Distinguishing integration and prediction accounts of ERP N400 modulations in language processing through experimental design

Francesco Mantegna[1], Florian Hintz[2], Markus Ostarek[2], Phillip M. Alday[2], & Falk Huettig[2,3]

[1]University of Trento, Italy

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

[3]Centre for Language Studies, Radboud University Nijmegen, The Netherlands

--- in press, *Neuropsychologia* ---

Corresponding author: Francesco Mantegna

Email: fmantegna93@gmail.com

Author note: All materials, data and analysis scripts are available on the Open Science Framework (OSF) at the following link: https://osf.io/rp4jy/?view_only=337e717ea5df489b80d5c6d96fd9e1d0

Abstract

Many theoretical accounts of prediction in language processing are based to a substantial amount on experimental evidence from electrophysiological studies measuring N400 target word modulations. A drawback of most of these studies is that lexical prediction accounts cannot be distinguished conclusively from (non-prediction) lexical integration ('bottom-up' activation) accounts. Here we explored whether it is possible to distinguish integration and prediction accounts of ERP N400 modulations in language processing through experimental design. By employing rhyming sentence completions, we kept the ease of integration constant across conditions that differed in word predictability only. This experimental design allowed us to attribute N400 target word effects across conditions to predictive language processing. We close by discussing recommendations for future electrophysiological studies on prediction in language.

Keywords: anticipation, ERPs, integration, N400, prediction

Introduction

Prediction of upcoming input is thought to be a main characteristic of language processing (e.g. Altmann & Mirkovic, 2009; Dell & Chang, 2014; Federmeier, 2007; Ferreira & Chantavarin, 2018; Pickering & Gambi, 2018; Hale, 2001; Hickok, 2012; Huettig 2015; Kuperberg & Jaeger, 2016; Levy, 2008; Norris, McQueen, & Cutler, 2016; Pickering & Garrod, 2013; Van Petten & Luka, 2012). One of the main pillars of experimental support for this notion comes from studies that have attempted to measure electrophysiological markers of prediction when participants read or listened to sentences ending in highly predictable words. The N400, a negative-going and centro-parietally distributed component of the ERP occurring approximately 400ms after (target) word onset, has been frequently interpreted as indexing prediction of the word (or the semantic representations and/or the phonological form of the predicted word, see Kutas & Federmeier, 2011; Nieuwland, 2019; Van Petten & Luka, 2012; for review). A major difficulty for interpreting N400 effects in language processing however is that it has been difficult to establish whether N400 target word modulations conclusively reflect prediction rather than (at least partly) ease of integration. In the present exploratory[1] study, we attempted to distinguish lexical prediction (i.e. 'top-down' activation) from lexical integration (i.e. 'bottom-up' activation) accounts of ERP N400 modulations in language processing.

Background

In a classic study, Kutas and Hillyard (1980) observed that when a word is semantically incongruent with preceding sentential context (e.g. "he spread the warm bread with socks") it elicits a larger N400 amplitude than when it is congruent (e.g. "he spread the warm bread with butter"). In accordance with this finding, N400 amplitude modulations have been traditionally interpreted as an index of violation of semantic expectations (Kutas & Hillyard, 1980; 1984). More recently, a number of studies

1 We use the term exploratory here in a statistical sense, i.e. future studies could usefully be confirmatory and pre-register empirical work testing the arguments we outline in the present paper.

suggested that N400 amplitude modulations may not be limited to semantic violations. Some evidence compatible with orthographic word form pre-activation comes from a study by Deacon et al. (2004). They observed that: (i) derivational pseudowords (i.e. derived from related root words) elicited N400 semantic priming effects similar to those obtained for words (indicating semantic activation of the root words); (ii) N400 repetition effects are seen even for pseudowords with little resemblance to known words. Similarly, Laszlo and Federmeier (2011) found influences of orthographic neighbors' number and frequency on the N400 amplitude in a regression analysis on single trial ERPs. Finally, there is also evidence compatible with pre-activation of a word's phonological form. Praamstra and Stegeman (1993) investigated the responsiveness of the N400 to phonological variables in a rhyme priming paradigm. They presented participants with rhyming and non-rhyming word-pairs and observed that non-rhyming words elicited larger N400 responses than rhyming words. Some electrophysiological studies have attempted to dissociate the influences of phonological expectations (e.g. ``the gambler had a streak of bad luggage'', expected phoneme 'lu-', semantically anomalous) from those of semantic expectations (e.g. ``Don caught the ball with his glove'', non-predicted phoneme, semantically appropriate). These studies observed distinct neural signatures for phonological mismatch (phonological mismatch negativity/N200) and semantic mismatch (N400 modulation) respectively (Connolly & Phillips, 1994; Van Den Brink, Brown & Hagoort, 2001).

All of the aforementioned studies have a crucial interpretation problem however: they measured the electrophysiological marker of anticipation (a reduced N400) during but not before the target word. It is possible that these studies measured ease of integration rather than prediction and that therefore readers (or listeners) may not have predicted proactively but instead integrated the bottom-up activated word (meaning) with its context post-lexically (cf. Baggio & Hagoort, 2011; Huettig, 2015; Huettig & Mani, 2016; Ito, Martin, & Nieuwland, 2017; Lau, Phillips, & Poeppel, 2008). A few electrophysiological studies have overcome this crucial interpretation problem by using a clever experimental manipulation measuring electrophysiological correlates of prediction before

the target word (De Long et al., 2005; Van Berkum et al., 2005; Wicha et al., 2004). De Long et al. (2005) for instance found N400 modulations on indefinite articles before the target word making use of the English language phonological rule that *a* is the indefinite article before consonant-initial words and *an* before vowel-initial words. A recent large multi-lab replication effort (N =334) however failed to replicate the crucial electrophysiological correlate of prediction before the target word (Nieuwland et al., 2018; a similar case is a failed replication of Otten & Van Berkum, 2009, see Kochari & Flecken, 2018). In short, there are some important uncertainties about the extent to which N400 (and other) ERP effects in the literature can be taken as reflecting anticipation of upcoming language input; however, there seems to be some consensus that the N400 at least partly reflects pre-activation (cf. "spreading activation" accounts such as Kutas & Federmeier 2011). Nevertheless, it remains the case that the design of almost all ERP studies investigating predictive language processing does not allow the strong conclusion that the N400 is an 'uncontaminated' electrophysiological marker of prediction. This is of course highly problematic, if we cannot be sure *to what extent* N400 modulations reflect prediction, then any conclusions about prediction in language processing drawn from such studies are at best considered unsafe. A recent study (Nieuwland et al., in press) presented some experimental evidence that prediction and integration processes may have distinct N400 spatiotemporal profiles. Here we chose a different approach, we attempted to distinguish integration and prediction accounts of ERP N400 modulations in language processing through experimental design.

Rhyme processing

Participants were presented with rhyming sentence completions. In the psycholinguistics literature, facilitation of rhyme processing has been well studied in both behavioral and electrophysiological rhyme priming experiments (Shulman, Hornak, & Sanders, 1978; Hillinger, 1980; Donnenwerth-Nolan, Tanenhaus, & Seidenberg, 1981; Rugg & Barrett, 1987; Praamstra, Meyer, & Levelt, 1994). In sentential contexts, processing facilitation due to phonological features of rhyming words has been studied to a lesser extent. In a behavioral experiment, Rapp and Samuel (2002) investigated whether

surface properties (i.e. phonology, stress patterns) can bias lexical selection in sentential context. Participants were asked to complete sentences like ``He'd gone to deposit his _____ and nearly broke his _____''. The authors observed faster response latencies for phonologically similar completions, such as ``neck'', preceded by rhyming primes (e.g. ``check''), as compared to phonologically dissimilar completions, such as ``ankle'', preceded by non-rhyming primes (e.g. ``payment''). Van Petten et al. (1999) investigated the differences in the N400 time window between plain incongruous words - semantically anomalous sentence completions with no phonological relationship to congruous completions - and rhyming words - semantically anomalous sentence completions sharing final phonemes with congruous completions (e.g. MUFFINS when DOLPHINS was expected). They observed no rhyming effect, i.e. the ERPs elicited by rhyming words did not differ from those elicited by incongruous words. N400 amplitude modulation elicited by rhyming versus non-rhyming words in contextually non-biasing sentential context, to the best of our knowledge, has not been directly investigated.

The current study

In order to distinguish lexical prediction from lexical integration accounts of ERP N400 modulations, it is necessary to define what we mean by prediction and integration. We consider prediction to be the pre-activation of upcoming words or representations (e.g. semantic, phonological) ahead of time. Integration, in contrast, we define as the combination of incoming words and representations (e.g. semantic, phonological) into a higher order (e.g. sentential) representation in absence of such pre-activation. Note that context can modulate both prediction and integration. Context and prediction are straightforwardly related because context can pre-activate upcoming words (and representations). Context can also affect integration because even without pre-activation a word may be more difficult to integrate with the preceding context *after* it has become activated (e.g. on reading the word) in a 'bottom-up' fashion. It is important to stress here that prediction and integration are necessarily related (e.g., predicted words can be more easily integrated than non-predicted words), the question

we address here is whether prediction and integration can be dissociated through experimental design to advance our understanding of the mechanisms of language comprehension.

In the present study, we used rhyming sentence completions as target words (e.g. the Dutch words *hart* and *start*, in English heart and start). Importantly, rhyme overlap in the critical conditions was the same. The rationale was that, in the absence of a biasing sentential context and keeping plausibility of sentence completion constant, the words *hart* and *start* should be equally easy to integrate. In other words, we designed the experiment in such a way that conditions did not differ according to the integration account; there is no reason why *hart* should be easier to integrate than *start* in absence of any pre-activation or any other potential confound such as word frequency or plausibility. The crucial difference across conditions was the lexical predictability of the rhyme word. Thus, the rationale was that by keeping the ease of integration constant across conditions, any difference in N400 effects across conditions we can attribute to lexical prediction.

In line with previous work, we used cloze probability as a proxy for lexical predictability. We designed a revised version of the cloze probability test (Taylor, 1953; Kutas & Hillyard, 1984) with a specific focus on rhymes in order to select 'lexically predictable' and 'lexically unpredictable' rhyming words for each item. Participants in pre-tests were asked to provide the rhyming word in the second sentence that first came to their mind given the final word in the first sentence (e.g. "Sven was niet bekend met de term kwart, Fleur dacht te weinig aan haar _____", "Sven did not know the term quarter, Fleur didn't think enough about her _____''). Sentence context was constructed so that it was semantically low-constraining (the semantic context did not point unequivocally to a specific rhyme completion). We grouped participants' responses in two separate classes: rhyming words which reached considerably high agreement (i.e. high-cloze probability, 'hart' in the example) and zero to low agreement (i.e. low-cloze probability, 'start', in the example). Based on pre-test results, we selected, for each sentence pair, a non-rhyming word that matched 'lexically predictable' and 'lexically unpredictable' words in terms of word frequency, phonological neighborhood density,

concreteness and semantic distance. In another pre-test, participants were asked to rate the plausibility of all sentence completions.

Experimental predictions

Given that most sentence final words were selected to rhyme we expect participants to build up an expectation to hear rhyming sentence completions. In the first condition (the congruent condition), if participants predict they will likely predict (according to our pre-test) that the final word (on hearing "Fleur dacht te weinig aan haar") would be *hart*. In the second condition (the intermediate condition), even if participants predict, they are very unlikely to predict (according to our pre-test) that the final word (on hearing "Fleur dacht te weinig aan haar") would be *start*. Similarly, in the third condition (the incongruent condition) participants cannot predict that the final word (on hearing "Fleur dacht te weinig aan haar") would be *vent* (a completely unrelated word). Thus, in both intermediate condition and incongruent condition, if participants predict, the prediction will be disconfirmed. Crucially, the completion *start* (in the intermediate condition) is *only* a mismatch if people *predict hart* because both *start* and *hart* should be equally easy to integrate in the semantically low-constraining 'poem-like' context of the sentences encountered. In other words, there is nothing 'wrong' with a *start* completion of the sentence, it is not semantically anomalous, it is grammatical and plausible and it rhymes (it is just not lexically predicted). According to this logic, differences in the target word N400 between congruent and intermediate conditions would therefore reflect prediction of lexical content rather than ease of integration of bottom-up encountered input.

Method

Participants

Thirty-one participants (mean age = 22.06 years old, range 18–30 years; 6 male) took part in the EEG experiment as paid volunteers. They were all right-handed native Dutch speakers. Participants had no history of psychiatric or neurological disorders, and they reported normal hearing. Informed consent

was obtained from all participants. Data collected from one participant were excluded from the analyses because of poor EEG recording quality (see below). The final set of participants thus consisted of 30 participants (mean age = 22.13 years old, range 18–30 years; 6 male).

Thirty participants took part in the first online experiment and another forty-five participants took part in the second online experiment. All participants were paid for their participation and were contacted via the MPI database. Participants involved in the online experiments were different from the ones who took part in the EEG experiment. Ethical approval to conduct the study was provided by the ethics board of the Faculty of Social Sciences at Radboud University.

Materials

Pre-test: rhyme cloze probability estimates

We ran an online experiment on an open source survey platform (LimeSurvey, GmbH, Hamburg, Germany. URL http://www.limesurvey.org) to collect rhyme cloze probability estimates for each sentence-pair. In this first online experiment, 30 participants were presented with 135 sentence-pairs, each trial ended with a blank space (e.g., The blonde lady is terribly afraid of wine, some of the posters show a _____). We instructed participants to provide "the rhyming word that fits best with the word at the end of the first sentence".

In order to dissociate prediction from integration accounts, sentences were designed to be semantically low-constraining. Thus, our carrier sentences were built in such a way that they allowed for multiple, equally acceptable completions - for instance "the solution for the crossword was _____'' or "the long story was about his _____''. Nonetheless, when participants were instructed to provide the rhyming word that fits best in a "poem-like" context, one (or few) words were typically mentioned frequently (around 50% of participants) while some other words were mentioned only very rarely (3.33% of participants) (see Figure 1). We grouped target words in two classes based on their frequency occurrences in the pre-test: high cloze probability (around 50% of agreement on

rhyming judgements), and low cloze probability (around 3% of agreement on rhyming judgements, (see Figure 1).

After that, for each linguistic item, we selected from the high cloze probability and the low cloze probability word classes - as obtained in the sentence completion pre-test – one high-/low- cloze probability word pair whose numerical attributes were similar for each control variable measure (see below). Then, we selected a non-rhyming word with zero cloze probability whose numerical attributes were, once again, similar to the high-/low- cloze probability word pair for each control variable measure. In doing so, we avoided any evident semantic relation between the first and the second lead-in sentence and between the prime and the target (i.e. between the first and the second final word). For each linguistic item, we also avoided any cohort overlap, both between the three alternative sentence completions and between the prime and the target, in order to avoid any cohort similarity effect (cf. Van Petten et al., 1999).

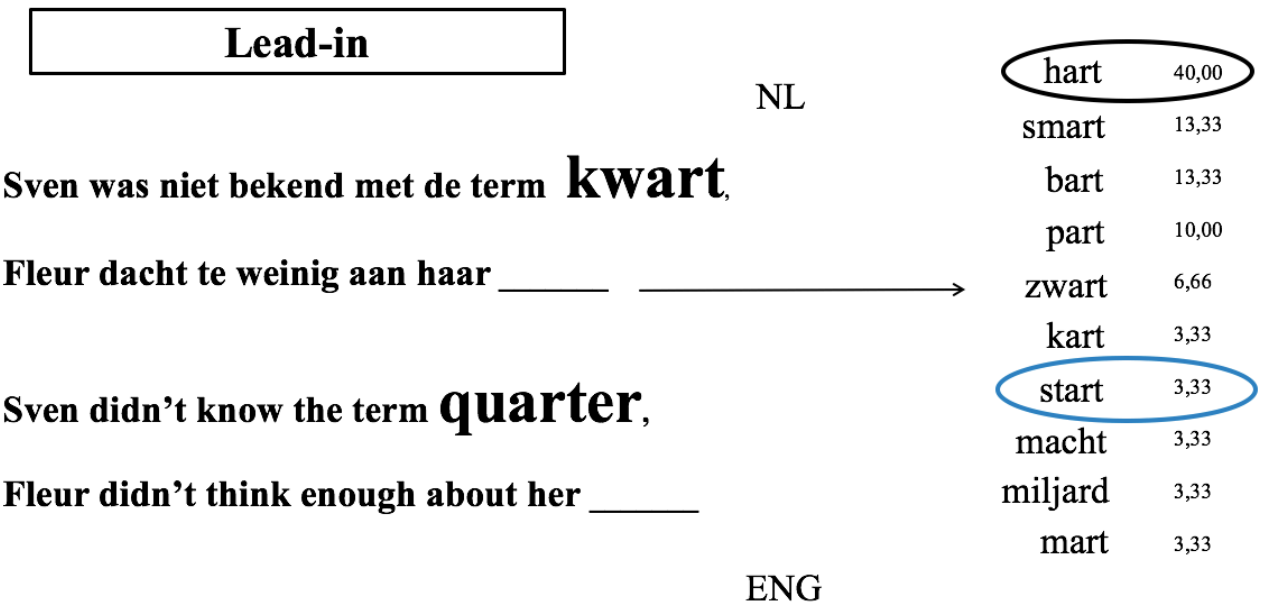

Figure 1. **Target selection**. Rhyme cloze probability values for one linguistic item as obtained in the sentence completion pre-test. The words *hart* (high-cloze probability) and *start* (low-cloze probability) were selected for the EEG experiment because of their match in terms of control variables.

We controlled for the following confounding variables: logarithmic word frequency (SUBTLEX-NL, Keuleers, Brysbaert & New, 2010), phonological neighborhood density based on CLEARPOND database (Marian, Bartolotti, Chabal & Shook, 2012), concreteness (Brysbaert, Warriner & Kuperman, 2014), semantic distance between word pairs and between target words and lead-in sentences (snaut, Mandera, Keuleers & Brysbaert, 2017). All these variables were found to modulate the amplitude of the N400 component (Van Petten & Kutas, 1990; Holcomb et al., 1999; Carrasco-Ortiz et al., 2017; Frank & Willems, 2017). Descriptive statistics for each of these control variables balanced across conditions are reported in Table 1. In order to evaluate the influence of stimuli characteristics on our dependent variable, we included confounding variables as covariates (Sassenhagen & Alday, 2016) in mixed-effects models (see below).

	CONGRUENT		**INTERMEDIATE**		**INCONGRUENT**	
	mean	sd	mean	sd	mean	sd
log word frequency	2.79	0.95	2.59	1.42	2.70	0.79
PND	15.67	8.91	13.12	8.29	13.73	9.43
concreteness	4.15	0.92	3.77	0.89	4.14	0.86
distance P–T	0.85	0.10	0.87	0.10	0.87	0.09
distance S–T	0.71	0.14	0.72	0.15	0.73	0.12
rhyme evaluation	4.57	0.27	4.41	0.31	1.07	0.11
plausibility	3.59	0.52	2.99	0.55	2.93	0.55

Table 1. **Control variables**: logarithmic word frequency, phonological neighborhood density, concreteness, semantic distance between prime and target, semantic distance between sentence and target, rhyme evaluation, plausibility evaluation.

Based on the results of the previous rhyme cloze probability test, we built the stimuli for the EEG experiment. The stimulus materials consisted of 135 sentence pairs presented auditorily. The experiment was in Dutch. Linguistic items consisted of sentence-pairs in which the final word of the second sentence (i.e. the target) either rhymed or did not rhyme with the final word of the first sentence (i.e. the prime). The target varied across three experimental conditions (see Figure 2). In the congruent condition, the rhyming target word was lexically predictable (i.e. high-cloze probability). In the intermediate condition, the rhyming target word was lexically unpredictable (i.e. low-cloze probability). In the incongruent condition, the target words did not rhyme (i.e. zero-cloze probability). The full stimulus materials with English translation can be found in the supplementary materials (Table A2.1. and A2.2.).

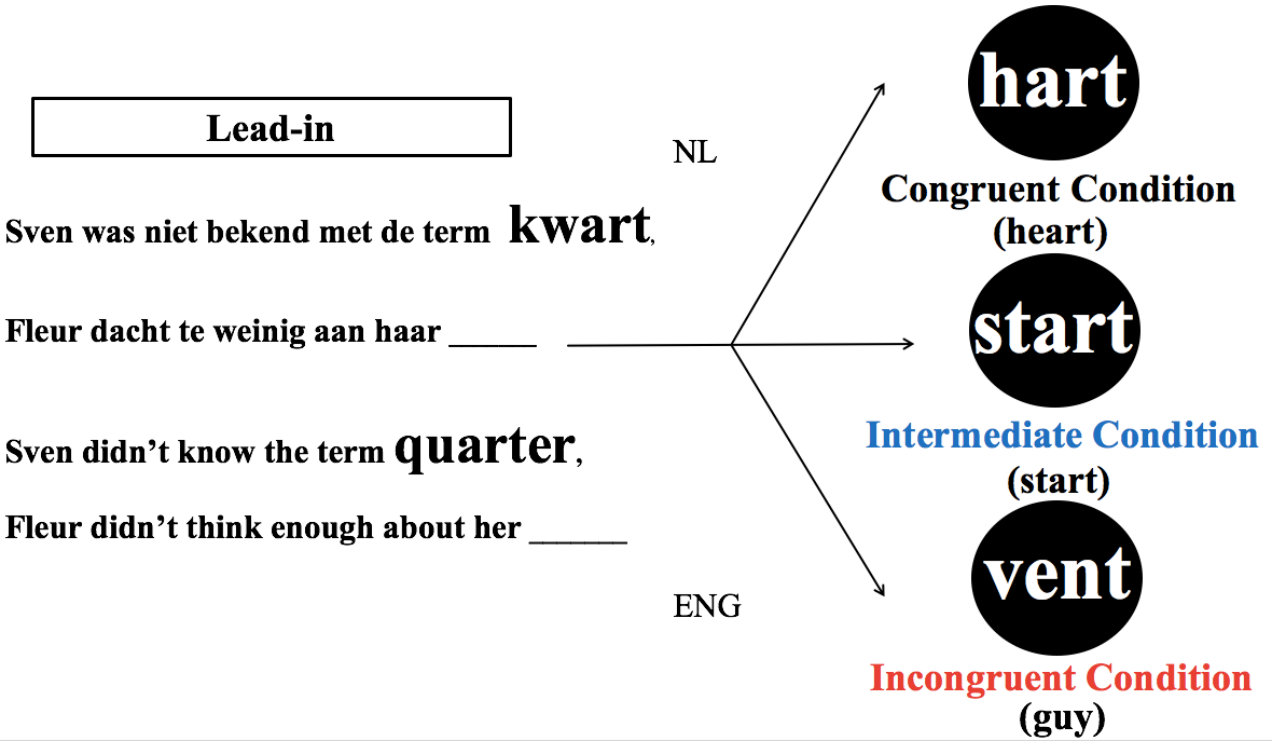

Figure 2. **Linguistic Item**. For each sentence-pair three alternative target words were selected. In the congruent condition, there was rhyme overlap and the target word was lexically predictable. In the intermediate condition, there was rhyme overlap but the target word was lexically unpredictable. In the incongruent condition, there was no rhyme overlap.

Both in the first and in the second lead-in sentence we inserted a 500ms silence period between the end of the sentence and the onset of the final word. This silent gap allowed for more isolated observations of brain activity evoked by sentence-final words. At the same time, it was expected to improve the chances that participants predict the critical word. It also further emphasizes the rhyme by placing more stress on the final word. Moreover, we opted for naturalistic speech and instructed the speaker to adopt a rhyming intonation in order to emphasize the phonological properties of the stimuli. Linguistic stimuli were recorded by a native Dutch speaker. For each linguistic item, we recorded the speaker reading aloud the first lead-in sentence followed by the second lead-in sentence. The speaker was instructed to repeat the second lead-in sentence three times, each time followed by a target word belonging to a different experimental condition. Then, we used a speech analysis software (Audacity®) to splice these former audio files into a longer speech sequence containing the two lead-in sentences while the three target words were split into three distinct audio files. During the experiment, we presented a common recording of the matrix sentence along with the recording of one of the three target words in order to avoid anticipatory co-articulation (Kuehn & Moll, 1972; Martin & Bunnell, 1981).

Pre-test: rhyme and plausibility evaluation

Finally, we ran another online experiment on LimeSurvey (GmbH, Hamburg, Germany. URL http://www.limesurvey.org) to collect rhyming and plausibility evaluations for our linguistic stimuli. In this second online experiment, 45 participants were asked to provide a rating on a scale from 1 (not at all) to 5 (very well) on ``how well did the last word rhyme with the last word in the previous sentence?''. Moreover, participants were asked to provide a rating on a different scale from 1 (highly implausible) to 5 (highly plausible) on ``how plausible you find the last word in the second sentence, regardless of whether the words rhyme?". We assumed that participants were able to distinguish plausible from implausible scenarios based on their everyday life experience (Chwilla & Kolk, 2005). Overall, participants provided close to highest rhyme evaluation for the congruent and the

intermediate conditions and the lowest rhyme evaluation for the incongruent condition (see Table 1). Plausibility ratings (see Table 1) were used as a covariate in mixed-effects models (see below).

Design

We split 135 target words for each condition into three balanced lists (405/3 = 135) using a Latin square design. Experimental conditions were counterbalanced. All 135 sentence-pairs within a list were pseudo-randomized in a single sequence for each participant using Mix (van Casteren & Davis, 2006). Trials were presented in 10 blocks, which consisted of 15 sentence pairs each with a short break between blocks. Since the number of trials was the same across conditions, only one third of the trials - belonging to the incongruent condition - did not rhyme. This frequency manipulation was intended to enhance participants' expectations to hear rhymes.

Procedure

Participants were seated in front of a computer screen (Samsung 691BF [R], 1280 x 1024, 60Hz). They were instructed to keep their eyes open and look at a fixation point during experimental trials. Acoustic stimuli were delivered via headphones. The experiment was self-paced, allowing participants to press a button to go ahead with the next trial. Participants were invited to move only when they could not see the fixation cross on the screen. There were longer breaks at the end of each experimental block. Participants were instructed to listen carefully and were prompted randomly after one-third of trials to rate how well the last word fit into the preceding context on a scale from 1 to 5. Before starting the experiment, participants performed 10 practice trials.

Data acquisition

The EEG was recorded in an electrically and acoustically shielded experiment room at a sample rate of 500 Hz using an active electrode system with a BrainAmp DC amplifier (Brain Products GmbH, Gilching, Germany). We used Presentation® software (Neurobehavioral Systems) for stimulus

delivery and EasyCap manufactured equidistant 64-electrodes montage consisting of 59 EEG channels, 4 EOG, and 2 mastoid electrodes. The electrooculography was recorded horizontally from the electrodes placed on the left and right outer canthi and vertically from the electrodes positioned above and below the left eye. Each electrode was referenced online to the left mastoid. Electrode impedance was kept below 25 kΩ.

Pre-Processing

The EEG analyses were performed using FieldTrip (v.20170414, Oostenveld, Fries, Maris & Schoffelen, 2011), Matlab release 2016b. Prior to data segmentation, data were re-referenced offline to the average of the two mastoids and subsequently filtered using a highpass filter with a frequency cut-off of 0.1 Hz and a lowpass filter with a frequency cut-off of 30 Hz. Both filters were zero-phase, two-pass Butterworth filters.

Three peripheral occipital channels (C22, C26, C54) were excluded from the analyses due to a high noise level. After channel exclusion, we applied a semi-automatic artifact rejection procedure to each time window separately. Trials were rejected based on three different criteria: an amplitude criterion of ± 100 μV, a gradient criterion (i.e. the maximum admissible voltage step between two adjacent time points) of 50 μV, and a peak-to-peak amplitude criterion (i.e., the maximum admissible absolute difference between two values within each epoch) of 100 μV (Luck, 2014). After that, we visually inspected the remaining epochs and we rejected those containing eye movements, blinks and large drifts in single electrodes. When less than 30 trials (i.e. 33.33%) per condition resisted the rejection procedure participants were excluded from further analyses. Overall, 10.89% of data was rejected. One participant was excluded leaving 30 participants for further analyses. On average, the remaining number of trials per condition was: 40.17 (sd 4.83) in the congruent condition, 39.90 (sd 5.10) in the intermediate condition, 40.23 (sd 4.77) in the incongruent condition.

Statistical analysis

Following pre-processing, we performed statistical analyses in an *a priori* determined N400 time window of 300-500 ms after target word onset (cf. Kutas & Federmeier, 2011). For the statistical analysis, a 200 ms baseline window preceding target word onset was not subtracted *a priori* but rather used as a covariate in linear mixed-effects models (Alday, in press) using the *lme4* package (v.1.1.19, Bates, Maechler, Bolker & Walker, 2014) in R (v.3.4.1, R Core Team, 2013). The mixed-effects models were computed using the single-trial mean-voltage in the N400 time window, with condition sequential difference coded and continuous covariates for word frequency, phonological neighborhood size, semantic distance, concreteness and plausibility. Sequential difference coding represents the differences between "sequential" conditions directly; here this means that the contrasts *intermediate* > *congruent* and *incongruent* > *intermediate* are directly represented in the model, with the *congruent* condition being implicitly encoded in the intercept. By using this coding scheme, the main effects in our model correspond directly to our pairwise hypotheses of interest across the entire the scalp. As such, all statistics reported here correspond directly to model coefficients and no post-hoc tests were necessary. Topography was modelled as continuous coordinates in three-dimensional space (for a similar approach, see Brilmayer et al. 2019). These three coordinates x, y, z, were allowed to interact with the predictors for condition and baseline. In this approach, topographical distribution of condition-related effects corresponds to interaction terms between the condition and the topographical predictors (x, y, z). For model parsimony, covariates were allowed to interact with condition but not with each other nor topography. Random effects consisted of by-participant and by-item intercepts and slopes for condition, corresponding to a parsimonious model that controls for variation along the effect of interest (Bates et al., 2015; Matuschek et al., 2017).

Results

Grand-averaged ERPs are shown in Figure 3 for three representative midline electrodes (whose position is indicated with red dots in the blank EEG montages represented next to each figure). A frontal electrode (i.e. `C58'), a central electrode (i.e. `C30') and a posterior electrode (i.e. `C28') are

represented. For each condition, the parametric confidence intervals (Wald, 1973) are represented with a semi-transparent ribbon surrounding the solid line (i.e. the mean). The zero time-point marked with a blue dotted line corresponds to the onset of the target word. The statistical test revealed significant differences in the N400 time window (i.e. 300-500 ms, Figure 5) between incongruent and intermediate ($t = -2.5$) as well as intermediate and congruent ($t = -2$) conditions (see mixed linear model summary, Table A3.1, supplementary materials). Difference topographies between conditions are represented in Figure 4. In both cases, a centro-parietal distribution can be observed which is consistent with the canonical view of the N400 (cf. Kutas & Federmeier, 2011).

In line with our hypotheses, we observed a significant difference, in the N400 window, not only between incongruent and congruent conditions - which differ in terms of their adherence to the rhyming scheme - but also between intermediate and congruent conditions - which differ in terms of their predictability. Importantly, the covariance analysis suggests that, although a few covariates have some moderating influence on the effect of condition, none of them really change its overall structure (see Figure 5).

A late positivity is also visible in the ERPs (Figure 3). This is in line with previous observations in N400 paradigms (cf. the P560 in Kutas & Hillyard 1980; Van Petten & Luka, 2012 for review) and, in language studies, it has been attributed to task effects when using a violation paradigm (cf. Sassenhagen et al. 2014).

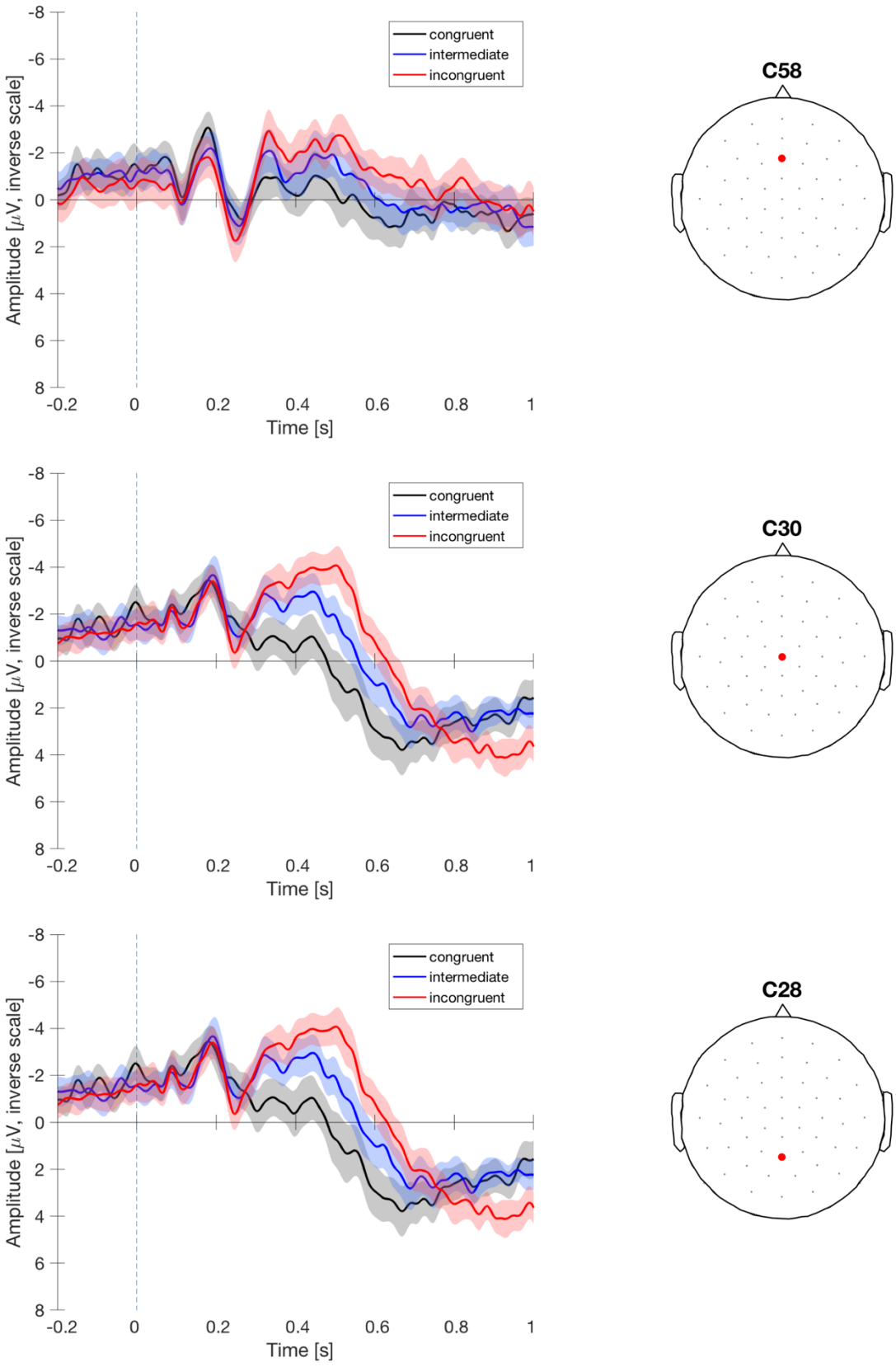

Figure 3. **Midline electrodes**. Grand-averaged ERPs for three representative electrodes (i.e. a frontal electrode above, a central electrode in between, a posterior electrode below) are represented.

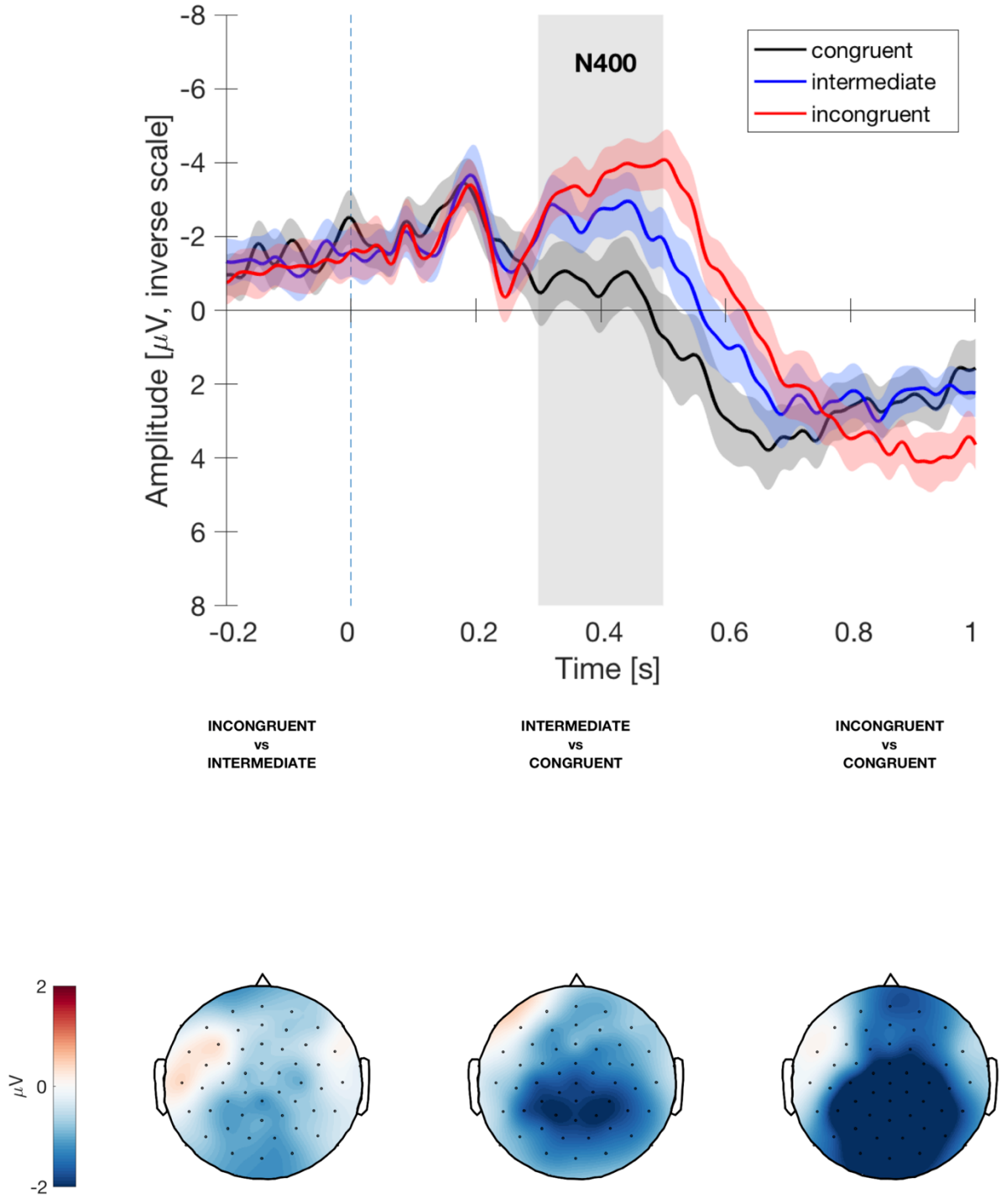

Figure 4. **The N400 effect**. Grand-averaged ERPs for posterior electrode 'C28' are shown in the upper part of the figure. The N400 time window (300-500 ms) is highlighted with a gray shaded rectangle. Difference topographies for each contrast are represented in the lower part of the figure.

	Chisq	Df	Pr(>Chisq)
condition	10.7478404	2	0.0046359
x	9.5818284	1	0.0019651
y	2.743738	1	0.0976361
z	48.9033525	1	< 0.001
word frequency	0.1291206	1	0.7193456
phonological neighborhood	2.3131605	1	0.128283
semantic distance	0.8676858	1	0.3515968
concreteness	1.6158538	1	0.2036709
plausibility	0.0367458	1	0.8479834
condition:x	0.1065219	2	0.9481325
condition:y	100.3809839	2	< 0.001
x:y	6.9870909	1	0.00821
condition:z	32.590449	2	< 0.001
x:z	15.6529886	1	< 0.001
y:z	0.0708424	1	0.7901142
condition:word frequency	0.7483737	2	0.6878484
condition:phonological neighborhood	0.5360686	2	0.7648816
condition:semantic distance	2.603896	2	0.2720014
condition:concreteness	6.1091003	2	0.0471439
condition:plausibility	5.7256462	2	0.0571073
condition:x:y	0.1143902	2	0.9444098
condition:x:z	3.4330719	2	0.1796875
condition:y:z	1.978454	2	0.371864
x:y:z	1.1186508	1	0.2902092
condition:x:y:z	1.255213	2	0.5338681

Table 2. **Analysis of Deviance (Type II Wald Chi-Square Tests)**. Type-II Wald tests for the N400 time window, analogous in interpretation to repeated measures ANOVA with Type-II sum of squares. The use of the Chi-square instead of the F statistic is an asymptotic approximation, equivalent to treating t values as z values (see Baayen, Davidson & Bates, 2008). These tests provide a convenient summary of the effects and asymptotically equivalent to likelihood-ratio tests. Terms related to the baseline covariate have been omitted here, but can be found in the full model summary in Appendix 3. Note the main effect for condition as well as the interaction with topographical factors (x=anterior-posterior axis, y=laterality, z=vertical axis), matching the graphical impression of graded negativity most prominent at posterior sites (see Figures 3-4).

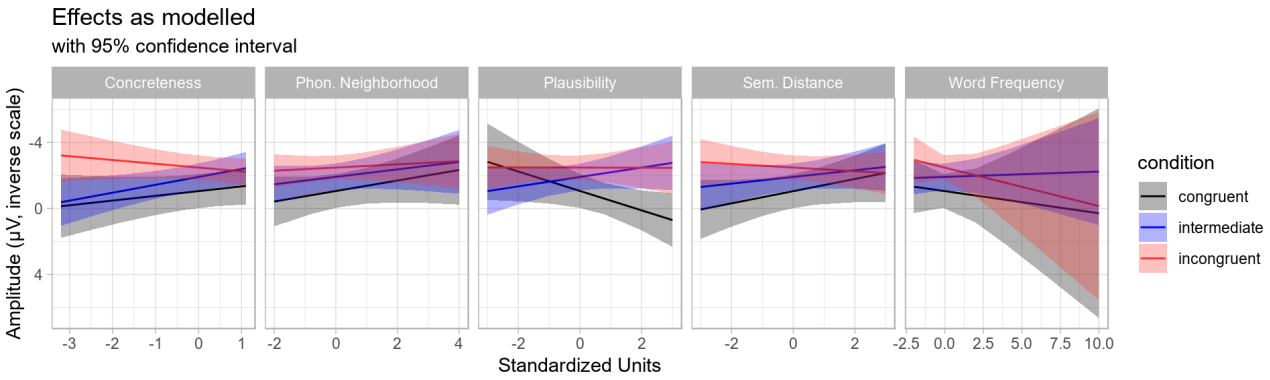

Figure 5. **Covariance Analysis**. The influence of concreteness, phonological neighborhood density, plausibility, semantic distance and word frequency on the N400 amplitude for each condition is represented. The different covariates have some moderating influence on the effect of condition, but none fundamentally change the overall structure, as the lines are largely parallel and the overall (vertical) order of the effects doesn't change (within the uncertainty given by the confidence intervals). The average trend across conditions is in line with previous findings (e.g. decreased plausibility leads to a more negative N400), although these effects are not particularly apparent due to the intentional choice of stimuli without much variation along these features.

Discussion

Prediction is an important feature of contemporary theories of language processing (Altmann & Mirkovic, 2009; Dell & Chang, 2014; Federmeier, 2007; Ferreira & Chantavarin, 2018; Pickering & Gambi, 2018; Hale, 2001; Hickok, 2012; Huettig 2015; Kuperberg & Jaeger, 2016; Levy, 2008; Norris, McQueen, & Cutler, 2016; Pickering & Garrod, 2013; Van Petten & Luka, 2012). Key evidence for prediction in language comes from electrophysiological studies. The vast majority of electrophysiological studies have interpreted a reduced N400 ERP component during the (potentially predicted) target word as an important electrophysiological marker of anticipation. It is indeed likely that this negative-going and centro-parietally distributed ERP component, which occurs approximately 400ms after target word onset, is partly indexing prediction of the word (or its 'constituent representations'). Previous EEG studies measuring N400 modulations on the target word however have been unable to distinguish lexical prediction (i.e. 'top-down' activation) from lexical integration (i.e. 'bottom-up' activation of 'incoming' words) accounts (though not all studies have acknowledged this interpretation problem explicitly).

In the present study, we used rhyming sentences in which the rhyme overlap in the critical conditions was the same. We also used sentential context that did not contextually bias towards the target word and kept the plausibility of sentence completion constant across conditions. This allowed us to distinguish prediction from integration accounts because the crucial conditions differed on lexical predictability but not plausibility of the rhyme word: by keeping the ease of integration constant across conditions, any difference in N400 effects across conditions can be attributed to prediction. EEG analyses revealed a robust difference in the N400 window between incongruent and congruent conditions (that differed in rhyme) and between intermediate and congruent conditions (that differed in predictability). In other words, we observed N400 target word modulations that we believe we can ascribe to prediction with a high degree of confidence (since the crucial conditions differed only in predictability but not plausibility). These conclusions are supported by our covariance analysis.

Some limitations of the present study deserve further exploration. As in most previous studies on prediction in language processing, we used cloze probability completions as a proxy for predictability. One of the shortcomings is that the cloze probability test (Taylor, 1953; Kutas & Hillyard, 1984) is a language production task and may not be a 'perfect' measure of predictability during language comprehension. An alternative would be to assess word predictability using corpus studies. Predictability measures derived from corpus studies however have the disadvantage that it is typically not verified experimentally how closely corpus-derived forward probabilities correspond to people's actual anticipatory language processing. It is reasonable to assume that both cloze probability and corpus-based measures are good proxies for word predictability but further empirical research could be directed at verifying these assumptions. Similarly, as most previous studies, we used sentence plausibility as a proxy for word integration. The assumption is that words that are more plausible are more easily integrated in a sentence interpretation than words that are less plausible. Future research could also further test this assumption.

The finding of a robust difference in the N400 window between intermediate and congruent conditions (which differed in final word predictability only) strongly suggests that it reflects predictive processes. The present study does not reveal whether language users routinely predict during language processing but minimally these results demonstrate that a) at least in some contexts, comprehenders can and do actively predict, and b) that the N400 can under certain (carefully controlled circumstances) be used as a neural marker of such prediction processes. It is also important to note that our study does not show at what representational level participants predicted the target words (e.g., to what extent phonological form was pre-activated). It is however likely that the N400 difference between intermediate and incongruent conditions reflects the partial phonological form overlap (i.e. the rhyme) with the predicted high cloze target word as plausibility was matched between intermediate and incongruent conditions (and the 3% difference in cloze probability is unlikely to account for this N400 difference). Given the recent failures to replicate experimental evidence for *routine* phonological form prediction (Nieuwland et al., 2018; cf. Nieuwland, 2019) further research is required to investigate the representational content of linguistic predictions during every day communicative interactions.

Is it possible to distinguish integration and prediction accounts of ERP N400 modulations in language processing through experimental design? We believe that the present exploratory study shows that the answer is yes. The EEG design we employed here may not be suitable to answer every question about prediction in language processing but we suggest that it demonstrates at least in principle that it is possible to distinguish lexical prediction from lexical integration accounts through experimental design. There is some recent experimental evidence that prediction and integration processes may have distinct N400 spatiotemporal profiles (Nieuwland et al., in press). We believe that both approaches (linking prediction and integration to distinct N400 spatiotemporal profiles as well as distinguishing both accounts through experimental design) are complementary and promising avenues for future research on prediction in language. Electrophysiological (target word N400) studies that neither distinguish prediction and integration accounts through experimental design nor

through spatiotemporal profiles will still be useful as a first step for exploring predictive processing but will have to be followed up with tighter experimental approaches.

References

- Alday, P. M. (in press, DOI 10.1111/psyp.13451). How much baseline correction do we need in ERP research? Extended GLM model can replace baseline correction while lifting its limits. *Psychophysiology*. *arXiv preprint arXiv:1707.08152*.
- Altmann, G. T., & Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive science*, 33(4), 583-609.
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. *R package version*, *1*(7), 1-23.
- Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.
- 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.
- 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.
- Brilmayer, I., Werner, A., Primus, B., Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2019). The exceptional nature of the first person in natural story processing and the transfer of egocentricity. *Language, Cognition and Neuroscience*, *34*(4), 411-427.
- Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods*, 46(3), 904-911.
- Carrasco-Ortiz, H., Midgley, K. J., Grainger, J., & Holcomb, P. J. (2017). Interactions in the neighborhood: Effects of orthographic and phonological neighbors on N400 amplitude. *Journal of Neurolinguistics*, 41, 1-10.
- Chwilla, D. J., & Kolk, H. H. (2005). Accessing world knowledge: Evidence from N400 and reaction time priming. *Cognitive Brain Research*, 25(3), 589-606.
- Connolly, J. F., & Phillips, N. A. (1994). Event-related potential components reflect phonological and semantic processing of the terminal word of spoken sentences. *Journal of Cognitive Neuroscience*, *6*(3), 256-266.
- Deacon, D., Dynowska, A., Ritter, W., & Grose-Fifer, J. (2004). Repetition and semantic priming of nonwords: Implications for theories of N400 and word recognition. *Psychophysiology*, 41(1), 60-74.
- 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 B: Biological Sciences*, *369*(1634), 20120394.
- 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.
- Donnenwerth-Nolan, S., Tanenhaus, M. K., & Seidenberg, M. S. (1981). Multiple code activation in word recognition: evidence from rhyme monitoring. *Journal of Experimental Psychology: Human Learning and Memory*, 7(3), 170.
- Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, *44*(4), 491-505.
- Ferreira, F., & Chantavarin, S. (2018). Integration and prediction in language processing: a synthesis of old and new. *Current Directions in Psychological Science*, *27*(6), 443-448.
- Frank, S. L., & Willems, R. M. (2017). Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Language, Cognition and Neuroscience*, 32(9), 1192-1203.
- Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies* (pp. 1-8). Association for Computational Linguistics.

- Hickok, G. (2012). The cortical organization of speech processing: Feedback control and predictive coding the context of a dual-stream model. *Journal of communication disorders*, *45*(6), 393-402.
- Hillinger, M. L. (1980). Priming effects with phonemically similar words. *Memory & Cognition*, 8(2), 115-123.
- Holcomb, P. J., Kounios, J., Anderson, J. E., & West, W. C. (1999). Dual-coding, context-availability, and concreteness effects in sentence comprehension: An electrophysiological investigation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(3), 721.
- 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., Martin, A. E., & Nieuwland, M. S. (2017). How robust are prediction effects in language comprehension? Failure to replicate article-elicited N400 effects. *Language, Cognition and Neuroscience*, *32*(8), 954-965.
- Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior research methods*, *42*(3), 643-650.
- Kochari, A. R., & Flecken, M. (2018). Lexical prediction in language comprehension: a replication study of grammatical gender effects in Dutch. *Language, Cognition and Neuroscience*, 1-15.
- Kuehn, D. P., & Moll, K. L. (1972). Perceptual effects of forward coarticulation. *Journal of Speech and Hearing Research*, *15*(3), 654-664.
- Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension?. *Language, cognition and neuroscience*, *31*(1), 32-59.
- Kutas, M., & Hillyard, S. A. (1980). Event-related brain potentials to semantically inappropriate and surprisingly large words. *Biological psychology*, *11*(2), 99-116.
- Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307(5947), 161.
- Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual review of psychology*, *62*, 621-647.
- Laszlo, S., & Federmeier, K. D. (2011). The N400 as a snapshot of interactive processing: Evidence from regression analyses of orthographic neighbor and lexical associate effects. *Psychophysiology*, *48*(2), 176-186.
- Lau, E., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (de)Constructing the N400. *Nature Reviews Neuroscience*, 9, 920-933.
- Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126-1177.
- Luck, S. J. (2014). *An introduction to the event-related potential technique*. MIT press.
- Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, 92, 57-78.
- Marian, V., Bartolotti, J., Chabal, S., & Shook, A. (2012). CLEARPOND: Cross-linguistic easy-access resource for phonological and orthographic neighborhood densities. *PloS one*, 7(8), e43230.
- Martin, J. G., & Bunnell, H. T. (1981). Perception of anticipatory coarticulation effects. *The Journal of the Acoustical Society of America*, *69*(2), 559-567.
- Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305-315.
- Nieuwland, M. S., Barr, D. J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D. I., Ferguson, H. J., Fu, X., Heyselaar, E., Huettig, F., Husband, E. M., Ito, A., Kazanina, N.,

Kogan, V., Kohút, Z., Kulakova, E., Mézière, D., Politzer-Ahles, S., Rousselet, G., Rueschemeyer, S.-A., Segaert, K., Tuomainen, J., & Von Grebmer Zu Wolfsthurn, S. (in press). Dissociable effects of prediction and integration during language comprehension: Evidence from a large-scale study using brain potentials. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*.

- Nieuwland, M. S. (2019). Do 'early' brain responses reveal word form prediction during language comprehension? A critical review. *Neuroscience and Biobehavioral Reviews, 96*, 367-400.
- Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., ... & Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife*, *7*, e33468.
- Norris, D., McQueen, J. M., & Cutler, A. (2016). Prediction, Bayesian inference and feedback in speech recognition. *Language, cognition and neuroscience*, *31*(1), 4-18.
- Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational intelligence and neuroscience*, *2011*, 1.
- Otten, M., & Van Berkum, J. J. (2009). Does working memory capacity affect the ability to predict upcoming words in discourse?. *Brain research*, *1291*, 92-101.
- Pickering, M. J., & Garrod, S. (2013). Forward models and their implications for production, comprehension, and dialogue. *Behavioral and Brain Sciences*, *36*(4), 377-392.
- Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: a theory and review. *Psychological Bulletin*.
- Praamstra, P., & Stegeman, D. F. (1993). Phonological effects on the auditory N400 event-related brain potential. *Cognitive Brain Research*, 1(2), 73-86.
- Praamstra, P., Meyer, A. S., & Levelt, W. J. (1994). Neurophysiological manifestations of phonological processing: Latency variation of a negative ERP component timelocked to phonological mismatch. *Journal of cognitive Neuroscience*, *6*(3), 204-219.
- Rapp, D. N., & Samuel, A. G. (2002). A reason to rhyme: Phonological and semantic influences on lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(3), 564.
- Rugg, M. D., & Barrett, S. E. (1987). Event-related potentials and the interaction between orthographic and phonological information in a rhyme-judgment task. *Brain and language*, 32(2), 336-361.
- Sassenhagen, J., Schlesewsky, M., & Bornkessel-Schlesewsky, I. (2014). The P600-as-P3 hypothesis revisited: Single-trial analyses reveal that the late EEG positivity following linguistically deviant material is reaction time aligned. *Brain and language*, *137*, 29-39.
- Sassenhagen, J., & Alday, P. M. (2016). A common misapplication of statistical inference: nuisance control with null-hypothesis significance tests. *Brain and language*, *162*, 42-45.
- Shulman, H. G., Hornak, R., & Sanders, E. (1978). The effects of graphemic, phonetic, and semantic relationships on access to lexical structures. *Memory & Cognition*, 6(2), 115-123.
- Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Bulletin*, 30(4), 415-433.
- Team, R. C. (2013). R: A language and environment for statistical computing.
- Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(3), 443.
- Van Casteren, M., & Davis, M. H. (2006). Mix, a program for pseudorandomization. *Behavior research methods*, 38(4), 584-589.
- Van Den Brink, D., Brown, C. M., & Hagoort, P. (2001). Electrophysiological evidence for early contextual influences during spoken-word recognition: N200 versus N400 effects. *Journal of cognitive neuroscience*, *13*(7), 967-985.

- Van Petten, C., & Kutas, M. (1990). Interactions between sentence context and word frequency in event-related brain potentials. *Memory & cognition*, 18(4), 380-393.
- 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.
- Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*(2), 176-190.
- Wald, A. (1973). Sequential analysis. Courier Corporation.
- 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. *Journal of cognitive neuroscience*, *16*(7), 1272-1288.