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

Laura Quante[a,b], Jens Bölte[a,b], Pienie Zwitserlood[a,b]

[a] Department of Psychology, University of Münster, Fliednerstraße 21, 48149 Münster, Germany

[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

1   **Abstract**

2   Late positive ERP components occurring after the N400, traditionally linked to reanalysis

3   due to syntactic incongruence, are increasingly considered to also reflect reanalysis and repair due

4   to semantic difficulty. Semantic problems can have different origins, such as a mismatch of

5   specific predictions based on the context, low plausibility, or even semantic impossibility of a

6   word in the given context. DeLong, Quante, and Kutas (*Neuropsychologia*, 2014) provided the

7   first direct evidence for topographically different late positivities for prediction mismatch (left

8   frontal late positivity for plausible but unexpected words) and plausibility violation (posterior-

9   parietal late positivity for implausible, incongruent words). The aim of the current study is twofold:

10   (1) to replicate this dissociation of ERP effects for plausibility violations and prediction mismatch

11   in a different language, and (2) to test an additional contrast within implausible words, comparing

12   impossible and possible sentence continuations. Our results replicate DeLong et al. (2014) with

13   different materials in a different language, showing graded effects for predictability and

14   plausibility at the level of the N400, a dissociation of plausible and implausible, anomalous

15   continuations in posterior late positivities and an effect of prediction mismatch on late positivities

16   at left-frontal sites. In addition, we found some evidence for a dissociation, at these left-frontal

17   sites, between implausible words that were fully incompatible with the preceding discourse and

18   those for which an interpretation is possible. We discuss the theoretical impact of our results in the

19   light of prediction, updating and integration of words into the discourse.

1 Introduction

24   The study of effects of context on language processing has a long tradition in

25   psycholinguistics, as modular (cf. Forster, 1981) and interactive (cf. McClelland & Rumelhart,

26   1981; Marslen-Wilson, 1987) theories of word recognition drastically differed with respect to the

27   role allotted to information stemming from sources other than the word itself. Proof for an impact

28   of contextual, top-down information on word recognition was already provided more than 30 years

29   ago, with priming paradigms and reaction time data (cf. Swinney, 1979; Schwaneflugel & Shoben,

30   1985). The advent of event-related potentials again fired the debate, because they allow insights

31   into the time-course of word recognition, which is difficult to come by with reaction times (but see

32   Zwitserlood, 1989). Ever since, a wealth of studies has shown that contextual information, when

33   constraining enough, has an early impact on lexical processing – even to the extent that upcoming

34   words are anticipated (Kutas & Federmeier, 2000; van Berkum, Zwitserlood, Hagoort, & Brown,

35   2003; DeLong, Urbach, & Kutas, 2005; van Berkum, Brown, Zwitserlood, & Hagoort, 2005). It

36   is thus not surprising that terminology has changed, and "anticipation" and "prediction" are now

37   used to refer to the impact, on lexical processing, of knowledge from sources other than the current

38   input (cf. Van Petten & Luka, 2012; Huettig & Janse, 2016; Kuperberg & Jaeger, 2016). Whereas

39   most researchers agree that (features of) upcoming words are predicted under certain

40   circumstances, it remains unresolved which factors promote (or prevent) predictive processing,

41   and what information about words (e.g., semantics, word forms) is predicted (see Ito et al., 2016;

42   Kuperberg & Jaeger, 2016).

43   To study effects of semantic context, expectation and prediction in language

44   comprehension, a particular event-related potential (ERP) component, the N400 (Kutas &

45   Hillyard, 1980), has been used extensively. The N400 is a negative-going wave peaking around

46   400 ms after stimulus onset, which is related to semantic processing (for a review, see Kutas &

47   Federmeier, 2011). For example, its amplitude is negatively correlated to a word's cloze

48   probability (proportion of respondents who completed a given context with this particular word),

49   a measure of semantic expectancy. Words with strong contextual support show a decrease in N400

50   amplitude relative to words that are less predictable or do not fit the context (Kutas & Federmeier,

51   2011). There is also evidence for ERP effects as a function of predictability in time windows

52   preceding the N400 (e.g., van Berkum, Zwitserlood, Hagoort, & Brown, 2003; Dikker &

53   Pylkkänen, 2011; Lau, Holcomb, & Kuperberg, 2013; Brothers, Swaab, & Traxler, 2015; see

54   Kuperberg & Jaeger, 2016, for an overview). However, evidence for the actual pre-activation or

55   anticipation of upcoming words, assessed before any of their input becomes available, is less

56   abundant (but see DeLong, Urbach, and Kutas, 2005; van Berkum, Brown, Zwitserlood, Kooijman

57   & Hagoort, 2005; Szewczyk & Schriefers, 2013; Ito, Corley, Pickering, Martin, & Nieuwland,

58   2016).

59   Our study uses ERPs and does not focus on prediction or expectation per se, but on the

60   consequences of prediction or expectation mismatch, and, more generally speaking, of contextual

61   mismatch.[i] Van Petten and Luka (2012) proposed that if listeners and readers predict upcoming

62   words, the EEG signal should reflect not only benefits of a confirmed prediction (visible as

63   attenuation of the N400) but also costs of a disconfirmed prediction. In their review article, they

64   assessed studies that compared congruent sentence completions with semantically anomalous

65   completions, and often observed a late positivity, about 600-900 ms after critical-word onset, with

66   a mainly parietal scalp topography. In addition, an anterior positivity was sometimes observed

67   when ERPs for unexpected but semantically congruent sentence completions were compared to

68   predictable, expected completions. It should be noted, however, that the 60+ studies included

69   showed a great variability in the post-N400 time window.

70   It thus seems that unexpected continuations that allow construction of a possible overall

71   sentence meaning differ from anomalous completions. Interestingly, studies that manipulate

72   semantic expectancy have predominantly used anomalous or unexpected plausible completions,

73   but rarely both. This motivated DeLong, Quante, and Kutas (2014) to contrast different levels of

74   plausibility within the same study, to determine how predictability and plausibility each contribute

75   to word recognition. As completions of highly constraining sentence pairs (*For the snowman's*

76   *eyes, the kids used two pieces of coal. For his nose, they used ....*), DeLong et al. compared ERPs

77   to highly predicable, expected (*a carrot*), unexpected but somewhat plausible (*a banana*), and

78   unexpected, implausible, anomalous (*a groan*) words. The unexpected but plausible continuations

79   should induce costs of disconfirmed prediction, combined with effort to integrate the unexpected

80   noun - signaled by frontal late positivity. This does not hold for anomalous continuations that

81   cannot be integrated with the current context. DeLong and colleagues observed a posterior late

82   positivity to anomalous completions, and an anterior late positivity to unexpected but plausible

83   completions, thus confirming Van Petten and Luka's (2012) conjecture. Corroboration for a

84   particular function of the frontal late positivity, also labeled frontal PNP (post-N400 positivity), in

85   prediction-related revision was recently provided by Swaab and colleagues (Boudewyn, Long, &

86   Swaab, 2015; Brothers, Swaab, & Traxler, 2015).

87   Predictability thus seems to influence early stages of processing, whereas plausibility

88   seems to affect late stages of processing, which is corroborated by eye-tracking studies (Staub,

89   2015, for an overview). Interestingly, Rayner, Warren, Juhasz, & Liversedge et al. (2004) and

90   Warren and McConnell (2007) further distinguished between plausibility and possibility, by

91   comparing words that result in implausible but possible meaning for the full sentence, to words

92   that induce an impossible overall sentence meaning, because they violate selection restrictions

93   (e.g., "inflate a carrot") for example. In both studies, effects of words leading to either impossible

94   or implausible sentence meaning were dissociable in eye-movement measures.

95   Processing differences between implausible and impossible sentence overall meaning are

96   also visible in EEG data. For example, Paczynski and Kuperberg (2012) showed that selection-

97   restriction violations evoked a posterior positivity between 700 to 900 ms after critical word onset,

98   whereas violations of world knowledge, which result in implausible but still possible sentence

99   meaning, did not differ from plausible sentences in this time window. Similar results were shown

100   by Kuperberg, Sitnikova, Caplan, and Holcomb (2003), Geyer, Holcomb, Kuperberg, and

101   Perlmutter (2006), and Paczynski, Kreher, Ditman, Holcomb, and Kuperberg (2006). When

102   Kuperberg (2007) evaluated factors evoking a late positivity, she concluded that none of the

103   following factors - the presence of selection-restriction violations, semantic associations between

104   the critical word and the preceding context, specific task instructions, or constraining context - by

105   themselves could explain all results. One hypothesis that she advanced was that the impossibility

106   to establish an overall meaning for the sentence might be the crucial factor inducing a late positivity

107   on the critical word.

108   Given the variable nature of late positivities, and given the dire need for replication studies

109   of phenomena that are rather new or for which evidence is scarce (see Nieuwland et al., 2017; see

110   also Dennis & Valacich, 2014), the present study aimed to replicate DeLong et al.'s (2014) second

111   experiment, with German stimuli presented to German native speakers. In addition, inspired by

112   suggestions made by Kuperberg (2007) and DeLong et al. (2014), we analyzed differences

113   between implausible word completions that resulted in either possible or impossible overall

114   sentence meaning. To create a condition of impossible sentence meaning, we divided the materials

115   into impossible and possible sets by means of subjective possibility ratings, collected in a pretest.

116   Following DeLong et al. (2014), we predicted a graded effect of contextual fit of critical

117   words at the level of the N400, with implausible continuations showing enhanced negativity

118   relative to unexpected but plausible words. Next, we expect a predictability effect, showing as an

119   anterior late positivity – relative to expected nouns - to unexpected but plausible sentence

120   completions, but not to implausible nouns. Next, we predict a plausibility effect, with a posterior

121   positivity only for implausible, anomalous sentence completions. If Kuperberg's (2007)

122   assumption is correct, we predict this posterior late positivity only for those sentence completions

123   that are truly anomalous and lead to an impossible overall sentence meaning, but not for those that

124   allow an integration of the critical word with the preceding discourse, resulting in a perhaps

125   implausible but nevertheless possible real-world meaning. This would constitute an effect of

126   possibility, which also might show in a difference between possible and impossible implausible

127   continuations in late positivity at anterior sites, with the possible continuations coinciding with

128   plausible but unexpected ones.

129

130   **2 Material and Methods**

131   **2.1 Stimuli**

132   Stimuli were 150 constraining German sentence pairs (mean contextual constraint = 0.77,

133   SD = 0.14, see cloze probability norming described below), which led to expectations for particular

134   sentence-medial words. Following the condition labels used in DeLong et al. (2014), each of the

135   150 contexts was completed by a) the semantically expected noun (with the highest cloze

136 probability for the specific context; EXP), b) an unexpected but somewhat plausible noun (USP),

137 and c) an unexpected, implausible noun (ANOM), resulting in a total of 450 sentence pairs (see

138 Table 1 for sample sentence pairs; the complete set of sentence pairs is provided in Supplemental

139 Table S1). To investigate whether the possible construction of overall sentence meaning was

140 crucial for late positivities, the materials in the ANOM condition were subdivided on the basis of

141 a pretest. Some of the sentences pairs in the ANOM condition contained critical nouns that allowed

142 for a possible real-life meaning (ANOM-Pos; 45 sentence pairs), the other sentence pairs did not

143 (ANOM-Impos; 105 sentence pairs). Fifty additional moderately constraining sentence pairs

144 completed by their expected critical noun were used as fillers to balance the proportion of sentence

145 pairs completed by expected versus unexpected nouns. Sentence material were either German

146 translations of stimuli used in DeLong et al. (2014) or constructed in the same fashion by the

147 experimenters. Where possible, critical nouns of the expected condition were re-used with

148 different sentences in the other two conditions (53.3% of critical words were used three times,

149 31.1% were used twice, and 15.6% were used only once). Since German nouns are coded for

150 gender (masculine, feminine, neutral), all three completions of a particular sentence pair had the

151 same grammatical gender. Written word frequency, word length and orthographic neighborhood

152 size of the critical nouns were matched between the three main conditions (see Table 2). Note that

153 the Pos and Impos items within the ANOM condition were not balanced with respect to these

154 factors.

155 **2.2 Cloze probability norming**

156 Stimulus norming for critical noun cloze probability was conducted in a separate sentence

157 completion task with 36 volunteers (native speakers of German, mainly students). They were

158 compensated with course credit and did not participate in the EEG study. Contexts were truncated

159   prior to the critical noun, and participants were asked to complete the second sentence with a single

160   noun that came to their mind first and fitted with the preceding context. Every context ended with

161   the three German indefinite articles, in the order masculine, feminine, and neuter, thus allowing

162   nouns of any grammatical gender. Cloze probability was calculated as the proportion of

163   participants who completed a particular sentence pair with a particular noun. The cloze probability

164   of the most frequent noun equals the contextual constraint of a given sentence pair. Sentence pairs

165   with a cloze probability of 50% or higher were considered highly constraining and included in the

166   study. Filler sentences had a cloze probability of 40% or higher. Table 2 shows mean cloze

167   probabilities for the experimental and filler conditions.

168   **2.3 Plausibility rating**

169   All 450 sentence pairs, truncated after the critical noun, were rated for plausibility ("How

170   plausible is the sentence pair's meaning…") on a scale of 1 (not plausible) to 5 (highly plausible)

171   by eight independent German raters who did not participate in the EEG study. Table 2 presents

172   mean plausibility ratings of all conditions. Following DeLong et al. (2014), mean plausibility was

173   greater than 1.5 in the EXP and USP conditions, and less than or equal to 1.5 in the ANOM

174   condition. The plausibility ratings differed significantly between all conditions (see Table 3).

175   **2.4 Possibility rating**

176   All 450 sentence pairs, truncated after the critical noun, were rated for possibility ("How

177   possible (in real-life) is the sentence pair's meaning…") by the same eight raters, on a scale of 1

178   (impossible) to 4 (possible). Table 2 specifies mean possibility ratings for all conditions. Similar

179   to the cut-off for Plausibility, mean possibility ratings were greater than 1.5 in the EXP, USP and

180   ANOM-Pos conditions, but equal to or less than 1.5 in the ANOM-Impos condition. The ratings

181 of all conditions differed significantly from each other (see Table 3). Participants performed both

182 plausibility and possibility ratings at the same time. No examples were provided to avoid biasing

183 the raters' judgements. Correlations between plausibility and possibility ratings are displayed in

184 Supplemental Analysis S1.

185   In the main experiment, each participant was presented with one of three 200-item lists,

186 with contexts and critical nouns used once per list (except for four critical nouns that occurred

187 twice per list, in different contexts). Lists 1, 2, and 3 were presented to 12, 10 and 10 participants,

188 respectively. Every list consisted of 50 predictable, expected nouns, 50 unexpected plausible

189 nouns, 50 unexpected implausible (ANOM) nouns, and 50 fillers. Approximately one third of the

190 ANOM nouns was rated possible (list 1: 17, list 2: 16, list 3: 12), the remaining two thirds were

191 rated impossible. Fifty comprehension questions followed 25 % of sentence pairs at random

192 intervals. Three additional sentence pairs preceded every list to familiarize participants with the

193 task. Sentence pairs within a list were randomized across subjects.

194 **2.5 ERP Participants**

195   Thirty-two students (23 f, 9 m) participated in the experiment after giving written informed

196 consent. They were compensated with course credit or cash (7.50 €/hour). Mean age was 25.3

197 years (19-34). All participants were monolingual native speakers of German and right-handed

198 (assessed via Edinburgh Handedness Inventory, Oldfield, 1971). Eight participants reported a left-

199 handed parent or sibling, one reported two left-handed relatives. All participants reported normal

200 or corrected-to-normal vision. One additional participant was tested but excluded from analysis

201 because of a technical problem during the experiment. The study protocol was conducted in

202 accordance with ethical standards of the Declaration of Helsinki and approved by the local ethics

203 committee of the University of Münster (approval number #2016-42-LQ).

2.6 Procedure

204

205   The experiment consisted of a single two-hour-session conducted in a quiet and dimly lit

206   room at the Westfälische Wilhelms-Universität Münster. Participants were seated approximately

207   1 m in front of a LED monitor (BenQ, model XL2420T, 144 Hz, 24"W) and read sentence pairs

208   for comprehension. The experiment was set up using Presentation software (NeuroBehavioral

209   Systems, Version 16.3). Stimuli were presented visually, in black type (RGB: 0, 0, 0; Arial 48 pt)

210   on a grey background (RGB: 148, 148, 148). The experiment was divided into eight blocks of

211   approximately six minutes length, with two-minute breaks between blocks. Every trial started with

212   a fixation cross (500 ms) in the center of the screen, followed by the first sentence of a pair

213   presented in its entirety. Participants advanced to the critical sentence via button press. This

214   sentence including the critical word was presented with a rapid serial visual presentation technique

215   (RSVP), each word presented centrally for 200 ms, with a stimulus onset asynchrony of 500 ms.

216   Yes/no comprehension questions followed 25 % of sentence pairs at random intervals. Participants

217   responded with two buttons on a response pad (Cedrus, model RB-830) with response buttons

218   counterbalanced across participants and lists. Comprehension questions appeared after the critical

219   noun sentence. In case of a question, participants' button press advanced to the next trial, otherwise

220   the next sentence pair appeared automatically after two seconds.

221   Material, design and procedure were almost identical to DeLong et al.'s second experiment

222   except for the following differences. First, DeLong et al.'s sentence material was translated into

223   German or constructed using the same sentence structure. Second, mean constraint of discourse

224   contexts and mean cloze probability of expected critical words were lower than in DeLong et al.

225   (.77 vs .89, respectively). Third, sentence pairs were not rated for possibility in DeLong's study.

226   Fourth, 32 participants completed the present EEG experiment, 24 students participated in DeLong

227   et al.'s second study.

228   **2.7 Electroencephalographic recording parameters**

229   EEG was recorded from 32 Ag/AgCl-electrodes attached to a WaveGuard 32-channel cap

230   (ANT, Advanced Neuro Technology). Electrodes were placed according to the International 10-

231   20 convention (Jasper, 1958), and an average reference was used (see Fig. 1 for scalp sites). Blinks

232   and vertical eye movements were monitored from electrodes placed above and below the left eye,

233   and horizontal eye movements were monitored from two electrodes placed on the outer canthi.

234   Impedances were kept below 5kΩ. The EEG was continuously recorded with ASA (Advanced

235   Source Analysis, version 4.7.3.1, ANT). Data collection and evaluation were controlled by ExMan

236   (Experiment Manager; MS Excel worksheet with active macros). EEG was amplified (ExG 20x,

237   fixed = 50 mV/V), low pass filtered (finite impulse response filter, cut-off frequency = .27 x

238   sampling rate) and continuously digitized at a sampling rate of 256 samples/second.

239   **2.8 Data analysis**

240   Before averaging, the EEG signal was filtered using a Butterworth half-amplitude bandpass

241   FIR-filter (0.1 Hz, 20 Hz, 12 db/oct). Vertical eye movements were corrected with principal

242   component analysis (Ille, Berg, & Scherg, 2002). Additionally, seven electrodes (0.7 %) were

243   interpolated (see Supplemental Table S2). The EEG was re-referenced offline to the algebraic

244   mean of left and right mastoids and averaged for each experimental condition, time-locked to the

245   critical noun onset. Before averaging, trials contaminated by artefacts (specified as voltage

246   changes exceeding ± 75 µV during the epoch) were rejected offline (on average 4.23 % of all trials,

247   SD = 4.80). ERPs were calculated for epochs extending from 500 ms pre- to 1500 ms post-stimulus

248   onset, thus using a pre-stimulus baseline of 500 ms.

249   In a first step, mass univariate analyses were conducted to compare spatial and temporal

250   properties of possible ERP effects found in the present experiment to the findings by DeLong et

251   al. (2014). ERPs from the three pairwise comparisons, [USP minus EXP], [ANOM minus EXP],

252   and [ANOM minus USP], were submitted to repeated measures, two-tailed t-tests at all sampled

253   time points between 250 ms and 1050 ms (206 total time points) at all 30 scalp electrodes, resulting

254   in 6180 total comparisons for each condition contrast. To control the number of false discoveries,

255   the Benjamini and Yekutieli (2001) procedure was applied using a false discovery rate level of

256   5%.

257   In a second step, mean amplitudes were analyzed by first conducting ANOVAs with three

258   levels of noun type (EXP, USP and ANOM) to compare the present results directly to those

259   reported in DeLong et al. (2014). These analyses were complemented by pairwise t-tests between

260   the four levels of noun type (EXP, USP, ANOM-Pos, and ANOM-Impos). ANOVAs were applied

261   to the data from three time windows: a) over all 30 electrode sites between 300 ms and 500 ms

262   (N400), b) over seven (left) anterior electrode sites [Fp1, Fpz, F7, F3, Fz, FC5, T7] between 600

263   ms and 1000 ms (frontal positivity), c) over seven posterior electrode sites [Cz, CP1, CP2, P3, Pz,

264   P4, POz] between 600 ms and 1000 ms (posterior positivity; see Fig. 1 for electrode placement).

265   Scalp regions and temporal windows were based on DeLong et al.'s (2014) second experiment.

266   To confirm the left lateralization of the anterior positivity, we extended the corresponding

267   ANOVA by the factor hemisphere (left, right), and included equivalent right hemisphere

268   electrodes (Fp2, F4, F8, Fc6, and T8) while excluding midline electrodes Fz and Fpz. If sphericity

269   was violated, ANOVA p-values and degrees of freedom were corrected using epsilon correction

270 (Greenhouse Geisser) for repeated measures with more than one degree of freedom. Significance

271 levels of pairwise t-tests were Bonferroni-adjusted (t-tests following ANOVAs with three levels

272 of noun type: $p_{boncor} < .0167$, t-tests comparing all four levels of noun type: $p_{boncor} < .0083$).

273

274 <h1 style="text-align:center">3 Results</h1>

275 **3.1 Behavioral results**

276 Participants correctly answered an average of 96.7% (median 97%, range = 90-100%) of

277 yes/no comprehension questions, suggesting they comprehended the sentence pairs during the

278 experiment.

279 **3.2 ERP results**

280 **3.2.1 Mass univariate analyses.**

281 The first mass univariate analysis focused on predictability, comparing USP versus EXP

282 nouns (see Fig. 2, top). There was a widespread N400 effect, with ERPs to USP nouns being more

283 negative than ERPs to EXP nouns. This negativity lasted from approximately 250 to 500 ms.

284 Starting shortly before the offset of the N400, between approximately 550 and 1000 ms, USP

285 nouns were more positive than EXP nouns, particular over left frontal and left lateral temporo-

286 parietal scalp sites.

287 The second mass univariate analysis looked at plausibility, comparing ANOM versus EXP

288 nouns (see Fig. 2, middle). Again, a widespread N400 effect emerged, with ERPs to ANOM nouns

289 being more negative than ERPs to EXP nouns, between approximately 250 ms and 500 ms. By

290   about 600 ms, a positivity of ANOM relative to EXP nouns emerged and continued up to the end

291   of the time window (1050 ms), being most prominent over central and posterior scalp sites.

292   The third mass univariate analysis compared ANOM versus USP nouns (see Fig. 2,

293   bottom). Between 300 ms and 450 ms, ERPs to ANOM nouns were more negative than ERPs to

294   USP nouns. In addition, from approximately 600 ms to the end of the time window (1050 ms),

295   ERPs to ANOM nouns were more positive than ERPs to USP nouns at posterior scalp locations.

296   In contrast, ERPs to USP nouns were more positive than ERPs to ANOM nouns over lateral frontal

297   scalp sites between approximately 700 to 900 ms. In all three mass-univariate analyses, significant

298   p-values are $p_{adj} < .05$.

299   **3.2.2 Analyses of variance.**

300   For visual inspection, Fig. 3 shows the grand average ERPs of all 32 participants over 30

301   scalp channels. Topographic scalp maps of ERP mean amplitude voltage differences can be seen

302   in Fig. 4, and four representative anterior and posterior channels are shown in Fig. 1. In line with

303   the results from mass univariate analyses described above, all figures reveal N400 effects for both

304   USP and ANOM nouns relative to EXP nouns, a post-N400 positivity for ANOM nouns over

305   posterior channels, and a post-N400 positivity for USP nouns over anterior channels. Early

306   components (P1, N1, and P2) do not differ as a function of noun type. Tables 4 and 5 provide mean

307   amplitudes of the four noun types and detailed results of pairwise t-tests between conditions.

308   ***3.2.2.1 300-500 ms.***

309   An ANOVA with three levels of noun type over all 30 electrode sites revealed a main effect

310    $[F(1.60, 49.46) = 74.98, p < .001, \varepsilon_{GG} = 0.80, \eta_p^2 = 0.39]$. ANOM nouns showed the largest

311   negativity (-2.89 μV), followed by USP nouns (-1.54 μV) and EXP nouns (0.83 μV). Post-hoc t-

312   tests are displayed in Fig. 5. The pairwise t-tests between all four noun types revealed significant

313   differences between all conditions ($t(31) \geq 4.38$, $p < .001$) except for the comparison of ANOM-

314   Pos (-3.01 µV) and ANOM-Impos nouns (-2.82 µV; $t(31) = -0.72$, $p = .479$).

315   ***3.2.2.2 600-1000 ms posterior scalp sites.***

316   An ANOVA with three levels of noun type, conducted over seven posterior electrode sites,

317   showed a main effect [$F(2, 62) = 21.06$, $p < .001$, $\eta_p^2 = 0.19$]. ANOM nouns had the largest

318   positivity (3.09 µV), followed by USP nouns (1.38 µV) and EXP nouns (0.98 µV). Post-hoc t-tests

319   are displayed in Fig. 5. Pairwise t-tests revealed significant differences between EXP and ANOM-

320   Pos nouns ($t(31) = -4.29$, $p < .001$), EXP and ANOM-Impos nouns ($t(31) = -5.40$, $p < .001$), USP

321   and ANOM-Pos nouns ($t(31) = -4.75$, $p < .001$), and USP and ANOM-Impos nouns ($t(31) = -4.87$,

322   $p < .001$). No reliable difference was found between EXP and USP nouns ($t(31) = -1.31$, $p = .200$)

323   and between the two types of ANOM nouns (ANOM-Pos = 2.95 µV, ANOM-Impos = 3.16 µV;

324   $t(31) = -0.72$, $p = .477$).

325   ***3.2.2.3 600-1000 ms anterior scalp sites.***

326   The extended ANOVA indicated a left lateralization of the effect (see Supplemental

327   Analysis S2). Therefore, we restricted our analysis to those electrodes analyzed in DeLong et al.

328   (2014). Over the seven left anterior electrodes, the ANOVA with three levels of noun type revealed

329   a main effect [$F(2, 62) = 6.13$, $p = .004$, $\eta_p^2 = 0.08$]. USP nouns showed the greatest positivity

330   (1.91 µV), followed by ANOM nouns (1.51 µV) and EXP nouns (0.79 µV). Post-hoc t-tests are

331   displayed in Fig. 5. Paired t-tests revealed that EXP nouns differed from USP nouns ($t(31) = -3.97$,

332   $p < .001$) and from ANOM-Pos nouns ($t(31) = -2.88$, $p = .007$). The differences between USP and

333   ANOM-Pos nouns (diff = 0.05 µV; $t(31) = -0.12$, $p = .904$) and between EXP and ANOM-Impos

334   nouns (diff = 0.54 µV; t(31) = -1.52, p = .138) were not significant. Although ANOM-Pos (1.96

335   µV) and ANOM-Impos nouns (1.33 µV) differed by 0.64 µV, this difference also did not reach

336   significance (t(31) = 1.89, p = .068).

337   Whereas there were no differences in other temporal and/or spatial analysis windows,

338   ANOM-Pos and ANOM-Impos thus seem to have a different impact on late anterior positivity.

339   Given that different items were compared in the ANOM-Pos and ANOM-Impos conditions, we

340   ran a regression analysis to assess possible effects of critical word characteristics that may cause

341   amplitude differences between nouns. An amplitude calculation for each item averaged over

342   participants is inadequate for exploring factors in multiple regression designs, because it disregards

343   interparticipants' variability. Therefore, we used the method suggested by Lorch and Myers

344   (1990). For every participant, we extracted amplitudes of individual words and fitted a linear

345   regression with factors plausibility, possibility, word length, word frequency, and orthographic

346   neighborhood size (without interaction terms). For every predictor, the resulting 32 t-values

347   entered a one-sample t-test. Only the effect of word length on amplitude was significantly different

348   from zero (t(31) = 3.25, p = .003). See Supplemental Table S3 for detailed results. Note that the

349   effect of Possibility failed significance, and even the Plausibility effect, which entails a within-

350   item comparison, is much weaker in this analysis.

4 General discussion

352   In this study with German materials and participants, we investigated the

353   electrophysiological signatures of different types of contextual fit – from predictable and thus

354   expected, to highly implausible continuations of short discourses consisting of sentences pairs.

355   Whereas no reliable differences were present before 300 ms, graded N400 effects for target nouns

356   were observed as a function of their predictability and plausibility in the discourse. Relative to

357 highly predictable nouns, amplitudes were more negative for unexpected but plausible

358 continuations, and again more negative for implausible continuations. Whether or not the

359 implausible noun was a somewhat strange but in principle possible continuation of the preceding

360 discourse had no impact on the N400. In the time window following the N400, positivities with

361 different scalp signatures were observed that differed as a function of noun type. At posterior

362 electrode sites, the 600 – 1000 ms time window revealed similar amplitudes for highly predictable

363 and unpredictable but plausible continuations. Relative to these two continuations, there was

364 enhanced positivity for both implausible noun types – which showed very similar amplitudes. At

365 anterior sites, predictable and completely impossible continuations had similar amplitudes, but

366 predictable nouns had a less positive amplitude than both unpredictable plausible and implausible,

367 but still possible continuations, which did not differ.

368   In the following, we compare our outcomes to the original study (experiment 2) by DeLong

369 and colleagues (2014) that we aimed to replicate, and evaluate our results against the predictions

370 made for continuations that are quite implausible, but for which a real-world interpretation can be

371 constructed given the discourse. We discuss the distinction between anterior and posterior late

372 positivities and the potential processing functions that may underlie them.

373 **4.1 Replication of DeLong, Quante, and Kutas (2014)**

374   The data patterns relevant for the replication of DeLong et al. (2014), with three conditions

375 (EXP, USP and ANOM) in three time windows (N400, Anterior Positivity, Posterior Positivity),

376 show a striking similarity between the two studies, illustrated in Fig. 5.

377   The data for the N400 from the two studies show a very similar pattern – ignoring the

378 position of the zero line. Relative to expected continuations, negativity is enhanced for unexpected

379   but plausible (USP) nouns, and again more so for implausible (ANOM) nouns. The results for the

380   N400 thus fully replicate the graded negativity reported by DeLong et al. Note that the possibility

381   to create an interpretation for some implausible continuations had no effect on N400 amplitude,

382   since our two anomalous conditions did not differ.

383   The posterior positivity after the N400 also shows the same pattern as obtained by

384   DeLong and colleagues (2014), with significant differences between the expected (EXP) and

385   anomalous continuations (ANOM), between unexpected (USP) and anomalous words, but not

386   between EXP and USP, the expected and unexpected continuations. Again, exactly the same

387   pattern with the same significances was observed in both studies. Moreover, the analysis with four

388   noun types showed no difference between the possible and impossible anomalous continuations.

389   Finally, the anterior positivity again showed a similar pattern in both studies, but with somewhat

390   different significances. Whereas in both studies, amplitudes for expected (EXP) and unexpected

391   plausible (USP) nouns differ, and amplitudes for expected and anomalous nouns do not differ, the

392   difference between the unexpected plausible and anomalous nouns that was reliable in DeLong et

393   al. failed significance in our data. The analysis with four noun levels gives an indication why this

394   might be the case. In this analysis, the anomalous nouns that have a possible interpretation given

395   the preceding discourse do show a significant difference to the expected nouns, thus coinciding

396   with the unpredictable but plausible nouns. The difference to the impossible anomalous nouns

397   remains insignificant. Note however that the post-hoc regression analysis, which takes into account

398   between-item differences in the analyses of possibility effects, questions whether these differences

399   can be attributed to possibility.

400   Thus, with one interesting exception we closely replicate Experiment 2 by DeLong and

401   colleagues (2014), with German materials – mainly but not exclusively translated from DeLong et

402   al. (2014), with somewhat lower predictability of the predictable, expected nouns, and with

403   German native speakers. We believe this replication of a dissociation between anterior and

404   posterior positivity in largely overlapping, post-N400 time windows to be an important

405   contribution to the growing evidence for a functional difference associated with these two late

406   positivities. As DeLong and colleagues, and unlike other studies, we show this relatively new

407   dissociation with the same population within one experiment. In the following, we discuss our

408   findings relative to data, hypotheses and models proposed by others.

409   **4.2 N400 and late positivities**

410   The N400 effects show that relative to an anomalous noun, an unexpected noun that is

411   nevertheless a perfectly plausible continuation shows a smaller negativity. This graded negativity,

412   relative to the predicted continuation, replicates findings from many studies that show amplitude

413   negativity to depend on the degree of deviation from the condition that serves as reference (see

414   Kutas & Federmeier, 2011, for an overview).

415   The late positivity observed in our data seems to come in two guises. There is a bilateral,

416   posterior positivity that separates expected and unexpected but plausible continuations from

417   implausible, anomalous continuations – with no difference between those for which a possible,

418   real-world meaning (ANOM-Pos) can be constructed and those for which this is not the case

419   (ANOM-Impos). A second late positivity, with anterior, left-lateralized scalp distribution, seems

420   to distinguish between nouns for which an interpretation in the given discourse is possible but

421   unexpected (USP and ANOM-Pos nouns) on the one hand, and predictable words on the other.

422   **4.3 Posterior late positivity**

423   The posterior late positivity observed in our data resembles the P600 that has been

424   commonly associated with syntactic violations (e.g., Hagoort, Brown, & Groothusen, 1993) or

425   syntactic complexity (e.g., Friederici, Hahne, & Saddy, 2002; Kaan & Swaab, 2003). This changed

426   some fifteen years ago, when late positivities were reported for words that constituted thematic-

427   role violations (e.g., *At breakfast the eggs would eat*) which are semantic in nature (Kolk, Chwilla,

428   van Herten, & Oor, 2003; Kuperberg, Sitnikova, Caplan, & Holcomb, 2003; Hoeks, Stowe, &

429   Doedens, 2004; see Brouwer, Fitz, & Hoeks, 2012, for an overview). Such "semantic illusions"

430   had no impact on the N400 but showed in late positivities, with a central-posterior/parietal scalp

431   topography that resembles the "syntactic" P600. This "semantic" P600 again fired the debate on

432   its functional significance. Most proposed models and views adhere to two processing streams –

433   semantic and syntactic – whose outputs can conflict with each other, which is reflected in the P600

434   (cf. Kim & Osterhout, 2005; Kolk & Chwilla, 2007; Kuperberg, 2007; Bornkessel-Schlesewsky

435   & Schlesewsky, 2008; Hagoort, Baggio, & Willems, 2009; Kos, Vosse, van den Brinke, &

436   Hagoort, 2010; Metzner, Malsburg, Vasishth, & Rösler, 2017; see Brouwer et al., 2017, for an

437   overview and a different model).

438   In their seminal review of data from about 60 studies, Van Petten and Luka (2012) conclude

439   that posterior late positivity is associated with attempts at reanalysis when a problem is detected –

440   be it a syntactic or semantic incongruency or anomaly. Our late posterior positivities for all

441   anomalous continuations fit this picture. Kuperberg (2013) prominently put late positivities into

442   the perspective of prediction, suggesting that the posterior late positivities reflect processing costs

443   when the incoming word disconfirms predicted events or event structure. This is the case even for

444   semantic illusions (e.g., *The cat that from the mice fled*, incoming word underlined, Kolk et al.,

445   2003) in which the incoming words semantically fit the event, but their thematic roles violate event

446   structure. In our data, all anomalous continuations show a posterior negativity. Clearly, impossible

447   continuations violate event structure (often but not always because of selection restriction

448   violations): "excuse" is not a viable candidate for a snowman's nose. This is different for the

449   unexpected but plausible continuations: lacking a carrot, a banana can serve as a snowman's nose,

450   and is thus compatible with the event of snowman construction. Consequently, our expected and

451   unexpected but plausible words do not differ in late posterior positivity. Note that relative to these

452   two conditions, and in contrast to our prediction, a clear late posterior positivity was evident for

453   both types of anomalous continuations, those that are completely impossible and those for which

454   an admittedly strange meaning could be constructed (e.g., *To save space, she bought herself a pig*

455   *(expected: a loft bed) in the store.*). Following the logic by Van Petten and Luka, both anomalies

456   initiate the reprocessing of prior input, and in Kuperberg's view, both anomalies seem severe

457   enough to violate event structure.

458   **4.4 Anterior late positivity**

459   Finally, we consider the anterior post-N400 positivity observed in our data. In the overall

460   analysis, unexpected but plausible continuations (*the banana as nose for the snowman*) and

461   implausible but still in some way possible continuations (*the woman who bought herself a pig to*

462   *save space*) group together. First, they both differ from expected, highly predictable continuations

463   (*the carrot for a snowman's nose, a loft bed to save space*) and second, both continuations allow

464   for a revision of the discourse on the basis of the meaning of the unexpected words. Note that the

465   differences observed here may be due to item characteristics, as the regression analysis indicated.

466   Although these data, given that they involve different items, should be treated with caution, it is

467   interesting that similar late positivities with a (left) frontal scalp distribution have been observed

468   when words are not predicted but semantically possible, given the preceding context (Federmeier,

469   Wlotko, De Ochoa-Dewald, & Kutas, 2007; DeLong, Urbach, Groppe, & Kutas, 2011; Thornhill

470   & Van Petten, 2012; Van Petten & Luka, 2012; DeLong, Quante, & Kutas, 2014). As noted by

471   Van Petten and Luka, and as is the case in our data, frontal late positivities follow an N400 – which

472   is not always the case for posterior positivities. This co-occurrence is taken as an index for the

473   sensitivity of frontal positivities to semantic predictability.

474   The exact functional significance of anterior late positivity is still under debate. Most

475   researchers agree that it signals disconfirmed lexical prediction or lexical "prediction error" (Van

476   Petten & Luka, 2012; Kuperberg, 2013, for overviews), and that the presence of a moderately or

477   highly constraining context that can trigger updating is a prerequisite (Boudewyn et al., 2015).

478   Note that both requirements apply to the two conditions in which we observed late frontal

479   positivity. Taking these constraints as given, it remains unclear what processing costs occur after

480   disconfirmed prediction. Do they involve inhibition of the predicted word – a hypothesis

481   formulated with quite some foresight by Marta Kutas (1993), or are processing costs due to

482   revising and updating working memory to integrate the unexpected continuation (Federmeier et

483   al. 2007; Kuperberg, 2013)? In an ingenious study, Brothers, Swaab, and Traxler (2016) observed

484   late frontal positivity for words that were not predicted by their participants – who were told to

485   actively predict continuations of sentences and who indicated afterwards whether the continuation

486   presented was the one they predicted or not. With full, trial by trial control of prediction, Brothers

487   et al., could distinguish between specific lexical prediction and general contextual support – which

488   our design does not allow. Given that they also observed early (pre-N400) effects of prediction,

489   the authors conclude that the left-lateralized anterior positivity reflects prediction-related, post

490   lexical update and revision mechanisms. Given the importance of such mechanisms for prediction

491   in language, such anterior late positivities should be investigated further, with better control over

492   item characteristics as is the case in our study.

493   **4.5 Limitations**

494   It is important to point out that our materials were not explicitly constructed for the

495   distinction between ANOM-Impos and ANOM-Pos and that materials were not balanced (45 vs.

496   105 sentence pairs). As the regression analysis showed, items differed in length, which had an

497   impact on the late anterior positivities. As DeLong et al.'s (2014) sentence pairs were not rated for

498   possibility, it is not clear whether the minor discrepancies between the results of the two studies

499   arose from potential differences of the nouns in the ANOM conditions. We also should note that

500   overall contextual constraint was slightly lower in our study than in DeLong et al. which we aimed

501   to replicate. Still, despite their post-hoc flavor, our results on (im)possibility provide an interesting

502   perspective on the possibility of contextual integration of even quite implausible continuations - a

503   good reason to consider this dimension in the future.

504   **4.6 Conclusions**

505   With German materials and participants, we replicated results of DeLong and colleagues

506   (2014) and showed an impact of three types of constraint in sentence processing: predictability,

507   plausibility and possibility. We observed graded effects on the N400, with the smallest negativity

508   for expected continuations, followed by plausible but not expected alternatives, and with the

509   largest negativity for implausible, anomalous continuations. Next, despite both being unexpected,

510   plausible and implausible words show different patterns of posterior late positivity, arguing for a

511   dissociation of predictability and plausibility. Finally, we believe that the distinction between

512   possible and impossible continuations, both being implausible, should be taken into account in

513   studies on prediction and processing words in context.

514   **Acknowledgements**

515   We are deeply grateful to Katherine DeLong and Marta Kutas for their invaluable support,

516   and thank Dan Ke, Christian Bürger, René Michel and Daniel Kluger for their assistance in data

517   collection and analysis.

518

519   **References**

520   Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple

521   testing under dependency. *Annals of Statistics*, 1165-1188.

522   Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2008). An alternative perspective on

523   "semantic P600" effects in language comprehension. *Brain Research Reviews*, *59*(1), 55-73.

524   doi:10.1016/j.brainresrev.2008.05.003

525   Boudewyn, M. A., Long, D. L., & Swaab, T. Y. (2015). Graded expectations: Predictive

526   processing and the adjustment of expectations during spoken language comprehension. *Cognitive,*

527   *Affective, & Behavioral Neuroscience*, *15*(3), 607-624. doi:10.3758/s13415-015-0340-0

528   Brothers, T., Swaab, T. Y., & Traxler, M. J. (2015). Effects of prediction and contextual

529   support on lexical processing: Prediction takes precedence. *Cognition*, *136*, 135-149.

530   doi:10.1016/j.cognition.2014.10.017

531   Brouwer, H., Crocker, M. W., Venhuizen, N. J., & Hoeks, J. C. (2017). A

532   Neurocomputational Model of the N400 and the P600 in Language Processing. *Cognitive Science*,

533   *41*(S6), 1318-1352. doi:10.1111/cogs.12461

534   Brouwer, H., Fitz, H., & Hoeks, J. (2012). Getting real about semantic illusions: rethinking

535   the functional role of the P600 in language comprehension. *Brain Research*, *1446*, 127-143.

536   doi:10.1016/j.brainres.2012.01.055

537   Coltheart, M., Davelaar, E. J., Jonasson, J. T., & Besner, D. (1977). Access to the Internal

538   Lexicon. In S. Dornic (Ed.), *Attention and Performance VI. Proceedings of the Sixth International*

539   *Symposium on Attention and Performance, Stockholm, Sweden, July 28-August 1, 1975* (pp. 535-

540   555). Hillsdale, NJ: Lawrence Erlbaum Associates.

541   DeLong, K. A., Quante, L., & Kutas, M. (2014). Predictability, plausibility, and two late

542   ERP positivities during written sentence comprehension. *Neuropsychologia, 61,* 150-162.

543   doi:10.1016/j.neuropsychologia.2014.06.016

544   DeLong, K. A., Urbach, T. P., Groppe, D. M., & Kutas, M. (2011). Overlapping dual ERP

545   responses to low cloze probability sentence continuations. *Psychophysiology, 48*(9), 1203-1207.

546   doi:10.1111/j.1469-8986.2011.01199.x

547   DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during

548   language comprehension inferred from electrical brain activity. *Nature Neuroscience, 8*(8), 1117-

549   1121. doi:10.1038/nn1504

550   Dennis, A. R., & Valacich, J. S. (2014). A replication manifesto. *AIS Transactions on

551   Replication Research*, *1*(1), 1. doi:10.17705/1atrr.00001

552   Dikker, S., & Pylkkänen, L. (2011). Before the N400: Effects of lexical-semantic violations

553   in visual cortex. *Brain and Language, 118*(1), 23-28. doi:10.1016/.bandl.2011.02.006

554   Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple

555   effects of sentential constraint on word processing. *Brain Research*, *1146*, 75-84.

556   doi:10.1016/j.brainres.2006.06.101

557   Forster, K. I. (198l). Priming and the effects of sentence and lexical contexts on naming

558   time: Evidence for autonomous lexical processing. *The Quarterly Journal of Experimental

559   Psychology, 33*(4), 465-496. doi:10.1080/14640748108400804

560   Friederici, A. D., Hahne, A., & Saddy, D. (2002). Distinct neurophysiological patterns

561   reflecting aspects of syntactic complexity and syntactic repair. *Journal of Psycholinguistic*

562   *Research*, *31*(1), 45-63.

563   Geyer, A., Holcomb, P. J., Kuperberg, G. R., & Perlmutter, N. (2006). Plausibility and

564   sentence comprehension. An ERP study. *Cogn. Neurosci. Suppl., Abstract*, 1-1.

565   Hagoort, P., Baggio, G., & Willems, R. M. (2009). Semantic unification. In *The Cognitive*

566   *Neurosciences, 4th ed.* (pp. 819-836). MIT press.

567   Hagoort, P., Brown, C., & Groothusen, J. (1993). The syntactic positive shift (SPS) as an

568   ERP measure of syntactic processing. *Language and Cognitive Processes*, *8*(4), 439-483.

569   doi:10.1080/01690969308407585

570   Hoeks, J. C., Stowe, L. A., & Doedens, G. (2004). Seeing words in context: the interaction

571   of lexical and sentence level information during reading. *Cognitive Brain Research*, *19*(1), 59-73.

572   doi:10.1016/j.cogbrainres.2003.10.022

573   Huettig, F., & Janse, E. (2016). Individual differences in working memory and processing

574   speed predict anticipatory spoken language processing in the visual world. *Language, Cognition*

575   *and Neuroscience*, *31*(1), 80-93. doi:10.1080/23273798.2015.1047459

576   Ille, N., Berg, P., & Scherg, M. (2002). Artifact correction of the ongoing EEG using spatial

577   filters based on artifact and brain signal topographies. *Journal of Clinical Neurophysiology, 19*(2),

578   113-124. doi:10.1097/00004691-200203000-00002

579   Ito, A., Corley, M., Pickering, M. J., Martin, A. E., & Nieuwland, M. S. (2016). Predicting

580   form and meaning: Evidence from brain potentials. *Journal of Memory and Language*, *86*, 157-

581   171. doi:10.1016/j.jml.2015.10.007

582   Jasper, H. H. (1958). Report of the committee on methods of clinical examination in

583   electroencephalography: 1957. *Electroencephalography and Clinical Neurophysiology, 10*(2),

584   370-375. doi:10.1016/0013-4694(58)90053-1

585   Kaan, E., & Swaab, T. Y. (2003). Repair, revision, and complexity in syntactic analysis:

586   An electrophysiological differentiation. *Journal of Cognitive Neuroscience*, *15*(1), 98-110.

587   doi:10.1162/089892903321107855

588   Kim, A., & Osterhout, L. (2005). The independence of combinatory semantic processing:

589   Evidence from event-related potentials. *Journal of Memory and Language, 52*(2), 205-225.

590   doi:10.1016/j.jml.2004.10.002

591   Kolk, H. H. J., & Chwilla, D. J. (2007). Late positivities in unusual situations. *Brain and*

592   *Language*, *100*(3), 257-261. doi:10.1016/j.bandl.2006.07.006

593   Kolk, H. H. J., Chwilla, D. J., Van Herten, M., & Oor, P. J. (2003). Structure and limited

594   capacity in verbal working memory: A study with event-related potentials. *Brain and Language*,

595   *85*(1), 1-36. doi:10.1016/s0093-934x(02)00548-5

596   Kos, M., Vosse, T., Van Den Brink, D., & Hagoort, P. (2010). About edible restaurants:

597   conflicts between syntax and semantics as revealed by ERPs. *Frontiers in Psychology*, *1*.

598   doi:10.3389/fpsyg.2010.00222

599   Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to

600   syntax. *Brain Research, 1146*, 23-49. doi:10.1016/j.brainres.2006.12.063

601   Kuperberg, G. R. (2013). The proactive comprehender: What event-related potentials tell

602   us about the dynamics of reading comprehension. In: *Unraveling the behavioral, neurobiological,*

603   *and genetic components of reading comprehension*. Miller, B., Cutting, L., & McCardle, P. (Eds):

604   Baltimore: Paul Brookes Publishing.

605   Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language

606   comprehension?   *Language,   Cognition   and   Neuroscience,   31*(1),   32-59.

607   doi:10.1080/23273798.2015.1102299

608   Kuperberg, G. R., Sitnikova, T., Caplan, D., & Holcomb, P. J. (2003). Electrophysiological

609   distinctions in processing conceptual relationships within simple sentences. *Cognitive Brain*

610   *Research, 17*(1), 117-129. doi:10.1016/s0926-6410(03)00086-7

611   Kutas, M. (1993). In the company of other words: Electrophysiological evidence for single-

612   word and sentence context effects. *Language and Cognitive Processes, 8*(4), 533-572.

613   doi:10.1080/01690969308407587

614   Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use

615   in language comprehension. *Trends in Cognitive Sciences, 4*(12), 463-470. doi:10.1016/S1364-

616   6613(00)01560-6

617   Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the

618   N400 component of the event related brain potential (ERP). *Annual Review of Psychology, 62*,

619   621-647. doi:10.1016/s1364-6613(00)01560-6

620   Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect

621   semantic incongruity. *Science, 207*(4427), 203-205. doi:10.1126/science.7350657

622   Lau, E. F., Holcomb, P. J., & Kuperberg, G. R. (2013). Dissociating N400 effects of

623   prediction from association in single-word contexts. *Journal of Cognitive Neuroscience, 25*(3),

624   484-502. doi:10.1162/jocn_a_00328

625   Lorch, R. F., & Myers, J. L. (1990). Regression analyses of repeated measures data in

626   cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition*,

627   16(1), 149–157. doi:10.1037/0278-7393.16.1.149

628   Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition.

629   *Cognition*, *25*(1), 71-102. doi:10.1016/0010-0277(87)90005-9

630   McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context

631   effects in letter perception: I. An account of basic findings. *Psychological Review*, *88*(5), 375.

632   doi:10.1037//0033-295x.88.5.375

633   Metzner, P., Malsburg, T., Vasishth, S., & Rösler, F. (2017). The importance of reading

634   naturally: Evidence from combined recordings of eye movements and electric brain potentials.

635   *Cognitive Science*, *41*(S6), 1232-1263. doi:10.1111/cogs.12384

636   Morey, R. D. (2008). Confidence intervals from normalized data: A correction to

637   Cousineau (2005). *Reason, 4*(2), 61-64. doi:10.20982/tqmp.04.2.p061

638   Nieuwland, M., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., ...

639   & Mézière, D. (2017). Limits on prediction in language comprehension: A multi-lab failure to

640   replicate evidence for probabilistic pre-activation of phonology. *bioRxiv*, 111807.

641   doi:10.1101/111807

642   Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh

643   inventory. *Neuropsychologia, 9*(1), 97-113. doi:10.1016/0028-3932(71)90067-4

644   Paczynski, M., Kreher, D. A., Ditman, T., Holcomb, P., & Kuperberg, G. R. (2006).

645   Electrophysiological evidence for the role of animacy and lexico-semantic associations in

646   processing nouns within passive structures. *Cogn. Neurosci. Suppl., Abstract*.

647   Paczynski, M., & Kuperberg, G. R. (2012). Multiple influences of semantic memory on

648   sentence processing: Distinct effects of semantic relatedness on violations of real-world event/state

649   knowledge and animacy selection restrictions. *Journal of Memory and Language*, *67*(4), 426-448.

650   doi:10.1016/j.jml.2012.07.003

651   Rayner, K., Warren, T., Juhasz, B. J., & Liversedge, S. P. (2004). The effect of plausibility

652   on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and*

653   *Cognition*, *30*(6), 1290.

654   Schwanenflugel, P. J., & Shoben, E. J. (1985). The influence of sentence constraint on the

655   scope of facilitation for upcoming words. *Journal of Memory and Language*, *24*(2), 232-252.

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

657   Staub, A. (2015). The effect of lexical predictability on eye movements in reading: Critical

658   review and theoretical interpretation. *Language and Linguistics Compass, 9*(8), 311-327.

659   doi:10.1111/lnc3.12151

660   Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re)consideration

661   of context effects. *Journal of Verbal Learning and Verbal Behavior*, *18*(6), 645-659.

662   doi:10.1016/s0022-5371(79)90355-4

663   Szewczyk, J. M., & Schriefers, H. (2013). Prediction in language comprehension beyond

664   specific words: An ERP study on sentence comprehension in Polish. *Journal of Memory and*

665   *Language, 68*(4), 297-314. doi:10.1016/j.jml.2012.12.002

666   Thornhill, D. E., & Van Petten, C. (2012). Lexical versus conceptual anticipation during

667   sentence processing: Frontal positivity and N400 ERP components. *International Journal of*

668   *Psychophysiology, 83*(3), 382-392. doi:10.1016/j.ijpsycho.2011.12.007

669   van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005).

670   Anticipating upcoming words in discourse: evidence from ERPs and reading times. *Journal of*

671   *Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 443. doi:10.1037/0278-

672   7393.31.3.443

673   van Berkum, J. J. A., Zwitserlood, P., Hagoort, P., & Brown, C. M. (2003). When and how

674   do listeners relate a sentence to the wider discourse? Evidence from the N400 effect. *Cognitive*

675   *Brain Research, 17*(3), 701-718. doi:10.1016/s0926-6410(03)00196-4

676   Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits,

677   costs, and ERP components. *International Journal of Psychophysiology, 83*(2), 176-190.

678   doi:10.1016/j.ijpsycho.2011.09.015

679   Warren, T., & McConnell, K. (2007). Investigating effects of selectional restriction

680   violations and plausibility violation severity on eye-movements in reading. *Psychonomic Bulletin*

681   *& Review*, *14*(4), 770-775. doi:10.3758/bf03196835

682   Zwitserlood, P. (1989). The locus of the effects of sentential-semantic context in spoken-

683   word processing. *Cognition*, *32*(1), 25-64. doi:10.1016/0010-0277(89)90013-9

684

685

686   **Notes**

ⁱ 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).

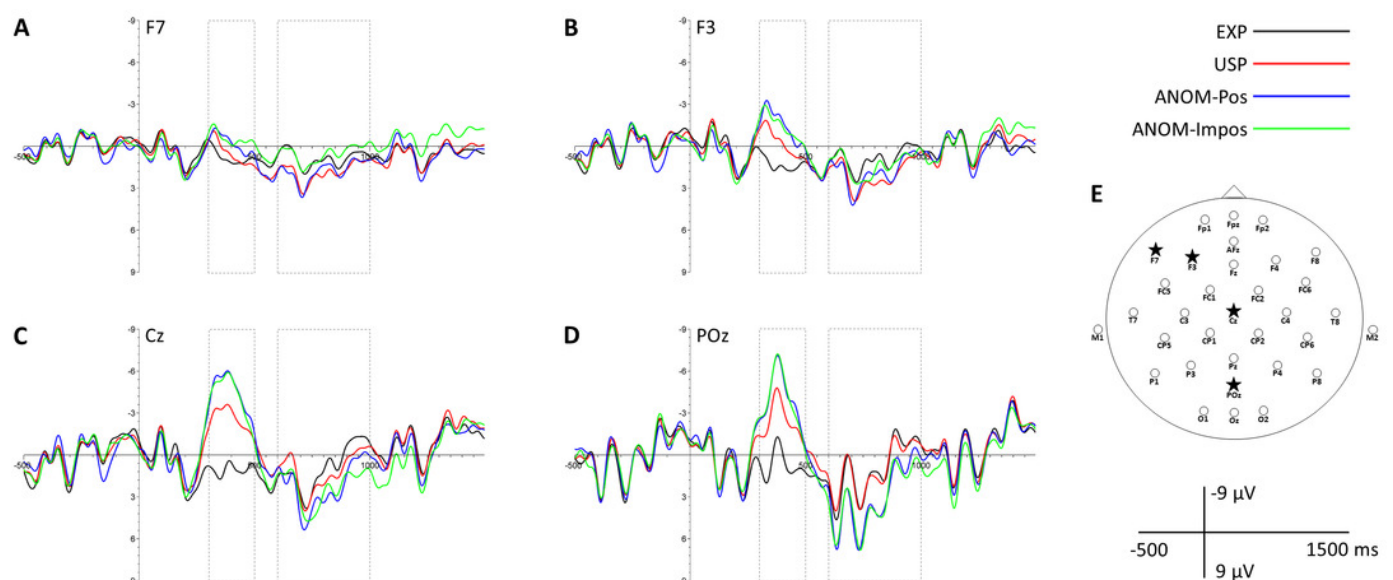
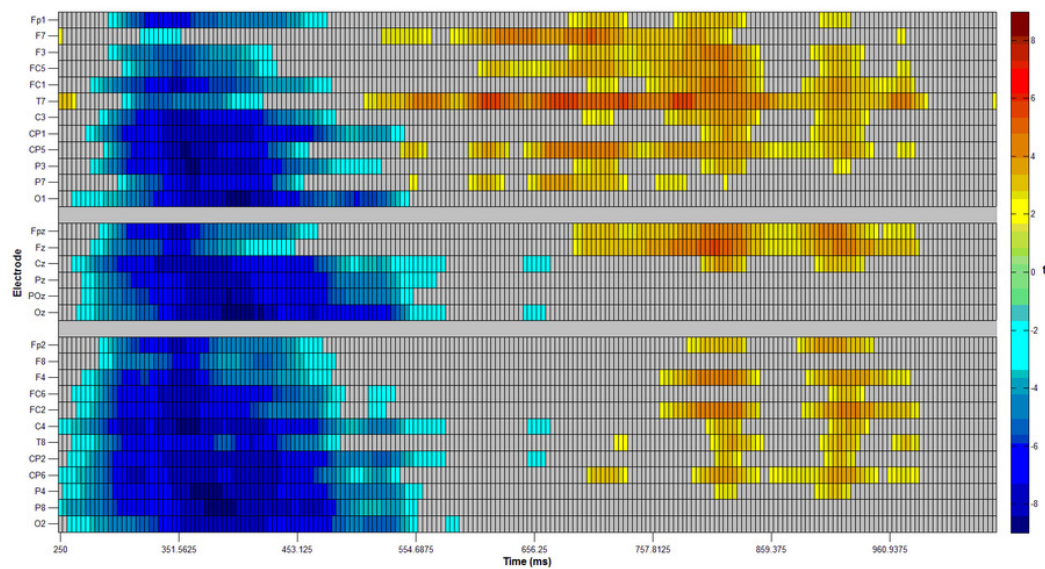
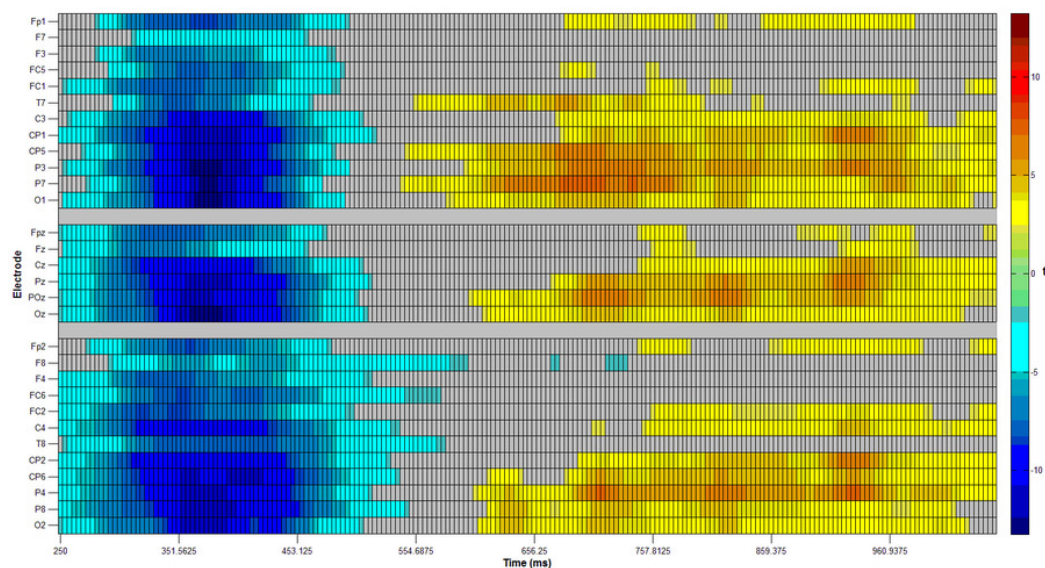



Figure 2

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.

A USP - EXP

B ANOM - EXP

C ANOM - USP

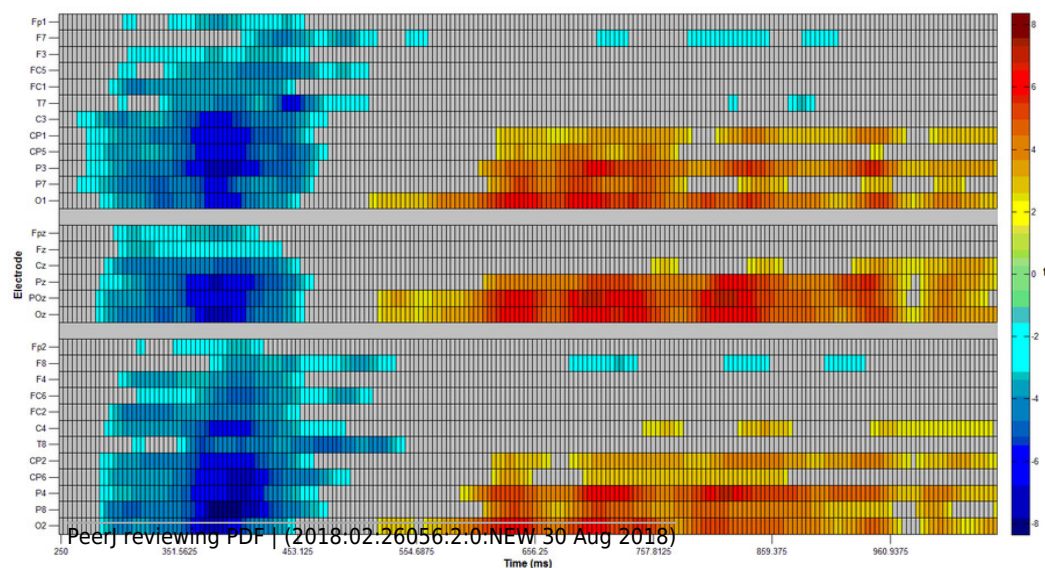



Figure 3

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

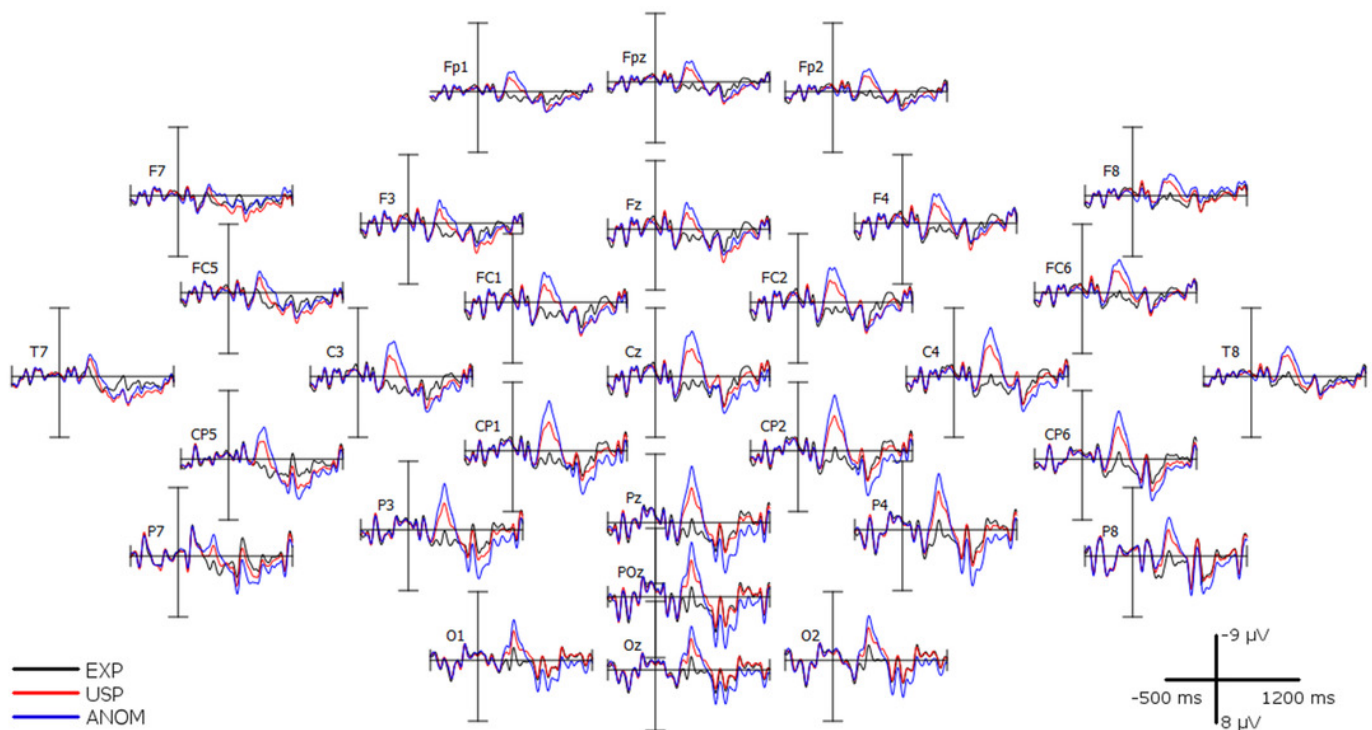



Figure 4

Topographic scalp maps.

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

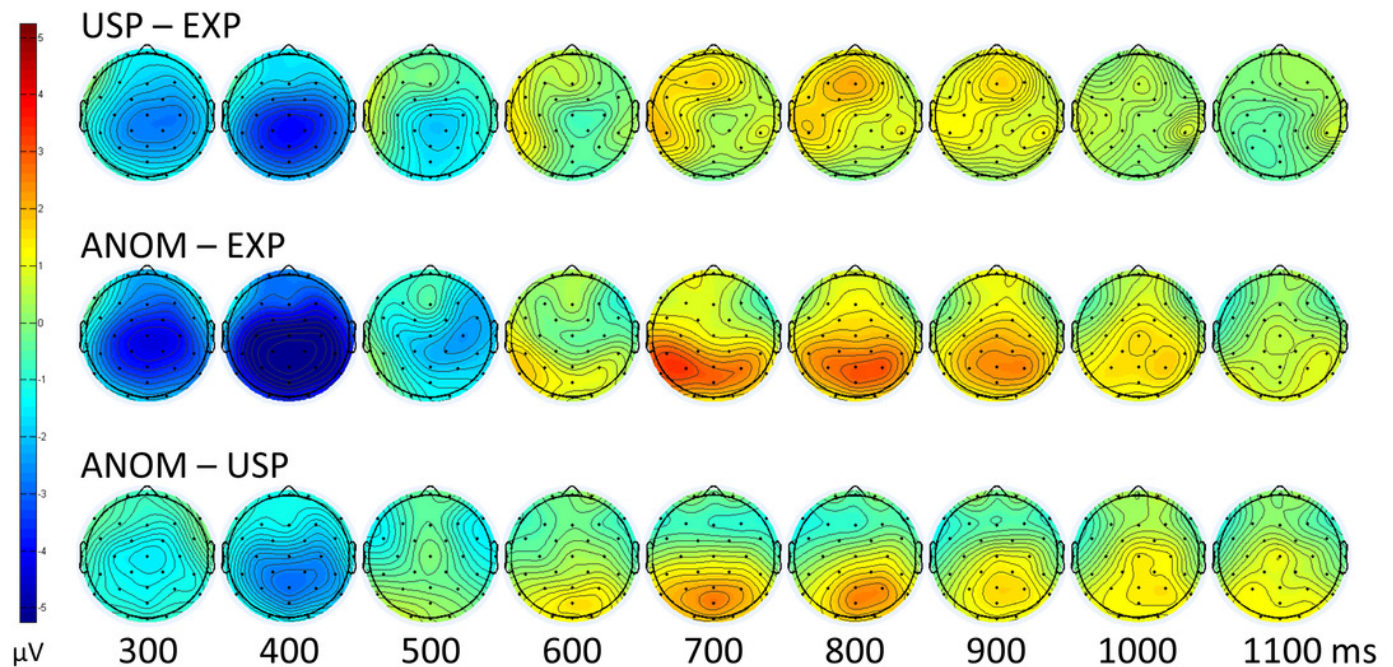
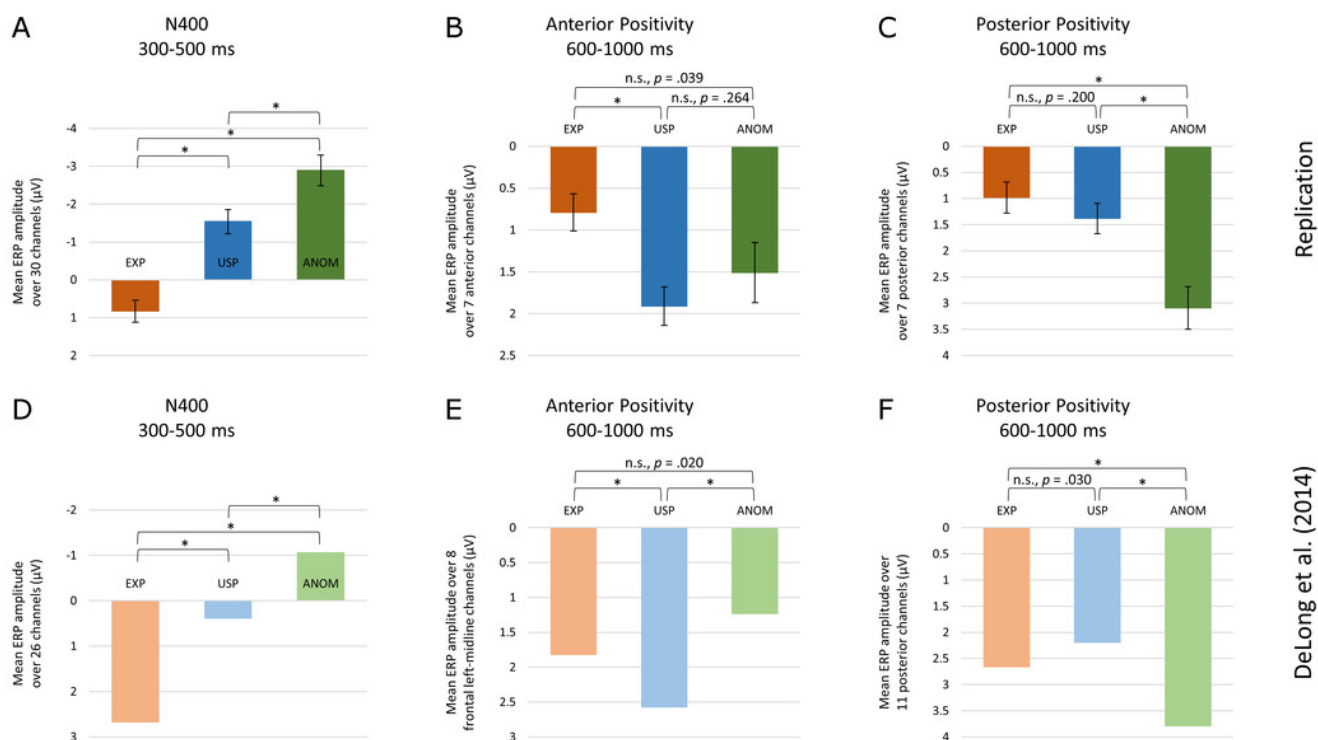



Figure 5

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 ($p_{boncor} < .0167$).

Sample sentence pairs

EXP, USP, ANOM-Impos

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 röntge 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.)*

EXP, USP, ANOM-Pos

4. Luisas neues WG-Zimmer war sehr klein, hatte aber hohe Decken. Um Platz zu sparen, kaufte 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.)*

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

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.)*

FILL

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.)*

Stimuli characteristics

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)[a]	Mean critical noun length (SD)	Mean critical orthographic neighborhood size (SD)[b]
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)	Low cloze/High plausibility	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/Low 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)

[a] Absolute annotated type frequencies according to dlexDB (http://www.dlexdb.de/)
[b] Orthographic neighborhood size (as defined by Coltheart, Davelaar, Jonasson, & Besner, 1977) according to dlexDB

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	-
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	-
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	-
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) = 0.34, *p* = .738

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.

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 (1.63)	-1.54 (1.82)	-2.89 (2.32)	-3.01 (2.41)	-2.82 (2.43)
Anterior Positivity	0.79 (1.24)	1.91 (1.32)	1.51 (2.05)	1.96 (2.19)	1.33 (2.27)
Posterior Positivity	0.98 (1.68)	1.38 (1.66)	3.09 (2.32)	2.95 (2.42)	3.16 (2.46)

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 ($p_{boncor} < .0083$)