relationship to corpus-based measures of word co-occurrence. *International Journal of Psychophysiology, 94*(3), 407-419. doi: 10.1016/j.ijpsycho.2014.10.012
Van Petten, C., Coulson, S., Rubin, S., Plante, E., & Parks, M. (1999). Time course of word identification and semantic integration in spoken language. *Journal of Experimental Psychology: Learning Memory and Cognition, 25*(2), 394-417. doi: Doi

10.1037//0278-7393.25.2.394
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
Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, *1*(1), 3-14.

Author Contributions

M. S. Nieuwland developed the study concept. Testing and data collection were performed by E. Heyselaar, E. Darley., S. Von Grebmer Zu Wolfsthurn, F. Bartolozzi , V. Kogan, A. Ito, S. Busch-Moreno, X. Fu., E. Kulakova, S. Politzer-Ahles, and Z. Kohút, under the supervision of M.S. Nieuwland, K. Segaert, N. Kazanina, G. Rousselet, H.J. Ferguson, J. Tuomainen, E. M. Husband, D.I. Donaldson, and S. Rueschemeyer. M. S. Nieuwland performed the data analysis and interpretation in consultation with D. J. Barr, N. Kazanina and S. Politzer-Ahles. M. S. Nieuwland drafted the manuscript, and D. J. Barr, F. Huettig, K. Segaert, A. Ito, N. Kazanina, S. Politzer-Ahles, H.J. Ferguson, J. Tuomainen, E. M. Husband, D. I. Donaldson, and S. Rueschemeyer provided commentary. All authors approved the final version of the manuscript for submission.

Acknowledgements

This work was partly funded by ERC Starting grant 636458 to H.J.F.