



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

How robust are prediction effects in language comprehension?

Citation for published version:

Ito, A, Martin, AE & Nieuwland, M 2017, 'How robust are prediction effects in language comprehension? Failure to replicate article-elicited N400 effects' *Language, Cognition and Neuroscience*, vol. 32, no. 8, pp. 954-965. DOI: 10.1080/23273798.2016.1242761

Digital Object Identifier (DOI):

[10.1080/23273798.2016.1242761](https://doi.org/10.1080/23273798.2016.1242761)

Link:

[Link to publication record in Edinburgh Research Explorer](#)

Document Version:

Peer reviewed version

Published In:

Language, Cognition and Neuroscience

Publisher Rights Statement:

This is an Accepted Manuscript of an article published by Taylor & Francis in *Language, Cognition and Neuroscience* on 18.10. 2018, available online:
<https://www.tandfonline.com/doi/full/10.1080/23273798.2016.1242761>

General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



**How robust are prediction effects in language comprehension? Failure
to replicate article-elicited N400 effects**

Aine Ito^a, Andrea E. Martin & Mante S. Nieuwland

*Department of Psychology
School of Philosophy, Psychology and Language Sciences
University of Edinburgh
Edinburgh, United Kingdom*

^a Corresponding author.

Aine Ito

Department of Psychology

School of Philosophy, Psychology and Language Sciences

University of Edinburgh

7 George Square

Edinburgh EH8 9JZ

Scotland, United Kingdom

Email: aine.ito@ed.ac.uk

Acknowledgements.

AEM was supported by a Future Research Leaders Grant from the Economic and Social Research Council of the United Kingdom [ES/K009095/1].

How robust are prediction effects in language comprehension? Failure to replicate article-elicited N400 effects

Current psycholinguistic theory proffers prediction as a central, explanatory mechanism in language processing (e.g., Pickering & Garrod, 2013). However, widely-replicated prediction effects may not mean that prediction is necessary in language processing (Huettig & Mani, 2016). As a case in point, Martin et al. (2013) reported ERP evidence for prediction in native speakers but not in non-native speakers. Articles mismatching an expected noun (e.g., “a” when people expect “*umbrella*”) elicited larger negativity in the N400 time window compared to articles matching the expected noun in native speakers, but not in non-native speakers. We attempted to replicate these findings in two experiments, but found no evidence for prediction irrespective of language nativeness. We argue that pre-activation of phonological form of upcoming nouns, as evidenced in article-elicited effects, may not be a robust phenomenon. We conclude that a view of prediction as a necessary computation in language comprehension must be re-evaluated.

Keywords: prediction; language comprehension; ERP; N400; bilingualism

Introduction

In current theories of language production and comprehension, prediction plays an outsized role as the mechanism by which language processing can occur quickly, incrementally, and rather effortlessly (e.g., Dell & Chang, 2014; Pickering & Garrod, 2013). While it is certainly the case that prediction effects in highly constrained sentential contexts have been widely reported, and even replicated, it is not clear that the existence of such effects warrant the conclusion that prediction is a necessary computation in language processing (see Huettig & Mani, 2016).

The debate about the functional role of prediction extends to current research on non-native language comprehension, where one of the central questions is whether, or under what circumstances, non-native speakers are able to successfully predict upcoming information like native speakers do (Foucart, Martin, Moreno, & Costa, 2014; Hopp, 2015; C. D. Martin et al., 2013). It is generally assumed that native speakers are able to pre-activate upcoming information based on the meaning they compute from linguistic input (Kutas, DeLong, & Smith, 2011). *Pre-activation*, a hypothesized incarnation of predictive processing, goes hand-in-hand with *incrementality* (i.e., the use of information as soon as it becomes available), the first being a possible mechanism by which rapid, ‘predictive’ computation of the context occurs, and the second being a processing heuristic that describes the computational architecture required to achieve language processing in real time (Altmann & Mirković, 2009). One reason that effective incremental use of information may be difficult to achieve for non-native speakers is that they read more slowly and more effortfully than native speakers (Clahsen & Felser, 2006; Havik, Roberts, Van Hout, Schreuder, &

Haverkort, 2009). As a result, non-native speakers may have fewer cognitive resources available to construct a rich mental representation of the context in an online fashion, rendering the prediction of upcoming information based on that context less likely (e.g., Grüter, Rohde, & Schafer, 2014; Ito, Martin, & Nieuwland, in press.; Kaan, Kirkham, & Wijnen, 2016). Consistent with this hypothesis, C. D. Martin et al. (2013) recently reported event-related potential (ERP) evidence that native speakers predict upcoming words but non-native speakers do not. By attempting to replicate the Martin et al. ERP results, the current study attempts to obtain additional support for differences in predictive processing in native and non-native speakers.

Martin et al.'s study itself was also based on another study, namely that of DeLong, Urbach, and Kutas (2005), who reported that native speakers predicted a specific upcoming word when the sentence was highly predictive. In DeLong et al., participants read predictive sentence contexts (e.g., "*The day was breezy so the boy went outside to fly...*"), one word at a time as is common in ERP research on written language comprehension, followed by an expected continuation (e.g., "*a kite*") or by an unexpected continuation (e.g., "*an airplane*"). Compared to unexpected nouns, expected nouns elicited smaller N400s, the ERP component associated with semantic processing costs (Kutas & Hillyard, 1980). The reduced semantic processing costs as suggested by the N400 effect could reflect the fact that participants pre-activated expected nouns before they appeared, leading to reduced processing costs, or that these nouns were simply easier to integrate into the sentence context. However, evidence for prediction was demonstrated by a similar N400 effect elicited by the preceding articles. Articles that preceded expected nouns elicited smaller N400s than articles that preceded expected nouns. DeLong et al. argued that these findings testify to pre-activation of the

nouns, with the article-elicited N400s being driven by whether the article is consistent or inconsistent with the prediction.

Martin et al. (2013) performed a similar study using native English speakers and Spanish-English bilinguals who had learned English after the age 8. In native speakers, unexpected articles elicited a negativity in the N400 time window¹ relative to expected articles, but no such difference was observed in non-native speakers. This was the case even though the reported offline cloze probability tests² suggested that the critical nouns were equally expected by native- and non-native speakers. Based on this finding, the authors claimed that non-native speakers did not pre-activate upcoming nouns like native speakers, instead they appeared to activate and integrate each word only after they encounter it. In both native- and non-native speakers, unexpected nouns elicited larger negativity in the N400 time window relative to expected nouns, although the difference was smaller and occurred later in non-native speakers than in native speakers. In native speakers, unexpected nouns additionally elicited post-N400 frontal positivity relative to expected nouns, but there was no such difference in non-native speakers. This frontal positivity effect was also reported by DeLong and colleagues in another study (DeLong, Urbach, Groppe, & Kutas, 2011) that used the same data as DeLong et al., and this effect was taken to reflect the extended processing consequences of a disconfirmed prediction. In the study by Martin and colleagues, this frontal positivity effect was absent in non-native speakers, which was taken as additional evidence that non-native speakers did not predict like native speakers.

While Martin et al. used a similar experimental rationale as DeLong et al. (2005), based on the *a/an* manipulation, the Martin et al. findings for native speakers are not a direct replication of the DeLong et al. findings. The two studies differ in several

important ways³, for example, the manner in which the sentences were presented.

Participants in DeLong et al. read each sentence word-by-word at a 500 ms Stimulus Onset Asynchrony (SOA), a standard procedure in ERP research on written language comprehension. Their critical nouns were always followed by subsequent words.

However, participants in Martin et al. first saw a sentence context as a chunk for as long as they wished, and then started the rest of the part, which was presented word-by-word at a 700 ms SOA.⁴ The critical noun was always the last word of the sentence. So participants had more time to read each word in Martin et al. We think that the procedure used by Martin et al., by focusing the attention of the participants on the final noun, may have been more likely to induce strategic anticipation than the procedure used by DeLong et al. In the current study, we use materials adapted from those of Martin et al., but our participants read them in the standard word by word presentation procedure used by DeLong et al.

In our own previous work, we also found evidence for prediction in native speakers (Ito, Corley, Pickering, Martin, & Nieuwland, 2016), but not in non-native speakers (Ito, Martin, & Nieuwland, in press). Importantly, these previous studies investigated predictive processing not by the *a/an* manipulation of DeLong et al. but by examining whether words that match the form or meaning of a highly expected word elicited reduced N400s compared to words that do not match form or meaning of an expected word (see also Federmeier & Kutas, 1999). In Ito et al. (2016), native speakers read predictive sentence contexts (e.g., “*The student is going to the library to borrow a...*”), followed by the predictable word (e.g., “*book*”), an implausible word that was related to the predictable word in form (e.g., “*hook*”) or meaning (e.g., “*page*”), or an implausible and unrelated word (e.g., “*sofa*”). All types of implausible words elicited

larger N400s than predictable words, but this N400 effect was reduced for form-related words and semantically related words relative to unrelated words. This indicates that processing of these words was facilitated as a result of information overlap with predictable words. Since the N400 reduction depended on cloze probability of predictable words, and could not be explained in terms of plausibility, the findings suggest that native speakers pre-activated both word form and meaning of predictable words. However, Ito et al. (in press) reported that, unlike the native speakers, non-native speakers did not show such a cloze-dependent N400 reduction.

Alongside the sentence materials described above, participants in Ito et al. (2016, submitted) also read materials with the *a/an* manipulation that were adapted from those of Martin et al. (2013). Through analysis of the ERP activity associated with those items, the current study attempts to replicate the findings of Martin et al. (2013). Our study counts as a conceptual, non-direct replication because our experiment differed from that of Martin et al. (2013) in several ways. Participants in our study read sentences with the *a/an* manipulation but also the sentences that manipulated whether words match the form or meaning of an expected word (Ito et al., 2016, in press), whereas participants in Martin et al. (2013) only read sentences with the *a/an* manipulation. Thus, unlike in Martin et al. (2013), our participants read both plausible and implausible sentences. We presented sentences word-by-word, instead of presenting a sentence context as a whole first and then the critical phrase as in Martin et al. By doing this, we keep the reading rate constant for each participant, and use the presentation procedure that is more similar to that of DeLong et al. and that is most common in ERP research on written language comprehension. However, to explore effects of presentation rate, we used two

different SOAs in separate experiments (500 ms in Experiment 1 and 700 ms in Experiment 2).⁵

Experiment 1

Methods

Participants

Twenty-three native English speakers (5 males and 18 females, age $M = 24$ years, $SD = 5$) and 21 Spanish-English late bilinguals (7 males and 14 females, age $M = 27$ years, $SD = 5$) participated in the experiment. All participants were right-handed and free from neurological or language disorders, and they were all tested at University of Edinburgh. Bilinguals reported their length of exposure to English ($M = 12.4$ years, $SD = 6.2$) and the age of acquisition ($M = 11.2$ years, $SD = 8.1$). We also asked bilinguals to rate their English proficiency on a 10-point scale. The average score for reading, writing, listening and speaking was 8.1 ($SD = .7$), which was similar to the self-rated proficiency reported in Martin et al. (2013) ($M = 7.6$, $SD = 1.0$ on a 10-point scale).

Stimuli

Critical stimuli were 64 predictive sentence contexts (e.g., “*As it's rainy, it's better to go out with...*”) which ended with either an expected noun phrase (e.g., “*an umbrella*”) or an unexpected noun phrase (e.g., “*a raincoat*”). We adapted stimuli from Martin et al. (2013), so that every item had only one sentence and had a similar number of words ($M = 11.9$ words, $SD = 1.4$). This final set was selected from 80 candidates based on cloze probability and plausibility pre-tests described below.

We conducted cloze tests using native English speakers and non-native English speakers whose native language was Spanish (native $N = 26$, non-native $N = 15$). The sentences were truncated before the article (e.g., “*As it's rainy, it's better to go out with...*”), and participants were instructed to complete the sentence fragment with the first plausible word that comes to mind. Sentences were excluded if the expected word did not have the highest cloze probability or if the cloze probability for the expected word was lower than 30%.

In a subsequent plausibility test with two counterbalanced lists, 24 native English speakers and 20 Spanish-English bilinguals evaluated the plausibility (1-completely implausible to 5-completely plausible) of each item in either the expected condition or the unexpected condition. These participants were recruited from similar population groups as those in the main ERP experiments. An item that had a plausibility rating lower than 3 was excluded.

In the final set, the mean native cloze probability was 67.3% ($SD = 18.7$) for expected nouns and 6.2% ($SD = 9.4$) for unexpected nouns, and the mean non-native cloze probability was 56.5% ($SD = 27.4$) for expected nouns and 9.7% ($SD = 15.0$) for unexpected nouns.⁶ The native cloze was similar to that in Martin et al. (69%), but the non-native cloze was lower than that in Martin et al. (65%). For articles, the mean native cloze probability was 75.1% ($SD = 17.9$) in the expected condition and 14.5% ($SD = 13.6$) in the unexpected condition, and the mean non-native cloze probability was 66.7% ($SD = 25.9$) in the expected condition and 21.9% ($SD = 18.6$) in the unexpected condition. Native speakers gave a mean plausibility of 4.6 ($SD = .2$) for the expected condition and 4.4 ($SD = .4$) for the unexpected condition, and non-native speakers gave 4.4 ($SD = .6$) for the expected condition and 4.5 ($SD = .4$) for the unexpected condition.

Native speakers rated the expected condition more plausible compared to the unexpected condition, $t(63) = 4.4, p < .001$ (paired t -test), although this difference was numerically small (mean difference of .2 on a 5-point scale). Non-native speakers rated both conditions similarly highly plausible, $p > .1$.

Procedure

The 64 sentences were divided into two counterbalanced lists so that each list contained the same number of expected and unexpected critical words. Because 30 expected nouns corresponded to “a” and 34 expected nouns to “an”, the number of *a/an* articles that appeared in each list was slightly unbalanced. The experimental stimuli included additional 160 filler sentences from Ito, Corley, Pickering, Martin, and Nieuwland (2016), 40 of which had predictable words and the other 120 had implausible words. These critical words were always followed by additional one word. Each participant thus saw a total of 104 plausible (72 expected and 32 unexpected) and 120 implausible sentences. The sentences were pseudo-randomised, such that participants did not see more than three successive sentences from the same condition.

During the EEG setup, the bilinguals completed a language background questionnaire. For the ERP experiment, participants silently read sentences presented visually word by word at 500 ms SOA (300 ms word duration, 200 ms inter-word interval; sentence-final words had a 600 ms duration). They were asked to minimize eye blinks and movements during the reading. A fixation cross followed each sentence, when the participants could move or blink, and they pressed a key to start the next sentence. Yes-No comprehension questions (which were not about the critical words)

appeared on 25% of the trials. The experiment was divided into five blocks, and the whole experiment took about 40 minutes.

Native speakers had a mean accuracy for comprehension questions of 96.2% ($SD = 3.6$; 6.1% of the data was excluded due to time-outs), and non-native speakers had a mean accuracy of 93.2% ($SD = 6.6$; 8.5% time-outs). Native speakers had 25 ($SD = 3$) artefact-free trials per condition on average and non-native speakers had 25 ($SD = 4$), with no difference between conditions. Participants who had less than 60% artefact-free trials were excluded from analyses (six participants).

Electroencephalogram (EEG) recording and data processing

The electroencephalogram (EEG) was recorded at a sample rate of 512 Hz and with 24-bit AD conversion using the Biosemi ActiveTwo system (BioSemi BV, Amsterdam, The Netherlands). This system's hardware is completely DC coupled and applies digital low pass filtering through its ADC's decimation filter (the hardware bandwidth limit), which has a 5th order sinc response with a -3 dB point at 1/5th of the sample rate (i.e., approximating a low-pass filter at 100 Hz). Data were recorded from 64 EEG, 4 EOG, and 2 mastoid electrodes using the standard 10/20 system. Offline, the EEG was re-referenced to the mastoid average⁷ and filtered further (0.01–30 Hz bandwidth filter with 12 dB slope plus 50 Hz Notch filter). Data was segmented into 1200 ms epochs (-200 to 1000 ms relative to critical word onset), corrected for eye-movements using the Gratton and Coles regression procedure as implemented in BrainVision Analyzer (Brain Products ©), baseline-corrected to -100 to 0 ms, automatically screened for movement- or electrode-artefacts (minimal/maximal allowed amplitude = -75/75 μ V), and averaged per condition per participant.

Statistical analysis

Following Martin et al. (2013), we computed mean ERP amplitudes per condition at three ROIs: Frontal (F3, Fz, F4, FC1, FCz, FC2), Central (C1, Cz, C2, CP1, CPz, CP2) and Parietal (P1, Pz, P2, PO3, POz, PO4), in the article N400 time window (250 – 400 ms), in the noun N400 time window (300 – 500 ms) and the noun post-N400 time window (500 – 900 ms). Effects of condition, language nativeness and scalp distribution were tested with a two (condition: expected, unexpected) by two (language group: native, non-native) by three (ROI: frontal, central, parietal) repeated-measures ANOVA. When appropriate, Greenhouse-Geisser corrections and corrected *F*-values are reported. Only statistical results with $p < .1$ are reported.

Results

Visual inspection of Figure 1 suggests that the expected and unexpected conditions did not differ in the article N400 time window, but differed in the noun N400 time window in native speakers, with unexpected nouns showing a larger (more negative) N400 ERP. This noun N400 effect extended into the post-N400 time window. There appears to be no effect of expectedness in any of the time windows in non-native speakers.

500 ms SOA

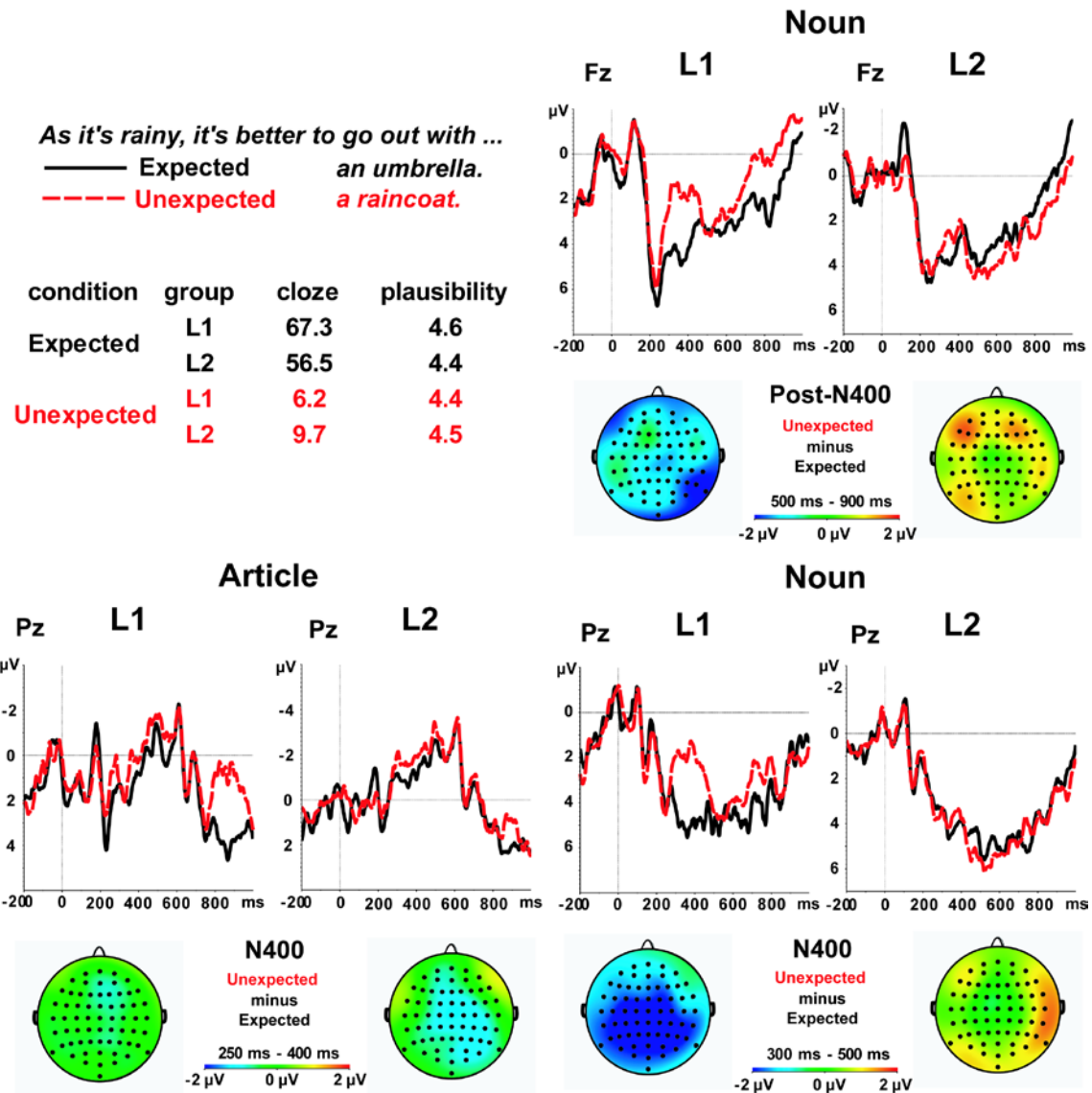


Figure 1. Results from Experiment 1. ERPs for expected and unexpected articles at Pz (bottom-left), for expected and unexpected nouns at Fz (top-right) and at Pz (bottom-right). Scalp distributions of the effect of expectedness (unexpected minus expected condition) are shown below each ERP plot in the N400 time window for articles (bottom-left) and for nouns (bottom-right), and in the post-N400 time window for nouns (top-right). Each panel shows results of native (L1) speakers on the left, and results of non-native (L2) speakers on the right.

Article N400 effects

Unexpected articles elicited numerically larger N400s ($M = -.46 \mu\text{V}$, $SD = 3.4$) than expected articles ($M = .25 \mu\text{V}$, $SD = 2.8$), but this difference was only marginally significant, $F(1, 42) = 3.6$, $p = .06$. There was no significant interaction involving condition, $ps > .1$. Since this statistical test is crucial to our conclusion, we also performed a linear-mixed effects model analysis using lme4 package (Bates, Maechler, Bolker, & Walker, 2014) in R (R Development Core Team, 2015). This approach generates more robust results because it simultaneously models variation at the participant-level and item-level (Baayen, Davidson, & Bates, 2008), and is also well suited to analyse data from non-native speakers where subject variability is relatively large (Cunnings, 2012). For this analysis, we exported N400 amplitude for each participant and for each trial in the same time window at the same sets of channels. The model evaluated N400 amplitude as predicted by condition (expected vs. unexpected), including random intercepts by participants and by items and random slopes of condition by participants and by items. The effect of condition was not significant, $|t| < 1$.⁸

Noun N400 effects

Unexpected nouns elicited larger N400s ($M = .3.0 \mu\text{V}$, $SD = 3.5$) than expected nouns ($M = 4.2 \mu\text{V}$, $SD = 2.8$), $F(1, 42) = 5.3$, $p < .05$. The ANOVA also revealed a significant interaction of condition by language group, $F(1, 42) = 6.6$, $p < .05$. Separate paired t -tests on each language group revealed that native speakers showed the N400 effect (unexpected $M = 2.1 \mu\text{V}$, $SD = 3.6$, expected $M = 4.4 \mu\text{V}$, $SD = 3.3$), $t(68) = 6.8$,

$p < .001$, but non-native speakers did not (unexpected $M = 4.0 \mu\text{V}$, $SD = 3.2$, expected $M = 3.9 \mu\text{V}$, $SD = 4.0$), $p > .1$.

Noun post-N400 frontal positivity

In the noun post-N400 time window, there was no difference between unexpected nouns ($M = 3.2 \mu\text{V}$, $SD = 3.8$) and expected nouns ($M = 3.5 \mu\text{V}$, $SD = 3.6$), $p > .1$. The ANOVA did not show any significant interaction involving condition either, $ps > .1$.

Discussion

We attempted to replicate the findings of Martin et al. (2013), who reported article-elicited N400 effects as evidence for predictive processing in native speakers (cf. DeLong et al. 2005) but no such effects for non-native speakers. In our study, neither native speakers nor non-native speakers showed a differential ERP effect at articles. Therefore, we did not replicate the effect that was previously taken as evidence for prediction (i.e., larger negativity in the N400 time window for unexpected articles relative to expected articles) in Martin et al. (2013) and in DeLong et al. (2005). Moreover, whereas native speakers showed an N400 effect for unexpected nouns relative to expected nouns, non-native speakers also did not show any difference for unexpected and expected nouns. Therefore, non-native speakers in our experiment appeared to be insensitive, at least in the initial stages of semantic processes reflected in N400 activity, to the predictability both of the article and of the noun.

Experiment 2

The failure to find an N400 effect of noun-predictability in non-native speakers in Experiment 1 could be because the reading rate of 500 ms SOA was too fast for them. We therefore performed the same experiment but with a slower SOA of 700 ms per word. This slower SOA was used in Martin et al. (2013), and is consistent with the use of slower SOAs in ERP studies on non-native language comprehension than on native language comprehension (e.g., Foucart et al., 2014; Tanner, McLaughlin, Herschensohn, & Osterhout, 2013; Tokowicz & MacWhinney, 2005). Non-native speakers may show more native-like processing when they have more time to comprehend each word. For example, in Ito et al. (in press), non-native speakers showed an LPC (P600) effect for words that were related in form to a predictable word, but only when words were presented at a 700 ms SOA and not at a 500 ms SOA. In contrast, native speakers showed the same effect at both SOAs (Ito et al., 2016). Since non-native speakers normally read more slowly than native speakers (Havik et al., 2009; Hopp, 2009), the longer 700 ms SOA is perhaps more likely to allow non-native speakers to read at a comfortable pace. Therefore, we hypothesised that non-native speakers may show an N400 effect for unexpected nouns similarly to native speakers in Experiment 2.

At a slower 700 ms SOA, native speakers may also show evidence for predictive processing. Several studies show that native speakers are also more likely to engage in predictive processing when reading more slowly (Ito et al., 2016; Wlotko & Federmeier, 2015). In Ito et al. (2016), native speakers showed evidence for prediction of word form at a 700 ms SOA, but not at a 500 ms SOA. This finding is relevant to the current study, which utilises the English *a/an* phonological rule. An N400 effect for unexpected

articles relative to expected articles would mean that participants pre-activated at least some aspect of word form (whether the expected word starts with a consonant or vowel). Therefore, Experiment 2 investigated whether native speakers (as well as non-native speakers) would show evidence of word form pre-activation at the slower SOA.

Methods

Participants

Twenty-three native English speakers (6 males and 17 females, age $M = 20$ years, $SD = 3$) and 23 Spanish-English late bilinguals (10 males and 13 females, age $M = 28$ years, $SD = 5$) participated in the experiment. Like in Experiment 1, all participants were right-handed and free from neurological or language disorders, and they were all tested at University of Edinburgh. Bilinguals' mean length of exposure to English was 15.8 years ($SD = 7.9$) and their mean age of acquisition was 10.5 years ($SD = 6.2$). The mean self-rated proficiency across the four measurements (reading, writing, listening and speaking) was 7.8 ($SD = 1.3$).

Stimuli and procedure

Stimuli and procedure were identical to Experiment 1, except that SOA was changed to 700 ms (500 ms word duration, 200 ms inter-word-interval; sentence final words had an 800 ms duration). Native speakers had a mean accuracy for comprehension questions of 94.8% ($SD = 4.1$; 4.0% of the data was excluded due to time-outs), and non-native speakers had a mean accuracy of 88.5% ($SD = 11.1$; 10.2% time-outs). Native speakers had 26 ($SD = 3$) artefact-free trials per condition on average

and non-native speakers had 27 ($SD = 3$), with no difference between conditions, except that there were more valid trials for unexpected nouns ($M = 28$) than for expected nouns ($M = 26$) in non-native speakers, $t(44) = -2.1, p < .05$. Participants who had less than 60% artefact-free trials were excluded from analyses (two participants).

Electroencephalogram (EEG) recording and data processing

EEG was recorded and processed in the same way as in Experiment 1.

Statistical analysis

The data were analysed as in Experiment 1.

Results

Like Experiment 1, visual inspection of Figure 2 suggests no effect of condition in the article N400 time window, but both native- and non-native speakers show N400 effects for unexpected nouns. In the post-N400 time window, native speakers but not non-native speakers show frontally distributed positivity for unexpected nouns.

700 ms SOA

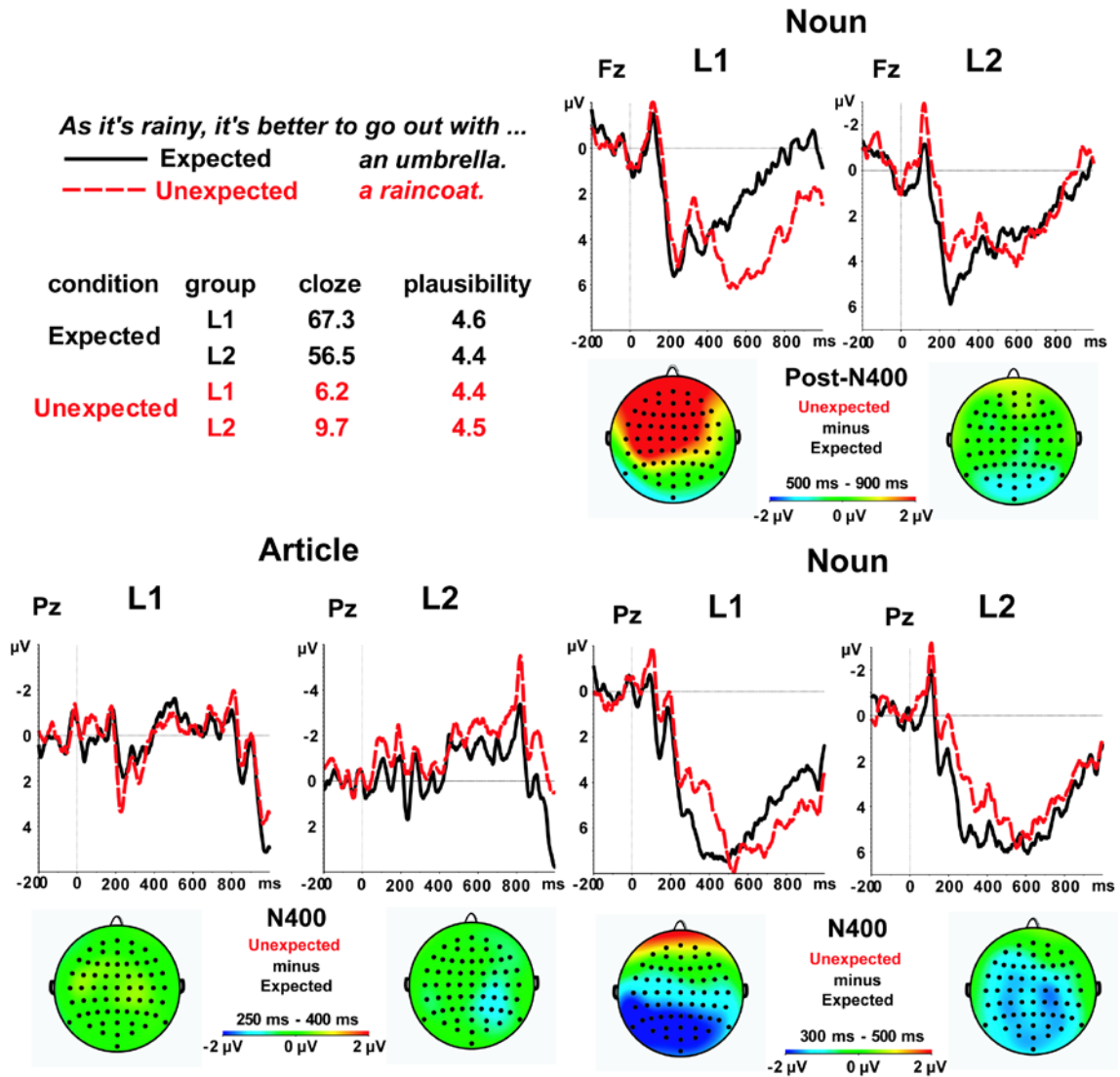


Figure 2. Results from Experiment 2. ERPs for expected and unexpected articles at Pz (bottom-left), for expected and unexpected nouns at Fz (top-right) and at Pz (bottom-right). Scalp distributions of the effect of expectedness (unexpected minus expected condition) are shown below each ERP plot in the N400 time window for articles (bottom-left) and for nouns (bottom-right), and in the post-N400 time window for nouns (top-right). Each panel shows results of native (L1) speakers on the left, and results of non-native (L2) speakers on the right.

Article N400 effects

N400s for unexpected articles ($M = -.17 \mu\text{V}$, $SD = 2.7$) and expected articles ($M = -.075 \mu\text{V}$, $SD = 2.5$) did not differ, $p > .1$. The ANOVA did not show any significant interaction involving condition either, $ps > .1$.

Noun N400 effects

Unexpected nouns elicited larger N400s ($M = 4.0 \mu\text{V}$, $SD = 3.4$) than expected nouns ($M = 5.3 \mu\text{V}$, $SD = 4.4$), $F(1, 44) = 8.2$, $p < .01$. The ANOVA additionally revealed a significant interaction of condition by ROI, $F(2, 88) = 3.4$, $p < .05$. We followed the interaction with paired t -tests on each ROI. The N400 effect was significant at central ROI (unexpected $M = 4.1 \mu\text{V}$, $SD = 3.6$, expected $M = 5.5 \mu\text{V}$, $SD = 4.5$), $t(45) = 2.8$, $p < .01$, and parietal ROI (unexpected $M = 4.6 \mu\text{V}$, $SD = 3.1$, expected $M = 6.4 \mu\text{V}$, $SD = 4.3$), $t(45) = 3.4$, $p < .01$, but not at frontal ROI (unexpected $M = 3.3 \mu\text{V}$, $SD = 3.5$, expected $M = 4.1 \mu\text{V}$, $SD = 4.2$), $p > .1$.

Noun post-N400 frontal positivity

The ANOVA on the noun post-N400 time window revealed a significant interaction of condition by language group, $F(1, 44) = 5.2$, $p < .05$, a significant interaction of condition by ROI, $F(1.4, 60.4) = 8.8$, $p < .01$, and a marginally significant three-way interaction of condition by language group by ROI, $F(2, 88) = 2.9$, $p = .06$. We followed these up with separate two (condition) by two (language group) ANOVAs on each ROI. At the frontal ROI, unexpected nouns elicited more positive ERPs ($M = 3.6 \mu\text{V}$, $SD = 4.1$) than expected nouns ($M = 2.2 \mu\text{V}$, $SD = 4.1$), $F(1, 44) = 6.7$, $p < .05$.

The ANOVA on the frontal ROI also showed a significant interaction of condition by language group, $F(1, 44) = 8.0, p < .01$. Follow-up paired t -tests on each language group showed that this frontal positivity effect was significant in native speakers (unexpected $M = 4.7 \mu\text{V}, SD = 4.4$, expected $M = 1.8 \mu\text{V}, SD = 4.7$), $t(22) = -3.5, p < .01$, but not in non-native speakers (unexpected $M = 2.5 \mu\text{V}, SD = 3.4$, expected $M = 2.6 \mu\text{V}, SD = 3.5$), $p > .1$. The ANOVA on the central ROI showed a significant interaction of condition by language group, $F(1, 44) = 4.9, p < .05$. Analogous follow-up t -tests showed a significant frontal positivity effect in native speakers (unexpected $M = 6.3 \mu\text{V}, SD = 5.0$, expected $M = 4.2 \mu\text{V}, SD = 5.4$), $t(22) = -2.5, p < .05$, but not in non-native speakers (unexpected $M = 3.6 \mu\text{V}, SD = 3.5$, expected $M = 4.0 \mu\text{V}, SD = 3.7$), $p > .1$. The ANOVA on the parietal ROI did not show any significant effect or interaction involving condition, $ps > .1$.

Between-experiment comparison: N400 and frontal positivity effects

We further tested whether the observed N400 and frontal positivity effects for unexpected nouns depended on SOA. The earlier analysis on noun N400 effects revealed that native speakers showed larger N400s for unexpected nouns than for expected nouns at both SOAs, whereas non-native speakers showed a similar N400 effect at 700 ms SOA only. We therefore carried out a two (condition) by two (language group) by two (SOA) ANOVA on the parietal ROI, where the N400 effect was strongest in the 700 ms SOA analysis. The ANOVA revealed a significant effect of condition, $F(1, 86) = 13.5, p < .001$, a significant interaction of condition by language group, $F(1, 86) = 5.5, p < .05$, but the three-way interaction was not significant, $p > .1$.

In the post-N400 time window, we found that native speakers but not non-native speakers showed the noun frontal positivity at 700 ms SOA, and both groups did not show such an effect at 500 ms SOA. This was confirmed by a significant three-way interaction of condition by language group by SOA, $F(1, 86) = 9.1, p < .01$, in the ANOVA on the frontal ROI.

Additional analysis using a global reference

We considered the potential effect of the reference procedure (average mastoid reference like DeLong et al., instead of the global average reference like Martin et al.) on our results. We therefore performed additional ANOVAs using the global reference, while keeping everything else identical to the current analyses. In Experiment 1, unexpected articles elicited numerically larger N400s ($M = -.016 \mu\text{V}, SD = 1.5$) than expected articles ($M = .22 \mu\text{V}, SD = 1.3$), but this difference was only marginally significant, $F(1, 41) = 3.9, p = .05$. There was no significant interaction involving condition, $ps > .1$, except that the interaction of condition by language group was marginally significant, $F(1, 41) = 2.9, p = .1$. Following the same procedure in the main analysis, we performed a linear mixed-effects model (random slopes were excluded because the model parallel to the main analysis did not converge). The effect of condition was not significant, $|t| < 1$. In Experiment 2, N400s for unexpected articles ($M = -.33 \mu\text{V}, SD = 1.6$) and expected articles ($M = -.44 \mu\text{V}, SD = 1.6$) did not differ, $p > .1$, and there was no significant interaction involving condition, $ps > .1$.

Discussion

Like Experiment 1, Experiment 2 did not find any differential ERP effect associated with expectedness of articles. However, unlike Experiment 1, Experiment 2 found larger N400s for unexpected nouns relative to expected nouns in both native- and non-native speakers. Experiment 2 additionally found that unexpected nouns elicited a larger frontal positivity in the post-N400 time window in native speakers. There was no such difference in non-native speakers. The post-N400 frontal positivity in native speakers replicated the effect reported in DeLong et al. (2011) and Martin et al. (2013). We discuss a possible interpretation of this effect in the General Discussion below.

General Discussion

We investigated whether native- and non-native speakers pre-activate word form information of predictable words (cf. DeLong et al., 2005; Martin et al., 2013). In previous studies by DeLong et al. and Martin et al., native speakers showed increased negativity in the N400 time window for unexpected articles relative to expected articles (e.g., *a* versus *an*, when *umbrella* was the most expected continuation of the sentence context “*As it's rainy, it's better to go out with...*”). Based on those findings, they concluded that native speakers predicted phonological word form of a predictable word (whether the word began with a consonant or with a vowel). In the current study, however, we did not replicate this finding. At neither the standard 500 ms SOA nor the slower 700 ms SOA, did we find a robust ERP difference between expected and unexpected articles. At the nouns, we replicated the N400 effect for unexpected nouns relative to expected nouns at both SOAs in native speakers and at 700 ms SOA in non-

native speakers. In native speakers, this N400 effect was followed by a frontal positivity at 700 ms SOA, while non-native speakers showed no such effect. This finding was also consistent with DeLong et al. and Martin et al.

No evidence for prediction: failure to replicate article-elicited N400 effects

Unlike DeLong et al. and Martin et al., the current study did not observe evidence for prediction. Although null results can never ‘prove a negative’ (i.e., that our participants did not predict upcoming words), we take our findings to suggest that lexical prediction does not always occur. By extension, we think that the article-elicited N400 effect associated with prediction may be obtained only under specific circumstances. As we discuss below, we think that sentence reading rate, the sentence presentation procedure, and the presence of filler sentences might bear upon predictive processing and therefore on the occurrence of the article N400-effect.

Our study differed from Martin et al. (2013) in several respects, and therefore constitutes an attempt at conceptual replication rather than exact replication. We emphasise that although we used a different EEG-channel reference procedure than Martin et al. (mastoid average reference instead of global average reference), we performed additional analyses that ruled out an explanation based on this difference. In our view, a more important difference was the presentation procedure. Our study used a word-by-word presentation instead of presenting a sentence context as a single chunk like in Martin et al. The procedure used by Martin et al. may have focused the attention of participants to the critical article-noun combinations, possibly causing participants to strategically expect certain words. However, our presentation procedure was identical to

DeLong et al., who did report article-elicited N400 effects, and, therefore, an explanation solely in terms of the presentation manner difference does not suffice.

Another important difference was that our participants saw 160 filler sentences intermixed with the 64 critical sentences, while DeLong et al. or Martin et al. did not use any filler sentences. The use of filler sentences is common practice in ERP studies on language comprehension, in particular because it can make the manipulation of interest less salient, thereby reducing the likelihood that participants pay specific attention to the manipulation of interest and that observed effects arise from this type of (implicit) task-demand, even when no explicit judgment task is imposed. This distraction by filler materials is particularly strong if the filler sentences contain strong and salient semantic or syntactic anomalies, as was the case in our study. When not seeing any filler sentences, participants in the studies by DeLong et al. and Martin et al. may have realised that the article-noun combination that appeared late in each sentence was always either highly predictable or relatively unpredictable. This may have caused participants to pay extra attention to the *a/an* manipulation, thereby inadvertently encouraging participants to engage in predictive processing. In the Martin et al. study, this implicit task demand may have been exacerbated by the presentation procedure which isolated and thereby emphasized the critical article-noun combination. If our explanation is correct, this would mean that article-elicited N400 effects are of *limited external validity* (i.e., hard to replicate in circumstances that differ from the original studies), which would suggest that preactivation of phonological word form is less common in natural language contexts than some theories assume (e.g., Dell & Chang, 2014; Pickering & Garrod, 2013).

While article-elicited prediction effects in the context of attention-grabbing filler materials could be taken as strong evidence that people routinely predict upcoming words, absence of prediction effects in the context of such filler materials could be taken to suggest that implausible or ungrammatical sentences interfere with regular comprehension and therefore with predictive processing. It is an empirical question whether we will observe article-elicited N400 effects if we use only plausible and grammatical fillers (or perhaps no fillers at all).

It is possible that article-elicited N400 effects are hard to replicate because such effects require comprehenders to pre-activate the actual word form information of the upcoming noun, rather than only its semantic features. In our previous work, we found evidence for word form prediction, in a design that did not involve the *a/an* manipulation, only in very high (94%) cloze sentences and only at a slower 700 ms SOA (Ito et al., 2016). Whether or not comprehenders pre-activate the word form of upcoming nouns may depend on the degree to which the article is a reliable cue to the noun. While Martin et al. and DeLong et al. did not discuss this issue, we think that cue-reliability may be an important factor in experiments without fillers, where each sentence contains an article that reliably confirms or disconfirms the sentence-final expected noun. In such an experimental setting, participants' realisation about this pattern might have boosted their sensitivity to the articles. However, this may not be the case in experiments where articles do not always occur, or when they do not reliably cue the upcoming noun because they match an adjective instead (e.g., *As it's rainy, it's better to go out with a big umbrella*). In our view, in natural language settings, articles may not be very reliable cues to upcoming nouns, which means that pre-activation of

word form may not be a common phenomenon (for similar arguments, Huettig & Mani, 2016; Huettig, 2015).

We emphasise that since the first observation of the article-N400 effect in native speakers by DeLong et al. (2005), the only published claim of replication of this effect, to our knowledge, has been Martin et al. (2013). Moreover, several crucial differences between these studies, as discussed previously in this paper, suggest that the Martin et al. findings may not have actually replicated the original DeLong et al. findings. These observations, along with the failure to replicate the article-effect in the current study suggests that article-elicited N400 effects may not have high external validity. That is, these effects may be obtained in specific, prediction-encouraging experimental settings, but they are perhaps not representative of how people understand language in natural settings (cf. Huettig, 2015).

Post-N400 positive ERP effect

In the post-N400 window, we only observed a robust effect of expectancy in native speakers at the slower SOA, in the form of a frontal positivity for unexpected nouns relative to expected nouns. Previous studies suggested that such effects reflect increased semantic integration difficulty after a prediction has been disconfirmed (DeLong et al., 2011; Federmeier, Wlotko, De Ochoa-Dewald, & Kutas, 2007; Martin et al., 2013; Thornhill & Van Petten, 2012; Van Petten & Luka, 2012). However, this interpretation does not straightforwardly explain the current data, since we did not find any evidence for prediction. A more parsimonious interpretation of the frontal positivity effect is that it reflects increased integration difficulty for unexpected nouns relative to expected nouns that is not the result of disconfirmed prediction. After all, expected

words have a better fit to the context than unexpected words, i.e. they are more plausible in relation to the context and they may be more semantically related to (or associated with) words in the context. This can render unexpected words, even when they are plausible, more difficult to integrate into the context compared to expected words. Therefore, we argue that the frontal positivity on its own does not demonstrate prediction.

Conclusion

The current study failed to replicate the findings of Martin et al. (2013) and DeLong et al. (2005), namely the elicitation of an N400 effect by unexpected articles compared to expected articles in native speakers. While null results cannot prove a negative (i.e. that participants did not predict), the current study failed to replicate this prediction effect at both standard and slower reading rates. We conclude that article-elicited N400 effects associated with prediction may have low external validity, and may be limited to specific experimental designs (see also Huettig, 2015), in particular experimental designs in which participants see a prediction-confirming or -disconfirming article in each sentence. Based on these points, we conclude that such prediction effects are perhaps not in fact representative of how people comprehend language in natural settings. Thus, in concord with Huettig and Mani (2016), we conclude that prediction may not be a necessary computation for language processing to occur, and that, minimally, its role in mechanistic theories of language processing must be carefully evaluated. In some processing architectures (see A. E. Martin, 2016), prediction effects would fall out naturally when the right confluences of cues and internal priors were activated, giving rise to an architectural form of predictive coding

that, while not necessary for comprehension to occur, may sometimes facilitate processing without the postulation of a separate, active prediction mechanism in the system. Regardless of this debate about processing architecture, however, our findings also highlight the need for further replication attempts on predictive processing, both direct and conceptual replications. There is no good reason to assume that the circumstances that gave rise to the replication crisis in psychology (Pashler & Wagenmakers, 2012) do not apply to psycholinguistics.

Acknowledgements

AEM was supported by a Future Research Leaders Grant from the Economic and Social Research Council of the United Kingdom [ES/K009095/1].

References

- Altmann, G. T. M., & Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science*, 33(4), 583–609.
- Baayen, H. R., 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.
- Bates, D. M., Maechler, M., Bolker, B. M., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4.
- Clahsen, H., & Felser, C. (2006). How native-like is non-native language processing? *Trends in Cognitive Sciences*, 10(12), 564–570.

- Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research*, 28(3), 369–382.
- Dell, G. S., & Chang, F. (2014). The P-chain: relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 369(1634), 20120394.
- DeLong, K. A., Urbach, T. P., Groppe, D. M., & Kutas, M. (2011). Overlapping dual ERP responses to low cloze probability sentence continuations. *Psychophysiology*, 48(9), 1203–1207.
- DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, 8(8), 1117–21.
- 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.
- Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research*, 1146, 75–84.
- Foucart, A., Martin, C. D., Moreno, E. M., & Costa, A. (2014). Can bilinguals see it coming? Word anticipation in L2 sentence reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(5), 1461–1469.
- Grüter, T., Rohde, H., & Schafer, A. J. (2014). The role of discourse-level expectations

- in non-native speakers' referential choices. In *Proceedings of the Annual Boston University Conference on Language Development* (Vol. 1, pp. 179–191).
- Havik, E., Roberts, L., Van Hout, R., Schreuder, R., & Haverkort, M. (2009). Processing subject-object ambiguities in the L2: A self-paced reading study with German L2 learners of Dutch. *Language Learning*, 59(1), 73–112.
- Hopp, H. (2009). The syntax–discourse interface in near-native L2 acquisition: Off-line and on-line performance. *Bilingualism: Language and Cognition*, 12(4), 463–483.
- Hopp, H. (2015). Semantics and morphosyntax in predictive L2 processing. *International Review of Applied Linguistics in Language Teaching*, 53(3), 277–306.
- Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research*, 1626, 118–135.
- Huettig, F., & Mani, N. (2016). Is prediction necessary to understand language? Probably not. *Language, Cognition, and Neuroscience*, 31(1), 19–31.
- 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.
- Ito, A., Martin, A. E., & Nieuwland, M. S. (in press). On predicting form and meaning in a second language. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.
- Kaan, E., Kirkham, J., & Wijnen, F. (2016). Prediction and integration in native and

second-language processing of elliptical structures. *Bilingualism: Language and Cognition*, 1–18.

Kutas, M., DeLong, K. A., & Smith, N. J. (2011). A look around at what lies ahead: Prediction and predictability in language processing. In *Predictions in the Brain: Using Our Past to Generate a Future* (pp. 190–207).

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

Martin, A. E. (2016). Language processing as cue integration: Grounding the psychology of language in perception and neurophysiology. *Frontiers in Psychology*, 7.

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.

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science a crisis of confidence? *Psychological Science*, 7(6), 528–530.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *The Behavioral and Brain Sciences*, 36(4), 329–47.

R Development Core Team. (2015). A language and environment for statistical computing. R Foundation for Statistical Computing.

- Tanner, D., McLaughlin, J., Herschensohn, J., & Osterhout, L. (2013). Individual differences reveal stages of L2 grammatical acquisition: ERP evidence. *Bilingualism: Language and Cognition*, *16*(2), 367–382.
- Thornhill, D. E., & Van Petten, C. (2012). Lexical versus conceptual anticipation during sentence processing: Frontal positivity and N400 ERP components. *International Journal of Psychophysiology*, *83*(3), 382–392.
- Tokowicz, N., & MacWhinney, B. (2005). Implicit and explicit measures of sensitivity to violations in second language grammar: An event-related potential investigation. *Studies in Second Language Acquisition*, *27*(2), 173–204.
- Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*(2), 176–190.
- 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.

Footnotes

¹ Martin et al. argued that the article effect constituted an N400 effect, but their effect did not have the posterior distribution typically observed for N400 effects. At parietal channels, where N400 differences are usually maximal, no difference was observed. Moreover, their N400 distribution was distorted by the reference procedure, because they used a common average reference rather than the average-mastoid procedure typically used in N400 research. Had they used the average-mastoid procedure, they could have observed a larger N400 for expected articles than for unexpected articles, which would be inconsistent with their interpretation.

² Cloze probability of a word is established in a cloze test as the proportion of participants who used that word to complete a given sentence fragment. It is commonly used as a proxy for predictability/expectedness.

³ There are also differences in terms of data analysis. For example, whereas DeLong et al. showed graded effects of predictability by reporting correlation analysis of the N400 waveform and the cloze value of the article, Martin et al. analysed their data using a factorial design with the factor expected/unexpected. In addition, DeLong et al. used a mastoid reference and filtered the data offline using 0.2 – 15 Hz filter, while Martin et al. used a global reference and filtered the data offline using a 30 Hz low pass filter. We will discuss further these differences in the results section when relevant to our own results.

⁴ Martin et al. reported that the SOA in their study was 500 ms, but it was actually 700 ms (as confirmed in a personal communication by Martin).

⁵ We initially used 500 ms SOA as was reported in Martin et al., but also used 700 ms SOA as this was the actual SOA used in their study.

⁶ Following Martin et al., we also conducted another cloze test with different participants (native $N = 24$, non-native $N = 12$), in which the sentences were truncated before the noun and always presented with an unexpected article (e.g., “*As it's rainy, it's better to go out with a...*”). Participants were reminded that sentences always ended with “a” or “an”, and they were instructed to take this, as well as the sentence content, into account when choosing nouns that fit in the context. In this cloze test, the mean native cloze probability was 16.8% ($SD = 8.6$) for expected nouns and 31.7% ($SD = 23.5$) for unexpected nouns, and the mean non-native cloze probability was 12.6% ($SD = 10.0$) for expected nouns and 30.5% ($SD = 29.7$) for unexpected nouns. In Martin et al., the mean non-native cloze probability was 3.5% for expected nouns and 37.4% for unexpected nouns. The non-native cloze probability for expected nouns was higher in our study, but the pattern of results was similar.

⁷ Martin et al. used a global reference instead of a mastoid reference. The global reference procedure is uncommon and sub-optimal in N400 research because the N400 can be broadly distributed across the scalp, in which case subtracting the average of all channels may diminish the observed effects. The global reference procedure may also lead to very different scalp distribution effects compared to the mastoid reference

procedure. In order to explore potential effects of the different reference, we also report results after using the global reference.

⁸ We followed a reviewer's suggestion and explored whether we would obtain an article-elicited N400 effect in other time windows. This additional analysis used the same ANOVAs run successive 50 ms time bins from 200 ms until 500 ms (200-250, 250-300, and so on), with a Bonferroni correction for the multiple comparisons. None of the time bins showed any significant effect of expectedness, $p > .1$.