Large-scale replication study reveals a limit on probabilistic prediction in language comprehension

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

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

[2] Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Kowloon, Hong Kong

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

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

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

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

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

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

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

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

[11] Psychology, Faculty of Natural Sciences, University of Stirling, Stirling, United Kingdom
[12] Department of Psychology, University of York, York, United Kingdom

Corresponding author:
Mante S. Nieuwland, Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD Nijmegen, The Netherlands. E-mail: mante.nieuwland@mpi.nl, phone: +31-24-3521911

ABSTRACT

In current theories of language comprehension, people routinely and implicitly predict upcoming words by pre-activating their meaning, morpho-syntactic features and even their specific phonological form. To date the strongest evidence for phonological prediction comes from a landmark 2005 Nature Neuroscience publication by DeLong, Urbach and Kutas, who observed a graded modulation of electrical brain potentials (N400) to nouns and preceding articles by the probability that people use a word to continue the sentence fragment ('cloze'). In a direct replication study spanning 9 laboratories (N=334), we failed to replicate the crucial article-elicited N400 modulation by cloze, while we successfully replicated the commonly-reported noun-elicited N400 modulation. We observed this pattern of failure and success in a pre-registered replication analysis, a pre-registered single-trial analysis, and exploratory Bayesian analyses. Contra the strong prediction view in which people routinely pre-activate the phonological word-form, our results suggest a more limited role for prediction during language comprehension.

INTRODUCTION

In the last decades, the idea that people routinely and implicitly predict upcoming words during language comprehension turned from a highly controversial hypothesis to a widely accepted assumption. Initial objections to prediction in language were based on a lack of empirical support (e.g., Zwitserlood, 1989), incompatibility with traditional bottom-up models and contemporary interactive models of language comprehension (e.g., Kintsch, 1988; Marslen-Wilson & Tyler, 1988), and the purported futility of prediction in a generative system where sentences can continue in infinitely many different ways (Jackendoff, 2002). Current theories of language comprehension, however, reject such objections and posit prediction as an integral and inevitable mechanism by which comprehension proceeds quickly and incrementally (e.g., Altmann & Mirkovic, 2009; Dell & Chang, 2014; Pickering & Garrod, 2013). Prediction, i.e., context-based pre-activation of an upcoming linguistic input, is thought to occur at all levels of linguistic representation (semantic, morpho-syntactic and phonological/orthographic) and serves to facilitate the integration of newly available bottom-up information into the unfolding sentence- or discourse-representation. In this line of thought, language is yet another domain in which the brain acts as a prediction machine (Clark, 2013; Van Berkum, 2010; see also Friston, 2005, 2010; Summerfield & De Lange, 2014), hard-wired to continuously match sensory inputs with top-down, grammatical or probabilistic expectations based on context and memory.

What promoted linguistic prediction from outlandish and deeply contentious to ubiquitous and somewhat anodyne? One of the key and most compelling pieces of empirical evidence for linguistic prediction to date comes from a landmark *Nature Neuroscience* publication by DeLong, Urbach and Kutas (2005), whose approach exploited a phonological rule of English whereby the indefinite article is realized as *a* before consonant-initial words and as *an* before vowel-initial words. In their experiment, participants read sentences of varying degree of contextual constraint that led to expectations for a particular consonant- or vowel-initial noun. This expectation was operationalized as a word's cloze probability (cloze), calculated in a separate, non-speeded sentence completion task as the percentage of continuations of a sentence fragment with that word (Taylor, 1953). For example, the sentence fragment "The day was breezy so the boy went outside to fly..." is continued with 'a' by 86% of participants, and ''The day was breezy so the boy went outside to fly a...", it is continued with 'kite' by 89% of participants. In the main experiment, word-by-word sentence presentation enabled DeLong and colleagues to examine electrical brain activity elicited by articles that were concordant with the highly expected but yet unseen noun ('a', followed by 'kite'), or by articles that were incompatible with the highly expected noun and heralded a less expected one ('an', followed by 'airplane'). The dependent measure was the amplitude of the N400[1] event-related potential (ERP), a negative ERP deflection that peaks approximately 400 ms after word onset and is maximal at centroparietal electrodes (Kutas & Hillyard, 1980). The N400 is elicited by every word of an unfolding sentence and its amplitude is smaller (less negative) with increasing ease of semantic processing (Kutas & Hillyard, 1984). DeLong et al. found that the N400 amplitude for a given word decreased as a function of increasing cloze probability, both for nouns and, critically, for articles. The systematic, graded N400 modulation by article-cloze was taken as strong evidence that participants activated the nouns in advance of their appearance, and that the disconfirmation of this

[1] In this article, we use "N400 amplitude" as a shorthand for "ERP amplitude in the time window associated with the N400"; this ERP amplitude is actually a sum of the N400 ERP component and other ERP components (reflecting other aspects of cognition) that overlap with it in time and space.

prediction by the less-expected articles resulted in processing difficulty (higher N400 amplitude at the article).

The results obtained with this elegant design warranted a much stronger conclusion than related results available at the time. Previous studies that employed a visual-world paradigm had revealed listeners' anticipatory eye-movements towards visual objects on the basis of probabilistic or grammatical considerations (e.g., Altmann & Kamide, 1999). However, predictions in such studies are scaffolded onto already-available visual context, and therefore do not measure purely pre-activation, but perhaps re-activation of word information previously activated by the visual object itself (Huettig, 2015). DeLong and colleagues examined brain responses to information associated with concepts that were not pre-specified and had to be retrieved from long-term memory 'on-the-fly'. Furthermore, DeLong and colleagues were the first to muster evidence for highly specific pre-activation of a word's phonological form, rather than merely its semantic (e.g., Federmeier & Kutas, 1999) or morpho-syntactic features (e.g., Van Berkum, Brown, Zwitserlood, Kooijman & Hagoort, 2005; Wicha, Moreno & Kutas, 2004). Crucially, as their demonstration involved semantically identical articles (function words) rather than nouns or adjectives (content words) that are rich in meaning, the observed N400 modulation by article-cloze is unlikely to reflect difficulty interpreting the articles themselves. Most notably, DeLong and colleagues were the first to examine brain activity elicited by a range of more- or less-predictable articles, not simply most- versus least-expected. Based on the observed correlation, they argued that pre-activation is not all-or-none and limited to highly constraining contexts, but occurs in a graded, probabilistic fashion, with the strength of a word pre-activation proportional to its cloze probability. Moreover, they concluded that prediction is an integral part of real-time language processing and, most likely, a mechanism for propelling the comprehension system to keep up with the rapid pace of natural language.

DeLong et al.'s study has had an immense impact on psycholinguistics, neurolinguistics and beyond. It is cited by authoritative reviews (e.g., Altmann & Mirkovic, 2009; Hagoort, 2017; Lau, Phillips & Poeppel, 2008; Pickering & Clark, 2014; Pickering & Garrod, 2007) as delivering decisive evidence for probabilistic prediction of words all way up to their phonological form. Moreover, as a demonstration of pre-activation of phonological form (sound) during reading, it is often cited as evidence for 'prediction through production' (e.g., Pickering & Garrod, 2013), the hypothesis that linguistic predictions are implicitly generated by the language production system. To date, DeLong et al. has received a total of 735 citations (Google Scholar), averaging to more than 1 citation per week over the past decade, with an increasing number of citations in each subsequent year. The results also played an important role in settling an ongoing debate in the neuroscience of language. It provided the clearest evidence that the N400 component, which for 25 years had been taken to directly index the high-level compositional processes by which people integrate a word's meaning with its context (Brown & Hagoort, 1993; Chwilla, Brown & Hagoort, 1995; Connolly & Phillips, 1994; Friederici, Steinhauer & Frisch, 1999; Van Berkum, Hagoort & Brown, 1999; Van Petten, Coulson, Rubin, Plante, & Parks, 1999), actually reflected non-compositional processes by which word information is accessed as a function of context.

But how robust are gradient effects of form prediction? In over a decade that has passed since the publication by DeLong and colleagues, there is still no published study that directly replicates their graded pattern of results (for an overview, see Ito, Martin & Nieuwland, 2017b). An alternative analysis of the same data by the authors did not yield a statistically significant result (DeLong, 2009), but was not mentioned in the published report. In at least three other unpublished data sets (DeLong, 2009; Miyamoto, 2016), DeLong and colleagues did not find a significant correlation between article-N400 and cloze probability. Martin, Thierry, Kuipers, Boutonnet, Foucart and Costa (2013) reported a successful conceptual replication in native speakers of English but not in bilinguals. However, their study did not test for a graded effect of cloze, and differed from the original in many crucial aspects of the experimental design, data-preprocessing and statistical analysis, clouding both a qualitative and quantitative comparison to the original results. Moreover, two attempts to replicate the Martin et al. results in English monolinguals failed to yield a reliable effect of cloze on article-ERPs (Ito, Martin & Nieuwland, 2017a,b).

As the tremendous scientific impact of the DeLong et al. findings is at odds with the apparent lack of replication attempts, we report here a direct replication study. Inspired by recent demonstrations for the need for large subject-samples in psychology and neuroscience research (Button et al., 2013; Open Science Collaboration, 2015), our replication spanned 9 laboratories each with a sample size equal to or greater than that of the original. In addition to duplicating the original analysis, our replication attempt also seeks to improve upon DeLong et al.'s data analysis. DeLong et al.'s original analysis reduced an initial pool of 2560 data points (32 subjects who each read 80 sentences) to 10 grand-average values, by averaging N400 responses over trials within 10 cloze probability decile-bins (cloze 0-10, 11-20, et cetera), per participant and then averaging over participants, even though these bins held greatly different numbers of observations (for example, the 0-10 cloze bin contained 37.5% of all data). These 10 values were correlated with the average cloze value per bin, yielding numerically high correlation coefficients with large confidence intervals (for example, the Cz electrode showed a statistically significant r-value of 0.68 with a 95% confidence interval ranging from 0.09 to 0.92). However, this analysis potentially compromises power by discretizing cloze probability into deciles and not distinguishing various sources of subject-, item-, bin-, and trial-level variation. Furthermore, treating subjects as fixed rather than random factor potentially inflates false positive rates, since the overall cloze effect is confounded with by-subject variation in the effect (Barr, Levy, Scheepers & Tily, 2013; Clark, 1973).

In our replication study, we followed two pre-registered analysis routes: a replication analysis that duplicated the DeLong et al. analysis, and a single-trial analysis that modelled variance at the level of item and subject. The effect of cloze on noun-elicited N400s (DeLong et al., 2015; Kutas & Hillyard, 1984) is a necessary but not sufficient evidence for the claim on pre-activation in language processing (as it is also compatible with the view that the noun's cloze probability correlates with the ease of integration of that noun into the context). It serves as a manipulation check to ensure that the experiment is able to successfully detect graded variation in N400 amplitude, but does not provide strong evidence for the prediction of phonological form. That evidence would come from the ERPs elicited by articles. Observing a reliable effect of cloze on article-elicited N400s in the replication analysis and, in particular, in the single-trial analysis, would constitute powerful evidence for the pre-activation of phonological form during reading.

RESULTS

We first obtained offline cloze probabilities for all target articles and nouns from a group of native English speakers. These values closely resembled those of the original study (see Methods for details). In the subsequent ERP experiment, a different group of participants (N=334) read the sentences word-by-word from a computer display at a rate of 2 words per second while we recorded their electrical brain activity at the scalp. The replication analysis and single-trial analysis described below were each pre-registered at https://osf.io/eyzaq/.

Replication analysis

We sorted the articles and nouns into 10 bins based on each word's cloze probability (e.g., items with 0-10% cloze were put in one bin, 10-20% in another, etc.). For each laboratory, we averaged ERPs per bin first within, then across, participants. No baseline correction was used, following the procedure described in the Methods section in DeLong et al (2005). We then correlated the averaged cloze values per bin with mean ERP amplitude in the N400 time window (200-500 ms) elicited by the nouns (for the noun analysis) or articles (for the article analysis)from the corresponding bin, yielding a Pearson correlation coefficient (r-value) per EEG channel. This analysis yielded a very different pattern than DeLong et al. observed (Fig. 1). In no laboratory did article-N400 amplitude at centro-parietal sites become significantly smaller (less negative) as article-cloze probability increased (in fact, in most laboratories the pattern went into the opposite direction). Only in one laboratory (Lab 2) did the correlation coefficient have a p-value below .05 in the predicted direction (positive) at any electrode (uncorrected for multiple comparisons), but this effect was observed at a few left-frontal electrodes, not at the central-parietal electrodes where DeLong et al found their N400 effects. Moreover, in two laboratories (Labs 3 and 5), a statistically significant effect was observed in the opposite direction, larger (more negative) article-N400 amplitude for articles with increasing cloze probability. For the nouns, the pattern was more similar to the DeLong et al. results. In six laboratories (Lab 2, 3, 4,6, 7, and 9), noun-N400 amplitude for nouns at central-parietal or parietal-occipital electrodes became smaller with increasing noun-cloze, and in two other laboratories (Lab 5 and 8) the effects clearly went in the expected direction without reaching statistical significance.

DeLong et al. recently mentioned using a 500 ms baseline correction procedure that was not mentioned in the published study (personal communication by DeLong, March 2017). In an exploratory analysis, we therefore recomputed the correlations based on data pooled from all laboratories using this baseline correction procedure (Fig 2.). This analysis also showed a lack of statistically significant positive correlations for the articles, but statistically significant positive correlations for the nouns. In exploratory Bayesian analyses reported below, we perform an analysis to establish whether these results are consistent with the size and direction of the effects reported by DeLong et al., regardless of statistical significance.

Single-trial analysis

We first performed baseline correction by subtracting the average amplitude in the 100 ms time window before word onset. Baseline-corrected ERPs for relatively expected and unexpected words and difference waveforms are shown in Fig. 3. Then, for the data pooled across all laboratories, we used linear mixed effects models to regress the N400 amplitude (in a spatiotemporal region of interest selected a priori based on the DeLong et al. results) on cloze probability. For the articles, the effect of cloze was not statistically significant at the α=.05 level, β = .29, CI [-.08, .67], $\chi^2(1) = 2.31$, p = .13 (see Fig. 4, left panel) [2], with β referring to the N400 difference in microvolts associated with stepping from 0% to 100% cloze. The effect of cloze on N400 amplitude at the article did not significantly differ between laboratories, $\chi^2(8) = 7.90$, p = .44. For the nouns, however, higher cloze values were strongly associated with smaller N400s, β = 2.22, CI [1.76, 2.69], $\chi^2(1) = 56.50$, $p < .001$ (see Fig. 4, right panel). This pattern did not significantly differ between laboratories, $\chi^2(8) = 11.59$, p = .17. The effect of cloze on noun-N400s was statistically different from its effect on article-N400s, $\chi^2(1) = 31.38$, $p < .001$.

Exploratory (i.e., not pre-registered) single-trial analyses

The effect of article-cloze did not significantly vary as a function of subject comprehension question accuracy, $\chi^2(1) = 0.45$, p = .50. In addition, the effect of article-cloze

[2] Unless otherwise indicated, p-values are two-tailed, and CIs are two-tailed 95% confidence intervals.

was also not statistically significant when subject comprehension accuracy was included in the analysis (100 ms baseline: β = .24, CI [-.17, .64], $\chi^2(1) = 1.27$, p = .26).

In our dataset, an analysis in the 500 to 100 ms time window before article-onset revealed a non-significant effect of cloze that resembled the pattern observed after article-onset, β = .16, CI [-.07, .39], $\chi^2(1) = 1.82$, p = .18. This suggested that a 500 ms baseline correction procedure, which was used but not reported in DeLong et al. (2005), would better correct for pre-article voltage-levels. We repeated our analysis with the 500 ms baseline correction procedure, the initially observed effect of article-cloze was numerically smaller and less significant than it was in the pre-registered analysis (β = .14, CI [-.25, .53], $\chi^2(1) = 0.46$, p = .50).

Exploratory Bayesian analyses

For the articles, our pre-registered replication analyses yielded non-significant p-values, indicating failure to reject the null-hypothesis that cloze has no effect on N400 activity. To better adjudicate between the null-hypothesis (H_0) and an alternative hypothesis (H_r), we performed exploratory replication Bayes factor analysis for correlations (Wagenmakers, Verhagen & Ly 2016). The obtained replication Bayes factor quantifies the evidence that there is an effect in the size and direction reported by DeLong et al. (see Fig. 5). For the articles, this yielded strong to extremely strong evidence for the null hypothesis that the effect of cloze is zero, with BF_{0r} values up to 154 (at the Cz electrode depicted by DeLong et al., BF_{0r} = 77), and strongest evidence at the posterior channels. For the nouns, we obtained extremely strong evidence for the alternative hypothesis that the effect is nonzero, particularly at posterior channels, with BF_{10} values up to 9,163,515 (at Cz, BF_{r0} = 10,725). The pattern of results was similar when the 500 ms pre-stimulus baseline was applied.

Next, we computed Bayesian mixed-effect model estimates (β) and 95% credible intervals (CrI) for our single-trial analyses, using priors based on the results from DeLong et al. In both of our article-analyses credible intervals included zero (100 ms baseline: β = .31, CrI [-.06 .69]; 500 ms baseline: β = .17, CrI [-.22 .55]). For the nouns, zero was not within the credible interval: β = 2.24, CrI [1.77 2.70]. These Bayesian analyses further demonstrate our failure to replicate the DeLong et al. article-effect alongside a successful replication of the noun-effect. The analyses suggest that the data (combined with prior assumptions about the effect) are not very consistent with the hypothesis that the article-effect is zero (further information and posterior summaries are available in Supplementary Figure 2), but also are extremely inconsistent with the hypothesis that the article-effect is as big as that observed by DeLong and colleagues (2005). The data are most consistent with an effect that is more likely to be positive than zero or negative, but is very small (so small that it was not detected at traditional significance levels in this large-scale experiment with substantially higher power than previous experiments).

Control experiment

Lack of a statistically significant, article-elicited prediction effect could reflect a general insensitivity of our participants to the phonologically conditioned variation of the English indefinite article, i.e., a/an alternation. We ruled out this alternative explanation in an additional experiment that followed the replication experiment as part of the same experimental session. Participants read 80 short sentences containing the same nouns as the replication experiment, preceded by a phonologically licit or illicit article (e.g., “David found a/an apple...”), presented in the same manner as before. In each laboratory, nouns following illicit articles elicited a late positive-going waveform compared to nouns following licit articles (see Fig. 6), starting at about 500 ms after word onset and strongest at parietal electrodes. This standard P600 effect (Osterhout & Holcomb, 1992) was confirmed in a single-trial analysis, $\chi^2(1) = 83.09$, $p < .001$, and did not significantly differ between labs, $\chi^2(8) = 8.98$, $p = .35$.

DISCUSSION

In a landmark study, DeLong, Urbach and Kutas observed a statistically significant, graded modulation of article- and noun-elicited electrical brain potentials (N400) by the pre-determined probability that people continue a sentence fragment with that word (cloze). They concluded that people routinely and probabilistically pre-activate upcoming words to a high level of detail, including whether a word starts with a consonant or vowel. Our direct replication study spanning 9 laboratories successfully replicated a statistically significant effect of cloze on noun-elicited N400 activity but, critically, failed to replicate such an effect of cloze on article-elicited N400 activity. This pattern of success and failure was observed in a pre-registered replication analysis that duplicated the original study's analysis, and a pre-registered single-trial analysis that modelled variance at the level of item and subject. Exploratory Replication Bayes Factor analyses confirmed that we successfully replicated the direction and size of the correlations reported by Delong et al. for the nouns, but not for the articles. Exploratory Bayesian mixed-effects model analyses suggested that, while there is some evidence that the true population-level effect may be in the direction reported by DeLong and colleagues, the effect is likely far smaller than what they reported. In fact, the effect is likely is too small to be meaningfully observed without very large sample sizes, hence of uncertain theoretical interest. Finally, a control experiment confirmed that our participants did respect the phonological alternation a/an of the article with nouns used in the replication experiment.

Our findings carry important theoretical implications by challenging a crucial cornerstone of the ‘strong prediction view’ held by current theories of language comprehension (e.g., Altmann & Mirkovic, 2009; Pickering & Garrod, 2013). The strong prediction view entails two key claims. The first is that people pre-activate words at all levels of representation in a routine and implicit (i.e., non-strategic) fashion. Pre-activation is not limited to a word’s meaning, but includes its grammatical features and even its orthographic and/or phonological form. This would put language on a par with other cognitive systems such as visual perception that attempt to predict the inputs to lower-level ones (Friston, 2005, 2010; Summerfield & De Lange, 2014). The second claim is that pre-activation occurs at all levels of contextual support and gradually increases in strength with the level of contextual support. When contextual support for a specific word is high, like at a 100% cloze value, the word’s form and meaning is strongly pre-activated. When contextual support for a word is low, like when it is one amongst 20 words each with a 5% cloze value, pre-activation is distributed across multiple potential continuations. However, even then, a word’s form and meaning are pre-activated, just weakly so. The strength of pre-activation is probabilistic, that is, linked to estimated probability of occurrence.

DeLong and colleagues, and subsequently other scientists (e.g., Dell & Chang, 2014; Pickering & Clark, 2013), took their results as the evidence to support both these claims. DeLong et al (2005) was – and still is - the only study to date that measured pre-activation at the prenominal articles a and an that do not differ in their semantic or grammatical content, and that observed a graded relationship between cloze and N400 activity across a range of low- and high-cloze words, rather than merely a difference between low- and high-cloze words. Given that the use of these articles depends on whether the next word starts with a vowel or consonant, their results were considered as powerful evidence that participants probabilistically pre-activated the initial sound of upcoming nouns.

However, we show that there is no statistically significant effect of cloze on article-elicited N400 activity, using a sample size more than ten times that of the original, and a statistical analysis that better accounts for sources of non-independence than the original averaging-based correlation approach. If an effect of cloze on article-N400s exists at all, its true effect size is so small that it cannot be reliably detected even in an expansive multi-laboratory approach, let alone in the typical sample size in psycholinguistic and neurolinguistic experiments (roughly, N= 30). This means that even if article-cloze is associated with a graded modulation of N400 amplitudes, this effect seems to be so small that it cannot be reliably measured with small samples, and thus the previous studies may not have contributed much reliable information to our understanding of this effect. Moreover, it is also possible that the effect is sensitive to specifics of the experimental procedure and context that it lacks generalizability. Current theoretical positions thus either require new strong evidence for phonological pre-activation or require revision. In particular, the strong prediction view that claims that pre-activation routinely occurs across all – including phonological – levels (Pickering & Garrod, 2013), can no longer be viewed as having strong empirical support. Our work impels the field think differently about what constitutes strong evidence within a theory, but also highlights the need for a theory of linguistic prediction to formulate quantitative predictions about the effect-size of to-be-observed effects.

By contrast, we observed a strong and statistically significant effect of cloze on noun-elicited activity in the majority of our analyses. Although three of the nine laboratories did not show statistically significant correlations between noun-cloze and N400s, data pooled all laboratories showed a strong and statistically significant noun-cloze effect, our Replication Bayes Factor analysis overwhelmingly replicated the direction and size of the noun-cloze effect of DeLong et al., and our more powerful single-trial analysis revealed a significant noun-cloze effect in each of the laboratories. These results are therefore consistent with the handful of studies that reported a graded relationship between noun-cloze and noun-N400s (DeLong et al., 2005; Kutas & Hillyard, 1984; Wlotko & Federmeier, 2012).

Where does this pattern of failure and success leave the strong prediction view? Following the experimental logic of DeLong et al, we do not have sufficient evidence to conclude that people routinely pre-activate the initial phoneme of an upcoming noun, or perhaps any other word form information. Without pre-activation of the initial phoneme, the specific instantiation of the article does not cause people to revise their prediction about the meaning of the upcoming noun, thus lacking any impact on processing. Crucially, this conclusion is incompatible with the strong prediction view, because it suggests that pre-activation does not occur to the level of detail that is often assumed. Our results are also incompatible with an alternative interpretation of the DeLong et al. findings that people predict the article itself together with the noun (Ito, Corley, Pickering, Martin & Nieuwland, 2016; Van Petten & Luka, 2012), and they pose a serious challenge to the theory that comprehenders predict upcoming words, including their initial phonemes, through implicit production (Pickering & Garrod, 2013). Crucially, the idea that prediction is probabilistic, rather than all-or-none, is now questionable, given that there is no other published report of a pre-activation gradient. Although other studies have claimed prediction of form (Ito et al., 2016) or a prediction gradient (Smith & Levy, 2013), no study has indisputably demonstrated graded pre-activation, i.e., graded effects occurring before the noun. Effects that are observed upon, rather than before the noun, do not purely index pre-activation but index a mixture of memory retrieval and semantic integration processes instigated by the noun itself (Baggio & Hagoort, 2011; Lau, Namyst, Fogel & Delgado, 2016; Otten & Van Berkum, 2008; Steinhauer, Royle, Drury & Fromont, 2017). Therefore, there is currently no clear evidence to support routine probabilistic pre-activation of a noun’s phonological form during sentence comprehension.

Our results, however, do not necessarily exclude phonological form pre-activation, and we temper our conclusion with a caveat stemming from the a/an manipulation. For this manipulation to ‘work’, people must specifically predict the initial phoneme of the next word, and revise this prediction when faced with an unexpected article. However, because articles are only diagnostic about the next word within the noun phrase, rather than about the head noun itself, an unexpected article does not refute the upcoming noun, it merely signals that another word would come first (e.g., ‘an old kite’). This opens up explanations for why the a/an manipulation ‘fails’. In addition, comprehenders may not predict the noun to follow immediately, but at a later point; the unexpected article then does not evoke a change in prediction. Predictions about a specific position may be disconfirmed too often in natural language to be viable. This idea is supported by corpus data (Corpus of Contemporary American English and British National Corpus,), showing a mere 33% probability that a/an is directly followed by a noun. Alternatively, people predict the noun to come next, but only revise their prediction about its linear position while retaining the prediction about its meaning. So perhaps a revision of the predicted meaning, not the position, is required to trigger differential ERPs. In both of these hypothetical scenarios, people do not revise their prediction about the upcoming noun’s meaning unless they must.

Our results can be straightforwardly reconciled with effects reported for other pre-nominal manipulations, such as those of Dutch or Spanish article-gender (e.g., Van Berkum et al., 2008; Otten, Nieuwland & Van Berkum, 2008; Otten & Van Berkum, 2009; Wicha et al., 2004). Unlike a/an articles, gender-marked articles can immediately disconfirm the noun, because article- and noun-gender agrees regardless of intervening words (e.g., the Spanish article ‘el’ heralds a masculine noun). Revising the prediction about the noun presumably results in a semantic processing cost, thereby modulating N400 activity. Although gender-marked articles do not consistently incur the exact same type of effect and have only been observed at very high cloze values, previous studies suggest that a noun's grammatical gender can be pre-activated along with its meaning. Compared to this gender-manipulation, DeLong et al's study based on the English a/an manipulation claimed a stronger version of the prediction view, namely that people predict which word comes next up to its phonological form and, make backwards prediction as to the phonological form of the preceding linguistic material even on the basis of probabilistic, graded information.

What do our results say about prediction during natural language processing? Like the conclusions by DeLong et al., ours are limited by the generalization from language comprehension in a laboratory setting. On one hand, a rich conversational or story context may enhance predictions of upcoming words, and listeners may be more likely to pre-activate the phonological form of upcoming words than readers. On the other hand, our laboratory setting offered particularly good conditions for prediction of the next word's initial sound to occur. Each article was always immediately followed by a noun, unlike in natural language. Moreover, compared to natural reading rates our word presentation rate was slow, which may facilitate predictive processing (Ito et al., 2016; Wlotko & Federmeier, 2015). In natural reading, articles are hardly fixated and often skipped (e.g., O'Regan 1979). In short, arguments can be made both for and against phonological form prediction in natural language settings, and novel avenues of experimentation are needed to settle this issue.

DeLong and colleagues recently stated an omission in the description of their data analysis, i.e., a baseline procedure was applied to the data but inadvertently omitted from the description (DeLong et al., 2005). We have shown that our conclusions hold regardless of the baseline procedure. In a recent commentary, Delong, Urbach, and Kutas (2017) also described filler-sentences in their experiment, which were omitted from their original report, and were neither provided nor mentioned to us by the authors upon our request for the stimuli. DeLong et al. used the existence of these filler-sentences to dismiss an alternative explanation of their original findings, namely that an unusual experimental context wherein every sentence contains an article-noun combination leads participants to strategically predict upcoming nouns. Following this logic, we failed to replicate their article-effects despite an experimental context that could inadvertently encourage strategic prediction (for demonstrations of experimental context boosting predictive processing, see Brothers, Swaab & Traxler, 2017; Lau, Holcomb & Kuperberg, 2013). Therefore, the presence of fillers in their experiment versus absence in ours cannot straightforwardly explain the different results, and may even strengthen our conclusions.

To conclude, we failed to replicate the main result of DeLong et al., a landmark study published more than ten years ago that has not been directly replicated since. Our results suggest that, if there is an effect of article-cloze probability on the amplitude of the N400, it is too small and/or too sensitive to unknown experimental design factors to have been meaningfully measured in previous small-sample-size experiments. We conclude that such an effect does not constitute strong evidence for current theoretical positions on the importance of prediction (e.g., Pickering & Garrod, 2013). Our findings thus challenge one of the pillars of the 'strong prediction view' in which people routinely and probabilistically pre-activate information at all levels of linguistic representation, including phonological form information such as the initial phoneme of an upcoming noun. Consequently, there is currently no convincing evidence that people pre-activate the phonological form of an upcoming noun during sentence comprehension, and we take our findings to suggest a more limited role for prediction during language comprehension. In addition, our findings further highlight the importance of direct replication, large sample size studies, transparent reporting and of pre-registration to advance reproducibility and replicability in the neurosciences.

MATERIALS AND METHODS

Experimental design and materials. Nieuwland twice requested all original materials from DeLong et al., including the questions and norms, with the stated purpose of direct replication (personal communication, November 4 and 19, 2015), upon which DeLong et al. made available the 80 sentences described in the original study. These sentences were then adapted from American to British spelling and underwent a few minor changes to ensure their suitability for British participants. The complete set of materials and the list of changes to the original materials are available online (Supplementary Table 1 and 2). The materials were 80 sentence contexts with two possible continuations each: a more or less expected indefinite article + noun combination. The noun was followed by at least one subsequent word. All article + noun continuations were grammatically correct. Each article + noun combination served once as the more expected continuation and the other time as the less expected continuation, in different contexts. We divided the 160 items in two lists of 80 sentences such that each list contained each noun only once. Each participant was presented with only one list (thus, each context was seen only once). One in four sentences was followed by a yes/no comprehension question, which yielded a mean response accuracy of 95% (after taking into account ambiguity in three of the questions, see Supplemental Table 2 and 3). While this percentage is very similar to that reported by DeLong et al., we note that this cannot be directly compared to the accuracy reported in DeLong et al., because we had to create new comprehension questions in the absence of the original ones. Regardless, because Delong et al. suggested that our results were due to poor language comprehension (DeLong, Urbach & Kutas, 2017), we describe an exploratory analysis in which we attempt to account for variation in response accuracy in the statistical model.

We obtained article cloze and noun cloze ratings from a separate group of native speakers of English who were students at the University of Edinburgh and did not participate in the ERP experiment. They were instructed to complete the sentence fragment with the best continuation that comes to mind (Taylor, 1953). We obtained article cloze ratings from 44 participants for 80 sentence contexts truncated before the critical article. Noun cloze ratings were obtained by first truncating the sentences after the critical articles, and presenting two different, counterbalanced lists of 80 sentences to 30 participants each, such that a given participant only saw each sentence context with the expected or the unexpected article. The obtained values closely resemble those of the original study, with the same range (0-100% for articles and nouns), slightly lower median values (for articles and nouns, 29% and 40%, compared to 31% and 46% in the original study), but slightly higher mean values (for articles and nouns, 41% and 46%, compared to 36% and 44%). Because the sentence materials we used describe common situations that can be understood by any English speaker, and because students at the University of Edinburgh come from across the whole of the UK, we had no a priori expectation that cloze ratings would differ substantially across laboratories, and thus we did not obtain cloze norms from other sites. Consistent with this assumption, nothing in our results suggests stronger cloze effects in University of Edinburgh students compared to other students, suggesting that our cloze norms are sufficiently representative for the other universities.

Participants. Participants were students from the University of Birmingham, Bristol, Edinburgh, Glasgow, Kent, Oxford, Stirling, York, or volunteers from the participant pool of University College London or Oxford University, who received cash or course credit for taking part in the ERP experiment. Participant information and EEG recording information per laboratory is available online (Supplementary Table 3). We pre-registered a target sample size of 40 participants per laboratory, which was thought to give at least 32 participants (the sample size of DeLong et al.) per laboratory after accounting for data loss, as was later confirmed. Due to logistic constraints, not all laboratories reached an N of 40. Because in two labs corruption of data was incorrectly assumed before computing trial loss, these laboratories tested slightly more than 40 participants. All participants (N = 356; 222 women) were right-handed, native English speakers with normal or corrected-to-normal vision, between 18–35 years (mean, 19.8 years), free from any known language or learning disorder. Eighty-nine participants reported a left-handed parent or sibling.

Procedure. After giving written informed consent, participants were tested in a single session. Sentences were presented visually in the center of a computer display, one word at a time (200 ms duration, followed by a blank screen of 300 ms duration[3]). Participants were instructed to read sentences for comprehension and answer yes/no comprehension questions by pressing hand-held buttons. The electroencephalogram (EEG) was recorded from at least 32 electrodes.

The replication experiment was followed by a control experiment, which served to detect sensitivity to the correct use of the a/an rule in our participants. Participants read 80 relatively short sentences (average length 8 words, range 5-11) that contained the same critical words as the replication experiment, preceded by a correct or incorrect article. As in

[3] Due to a programming error, in four labs (1, 3 , 5 and 8, which used E-prime scripts) the critical articles and nouns, but not other words, were followed by a 380 ms blank instead of the intended 300 ms. However, this delay is unlikely to have affected the results because if it was noticed at all, which is unlikely, it appeared after the N400 window associated with the article. Moreover, if anything, longer duration facilitates language comprehension and predictive processing (Camblin, LeDoux, Boudewyn, Gordon & Swaab, 2007; Wlotko & Federmeier, 2015; Ito et al., 2016), making it more, not less likely to find an effect of cloze on the article-ERPs. Of note, the qualitative pattern of the results from the pre-registered single-trial analysis did not change when we removed these labs from the analysis.

the replication experiment, each critical word was presented only once, and was followed by at least one more word. All words were presented at the same rate as the replication experiment. There were no comprehension questions in this experiment. After the control experiment, participants performed a Verbal Fluency Test and a Reading Span test; the results from these tests are not discussed here. All stimulus presentation scripts are publicly available in two different software packages (E-Prime and Presentation) on https://osf.io/eyzaq.

Data processing. Data processing was performed in BrainVision Analyzer 2.1(Brain Products, Germany). We performed one pre-registered replication analysis that followed the DeLong et al. analysis as closely as possible and one pre-registered single-trial analysis (Open Science Framework, https://osf.io/eyzaq). All non-pre-registered analyses are considered as exploratory. First, we interpolated bad channels from surrounding channels, and downsampled to a common set of 22 EEG channels per laboratory which were similar in scalp location to those used by DeLong et al. One laboratory did not have 12 of the selected 22 channels in its EEG channel montage, and we matched the full 22-channel layout used for other laboratories by creating 12 virtual channels from neighbouring channels using topographic interpolation by spherical splines. We then applied a 0.01-100 Hz digital band-pass filter (including 50 Hz Notch filter), re-referenced all channels to the average of the left and right mastoid channels (in a few participants with a noisy mastoid channel, only one mastoid channel was used), and segmented the continuous data into epochs from 500 ms before to 1000 ms after word onset. We then performed visual inspection of all data segments and rejected data with amplifier blocking, movement artifacts, or excessive muscle activity. Subsequently, we performed independent component analysis (Jung et al., 2000) on a 1-Hz high-pass filtered version of the data, and applied the obtained weightings to the original data to correct for blinks, eye movements or steady muscle artefacts. After this, we automatically rejected segments containing a voltage difference of over 120 µV in a time window of 150 ms or containing a voltage step of over 50 µV/ms. Participants with fewer than 60/80 article trials or 60/80 noun trials were removed from the analysis, leaving a total of 334 participants (range across laboratories 32-42, and therefore each lab had a sample size at least as large as DeLong et al.). On average, participants had 77 article trials and 77 noun trials.

Pre-registered replication analysis. We applied a 4th-order Butterworth band-pass filter at 0.2-15 Hz to the segmented data, averaged trials per participant within 10% cloze bins (0-10, 11-20, etc. until 91-100), and then averaged the participant-wise averages separately for each laboratory. Because the bins did not contain equal numbers of trials (the intermediate bins contained fewest trials), like in DeLong et al., not all participants contributed a value for each bin to the grand average per laboratory. For nouns and articles separately, and for each EEG channel, we computed the correlation between ERP amplitude in the 200-500 ms time window per bin with the average cloze probability per bin.

Pre-registered single-trial analysis. In this analysis, we did not apply the 0.2-15 Hz band-pass filter, which carries the risk of inducing data distortions (Luck, 2014; Tanner, Morgan-Short & Luck, 2015). However, we deemed it necessary to perform a baseline correction of the data. This procedure corrects for spurious voltage differences before word onset, generating confidence that observed effects are elicited by the word rather than differences in brain activity that already existed before the word and is a standard procedure in ERP research (Luck, 2014). DeLong et al (2005) did not report a baseline correction, nor did any of the related work from DeLong and colleagues that was reported in DeLong (2009). Yet baseline correction has been used in many other publications from the Kutas Cognitive Electrophysiology Lab. We chose a 100 ms pre-stimulus baseline as the most frequently used one both in other studies from Kutas lab and in similar studies from other labs. For each trial, we performed baseline correction by subtracting the mean voltage of the -100 to 0 ms time window from each data point in the epoch.

Instead of averaging N400 data across trials and participants for subsequent statistical analysis, we performed linear mixed effects model analysis (Baayen, Davidson & Bates, 2008) of the single-trial N400 data, using the “lme4” package (Bates, Maechler, Bolker & Walker, 2014) in the R software (R CoreTeam, 2014). This approach simultaneously models variance associated with each subject and with each item. Using a spatiotemporal region-of-interest approach based on the DeLong et al. results, our dependent measure (N400 amplitude) was the average voltage across 6 centro-parietal channels (Cz/C3/C4/Pz/P3/P4) in the 200-500 ms window for each trial. Analysis scripts and data to run these scripts are publicly available on https://osf.io/eyzaq.

For articles and nouns separately, we used a maximal random effects structure as justified by the design (Barr et al., 2013), which did not include random effects for ‘laboratory’ as there were only 9 laboratories. Z-scored cloze was entered in the model as a continuous variable, and laboratory was entered as a deviation-coded nuisance predictor. We tested the effects of ‘laboratory’ and ‘cloze’ through model comparison with a χ^2 log-likelihood test. We tested whether the inclusion of a given fixed effect led to a significantly better model fit. The first model comparison examined laboratory effects, namely whether the cloze effect varied across laboratories (cloze-by-laboratory interaction) or whether the N400 magnitudes varied over laboratory (laboratory main effect). If laboratory effects were nonsignificant, we dropped them from the analysis because they were not of theoretical interest. For the articles and nouns separately, we compared the subsequent models below. Each model included the random effects associated with the fixed effect ‘cloze’ (see Barr et al., 2014). All output β estimates and 95% confidence intervals (CI) were transformed from z-scores back to raw scores, and then back to the 0-100% cloze range, so that the voltage estimates represent the change in voltage associated with a change in cloze probability from 0 to 100.

Model 1: N400 ~ cloze * laboratory + (cloze | subject) + (cloze | item)

Model 2: N400 ~ cloze + laboratory + (cloze | subject) + (cloze | item)

Model 3: N400 ~ cloze + (cloze | subject) + (cloze | item)

Model 4: N400 ~ (cloze | subject) + (cloze | item)

In an analysis that included the data from both articles and nouns, we also tested the differential effect of cloze on article ERPs and on noun ERPs by comparing models with and without an interaction between cloze and the deviation-coded factor ‘wordtype’ (article/noun). Random correlations were removed for the models to converge.

Model 1: N400 ~ cloze * wordtype + (cloze * wordtype || subject) + (cloze * wordtype || item)

Model 2: N400 ~ cloze + wordtype + (cloze * wordtype || subject) + (cloze * wordtype || item)

Exploratory correlation analysis. Of note, DeLong et al. have recently described using a 500 ms baseline correction procedure that they failed to mention in DeLong et al. (2005). Using this baseline correction procedure, we recomputed the correlations that we obtained in our Replication analysis. To compare our results most directly with those reported in Figure 1C of DeLong el al. (2005), we pooled data from all the laboratories so that we would have a single r-value for each EEG-channel.

Exploratory single-trial analyses. We performed an exploratory analysis in the 500 to 100 ms time window before the article, using the originally (-100 to 0 ms) baselined data, using Model 3 and 4 from the article analysis. This window covers the first 400 ms of the word that preceded the article. Analysis in this window yielded a similar pattern as in the pre-registered analysis, which indicates that a baseline correction procedure covering the entire

500 ms pre-stimulus window would account better for pre-article voltage levels. We performed this additional analysis, the results of which did not change our conclusions and are shown in Supplementary Figure 1.

We also performed an exploratory analysis in which we control for a potential influence of response accuracy, taken as a proxy for the subject's attention to the task, on predictive processing of linguistic input. We entered the (z-transformed) average response accuracy of each subject in our model, and compared the models below. Comparison of Model 1 and 2 tested whether the effect of cloze on the article-N400s depended subject accuracy. Comparison of Model 2 and 3 tested whether there was a significant effect of cloze on article-N400s when subject accuracy was included in the model.

Model 1: N400 ~ accuracy * cloze + (cloze | subject) + (cloze | item)

Model 2: N400 ~ accuracy + cloze + (cloze | subject) + (cloze | item)

Model 3: N400 ~ accuracy + (cloze | subject) + (cloze | item)

Exploratory Bayesian analyses. Supplementing the Replication analysis, we performed a Replication Bayes factor analysis for correlations (Wagenmakers et al., 2016) using as prior the size and direction of the effect reported in the original study. We performed this test for each electrode separately, after collapsing the data points from the different laboratories. Because we had no articles in the 40-50 % cloze bin, there was a total of 9 and 10 data points per laboratory for the articles and nouns, respectively. Our analysis used priors estimated from the DeLong et al results matched as closely as possible to our electrode locations. A Bayes factor between 3 and 10 is considered moderate evidence, between 10-30 is considered strong evidence, 30-100 is very strong evidence, and values over 100 are considered extremely strong evidence (Jeffreys, 1961). In addition to using a 100 ms pre-stimulus baseline, we also computed the replication Bayes factors using the 500 ms pre-stimulus time window for baseline correction. Results are shown in Figure 5.

Supplementing the pre-registered single-trial analyses, we performed an exploratory Bayesian mixed-effects model analysis using the brms package for R (Buerkner, 2016), which fits Bayesian multilevel models using the Stan programming language (Stan Development Team, 2016). Nieuwland requested to use the results of a mixed-effects model reanalysis of the DeLong et al. data as an appropriate prior (personal communication from Nieuwland, November 14 and 22 2017); this request was declined by DeLong and colleagues. We were therefore limited to using a prior centered on a point estimate based on the Delong et al. correlation analysis, namely our estimate of the observed effect size at Cz for a difference between 0% cloze and 100% cloze (1.25 μV and 3.75 μV for articles and nouns, respectively, based on visual inspection of the graphs) and a prior centered on zero for the intercept. Both priors had a normal distribution and a standard deviation of 0.5 (given the a priori expectation that average ERP voltages in this window generally fluctuate on the order of a few microvolts; note that these units are expressed in terms of the z-scored cloze values, rather than the original cloze values, such that μ for the cloze prior was 0.45, which corresponds to a raw cloze effect of 1.25). We computed estimates and 95% credible intervals for each of the mixed-effects models we tested, and transformed these back into raw cloze units. The credible interval is the range of values such that one can be 95% certain that it contains the true effect, given the data, priors and the model. The results from these analyses are shown in Supplementary Figure 2; the analyses suggest that, while there may be a small positive association between article cloze and ERP amplitude elicited by the articles, the effect is substantially smaller than that estimated by Delong and colleagues (2005) and likely is too small to be meaningfully observed without very large sample sizes, hence of uncertain theoretical interest.

Control experiment. Analysis of the control experiment involved a comparison between a model with the categorical factor 'grammaticality' (grammatical/ungrammatical) and a model without. Our dependent measure (P600 amplitude; Osterhout & Holcomb, 1992) was the average voltage across 6 centro-parietal channels (Cz/C3/C4/Pz/P3/P4) in the 500-800 ms window for each trial. Results are shown in Figure 6.

Model 1: P600 ~ grammaticality + (grammaticality | subject) + (grammaticality | item)

Model 2: P600 ~ (grammaticality | subject) + (grammaticality | item)

Acknowledgements

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

Competing financial interests

The authors declare no competing financial interests.

Author contributions

M.S.N. and F.H. designed the research, M.S.N., D.J.B., G.R., and S.P.-A. planned the analysis. E.H., E.D., S.V.G.Z.W., F.B., V.K., A.I., S.B.-M., Z.F., E.K., S.P-A., and Z.K. collected data. M.S.N., K.S., N.K., G.R., H.J.F., J.T., E.M.H., D.I.D., and S.R supervised data collection. M.S.N. and S.P.-A. analyzed the data. M.S.N. drafted the manuscript and received comments from S.P.-A., N.K., K.S., D.J.B., H.J.F., E.M.H, and F.H.

REFERENCES

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. Cognition, 73(3), 247-264.

Altmann, G. T., & Mirkovic, J. (2009). Incrementality and Prediction in Human Sentence Processing. Cogn Sci, 33(4), 583-609. doi: 10.1111/j.1551-6709.2009.01022.x

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language, 59(4), 390-412. doi: 10.1016/j.jml.2007.12.005

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

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language, 68(3), 255-278. doi: 10.1016/j.jml.2012.11.001

Brothers, T., Swaab, T. Y., & Traxler, M. J. (2017). Goals and strategies influence lexical prediction during sentence comprehension. Journal of Memory and Language, 93, 203-216. doi: 10.1016/j.jml.2016.10.002

Brown, C., & Hagoort, P. (1993). The processing nature of the n400: evidence from masked priming. J Cogn Neurosci, 5(1), 34-44. doi: 10.1162/jocn.1993.5.1.34

Brown, C. M., Hagoort, P., & Chwilla, D. J. (2000). An event-related brain potential analysis of visual word priming effects. Brain Lang, 72(2), 158-190. doi: 10.1006/brln.1999.2284

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafo, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci, 14(5), 365-376. doi: 10.1038/nrn3475

Chwilla, D. J., Brown, C. M., & Hagoort, P. (1995). The N400 as a function of the level of processing. Psychophysiology, 32(3), 274-285.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav Brain Sci, 36(3), 181-204. doi: 10.1017/S0140525X12000477

Clark, H. H. (1973). Language as Fixed-Effect Fallacy - Critique of Language Statistics in Psychological Research. Journal of Verbal Learning and Verbal Behavior, 12(4), 335-359. doi: Doi 10.1016/S0022-5371(73)80014-3

Connolly, J. F., & Phillips, N. A. (1994). Event-related potential components reflect phonological and semantic processing of the terminal word of spoken sentences. J Cogn Neurosci, 6(3), 256-266. doi: 10.1162/jocn.1994.6.3.256

Dell, G. S., & Chang, F. (2014). The P-chain: relating sentence production and its disorders to comprehension and acquisition. Philos Trans R Soc Lond B Biol Sci, 369(1634), 20120394. doi: 10.1098/rstb.2012.0394

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

DeLong, K.A., Urbach, T.P., & Kutas, M. (2017). Is there a replication crisis? Perhaps. Is this an example? No: a commentary on Ito, Martin, and Nieuwland (2016), Language, Cognition and Neuroscience, DOI: 10.1080/23273798.2017.1279339

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

Friederici, A. D., Steinhauer, K., & Frisch, S. (1999). Lexical integration: sequential effects of syntactic and semantic information. Mem Cognit, 27(3), 438-453.
Friston, K. (2005). A theory of cortical responses. Philos Trans R Soc Lond B Biol Sci, 360(1456), 815-836. doi: 10.1098/rstb.2005.1622
Friston, K. (2010). The free-energy principle: a unified brain theory? Nat Rev Neurosci, 11(2), 127-138. doi: 10.1038/nrn2787
Hagoort, P. (2017). The core and beyond in the language-ready brain. Neurosci Biobehav Rev. doi: 10.1016/j.neubiorev.2017.01.048
Hauk, O., Davis, M. H., Ford, M., Pulvermuller, F., & Marslen-Wilson, W. D. (2006). The time course of visual word recognition as revealed by linear regression analysis of ERP data. Neuroimage, 30(4), 1383-1400. doi: 10.1016/j.neuroimage.2005.11.048
Huettig, F. (2015). Four central questions about prediction in language processing. Brain Res, 1626, 118-135. doi: 10.1016/j.brainres.2015.02.014
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
Ito, A., Martin, A. E., & Nieuwland, M. S. (2017a). How robust are prediction effects in language comprehension? Failure to replicate article-elicited N400 effects. Language Cognition and Neuroscience, 32(8), 954-965. doi: 10.1080/23273798.2016.1242761
Ito, A., Martin, A. E., & Nieuwland, M. S. (2017b). Why the A/AN prediction effect may be hard to replicate: a rebuttal to Delong, Urbach, and Kutas (2017). Language Cognition and Neuroscience, 32(8), 974-983. doi: 10.1080/23273798.2017.1323112
Jackendoff, R. (2002). Foundations of language : brain, meaning, grammar, evolution. Oxford ; New York: Oxford University Press.
Jung, T. P., Makeig, S., Humphries, C., Lee, T. W., McKeown, M. J., Iragui, V., & Sejnowski, T. J. (2000). Removing electroencephalographic artifacts by blind source separation. Psychophysiology, 37(2), 163-178.
Kintsch, W. (1988). The role of knowledge in discourse comprehension: a construction-integration model. Psychol Rev, 95(2), 163-182.
Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). Annu Rev Psychol, 62, 621-647. doi: 10.1146/annurev.psych.093008.131123
Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. Science, 207(4427), 203-205.
Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. Nature, 307(5947), 161-163.
Lau, E. F., Holcomb, P. J., & Kuperberg, G. R. (2013). Dissociating N400 Effects of Prediction from Association in Single-word Contexts. J Cogn Neurosci, 25(3), 484-502.
Lau, E., Namyst, A., Fogel, A., & Delgado, T. (2016). A direct comparison of N400 effects of predictability and incongruity in adjective-noun combination. Collabra: Psychology, 2(1).
Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (de)constructing the N400. Nat Rev Neurosci, 9(12), 920-933. doi: 10.1038/nrn2532
Marslen-Wilson, W., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. Cognition, 8(1), 1-71.
Martin, C. D., Thierry, G., Kuipers, J. R., Boutonnet, B., Foucart, A., & Costa, A. (2013). Bilinguals reading in their second language do not predict upcoming words as native readers do. Journal of Memory and Language, 69(4), 574-588. doi: 10.1016/j.jml.2013.08.001

Open Science, C. (2015). PSYCHOLOGY. Estimating the reproducibility of psychological science. Science, 349(6251), aac4716. doi: 10.1126/science.aac4716

O'Regan, K. (1979). Saccade size control in reading: evidence for the linguistic control hypothesis. Percept Psychophys, 25(6), 501-509.

Osterhout, L., & Holcomb, P. J. (1992). Event-Related Brain Potentials Elicited by Syntactic Anomaly. Journal of Memory and Language, 31(6), 785-806. doi: Doi 10.1016/0749-596x(92)90039-Z

Otten, M., Nieuwland, M. S., & Van Berkum, J. J. (2007). Great expectations: specific lexical anticipation influences the processing of spoken language. BMC Neurosci, 8, 89. doi: 10.1186/1471-2202-8-89

Otten, M., & Van Berkum, J. (2008). Discourse-Based Word Anticipation During Language Processing: Prediction or Priming? Discourse Processes, 45(6), 464-496. doi: Pii 90598764910.1080/01638530802356463

Otten, M., & Van Berkum, J. J. A. (2009). Does working memory capacity affect the ability to predict upcoming words in discourse? Brain Res, 1291, 92-101. doi: 10.1016/j.brainres.2009.07.042

Pickering, M. J., & Clark, A. (2014). Getting ahead: forward models and their place in cognitive architecture. Trends Cogn Sci, 18(9), 451-456. doi: 10.1016/j.tics.2014.05.006

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. Behav Brain Sci, 36(4), 329-347. doi: 10.1017/S0140525X12001495

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. Cognition, 128(3), 302-319. doi: 10.1016/j.cognition.2013.02.013

Steinhauer, K., Royle, P., Drury, J. E., & Fromont, L. A. (2017). The priming of priming: Evidence that the N400 reflects context-dependent post-retrieval word integration in working memory. Neurosci Lett, 651, 192-197. doi: 10.1016/j.neulet.2017.05.007

Summerfield, C., & de Lange, F. P. (2014). Expectation in perceptual decision making: neural and computational mechanisms. Nat Rev Neurosci, 15(11), 745-756. doi: 10.1038/nrn3838

Tanner, D., Morgan-Short, K., & Luck, S. J. (2015). How inappropriate high-pass filters can produce artifactual effects and incorrect conclusions in ERP studies of language and cognition. Psychophysiology, 52(8), 997-1009. doi: 10.1111/psyp.12437

Taylor, W. L. (1953). "Cloze Procedure": A New Tool For Measuring Readability. Journalism Quarterly, 30(4), 415-433.

Van Berkum, J. J. (2010). The brain is a prediction machine that cares about good and bad-any implications for neuropragmatics?. Italian Journal of Linguistics, 22, 181-208.

Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from ERPs and reading times. J Exp Psychol Learn Mem Cogn, 31(3), 443-467. doi: 10.1037/0278-7393.31.3.443

Van Berkum, J. J., Hagoort, P., & Brown, C. M. (1999). Semantic integration in sentences and discourse: evidence from the N400. J Cogn Neurosci, 11(6), 657-671.

Van Petten, C., Coulson, S., Rubin, S., Plante, E., & Parks, M. (1999). Time course of word identification and semantic integration in spoken language. Journal of Experimental Psychology-Learning Memory and Cognition, 25(2), 394-417. doi: Doi 10.1037//0278-7393.25.2.394

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: benefits, costs, and ERP components. Int J Psychophysiol, 83(2), 176-190. doi: 10.1016/j.ijpsycho.2011.09.015

Wagenmakers, E. J., Verhagen, J., & Ly, A. (2016). How to quantify the evidence for the absence of a correlation. Behav Res Methods, 48(2), 413-426. doi: 10.3758/s13428-015-0593-0

Wicha, N. Y., Moreno, E. M., & Kutas, M. (2004). Anticipating words and their gender: an event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. J Cogn Neurosci, 16(7), 1272-1288. doi: 10.1162/0898929041920487

Wlotko, E. W., & Federmeier, K. D. (2012). So that's what you meant! Event-related potentials reveal multiple aspects of context use during construction of message-level meaning. Neuroimage, 62(1), 356-366. doi: 10.1016/j.neuroimage.2012.04.054

Wlotko, E. W., & Federmeier, K. D. (2015). Time for prediction? The effect of presentation rate on predictive sentence comprehension during word-by-word reading. Cortex, 68, 20-32. doi: 10.1016/j.cortex.2015.03.014

Zwitserlood, P. (1989). The locus of the effects of sentential-semantic context in spoken-word processing. Cognition, 32(1), 25-64.

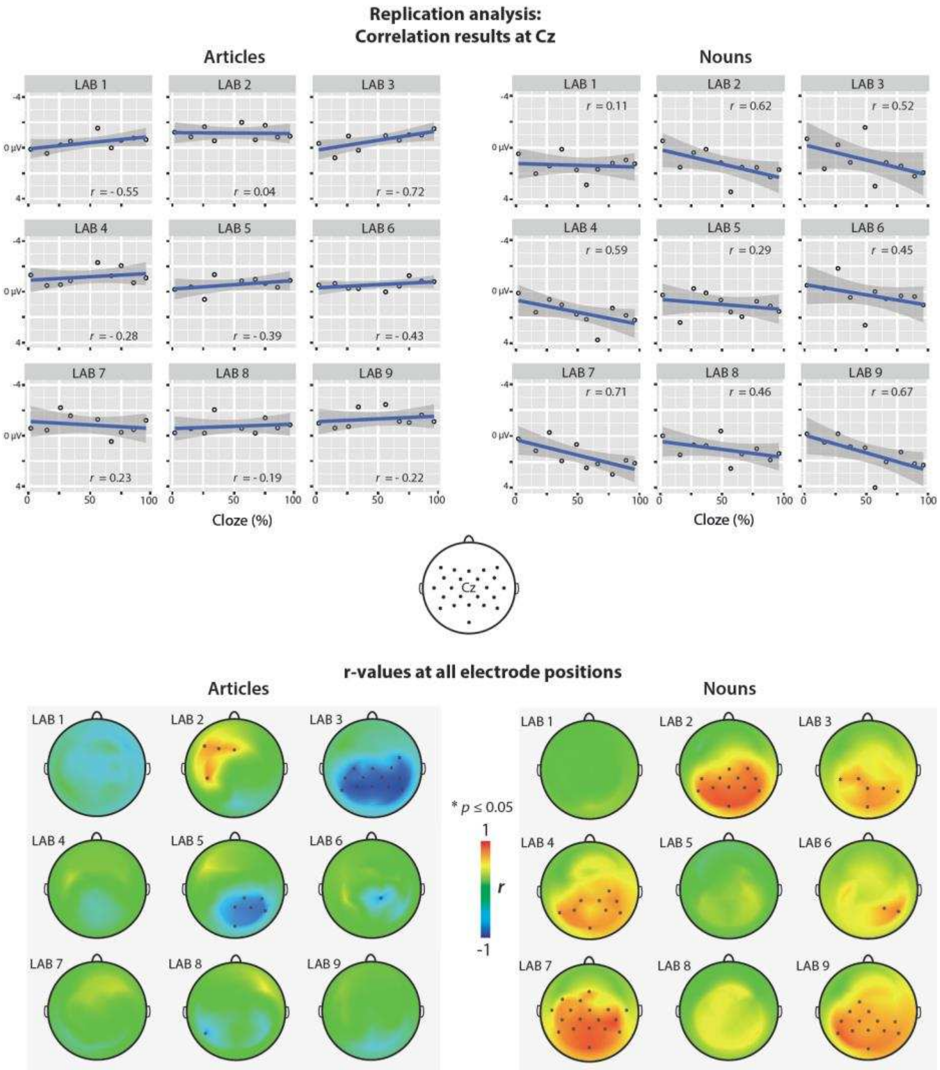


Figure 1. Replication analysis. Correlations between N400 amplitude and article/noun cloze probability per laboratory. N400 amplitude is the mean voltage in the 200-500 ms time window after word onset. A positive value corresponds to the canonical finding that N400 amplitude became smaller (less negative—more positive) with increasing cloze probability. Here and in all further plots, negative voltages are plotted upwards. Upper graph: Scatter plots showing the correlation between cloze and N400 activity at electrode Cz, for each lab. The position of Cz and the other electrodes is displayed in the head plot in between the upper and lower graph. Lower graph: Scalp distribution of the r-values for each lab. Asterisks (*) indicate electrodes that showed a statistically significant correlation (two-tailed $p < 0.05$, not corrected for multiple comparisons). Exact r- and p-values for each laboratory and EEG channel are available on https://osf.io/eyzaq.

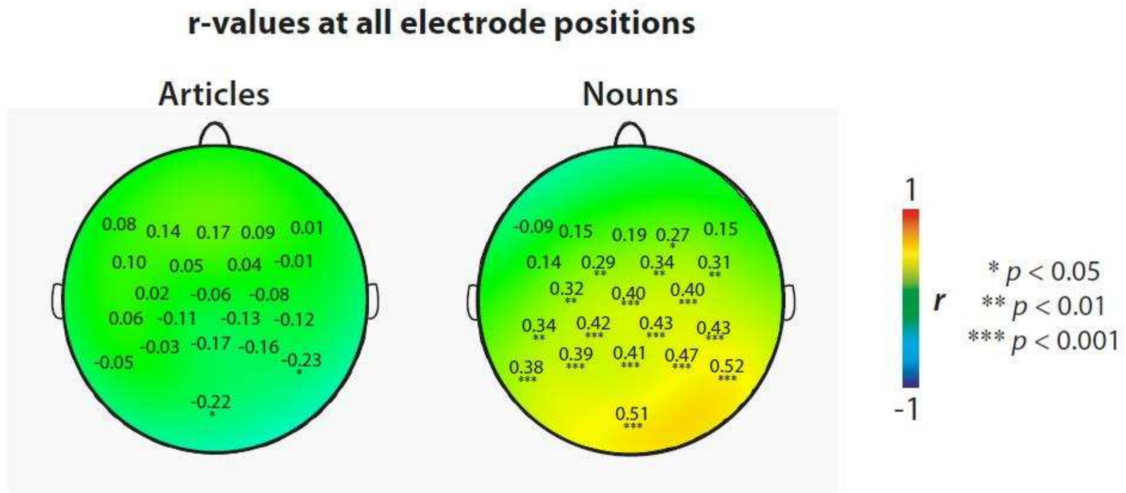


Figure 2. Replication analysis. Scalp distribution and r-values at each channel based on data pooled from all laboratories, using a 500 ms baseline correction procedure as used by DeLong et al (2005). Asterisks (*) indicate electrodes that showed a statistically significant correlation (two-tailed, not corrected for multiple comparisons).

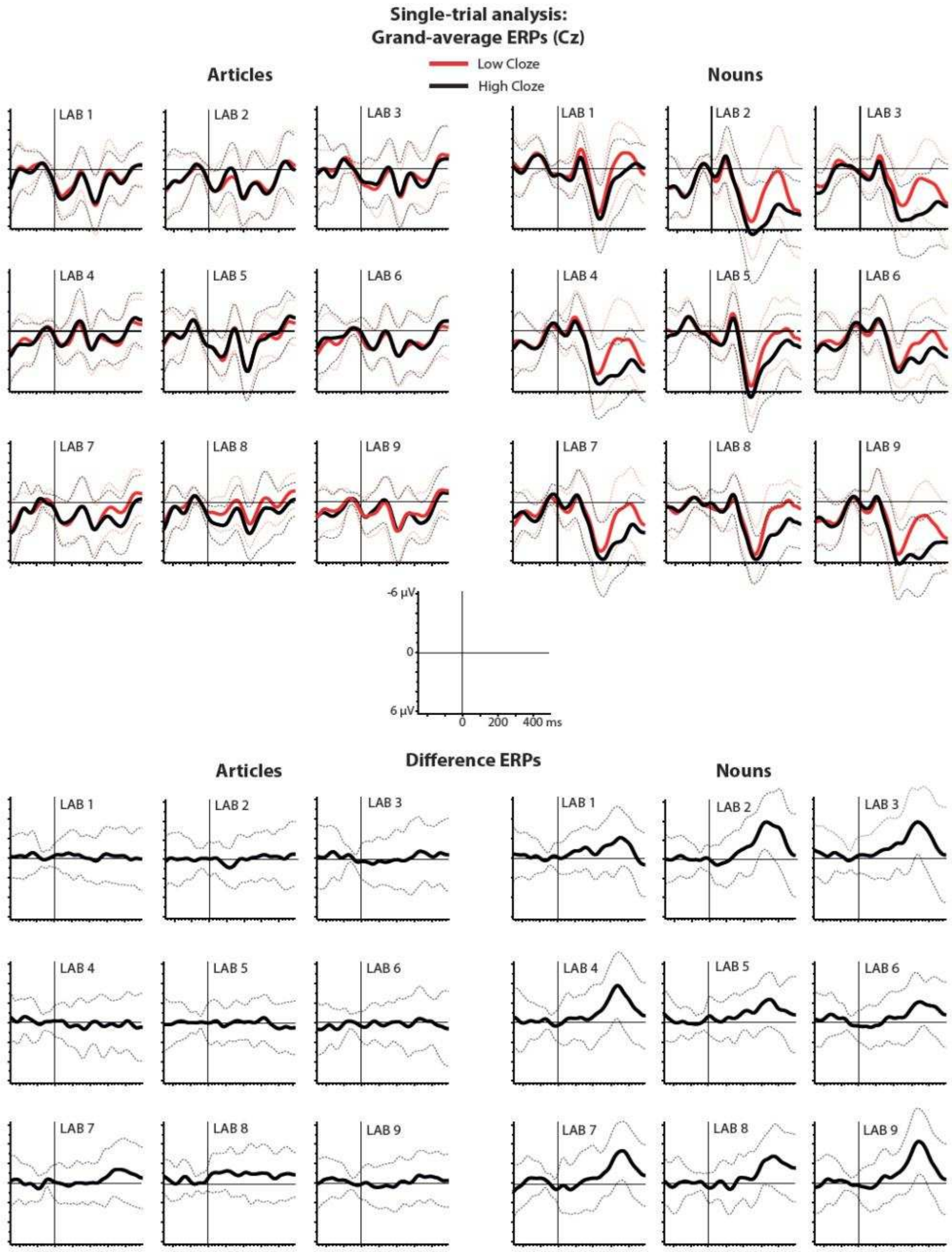


Figure 3. Single-trial analysis. Grand-average ERPs elicited by relatively expected and unexpected words (cloze higher/lower than 50%) and the associated difference waveforms (low minus high cloze) at electrode Cz. Dotted lines indicate 1 standard deviation above or below the grand average.

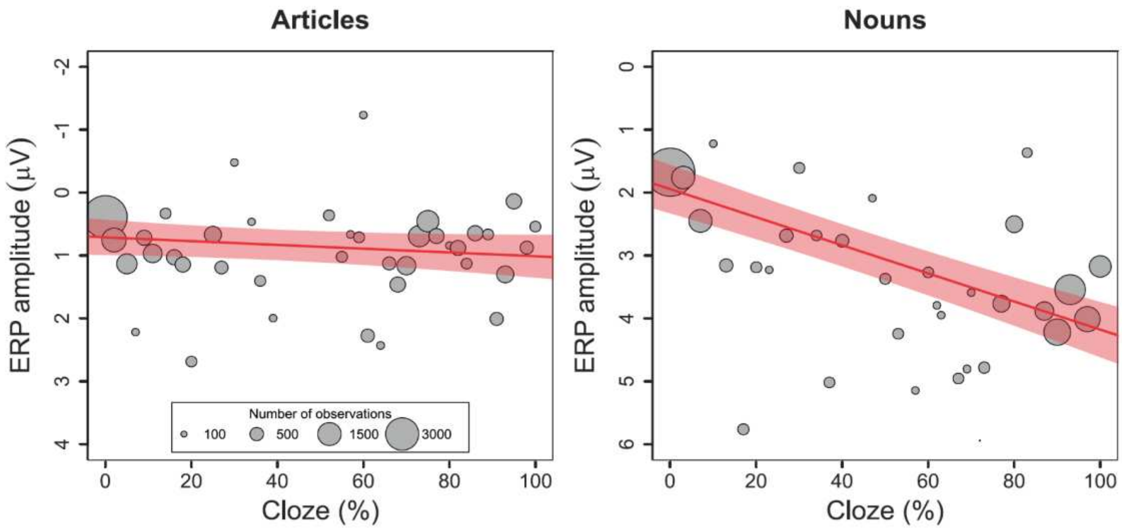


Figure 4. Single-trial analysis. Relationship between cloze and ERP amplitude for articles and nouns in the N400 spatiotemporal window, as illustrated by the mean ERP values per cloze value (number of observations reflected in circle size), along with the regression line and 95% confidence interval. A change in article cloze from 0 to 100 is associated with a change in amplitude of 0.296 µV (95% confidence interval: -.08 to .67). A change in noun-cloze from 0 to 100 is associated with a change in amplitude of 2.22 µV (95% confidence interval: 1.75 to 2.69). The data for these analyses was pooled across all 9 labs.

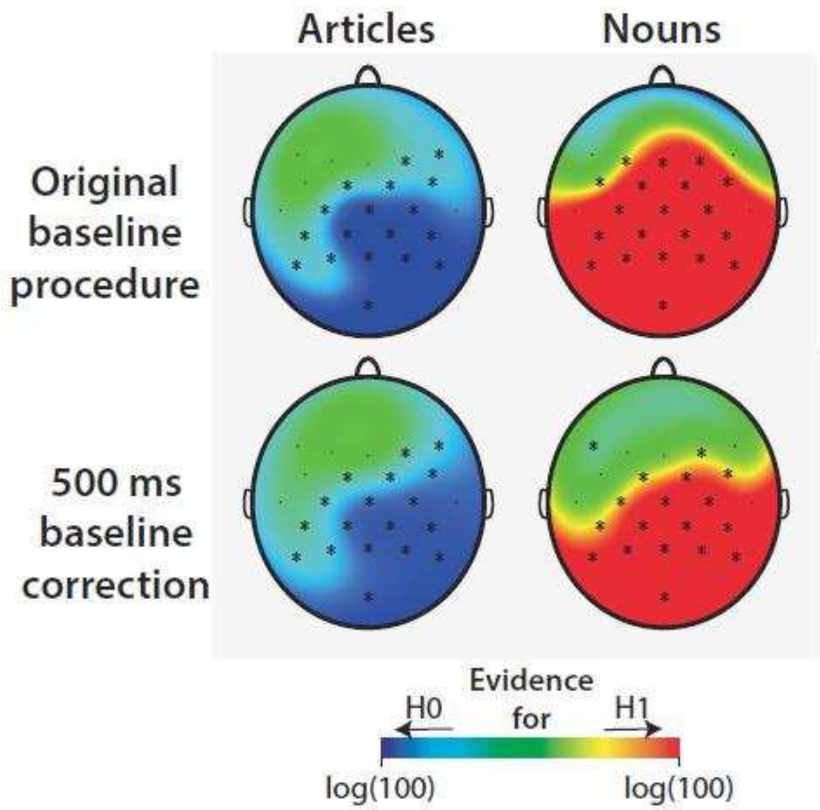


Figure 5. Exploratory Bayes factor analysis associated with the replication analysis, quantifying the obtained evidence for the null hypothesis (H_0) that N400 is not impacted by cloze, or for the alternative hypothesis (H_1) that N400 is impacted by cloze with the direction and size of effect reported by DeLong et al. Scalp maps show the common logarithm of the replication Bayes factor for each electrode, capped at log(100) for presentation purposes. Electrodes that yielded at least moderate evidence for or against the null hypothesis (Bayes factor of ≥ 3) are marked by an asterisk. At posterior electrodes where DeLong et al. found their effects, our article data yielded strong to extremely strong evidence for the null hypothesis, whereas our noun data yielded extremely strong evidence for the alternative hypothesis (upper graphs). These results were obtained with the procedure described in DeLong et al. (no baseline correction), and with a 500 ms pre-word baseline correction (lower graphs), the procedure later described by DeLong and colleagues.

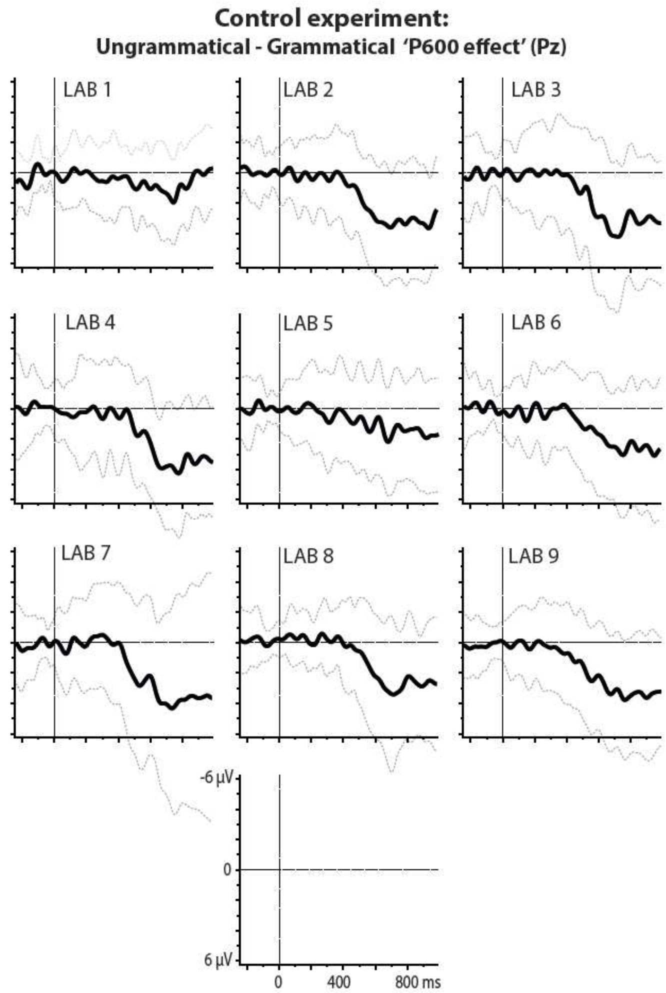
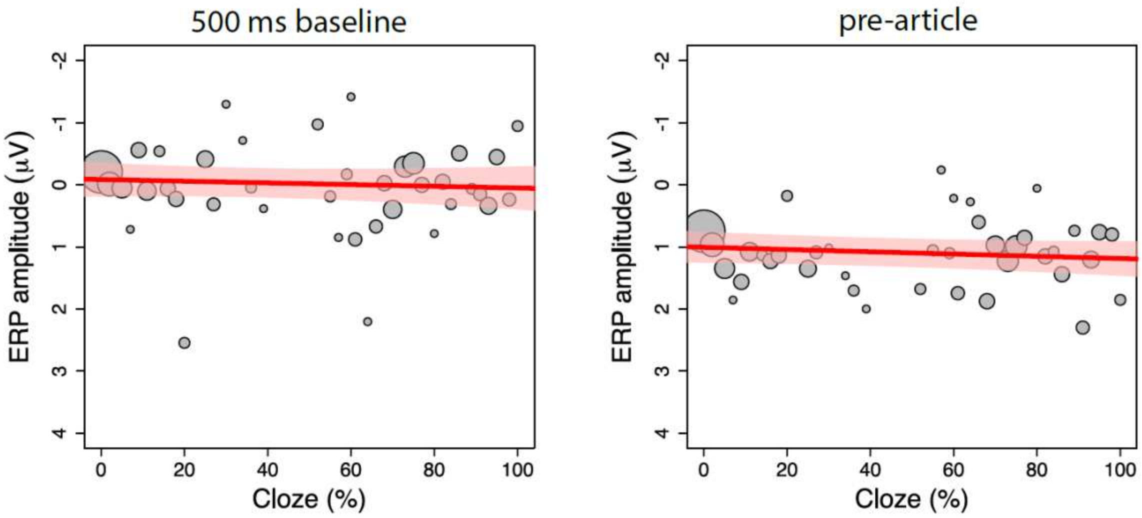
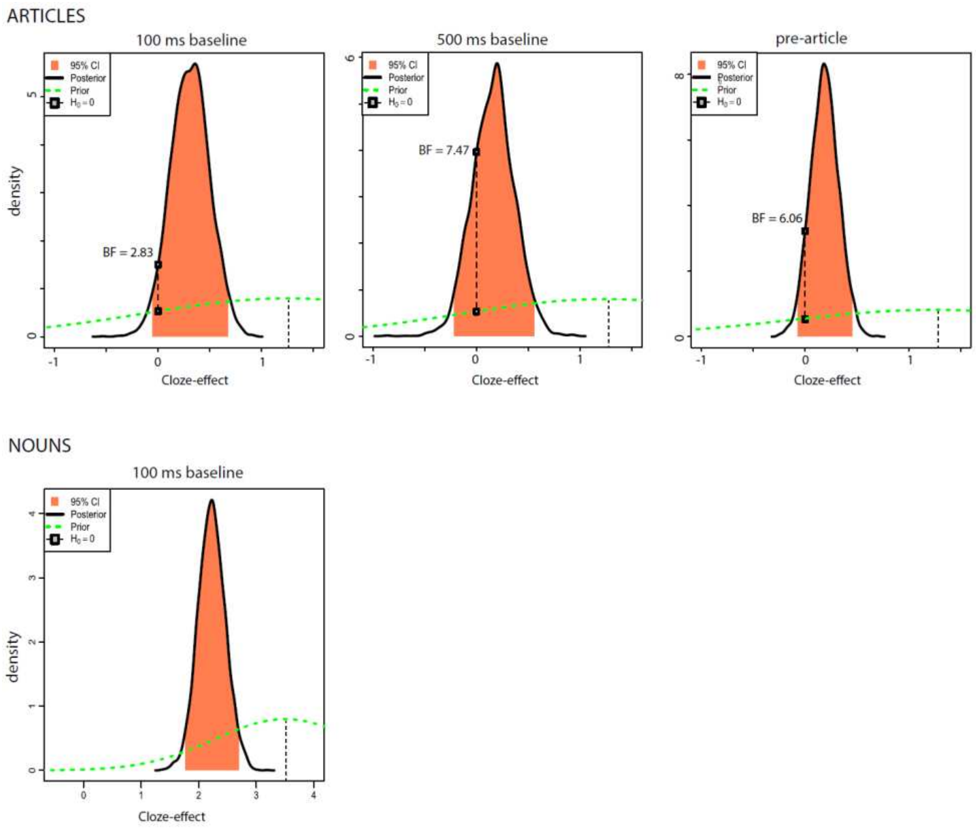


Figure 6. Control experiment. P600 effects at electrode Pz per lab associated with flouting of the English a/an rule. Plotted ERPs show the grand-average difference waveform and standard deviation for ERPs elicited by ungrammatical expressions ('an kite') minus those elicited by grammatical expressions ('a kite').



Supplementary Figure 1. **Exploratory single-trial analyses**: The relationship between cloze and ERP amplitude as illustrated by the mean ERP values per cloze value (number of observations reflected in circle size), along with the regression line and 95% confidence interval, from two exploratory analyses. We performed a test which used a longer baseline time windows (500 ms, left panel) to better control for pre-article voltage levels. This test reduced the initially observed effect of article-cloze, β = .14, CI [-.25, .53], $\chi^2(1) = 0.46$, p = .50). An analysis in the 500 to 100 ms time window before article-onset (right panel) revealed a non-significant effect of cloze that resembled the pattern observed after article-onset, β = .16, CI [-.07, .39], $\chi^2(1) = 1.82$, p = .18, shedding doubt on the conclusion that the observed results are due to the presentation of the articles.



Supplementary Figure 2. Results from exploratory Bayesian mixed-effects model analyses, represented by posterior distributions for the effect of cloze on ERP amplitudes in the N400 window. The x-axis shows cloze effect sizes (i.e., changes in microvolts associated with an increase from 0% cloze probability to 100% cloze probability). The black line indicates the posterior distribution of effects; higher values of the posterior density at a given effect size indicate higher probability that this is the true effect size in the population. The peak of the posterior distribution roughly corresponds to the point estimate of the effect size (the regression coefficient) fitted from the Bayesian mixed effect model, i.e., the most likely value of the true effect size. The middle 95% of the posterior distribution, shaded in pink, corresponds to a two-tailed 95% credible interval for the effect size—i.e., an interval that we can be 95% confident contains the true effect. The green dotted line indicates the prior distribution (i.e., our expectation about where the true effect would lie before the data were collected). For the articles, this prior is centred on 1.25μV, an approximation of the effect observed by Delong and colleagues (2005), and for the nouns it is centred on 3.5μV. The black connected dots illustrate the ratio between the posterior and prior distribution (i.e., the Bayes Factor) at the effect size of 0μV; for example, a Bayes Factor of 4 suggests we can be 4 times more certain that the true effect is zero after having conducted this experiment than before, or, in other words, that the data increased our confidence in the null effect of zero fourfold. We performed these analyses for each of the linear mixed-effects model analyses we performed. We note that in all the article-analyses, the posterior probability of the estimated effect being greater than zero is around 80 or 90%, although this is also true for the pre-stimulus variable, shedding doubt that the observed results are due to presentation of the articles. In none of our article-analyses did zero lie outside the obtained credible interval, whereas for the nouns, zero lay outside the credible interval. These results are consistent with a failure to replicate the size of the article-effect reported by DeLong et al. article-effect and successful replication of the noun-effect.