Dissociating predictability, plausibility and possibility of sentence continuations in reading: Evidence from late-positivity ERPs

Laura Quante Corresp., 1, 2, Jens Bölte 1, 2, PieNie Zwitserlood 1, 2

1 Department of Psychology, University of Münster, Münster, Germany
2 Otto-Creutzfeldt-Center for Cognitive and Behavioral Neuroscience, University of Münster, Münster, Germany

Corresponding Author: Laura Quante
Email address: l.quante@uni-muenster.de

Late positive ERP components occurring after the N400, traditionally linked to reanalysis due to syntactic incongruence, are increasingly considered to also reflect reanalysis and repair due to semantic difficulty. Semantic problems can have different origins, such as a mismatch of specific predictions based on the context, low plausibility, or even semantic impossibility of a word in the given context. DeLong, Quante, and Kutas (Neuropsychologia, 2014) provided the first direct evidence for topographically different late positivities for prediction mismatch (left frontal late positivity for plausible but unexpected words) and plausibility violation (posterior-parietal late positivity for implausible, incongruent words). The aim of the current study is twofold: (1) to replicate this dissociation of ERP effects for plausibility violations and prediction mismatch in a different language, and (2) to test an additional contrast within implausible words, comparing impossible and possible sentence continuations. Our results replicate DeLong et al. (2014) with different materials in a different language, showing graded effects for predictability and plausibility at the level of the N400, a dissociation of plausible and implausible, anomalous continuations in posterior late positivities and an effect of prediction mismatch on late positivities at left-frontal sites. In addition, we found some evidence for a dissociation, at these left-frontal sites, between implausible words that were fully incompatible with the preceding discourse and those for which an interpretation is possible. We discuss the theoretical impact of our results in the light of prediction, updating and integration of words into the discourse.
Dissociating predictability, plausibility and possibility of sentence continuations in reading: Evidence from late-positivity ERPs

Laura Quante\textsuperscript{a,b}, Jens Bölte\textsuperscript{a,b}, Pienie Zwitserlood\textsuperscript{a,b}

\textsuperscript{a}Department of Psychology, University of Münster, Fliednerstraße 21, 48149 Münster, Germany

\textsuperscript{b}Otto-Creutzfeldt-Center for Cognitive and Behavioral Neuroscience, University of Münster, Fliednerstraße 21, 48149 Münster, Germany

Corresponding author:

Laura Quante

Email address: l.quante@uni-muenster.de
Abstract

Late positive ERP components occurring after the N400, traditionally linked to reanalysis due to syntactic incongruence, are increasingly considered to also reflect reanalysis and repair due to semantic difficulty. Semantic problems can have different origins, such as a mismatch of specific predictions based on the context, low plausibility, or even semantic impossibility of a word in the given context. DeLong, Quante, and Kutas (Neuropsychologia, 2014) provided the first direct evidence for topographically different late positivities for prediction mismatch (left frontal late positivity for plausible but unexpected words) and plausibility violation (posterior-parietal late positivity for implausible, incongruent words). The aim of the current study is twofold: (1) to replicate this dissociation of ERP effects for plausibility violations and prediction mismatch in a different language, and (2) to test an additional contrast within implausible words, comparing impossible and possible sentence continuations. Our results replicate DeLong et al. (2014) with different materials in a different language, showing graded effects for predictability and plausibility at the level of the N400, a dissociation of plausible and implausible, anomalous continuations in posterior late positivities and an effect of prediction mismatch on late positivities at left-frontal sites. In addition, we found some evidence for a dissociation, at these left-frontal sites, between implausible words that were fully incompatible with the preceding discourse and those for which an interpretation is possible. We discuss the theoretical impact of our results in the light of prediction, updating and integration of words into the discourse.
1 Introduction

The study of effects of context on language processing has a long tradition in psycholinguistics, as modular (cf. Forster, 1981) and interactive (cf. McClelland & Rumelhart, 1981; Marslen-Wilson, 1987) theories of word recognition drastically differed with respect to the role allotted to information stemming from sources other than the word itself. Proof for an impact of contextual, top-down information on word recognition was already provided more than 30 years ago, with priming paradigms and reaction time data (cf. Swinney, 1979; Schwaneflugel & Shoben, 1985). The advent of event-related potentials again fired the debate, because they allow insights into the time-course of word recognition, which is difficult to come by with reaction times (but see Zwitserlood, 1989). Ever since, a wealth of studies has shown that contextual information, when constraining enough, has an early impact on lexical processing – even to the extent that upcoming words are anticipated (Kutas & Federmeier, 2000; van Berkum, Zwitserlood, Hagoort, & Brown, 2003; DeLong, Urbach, & Kutas, 2005; van Berkum, Brown, Zwitserlood, & Hagoort, 2005). It is thus not surprising that terminology has changed, and “anticipation” and “prediction” are now used to refer to the impact, on lexical processing, of knowledge from sources other than the current input (cf. Van Petten & Luka, 2012; Huettig & Janse, 2016; Kuperberg & Jaeger, 2016). Whereas most researchers agree that (features of) upcoming words are predicted under certain circumstances, it remains unresolved which factors promote (or prevent) predictive processing, and what information about words (e.g., semantics, word forms) is predicted (see Ito et al., 2016; Kuperberg & Jaeger, 2016).

To study effects of semantic context, expectation and prediction in language comprehension, a particular event-related potential (ERP) component, the N400 (Kutas & Hillyard, 1980), has been used extensively. The N400 is a negative-going wave peaking around
400 ms after stimulus onset, which is related to semantic processing (for a review, see Kutas & Federmeier, 2011). For example, its amplitude is negatively correlated to a word’s cloze probability (proportion of respondents who completed a given context with this particular word), a measure of semantic expectancy. Words with strong contextual support show a decrease in N400 amplitude relative to words that are less predictable or do not fit the context (Kutas & Federmeier, 2011). There is also evidence for ERP effects as a function of predictability in time windows preceding the N400 (e.g., van Berkum, Zwitserlood, Hagoort, & Brown, 2003; Dikker & Pylkkänen, 2011; Lau, Holcomb, & Kuperberg, 2013; Brothers, Swaab, & Traxler, 2015; see Kuperberg & Jaeger, 2016, for an overview). However, evidence for the actual pre-activation or anticipation of upcoming words, assessed before any of their input becomes available, is less abundant (but see DeLong, Urbach, and Kutas, 2005; van Berkum, Brown, Zwitserlood, Kooijman & Hagoort, 2005; Szewczyk & Schriefers, 2013; Ito, Corley, Pickering, Martin, & Nieuwland, 2016).

Our study uses ERPs and does not focus on prediction or expectation per se, but on the consequences of prediction or expectation mismatch, and, more generally speaking, of contextual mismatch.\(^1\) Van Petten and Luka (2012) proposed that if listeners and readers predict upcoming words, the EEG signal should reflect not only benefits of a confirmed prediction (visible as attenuation of the N400) but also costs of a disconfirmed prediction. In their review article, they assessed studies that compared congruent sentence completions with semantically anomalous completions, and often observed a late positivity, about 600-900 ms after critical-word onset, with a mainly parietal scalp topography. In addition, an anterior positivity was sometimes observed when ERPs for unexpected but semantically congruent sentence completions were compared to
predictable, expected completions. It should be noted, however, that the 60+ studies included showed a great variability in the post-N400 time window.

It thus seems that unexpected continuations that allow construction of a possible overall sentence meaning differ from anomalous completions. Interestingly, studies that manipulate semantic expectancy have predominantly used anomalous or unexpected plausible completions, but rarely both. This motivated DeLong, Quante, and Kutas (2014) to contrast different levels of plausibility within the same study, to determine how predictability and plausibility each contribute to word recognition. As completions of highly constraining sentence pairs (*For the snowman’s eyes, the kids used two pieces of coal. For his nose, they used ....*), DeLong et al. compared ERPs to highly predicable, expected (*a carrot*), unexpected but somewhat plausible (*a banana*), and unexpected, implausible, anomalous (*a groan*) words. The unexpected but plausible continuations should induce costs of disconfirmed prediction, combined with effort to integrate the unexpected noun - signaled by frontal late positivity. This does not hold for anomalous continuations that cannot be integrated with the current context. DeLong and colleagues observed a posterior late positivity to anomalous completions, and an anterior late positivity to unexpected but plausible completions, thus confirming Van Petten and Luka’s (2012) conjecture. Corroboration for a particular function of the frontal late positivity, also labeled frontal PNP (post-N400 positivity), in prediction-related revision was recently provided by Swaab and colleagues (Boudewyn, Long, & Swaab, 2015; Brothers, Swaab, & Traxler, 2015).

Predictability thus seems to influence early stages of processing, whereas plausibility seems to affect late stages of processing, which is corroborated by eye-tracking studies (Staub, 2015, for an overview). Interestingly, Rayner, Warren, Juhasz, & Liversedge et al. (2004) and Warren and McConnell (2007) further distinguished between plausibility and possibility, by
comparing words that result in implausible but possible meaning for the full sentence, to words
that induce an impossible overall sentence meaning, because they violate selection restrictions
(e.g., “inflate a carrot”) for example. In both studies, effects of words leading to either impossible
or implausible sentence meaning were dissociable in eye-movement measures.

Processing differences between implausible and impossible sentence overall meaning are
also visible in EEG data. For example, Paczynski and Kuperberg (2012) showed that selection-
restriction violations evoked a posterior positivity between 700 to 900 ms after critical word onset,
whereas violations of world knowledge, which result in implausible but still possible sentence
meaning, did not differ from plausible sentences in this time window. Similar results were shown
by Kuperberg, Sitnikova, Caplan, and Holcomb (2003), Geyer, Holcomb, Kuperberg, and
Perlmutter (2006), and Paczynski, Kreher, Ditman, Holcomb, and Kuperberg (2006). When
Kuperberg (2007) evaluated factors evoking a late positivity, she concluded that none of the
following factors - the presence of selection-restriction violations, semantic associations between
the critical word and the preceding context, specific task instructions, or constraining context - by
themselves could explain all results. One hypothesis that she advanced was that the impossibility
to establish an overall meaning for the sentence might be the crucial factor inducing a late positivity
on the critical word.

Given the variable nature of late positivities, and given the dire need for replication studies
of phenomena that are rather new or for which evidence is scarce (see Nieuwland et al., 2017; see
also Dennis & Valacich, 2014), the present study aimed to replicate DeLong et al.’s (2014) second
experiment, with German stimuli presented to German native speakers. In addition, inspired by
suggestions made by Kuperberg (2007) and DeLong et al. (2014), we analyzed differences
between implausible word completions that resulted in either possible or impossible overall
sentence meaning. To create a condition of impossible sentence meaning, we divided the materials into impossible and possible sets by means of subjective possibility ratings, collected in a pretest.

Following DeLong et al. (2014), we predicted a graded effect of contextual fit of critical words at the level of the N400, with implausible continuations showing enhanced negativity relative to unexpected but plausible words. Next, we expect a predictability effect, showing as an anterior late positivity – relative to expected nouns - to unexpected but plausible sentence completions, but not to implausible nouns. Next, we predict a plausibility effect, with a posterior positivity only for implausible, anomalous sentence completions. If Kuperberg’s (2007) assumption is correct, we predict this posterior late positivity only for those sentence completions that are truly anomalous and lead to an impossible overall sentence meaning, but not for those that allow an integration of the critical word with the preceding discourse, resulting in a perhaps implausible but nevertheless possible real-world meaning. This would constitute an effect of possibility, which also might show in a difference between possible and impossible implausible continuations in late positivity at anterior sites, with the possible continuations coinciding with plausible but unexpected ones.

2 Material and Methods

2.1 Stimuli

Stimuli were 150 constraining German sentence pairs (mean contextual constraint = 0.77, SD = 0.14, see cloze probability norming described below), which led to expectations for particular sentence-medial words. Following the condition labels used in DeLong et al. (2014), each of the 150 contexts was completed by a) the semantically expected noun (with the highest cloze
probability for the specific context; EXP), b) an unexpected but somewhat plausible noun (USP), and c) an unexpected, implausible noun (ANOM), resulting in a total of 450 sentence pairs (see Table 1 for sample sentence pairs; the complete set of sentence pairs is provided in Supplemental Table S1). To investigate whether the possible construction of overall sentence meaning was crucial for late positivities, the materials in the ANOM condition were subdivided on the basis of a pretest. Some of the sentences pairs in the ANOM condition contained critical nouns that allowed for a possible real-life meaning (ANOM-Pos; 45 sentence pairs), the other sentence pairs did not (ANOM-Impos; 105 sentence pairs). Fifty additional moderately constraining sentence pairs completed by their expected critical noun were used as fillers to balance the proportion of sentence pairs completed by expected versus unexpected nouns. Sentence material were either German translations of stimuli used in DeLong et al. (2014) or constructed in the same fashion by the experimenters. Where possible, critical nouns of the expected condition were re-used with different sentences in the other two conditions (53.3% of critical words were used three times, 31.1% were used twice, and 15.6% were used only once). Since German nouns are coded for gender (masculine, feminine, neutral), all three completions of a particular sentence pair had the same grammatical gender. Written word frequency, word length and orthographic neighborhood size of the critical nouns were matched between the three main conditions (see Table 2). Note that the Pos and Impos items within the ANOM condition were not balanced with respect to these factors.

2.2 Cloze probability norming

Stimulus norming for critical noun cloze probability was conducted in a separate sentence completion task with 36 volunteers (native speakers of German, mainly students). They were compensated with course credit and did not participate in the EEG study. Contexts were truncated
prior to the critical noun, and participants were asked to complete the second sentence with a single noun that came to their mind first and fitted with the preceding context. Every context ended with the three German indefinite articles, in the order masculine, feminine, and neuter, thus allowing nouns of any grammatical gender. Cloze probability was calculated as the proportion of participants who completed a particular sentence pair with a particular noun. The cloze probability of the most frequent noun equals the contextual constraint of a given sentence pair. Sentence pairs with a cloze probability of 50% or higher were considered highly constraining and included in the study. Filler sentences had a cloze probability of 40% or higher. Table 2 shows mean cloze probabilities for the experimental and filler conditions.

2.3 Plausibility rating

All 450 sentence pairs, truncated after the critical noun, were rated for plausibility (“How plausible is the sentence pair’s meaning…”) on a scale of 1 (not plausible) to 5 (highly plausible) by eight independent German raters who did not participate in the EEG study. Table 2 presents mean plausibility ratings of all conditions. Following DeLong et al. (2014), mean plausibility was greater than 1.5 in the EXP and USP conditions, and less than or equal to 1.5 in the ANOM condition. The plausibility ratings differed significantly between all conditions (see Table 3).

2.4 Possibility rating

All 450 sentence pairs, truncated after the critical noun, were rated for possibility (“How possible (in real-life) is the sentence pair’s meaning…””) by the same eight raters, on a scale of 1 (impossible) to 4 (possible). Table 2 specifies mean possibility ratings for all conditions. Similar to the cut-off for Plausibility, mean possibility ratings were greater than 1.5 in the EXP, USP and ANOM-Pos conditions, but equal to or less than 1.5 in the ANOM-Impos condition. The ratings
of all conditions differed significantly from each other (see Table 3). Participants performed both
plausibility and possibility ratings at the same time. No examples were provided to avoid biasing
the raters’ judgements. Correlations between plausibility and possibility ratings are displayed in
Supplemental Analysis S1.

In the main experiment, each participant was presented with one of three 200-item lists,
with contexts and critical nouns used once per list (except for four critical nouns that occurred
twice per list, in different contexts). Lists 1, 2, and 3 were presented to 12, 10 and 10 participants,
respectively. Every list consisted of 50 predictable, expected nouns, 50 unexpected plausible
nouns, 50 unexpected implausible (ANOM) nouns, and 50 fillers. Approximately one third of the
ANOM nouns was rated possible (list 1: 17, list 2: 16, list 3: 12), the remaining two thirds were
rated impossible. Fifty comprehension questions followed 25% of sentence pairs at random
intervals. Three additional sentence pairs preceded every list to familiarize participants with the
task. Sentence pairs within a list were randomized across subjects.

2.5 ERP Participants

Thirty-two students (23 f, 9 m) participated in the experiment after giving written informed
consent. They were compensated with course credit or cash (7.50 €/hour). Mean age was 25.3
years (19-34). All participants were monolingual native speakers of German and right-handed
(assessed via Edinburgh Handedness Inventory, Oldfield, 1971). Eight participants reported a left-
handed parent or sibling, one reported two left-handed relatives. All participants reported normal
or corrected-to-normal vision. One additional participant was tested but excluded from analysis
because of a technical problem during the experiment. The study protocol was conducted in
accordance with ethical standards of the Declaration of Helsinki and approved by the local ethics
committee of the University of Münster (approval number #2016-42-LQ).
2.6 Procedure

The experiment consisted of a single two-hour-session conducted in a quiet and dimly lit room at the Westfälische Wilhelms-Universität Münster. Participants were seated approximately 1 m in front of a LED monitor (BenQ, model XL2420T, 144 Hz, 24”W) and read sentence pairs for comprehension. The experiment was set up using Presentation software (NeuroBehavioral Systems, Version 16.3). Stimuli were presented visually, in black type (RGB: 0, 0, 0; Arial 48 pt) on a grey background (RGB: 148, 148, 148). The experiment was divided into eight blocks of approximately six minutes length, with two-minute breaks between blocks. Every trial started with a fixation cross (500 ms) in the center of the screen, followed by the first sentence of a pair presented in its entirety. Participants advanced to the critical sentence via button press. This sentence including the critical word was presented with a rapid serial visual presentation technique (RSVP), each word presented centrally for 200 ms, with a stimulus onset asynchrony of 500 ms. Yes/no comprehension questions followed 25 % of sentence pairs at random intervals. Participants responded with two buttons on a response pad (Cedrus, model RB-830) with response buttons counterbalanced across participants and lists. Comprehension questions appeared after the critical noun sentence. In case of a question, participants’ button press advanced to the next trial, otherwise the next sentence pair appeared automatically after two seconds.

Material, design and procedure were almost identical to DeLong et al.’s second experiment except for the following differences. First, DeLong et al.’s sentence material was translated into German or constructed using the same sentence structure. Second, mean constraint of discourse contexts and mean cloze probability of expected critical words were lower than in DeLong et al. (.77 vs .89, respectively). Third, sentence pairs were not rated for possibility in DeLong’s study.
Fourth, 32 participants completed the present EEG experiment, 24 students participated in DeLong et al.’s second study.

2.7 Electroencephalographic recording parameters

EEG was recorded from 32 Ag/AgCl-electrodes attached to a WaveGuard 32-channel cap (ANT, Advanced Neuro Technology). Electrodes were placed according to the International 10-20 convention (Jasper, 1958), and an average reference was used (see Fig. 1 for scalp sites). Blinks and vertical eye movements were monitored from electrodes placed above and below the left eye, and horizontal eye movements were monitored from two electrodes placed on the outer canthi. Impedances were kept below 5kΩ. The EEG was continuously recorded with ASA (Advanced Source Analysis, version 4.7.3.1, ANT). Data collection and evaluation were controlled by ExMan (Experiment Manager; MS Excel worksheet with active macros). EEG was amplified (ExG 20x, fixed = 50 mV/V), low pass filtered (finite impulse response filter, cut-off frequency = .27 x sampling rate) and continuously digitized at a sampling rate of 256 samples/second.

2.8 Data analysis

Before averaging, the EEG signal was filtered using a Butterworth half-amplitude bandpass FIR-filter (0.1 Hz, 20 Hz, 12 db/oct). Vertical eye movements were corrected with principal component analysis (Ille, Berg, & Scherg, 2002). Additionally, seven electrodes (0.7 %) were interpolated (see Supplemental Table S2). The EEG was re-referenced offline to the algebraic mean of left and right mastoids and averaged for each experimental condition, time-locked to the critical noun onset. Before averaging, trials contaminated by artefacts (specified as voltage changes exceeding ± 75 µV during the epoch) were rejected offline (on average 4.23 % of all trials,
ERPs were calculated for epochs extending from 500 ms pre- to 1500 ms post-stimulus onset, thus using a pre-stimulus baseline of 500 ms.

In a first step, mass univariate analyses were conducted to compare spatial and temporal properties of possible ERP effects found in the present experiment to the findings by DeLong et al. (2014). ERPs from the three pairwise comparisons, [USP minus EXP], [ANOM minus EXP], and [ANOM minus USP], were submitted to repeated measures, two-tailed t-tests at all sampled time points between 250 ms and 1050 ms (206 total time points) at all 30 scalp electrodes, resulting in 6180 total comparisons for each condition contrast. To control the number of false discoveries, the Benjamini and Yekutieli (2001) procedure was applied using a false discovery rate level of 5%.

In a second step, mean amplitudes were analyzed by first conducting ANOVAs with three levels of noun type (EXP, USP and ANOM) to compare the present results directly to those reported in DeLong et al. (2014). These analyses were complemented by pairwise t-tests between the four levels of noun type (EXP, USP, ANOM-Pos, and ANOM-Impos). ANOVAs were applied to the data from three time windows: a) over all 30 electrode sites between 300 ms and 500 ms (N400), b) over seven (left) anterior electrode sites [Fp1, Fpz, F7, F3, Fz, FC5, T7] between 600 ms and 1000 ms (frontal positivity), c) over seven posterior electrode sites [Cz, CP1, CP2, P3, Pz, P4, POz] between 600 ms and 1000 ms (posterior positivity; see Fig. 1 for electrode placement).

Scalp regions and temporal windows were based on DeLong et al.’s (2014) second experiment. To confirm the left lateralization of the anterior positivity, we extended the corresponding ANOVA by the factor hemisphere (left, right), and included equivalent right hemisphere electrodes (Fp2, F4, F8, Fc6, and T8) while excluding midline electrodes Fz and Fpz. If sphericity was violated, ANOVA p-values and degrees of freedom were corrected using epsilon correction.
(Greenhouse Geisser) for repeated measures with more than one degree of freedom. Significance levels of pairwise t-tests were Bonferroni-adjusted (t-tests following ANOVAs with three levels of noun type: \( p_{\text{boncor}} < .0167 \), t-tests comparing all four levels of noun type: \( p_{\text{boncor}} < .0083 \)).

3 Results

3.1 Behavioral results

Participants correctly answered an average of 96.7% (median 97%, range = 90-100%) of yes/no comprehension questions, suggesting they comprehended the sentence pairs during the experiment.

3.2 ERP results

3.2.1 Mass univariate analyses.

The first mass univariate analysis focused on predictability, comparing USP versus EXP nouns (see Fig. 2, top). There was a widespread N400 effect, with ERPs to USP nouns being more negative than ERPs to EXP nouns. This negativity lasted from approximately 250 to 500 ms. Starting shortly before the offset of the N400, between approximately 550 and 1000 ms, USP nouns were more positive than EXP nouns, particular over left frontal and left lateral temporo-parietal scalp sites.

The second mass univariate analysis looked at plausibility, comparing ANOM versus EXP nouns (see Fig. 2, middle). Again, a widespread N400 effect emerged, with ERPs to ANOM nouns being more negative than ERPs to EXP nouns, between approximately 250 ms and 500 ms. By
about 600 ms, a positivity of ANOM relative to EXP nouns emerged and continued up to the end of the time window (1050 ms), being most prominent over central and posterior scalp sites.

The third mass univariate analysis compared ANOM versus USP nouns (see Fig. 2, bottom). Between 300 ms and 450 ms, ERPs to ANOM nouns were more negative than ERPs to USP nouns. In addition, from approximately 600 ms to the end of the time window (1050 ms), ERPs to ANOM nouns were more positive than ERPs to USP nouns at posterior scalp locations. In contrast, ERPs to USP nouns were more positive than ERPs to ANOM nouns over lateral frontal scalp sites between approximately 700 to 900 ms. In all three mass-univariate analyses, significant p-values are \( p_{adj} < .05 \).

### 3.2.2 Analyses of variance.

For visual inspection, Fig. 3 shows the grand average ERPs of all 32 participants over 30 scalp channels. Topographic scalp maps of ERP mean amplitude voltage differences can be seen in Fig. 4, and four representative anterior and posterior channels are shown in Fig. 1. In line with the results from mass univariate analyses described above, all figures reveal N400 effects for both USP and ANOM nouns relative to EXP nouns, a post-N400 positivity for ANOM nouns over posterior channels, and a post-N400 positivity for USP nouns over anterior channels. Early components (P1, N1, and P2) do not differ as a function of noun type. Tables 4 and 5 provide mean amplitudes of the four noun types and detailed results of pairwise t-tests between conditions.

#### 3.2.2.1 300-500 ms.

An ANOVA with three levels of noun type over all 30 electrode sites revealed a main effect [\( F(1.60, 49.46) = 74.98, p < .001, \varepsilon_{GG} = 0.80, \eta_p^2 = 0.39 \)]. ANOM nouns showed the largest negativity (-2.89 \( \mu V \)), followed by USP nouns (-1.54 \( \mu V \)) and EXP nouns (0.83 \( \mu V \)). Post-hoc t-
The pairwise t-tests between all four noun types revealed significant differences between all conditions ($t(31) \geq 4.38$, $p < .001$) except for the comparison of ANOM-Pos (-3.01 $\mu$V) and ANOM-Impos nouns (-2.82 $\mu$V; $t(31) = -0.72$, $p = .479$).

### 3.2.2.2 600-1000 ms posterior scalp sites.

An ANOVA with three levels of noun type, conducted over seven posterior electrode sites, showed a main effect [$F(2, 62) = 21.06$, $p < .001$, $\eta^2_p = 0.19$]. ANOM nouns had the largest positivity (3.09 $\mu$V), followed by USP nouns (1.38 $\mu$V) and EXP nouns (0.98 $\mu$V). Post-hoc t-tests are displayed in Fig. 5. Pairwise t-tests revealed significant differences between EXP and ANOM-Pos nouns ($t(31) = -4.29$, $p < .001$), EXP and ANOM-Impos nouns ($t(31) = -5.40$, $p < .001$), USP and ANOM-Pos nouns ($t(31) = -4.75$, $p < .001$), and USP and ANOM-Impos nouns ($t(31) = -4.87$, $p < .001$). No reliable difference was found between EXP and USP nouns ($t(31) = -1.31$, $p = .200$) and between the two types of ANOM nouns (ANOM-Pos = 2.95 $\mu$V, ANOM-Impos = 3.16 $\mu$V; $t(31) = -0.72$, $p = .477$).

### 3.2.2.3 600-1000 ms anterior scalp sites.

The extended ANOVA indicated a left lateralization of the effect (see Supplemental Analysis S2). Therefore, we restricted our analysis to those electrodes analyzed in DeLong et al. (2014). Over the seven left anterior electrodes, the ANOVA with three levels of noun type revealed a main effect [$F(2, 62) = 6.13$, $p = .004$, $\eta^2_p = 0.08$]. USP nouns showed the greatest positivity (1.91 $\mu$V), followed by ANOM nouns (1.51 $\mu$V) and EXP nouns (0.79 $\mu$V). Post-hoc t-tests are displayed in Fig. 5. Paired t-tests revealed that EXP nouns differed from USP nouns ($t(31) = -3.97$, $p < .001$) and from ANOM-Pos nouns ($t(31) = -2.88$, $p = .007$). The differences between USP and ANOM-Pos nouns (diff = 0.05 $\mu$V; $t(31) = -0.12$, $p = .904$) and between EXP and ANOM-Impos
nouns (diff = 0.54 μV; t(31) = -1.52, p = .138) were not significant. Although ANOM-Pos (1.96 μV) and ANOM-Impos nouns (1.33 μV) differed by 0.64 μV, this difference also did not reach significance (t(31) = 1.89, p = .068).

Whereas there were no differences in other temporal and/or spatial analysis windows, ANOM-Pos and ANOM-Impos thus seem to have a different impact on late anterior positivity. Given that different items were compared in the ANOM-Pos and ANOM-Impos conditions, we ran a regression analysis to assess possible effects of critical word characteristics that may cause amplitude differences between nouns. An amplitude calculation for each item averaged over participants is inadequate for exploring factors in multiple regression designs, because it disregards interparticipants’ variability. Therefore, we used the method suggested by Lorch and Myers (1990). For every participant, we extracted amplitudes of individual words and fitted a linear regression with factors plausibility, possibility, word length, word frequency, and orthographic neighborhood size (without interaction terms). For every predictor, the resulting 32 t-values entered a one-sample t-test. Only the effect of word length on amplitude was significantly different from zero (t(31) = 3.25, p = .003). See Supplemental Table S3 for detailed results. Note that the effect of Possibility failed significance, and even the Plausibility effect, which entails a within-item comparison, is much weaker in this analysis.

4 General discussion

In this study with German materials and participants, we investigated the electrophysiological signatures of different types of contextual fit – from predictable and thus expected, to highly implausible continuations of short discourses consisting of sentences pairs. Whereas no reliable differences were present before 300 ms, graded N400 effects for target nouns were observed as a function of their predictability and plausibility in the discourse. Relative to
highly predictable nouns, amplitudes were more negative for unexpected but plausible continuations, and again more negative for implausible continuations. Whether or not the implausible noun was a somewhat strange but in principle possible continuation of the preceding discourse had no impact on the N400. In the time window following the N400, positivities with different scalp signatures were observed that differed as a function of noun type. At posterior electrode sites, the 600 – 1000 ms time window revealed similar amplitudes for highly predictable and unpredictable but plausible continuations. Relative to these two continuations, there was enhanced positivity for both implausible noun types – which showed very similar amplitudes. At anterior sites, predictable and completely impossible continuations had similar amplitudes, but predictable nouns had a less positive amplitude than both unpredictable plausible and implausible, but still possible continuations, which did not differ.

In the following, we compare our outcomes to the original study (experiment 2) by DeLong and colleagues (2014) that we aimed to replicate, and evaluate our results against the predictions made for continuations that are quite implausible, but for which a real-world interpretation can be constructed given the discourse. We discuss the distinction between anterior and posterior late positivities and the potential processing functions that may underlie them.

4.1 Replication of DeLong, Quante, and Kutas (2014)

The data patterns relevant for the replication of DeLong et al. (2014), with three conditions (EXP, USP and ANOM) in three time windows (N400, Anterior Positivity, Posterior Positivity), show a striking similarity between the two studies, illustrated in Fig. 5.

The data for the N400 from the two studies show a very similar pattern – ignoring the position of the zero line. Relative to expected continuations, negativity is enhanced for unexpected
but plausible (USP) nouns, and again more so for implausible (ANOM) nouns. The results for the N400 thus fully replicate the graded negativity reported by DeLong et al. Note that the possibility to create an interpretation for some implausible continuations had no effect on N400 amplitude, since our two anomalous conditions did not differ.

The posterior positivity after the N400 also shows the same pattern as obtained by DeLong and colleagues (2014), with significant differences between the expected (EXP) and anomalous continuations (ANOM), between unexpected (USP) and anomalous words, but not between EXP and USP, the expected and unexpected continuations. Again, exactly the same pattern with the same significances was observed in both studies. Moreover, the analysis with four noun types showed no difference between the possible and impossible anomalous continuations. Finally, the anterior positivity again showed a similar pattern in both studies, but with somewhat different significances. Whereas in both studies, amplitudes for expected (EXP) and unexpected plausible (USP) nouns differ, and amplitudes for expected and anomalous nouns do not differ, the difference between the unexpected plausible and anomalous nouns that was reliable in DeLong et al. failed significance in our data. The analysis with four noun levels gives an indication why this might be the case. In this analysis, the anomalous nouns that have a possible interpretation given the preceding discourse do show a significant difference to the expected nouns, thus coinciding with the unpredictable but plausible nouns. The difference to the impossible anomalous nouns remains insignificant. Note however that the post-hoc regression analysis, which takes into account between-item differences in the analyses of possibility effects, questions whether these differences can be attributed to possibility.

Thus, with one interesting exception we closely replicate Experiment 2 by DeLong and colleagues (2014), with German materials – mainly but not exclusively translated from DeLong et
al. (2014), with somewhat lower predictability of the predictable, expected nouns, and with
German native speakers. We believe this replication of a dissociation between anterior and
posterior positivity in largely overlapping, post-N400 time windows to be an important
contribution to the growing evidence for a functional difference associated with these two late
positivities. As DeLong and colleagues, and unlike other studies, we show this relatively new
dissociation with the same population within one experiment. In the following, we discuss our
findings relative to data, hypotheses and models proposed by others.

4.2 N400 and late positivities

The N400 effects show that relative to an anomalous noun, an unexpected noun that is
nevertheless a perfectly plausible continuation shows a smaller negativity. This graded negativity,
relative to the predicted continuation, replicates findings from many studies that show amplitude
negativity to depend on the degree of deviation from the condition that serves as reference (see
Kutas & Federmeier, 2011, for an overview).

The late positivity observed in our data seems to come in two guises. There is a bilateral,
posterior positivity that separates expected and unexpected but plausible continuations from
implausible, anomalous continuations – with no difference between those for which a possible,
real-world meaning (ANOM-Pos) can be constructed and those for which this is not the case
(ANOM-Impos). A second late positivity, with anterior, left-lateralized scalp distribution, seems
to distinguish between nouns for which an interpretation in the given discourse is possible but
unexpected (USP and ANOM-Pos nouns) on the one hand, and predictable words on the other.

4.3 Posterior late positivity
The posterior late positivity observed in our data resembles the P600 that has been commonly associated with syntactic violations (e.g., Hagoort, Brown, & Groothusen, 1993) or syntactic complexity (e.g., Friederici, Hahne, & Saddy, 2002; Kaan & Swaab, 2003). This changed some fifteen years ago, when late positivities were reported for words that constituted thematic-role violations (e.g., *At breakfast the eggs would eat*) which are semantic in nature (Kolk, Chwilla, van Herten, & Oor, 2003; Kuperberg, Sitnikova, Caplan, & Holcomb, 2003; Hoeks, Stowe, & Doedens, 2004; see Brouwer, Fitz, & Hoeks, 2012, for an overview). Such “semantic illusions” had no impact on the N400 but showed in late positivities, with a central-posterior/parietal scalp topography that resembles the “syntactic” P600. This “semantic” P600 again fired the debate on its functional significance. Most proposed models and views adhere to two processing streams – semantic and syntactic – whose outputs can conflict with each other, which is reflected in the P600 (cf. Kim & Osterhout, 2005; Kolk & Chwilla, 2007; Kuperberg, 2007; Bornkessel-Schlesewsky & Schlesewsky, 2008; Hagoort, Baggio, & Willems, 2009; Kos, Vosse, van den Brinke, & Hagoort, 2010; Metzner, Malsburg, Vasishth, & Rösler, 2017; see Brouwer et al., 2017, for an overview and a different model).

In their seminal review of data from about 60 studies, Van Petten and Luka (2012) conclude that posterior late positivity is associated with attempts at reanalysis when a problem is detected – be it a syntactic or semantic incongruency or anomaly. Our late posterior positivities for all anomalous continuations fit this picture. Kuperberg (2013) prominently put late positivities into the perspective of prediction, suggesting that the posterior late positivities reflect processing costs when the incoming word disconfirms predicted events or event structure. This is the case even for semantic illusions (e.g., *The cat that from the mice fled*, incoming word underlined, Kolk et al., 2003) in which the incoming words semantically fit the event, but their thematic roles violate event
structure. In our data, all anomalous continuations show a posterior negativity. Clearly, impossible continuations violate event structure (often but not always because of selection restriction violations): “excuse” is not a viable candidate for a snowman’s nose. This is different for the unexpected but plausible continuations: lacking a carrot, a banana can serve as a snowman’s nose, and is thus compatible with the event of snowman construction. Consequently, our expected and unexpected but plausible words do not differ in late posterior positivity. Note that relative to these two conditions, and in contrast to our prediction, a clear late posterior positivity was evident for both types of anomalous continuations, those that are completely impossible and those for which an admittedly strange meaning could be constructed (e.g., *To save space, she bought herself a pig*(expected: *a loft bed* *in the store*). Following the logic by Van Petten and Luka, both anomalies initiate the reprocessing of prior input, and in Kuperberg’s view, both anomalies seem severe enough to violate event structure.

4.4 Anterior late positivity

Finally, we consider the anterior post-N400 positivity observed in our data. In the overall analysis, unexpected but plausible continuations (*the banana as nose for the snowman*) and implausible but still in some way possible continuations (*the woman who bought herself a pig to save space*) group together. First, they both differ from expected, highly predictable continuations (*the carrot for a snowman’s nose, a loft bed to save space*) and second, both continuations allow for a revision of the discourse on the basis of the meaning of the unexpected words. Note that the differences observed here may be due to item characteristics, as the regression analysis indicated. Although these data, given that they involve different items, should be treated with caution, it is interesting that similar late positivities with a (left) frontal scalp distribution have been observed when words are not predicted but semantically possible, given the preceding context (Federmeier,
As noted by Van Petten and Luka, and as is the case in our data, frontal late positivities follow an N400 – which is not always the case for posterior positivities. This co-occurrence is taken as an index for the sensitivity of frontal positivities to semantic predictability.

The exact functional significance of anterior late positivity is still under debate. Most researchers agree that it signals disconfirmed lexical prediction or lexical “prediction error” (Van Petten & Luka, 2012; Kuperberg, 2013, for overviews), and that the presence of a moderately or highly constraining context that can trigger updating is a prerequisite (Boudewyn et al., 2015). Note that both requirements apply to the two conditions in which we observed late frontal positivity. Taking these constraints as given, it remains unclear what processing costs occur after disconfirmed prediction. Do they involve inhibition of the predicted word – a hypothesis formulated with quite some foresight by Marta Kutas (1993), or are processing costs due to revising and updating working memory to integrate the unexpected continuation (Federmeier et al. 2007; Kuperberg, 2013)? In an ingenious study, Brothers, Swaab, and Traxler (2016) observed late frontal positivity for words that were not predicted by their participants – who were told to actively predict continuations of sentences and who indicated afterwards whether the continuation presented was the one they predicted or not. With full, trial by trial control of prediction, Brothers et al., could distinguish between specific lexical prediction and general contextual support – which our design does not allow. Given that they also observed early (pre-N400) effects of prediction, the authors conclude that the left-lateralized anterior positivity reflects prediction-related, post-lexical update and revision mechanisms. Given the importance of such mechanisms for prediction
in language, such anterior late positivities should be investigated further, with better control over item characteristics as is the case in our study.

4.5 Limitations

It is important to point out that our materials were not explicitly constructed for the distinction between ANOM-Impos and ANOM-Pos and that materials were not balanced (45 vs. 105 sentence pairs). As the regression analysis showed, items differed in length, which had an impact on the late anterior positivities. As DeLong et al.’s (2014) sentence pairs were not rated for possibility, it is not clear whether the minor discrepancies between the results of the two studies arose from potential differences of the nouns in the ANOM conditions. We also should note that overall contextual constraint was slightly lower in our study than in DeLong et al. which we aimed to replicate. Still, despite their post-hoc flavor, our results on (im)possibility provide an interesting perspective on the possibility of contextual integration of even quite implausible continuations - a good reason to consider this dimension in the future.

4.6 Conclusions

With German materials and participants, we replicated results of DeLong and colleagues (2014) and showed an impact of three types of constraint in sentence processing: predictability, plausibility and possibility. We observed graded effects on the N400, with the smallest negativity for expected continuations, followed by plausible but not expected alternatives, and with the largest negativity for implausible, anomalous continuations. Next, despite both being unexpected, plausible and implausible words show different patterns of posterior late positivity, arguing for a dissociation of predictability and plausibility. Finally, we believe that the distinction between
possible and impossible continuations, both being implausible, should be taken into account in studies on prediction and processing words in context.
Acknowledgements

We are deeply grateful to Katherine DeLong and Marta Kutas for their invaluable support, and thank Dan Ke, Christian Bürger, René Michel and Daniel Kluger for their assistance in data collection and analysis.
References

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 1165-1188.

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

Boudewyn, M. A., Long, D. L., & Swaab, T. Y. (2015). Graded expectations: Predictive processing and the adjustment of expectations during spoken language comprehension. *Cognitive, Affective, & Behavioral Neuroscience*, 15(3), 607-624. doi:10.3758/s13415-015-0340-0

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., Crocker, M. W., Venhuizen, N. J., & Hoeks, J. C. (2017). A Neurocomputational Model of the N400 and the P600 in Language Processing. *Cognitive Science*, 41(S6), 1318-1352. doi:10.1111/cogs.12461

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

Coltheart, M., Davelaar, E. J., Jonasson, J. T., & Besner, D. (1977). Access to the Internal Lexicon. In S. Dornic (Ed.), *Attention and Performance VI. Proceedings of the Sixth International*
Symposium on Attention and Performance, Stockholm, Sweden, July 28-August 1, 1975 (pp. 535-555). Hillsdale, NJ: Lawrence Erlbaum Associates.

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., Groppe, D. M., & Kutas, M. (2011). Overlapping dual ERP responses to low cloze probability sentence continuations. Psychophysiology, 48(9), 1203-1207. doi:10.1111/j.1469-8986.2011.01199.x

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

Dennis, A. R., & Valacich, J. S. (2014). A replication manifesto. AIS Transactions on Replication Research, 1(1), 1. doi:10.17705/1atrr.00001

Dikker, S., & Pylkkänen, L. (2011). Before the N400: Effects of lexical-semantic violations in visual cortex. Brain and Language, 118(1), 23-28. doi:10.1016/j.bandl.2011.02.006

Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. Brain Research, 1146, 75-84. doi:10.1016/j.brainres.2006.06.101

Forster, K. I. (1981). Priming and the effects of sentence and lexical contexts on naming time: Evidence for autonomous lexical processing. The Quarterly Journal of Experimental Psychology, 33(4), 465-496. doi:10.1080/14640748108400804
Friederici, A. D., Hahne, A., & Saddy, D. (2002). Distinct neurophysiological patterns reflecting aspects of syntactic complexity and syntactic repair. *Journal of Psycholinguistic Research, 31*(1), 45-63.

Geyer, A., Holcomb, P. J., Kuperberg, G. R., & Perlmutter, N. (2006). Plausibility and sentence comprehension. An ERP study. *Cogn. Neurosci. Suppl., Abstract*, 1-1.

Hagoort, P., Baggio, G., & Willems, R. M. (2009). Semantic unification. In *The Cognitive Neurosciences, 4th ed.* (pp. 819-836). MIT press.

Hagoort, P., Brown, C., & Groothusen, J. (1993). The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Language and Cognitive Processes, 8*(4), 439-483. doi:10.1080/01690969307585

Hoeks, J. C., Stowe, L. A., & Doedens, G. (2004). Seeing words in context: the interaction of lexical and sentence level information during reading. *Cognitive Brain Research, 19*(1), 59-73. doi:10.1016/j.cogbrainres.2003.10.022

Huettig, F., & Janse, E. (2016). Individual differences in working memory and processing speed predict anticipatory spoken language processing in the visual world. *Language, Cognition and Neuroscience, 31*(1), 80-93. doi:10.1080/23273798.2015.1047459

Ille, N., Berg, P., & Scherg, M. (2002). Artifact correction of the ongoing EEG using spatial filters based on artifact and brain signal topographies. *Journal of Clinical Neurophysiology, 19*(2), 113-124. doi:10.1097/00004691-200203000-00002
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

Jasper, H. H. (1958). Report of the committee on methods of clinical examination in electroencephalography: 1957. *Electroencephalography and Clinical Neurophysiology, 10*(2), 370-375. doi:10.1016/0013-4694(58)90053-1

Kaan, E., & Swaab, T. Y. (2003). Repair, revision, and complexity in syntactic analysis: An electrophysiological differentiation. *Journal of Cognitive Neuroscience, 15*(1), 98-110. doi:10.1162/089892903321107855

Kim, A., & Osterhout, L. (2005). The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of Memory and Language, 52*(2), 205-225. doi:10.1016/j.jml.2004.10.002

Kolk, H. H. J., & Chwilla, D. J. (2007). Late positivities in unusual situations. *Brain and Language, 100*(3), 257-261. doi:10.1016/j.bandl.2006.07.006

Kolk, H. H. J., Chwilla, D. J., Van Herten, M., & Oor, P. J. (2003). Structure and limited capacity in verbal working memory: A study with event-related potentials. *Brain and Language, 85*(1), 1-36. doi:10.1016/s0093-934x(02)00548-5

Kos, M., Vosse, T., Van Den Brink, D., & Hagoort, P. (2010). About edible restaurants: conflicts between syntax and semantics as revealed by ERPs. *Frontiers in Psychology, 1*. doi:10.3389/fpsyg.2010.00222
Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research, 1146*, 23-49. doi:10.1016/j.brainres.2006.12.063

Kuperberg, G. R. (2013). The proactive comprehender: What event-related potentials tell us about the dynamics of reading comprehension. In: *Unraveling the behavioral, neurobiological, and genetic components of reading comprehension*. Miller, B., Cutting, L., & McCardle, P. (Eds): Baltimore: Paul Brookes Publishing.

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

Kuperberg, G. R., Sitnikova, T., Caplan, D., & Holcomb, P. J. (2003). Electrophysiological distinctions in processing conceptual relationships within simple sentences. *Cognitive Brain Research, 17*(1), 117-129. doi:10.1016/s0926-6410(03)00086-7

Kutas, M. (1993). In the company of other words: Electrophysiological evidence for single-word and sentence context effects. *Language and Cognitive Processes, 8*(4), 533-572. doi:10.1080/01690969307587

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences, 4*(12), 463-470. doi:10.1016/S1364-6613(00)01560-6

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:10.1016/s1364-6613(00)01560-6
Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science, 207*(4427), 203-205. doi:10.1126/science.7350657

Lau, E. F., Holcomb, P. J., & Kuperberg, G. R. (2013). Dissociating N400 effects of prediction from association in single-word contexts. *Journal of Cognitive Neuroscience, 25*(3), 484-502. doi:10.1162/jocn_a_00328

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

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition, 25*(1), 71-102. doi:10.1016/0010-0277(87)90005-9

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review, 88*(5), 375. doi:10.1037//0033-295x.88.5.375

Metzner, P., Malsburg, T., Vasisht, S., & Rösler, F. (2017). The importance of reading naturally: Evidence from combined recordings of eye movements and electric brain potentials. *Cognitive Science, 41*(S6), 1232-1263. doi:10.1111/cogs.12384

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Reason, 4*(2), 61-64. doi:10.20982/tqmp.04.2.p061

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.

doi:10.1101/111807

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia, 9*(1), 97-113. doi:10.1016/0028-3932(71)90067-4

Paczynski, M., Kreher, D. A., Ditman, T., Holcomb, P., & Kuperberg, G. R. (2006). Electrophysiological evidence for the role of animacy and lexico-semantic associations in processing nouns within passive structures. *Cogn. Neurosci. Suppl., Abstract.*

Paczynski, M., & Kuperberg, G. R. (2012). Multiple influences of semantic memory on sentence processing: Distinct effects of semantic relatedness on violations of real-world event/state knowledge and animacy selection restrictions. *Journal of Memory and Language, 67*(4), 426-448.

doi:10.1016/j.jml.2012.07.003

Rayner, K., Warren, T., Juhasz, B. J., & Liversedge, S. P. (2004). The effect of plausibility on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(6), 1290.

Schwanenflugel, P. J., & Shoben, E. J. (1985). The influence of sentence constraint on the scope of facilitation for upcoming words. *Journal of Memory and Language, 24*(2), 232-252.

doi:10.1016/0749-596x(85)90026-9

Staub, A. (2015). The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation. *Language and Linguistics Compass, 9*(8), 311-327.

doi:10.1111/lnc3.12151
Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior, 18*(6), 645-659. doi:10.1016/s0022-5371(79)90355-4

Szewczyk, J. M., & Schriefers, H. (2013). Prediction in language comprehension beyond specific words: An ERP study on sentence comprehension in Polish. *Journal of Memory and Language, 68*(4), 297-314. doi:10.1016/j.jml.2012.12.002

Thornhill, D. E., & Van Petten, C. (2012). Lexical versus conceptual anticipation during sentence processing: Frontal positivity and N400 ERP components. *International Journal of Psychophysiology, 83*(3), 382-392. doi:10.1016/j.ijpsycho.2011.12.007

van Berkum, J. J. A., 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. doi:10.1037/0278-7393.31.3.443

van Berkum, J. J. A., Zwitserlood, P., Hagoort, P., & Brown, C. M. (2003). When and how do listeners relate a sentence to the wider discourse? Evidence from the N400 effect. *Cognitive Brain Research, 17*(3), 701-718. doi:10.1016/s0926-6410(03)00196-4

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

Warren, T., & McConnell, K. (2007). Investigating effects of selectional restriction violations and plausibility violation severity on eye-movements in reading. *Psychonomic Bulletin & Review, 14*(4), 770-775. doi:10.3758/bf03196835
Zwitserlood, P. (1989). The locus of the effects of sentential-semantic context in spoken-word processing. *Cognition, 32*(1), 25-64. doi:10.1016/0010-0277(89)90013-9
We use “predictable” and “expected” interchangeably to characterize continuations that are highly expected given the preceding discourse, with predictability assessed by means of a cloze procedure.
Figure 1

Representative anterior and posterior scalp channels.

ERPs for EXP, USP, ANOM-Pos and ANOM-Impos nouns. Displayed channels are marked as stars on the electrode montage mapping (E). Dashed-line boxes indicate analyzed time windows (N400: 300-500 ms; post-N400 positivity: 600-1000 ms).
Figure 2

Mass univariate analyses.

Raster plots of t-values with control for false discovery rates in two dimensional grids of the following comparisons: (A) USP nouns minus EXP nouns, (B) ANOM nouns minus EXP nouns, and (C) ANOM nouns minus USP nouns. Results are plotted in 4 ms lags. Left scalp electrodes are displayed in the upper section, midline scalp electrodes in the center, and right scalp electrodes in the lower section of each panel. Red (blue) indicates that ERPs to the first noun type are more positive (negative) than ERPs to the second noun type. See Fig. 1E for electrode placement.
Figure 3

Grand average (n=32) recorded over 30 scalp channels.
Figure 4

Topographic scalp maps.

ERP mean voltage differences of the three main comparisons for time points 300 to 1100 ms.
Figure 5

Comparison of results

EEG results from the current study (A-C) and Experiment 2 by DeLong, Quante, and Kutas (2014) (D-F), for three time windows (N400, Anterior Positivity, Posterior Positivity) and three noun conditions (EXP, USP, ANOM). Significance levels of pairwise t-tests were Bonferroni-adjusted ($p_{boncor} < .0167$).
Table 1 (on next page)

Sample sentence pairs
1. Peter stand bei Morgendämmerung auf, fuhr den ganzen Tag Traktor und fütterte abends seine Kühe. An manchen Tagen wäre er aber lieber kein [Bauer, Erwachsener, Trick] sondern ein unbekümmertes Kind. (Peter gets up at dawn, drives the tractor all day and feeds his cows in the evening. On some days he would rather not be a [farmer, adult, trick] but a carefree child.)

Comprehension question: Does Peter have cows?

2. Alice brach sich ihr Bein im Wanderurlaub. Der Arzt röntgte ihr Bein und legte es in einen [Gips, Rollstuhl, Vogel] für zehn Wochen. (Alice broke her leg while hiking. The doctor x-rayed her leg and put it in a [cast, wheelchair, bird] for ten weeks.)

3. Anne schrieb gerade ihre Masterarbeit und brauchte noch weitere Quellen für ihre Annahmen. Deshalb machte sie sich auf den Weg in eine [Bibliothek, Lehrbuchsammlung, Feder] für ihren Fachbereich. (Anne was writing her master’s thesis and needed more sources for her assumptions. Therefore, she made her way to a [library, textbook collection, feather] for her department.)

Comprehension question: Was Luisa’s new room very small?

4. Luisas neues WG-Zimmer war sehr klein, hatte aber hohe Decken. Um Platz zu sparen, kauft sie sich deshalb ein [Hochbett, Aufbewahrungssystem, Schwein] im Baumarkt. (Luisa’s new room was very small but had high ceiling. To save space, she bought herself a [loft bed, storage system, pig] in the store.)

5. Frank hält sich selbst für einen Komiker. Trotzdem kennt er nicht einen [Witz, Schauspieler, Anzug] oder Sketch, über den sein Publikum lachen würde. (Frank considers himself quite a comedian. But he doesn’t know a [joke, actor, suit] or sketch his audience would laugh about.)

6. Marleen war schüchtern und konnte nicht gut mit Lob umgehen. Sie war peinlich berührt durch ein [Kompliment, Tattoo, Bügeleisen] ihres Vorgesetzten. (Marleen was very shy and could not handle praise well. She was embarrassed by a [compliment, tattoo, iron] from her supervisor.)

7. Marina war viel auf Reisen und erlebte fast jeden Tag etwas Neues. Um sich an alles zu erinnern, schrieb sie ein [Tagebuch] und klebte Fotos dazu. (Marina travels a lot and has new experiences almost every day. To remember everything, she writes a [diary] and adds pictures.)
**Table 2** (on next page)

Stimuli characteristics

| Stimulus | Characteristics | Value |
|----------|-----------------|-------|
| Feature 1 | Description 1  | Value 1 |
| Feature 2 | Description 2  | Value 2 |
| Feature 3 | Description 3  | Value 3 |
| Condition label | Condition                        | Number of items | Mean critical noun cloze probability (SD), Range: 0-1 | Mean context + noun plausibility rating (SD), Range: 1-5 | Mean context + noun possibility rating (SD), Range: 1-4 | Mean contextual constraint (SD), Range: 0-1 | Mean critical noun written frequency (SD)* | Mean critical noun length (SD)* | Mean critical orthographic neighborhood size (SD)^
|-----------------|---------------------------------|-----------------|-------------------------------------------------------|--------------------------------------------------------|--------------------------------------------------------|---------------------------------|---------------------------------|-------------------------------|--------------------------------|
| Expected        | High cloze/High plausibility    | 150             | 0.77 (0.14)                                          | 4.68 (0.36)                                             | 3.77 (0.33)                                             | 0.77 (0.14)                                  | 2148.27 (4233.85)                   | 6.91 (2.55)                      | 11.99 (16.08)                    |
| Unexpected      | Somewhat Plausible (USP)        | 150             | <0.01 (<0.01)                                        | 2.96 (0.99)                                             | 3.19 (0.56)                                             | 0.77 (0.14)                                  | 2551.02 (5749.39)                   | 7.38 (3.15)                      | 12.76 (17.99)                    |
| ANOMalous       | Low cloze/High plausibility     | 150             | <0.01 (<0.01)                                        | 1.05 (0.13)                                             | 1.44 (0.45)                                             | 0.77 (0.14)                                  | 2278.32 (4278.45)                   | 6.95 (2.75)                      | 12.89 (18.00)                    |
| - ANOM-Impos    | ANOM + impossible meaning      | 105             | <0.01 (<0.01)                                        | 1.02 (0.11)                                             | 1.20 (0.18)                                             | 0.77 (0.14)                                  | 2659.35 (4840.02)                   | 6.57 (2.45)                      | 14.89 (19.81)                    |
| - ANOM-Pos      | ANOM + possible meaning         | 45              | <0.01 (<0.01)                                        | 1.11 (0.14)                                             | 2.00 (0.38)                                             | 0.78 (0.14)                                  | 1369.05 (2268.36)                   | 7.84 (3.02)                      | 8.14 (11.53)                     |
| FILL            | High cloze/High plausibility    | 50              | 0.76 (0.18)                                          | -                                                      | -                                                      | 0.76 (0.18)                                  | 3276.12 (5290.63)                   | 7.20 (2.86)                      | 9.62 (11.36)                     |

* Absolute annotated type frequencies according to dlexDB (http://www.dlexdb.de/)

Ornographic neighborhood size (as defined by Coltheart, Davelaar, Jonasson, & Besner, 1977) according to dlexDB
Table 3 (on next page)

Differences between conditions
| Comparison | Plausibility | Possibility | Word frequency | Orthographic neighbors | Word length | Contextual constraint |
|------------|--------------|-------------|----------------|------------------------|-------------|----------------------|
| EXP vs. USP | $t(298) = 19.89$, $p < .001$ | $t(298) = 10.81$, $p < .001$ | $t(298) = -0.63$, $p = .528$ | $t(298) = -0.30$, $p = .765$ | $t(298) = -1.43$, $p = .154$ | $t(298) =$ |
| EXP vs. ANOM | $t(298) = 116.21$, $p < .001$ | $t(298) = 51.10$, $p < .001$ | $t(298) = -0.26$, $p = .792$ | $t(298) = -0.45$, $p = .650$ | $t(298) = -0.15$, $p = .879$ | $t(298) =$ |
| USP vs. ANOM | $t(298) = 23.47$, $p < .001$ | $t(298) = 29.96$, $p < .001$ | $t(298) = 0.41$, $p = .684$ | $t(298) = -0.15$, $p = .882$ | $t(298) = 1.25$, $p = .213$ | $t(298) =$ |
| ANOM-Pos vs. ANOM-Impos | $t(67.58) = 3.88$, $p < .001$ | $t(53.11) = 13.62$, $p < .001$ | $t(145.15) = -2.21$, $p = .028$ | $t(131.79) = 2.60$, $p = .011$ | $t(67.00) = 2.38$, $p = .020$ | $t(85.06) =$ |

Because of unequal group sizes, a Welch-test was conducted in case of ANOM-Pos vs. ANOM-Impos. Significant p-values are marked in bold.
Table 4 (on next page)

Mean amplitude and standard deviation (μV) of the four noun types across time windows and different scalp sites
|                | EXP   | USP   | ANOM  | ANOM-Pos | ANOM-Impos |
|----------------|-------|-------|-------|----------|------------|
| N400           | 0.83  | -1.54 | -2.89 | -3.01    | -2.82      |
| Anterior Positivity | 0.79  | 1.91  | 1.51  | 1.96     | 1.33       |
| Posterior Positivity | 0.98  | 1.38  | 3.09  | 2.95     | 3.16       |
Table 5 (on next page)

Pairwise t-tests between the four noun types
|                          | Mean of the differences [μV] | t(31) | p         | 95% confidence interval |
|--------------------------|-----------------------------|-------|-----------|-------------------------|
| **300-500 ms, all scalp sites (N400)** |                             |       |           |                         |
| EXP vs. USP              | 2.37                        | 8.48  | <.001*    | [1.80; 2.94]            |
| EXP vs. ANOMI            | 3.65                        | 9.26  | <.001*    | [2.85; 4.46]            |
| EXP vs. ANOMP            | 3.84                        | 9.83  | <.001*    | [3.04; 4.64]            |
| USP vs. ANOMI            | 1.28                        | 4.38  | <.001*    | [0.68; 1.88]            |
| USP vs. ANOMP            | 1.47                        | 5.83  | <.001*    | [0.95; 1.98]            |
| ANOMP vs. ANOMI          | -0.19                       | -0.72 | .479      | [-0.72; 0.35]           |
| **600-1000 ms, posterior scalp sites** |                             |       |           |                         |
| EXP vs. USP              | -0.40                       | -1.31 | .200      | [-1.02; 0.22]           |
| EXP vs. ANOMI            | -2.17                       | -5.40 | <.001*    | [-2.99; -1.35]          |
| EXP vs. ANOMP            | -1.97                       | -4.29 | <.001*    | [-2.90; -1.03]          |
| USP vs. ANOMI            | -1.77                       | -4.87 | <.001*    | [-2.52; -1.03]          |
| USP vs. ANOMP            | -1.57                       | -4.75 | <.001*    | [-2.24; -0.89]          |
| ANOMP vs. ANOMI          | -0.21                       | -0.72 | .477      | [-0.80; 0.38]           |
| **600-1000 ms, anterior scalp sites** |                             |       |           |                         |
| EXP vs. USP              | -1.12                       | -3.97 | <.001*    | [-1.70; -0.55]          |
| EXP vs. ANOMI            | -0.54                       | -1.52 | .138      | [-1.27; 0.18]           |
| EXP vs. ANOMP            | -1.17                       | -2.88 | .007*     | [-2.00; -0.34]          |
| USP vs. ANOMI            | 0.58                        | 1.50  | .145      | [-0.21; 1.37]           |
| USP vs. ANOMP            | -0.05                       | -0.12 | .904      | [-0.84; 0.74]           |
| ANOMP vs. ANOMI          | 0.63                        | 1.89  | .068      | [-0.05; 1.31]           |

* Significant after Bonferroni adjustment ($p_{boncor} < .0083$)