Neuroscience and Biobehavioral Reviews 96 (2019) 367–400

Contents lists available at ScienceDirect

Neuroscience and Biobehavioral Reviews

journal homepage: www.elsevier.com/locate/neubiorev

Review article

Do 'early' brain responses reveal word form prediction during language comprehension? A critical review

Mante S. Nieuwland[a,b,*]

[a] Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands
[b] Donders Institute for Brain, Cognition and Behaviour, Nijmegen, the Netherlands

ABSTRACT

Current theories of language comprehension posit that readers and listeners routinely try to predict the meaning but also the visual or sound form of upcoming words. Whereas most neuroimaging studies on word prediction focus on the N400 ERP or its magnetic equivalent, various studies claim that word form prediction manifests itself in 'early', pre-N400 brain responses (e.g., ELAN, M100, P130, N1, P2, N200/PMN, N250). Modulations of these components are often taken as evidence that word form prediction impacts early sensory processes (*the sensory hypothesis*) or, alternatively, the initial stages of word recognition before word meaning is integrated with sentence context (*the recognition hypothesis*). Here, I comprehensively review studies on sentence- or discourse-level language comprehension that report such effects of prediction on early brain responses. I conclude that the reported evidence for the sensory hypothesis or word recognition hypothesis is weak and inconsistent, and highlight the urgent need for replication of previous findings. I discuss the implications and challenges to current theories of linguistic prediction and suggest avenues for future research.

1. Introduction

It is well-established that people sometimes implicitly predict upcoming information during language comprehension, and maybe even specific words. Such predictions can involve the activation of the semantic, grammatical and/or form features of a word before it appears, and are thought to facilitate processing once the word is encountered (e.g., Kutas et al., 2011; Pickering and Garrod, 2007, 2013). Prediction is sometimes viewed as an integral mechanism of language comprehension that allows it to operate efficiently and incrementally (e.g., Altmann and Mirkovic, 2009), although there is ongoing debate about the details of this mechanism. For example, it is unclear whether prediction involves actual active hypothesis generation about the upcoming occurrence of specific input words, a more passive pre-activation of semantic content that naturally emerges from a representation of the context, or both (for discussion, see Baggio, 2018; Kutas et al., 2011; Van Berkum, 2009; Van Petten and Luka, 2012). Demonstrations of linguistic prediction are rapidly accumulating in the scientific literature, often involving scalp-recordings of the brain's electric or magnetic activity (electro- or magneto-encephalography, EEG/MEG). Many of those demonstrations involve the N400 (Kutas and Hillyard, 1980, 1984), the event-related potential (ERP) component commonly associated with semantic processing (for reviews, see Kutas and Federmeier, 2011; Lau et al., 2009; Van Berkum, 2009), and, more recently, the frontal post-N400 positivity or PNP (Van Petten and Luka, 2012). However, a growing number of studies report effects of prediction that occur 'early', which usually means that the effect occurs before the peak of the N400 component at approximately 400 ms after word onset. These early effects often feature in passing in prominent reviews as clear-cut evidence for prediction of a specific word form (e.g., Hagoort, 2017; Kuperberg and Jaeger, 2016; Pickering and Garrod, 2013; Pickering and Gambi, 2018), but, to the best of my knowledge, have not yet been subjected to a dedicated and in-depth review.

To fill this gap, this review provides a comprehensive overview, approximate classification and in-depth discussion of early prediction effects reported in the psycholinguistic literature[1], roughly following the temporal order in which these effects occur (see Table 1). I will briefly describe the rationale of each study, the type of prediction it addresses, aspects of the experimental design and data analysis that are relevant for interpretation, and its conclusions. My review is 'critical' in the sense that I discuss potential limitations to the reported conclusions and pose questions that remain to be addressed in future research. I also

* Correspondence to: Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD Nijmegen, the Netherlands.
E-mail address: mante.nieuwland@mpi.nl.

[1] By way of disclaimer: I am only human and my literature search is likely to be imperfect. I may have unintentionally omitted relevant studies, for which I apologize in advance. In addition, the P300 ERP is not included in this review because it is primarily indexes task-related processes rather than regular comprehension processes, and because it is not generally considered to be an early, pre-N400 component associated with the prediction of linguistic form (for relevant discussion on the P300, see Folstein and Van Petten, 2011; Osterhout and Hagoort, 1999; Van Petten and Luka, 2012). However, the P300 is briefly discussed in relation to a prediction-task in the N250 section.

https://doi.org/10.1016/j.neubiorev.2018.11.019
Received 18 May 2018; Received in revised form 1 October 2018; Accepted 27 November 2018

Table 1
Overview of the EEG/MEG components/effects reviewed in this article. The table lists the effect name, main citation for the effect, sample size, what type of prediction was studies ('sensory' or 'recognition' refers whether the effect is relevant for the sensory or recognition hypothesis), description of the observed effects, information about their approximate timing (or the window of statistical testing), the modality (with word presentation rate in ms for written studies) and language for which they have been observed, the used EEG-reference and filtering procedure, the employed secondary task, and known close or direct replication attempts.

Effect Name	Study	Analyzed participants[b] (N)	Prediction type	Description	Timing (ms)	Modality (written word rate)	Language	Reference	Filter (Hz)	Secondary Task	Replication[d]
Early Inflection Positivity[a]	Van Berkum et al. (2005)	24	Lexical, Grammatical gender	Effect of prediction-inconsistent inflections on prenominal adjectives	50-250	Spoken	Dutch	Left Mastoid	0.02-70	–	No, 2 attempts[c]
Early Cloze Positivity[a]	Dambacher et al. (2009)	32	Lexical (sensory)	Effect of predictability (cloze probability) at occipital EEG channels	50-90	Written (280)	German	Global average	0.1-30	Comprehension questions	No, 2 attempts[e]
ELAN	Lau et al. (2006)	32	Word category	Early Left-Anterior Negativity elicited by word category violations	100-300[f]	Written (500)	English	Global average	0.1-70	Grammaticality judgment	No, 1 attempt[g]
M100 (MEG)	Dikker et al. (2009)	13 (Exp. 1) 12 (Exp. 2)	Word category (sensory)	Effect of word category violation on M100 source-reconstructed activity in visual cortex	130-145	Written (600)	English	n/a	1-40	Grammaticality judgment	Yes/Partly, 1 successful self-replication for participles[h]
	Dikker et al. (2010)	15	Word category (sensory)	Effect of phonological typicality mismatch on M100 source-reconstructed activity in visual cortex	135-150	Written (600)	English	n/a	1-40	Grammaticality judgment	No known attempts
	Dikker and Pylkkanen (2011)	15	Lexical (sensory)	Effect of mismatch between pictures and subsequent words on the M100 (not source-reconstructed)	92-102	Written (600)	English	n/a	1-40	Picture-word matching	No known attempts
P130 (Occipital P1)	Kim and Lai (2012)	20	Lexical (sensory)	Occipital effect of orthographic similarity between pseudowords and expected words	125-145	Written (550)	English	Mastoid average	0.1-30	Comprehension questions	No, 2 attempts
	Kim and Gilley (2013)	26/26 (between-subjects)	Lexical, Word category (sensory)	Occipital effect of predictable syntactic anomaly	125-145	Written (500)	English	Global average	0.1-100	Comprehension questions	No known attempts
N1/P2	Penolazzi et al. (2007)	17	Lexical (recognition)	Interaction between cloze probability and word length on N1 and P2 peaks	110-130, 170-190	Written (700)	English	Mastoid average	0.1-30	Comprehension questions	No known attempts
	Lee et al. (2012)	21		Interaction between cloze probability and frequency on N1, and effect of cloze probability on P2	120-150, 200-250	Written (700)	Mandarin Chinese	Mastoid average	0.1-30	Comprehension questions	No known attempts
N200, PMN (Phonological Mismatch negativity)	Connolly and Phillips (1994)	20	Lexical, Phonological (recognition)	Evenly distributed N200 effect of unexpected word-initial phoneme	275 ms	Spoken	English	Linked ears	0.01-30	–	No[i]
	Hagoort and Brown (2000)	12 (Exp. 1) 12 (Exp. 2)		Evenly distributed N200 effect of unexpected word-initial phoneme	200-300 ms	Spoken	Dutch	Left mastoid	0.02-30	–	
	Van den Brink et al. (2001)	21		Evenly distributed N200 effect of unexpected word-initial phoneme	150-250	Spoken	Dutch	Left mastoid	0.02-30	–	
	van den Brink and Hagoort, (2004)	21		Posterior N200 effect of unexpected word-initial phoneme	150-250	Spoken	Dutch	Left mastoid	0.02-30	–	
	Boudewyn et al. (2015)	20		Frontal N200 effect of unexpected word-initial phoneme	200-300	Spoken	English	Mastoid average	0.01-25	Comprehension questions	

(*continued on next page*)

Table 1 (*continued*)

Effect Name	Study	Analyzed participants[b] (N)	Prediction type	Description	Timing (ms)	Modality (written word rate)	Language	Reference	Filter (Hz)	Secondary Task	Replication[d]
N250	Brothers et al. (2015)	24	Lexical (recognition)	Negativity for unpredicted words compared to predicted words of similar cloze probability	200-300	Written (600)	English	Mastoid average	0.05-30	Sentence-final word prediction task	Yes, 2 successful self-replications[j]

[a] This name is given for ease of exposition; it was not used by the authors reporting this effect.

[d] This refers to the sample size upon which the statistical results are based, it does not always correspond to the number of tested participants. In some of the reported studies 1/3 of all tested participants were excluded from the analysis.

[c] These follow-up studies with the same manipulation generated other, later ERP effects ((Otten and Van Berkum, 2008; Otten et al., 2007), and are not included in this review.

[d] Replication is loosely defined as whether another study found a similar and statistically significant effect involving the same manipulation. The known replication attempts are usually not fully direct (identical) to the original in terms of design, and none report quantitative comparisons of reported effect size.

[e] Replication attempts by the same authors, with the same stimuli but slower presentation procedures and different analyses (Dambacher et al., 2012).

[f] Lau et al. observed an ELAN effect in the 200–400 ms time window after critical word onset, but the ELAN is sometimes observed as early as 100 ms.

[g] Kaan et al. (2016).

[h] A similar M1 effect was observed for participle target words in Experiment 1 and 2. Participle target words were also included in the design of Dikker et al. (2010) but the associated results were not reported.

[i] Replication is hard to establish because the studies have defined and tested N200 activity in different ways. Several other studies have not reported or failed to find N200 activity in related paradigms (e.g., Diaz and Swaab, 2007; Van Petten et al., 1999), which are discussed in the N200/PMN section.

[j] In the studies that replicated this effect, however, this effect is not labelled and interpreted/discussed as an N250 effect but as an N400 effect (Brothers et al., 2017; Dave et al., 2018).

dedicate discussion to the replicability of the reported effects, which is currently an important issue not only in psychology (e.g., Chambers, 2017; Open Science Collaboration, 2015; Zwaan et al., 2018), but is also receiving increased attention in language neuroscience (e.g., Nieuwland et al., 2018a; Siegelman et al., 2017).

The current review does not challenge the generally accepted view that prediction contributes to language comprehension in some form or another. Linguistic and non-linguistic context can activate semantic and possibly form features of a word before it appears. This can happen due to spreading activation between individual words because they share semantic features (e.g., understanding the word ‘chocolate’ activates semantic information that facilitates understanding the word ‘candy’, a phenomenon called semantic priming; Neely, 1977), or due to the unfolding interpretation of the message conveyed by the context. These instantiations of semantic prediction are well-documented and extensively reviewed elsewhere (e.g., Kutas et al., 2011; Van Berkum, 2009; Van Petten and Luka, 2012), although there is still discussion about to extent to which these prediction are actively generated (as an explicit and specific hypothesis about what comes next) or just reflect a passive pre-activation that results from processing of the context (Baggio, 2018; Kuperberg and Jaeger, 2016). The current review also does not negate that people sometimes expect a specific word to appear (lexical prediction). During conversational interactions, for example, people are likely to expect a specific word that refers to whatever or whomever the conversation is about (e.g., Altmann and Kamide, 1999; Kehler and Rohde, 2013) or that refers to a highly salient person or object in the non-linguistic, perceptual environment. This review also does not negate that it is possible, in principle, to strategically predict the specific perceptual features of a word, for example, if one is instructed to do so or if it benefits performance on a specific task, but these predictions may not be representative of routine predictive processing in natural language use.

The current review does, however, scrutinize the reported evidence from early brain responses for the prediction of a specific word form, and considers this evidence in light of architectural claims about the language system. Prediction-related effects on early brain responses are sometimes taken as evidence that prediction directly facilitates the non-linguistic, perceptual analysis of linguistic input (e.g., Dikker et al., 2009) or that prediction facilitates word recognition from linguistic, form-based representations (e.g., Lee et al., 2012). In both these views, prediction facilitates the processes that are initiated upon encountering a word and that precede the contextual integration of word meaning. For that reason, prediction effects observed on early brain responses are often contrasted with effects observed on ‘late’ components like the N400 and the PNP, which are sometimes thought to reflect the contextual integration of semantic information. Many studies have used the N400 to demonstrate the pre-activation of meaning and even the grammatical gender of expected words (e.g., Federmeier and Kutas, 1999; Freunberger and Roehm, 2016; Ito et al., 2016; Maess et al., 2016; Wlotko and Federmeier, 2015; Nieuwland et al., 2018a; Ito et al., 2016, 2017; Otten and Van Berkum, 2008). However, a reduced N400 component to a predictable word, compared to an unpredictable word, itself may not be clear-cut evidence that people activate the actual form of a word before it appears (e.g., Federmeier and Kutas, 1999, for discussion). N400 activity may reflect merely the extent to which an incoming word matches the prediction of a word’s meaning (i.e. semantic features) rather than its actual word form. The current literature does not appear to raise such caveats against activity of earlier ERP components because they are not associated with processing word meaning.

This review covers as many as 8 different early EEG or MEG effects observed in an even larger number of studies. The scope of this review is limited to studies from the EEG/MEG literature on the role of prediction in sentence- or discourse-level language comprehension. Not covered in this review are behavioral effects of predictability (e.g., Luke and Christianson, 2016; Staub and Clifton, 2006; for a review, see Staub, 2015), early EEG/MEG predictability effects in studies that use single-word or word-pair stimuli[2] (Sohoglu et al., 2012; Ettinger et al., 2014; Fruchter et al., 2015; Gagnepain et al., 2012), or EEG/MEG effects that occur before a predicted word is presented but that are considered 'late' like the N400 and P600 (e.g., Otten et al., 2007; Otten and Van Berkum, 2008; for a review, see Kutas et al., 2011; Van Berkum et al., 2008). The studies covered in this review all minimally required participants to recruit syntactic and semantic combinatorial processes typically involved in comprehension of sentences or stories. The participants in these studies read or listened to sentences or stories in their native language that were complete (i.e., no missing or degraded stimuli), although many of these studies presented participants with semantically, syntactically or orthographically anomalous phrases. The reviewed studies used rather different task instructions and different manipulations, but all manipulated word (category) predictability with a cloze probability norming test[3], under the assumption that high cloze words are likely to be predicted during online comprehension. I will first discuss the general theoretical background for these studies, and then discuss each study in the order in which they appear in Table 1.

1.1. The sensory hypothesis and the word recognition hypothesis

The studies and associated early EEG/MEG effects included in this review broadly fall into two groups that are associated with different questions and theoretical implications. The first group asks whether prediction of a specific word form is implemented at the level of early perceptual processes in primary sensory cortices (*the sensory hypothesis*[4]), and includes some of the earliest effects (Early Cloze Positivity, M100, P130). The second group asks whether prediction or predictability facilitates the recognition of a word (*the word recognition hypothesis)*, and assumes this process to be separate from the semantic integration of word meaning with sentence context. The effects in this group are the N1, P2, N200/PMN and N250, which are associated with visual/orthographic or auditory/phonological word form prediction, but not explicitly considered to be perceptual in nature (i.e., reflecting processes in primary sensory cortices).

The difference between the sensory hypothesis and the word recognition hypothesis traces back to the traditional distinction between pre-lexical, lexical and post-lexical stages of processing in theories of word recognition (e.g., Marslen-Wilson, 1987; Morton, 1979; for a review, see Balota et al., 2006; Rastle, 2007; Dahan and Magnuson, 2006). Pre-lexical processes operate on acoustic or visual features of the input and result in the activation of a lexical representation (i.e., the orthographic, phonological, syntactic and semantic information associated with a specific word form). Once activated, lexical information is integrated into the sentence- or discourse-context in a post-lexical stage of processing. Theories of visual word recognition additionally distinguish between pre-lexical processes operating on basic (non-linguistic) visual features or on letters (e.g., Grainger and Jacobs, 1996). Theories of spoken word recognition sometimes distinguish between pre-lexical processes operating on acoustic features or on phonemic representations (for a review, see Dahan and Magnuson, 2006; Kazanina et al., 2018; Monahan, 2018). The sensory hypothesis holds that linguistic predictions impact the lowest, pre-lexical level of processing that takes basic perceptual features as its input, whereas the word recognition hypothesis holds that predictions impact the lexical stage of processing, where information associated with a specific word form is accessed (lexical access).

The sensory hypothesis (e.g., Dambacher et al., 2009; Dikker et al., 2009, 2010) is derived from the predictive coding framework. The predictive coding framework originated in perception sciences but pervades many domains of cognitive psychology (Bubic et al., 2010; Clark, 2013; Friston, 2005, 2008; Friston and Kiebel, 2009; Hickok, 2012; Kok et al., 2012). This framework delineates the transfer of predictions generated by higher-order cortical areas to lower-order cortical areas through feedback connections (e.g., Friston, 2005, 2010; Rao and Ballard, 1999; Summerfield et al., 2006). Lower-order areas match these predictions (also called perceptual templates) against incoming sensory input, and transfer the difference between the predicted and received input (i.e., 'prediction error') back to higher-order cortical areas through feedforward connections. Effects of prediction error can be observed on neural responses associated with perceptual and attention within the first 200 ms after stimulus onset (e.g., Den Ouden et al., 2012; Friston, 2005, 2010; Rauss et al., 2011). This dynamic interplay between higher- and lower-order cortical areas may underlie efficient hierarchical perceptual processing, because while perceptual expectations reduce overall response amplitude in primary cortex, they can facilitate perception by sharpening the sensitivity to differences between expected and unexpected input (e.g., Kok et al., 2012, 2014; Summerfield and De Lange, 2014). The sensory hypothesis applies this framework to language comprehension: higher-order language areas generate predictions and implement them as perceptual templates in primary auditory or visual cortex, namely as a representation of the sound of an upcoming spoken word or the visual appearance of an upcoming written word.

According to the word recognition hypothesis (e.g., Connolly and Phillips, 1994; Lee et al., 2012; Van den Brink et al., 2001,), effects of prediction manifest themselves not during pre-lexical perceptual processing but during the recognition of a specific word form, also called lexical access or lexical processing (e.g., Marslen-Wilson, 1987; Norris et al., 2000). Lexical processing makes available semantic and grammatical information associated with a word in long-term memory. Post-lexical processes then integrate the associated information with the sentence context to generate a sentence-level interpretation. ERP studies typically define pre-lexical, lexical and post-lexical processes in terms of sensitivity to a certain level of manipulation (e.g., Penolazzi et al., 2007; Molinaro et al., 2010). For example, an ERP component is thought to index pre-lexical processes if it shows sensitivity to physical differences (e.g., number of characters) but not to lexical factors (e.g., lexical frequency), to index lexical processes if it shows sensitivity to lexical factors but not to sentence-level manipulations (e.g., semantic incongruity), and to index post-lexical, semantic integration processes if it shows sensitivity to sentence congruity. According to the word recognition hypothesis, word form prediction causes people to recognize a word faster ('facilitated lexical access'), and this prediction effect manifests itself on early, pre-N400 ERP components that are associated with lexical processing and that are not sensitive to sentence congruity. Importantly, this hypothesis assumes that the N400 reflects a post-lexical, integration process that combines word meanings into a sentence-level representation (the "integration view"; Brown and Hagoort, 1993). However, an influential alternative view is that the N400 reflects a process of lexical-semantic access (e.g., Kutas and Federmeier, 2011; Kutas and Hillyard, 1984). The latter, alternative view has important implications for the interpretation of early, pre-N400 effects, and I will return to this issue in the Discussion section.

The distinction between the sensory hypothesis and the word recognition hypothesis is useful for the purpose of this review, because it allows for a basic categorization based on the locus of the purported

[2] One exception is a study by Dikker and Pylkkänen (2011), which investigated effects of picture context on comprehension of subsequent phrases, and was included because it followed-up on previous work by Dikker and colleagues.

[3] In a cloze probability test, participants complete sentences truncated before the target word with the first plausible completion that comes to mind. The cloze probability, or predictability of a word is the percentage of responses with that word. Some studies in this review compute word category cloze probability, the percentage of responses using a particular word category. The participants in the cloze test do not partake in the main EEG/MEG study.

[4] The term 'sensory hypothesis' is used because it was coined by Dikker et al. (2009). N.B., this hypothesis is about the effect of prediction on perceptual processes, not on actual sensory processes that detect initial physical stimulation.

predictability effect (i.e., an effect on non-linguistic/perceptual processing or on linguistic processing). Some of the reviewed effects are not associated with one or either of these hypotheses (the Early Inflection Positivity and ELAN), but they are included regardless just because they report early ERP effects of prediction. After reviewing the studies on prediction-related early brain responses listed in Table 1, I discuss some commonalities and discrepancies between these studies. I will then summarize the evidence from these studies for the sensory hypothesis and the word recognition hypothesis, and for the general hypothesis that people routinely predict word form. Finally, I discuss a number of open methodological and theoretical questions and suggest avenues for further research.

2. The early inflection positivity

In a benchmark ERP study on prediction during spoken story comprehension, Van Berkum et al. (2005) reported a very early ERP effect of prediction. This study derived support for noun-prediction from an observed ERP effect elicited by the noun-preceding adjectives, not from the early onset of this effect. However, the unusually early effect-onset has been interpreted by other researchers as supporting the sensory hypothesis (e.g. Dambacher et al., 2009), based on the assumption that effects observed within the first 100–150 ms after a stimulus must reflect early perceptual processing. Participants in Van Berkum et al. listened to Dutch mini-stories of two sentences that led to an expectation for a particular noun like ‘painting’ (1), with an average cloze probability of 86%. The stories either contained the expected noun or a semantically coherent alternative like ‘bookcase’. Van Berkum et al. capitalized on the Dutch grammatical rule whereby the grammatical gender of indefinite nouns determines the presence of an inflectional suffix on pre-nominal adjectives. In the Dutch equivalent of ‘big bookcase’, the adjective ‘grote’ (big$_{\text{com}}$) has the *e*-suffix in agreement with the common-gender noun ‘boekenkast’ (bookcase), whereas in ‘big painting’ the adjective ‘groot’ (big$_{\text{neu}}$) has no suffix before the neuter-gender noun ‘schilderij’ (painting). The inflection on ‘grote’ (big$_{\text{com}}$) therefore signals that the predicted word ‘painting’ is not going to follow. Van Berkum et al. examined ERPs that were time-locked to the onset of the adjective, to the onset of the inflection, and onset of the noun (Fig. 1).

(1) Example materials from Van Berkum et al. (2005), translated from Dutch.

Context Sentence: *The burglar had no trouble locating the secret family safe.*

Prediction-Consistent: *Of course, it was situated behind a* ***big$_{neu}$*** *but unobtrusive* ***painting$_{neu}$****.*

Prediction-Inconsistent: *Of course, it was situated behind a* ***big$_{com}$*** *but unobtrusive* ***bookcase$_{com}$****.*

When the ERPs were time-locked to word onset, prediction-inconsistent adjectives elicited a positive deflection in roughly the 500–800 ms time window (as can be seen in Fig. 1a, this window overlaps with noun onset in some of the items), and a positivity in the 300–400 ms range. However, neither effect was statistically significant in any of the reported analyses. Van Berkum et al noted that this lack of a significant effect could have been caused by variability in inflection-onset between items, which would smear out an effect that is presumably associated with inflection onset. In an analysis of ERPs time-locked to inflection-onset (Fig. 1b), prediction-consistency elicited a very early positive deflection, visible already at inflection-onset. Based on visual inspection, this effect was tested in the 50–250 ms time window, revealing significant effects in midline electrodes and in 3 electrode quadrants (all *p*-values in the 0.01-0.05 range). The ERPs elicited by the nouns showed a standard N400 effect of consistency (Fig. 1c). They also report an early negative peak in the 100–200 ms time window following the nouns. They briefly discuss this effect and argue that, based on its N400-like spatial distribution, the effect could have been the first part of the ascending flank of an early N400 effect (e.g., Van Petten et al., 1999), separated by the rest of the flank by a dip at approximately 200 ms presumably caused by residual alpha ‘noise’ (i.e., activity in the 8–12 Hz frequency band that is unrelated to the manipulation; alpha-band activity is indeed clearly visible in the ERP waveforms).

Van Berkum et al. also performed a control experiment wherein participants listened to the target sentences without the context sentences, so that no particular noun would be expected during comprehension. Consistent with the Van Berkum et al. conclusions, no reliable effect at the inflections or the nouns was observed.

2.1. Discussion

Van Berkum et al. (2005) is one of the first and best-cited ERP studies on linguistic prediction. It also stands out by using fluent and naturally-spoken stories that did not contain syntactically or semantically anomalous or implausible sentences. Its main claim about pre-activation of upcoming nouns was supported by the differential effect of prediction-consistency at the prenominal adjective. The specific latency and even the specific morphology of this effect is, in principle, of little consequence to that claim. But for this review the latency of the effect is of interest, and the question can be raised why this effect is indeed so early and whether its early latency constitutes evidence for the sensory hypothesis (e.g., Dambacher et al., 2009).

2.1.1. Why is this effect so early?

Van Berkum et al. already gave a comprehensive answer to this question, which is worth repeating here (see also Otten and Van Berkum, 2008). The onset of the inflection had been defined, by a trained native speaker of Dutch, as the point where the acoustic signal of the phoneme marking the inflection or no-inflection started to differ. But as Van Berkum et al. note, there might be earlier acoustical differences that Dutch speakers can pick up on, such as the co-articulatory and durational changes in the word stem, and syllable boundary cues (Zwitserlood, 2004). They estimated that critical inflectional information could have been available about 100–150 ms before the phoneme-based estimation. Because they used a 150 ms pre-inflection time window for baseline correction (subtracting average activity from a pre-target time window from the data), an earlier effect would not have been visible in the analysis. It could explain why the effect appeared as early as 0–50 ms after inflection onset. The true, underlying effect could have started at about 200–300 ms after the moment at which the acoustic signal started to differ between consistent and inconsistent conditions.

Van Berkum et al. note that a 200–300 ms effect onset latency is not unlike that of P600 effects, including those reported for prenominal gender-manipulations (e.g. Wicha et al., 2004). In spoken language studies, ERPs are typically time-locked to word onset (e.g., Hagoort and Brown, 2000) rather than to the time point where the acoustical information starts to differ between two conditions. Spoken language P600-effects may thus have a much earlier latency than is typically reported. Van Berkum et al. did not rule out that the observed early positivity was in fact similar to a ‘regular’ P600 effect, and noted that the results were inconclusive in this respect.

Follow-up studies by Van Berkum and colleagues with the same adjective-inflection manipulation do not corroborate the observed effect latency, and found a different effect of prediction-consistency appearing before the noun. In a written comprehension study, Otten and Van Berkum (2008) observed a late negativity in the 900–1200 ms time window. In a spoken comprehension study, Otten et al. (2007) observed a right-anterior negativity in the 300–600 ms time window. However, this effect was time-locked to adjective-onset, not to inflection-onset.

2.2. Conclusion

Van Berkum et al. (2005) is a benchmark ERP study that is widely cited among major theoretical review papers on linguistic prediction (Pickering and Garrod, 2007, 2013). It stands out as one of the more

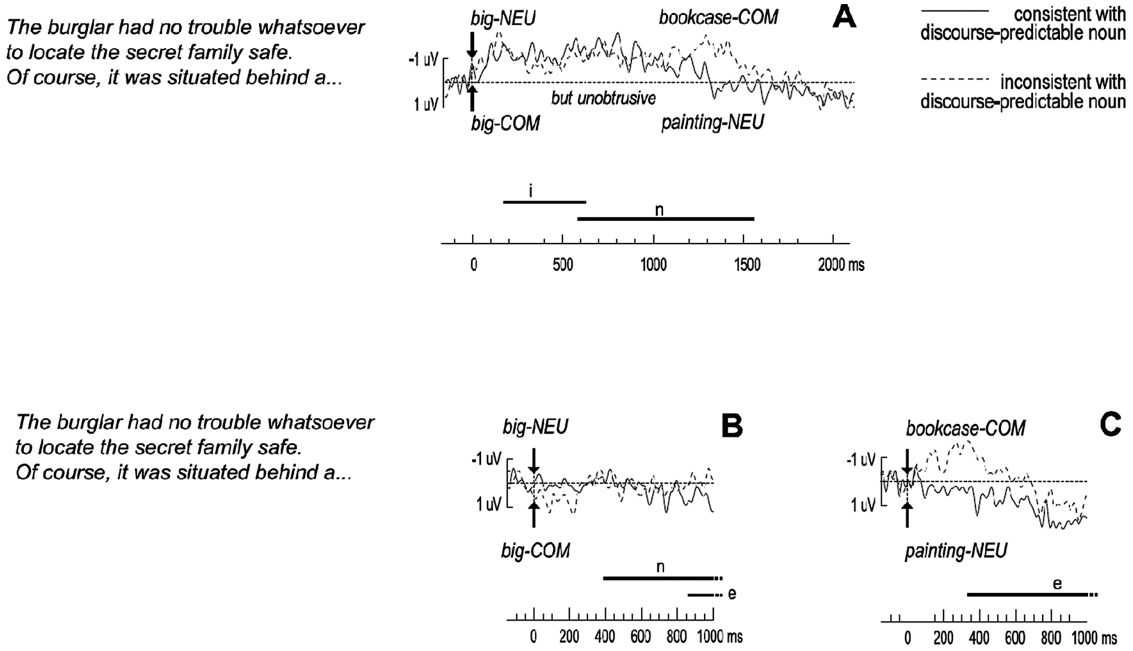

Fig. 1. Main results from Van Berkum et al. (2005). ERPs time-locked to onset of prediction-consistent or –inconsistent (A) adjective, (B) inflections and (C) nouns. Range and mean of the inflection onset and noun onset is marked with (i) or (n).

elegant demonstrations of prediction, wherein people listened to coherent and naturally-spoken stories (see also Otten and Van Berkum, 2009), devoid of the implausible or anomalous sentences often used in psycholinguistic experiments. The authors found a very early positive ERP effect (Early Inflection Positivity) elicited by prediction-inconsistent inflections on a prenominal adjective. They concluded that the participants had activated the nouns before they appear such that the nouns could impact the ongoing parsing operations.

As discussed by Van Berkum et al., the latency of the Early Inflection Positivity comes with an important caveat about the time-locking procedure. The effects were time-locked to the onset of the inflection but this may not be the earliest time point where prediction-consistent and -inconsistent conditions started to differ in their acoustical information. Van Berkum et al. suggested that their early effect may in fact not have been unusually early, but similar to P600 effects reported in the literature for grammatical gender manipulations (e.g., Hagoort and Brown, 2000). In other words, the Early Inflection Positivity supports the hypothesis that people can predict upcoming words but does not support the hypothesis that such predictions are instantiated at the earliest perceptual level, as has been argued (Dambacher et al., 2009), and has yet to be replicated.

3. The early cloze positivity

The earliest ERP effect observed on predictable words was reported by Dambacher et al. (2009). Their participants read German sentence-pairs with high- or low-frequency critical words in each second sentence (2). Based on the context sentence, the high-frequent ('ship') and low-frequent ('scepter') words were either high predictable (minimum cloze value of 50%, average cloze value 84%) or low predictable (maximum cloze value 10%, average cloze value 1%). The context sentence was presented as a whole, and each target sentence was presented one word at a time at a relatively fast pace of one word every 280 ms (the 'standard' word presentation rate in written experiments is 500 ms per word).

(2) Example item from Dambacher et al. (2009), translated from German

Context 1: *The man on the picture fiddled around with models of Columbus' fleet.*

Context 2: *The man on the picture wore a golden crown and sat stately on a throne.*

Target sentence: *In his right hand, he held a* ***ship/scepter*** *of considerable length.*

The authors reported that the ERP waveforms for low and high predictability started to diverge at about 50 ms after word onset (see Fig. 2). Compared to high predictable words, low predictable words elicited enhanced positivity at left-occipital channels and enhanced negativity at right-frontal channels, irrespective of word frequency. Because of this early time course, and the occipital scalp distribution of part of the positive effect, the authors concluded that these effects demonstrate rapid matching of form-specific, perceptual predictions with incoming visual input.

3.1. Discussion

The Dambacher et al. findings are intriguing because they suggest an unusually rapid impact of predictability on word processing. The authors stated that the observed effect latency was "considerably faster than most previous reports of interactions between top-down and bottom-up information in visual perception". Together with the occipital distribution of the effect, the early latency led the authors to claim evidence for the rapid verification of predicted visual input in the visual domain, in line with the sensory hypothesis.

The approach taken by Dambacher and colleagues has several noteworthy features. They used a counterbalanced design that compared the same word in different conditions, thus minimizing the impact of lexical variables. All materials were semantically and grammatically correct and their participants 'read for comprehension', that is, they read without performing a judgment task related to the manipulation but answering occasional comprehension questions. Dambacher et al. presented words at a fast rate in an attempt to 'mimic' a natural reading rate[5]. The Dambacher et al. results raise some important questions for future research.

[5] In natural reading, people read faster than the typical 500 ms per word in serial visual presentation procedures (e.g., Rayner, 1998). However, natural reading allows for parafoveal preview and for backtracking or slowing down when comprehension difficulty occurs. This is not possible in serial visual presentation. In some studies, participants reported that a pace of 300 ms per word was too fast to keep up with the unfolding sentence meaning (e.g., Camblin et al., 2007).

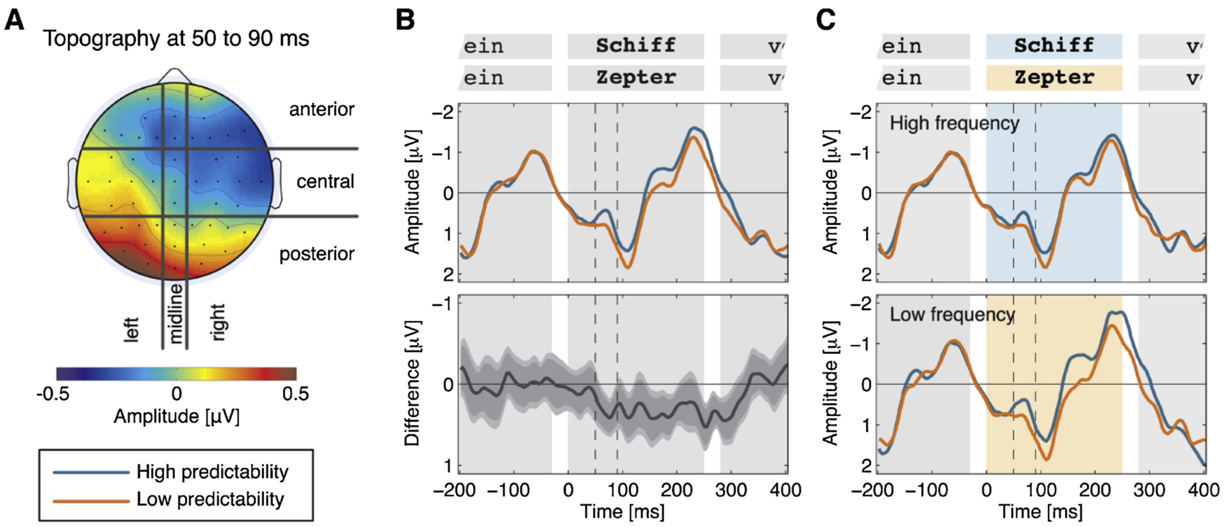

Fig. 2. Results from Dambacher et al. (2009; Fig. 4). Left graph (A): Topography of the ERP difference between high and low predictable words. Middle graphs (B): Condition and difference ERP waveforms for high and low predictable words. Right graphs (C): Condition ERP waveforms by critical word frequency.

3.1.1. *What does the early cloze positivity look like?*

Unlike most ERP studies on language comprehension, Dambacher et al. used a global average reference procedure, which subtracts the average activity across all channels from each individual channel for each time point. This method causes amplitudes to be smaller for centrally located electrodes than for peripheral electrodes (e.g., Curran et al., 1993), and can lead to effects of reversed polarity at opposite sides of the scalp (the polar average reference effect; Junghofer et al., 1999), as demonstrated for a P600 effect in Fig. 3.

The results of Dambacher et al also show a polar average reference effect (Fig. 2A). How problematic is this? At the least, it means that the scalp distribution, waveform morphology and effect-size of the Early Cloze Positivity cannot be directly and meaningfully compared to the results of studies that use a common mastoid-reference (i.e., the more typical procedure in psycholinguistics). While the global average reference is uncommon in psycholinguistic studies, it is common in studies on early perceptual ERP effects. In such studies, a mastoid reference is not optimal because it leads to dampened amplitudes at temporal and occipital electrodes of interest, because these electrodes are close to the mastoids (and are therefore more likely to have activity in common with the mastoids). In light of the polar average reference effect, however, more caution is needed in concluding that the effects are generated by sensory brain regions, since the effect could be an artefact of the reference procedure, and it would be wise to present results from different procedures. In addition, some authors warn that the common average reference procedure can yield unstable and distorted results when recordings are done with low-density (< 128) electrode arrays (e.g., Dien, 1998; Junghofer et al., 1999), and with unevenly distributed electrodes.

3.1.2. *How reliable and replicable is the early cloze predictability?*

The pursued analysis focused primarily on estimating the first time-point at which predictability generated an effect. In a mass univariate analysis, they tested the effect of predictability at every time-point and at every channel, which amounted to as many as roughly 9000 data-points (50 channels, 700 ms time window with 256 Hz sampling rate). Such an analysis calls for an appropriate correction for multiple comparisons. Dambacher et al. used an ad-hoc approach to multiple comparison correction in which they labeled an effect at a given time point as statistically significant if 3 consecutive time-points (> 10 ms) each showed a significant effect. However, this approach can be anti-conservative and give poor control of false positive errors (see Rousselet and Pernet, 2011; Groppe et al., 2011).

The Early Cloze Positivity has yet to be replicated. Two subsequent experiments with the same materials did not replicate this early effect (Dambacher et al., 2012, briefly discussed in the next section), although the analysis differed from the one used in Dambacher et al. (2009).

3.1.3. *What does the early cloze positivity reflect?*

Dambacher et al. tested for potential interaction effects of frequency and predictability. They averaged activity in the 50–90 ms post-noun time window based on the mass univariate test results[6] and reported a lack of a significant interaction in that time window. However, they did not test for the interaction pattern in later time windows. As is visible in Fig. 2, the effect of predictability appears larger for low-frequent words in that later window, and a later re-analysis of the same data set (Dambacher et al., 2012) reports an early interaction effect between frequency and predictability. This raises the possibility that the effects are not, or not only due to predictability.

One question is whether the Early Cloze Positivity was elicited by the critical words or whether it started earlier and continued into the time windows of interest. The pre-target words were identical for low- and high-predictable contexts, therefore it is unclear why such a pre-target effect would appear, but the ERPs for the two conditions did visually diverge in the pre-target window just before target onset (Fig. 2B in Dambacher et al.), possibly reflecting N400 differences. Effects tested in the pre-target window did not reach statistical significance (although p-values of .15 were reported), but this analysis used 40 ms bins instead of the by time-point analysis with liberal statistical threshold used for the target words.

The fact that the Early Cloze Positivity does not seem to be the modulation of one specific ERP component but rather presents a slowly developing divergence is compatible with an effect arising from pre-target N400 activity. If this alternative account is correct, then the Early Cloze Positivity may not occur at a slower presentation rate like 500 ms per word, even though such a slower rate can boost predictive processing (Ito et al., 2016, 2017; Tanner et al., 2017; Wlotko and Federmeier, 2015). At such a rate, differential activity elicited by pre-target words could be over by the start of the target word, or a baseline correction procedure would effectively minimize differences due to pre-target word N400 activity[7]. Consistent with this alternative account,

[6] This is a non-independent test procedure that can inflate the obtained results (e.g., Kriegeskorte et al., 2009; Vul and Pashler, 2012).

[7] At a rate of 500 ms per word, a pre-target baseline window of 200 ms minimizes differences due to activity in the 300-500 ms time window after pre-target onset.

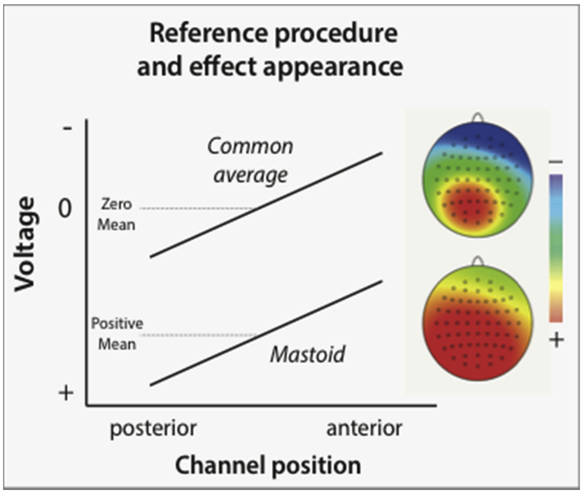

Fig. 3. Graphical demonstration of the polar average reference effect, applied to a hypothetical P600 effect. The mastoid reference procedure subtracts the average activity of left and right mastoid channels from all channels. With this common procedure, the widely distributed P600 effect is positive at all electrodes ('positive mean') but more positive at posterior channels than at anterior channels. The global average reference procedure subtracts the mean of all channels from each channel ('zero mean'). As a result, posterior channels show a smaller positive effect than before, whereas anterior channels now show a negative effect. This example focuses on the posterior to anterior dimension, but the principle holds for any dimension along the scalp (e.g., left to right hemisphere).

Dambacher et al. (2012) did not find early prediction effects at slower rates (490 or 700 ms per word).

3.2. Summary

Dambacher et al. (2009) reported the earliest effect to date of predictability on ERPs elicited by predictable words. They found an effect of word predictability as early as 50 ms after word onset (the Early Cloze Positivity). Based on this early latency, and the occipital distribution of the effect, they concluded that prediction impacts the early perceptual processes initiated by a word, in accordance with the sensory hypothesis.

Their findings suggest a remarkably early impact of predictability on processing, as no other written language study has reported prediction effects as early as 50 ms after word onset. The interpretation of their results is aided by several strong features of the experimental approach. They used a counterbalanced design to minimize the impact of lexical variables, and their participants read semantically and grammatically correct two-sentence stories, and a fast pace to mimic a natural reading rate. But some key aspects of the Early Cloze Positivity remain unclear and need to be addressed in dedicated follow-up.

It is unclear whether this effect reflects modulation of a specific ERP component or a slowly developing effect spanning multiple components. It is unclear whether this effect occurs only at rapid presentation procedures. No other studies have reported such early cloze effects, and two follow-up studies by the same authors (Dambacher et al., 2012), with the same materials but a slower presentation procedure, failed to find such early effects. Dambacher et al. (2009) explained the early onset of their effect in terms of their high-demanding, fast reading procedure, arguing that this procedure enhanced participants' use of perceptual information. However, this account is tentative and inconsistent with a larger body of ERP studies which suggests that faster presentation rates hamper predictive processing compared to slower rates (Camblin et al., 2007; Dambacher et al., 2012; Ito et al., 2016, 2017; Tanner et al., 2017; Wlotko and Federmeier, 2015). It remains unclear to what extent the Early Cloze Positivity results from condition differences in neural activity in the pre-target window.

In sum, the Early Cloze Positivity reported by Dambacher et al. (2009) seems to be an isolated finding that is at odds with the results of various other studies, and does not seem to offer straightforward support for the sensory hypothesis.

4. The early left-anterior negativity (ELAN)

The ELAN is a well-known early ERP effect often associated with the detection of phrase structure/word category violations (e.g., "The scientist criticized Max's of"; Friederici et al., 1993; Hahne and Friederici, 1999; Neville et al., 1991). The ELAN appears around 100–200 ms after word onset at left-anterior channels (but see Osterhout et al., 2004; Steinhauer and Drury, 2012; Tanner, 2015, for a critical discussion of the nature and interpretation of ELAN effects). The ELAN's early onset has been taken as evidence that the language system very rapidly detects word category violations, and that structure building operations take place even before people access the meaning or morphological features of a word (e.g., Friederici, 2002).

4.1. Lau et al. (2006)

Lau et al. (2006) proposed an explanation for the ELAN phenomenon wherein people generate online predictions about word category. If sentence context allows people to predict word category, then a diagnosis of grammaticality only requires a quick check whether the incoming word matches the predicted category. Such a quick check could happen before the meaning of a word is retrieved and before agreement-relationships with context words are computed, allowing for very rapid violation detection. To test their hypothesis, Lau et al. used an elegant design that manipulated word category predictions via the use of ellipsis (3), crossing the contextual availability of ellipsis with grammaticality of the target word ('of' in the example). All ungrammatical sentences had the possessor + function word manipulation (Max's of) originally used by the classic study by Neville et al. (1991). The experimental logic was as follows: When there is a possessor in the first phrase (ellipsis conditions), the possessor in the second phrase allows for an elliptical reading that would be a grammatical end-of-phrase ('Dana's' means 'Dana's mother'). Because the word after the possessor therefore is not required to be a noun, the possessor is unlikely to trigger a word category prediction. Without possessor in the first phrase (no-ellipsis conditions), a possessor in the second phrase requires an overt noun to continue the sentence in a grammatical way, which could trigger the prediction of an overt noun. Lau et al. reasoned that if the ELAN reflects the violation of a word category prediction, then they should observe a reduction of the ELAN effect in the ellipsis-conditions because people entertain weaker noun-predictions.

(3) Example item in all four conditions from Lau et al. (2006)

Ellipsis-Grammatical:

Although Erica kissed Mary's mother, she did not kiss the daughter ***of*** *the bride*

Ellipsis-Ungrammatical:

Although Erica kissed Mary's mother, she did not kiss Dana's ***of*** *the bride*

No Ellipsis-Grammatical:

Although the bridesmaid kissed Mary, she did not kiss the daughter ***of*** *the bride*

No Ellipsis-Ungrammatical:

Although the bridesmaid kissed Mary, she did not kiss Dana's ***of*** *the bride*

In the ERP experiment, participants read 128 such sentences along with a large number of filler sentences so that only 25% of all sentences were ungrammatical. Participants judged the grammaticality of each sentence after it finished. The ERPs elicited by the target words were tested statistically in 4 time windows (0–200 ms, 200–400 ms for the ELAN, 300–500 ms for the N400, and 600–1000 ms for the P600). In the

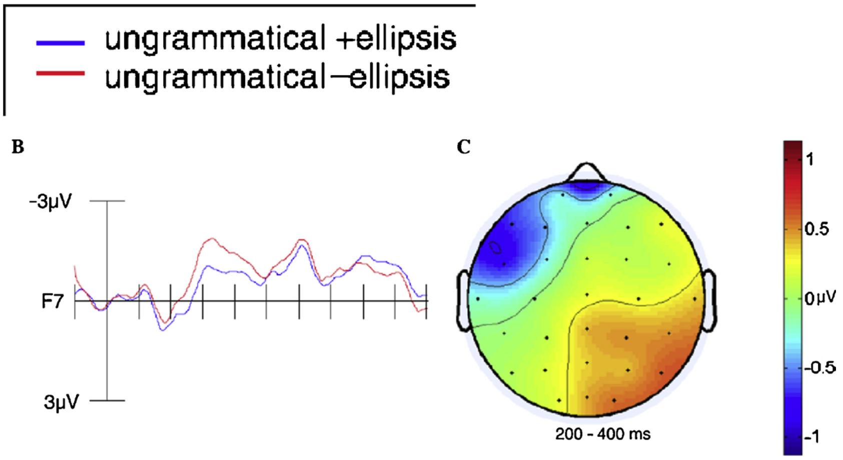

Fig. 4. Main result from Lau et al. (2006), the ELAN effect observed at left-anterior channel F7 for ungrammatical sentences without ellipsis compared to ungrammatical sentences with ellipsis, accompanied with a positive effect at right-posterior channels (consistent with a polar average reference effect).

first-reported analysis, the ERP data were re-referenced to the global average of all electrodes, but results from a mastoid reference procedure were reported as an additional analysis. Unlike previous ELAN studies that tested for an effect of grammaticality (e.g., Neville et al., 1991; Hahne and Friederici, 1999), Lau et al. only tested for the effect of ellipsis in either the grammatical or the ungrammatical sentences, because they observed effects of grammaticality already at the pre-critical word ('Dana's' or 'daughter' in the example), which they argued came from the comparison of ERPs elicited by different lexical items.

The critical words ('of') elicited a left anterior negativity in the 200–400 ms time window (ELAN) in no ellipsis-ungrammatical sentences compared to ellipsis-ungrammatical sentences (see Fig. 4). The ELAN finding supported their hypothesis that availability of ellipsis reduces word category-prediction and therefore weakens the ELAN. For the grammatical sentences, the authors reported a 'scarcity of reliable contrasts' (they observed some statistical significant effects but no clear pattern emerged), and they concluded that ellipsis did not meaningfully affect online processing in grammatical sentences.

4.2. Discussion

Lau et al. were the first to investigate effects of syntactic context on the ELAN, a well-known syntactic ERP effect associated with rapid and automatic detection of word category violations. They used an ingenious design based on the canonical study by Neville et al. (1991), manipulating the word category expectations through availability of an elliptical interpretation. A major strength of this study lies in its controlled design which compared two conditions with the exact same words in the subordinate clause. As discussed by Lau and colleagues, a weakness of their study (or of their data) was that they could not meaningfully compare grammatical and ungrammatical sentences. This lack of standard grammaticality comparison, along with other aspects of the study, complicate a direct comparison with previous research and therefore an identification of an ELAN effect.

4.2.1. Is this really an ELAN effect?

Like the Dambacher et al. (2009) study, Lau et al. relied on the global average reference procedure. This procedure can create a polar average reference effect (Fig. 3; Junghofer et al., 1999), which is also visible for the Lau et al. result (Fig. 4). The global average reference procedure could have transformed a more widely distributed effect with a posterior focus into an ELAN effect plus a positive ERP effect limited to peripheral posterior electrodes. Lau et al. also report ERPs from mastoid reference data (Appendix, Fig. 4 in Lau et al.), which showed no effect at left-anterior sites, ELAN or otherwise, but showed that the two ungrammatical conditions start to diverge from as early as 0 ms, particularly at posterior channels. Lau et al. dismissed the results from the mastoid reference procedure by arguing that it would not be sensitive to pick up condition effects if the mastoid electrodes themselves showed such effects. This may be true, but this would be true for earlier ELAN studies as well, and left-anterior channels are generally not strongly affected because they are far away from the mastoid. Furthermore, the global average reference procedure is not recommended for low-density montages (Dien, 1998), especially in a research field where a mastoid reference procedure is common, such as in previous ELAN studies (Friederici et al., 1993; Hahne and Friederici, 1999; Neville et al., 1991). Also of importance, the global average reference procedure in Lau et al. masks P600 effects of grammaticality, whereas such effects are consistently observed in other studies (e.g., Neville et al., 1991). Lau et al. did not test for effects of grammaticality, but visual inspection of their ERP plots with global average reference suggests that ungrammatical sentences barely elicited P600 effects compared to grammatical sentences (Fig. 5A versus 5B). Although the mastoid-referenced ERPs for the grammatical conditions were not shown, an enhanced P600 for the ungrammatical conditions is clearly visible with the mastoid reference (Fig. 5C).

In sum, the ELAN-like effect is only visible with the global average reference procedure, and there is no sign of a 'standard' P600 effect of grammaticality. Both these aspects of the results do not correspond to previous results and suggest alternative interpretations.

4.2.2. What is the role of the pre-critical words and the task?

Irrespective of reference procedure, Lau et al. also found effects of ellipsis in the 0–200 ms time window. ERPs for the no-ellipsis-ungrammatical sentences were immediately more positive than those for ellipsis-ungrammatical sentences. Effects starting at the onset of the target word are implausible and suggest that the conditions differed before the onset of the target word.

Lau et al. also analyzed effects elicited by the pre-critical words. These results are also consistent with a polar average reference effect: in the 300–500 ms time window, ungrammatical sentences elicited a negativity at anterior channels and a positivity at posterior channels. Lau et al. argued that these effects reflected the different ERPs associated with different lexical items (e.g., 'Dana's/daughter'). But they also found a small but significant main effect of ellipsis, even though that comparison involves identical lexical items. There was no significant interaction between ellipsis and grammaticality, but the ellipsis effect was very small and only visible for grammatical sentences, not for ungrammatical sentences (the ones they were most interested in). However, one should keep in mind that the global average reference procedure could have masked more widespread differences between conditions in this time window.

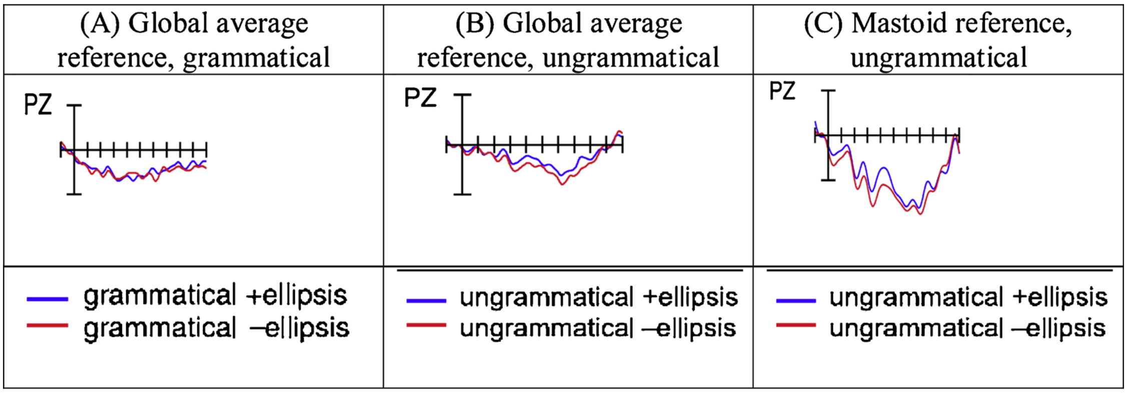

Fig. 5. Effect of reference procedure on P600 activity at electrode Pz in Lau et al. (2006). For each ERP graph, the x-axis spans from -3 to 3 μV, and the y-axis spans from -100 to 1000 ms relative to critical word onset.

Effects at the pre-critical words matter because these words were relevant for the grammaticality judgment task. The participants could have strategically paid attention to the possessor words in the second phrase (Dana's) because they cued impending sentence ungrammaticality, and this attention-strategy could have played out differently depending on whether the first phrase also contained a possessor word. This is important because task-relevant words can elicit ERP effects that one would not observe without task (e.g., Baggio, 2012; Roehm et al., 2007; Vega-Mendoza et al., 2017), effects that could reflect evaluation of the materials, decision-making, and possibly response planning (Polich and Kok, 1995). Moreover, task-associated ERPs can be of fairly long duration and spill-over into the time window of the critical word, causing differences to appear early as 0 ms after word onset.

4.3. Summary

The study by Lau et al. is a well-known ERP demonstration of word-category prediction, and it features in prominent reviews on linguistic prediction (e.g., Hagoort, 2017; Pickering and Garrod, 2013). It followed up on the seminal ELAN paper by Neville et al. (1991) with a clever ellipsis-based manipulation. A possessor-noun combination (e.g., Mary's mother) in the first phrase of a sentence presumably led to weaker expectations that a new possessor (Dana's) was also followed by a noun. Lau et al. reported that such weakening of word category expectations also reduced the ELAN. They concluded that word category predictions allow people to reduce a word's grammaticality diagnosis to a mere check of its category, which could explain why ELAN effects occur so early, and why violations of word categories are detected so rapidly. As I discussed in this review, however, the support for this conclusion is equivocal and several questions remain to be addressed.

While other ELAN studies (e.g., Neville et al., 1991) found ELAN effects and clear P600 effects of grammaticality with a mastoid-reference procedure, Lau et al. did not, which complicates the comparison with previous literature. The conclusion about the ELAN rests entirely on which reference procedure one thinks should be used, and – in this case - the arguments *for* using a mastoid-reference are stronger than the argument *against*.

With mastoid-reference, Lau et al. found an early-onset positive-going ERP waveform for ungrammatical sentences that started to differ as a function of ellipsis immediately at word onset (0 ms). This effect is too early to be a plausible reflection of violation detection, and could be a spill-over effect of ERP activity elicited by the pre-critical possessor words. In this particular design, exacerbated by the demands of the grammaticality judgment task, the possessor words may have generated condition-related effects of their own. This raises the question whether similar results would be obtained in a design where the possessor words would not be task-relevant.

Notably, a recent study by Kaan et al. (2016) failed to replicate the Lau et al. ELAN effects in a similar paradigm, regardless of reference procedure. Therefore, the Lau et al. have yet to be replicated and the support from this paradigm for word category prediction remains to be established.

5. The visual M100

The visual M100 is an early magnetic brain response that originates from primary visual cortex (V1) and is associated with processing of basic visual features like luminance and stimulus size, regardless of stimulus category (Tarkiainen et al., 1999, 2002). The M100 is distinct from the M170, the earliest brain response that shows sensitivity to words and is thought to originate from the visual word form area (e.g., Carreiras et al., 2014; Dehaene et al., 2005; Solomyak and Marantz, 2009). In a series of MEG studies, Dikker and colleagues (Dikker et al., 2009, 2010; Dikker and Pylkkanen, 2011) investigated the effects of linguistic prediction on early perceptual processes as indexed by the M100 response originating from primary visual cortex. The first two studies tested for early effects of word category prediction and the third study tested for early effects of lexical prediction. I will describe each study separately and then discuss them together.

5.1. Dikker et al. (2009)

Like Lau et al. (2006), Dikker et al. (2009) asked whether ELAN effects appear early because people predict word category. Dikker et al. posited the *sensory hypothesis:* brain activity in the first 100–150 ms after visual stimulus onset is known to be dominated by perceptual processing, and because the early-appearing ELAN[8] reflects word category predictions (Lau et al., 2006), these predictions must play out at the level of primary perceptual processes, such that ELAN-like effects in written language studies are generated in the occipital lobes. To test this hypothesis, their first experiment contrasted different types of word category violations (4). In their design, the critical word was from the expected category (grammatical) or from an unexpected category (ungrammatical). They predicted enhanced visual M1/M170 responses to these violations only if word category membership could be quickly established by detection of a high-frequent, closed-class morpheme. Specifically, they predicted these responses to prepositions ('about', 'of') or verbs with regular verb-inflection ('ed' in 'reported'), but not to nouns without overt category marking ('report'). Of note, the participle and bare stem target words were always sentence-final, prepositions were never sentence-final.

[8] The ELAN effect reported by Lau et al. (2006) started at 200 ms after word onset, at left-anterior channels. Dikker et al. did not mention whether they considered this latency and scalp distribution to be indicative of perceptual processing.

(4) Example materials from Dikker et al. (2009; Experiment 1)

Preposition-Expected: *The boys heard Joe's stories* ***about*** *Africa.*

Preposition-Unexpected: *The boys heard Joe's* ***about*** *stories Africa.*

Participle-Expected: *The discovery was* ***reported****.*

Participle-Unexpected: *The discovery was in the* ***reported****.*

Bare stem-Expected: *The discovery was in the* ***report****.*

Bare stem-Unexpected: *The discovery was* ***report****.*

The participants each read a total of 540 sentences (300 target sentences and 240 filler sentences), and judged the grammaticality of each sentence. It is important to note a few aspects of the design: the sentence context before the critical word always differed between expected and unexpected words (for relevant discussion, see Lau et al., 2006; Osterhout et al., 2004; Steinhauer and Drury, 2012), the design was fully within-subject so that each participant saw each item in both the expected and the incorrect condition, and the preposition sentences had only two possible target words ('about' or 'of'). In a separate cloze completion task, Dikker et al. established the expectedness of each word category. An independent group of participants completed versions of the sentences that were truncated before the target word. The results confirmed that the first word of the completion was almost never from the same word category as the unexpected target word, whereas 44% of the preposition sentence completions started with a preposition, 29% of the participle sentence completions started with a participle, and 91% of the bare-stem sentence completions started with a noun.

Dikker et al. measured early visual activity using a neural source reconstruction procedure called dipole modelling. This involved an estimation of the sources underlying sensor-level activity in the 200 ms after target word onset, and subsequently reconstructing the activity associated with these sources (M100 or M170) and testing for activity differences between conditions. They tested the effects of the conditions on M100 amplitude in a 15 ms time window around the M100 peak. M170 responses from a separate source reconstruction were also tested but did not generate a clear pattern. Unexpected prepositions and participles elicited larger M100 s than expected ones, whereas no such effect was found for bare stems (Fig. 6, left graphs). For participles, an effect on the post-M100 deflection, perhaps related to the M170 response, was visible but not analyzed. In addition to dipole-modelling, Dikker et al. performed analysis at sensor-level data, using the most posterior channels close to primary visual cortex. This analysis did not reveal a significant interaction between word type and expectedness. Based on the bilateral distribution of the M100 response, they tested left and right sections separately. However, they did not report an interaction effect involving hemisphere, and in the source analysis they had modeled the M100 with a single dipole for activity from both hemispheres. In the left hemisphere, they found an expectedness effect that did not differ significantly between word types, but when they tested the effect of expectedness for each word type separately, the effect was significant for prepositions and participles but not for bare stems.

The authors took these results as "a rather strong confirmation of the sensory hypothesis" (p. 360), but they also discussed two potential confounds that led them to perform a second experiment. The bare stems in Experiment 1 possibly failed to generate an expectedness effect because their word category was ambiguous between nouns and verbs (e.g., 'report'), and/or because unexpected base stems did not mismatch a very strong category prediction (only 29% of completions of these context sentences like 'The discovery was', used participles).

To address these confounds, Experiment 2 again included the bare stem sentences and participle sentences (but not the preposition sentences), and also included sentences with category-unambiguous nouns (e.g., 'tree'), as in the example in (5). In addition, for each type of sentence a manipulation of prediction strength was included by the presence or absence of an adverb or adjective before the critical word. Cloze completions confirmed the increase in participle-expectancy associated with the adverbs (with adverbs 79% of the completions were participles, without adverbs only 29%). The participants each read 720 sentences in total: 60 per cell in the design that crossed word type with expectedness and prediction strength. No filler sentences were included. Of note, in this design the presence of "in the" and an adverb was enough to know the upcoming category violation.

(5) Example materials from Dikker et al. (2009; Experiment 2), prediction strength was considered 'strong' when sentences included an adverb or adjective, here in parentheticals, and 'weak' if no adverb or adjective was presented.

Participle-Expected: *The discovery was (solemnly)* ***reported****.*

Participle-Unexpected: *The discovery was in the (solemnly)* ***reported****.*

Bare stem-Expected: *The discovery was in the (solemn)* ***report****.*

Bare stem-Unexpected: *The discovery was (solemnly)* ***report****.*

Noun-Expected: *The owl was in the (high)* ***tree****.*

Noun-Unexpected: *The owl was (highly)* ***tree****.*

The analysis for Experiment 2 was similar to that of Experiment 1, with dipole-model reconstruction of M100 source activity, and amplitude around the M100 peak as the dependent measure in an ANOVA test. As in Experiment 1, they found a statistically significant interaction between expectedness and word type (Fig. 6, right graphs). Although no statistically significant 3-way interaction was reported, the authors proceeded by testing for the 2-way interaction between prediction strength and expectedness in each word type. Unexpected participles elicited slightly enhanced M100 s compared to expected participles, but prediction strength did not modulate this effect. Overall, critical words in the strong prediction sentences elicited enhanced M100 s compared to those in weak prediction sentences, but this prediction strength effect did not interact with word type or correctness.

In the two experiments, Dikker et al. found that word category violations only elicited an M100 effect if the target word category is marked by a closed-class morpheme like '-ed' in participles. They concluded that people predict the basic visual features of the category that words belongs to ('form-based estimates'). Using these form-based estimates, primary visual cortex can rapidly detect the presence of overt function morphemes. Through this prediction-based detection, morphemes rapidly disconfirm the predicted word category. Violation detection therefore does not require any deep semantic analysis, but operates rapidly through analysis of basic visual features.

5.2. Dikker et al. (2010)

In a follow-up study, Dikker et al. (2010) examined whether closed-class morphemes are indeed critical in generating early M100 effects of word category violations, or whether the M100 response in visual cortex is more generally sensitive to any form-features that cue a particular word category. They addressed this issue by testing for effects of phonological typicality, the phenomenon that some nouns and verbs contain phonological form features that are typical of their respective category, while other, neutral nouns and verbs have features that are equally common in each category (see Farmer et al., 2006). Phonological typicality may influence written language comprehension (but see Staub et al., 2009, 2011), presumably because it is correlated with orthographical typicality. Their design included expected/unexpected target sentences (6), which contained bimorphemic nouns with closed-class category-marking morphemes (e.g., 'princess'), monomorphemic nouns that were phonologically typical (e.g., 'soda'), or neutral nouns (e.g., 'infant'). No cloze-test was reported to establish category expectancy of the materials. Participants read 480 sentences, 240 target sentences (40 items in each of these 6 conditions) and 240 filler sentences that contained an expected or unexpected participle verb (like in Dikker et al., 2009) to counterbalance expectancy of nouns and participles. Participants evaluated the grammaticality of each sentence in a judgment task.

(6) Example materials from Dikker et al. (2010)

Bimorphemic Nouns, Expected/Unexpected: *The beautiful/beautifully* ***princess*** *was painted.*

Typical Nouns, Expected/Unexpected: *The tasteless/tastelessly* ***soda*** *was marketed.*

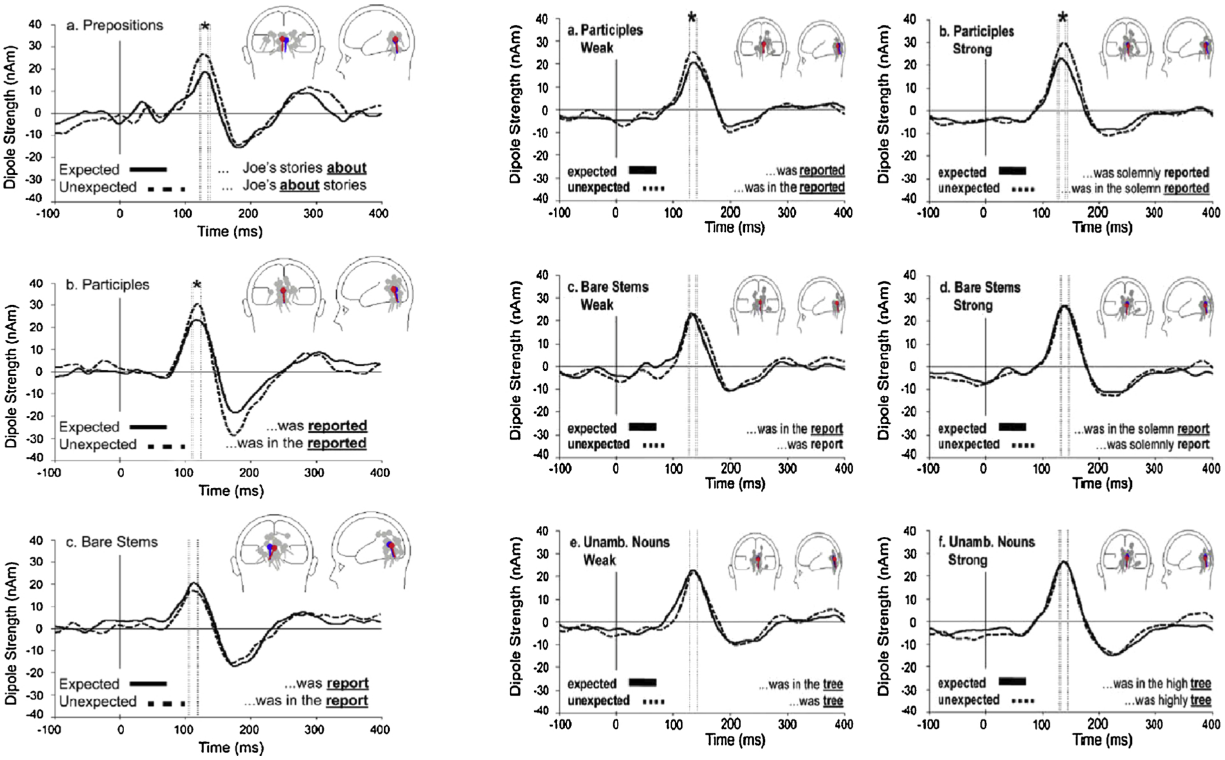

Fig. 6. M100 results from Experiment 1 (left graphs) and Experiment 2 (right graphs) of Dikker et al. (2009). Left graphs, from top to bottom: M100 effects of word category expectedness for prepositions, participles and bare stems. Right graphs, from top to bottom: M100 effects of word category expectedness in weak and strong predictive sentences, for participles, bare stems and unambiguous nouns.

Neutral Nouns, Expected/Unexpected: *The cute/cutely* ***infant*** *was dressed.*

No sensor-level analysis was reported. Dipole modelling followed the procedure of Dikker et al. (2009). A categorical ANOVA showed that the effect of expectedness on M100 amplitude depended on word type: bimorphemic and typical nouns elicited a statistically significant effect of expectedness but neutral nouns did not (see Fig. 7). Effects of expectedness on the participle filler-sentences were not analyzed or reported. In a single-trial, mixed-effects model analysis[9], the authors showed that M100 amplitude gradually increased with decreasing form typicality (the numerical distance between noun form typicality and the average form typicality of all nouns from a corpus count). The results suggested that there was no additional effect of a closed-class morpheme, and that the M100 effects were primarily a function of the mismatch of the item with the phonological typicality of the expected word category.

Dikker et al. (2010) concluded that readers predict word category at the level of very basic (non-lexical) visual form features that are probabilistically associated with that category. Another interpretation, wherein M100 responses in visual cortex directly reflect word category detection, was deemed unlikely based on established insensitivity of the M100 to lexical variables (Salmelin, 2007; Tarkiainen et al., 1999). These results thus extend the Dikker et al. (2009) results by showing that while presence of a closed-class morpheme seems sufficient to elicit the M100 violation effect, it is not a necessary condition. Whether Dikker et al. (2010) replicated the M100 expectedness effect that they previously found for participles is unknown because those results were not reported.

5.3. *Dikker and pylkkänen (2011)*

Dikker and Pylkkänen (2011) investigated whether people predict visual form features of specific lexical items. In a picture-word matching task, the participants first saw a picture and then read a noun phrase referring to an object that was or was not in the picture. In the strong prediction condition, the picture was of a single, clearly identifiable object. In the weak prediction condition, the picture was either of a grocery bag with groceries or an ark with animals to represent the semantic category 'food' or 'animals', respectively, to which each target word could belong. The first half of the experiment only contained strong prediction trials, but the second half mixed weak and strong prediction trials.

Their analysis did not involve source-reconstructed M100 activity, but only sensor-level activity. Sensors were selected based on their overall M100 responsiveness, which were 3 right-hemisphere and 12 left-hemisphere sensors. A weighted averaged was then computed between hemispheres after resigning sensors from one hemisphere so that left and right hemisphere M100 responses were of the same sign. Analysis was based on a 10 ms window around the peak at 97 ms.

An analysis of the strong prediction items revealed that lexical mismatch elicited stronger M100 responses than match (Fig. 8), regardless of experiment-half. An analysis of the weak prediction items revealed no effect of mismatch. In a later time-window after word onset (250–400 ms), a statistically significant, N400-like effect of lexical match was found for strong prediction items but not for weak prediction items (but no interaction analysis was reported). The N400-like activity for the matching or mismatching weak prediction items was similar to that of the mismatching strong prediction items. Based on their results, Dikker and Pylkkänen (2011) extended the sensory hypothesis and argued that people predict the visual form features of specific lexical items when possible (although they did not specify the conditions that must be met for such predictions).

[9] This analysis did not contain random slopes for the effects of interest, which assumes that the effect of predicted typicality mismatch (or of other factors included in the analysis) did not vary between subjects and items (for discussion, see Barr et al., 2013).

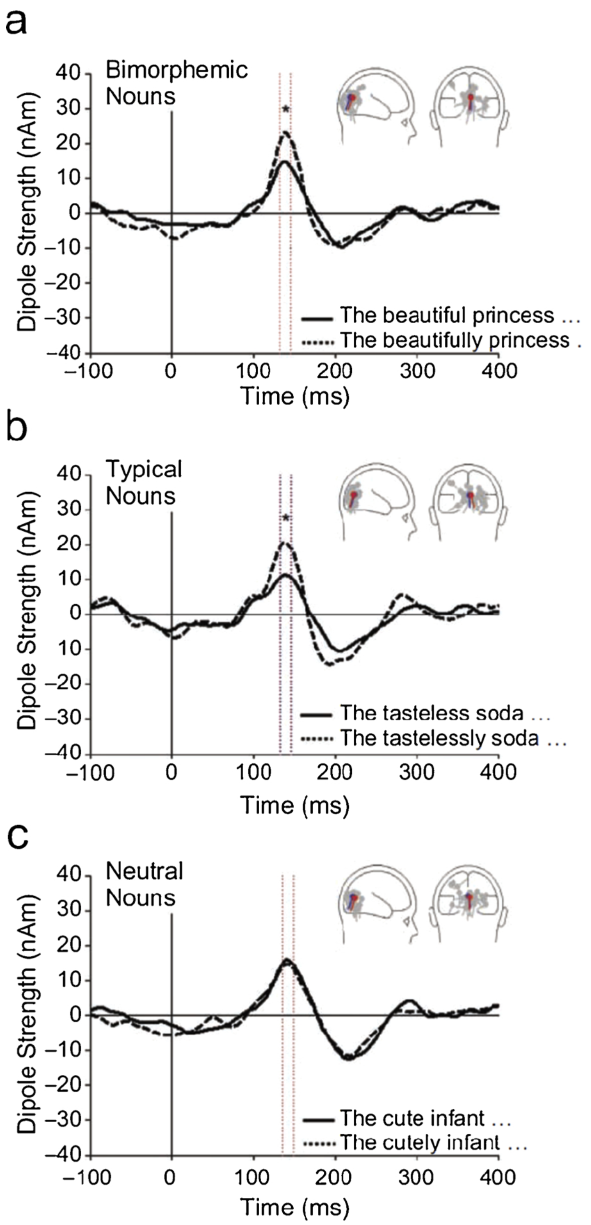

Fig. 7. M100 results from Dikker et al. (2010). From top to bottom: M100 effects of word category expectedness for bimorphemic nouns, typical nouns and neutral nouns.

5.4. Discussion

Dikker and colleagues conducted an impressive series of cleverly designed experiments to test the *sensory hypothesis*, the strongest hypothesis to date regarding the level of representational detail at which linguistic predictions play out. According to the sensory hypothesis, predictions involve top-down pre-activation of basic visual features as represented in primary visual cortex. Word category predictions pre-activate visual features associated with that category, and, likewise, lexical predictions pre-activate the basic visual features of a specific word. Dikker et al. reported M100 evidence that is consistent with the sensory hypothesis. Because previous research has shown that M100 activity does not distinguish between words and non-words (e.g., Salmelin, 2007), Dikker et al. argued that the visual form features that are pre-activated must be of a sub-lexical, perceptual nature. These results are often cited as strong evidence that predictions propagate all the way down to the lower-level of the processing hierarchy (Clark, 2013; Lupyan, 2015; Pickering and Garrod, 2013).

The M100 studies by Dikker and colleagues are highly innovative and have delivered captivating results. Hopefully, these results will spark further research to elucidate the perceptual correlates of linguistic prediction. Here, I discuss a few relevant questions that remain to be addressed.

5.4.1. What is the role of prediction strength?

Dikker and colleagues conclude that their M100 results reflect the top-down effects of prediction on visual feature process, not the process of anomaly detection itself. Following this argumentation, stronger predictions should generate stronger effects. However, results from the three studies do not directly support this.

In Dikker et al. (2009), stronger word category predictions did not elicit greater M100 expectedness effects, thus failing to support the sensory prediction hypothesis. In fact, strong category predictions led to reliably higher amplitude M100 responses overall. This seems inconsistent with the predictive coding framework, which holds that predictions should reduce neural response amplitudes (e.g., Kok et al., 2012).

In Dikker et al. (2010), there is no measure or manipulation of prediction strength. No cloze values were reported, so it is unclear whether the sentences with adverbs led to strong expectations for participles, and whether the different adverbs in the three noun type conditions led to equally strong expectations. This is relevant because some adverbs might lead people to expect adjectives instead of participles (e.g., ‘fashionably late/dressed’, ‘finally ready/ended’). The expectedness of the nouns in the three conditions is also unknown. Category cloze values are also relevant for the single-trial analysis. There, the relevant measure of phonological typicality is the distance between the target word's typicality and the mean typicality score of the expected word category. However, this assumes a word category expectedness of 100%. An alternative, and perhaps more appropriate measure would scale the mean category typicality score to the strength of expectation for that category.

Dikker & Pylkkänen (2011) argued that a mismatch M100 effect was observed only when a single word could be predicted (strong prediction conditions). However, they did not report an analysis that tested the impact of prediction strength on the mismatch M100 effect. They reported a significant mismatch effect (on selected electrodes) in strong prediction conditions and no significant effect in weak prediction conditions, but no interaction test was performed, and the difference between significant and non-significant itself may not be significant (see Nieuwenhuis et al., 2011; Gelman and Stern, 2006). Furthermore, the M100 responses to matching words did not visibly differ between strong and weak prediction conditions (although there was no statistical test reported), failing to generate support for their conclusion.

In sum, while the Dikker et al. studies reported effects of the match between an incoming word and the sentence or picture context, the role of prediction strength in these experiments remains elusive.

5.4.2. What does it mean to predict visual features?

The conclusion that readers predict basic visual features was based on the established sensitivity of the M100 to low-level visual features, such as the noise-level and size of letter-strings, and insensitivity to stimulus content or lexical features (Tarkiainen et al., 1999; Salmelin, 2007). But Dikker and colleagues themselves do not test or demonstrate sensitivity to low-level features like word length. In Dikker et al. (2009, 2010) there were large differences between word types in average word length, but in neither of the experiments did the M100 increase overall with word length. In Dikker et al. (2010), the single-trial analysis did not yield a reliable effect of word length on the M100[10], even though the critical words varied considerably in length.

[10] Several variables were included that may correlate with word length (morpheme presence, orthographic length, number of syllables, phonological length).

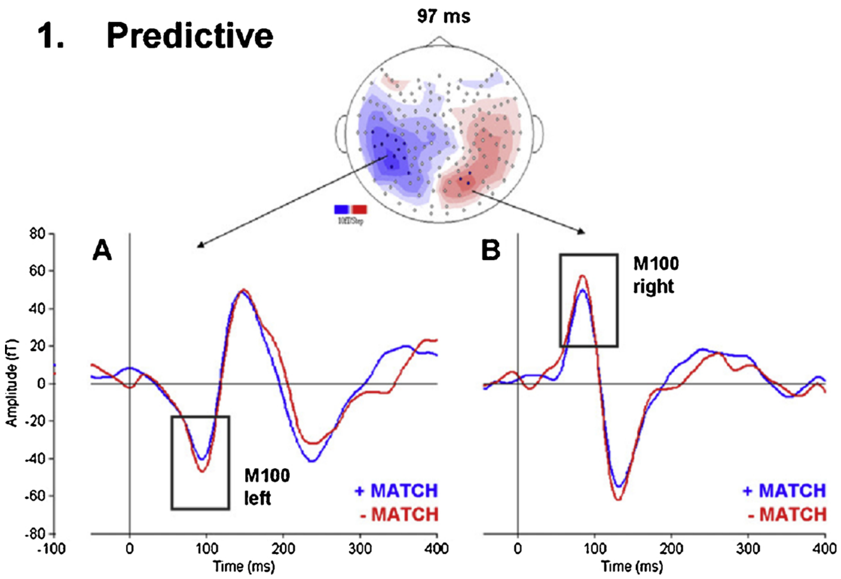

Fig. 8. Main results from Dikker & Pylkkänen (2011). Sensor-level M100 activity for picture-matching and –mismatching words, for predictive and non-predictive contexts.

This leads to a somewhat strained conclusion that M100 is more sensitive to presumably very fine-grained visual differences between predicted category features and the visual input, than to rather coarse-grained, string-size differences between actual inputs. Maybe this requires a revised theory of what the M100 reflects, or perhaps the design was simply not able to detect string-size effects. But the argument that prediction impacts basic perceptual processes *because* the observed effects occur on a component that is only sensitive to perceptual features, would be stronger with accompanying evidence for this sensitivity in the specific experiment. At the same time, potentially confounding effects of eye-movements must be ruled out, given the relationship between M1 activity and saccadic eye-movements (McDowell et al., 2008), as well the relationship between saccadic activity and orthographic regularity (e.g., Radach et al., 2004).

Lack of sensitivity to visual features in these experiments, and lack of downstream consequences on linguistically-relevant visual processing (e.g., on the M170 component, which is sensitive to lexical features; Solomyak and Marantz, 2009), raises the question of which visual features were predicted. Which basic visual features of phonologically typical nouns –as a category - differ from those of atypical nouns? And is the most relevant measure not perhaps orthographical typicality (bigram frequency; e.g., Hauk et al. (2006)) rather than phonological typicality? The sensory hypothesis remains rather underspecified on these issues.

5.4.3. *What happens after the M100?*

Theories of prediction assume that prediction facilitates comprehension when the prediction is correct (Altmann and Mirkovic, 2009; Pickering and Garrod, 2013). Likewise, Dikker and colleagues state that the effect of word category predictions on early visual processing is what allows for rapid violation detection. But the Dikker et al. studies only show reliable effects on the M100, there is no evidence for downstream consequences. If the brain is able to use visual cues to word category, does this lead to reduced downstream processing costs? Does it lead to faster detection responses?

Dikker et al. (2009) argue that visual form prediction could impact the M100 and the subsequent M170 component, with the latter component more likely to show any effect given its established role in lexical processing (Carreiras et al., 2014; Dehaene et al., 2005; Solomyak and Marantz, 2009). In Experiment 1, they find effects on the M100 but not the M170. The authors state that the null-effect on the M170 component is hard to interpret given that modelling the bilateral M170 sources is a noise-prone procedure. As visible in their Fig. 6, unexpected participles elicit at clear effect at 170 ms on the peak that follows the M100, while unexpected prepositions do not, but this activity was not analyzed. In Experiment 2, however, no analysis of M170 activity was reported, and the M170 was not discussed any further. In Dikker et al. (2010), a post-M100 effect was visible on the M170 component for typical nouns but not for the other nouns. It is not clear whether this is a reliable or meaningful effect, and post-M100 activity was not analyzed or reported.

It is unclear whether a word category violation mismatch does anything beyond the M100 response. In both studies, the statistical analysis focuses on M100 activity, but the preprocessing involved data segments from 300 ms before to 900 ms after target word onset (would allowed for an analysis of P600 activity associated with syntactic violations). If visual form predictions facilitate and expedite language processing, as Dikker and colleagues claim, does this yield an equivalent pattern downstream from the M100?

Interestingly, the N400-like responses in Dikker and Pylkkänen (2011) were smaller for matching words than for mismatching words, only in the strong prediction condition. These results thus suggested facilitation of predicted words. However, there was no match-effect in weak prediction items at all, which is surprising because the pictures were intended to prime the semantic category, and N400 activity is typically highly sensitive to semantic category priming (e.g., Kutas and Federmeier, 2000, 2011).

5.4.4. *What is the impact of the 1 Hz high-pass filter?*

In all three studies, Dikker and colleagues use a 1–40 Hz bandpass filter, which removes low and high frequency activity. Such filters are typically used to increase the signal to noise ratio, which facilitates source localization. While high-pass filter settings like 1 Hz are common in some MEG laboratories, many researchers warn for potentially strong distortions of the signal (e.g., Acunzo et al., 2012; Rousselet, 2012; Tanner et al., 2015). This is particularly important for early effects (e.g., Acunzo et al., 2012), because a 1 Hz high-pass filter can distort the signal backwards in time[11], causing later effects to appear as earlier components (Fig. 9). In the Dikker et al. studies, M100 activity is arguably followed by other effects of anomaly detection, but how much

[11] This can happen with a backward filter and a zero-phase shift filter that is applied forward and backwards in time.

the filter settings matter for prediction-related visual M100 effects remains to be seen.

5.4.5. How strong is the evidence for the sensory hypothesis?

Dikker and colleagues took their results as strong evidence that the brain predictions visual features of an upcoming word or visual features associated with a word category. But the obtained evidence may not be particularly strong. In Dikker et al. (2009), the crucial findings in both experiments in support of the sensory hypothesis are associated with p-values in the 0.01-0.05 range. In comparison, the effect of prediction strength in Experiment 2, which was inconsistent with the sensory hypothesis, was associated with $p = 0.003$. In Dikker et al. (2010), the crucial effects are also associated with p-values in the 0.01-0.05 range. In Dikker and Pylkkänen (2011), there was relatively strong evidence for a mismatch effect in the strong predictive condition ($p = 0.004$), but, as mentioned earlier, no results reported for the interaction between prediction strength and mismatch. Perhaps these patterns signal that the observed effects are relatively small M100 fluctuations, which are hard to detect in MEG data especially with small sample sizes ($N < 15$). Dedicated and pre-registered replication attempts are needed to gather stronger support for the sensory hypothesis.

Pre-registration is particularly important, but also daunting, because of the complexity of the analyses by Dikker and colleagues and the many steps during analysis that were taken to ensure data quality control. In each study, participants were removed from the analysis if they did not show clear visual M100 responses based on a visual inspection. In the dipole modelling procedure, models were constructed for each condition of each participants, and then selected based on their visual fit to the sensor-level data and minimum norm estimates of M100 activity. There are differences in how the data entering the statistical results were obtained. In the studies with dipole modelling (Dikker et al., 2009, 2010), for the categorical analysis a 15 ms time window of analysis was selected around the M100 peak detected in the individual conditions, but while Dikker et al. (2009) performed this peak-detection on the grand-average per condition across subjects, Dikker et al. (2010) used the average per condition per subject. In Dikker and Pylkkänen (2011), no dipole modeling was performed, the sensor-level analysis was now performed on a smaller window around the peak (10 ms) than before, based on a selection of sensors. It is unclear whether these differences in procedure were decided on after visual inspection of the data.

The question thus arises whether or not the results from the Dikker et al. studies hinge on the specific choices that were made during the analysis, and whether the results will generalize to novel observations.

5.5. Summary

Dikker and colleagues posited the sensory hypothesis of linguistic prediction, in which syntactic and semantic predictions during reading have basic (non-linguistic) visual correlates, and the brain generates estimates of the likely physical appearance of an upcoming word. The MEG results from their three pioneering studies were taken to support the sensory hypothesis.

I have highlighted some aspects of their results that warrant further investigation, in particular the role of prediction strength, the sensitivity of the M100 to basic visual features (string size), repetition of words and word stems, the downstream consequences of visual prediction mismatch, and the strength and replicability of the obtained evidence. It is also important to establish that early occipital effects in these paradigms are not ‘merely’ driven by saccadic activity or by visual attention (e.g., Leopold and Logothetis, 1998; Coffman et al., 2013; McDowell et al., 2008).

Along with these specific issues, the Dikker et al. studies also raise general questions. For example, it may seem relatively clear that one could generate a visual prediction for one highly expected word, but how is the prediction of an entire word category translated into a perceptual template? And how does this rhyme with the M100 sensitivity to rather crude visual features such as luminance and string size? Dikker et al. (2010) argue that M100 modulation by phonological typicality demonstrates that people generated a visual feature prediction associated with word category. This implicitly assumes that phonological typicality of a word category was correlated with a visually-relevant variable that was not measured, namely orthographic typicality. The patterns observed by Dikker et al. may thus be stronger in language with a regular correspondence between sound and spelling (e.g., Italian) than in languages where this correspondence is irregular, suggesting a potential avenue for follow-up research. However, while orthographic typicality has been associated with early occipital activity around 100 ms (Hauk et al., 2006), it is unclear whether adverbs and nouns differ in their orthographical typicality, and whether they do so in a sufficiently consistent manner to be translated into a basic visual prediction.

Another general question is whether the observed effects generalize to other, more naturalistic experimental circumstances (see also Willems, 2015). The results of Dikker and Pylkkänen (2011) suggest prediction of a specific word form when participants see a picture of the object, but participants may have visually imagined the word they thought would appear next. This does not necessarily mean people predict the specific visual form of words during regular reading. The results of Dikker et al. (2009, 2010) suggest that visual prediction is more common, as it would happen as soon as a reasonably strong prediction can be formed about the upcoming word category. Therefore, replicating these results in different experimental circumstances (less repetition, more varied sentences, different task demands) would be an important step towards an understanding of the predictive nature of the language system.

6. The P130

The P130 ERP, or occipital P1, is an early positive ERP component that can observed at occipital electrodes as a response to visual stimuli (e.g., Johannes et al., 1995). Of note, this component is often described as the ERP equivalent of the magnetic M100 (e.g., Mangun et al., 1998; Pitcher et al., 2011). The P130 is sensitive to exogenous and endogenous factors such as luminance and spatial attention (e.g., Awh and Jonides, 2001). I will first discuss a study by Kim and Lai (2012), who reported P130 evidence for word form prediction using a pseudoword/nonword design, and then a study by Kim and Gilley (2013), who reported P130 evidence for the prediction of syntactically anomalous word forms.

6.1. Kim and lai (2012)

Kim and Lai (2012) tested a hypothesis based on the interactive activation view of visual word recognition (McClelland and Rumelhart, 1981), which assumes a dynamic interaction between feedforward and feedback processes. They hypothesized that prediction pre-activates lexical and word form features, and those features are further activated by a pseudoword that is visually similar to the predicted word, e.g., ‘cake’ by ‘ceke’ (7). This activation boost rapidly sensitizes the word recognition system to detect lower-level visual discrepancy between predicted word form input and actual input, possibly resulting in early effects in visual cortex (Occipital P1 or N170). This feedforward-feedback dynamic does not take place for pseudowords that are not similar to the anticipated word (e.g., ‘tont’) or for nonwords (e.g., ‘srdt’). In their experiment, predictable control words had a high average cloze probability of 90%, similar pseudowords were created by replacing a single non-initial letter of each control word, nonsimilar pseudowords were created by swapping the similar pseudowords between two items, and the nonwords were consonant strings.

(7). Example materials from Kim and Lai (2012)

Control: *She measured flour so she could bake a* ***cake***…

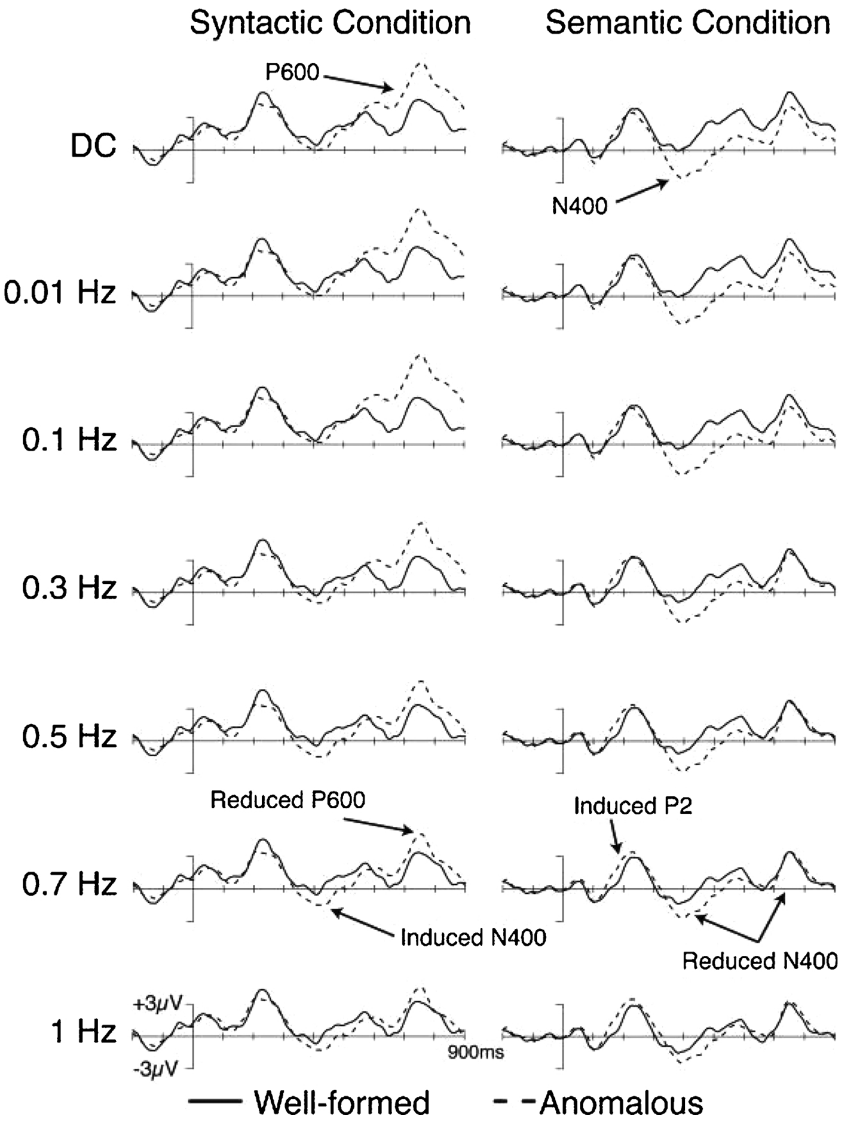

Fig. 9. Demonstration of signal distortions by high-pass filtering, adapted from Tanner et al. (2015). The top graphs show P600 effects of syntactic anomaly and N400 effects of semantic anomaly when there is no filter (DC) and increasing high-pass filter settings up to 1 Hz (bidirectional Butterworth filter). At higher filter settings, a reduced P600 effect of syntactic anomaly is accompanied by an artefactual N400-like effect, and a reduced N400 effect of semantic anomaly is accompanied by an artefactual P2-like effect.

Similar pseudoword: *She measured flour so she could bake a* ***ceke****…*

Nonsimilar pseudoword: *She measured flour so she could bake a* ***tont****…*

Nonword: *She measured flour so she could bake a* **srdt***…*

Participants read 45 sentences per condition and answered occasional yes/no comprehension questions. No filler sentences were included, therefore 75% of all sentences contained a pseudoword or nonword. Analysis focused on the P130 and the N170, with time windows chosen through visual inspection of the component peaks (125–145 and 170–205 ms respectively), and on the N400 the P600, using typical time windows based on previous literature (300–500 and 500–700 ms respectively). Selection of data channels for analysis was based on visual inspection of the scalp distributions of the effects and previous literature.

They found that P130 activity reliably differed between the conditions (Fig. 10): consistent with their hypothesis, similar pseudowords elicited enhanced P130 ERPs compared to all other conditions (all p-values in the 0.01-0.05 range), whereas the other conditions did not differ from each other. The occipital-temporal N170 also differed between conditions: nonsimilar pseudowords and nonwords elicited enhanced N170 compared to control words, whereas similar pseudowords did not. Kim and Lai interpreted the N170 effects as visual form processing difficulty due to a lack of activated lexical representations.

6.2. Discussion

Kim and Lai (2012) took an innovative approach by combining a traditional sentence reading paradigm with a typical visual word recognition paradigm using pseudowords and nonwords. Their P130 results were consistent with the rapid detection of a similarity-based conflict between predicted and actual visual input. These effects are not easily explained by physical differences, because the similar and nonsimilar pseudowords were counterbalanced between conditions. A very interesting implication of their findings is that early visual processing costs are greater for words that are similar to the predicted word than for dissimilar words, i.e. costs do not increase but decrease with prediction error. This pattern seems incompatible with the conclusion reached by Dikker and colleagues based on their M100 findings, and raises a challenge for future research because the M100 and P130 are thought to arise from the same neural generator (e.g., Mangun et al., 1998). As already discussed by Kim and Lai, however, previous ERP studies with related designs did not generate such early effects, and this raises an important question.

6.2.1. Do these results generalize to real words?

A high percentage (75%) of the sentences that each participant read contained a nonword or pseudoword. Because all sentences also led to very strong expectations for a specific word (i.e. cloze probability in the 85–100 range), it could be the case that participant learned to predict the occurrence of a nonword, or paid more attention to orthography than they would during regular reading. The P130 component may be boosted by attention to orthography like in a letter detection task (e.g., Proverbio and Adorni, 2009), and by increases in visual selective attention (e.g., Hillyard and Anllo-Vento, 1998; Johannes et al., 1995). This could mean that the P1 pattern may be harder to find with sentences that contain only real words, if participants are more likely to process for word meaning than for visual form. If the Kim and Lai results are limited to nonword paradigms with very high-constraint sentences, this raises the question whether results reflect regular reading processes, or whether they, alternatively, reflect task-based visual attention effects or perhaps increased intraword-saccadic activity for similar pseudowords (e.g., Meyberg et al., 2015).

6.2.2. Are the P130 results replicable?

The results of Kim and Lai (2012) have not been observed in similar experiments with pseudowords (e.g., Bulkes et al., 2018; Laszlo and Federmeier, 2009; Vissers et al., 2006), and have not been observed elsewhere. Thus far unpublished studies have not supported the P130-based conclusions of Kim and Lai. A direct replication from the Kim laboratory (Wittenberg, 2012) failed to yield the same pattern of effects. Two recent close-replication studies with the same supported-pseudoword condition did not find any P130 effects (Bulkes et al., 2018). Further direct replication is needed to establish whether the Kim and Lai (2012) P130 effect is an isolated finding or a generalizable phenomenon.

6.3. Summary

Kim and Lai (2012) reported that pseudowords elicited a P130 (occipital P1) effect only when they were visually similar to words that were highly expected given the sentence context. They hypothesized that lexical and word form features of expected control words are pre-activated during reading, and that their activation is further boosted by similar pseudowords, which in turn sensitizes the visual feature processing system to detect small discrepancies between predicted and actual input. This is a very interesting hypothesis, and raises an important challenge to prediction-based theories of language comprehension (Dell and Chang, 2014; Pickering and Clark, 2014). Such theories assume that processing costs increase with prediction error (the difference between current input and predicted input), whereas the Kim and Lai results suggest the opposite. However, the pseudoword P130 results have yet to be replicated and future research should test their implications for regular reading.

6.4. Kim and Gilley (2013)

Kim and Gilley (2013) followed-up on the word category violation studies by Lau et al. (2006) and Dikker et al. (2009, 2010). Like those previous studies, they argued that such early effects could signal the involvement of predictive processing and that word category violations can have an effect on early perceptual processing. They used an experimental design that manipulated the variability of the word category violation condition (8). One group of participants read sentences in a correct control condition where the critical word was always ‘the’, and in a low-variability anomalous condition with a word category violation that was always the word ‘for’. Another group of participants read the same control sentences and sentences in a high-variability anomalous condition with a word category violation that could be one of 7 different function words. Kim and Gilley hypothesized that participants in the low-variability group would engage more in predictive processing than participants in the high-variability group because of the distributional regularities in the experiment. Word form prediction could therefore speed up word recognition and lead to an earlier sensitivity to the anomaly.

(8). Example materials from Kim and Gilley (2013)

Correct control: *The thief was caught by* ***the*** *police …*

Low-variability anomaly: *The thief was caught by* ***for*** *police …*

High-variability anomaly: *The thief was caught by* ***at/for/of/on/from/over/with*** *police …*

Participants read 35 target sentences per condition and answered occasional yes/no comprehension questions. The sentences were mixed with 105 filler sentences that could be correct or containing agreement violations (40% of all sentences in the experiment were anomalous). EEG recording was performed with the same system as in Kim and Lai (2012), but the analysis procedure for P130 and N170 activity differed from Kim and Lai (2012) in a few ways. A global average reference was used instead of the mastoid average. P1 activity was quantified as the peak activity in a 125–145 ms time window, rather than average activity, and the selection of occipital channels differed from that of Kim and Lai (2012). The time windows of analysis for the N170, N400 and P600 differed from Kim & Lai (here, 170–270, 350–450, and 500–800, respectively). Supporting their hypothesis, Kim and Gilley reported a left-lateralized, occipital P130 effect for word category violations in the low-variability condition but not in the high-variability condition (Fig. 11). Subsequent N170 and P600 effects did not differ between these conditions. Source-localization of the P130 and N170 effects revealed occipital sources but are not discussed in detail here as these findings are not pertinent to the conclusions.

6.5. Discussion

Kim and Gilly (2013) used an innovative paradigm that manipulated the variability of a syntactic anomaly. They observed earlier anomaly effects (P130) for highly predictable anomalies than for less predictable ones. They concluded that early sensitivity to word category anomalies results from prediction, and that long-term syntactic knowledge can be rapidly adapted to recent linguistic experience, through the learning of distributional patterns specific to the experiment. The former conclusion is consistent with Lau et al. (2006) and Dikker et al. (2009). The latter conclusion seems consistent with recent demonstrations of adaptation to the within-experiment probability of a certain linguistic phenomenon (e.g., Fine et al., 2013; but see Stack et al., 2018). The Kim and Gilley results raise several important questions.

6.5.1. Is this purely a P130 effect?

Kim and Gilley tested prediction effects on the same components as Kim and Lai (2012), but with potentially relevant differences in data processing and analysis. Using the global average reference procedure instead of the mastoid reference procedure in Kim and Lai (2012) could be relevant, as I discussed earlier in this review. Is the observed P130 effect qualitatively similar to that observed by Kim and Lai? Do the Kim and Gilley results contain a polar average reference effect (for discussion, see the section on the Early Cloze Positivity and Fig. 3). More generally, if the changes in pre-processing and the analysis procedure (e.g., channel selection, peak measurement) were based on visual inspection of the data, the newly observed effects are likely inflated and less likely to replicate (Luck and Gaspelin, 2017).

Another potential issue with the P130 interpretation is the existence of a differential effect of anomaly in the low-variability condition starting as early as word onset (Fig. 10). Such an effect seems implausible early and could reflect residual noise (random activity

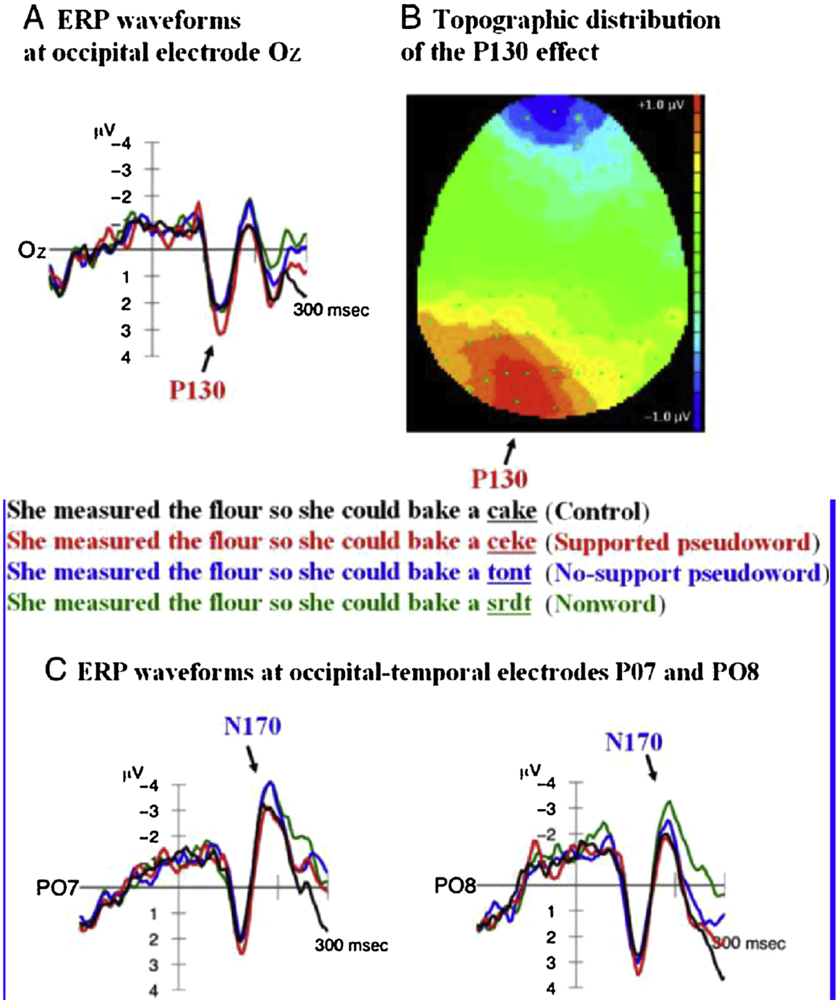

Fig. 10. Results from Kim and Lai (2012). Left graphs (A,B) show the P130 (Occipital P1) effect for similar/supported pseudowords, right graphs (C) show the N170 effect for nonsimilar/nonsupported pseudowords and nonwords.

fluctuations that are associated with a low sample dataset). Although the differential effect seems to disappear before the P130 component, it is still unclear whether the P130 effect is purely a modulation of the P130 component or due to more positive amplitudes overall in the first few hundreds of milliseconds. Kim and Gilley reported lack of a statistically significant effect of anomaly in the 60–80 ms time window, although it is unclear why this specific window was chosen, as it did not capture the early nature of the effect and does not capture a potential influence of the early difference on effects in the P130 window. Another approach would subtract activity in the early post-onset window from the data segment ('post-onset baseline correction', e.g., Hagoort and Brown, 2000; Ito et al., 2017; Kaan, 2002; Osterhout et al., 1994; Tanner et al., 2013)

6.5.2. Do repeated anomalies boost prediction?

Kim and Gilley claim that participants in the low-variability condition were more likely to predict the control word 'the' than participants in the high-variability condition, even though control words were equally likely in the two conditions. If low-variability boosted prediction of control words, one would expect a reduced P130 component. One would also expect variability to modulate downstream anomaly effects on the N400 or P600, in particular P600 effects are sensitive to probability and saliency manipulation (e.g., Coulson et al., 1998). However, variability did not impact the P130 component elicited by control words, only that elicited by anomalies, and variability did not have any impact on N400 or P600 anomaly effects. These patterns therefore do not support the conclusion that repeating anomalies boosts prediction of correct words. An alternative interpretation is that low-variability did not cause more prediction overall but increased participants' visual selective attention to the anomalous word form 'for' (e.g., Hillyard and Anllo-Vento, 1998; Proverbio and Adorni, 2009).

6.6. Summary

Kim and Gilley (2013) reported an enhanced P130 effect of syntactic anomaly when that anomaly occurred on the same lexical item throughout the experiment, compared to anomalies on different lexical items. They concluded that low-variability of anomalies boosted prediction of the correct control word (which was always 'the' regardless of the anomaly-variability), by creating higher 'affordances' for prediction when fewer words could appear in the target position. In addition, they presented their results as further evidence that word form predictions play out at a low-level, sensory level of representation.

The Kim and Gilley study took a refreshing, innovative approach and their P130 effects may indeed be the earliest ERP response to syntactic anomaly in the literature. However, as I discussed in this review, some aspects of the data and analysis complicate the identification and interpretation of the P130 effects and their comparison to the P130 effects reported by Kim and Lai (2012). In addition, the Kim and Gilley results do not yield clear evidence that low-variability of anomalies indeed boosted prediction of correct control words, as variability seemed to modulate the P130 to anomalies, but not to control words.

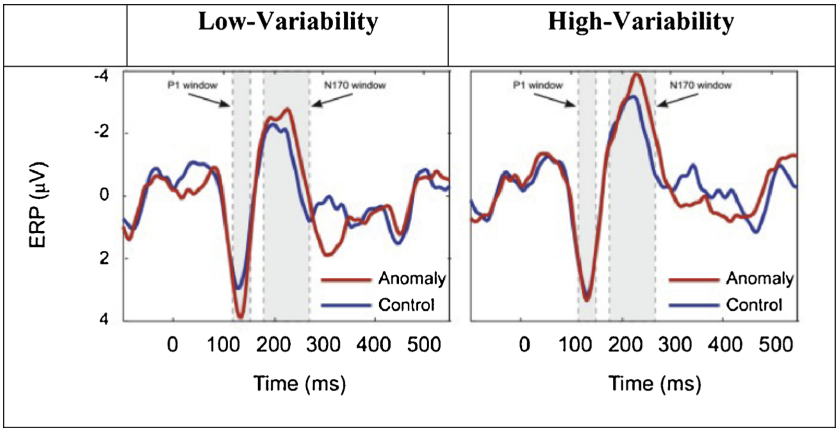

Fig. 11. Early ERP effects of syntactic anomaly in low-variability and high-variability conditions at channels PO7 (Kim and Gilley, 2013).

The variability manipulation in Kim and Gilley is certainly interesting and novel, but their interpretation in terms of 'predictive affordances' raises further questions. Why is it assumed that repeated anomalies boost prediction of correct words but not of anomalous word forms? Would any regularity in the experiment, linguistic or non-linguistic, boost prediction? More generally, it remains to be seen whether Kim and Gilley demonstrates the effect of task-demands imposed by a rather artificial and repetitive anomaly design, or of a more general phenomenon of language comprehension.

7. The N1/P2

The N1 and P2 components together form the N1/P2 complex, a biphasic ERP response to visual stimuli such as written words (e.g., Hillyard et al., 1998). The N1 is primarily associated with the correct allocation of attention, and the selection of perceptual features (e.g., color) for further processing (e.g., Hillyard et al., 1998). The P2 is sensitive to visual stimulus repetition and therefore often thought to index the matching of immediately-available visual information to perceptual representations stored in memory ('perceptual matching'; Luck and Hillyard, 1994).

7.1. Penolazzi et al. (2007)

A study by Penolazzi et al. (2007) investigated early ERP effects of predictability (semantic context) during sentence reading. They did not explicitly test whether prediction impacts early ERP components, but they used a low/high cloze probability manipulation that is typical of studies on prediction. They reasoned that previous studies missed early effects of predictability by not taking into account pre-lexical and lexical variables (word length and frequency), and predicted that early ERP effects of predictability would interact with such variables. Their participants read sentences with target words that could vary in three dimensions (9): short/long (on average, 4 or 6.2 letter respectively), low/high frequency, low/high cloze probability (0 or higher than 50). Probability was manipulated via changing one word in a context. The target words were always nouns in the same, non-final sentence-position.

(9). Example materials from Penolazzi et al. (2007). Context words (underlined) for high/low cloze targets words (bold) are separated by a forward slash.

Long, high-frequent: *The extremists planned a* <u>*terrorist/peaceful*</u> ***attack*** *on the government.*

Short, high-frequent: *He drinks* <u>*cocktails/tea*</u> *in the* ***bar*** *down the road.*

Long, low-frequent: *The* <u>*boiler/contraption*</u> *was a water* ***heater*** *used in houses.*

Short, low-frequent: *Children listened to the* <u>*fairy/angry*</u> ***tale*** *in the classroom.*

Participants read a total of 280 sentences (35 from each of the 8 conditions) along with 70 filler sentences. The analysis was different from what is usually done in ERP studies: the authors computed root mean square (RMS) values from the grand-average ERPs across participants and conditions and electrodes, from which they selected peaks in the ERP waveform for subsequent analysis (110–130, 170–190 and 280–320), which roughly correspond to N1, P2[12] and early N4. Analysis of variance were performed with the three factors in the design and another two distributional factors associated with the EEG channels (left/right, anterior/posterior). For each peak, the authors found that cloze probability interacted with one or more of the other factors (Fig. 12). Main effects of probability were observed only in later N400 windows.

7.2. Discussion

The study by Penolazzi et al. is a well-cited report of early interaction between lexical processes and semantic integration processes. The authors argued against a temporal division between the impact of lexical factors and contextual integration as assumed in some theories of language comprehension (e.g., Friederici, 2002), and emphasized the importance to investigate semantic processes taking place before the N400. The results also demonstrate the importance of taking into account word length and frequency when testing early semantic effects. Although Penolazzi and colleagues did not interpret their results in terms of prediction, the early latency of the effects could be indicative of predictive processing, and therefore raise some new questions.

7.2.1. What do these effects mean for prediction?

The primary focus of Penolazzi et al. was to find out whether cloze probability interacted with lower-level factors. This is indeed what their results suggested but the specific pattern that they observed remains unclear, and the authors did not offer a hypothesis or conclusion about whether, for example, cloze probability has an earlier effect for longer words than shorter ones. The results suggest that when testing for early ERP effects of prediction, variance in word length in the stimulus materials should be minimized.

7.2.2. What is the role of pre-target words?

One potential limitation to the conclusions of Penolazzi et al. is the influence of the pre-target words. They carefully controlled length, frequency and cloze probability of the target words, but the cloze probability manipulation was created by changing words in the context.

[12] Penolazzi et al. referred to the second peak as P1 as it was the first positive peak, but the time course of that peak is consistent with what could be an early P2, which follows the N1.

These changes sometimes or often involved the words before the critical word, dependent on condition[13]. This is relevant because the high- and low-cloze incurring words differed strongly in meaning and their relation to the sentence context. For example, in the example in (9) 'peacef..' might be harder to integrate than 'terrori..' into a sentence context about extremists. Moreover, even if the context manipulation was not on the pre-target word, the pre-target word might be easier or harder to integrate depending on the cloze manipulation, e.g. in the example in (9) 'wat..' might be easier to integrate with 'The boiler was a..' than with 'The contraption was a..'. Differential effects of the pre-target words may have spilled-over into the beginning of the target-word window, given that N400 effects of semantic integration can last until 1000 ms after word onset (e.g., Van Berkum et al., 1999). Whether and to what extent pre-target differences influenced the Penolazzi et al. findings is unknown, but some caution is warranted in interpreting their results. The Penolazzi et al. results have yet to be replicated, a dedicated follow-up is needed that overcomes potential issues with pre-target activity.

7.3. Summary

Penolazzi et al. reported an ERP study that carefully controlled sentence position, cloze probability, word length and frequency of the critical nouns. They found very early ERP effects of cloze probability, roughly corresponding to the N1 and P2 component of the word-elicited ERP waveform. These effects depended on lexical/pre-lexical variables, however, which led to the conclusion that early contextual integration processes are "modulated by physical, phonological or orthographic processes triggered by a target word". Although the authors did not present their study as being about prediction, their cloze manipulation was similar to that used in prediction studies and could be construed as a test of the word recognition hypothesis. The results of this study have not been replicated, to the best of my knowledge. A follow-up study would need to tease apart effects of pre-target and target words and offer more specific insights into the early interaction between lexical and contextual processes.

7.4. Lee et al. (2012)

Lee et al. (2012) also tested for interactions between cloze probability and frequency on early ERP components, and framed their study explicitly in terms of prediction. They argued that context information can be used in a predictive manner, so that perceptual features of an upcoming word are pre-activated and facilitate visual word recognition during the stage of processing where lexical variables also matter. Their study involved the Mandarin Chinese writing system, in which 76% of the words consist of 2 or 3 characters, so there is little variation in word length. In a 2 by 2 design, they tested for effects of cloze probability (high/low) and frequency (high/low) of two-character words embedded in a sentence context, with average cloze values of 75–80% for high predictable sentences and of 0–5% for low predictable sentences. Participants saw each word in a high and low cloze context in two separate experimental sessions 2 weeks apart. ERP analysis was performed after visual inspection of relevant peaks in the data (N1: 120–150, P2: 200–250, N4: 300–500), see Fig. 13.

The authors report a significant interaction, with low predictable, high frequent words eliciting a larger N1 than high predictable, high frequent words, but no such difference for low frequent words. There was no main effect of predictability on the N1. For the P2 peak, they found an interaction between predictability, frequency and repetition (referring to the two sessions each participant did), and therefore only analyzed the first session for each participant. They reported enhanced central-posterior P2 peaks for high predictable sentence, but not consistently across electrodes.

7.5. Discussion

Lee et al. reported an interaction between Chinese word frequency and predictability on an anterior N1 component, and a subsequent effect of predictability on the P2 component. They concluded from these results that contextual information facilitates visual-feature and orthographic processing in the early stage of word recognition, and semantic integration in the later N400 stage. This conclusion is similar to that of Penolazzi et al. (2009), although the specific conclusions differed, as Penolazzi et al. did not observe an early interaction between cloze probability and frequency but between cloze probability and word length. Here, too, it is not entirely clear what the results mean beyond demonstrating an early interaction between lower- and higher-level factors. Lee et al. suggest that the N1 effects could reflect a top-down prediction effect on early visual feature processing (e.g., Dambacher et al., 2009; Dikker et al., 2009; Kim and Lai, 2012), but the effect had a frontal distribution, not an occipital one. Lee et al. also suggested an alternative interpretation in terms of directed attention towards visual features to enhance perception of the expected stimuli. More generally, both these explanations are tentatively based on previous interpretations of N1 activity, but do not specifically address why the effects occur only for high-frequent words, and why low predictable words elicit a larger N1 than high predictable words.

7.5.1. What is the role of pre-target words?

Lee et al. carefully controlled the pre-targets for frequency and word class across high and low cloze probability. But high and low cloze sentences did differ up to the target word, and therefore it is possible that pre-target word integration effects spill-over into the target word time window like in Penolazzi et al.

7.5.2. Is the P2 effect an early N400 effect?

As visible in Fig. 13 (right graphs), the relevant P2 effect is directly followed by a clear N400 effect. In such a case, it is hard to distinguish the two effects, especially because the P2 effect had a scalp distribution very similar to that of the N400 effect. What is labelled as a P2 effect may in fact well be the early onset of an N400 effect, which can occur at 200 ms after word onset (e.g., Molinaro et al., 2010; Nieuwland et al., 2018a). This would also offer a more straightforward interpretation to the direction of the effects (low predictable words elicit a larger N400, therefore a smaller P2). If the effect is purely a P2 modulation, the question arises why low predictable words elicit a smaller P2 than high predictable words, while the effect on the N1 went in the opposite direction.

7.6. Summary

Lee et al. reported an interesting ERP study that examined early ERP effects of cloze probability and word frequency, while carefully controlling for word length. They found that low cloze words elicited an enhanced anterior N1 effect compared to high cloze words, but only when the words were high-frequent. They also found an enhanced P2 component to high cloze words compared to low cloze words. These results raise interesting questions about the potential interaction between predictive processing and lexical variables. The results of this study have not been replicated, to the best of my knowledge, and the impact of pre-target words on the observed early effects remains unclear. A follow-up study would need to offer further insights why early prediction effects may differ for high and low frequent words.

[13] For the long and high-frequent, short and high-frequent, long and low-frequent, and short and low-frequent conditions, this was the case in 30, 12, 19, and 21 out of 70 items, respectively.

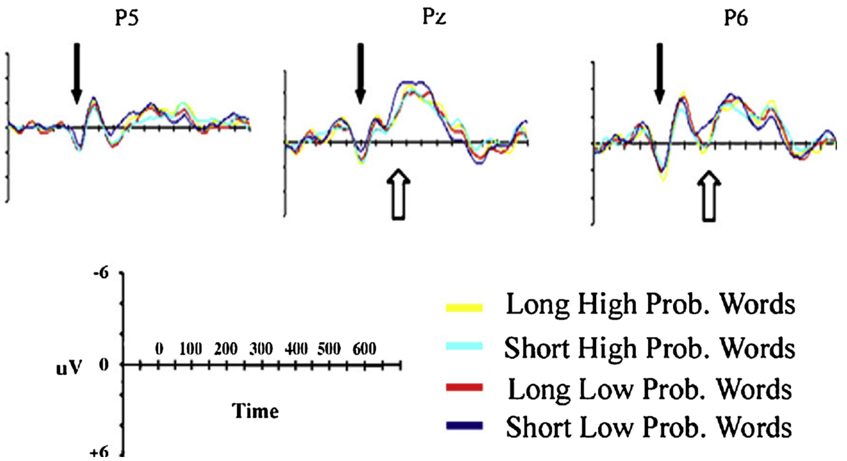

Fig. 12. Grand-average ERP waveforms for long/short words with high/low cloze probability, adapted from Penolazzi et al. (2009). The black arrow indicate the first peak (P2) where the first word length by cloze interaction was observed. The white arrow indicates the first peak (early N4) where the first main effects of length or cloze were observed.

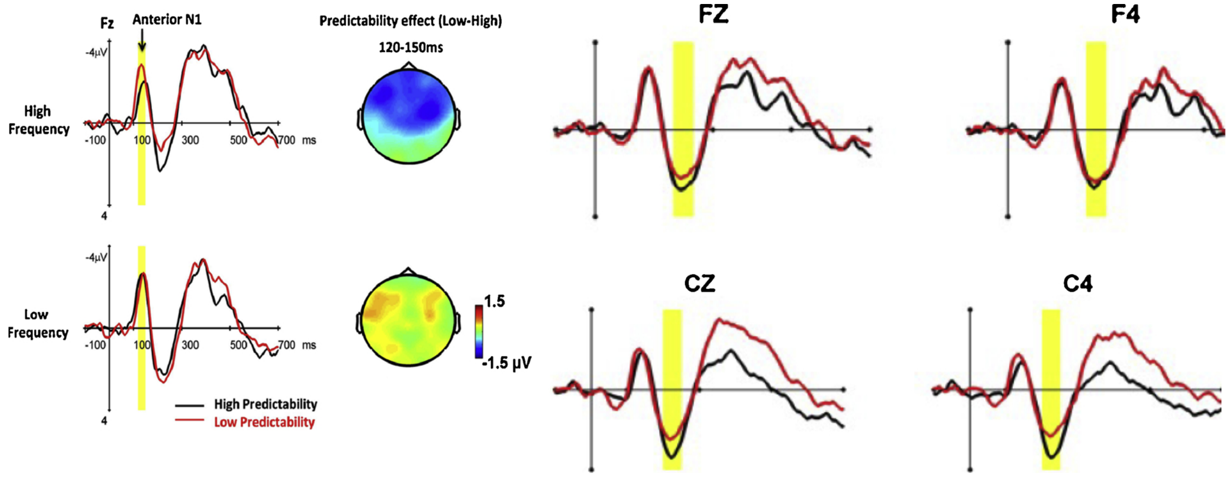

Fig. 13. Effect of predictability on high and low frequent words on the anterior N1 peak (left graphs), effect of predictability on the P2 peak (right graphs), figure adapted from Lee et al.

8. The N200 or phonological mismatch negativity (PMN)

In the prediction literature reviewed here, the N200 and PMN refer to a related ERP phenomenon, a frontocentral, negative wave peaking between 200 and 350 ms after stimulus onset that is associated with phonological or ‘sound form’ prediction. Studies that report this peak in both expected and unexpected conditions refer to a modulation of the N200 component (i.e. an N200 effect; e.g., Hagoort and Brown, 2000), whereas studies that report this peak in unexpected conditions only refer to a PMN elicited by sounds that disconfirm a prediction. The N200 and PMN labels both originate from a rich literature on the detection of novelty or mismatch in ‘oddball’ designs (for reviews, see Folstein and Van Petten, 2008; Naatanen et al., 2007). For simplicity, I will refer to N200 in teh general sections of this review when discussing effects labelled in specific studies as either PMN (Connolly and Phillips, 1994), N250 (Hagoort and Brown, 2000) or N200 (all other studies).

Several ERP studies have tested the word recognition hypothesis in the auditory modality via the N200. Listeners may be able to use the initial phoneme of an encountered word to rapidly detect a mismatch with an expected word (Connolly and Phillips, 1994) or to select a word from several lexical candidates (van den Brink et al., 2001, 2004). The reviewed studies argue that this early process is initialized before the meaning of the word is integrated with the sentence context, which they assumed is reflected in N400 activity. Here, I review the studies reported support from the N200 for this hypothesis, and then discuss the implications of these studies for the N200 together, along with studies that have failed to find N200 effects.

8.1. Connolly and Phillips (1994)

Connolly and Phillips (1994) investigated whether the N200/PMN reflects an early phonological processing function that is sensitive to context-based predictions. The N200 had been observed in previous experiments (e.g., Connolly et al., 1992), but it was unclear whether this effect reflected a purely phonological process or a semantic process. Connolly and Phillips argued that if people use sentence context to predict both the phonological form and the meaning of a word, then the N200 and N400 are differentially sensitive to whether the presented word matches the expected word in terms of its initial phoneme and/or its meaning. Their experiment thus aimed to differentiate N200 and N400 activity using sentence-final words in high constraint sentences (10) that matched or mismatched the initial phoneme of an expected word, or that matched or mismatched its meaning. Words with a mismatching phoneme but matching meaning always had a lower cloze probability than the expected word.

(10) Example materials from Connolly and Phillips (1994). Expected, highest-cloze word in parenthesis.

Phoneme Match, Semantic Match: *At night, the old woman locked the door.* (door)

Phoneme Match, Semantic Mismatch: *Phil put some drops in his icicles.* (eyes)

Phoneme Mismatch, Semantic Match: *They left the dirty dishes in the kitchen.* (sink)

Phoneme Mismatch, Semantic Mismatch: *Joan fed her baby some warm nose.* (milk)

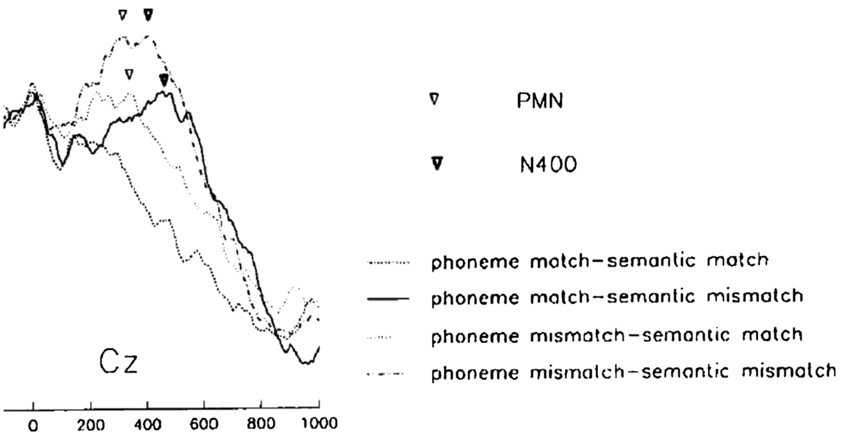

Fig. 14. Grand-average ERPs per condition at electrode Cz, adapted from Fig. 1 of Connolly and Phillips (1994). N200 and N400 peaks of phoneme mismatch and semantic mismatch are indicated by open and closed triangles, respectively.

Participants listened to 160 sentences (40 per condition), each consisting of 6–8 words. N200/PMN amplitude and latency per participant was scored by taking the most negative point between 150 and 350 ms after word onset, and the N400 was scored in the 350–600 ms time window. As shown in Fig. 14, the phoneme mismatch conditions showed enhanced negativity in the N200 window, and the semantic mismatch conditions showed enhanced negativity in the N400 time window. The N200 and N400 effects were evenly distributed and did not differ as a function of scalp location (5 different electrodes were analyzed).

In other words, an N200 peak was only detectable when the initial phoneme mismatched the predicted phoneme, and an N400 peak was only detectable when the word mismatched the meaning of the predicted word. Connolly and Phillips took this functional differentiation between the N200 and the N400 to support the hypothesis that word form predictions have an impact on the initial phonological analysis by which a word is recognized. They did not explain why the double mismatch condition elicited a larger amplitude N200 than the phoneme mismatch only condition.

8.2. Hagoort and Brown (2000)

Hagoort and Brown (2000) performed two spoken language experiments where they compared ERPs elicited by sentence-final or sentence-medial words that were either highly expected or semantically anomalous, which did not share initial sounds. They found enhanced negativity for anomalous words (Fig. 15), with a small peak around 250 ms after word onset and a larger subsequent N400 peak. They labelled the first peak as an N250 effect, and although the N250 and N400 effects did not differ in scalp distributions, they tentatively concluded these effects were functionally separate, and related to the N200/PMN effect reported by Connolly and Phillips (1994). However, whereas Connolly and Phillips argued that their N200/PMN effect reflected the phonological mismatch between encountered input with expected input, Hagoort and Brown interpreted their N250 effect within the Cohort model. A word-initial phoneme generates several lexical candidates through a strictly bottom-up, form-driven process, after which context has a top-down effect on selecting the candidate that is optimally compatible with both form and content constraints. This lexical selection process takes place for all words but it is more difficult if context does not support the lexical candidates that are available through form-based activation, leading to N250 effects. Once a word is recognized, the language system integrates its meaning with the sentence context, generating an N400 component.

8.3. Van den Brink et al. (2001)

Van den Brink et al. (2001) conducted an ERP study on spoken Dutch sentence comprehension, using a design that was similar to that of Conolly and Phillips (1994). High-constraint sentences ended with the expected word (Fully Congruent Condition), an anomalous word that started with the same phoneme as the expected word (Initially Congruent Condition), or an anomalous word that did not start with that phoneme (Fully Incongruent Condition). Analysis was performed in the 150–250 ms (N200) and the 300–500 ms (N400) time windows, which were based on visual inspection of the grand average ERPs (Fig. 16).

In the N200 window, they found a main effect of condition: Fully Incongruent words elicited enhanced negativity compared to the Fully Congruent words, and no such difference was observed between Initially Congruent and Fully Congruent words. The pairwise comparison between the Fully Incongruent condition and the Initially Congruent condition was missing. Only at fronto-central channels was a separate peak discernible for the three condition in the N200 time window, but analysis of the scalp distributions revealed that the N200 effect was widespread. The N400 effect showed a typical central-posterior distribution.

Like Connolly and Phillips (1994) and Hagoort and Brown (2000), van den Brink et al. thus conclude that the N200/N250 and N400 peaks reflect functionally distinct processes. But following Hagoort and Brown (2000), they concluded that the N200 reflected a lexical selection process taking place in all conditions, rather than the detection of a phonological mismatch (cf. Connolly and Phillips, 1994).

8.4. Van den brink and Hagoort (2004)

Van den Brink and Hagoort (2004) performed a follow-up ERP study that was very similar to van den Brink et al. (2001), with the same three conditions. However, the semantically incongruent words were also syntactically anomalous. For this purpose of this review, this difference between the experiments is not crucial, because the syntactic anomaly only became apparent at the end of the words (through an inflectional suffix that renders the word a verb instead of the expected noun). Analysis was performed in the same N200/N400 time windows as in van den Brink et al. (2001).

All three conditions again elicited a negative peak at frontal channels in the N200 time window (see Fig. 17). They observed a significant main effect of condition in this window: the Fully Incongruent condition elicited more negative ERPs than both the Initially Congruent and the Fully Congruent condition, which did not differ from each other. A

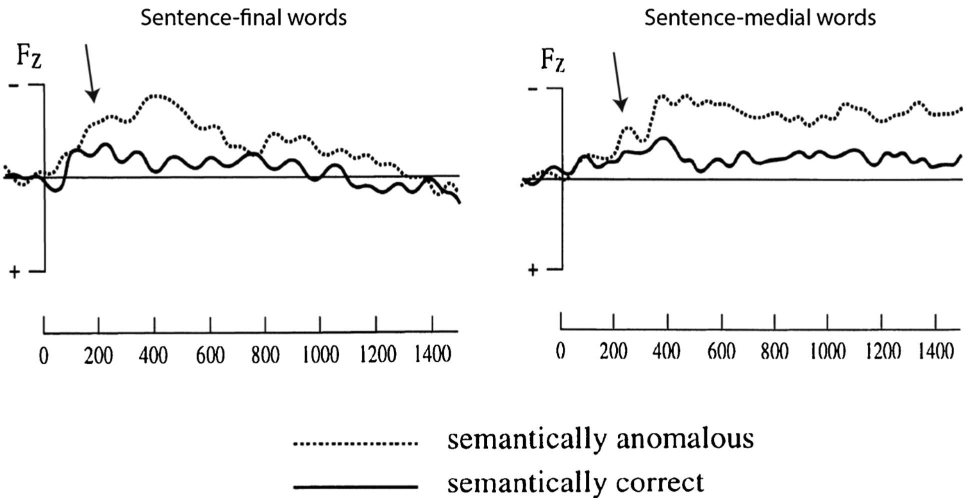

Fig. 15. Grand-average ERPs at electrode Fz elicited by semantically anomalous and correct words in sentence-final (left graphs) and sentence-medial (right graphs) position, adapted from Hagoort and Brown (2000). Arrows point to the negative deflection interpreted as N250 effect preceding the N400 peak.

distributional analysis revealed that the enhanced negativity for the Fully Incongruent condition was larger over posterior channels, as is also visible in Fig. 16. No follow-up analysis was reported on whether there was a statistically significant difference at frontal channels. Van den Brink and colleagues concluded that their results demonstrated a similar lexical selection process as they had argued before (Hagoort and Brown, 2000; Van den Brink et al., 2001).

8.5. Boudewyn et al. (2015)

Boudewyn et al. (2015) investigated lexical form prediction in an experiment where participants listened to mini-stories (11) that contained a critical word that was either predictable or unpredictable and either locally consistent or inconsistent (i.e., semantically compatible/incompatible with preceding feature words, like ‘cake’ with ‘sweet and tasty’ or ‘healthy and tasty’). A cloze test confirmed the high or low predictability of the target words after the context sentence (i.e., the feature words ‘sweet/healthy and tasty’ were not included in the cloze test, which means that the cloze values do not fully correspond to the sentences used in the experiment).

(11) Example materials from Boudewyn et al. (2015).

Context: *Frank was throwing a birthday party, and he had made the dessert from scratch. After everyone sang, he sliced up some*

Globally Predictable, Locally Consistent: *sweet and tasty* ***cake*** *that looked delicious.*

Globally Predictable, Locally Inconsistent: *healthy and tasty* ***cake*** *that looked delicious.*

Globally Unpredictable, Locally Consistent: *healthy and tasty* ***veggies*** *that looked delicious.*

Globally Unpredictable, Locally Inconsistent: *sweet and tasty* ***veggies*** *that looked delicious.*

Grand-average ERP waveforms for each condition are shown in Fig. 18. Of note, there was no negative peak visible in the selected N200 time window at any of the channels and in any of the conditions, and the N400 effect onset for the globally unpredictable conditions was visible around 200 ms. The authors defined N200 activity as activity taking place in the 200–300 ms time window, and N400 activity as activity in the 300–600 ms time window. In the N200 time window, there was a significant interaction between global predictability and local consistency: there was no significant effect of consistency for globally predictable words, whereas for globally unpredictable words, local consistency was associated with less negative ERPs than local inconsistency, but only at right-anterior channels. A scalp distribution comparison of the N200 activity and N400 activity was performed on the difference between the globally unpredictable and locally inconsistent condition and the average of the other three conditions, which showed a less posterior distribution for the N200 than for the N400.

Based on the observed differences in the 200–300 ms time window, and the different distribution of activity in this window compared to the N400 window, Boudewyn et al. concluded that they had observed an N200 effect that reflected the mismatch between the expected and received word form.

8.6. Discussion

The N200 studies reviewed here involved comprehension of naturally spoken sentences or mini-stories. The major strength of the auditory modality is that it allows researchers to test the neural effects of the smallest meaningful unit of spoken language, the phoneme, well before the word in which that phoneme appears is complete or uniquely identifiable. This allows for a much more fine-grained hypothesis test than written comprehension studies, where effects of whole words are tested in an unnatural reading procedure. The N200 studies reviewed here compared ERP activity time-locked to word-initial phonemes that mismatched or matched an expected word. Mismatch was associated with an enhanced negativity in the 100–300 ms time window after word onset (N200) compared to match, and also enhanced negativity in the later, N400 time window. The two main interpretations for these effects relate to the lexical selection stage in the Cohort model of word recognition (e.g., Marslen-Wilson and Tyler, 1980). Connolly and Phillips (1994) and Boudewyn et al. (2015) argued that listeners predict the phonological form of the critical word, and that a PMN is elicited if phonological analysis of the first phoneme yields a mismatch with the expected phoneme and the predicted candidate cannot be selected. Hagoort and Brown (2000) and van den Brink et al. (2001, 2004) did not take their results to demonstrate lexical form prediction, but concluded that the observed N200 effects are the same as the PMN and instead reflect lexical selection processes. They argued that N200 modulations reflect the relative difficulty in lexical selection from the combination of bottom-up, phonological information and top-down contextual information. Both interpretations thus assume a functional distinction between the N200 component and the N400 component. However, these components can only be considered functionally distinct if they clearly and reliably dissociable across studies. There are reasons to doubt that they are.

8.6.1. Is the N200 truly different from the N400?

From the reviewed studies, it remains unclear whether the N200

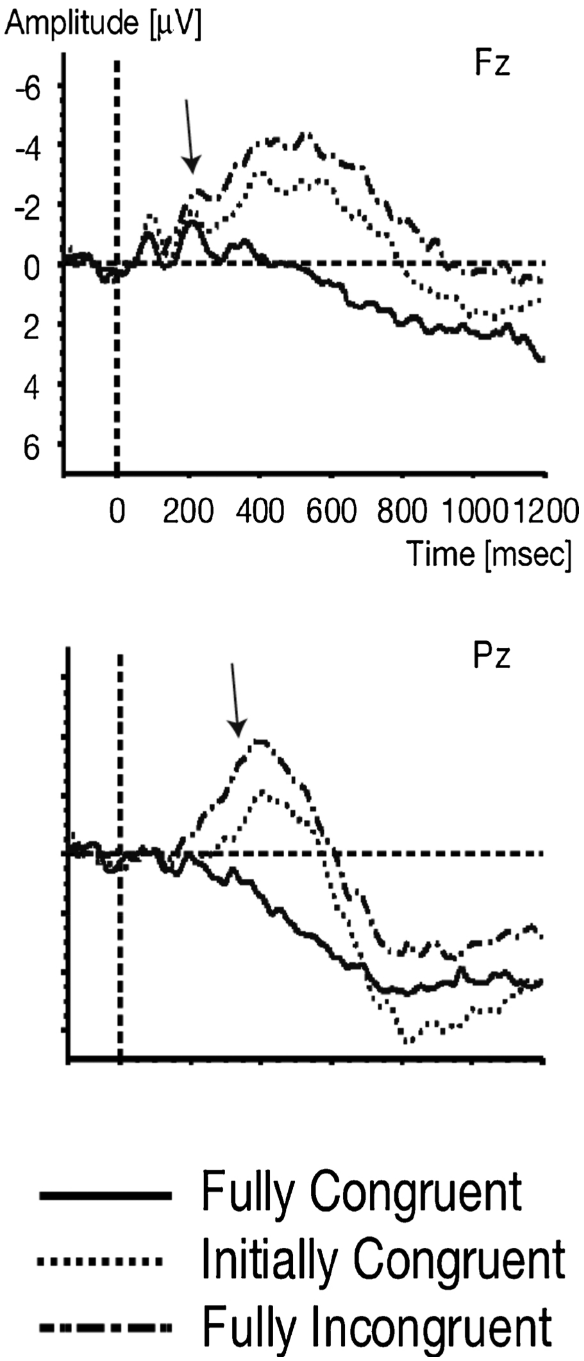

Fig. 16. Grand-average ERPs at Fz and Pz elicited by Fully Congruent, Initially Congruent and Fully Incongruent sentence-final words (adapted from van den Brink et al., 2001). Arrows indicate the peak of the N200 component at Fz, which was only visible at frontal central channels, and the peak of the N400 component at Pz.

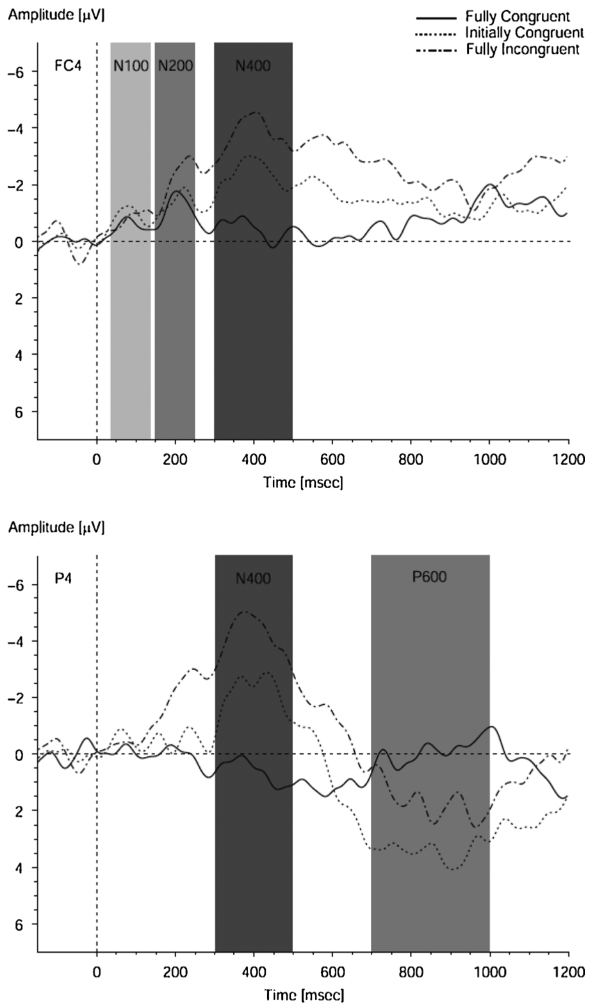

Fig. 17. Grand-average ERPs at FC4 and P4 elicited by Fully Congruent, Initially Congruent and Fully Incongruent sentence-final words (adapted from van den Brink and Hagoort, 2004). Grey bars show the time windows for statistical testing.

effect truly reflects the modulation of a dissociable N200 component or a divergence in the upward flank of two N400 components. Many other spoken language studies have reported N400 effects that start as early as 150–200 ms after word onset (Van Petten et al., 1999; van Berkum et al., 2003, 2005, 2008) without observing (or reporting) any N200 activity, so it seems safe to assume N400 activity can take place in the time window in which some studies report an N200 effect. The clearest evidence for a dissociable N200 effect would be if one observed a modulation of a clearly distinct negative peak with a consistent timing across conditions and experiments that dissipated before the onset of the N400 component. However, that is not what the studies thus far show, and perhaps it is unlikely to begin with due to component overlap when modulations of both the N200 and N400 both occur. Moreover, the criteria to establish N200 activity differ between studies, which makes it very difficult to tell if reports of N200 effects are indeed showing the same pattern in the available studies. Some studies rely on detection of separate N2 and N4 peaks regardless of scalp distribution (e.g., Connolly and Phillips, 1994; Hagoort and Brown, 2000), whereas other rely on solely on scalp distribution (e.g., Boudewyn et al., 2015).

Connolly and Phillips (1994) measured the N200 as the most negative point between 150–350 ms after word onset. This definition assumed that any peak before 350–400 ms after word onset must be an N200 because N400 effects do not peak that early. However, N400 s sometimes peak between 350–400 ms (e.g., Van Berkum et al., 2005), and the N200-definition by Connolly and Phillips makes any most negative point in that window into the 'peak' of a component, even if that negativity was just a random fluctuation (noise) in the ERP signal and no clear N200 component is discernible from the grand-average ERP (let alone subject-average ERPs or single trial ERPs). This definition also forces a functional distinction between two small peaks that appear in close succession during a larger ERP modulation (e.g., N2/N4 peaks in the double mismatch condition), and may just result from random fluctuations (noise) in the ERP signal. Also, no scalp distribution differences were found for the N2 and N4. An alternative account, wherein all anomalous conditions elicit only an N400 component, is plausible:

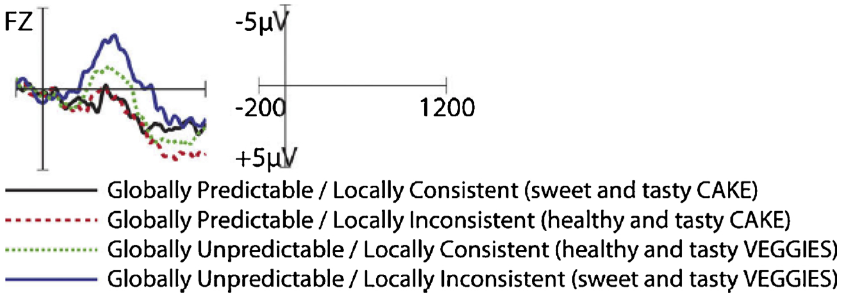

Fig. 18. Grand-average ERPs per condition at electrode Fz, adapted from Boudewyn et al. (2005).

the largest N400 for the double mismatch condition, a smaller N400 for the phonemic mismatch and semantic match condition, and a delayed-onset N400 for the phonetic match and semantic mismatch condition.

Hagoort and Brown (2000) defined N200 activity through visual inspection of peak activity. They observed peaks in the 200–300 ms time window in correct and anomalous conditions, and interpreted them as N200 activity. However, the grand-average ERP waveforms clearly show more fluctuations throughout the whole ERP waveform between 0 and 1500 ms after word onset, not just in the N200 and N400 window. These fluctuations were around the 10 Hz range, which suggests the presence of *residual alpha activity*, that is, alpha activity that is not related to the experimental manipulation. Alpha activity can be of very high amplitude, and excessive alpha is often taken as a reason to exclude a subject from further analysis (e.g., van Berkum et al., 2003; van den Brink et al., 2001, 2004). Even if some subjects with excessive alpha activity are excluded from the analysis, remaining alpha activity in other subjects can still show up in the morphology of the ERP waveform, especially in experiments with low trial numbers (with infinite trial numbers, averaging will cancel out alpha activity that is not functionally related to the manipulation). If the waveform contains residual alpha activity, at least one peak is to be expected every 100 ms because of its approximately 10 Hz frequency. The combination of an upward N400 flank with alpha-related peak activity can thus appear like an N200 peak in the 200–300 ms window (see also Brunelliere and Soto-Faraco, 2013; Groppe et al., 2010; for related discussion). Perhaps that this problem with overlapping alpha activity can be addressed by including a measure of individual alpha into the analysis as a covariate.

An account in terms of residual alpha may also explain an apparent paradox in the results of van den Brink et al. (2001, 2004). In both studies, the peak that was labelled as an N200 component was frontal, but the difference between anomaly and correct words in this early window was visibly larger at posterior channels where no N200-like peak was discernible and where the upward N400 flank to the Fully Incongruent conditions started well before 200 ms after word onset (Fig. 16 and 17). In Van den Brink et al. (2001), the distinction between N200 and N400 was supported by a scalp distribution analysis that revealed a more evenly distributed N200 effect compared to the posterior N400 effect. However, van den Brink et al. (2004) found that the N200 effect was as posterior as the N400 and acknowledged that the N200 effect may well be an N400 effect. The apparent paradox that the differential effect is not where the N200 component is visible disappears if the N200 component is identified as frontal alpha activity whereas the effect in this time window is driven by the early onset of an N400 effect with a more posterior distribution.

In the study by Boudewyn et al. (2015), the grand-average ERPs did not reveal any discernible peak in the N200 time window. They tested for N200 activity in the 200–300 ms time window, and supported the distinction between N200 and N400 activity with a scalp distribution analysis, which revealed more evenly distributed effects in the N200 time window and more posterior effects in the N400 time window (see also van den Brink et al., 2001, although this involved a different time window). This scalp distributional analysis compared one condition (globally unpredictable and locally inconsistent, ‘*sweet and tasty veggies’*) to the average of the three other conditions. However, this analysis does not allow a clear view on the comparison that is strongest and most relevant (globally predictable and locally consistent versus globally unpredictable and locally inconsistent, ‘*sweet and tasty cake/veggies*’, which showed the biggest N400 difference), and cleanest in the sense that the normed cloze values presumably best correspond to the sentences in the experiment. The single-condition ERPs (Fig. 2 in Boudewyn et al.) suggest that the topographical differences in the N200 and N400 window are in part caused by the globally predictable and locally inconsistent condition and the globally unpredictable and locally consistent condition (red and green lines, respectively), but the sentence constraints of these conditions are unclear and one of those conditions also contained unpredictable words (‘veggies’).

In sum, a closer look at the N200 studies suggests that each of these studies defines the N200 in a different way and observes an N200 effect in very differently appearing effects. In all the reported studies, N400 activity is visibly taking off in what is defined as the N200 window, as early as 150–200 ms after word onset. In some studies, the ERP waveforms clearly show 10 Hz alpha activity that could easily be mistaken as N200 peaks. Sometimes a peak in the early time window suffices to label the effect as an N200 effect (regardless of scalp distribution), sometimes a different scalp distribution in the early time window compared to the N400 window suffices to label the effect as an N200 effect (regardless of any observable peak). In light of these concerns, a more parsimonious account of the results is that there is no separate N200 modulation, but that all these studies demonstrate the rapid onset of the N400 component during spoken language comprehension (e.g., Van Petten et al., 1999). I return to this issue and its implications in the general discussion.

8.6.2. *Have other spoken language studies also found N200 effects?*

If N200 effects are ‘real’, then one would expect to observe such effects in any spoken language study with a strong cloze manipulation and expected and unexpected critical words that differ in their first phoneme. However, there are various such studies that do not find/report an N200 effect. For example, a canonical study by Van Petten et al. (1999) found no N200 effect but reported that N400 activity started to diverge for contextually expected and unexpected words as early as 150–200 ms after word onset and well before the words could be uniquely identified. Diaz and Swaab (2007) performed a study specifically to test the N200 hypothesis but failed to find N200 activity during sentence comprehension. Like Van Petten et al. (1999) and Diaz and Swaab (2007), spoken language studies often find early N400 effects of expectancy but no discernible N200 effects (e.g., Corley et al., 2007; Federmeier et al., 2002; MacGregor et al., 2009; Van Petten and Luka, 2012; Van Berkum et al., 2005, 2008; cf. MacGregor et al., 2010). Moreover, if the N200 effect is indeed an index of phonological prediction, the N200 effect (either as a peak or as a scalp distribution that differs from an N400 effect) would occur only in high constraint sentences (high cloze sentences ending with a predictable/unpredictable word), not in low constraint sentences (e.g., low cloze sentences ending with a congruent or anomalous word). Previous studies have not yet made such direct comparisons.

Of course, it is possible that some spoken language studies have overlooked potential N200 effects or did not consider such effects important enough to pursue in their analysis and discussion. Several studies allude to negative peaks in the 100–300 ms time window as potential N200 effects but did not analyze or report these effects in full (e.g., Van Berkum et al., 2005; van den Brink et al., 2006; Rommers et al., 2013). However, the concerns with N200 effects described in this review (i.e., with definition and identification of N200 results, possible influences of residual noise/alpha activity) also apply to those results. Without a full analysis and report, it is unclear whether such hints of N200 activity strengthen or weaken the case for a N200 form-prediction effect that is consistent in time and scalp distribution. For example, van den Brink et al. (2006) performed a very similar study as van den Brink et al. (2001, 2004) and, while they described an N200 component to be visible in their ERP waveforms, they did not analyze that time window. Like in van den Brink et al. (2004), the N200 resembled alpha activity at frontal channels, while the difference between expected and unexpected words was largest at posterior channels, suggesting that, here too, the authors would not be able to meaningfully separate N200 from N400 activity.

8.7. Summary

Several spoken language comprehension studies tested the hypothesis that people can predict the phonological form of upcoming words, by investigating whether people detect a deviation from an expected word upon hearing a single phoneme, the smallest meaningful unit of spoken language. More specifically, these studies tested whether word-initial phonemes that deviate from expected words elicit a N200 effect, an effect that is originally associated with auditory ‘oddballs’ (e.g., Patel and Azzam, 2005, for a review) and therefore considered distinct from the N400 effect associated with semantic processing (Connolly and Phillips, 1994). The current review has highlighted the difficulty in disentangling the N200 effect from the N400 effect, echoing some of the concerns that have been raised previously (e.g., Diaz and Swaab, 2007; Groppe et al., 2010; Van Petten et al., 1999). A major issue is that some claims of N200 effects depend on identifying an N200 peak (regardless of scalp distribution), whereas other claims are based on topographical differences between the early time window and the N400 time window (regardless of detectable peak). Detection of a clear and meaningful N200 peak may not be necessary to infer N200 activity (e.g., Luck, 2014). However, claims solely based on topographical differences are not straightforward either. For example, such topographical differences would need to appear consistently across studies, which they do not. In addition, they would lend evidential support for prediction if they do not occur in manipulations without prediction mismatch, for example in low-constraint sentences. As far as I know, no such comparison has been made in the available literature.

Therefore, a more parsimonious account can be considered, namely that phonological deviations from expected words elicit rapid-onset N400 effects but no distinct N200 effect (e.g., Van Petten et al., 1999). This raises the question how N400 activity unfolds over time and corresponds to the amount of word information available at a given time.

9. The N250

The N250 ERP component is a negative deflection that peaks at about 250 ms after visual word onset and that is often observed in masked priming studies on visual word recognition (e.g., Chauncey et al., 2008; Holcomb and Grainger, 2006; Grainger et al., 2006; Kiyonaga et al., 2007; for a review, see Grainger and Holcomb, 2009). N250 amplitude is larger for target words that differ from preceding stimuli by a single letter compared to targets that completely overlap with their primes (e.g., teble-TABLE compared to table-TABLE). The N250 component is thought to reflect word form processing, and may therefore be sensitive to deviations from an expected word form.

9.1. Brothers et al. (2015)

An ERP study by Brothers et al. (2015) reported an N250 effect of sentence-level prediction. Brothers et al. used a novel paradigm to overcome a methodological limitation of previous N400 studies on prediction. They argued that previous studies had not resolved the issue of whether unexpected words elicit enhanced N400 s because they mismatch the predicted word or because they are less contextually supported (e.g., less plausible or less strongly related to words on the sentence context). In Brothers et al., participants read medium-constraint sentences (cloze range 40–60%) that ended with either low cloze or medium cloze words (1 or 50% cloze probability, respectively), and were asked to predict the last word of the sentence, and to indicate after each sentence whether their prediction had been correct. Based on the participants’ responses, medium-cloze words were categorized as predicted or unpredicted despite having a similarly good contextual fit, whereas low-cloze words were unpredicted and had a poor contextual fit. They hypothesized that if lexical pre-activation facilitates early stages of orthographic processing, lexical prediction would modulate the N250 since this component reflects processing of word form (Grainger and Holcomb, 2009).

As shown in Fig. 19, predicted words elicited a widespread, steep positivity in the 200–600 ms time window, unpredicted medium cloze words elicited a much less steep positivity and a clear N400 peak at frontal channels, and unpredicted low cloze words elicited clear N400 component at all channels. Analysis was performed in the 200–300 ms (N250) and 300–500 ms (N400) time windows. The main finding was that in the N250 window, predicted words elicited significantly more positive ERPs than the other two conditions, which did not differ from each other. In the N400 window, predicted words elicited more positive ERPs than the two other conditions, and unpredicted medium cloze words elicited more positive ERPs (i.e., smaller N400 s) than unpredicted low cloze words.

Based on these results, Brothers et al. concluded that the early negativity effect of correct prediction precedes the N400 effect of contextual support by about 100 ms. They interpreted the early negativity as an N250 effect, that is, a reduction of the N250 for correctly predicted words, and took this as evidence for the effect of lexical preactivation on early orthographical processes.

9.2. Discussion

Brothers et al. used an innovative paradigm to address an old research question. Ever since the landmark study by Kutas and Hillyard (1984) that showed the effect of cloze probability on the N400, researchers have asked whether and to what extent N400 modulations are driven by prediction or by non-predictive processes such as contextual integration (e.g., Lau et al., 2009, 2016; Ito et al., 2016; Nieuwland et al., 2018a; for a review, see Kutas and Federmeier, 2000, 2011; Lau et al., 2008; Van Berkum et al., 2008; Van Petten and Luka, 2012). But Brothers et al. concluded that lexical prediction effects are reflected in an effect that is distinct from the N400, although the N250 and N400 have some spatial and temporal overlap. While the paradigm of Brothers et al. is certainly innovative and their results seem to be replicable, the morphology of the ERP waveforms shed some doubt on whether the effect of prediction should be considered an N250 effect, especially in light of the prediction-task.

9.2.1. Does the N250 effect reflect a modulation of the N250 component?

Brothers et al. interpret their prediction-related negativity as an N250 effect based on the effect onset around 250 ms (as visible in the difference waveform in Fig. 19), which was earlier than the onset of the context-related N400 effect. However, an effect onset at 250 ms and a

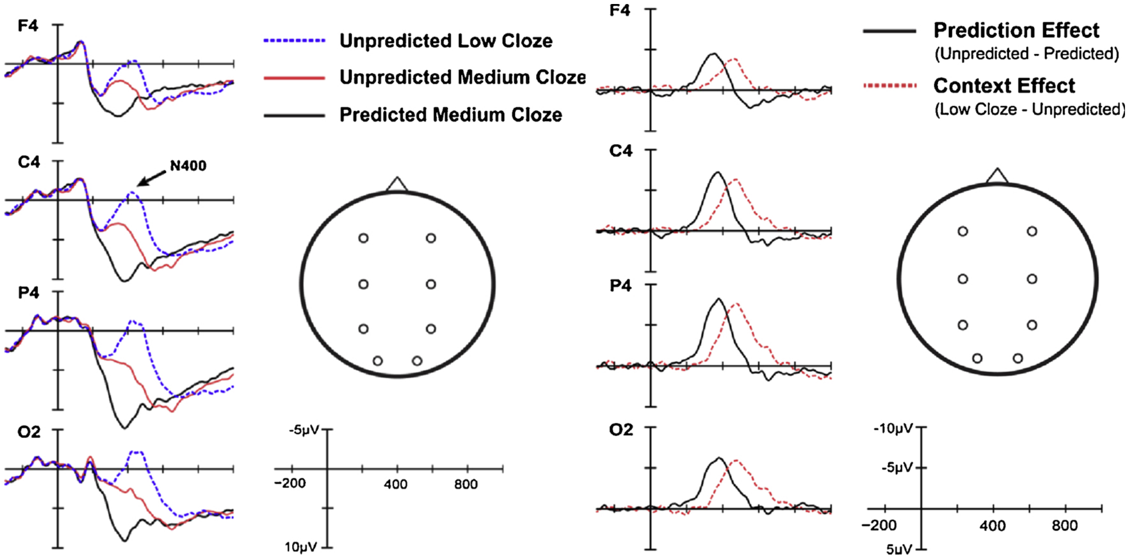

Fig. 19. ERP results of Brothers et al. (2015). Left graphs: Grand-average ERPs elicited by predicted words and by unpredicted words of low or medium cloze. Right graphs: Difference waveform of the prediction effect (unpredicted medium cloze– predicted medium cloze) and the context effect (unpredicted low cloze – unpredicted medium cloze).

maximum difference just before 400 ms is not unusual for an N400 effect even in written language studies (e.g., Van Berkum et al., 1999). In that sense, the early effect they report is compatible with an N400 effect (see also Nieuwland et al., 2018b). The scalp distribution of the N250 effect was found to be more even than the posterior N400 effect (but see Dave et al., 2018, who did not find such a difference). In Brothers et al., no discernible negative peak was visible in the waveforms between 200–300 ms. In fact, in that time window, the ERPs are positive-going. In other words, there is an amplitude difference in the time period where the N250 component can be observed, but there are no clear N250 components. One can argue that a clear N250 component in the single-condition ERPs is not absolutely required to conclude that there is a specific effect that differs from a modulation of the N400. However, evidence for an N250 effect (and therefore for the facilitation of orthographic processing by prediction) would be stronger if there are two N250 peaks that differ in amplitude. Brothers et al. based their conclusions on the orthographic priming literature, but such priming studies typically compare the amplitude of two clear N250 components elicited by identical stimuli as a function of the prime (e.g., Grainger and Holcomb, 2009).

This issue can be further demonstrated with a hypothetical example using the well-known P300 and N400 components, which have a similar scalp distribution and also overlap in time. If condition A elicits a P300 component that is more positive than the P300 component for condition B and the difference between A and B peaks around 400 ms, then we would conclude that A elicits a P300 effect compared to B, not that B elicited a negative (N400) effect compared to A. Similarly, the ERP patterns in Brothers et al. do not lend strong evidence for a prediction-based modulation of the N250 component because there appears to be no N250 component (or it would have to be one that is not visible from the grand-average ERPs), and the effect is driven by a strong positive ERP deflection to predicted words. The question then becomes why the predicted words in Brothers et al. elicited a distinct positive component, which is visible in the ERPs. Relevant to this question may be the role of the prediction-task.

9.2.2. What is the influence of the prediction task?

Participants were instructed to try to predict sentence-final words and indicate by button-press whether they had predicted correctly. This procedure was necessary to separate predicted and unpredicted trials. However, this prediction task is completely explicit and caution is warranted in trying to generalize from such a task to predictive processing during natural language comprehension. Moreover, the prediction task may have elicited ERP activity effect that one would not observe without the task, in particular modulations of the P300 or P300b, the ERP component commonly associated with task-related decision processes (Polich, 2007). The P300 is enhanced for stimuli that stand out in terms of their relevance for the task (targets), and its latency is dependent on the time it takes participant to evaluate and categorize stimuli (Kutas et al., 1977; McCarthy and Donchin, 1981). A re-interpretation of the Brothers et al. results is that the early difference in the N250 window is in fact, at least in part, a P300 effect. Due to their task-relevance, predicted words elicited a rapid and strong P300 response (Fig. 17). The two unpredicted conditions may also elicit task-related P300 responses, but those may have a less rapid downward flank and peak later, and are counteracted by the N400 peak (for medium cloze word, the N400 peak was only visible at frontal channels). While the effect of prediction involves a comparison between one condition that elicits a rapid P300 peak and one condition that elicits an N400 (and a later P300), yielding an early effect, the effect of context involves a comparison between two conditions that both show a clear N400 flank, yielding a later effect.

The suggested P300 interpretation is further illustrated with a visual comparison to the results of Roehm et al. (2007). Roehm et al. investigated the effect of different task demands on processing of semantic relations between word pairs (antonymy, semantically related or unrelated). In Experiment 1 (Fig. 20, top graphs), where participants performed a sensicality judgment after reading high-constraint sentences ("The opposite of black is.."), antonyms ('white') elicited a rapid and strong P300 peak and no N400 peak, whereas related words ('yellow') and unrelated words ('nice') showed later P300 peaks and clear N400 peaks. The Brothers et al. results look very similar to these results in the 200–500 ms window. In Experiment 2 (Fig. 20, lower graphs), participants performed a lexical decision task on word pairs presented out of sentence context, and all three conditions elicited an N400 peak but no clear P300 peak (a P300 peak was observed for pseudowords). Of note, Roehm et al. did not argue that the P300 is a component that is associated with linguistic prediction during regular language comprehension[14], and suggested that ERP studies on language processing require a much more detailed screening for possible task- or

[14] Some authors associate the P300 with prediction in highly idiomatic expressions (Vespignani et al., 2010). Such studies warrant replication and further investigation, but do not indicate a default involvement of P300 activity in language comprehension.

strategy-related positivity effects than previously assumed. I concur with that conclusion, and the Brothers et al. study may be one such study where the potential impact of P300 activity was not considered in the conclusions.

This illustrates a more general problem with experimental designs that elicit both P300 and N400 activity (or other components with spatiotemporal overlap), namely that it becomes very hard to dissociate the two components meaningfully. The components overlap substantially in time and scalp distribution, and their negative and positive voltage cancel each other out at the scalp surface, which can lead to unpredictable patterns since the timing of the P300 peak depends on the timing of decision processes (Kutas et al., 1977).

9.3. Summary

The Brothers et al. study used a novel approach to tackle an old question about when the effects of prediction and contextual becomes visible during semantic processing. Participants were instructed to predict sentence-final words and indicate whether they had done so correctly after each sentence had ended. By splitting trials on prediction accuracy, Brothers et al. were able to separate the ERP effect of correct prediction (predicted versus unpredicted words of similar cloze probability) and that of contextual support (unpredicted words of low versus medium cloze probability). They reported an N250 prediction effect and an N400 contextual support effect. They took the N250 prediction effect as a modulation of the N250 ERP component that is associated with orthographic processing (Grainger and Holcomb, 2009), thus concluding that prediction impacts the early stages of visual word recognition, preceding the integration of a word's meaning with the sentence context.

This is an interesting hypothesis, but their conclusion is limited by the fact that they did not observe a clear N250 component, which hampers their conclusion about an N250 effect. In two follow-up studies with the same prediction instruction (Brothers et al., 2017; Dave et al., 2018), a similar patterns of results was observed but labelled as an early N400 effect instead of an N250 effect, perhaps because no support from scalp topography was obtained for a functional distinction between activity in the N250 and N400 time windows. This could suggest that Brothers and colleagues have revised their interpretation of what their observed pattern reflects. In all these studies, therefore, the effects of prediction (as defined in this task) appear earlier than those of context support. However, ultimately it isn't clear from these studies whether prediction indeed impacts orthographical processing because this conclusion was premised on the observation of an N250 effect.

I have described a plausible alternative interpretation of their results, namely that the early prediction effect arose partly due to a strong P300 response to predicted words, which were by definition most relevant to the prediction task. This P300 response started at about 200–250 ms after word onset, at about the same time as the standard P2 component associated with onset of visual stimuli, and this P300 response may have not occurred, or not as early, for unpredicted words. Low and medium cloze unpredicted words elicited an N400 component that initially shared an upwards flank, therefore leading to a later difference. More cautiously, the ERPs in their experiment are an unknown mix of P300 activity elicited by the task and N400 activity elicited by the words (and possibly modulated by the task). The observed P300 activity seems to reflect the response to a recognized word, and as such can be said to show the early effect of prediction on recognition (participants recognize a word as being the one they predicted). However, it is unclear whether the observed P300 activity reflects a recognition process or word-form analysis itself or only a decision-related process after recognition has occurred. In addition, because the observed results can only be obtained with a prediction task, the conclusions do not readily generalize to comprehension processes in absence of a prediction task.

10. General discussion

This literature review covers a range of early brain responses that are associated with linguistic prediction, with 'early' loosely defined as occurring before N400 peak amplitude. The motivation for writing this review was to offer an in-depth discussion of these effects to fill a gap in knowledge left by previous literature reviews on linguistic prediction, which have focused primarily on 'late' effects such as the N400 and late positive component (Kutas et al., 2011; Van Berkum, 2009; Van Petten and Luka, 2012). Early prediction effects stand out because they are sometimes obtained on well-known brain responses associated with non-linguistic (i.e. perceptual or attentional) processing. For that reason, these effects are sometimes taken as evidence that linguistic predictions are implemented in terms of perceptual processes (the sensory hypothesis) or that linguistic predictions impact the recognition of a specific word form before its meaning can be integrated with sentence context (the recognition hypothesis). I reviewed 8 different components/effects, and this variety reflects the many differences between the reviewed studies in the used stimuli, modality and experimental approach. I have discussed the details of the design of each of these studies, summarized their results and conclusions, and discussed their strengths and potential limitations.

In the title of this review, I posed a question: Do 'early' brain responses reveal word form prediction during language comprehension? One possible, optimistic answer is "yes, because we can take the reported results at face value, which gives us a large body of published studies that report statistically significant effects on early brain responses, and those collectively support the conclusion that people can predict the form of upcoming words (or word categories)". An alternative, more cautious answer is "no or too early to say, because many of these effects do not stand up to close scrutiny and they have either failed to replicate or have yet to be replicated, and we cannot take these isolated findings as clear-cut, convergent support for the same conclusion." As is probably clear by now, I am inclined towards the more cautious answer, because evidence for word form prediction from early brain responses is not nearly as strong or straightforward as it is sometimes portrayed (Hagoort, 2017; Kuperberg and Jaeger, 2016; Pickering and Gambi, 2018). This body of literature often features in short summary form in theoretical reviews that argue for the importance of prediction (Christiansen and Chater, 2016; Kuperberg and Jaeger, 2016; Pickering and Garrod, 2013). However, such summaries might suffer from confirmation bias and there has been hardly any post-publication, critical review of methods, data and conclusions, or attempts at replication.

Each of the reviewed studies has taken a different and often very creative approach to investigate a similar issue, namely the issue of whether the effects of context-based predictability can be observed on early components that are thought to reflect either non-linguistic processing or 'pre-semantic' linguistic processing of word form. But the reviewed studies also suffer from similar obstacles to interpretation. Some of these obstacles are fairly specific to EEG/MEG research, such as the lack of a clear or consistent definition and functional interpretation of an EEG/MEG component/effect, potential data distortions resulting from pre-processing procedures, difficulty in meaningfully teasing apart different EEG/MEG components/effects that overlap in time or space, and potential distortions by EEG/MEG effects of a judgment task. Other obstacles are much more general because they apply to other domains of research, but they are possibly even more important. These include the lack or failure of replication (Zwaan et al., 2018), small sample sizes (Button et al., 2013), statistical analysis techniques with non-maximal generalization performance (Barr et al., 2013), multiple comparison problems (e.g., Luck and Gaspelin, 2017), and large numbers of researcher degrees of freedom combined with analysis strategies that are often contingent on the data (Gelman and Loken, 2013; Luck and Gaspelin, 2017). When considering all these potential issues in conjunction with known publication bias (see also Forstmeier et al., 2017),

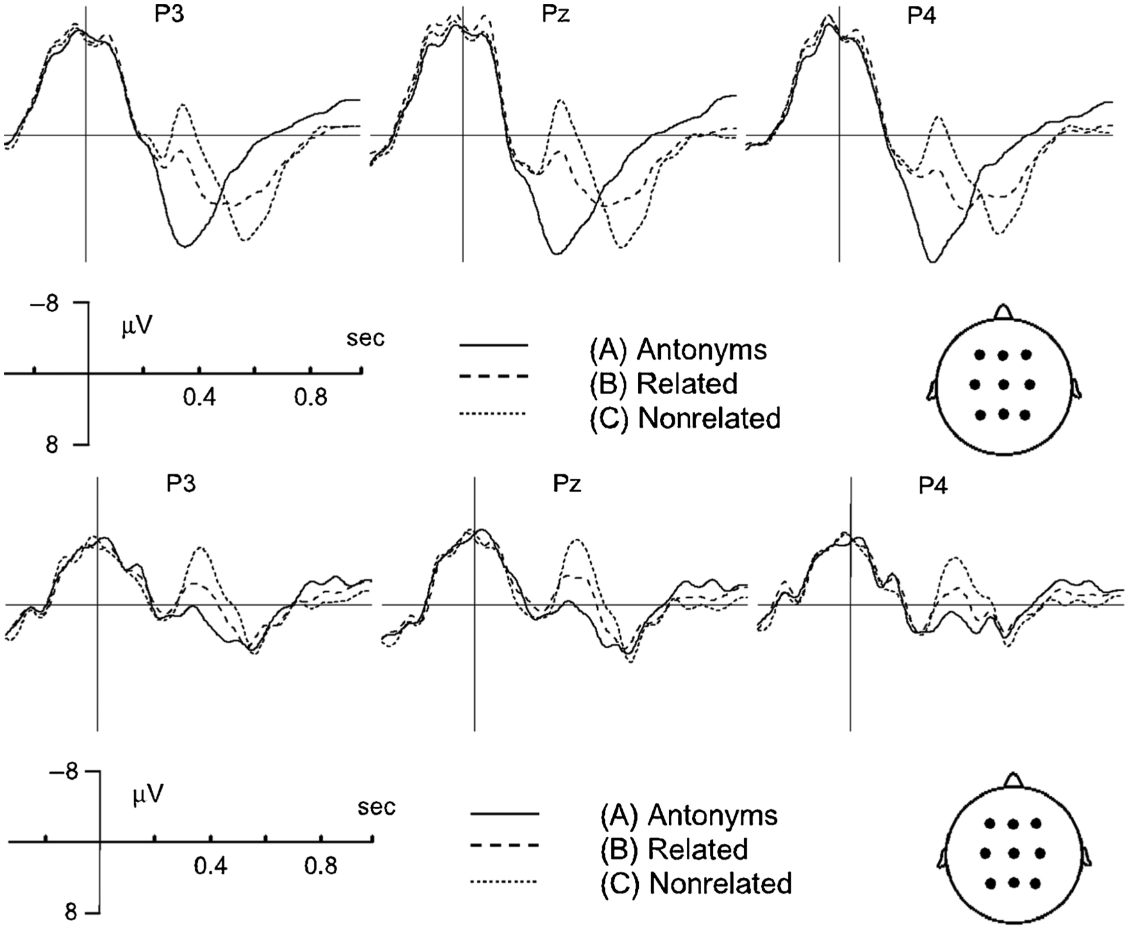

Fig. 20. ERPs associated with antonyms, related words and unrelated words from Roehm et al. (2007). Top graphs: Experiment 1, the words appeared in a constraining sentence context. Lower graphs: Experiment 2, the words appeared as word pairs in a lexical decision task.

one could doubt whether there is convincing evidence for word form prediction from the reviewed studies. While the studies have yielded statistically significant effects in support of their hypothesis, statistical significance conveys little information when measurements are noisy, such as in low-sample neuroimaging experiments where there are many possible ways to analyze the data. Using data-contingent analyses, such studies yield exaggerate estimates of effect size that are unlikely to replicate (Gelman and Loken, 2014; Loken and Gelman, 2017; Vasishth et al., 2018). This will lead a field astray, in particular when direct replication research is rather rare in that field.

I wish to emphasize that I am not arguing that the reviewed studies involve questionable research practices, nor do I think that the approaches in the reviewed studies are unusual for the field of cognitive neuroscience of language. In fact, some of the reported approaches are quite common (for discussion, see Luck and Gaspelin, 2017), and several of the criticisms in this review can be raised against some of my own work. As Gelman and Loken (2014) note, “whereas research is hard, criticism is easy and flaws can be found in any research design if you look hard enough.” I am also not arguing against a role for prediction in regular language comprehension. My main point is that the results of isolated studies should be treated with more caution when making major theoretical claims about prediction. Several highly influential studies garner a steady stream of citations as ‘strong evidence’ for form prediction despite not having been replicated or having failed to replicate (for discussion, see also Nieuwland et al., 2018a). This bring me to the question of where to go from here, and I would like to make a few methodological recommendations.

10.1. Where do we go from here?

Fortunately, solutions to address the general criticisms are well-known and relatively straightforward: trying to establish the (direct/indirect) replicability of the reviewed effects (Zwaan et al., 2018), standard pre-registration of analysis procedures to reduce data-contingent analyses and collecting larger numbers of observations to obtain more realistic effect sizes (Button et al., 2013; Munafo et al., 2017), applying current statistical procedures across the board to better account for known sources of variance (Barr et al., 2013), include more distributional information and measures of uncertainty/confidence in the visualization of the results (Allen et al., 2012; Rousselet et al., 2016), make data and analysis scripts readily and publicly available such that alternative explanations for a given finding can be explored and further tested (Wicherts et al., 2011), and reduce the distorting influence of the arbitrary cut-off of ‘statistical significance’ (e.g., Bakker et al., 2012; Gelman and Stern, 2006; Loken and Gelman, 2017; Vasishth et al., 2018). In addition, mainstream science journals should do more to facilitate and embrace replication of research they publish.

These ‘open-science’ practices are slowly becoming more common in other domains of psychological research (Asendorpf et al., 2013; Forstmeier et al., 2017), and I cannot think of a good reason why they cannot be adapted to improve the neurobiological study of language (e.g., see Nieuwland et al., 2018).

Solutions to the EEG/MEG-specific criticisms are perhaps not always straightforward, but some of my suggestions have already featured in the effect-specific sections, and I am not certainly the first to make them. Some suggestions involve improvements of experimental design. For example, minimize or better account for effects arising from stimuli occurring before the critical words (i.e., 'baseline problems'). Minimizing differences in stimulus materials right before the critical words is important because such differences may generate effects that spill-over into the earliest time windows of the critical word. The controversy around the well-known ELAN effect demonstrates this point (see Steinhauer and Drury, 2012; Osterhout et al., 2004; Tanner, 2015). Taking steps to avoid baseline problems is important, but unfortunately not a guarantee for success. Baseline problems can arise from structural differences between conditions in terms of pre-target content, but they can also arise from noise (random data fluctuations) and are more likely to occur in data with relatively few observations. Another suggestion is to minimize the impact of a secondary task, by avoiding use of a task altogether or avoiding use of a task that focuses attention on the manipulation of interest. EEG/MEG bypasses the need for an explicit behavioral task and can detect rapid and/or brief quantitative and qualitative changes in neural activity related to a stimulus. This allow for experiments wherein language comprehension is itself is the primary, implicit task. A secondary, explicit task (e.g., acceptability judgments) is sometimes very useful and informative (e.g., to determine whether two groups of participants evaluate sentences equally accurately), and sometimes boosts some of the processes involved in comprehension. However, such a task itself can generate brain responses that distort the effects of interest and induce strategic behaviors that are not representative of the processes that one is trying to explain (Roehm et al., 2007; Vega-Mendoza et al., 2017). Demonstrations of the broader relevance of prediction are most convincing when participants are simply reading or listening to naturalistic materials under the instruction to comprehend the materials to their best ability, or perhaps when they are instructed to answer questions about the content of the stimulus materials that have nothing to do with the manipulation of interest (for further reading, see Willems, 2015).

Other methodological suggestions involve how data are processed, presented and interpreted. For example, EEG studies should use a reference and filter procedure that is suitable for studying the component of interest and matches the ones used in relevant previous studies, or present data from different procedures to show that the claim is supported *irrespective* of the reference and filter. This is crucial for a quantitative and qualitative comparison of effects across studies. A suggestion for presentation is to re-popularize the graphical presentation of difference waveforms. Many studies do not depict difference waveforms (including some of my own, I should add), possibly because the difference waveform sometimes seem redundant. However, people may have difficulty in mentally subtracting two time courses, especially on strongly sloped lines (e.g., Rousselet, 2016). Plotting difference waveforms is then a straightforward way to inspect whether a difference between conditions is indeed visible on multiple, separable components or whether the difference is extended and slowly developing across multiple components but not actually a clear separate modulation of two components. A related suggestion is to be careful with inferring the modulation of a specific component from an observed difference between conditions (see Luck, 2014). On one hand, a difference in a specific time window is not itself evidence for an effect on a component, which means that the interpretation should be guided by inspection of the waveform morphology and/or scalp distributions. On the other hand, sometimes even an ostensibly clear definition such as "the most negative peak in this time window" can be problematic, and sometimes there are no discernible peaks in the ERP waveform (in which case analysis of scalp topography can help to distinguish two ERP effects). I have noted this problem in particular for the distinction between the N200 and N400, and the potential issue with overlay of alpha activity and N400 activity. Novel avenues of exploring and defining brain responses are needed to advance this line of research, for example by better taking into account spatial characteristics of components via source modelling, or by basing definitions on output from dimensionality reduction techniques such as principal/independent component analysis (PCA/ICA).

10.2. Do people routinely predict word form during language comprehension?

Current theories of language comprehension assume that people constantly try to generate predictions about upcoming words, even about the details of their physical form, in order to keep up with rapidly unfolding linguistic input. For the sake of the argument, I will play devil's advocate and posit that while the reviewed effects are suggestive, it is too early to conclude that they are 'true' effects or truly the effects claimed to be, either because the effect has not been replicated, or because there are plausible, alternative interpretations that cannot (yet) be excluded. Let us consider the implications.

First, the reviewed studies may offer little or no clear evidence for the sensory hypothesis, which states that people generate predictions about the perceptual attributes of words (or word categories) and these predictions are implemented as low-level perceptual representations in primary visual/auditory cortex. These studies mostly concern the Early Cloze Positivity, ELAN, M100, N1 and the P130 effects. The sensory hypothesis follows from the predictive coding framework, which already pervades many domains of cognitive psychology (Bubic et al., 2010; Clark, 2013; Hickok, 2012; Kilner et al., 2007), and may seem like a promising extension of prediction-based theories of language comprehension (Pickering and Garrod, 2013). However, lack of clear evidence from the reviewed studies means that support for the sensory hypothesis must be sought elsewhere. Some studies, for example, show the visual presentation of a word impacts early perceptual processing of a spoken version of that word (e.g., Sohoglu et al., 2012). But such evidence, alike current evidence for predictive coding in perception sciences, comes from repetitive task designs wherein participants make explicit judgments about a very large number of highly similar, sometimes perceptually-ambiguous stimuli. Such designs, which focus on speech perception rather than speech comprehension, might encourage participants to generate sensory predictions as they are relevant to the task. But does it make sense for the language comprehension system to continuously try predicting upcoming information all the way down at the primary perceptual level? Whereas pre-activation of aspects of word meaning can emerge naturally from a representation of the context itself (see Baggio, 2018; Kutas et al., 2011, for discussion), an active prediction of lexical detail would make sense if the prediction of one unique word is sufficiently strong that it can materialize into a specific visual or auditory form representation. This might be the case in some experiments, for example, if participants see a written word before it is spoken (or a picture before a written word, as in Dikker and Pylkkanen, 2011). In everyday language situations, some words might be highly predictable because they occur repeatedly throughout a conversation, but the highly variable context of language use may not generally permit very strong lexical predictions often enough to make sensory prediction worthwhile. It is not clear that sensory word prediction is an integral or even important mechanism of the language system, whose task is primarily to transfer and infer meaning (see also Baggio, 2018; Martin, 2016). Sensory prediction, if reliably demonstrated, may merely be an epiphenomenon of accessing word meaning in a highly perception-oriented and repetitive task-based environment.

It remains unclear what the reviewed results say about regular language comprehension. Demonstrations that perceptual prediction *can* happen in an experiment do not mean it *will* happen outside of the experiment, and the ecological validity of many (if not all of) the reviewed studies can be questioned, in particular the studies where participants have the explicit instruction of predicting words or of detecting anomalies in highly repetitive sentence constructions and even repetition of the same sentence contexts. Such factors may have caused participants to engage in strategic prediction. How these factors influence reading strategies, saccadic behavior and perhaps early neural activity is unclear, and whether the observed effects can be replicated in circumstances more similar to natural reading and listening and conversation remains to be seen (e.g., Willems, 2015).

A second implication is that there may be little or no clear evidence for the recognition hypothesis, which holds that effects of prediction on lexical access precede and are distinct from N400 activity. These studies concern the P2, N200/PMN and N250. But the recognition hypothesis, at least how it is formulated in some of the reviewed studies, may be problematic to begin with. It is entirely predicated on an older view of N400 activity as reflecting purely postlexical processing (Brown and Hagoort, 1993), taking place after the word has been recognized and its meaning has been accessed. However, this post-lexical view has fallen out of fashion because a large body of literature suggests that N400 activity reflects semantic access processes (e.g., Barber and Kutas, 2007; Kutas and Federmeier, 2011; Lau et al., 2009, 2009) and those processes may be decoupled from word recognition. For example, semantic access processes, as reflected in N400 activity, can be initiated before a word is recognized (e.g., Van Petten et al., 1999), and N400 activity is also generated by non-words (which cannot be recognized as words; see Kutas and Federmeier, 2011, for discussion). Moreover, the reviewed studies seem to have great difficulty in reliably differentiating between activity of the N400 and of other, 'earlier' components. This difficulty stems from the timing of N400 activity, which can start in the same time frame as reported for the 'early' components, namely as early as 200 ms after onset of a written word (e.g., Kutas and Federmeier, 2011; Nieuwland et al., 2018a) or 150–200 ms after spoken word onset (i.e. well before a spoken word is uniquely identifiable; Van Petten et al., 1999).

Thus, the available results perhaps demonstrate that prediction facilitates the lexico-semantic access processes associated with early N400 activity rather than recognition processes associated with a unique pre-N400 component (see also Van Petten et al., 1999, for discussion). Early N400 effects are not indisputable evidence for actual prediction of word form. Maybe people predict an entire word meaning without its specific form (i.e., prediction of a lemma but not a lexeme, to borrow terms from the language production literature; Roelofs et al., 1998), and then use the first part of a spoken word to establish whether or not a different meaning needs to be retrieved. However, early N400 effects of prediction are also not incompatible with word form prediction, and solving this issue would require a better understanding of the processes occurring within the first 100–300 ms after onset of a spoken word in a sentence context. Major advances in mapping the spatio-temporal dynamics of early lexical processing have thus far not involved sentence-level comprehension (e.g., MacGregor et al., 2012; Shtyrov and Pulvermüller, 2007). In addition to trying to disentangle early N400 activity from other components, further research could thus zoom in on the comparison between earlier and later aspects of prediction-related N400 activity, their underlying neural generators, and their sensitivity to different information carried by the input (for discussion, see also Baggio and Hagoort, 2011; Nieuwland et al., 2018a; Pylkkanen and Marantz, 2003).

11. Conclusions

This review covers studies on sentence- or discourse-level language comprehension which claim that prediction of word form can be demonstrated on 'early' (pre-N400) brain responses. Some of these studies claim support for the sensory hypothesis, which holds that word form prediction are represented as activity in primary sensory cortices. Some of the studies have claimed support for the word recognition hypothesis, wholds holds that form predictions facilitate the processes by which words are recognized, a process that takes place before its meaning is related to the context. In my in-depth review of these studies, I conclude that the evidence for the sensory hypothesis and word recognition hypothesis from these studies is weak and inconsistent, and I caution against a strong reliance on well-cited, but not-replicated or not-yet-replicated findings in theories of predictive processing. There is a desperate need for replication of previous findings, both direct and indirect, before the theoretical importance of word form prediction during sentence or discourse comprehension can be meaningfully evaluated.

References

Acunzo, D.J., MacKenzie, G., van Rossum, M.C.W., 2012. Systematic biases in early ERP and ERF components as a result of high-pass filtering. J. Neurosci. Methods 209 (1), 212–218. https://doi.org/10.1016/j.jneumeth.2012.06.011.

Allen, E.A., Erhardt, E.B., Calhoun, V.D., 2012. Data visualization in the neurosciences: overcoming the curse of dimensionality. Neuron 74 (4), 603–608. https://doi.org/10.1016/j.neuron.2012.05.001.

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

Altmann, G.T., Mirkovic, J., 2009. Incrementality and prediction in human sentence processing. Cogn. Sci. 33 (4), 583–609. https://doi.org/10.1111/j.1551-6709.2009.01022.x.

Asendorpf, J.B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J.J.A., Fiedler, K., et al., 2013. Recommendations for increasing replicability in psychology. Eur. J. Pers. 27 (2), 108–119. https://doi.org/10.1002/per.1919.

Awh, E., Jonides, J., 2001. Overlapping mechanisms of attention and spatial working memory. Trends Cogn Sci 5 (3), 119–126.

Baggio, G., 2012. Selective alignment of brain responses by task demands during semantic processing. Neuropsychologia 50 (5), 655–665.

Baggio, G., 2018. Meaning in the Brain. MIT Press ISBN 9780262038126.

Baggio, G., Hagoort, P., 2011. The balance between memory and unification in semantics: a dynamic account of the N400. Lang. Cogn. Process. 26 (9), 1338–1367. https://doi.org/10.1080/01690965.2010.542671.

Bakker, M., van Dijk, A., Wicherts, J.M., 2012. The rules of the game called psychological science. Perspect. Psychol. Sci. 7 (6), 543–554. https://doi.org/10.1177/1745691612459060.

Balota, D.A., Yap, M.J., Cortese, M.J., 2006. Visual word recognition: the journey from features to meaning (a travel update). Handbook of Psycholinguistics, second edition. pp. 285–375.

Barber, H.A., Kutas, M., 2007. Interplay between computational models and cognitive electrophysiology in visual word recognition. Brain Res. Rev. 53 (1), 98–123. https://doi.org/10.1016/j.brainresrev.2006.07.002.

Barr, D.J., Levy, R., Scheepers, C., Tily, H.J., 2013. Random effects structure for confirmatory hypothesis testing: keep it maximal. J. Mem. Lang. 68 (3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001.

Boudewyn, M.A., Long, D.L., Swaab, T.Y., 2015. Graded expectations: predictive processing and the adjustment of expectations during spoken language comprehension. Cogn. Affect. Behav. Neurosci. 15 (3), 607–624. https://doi.org/10.3758/s13415-015-0340-0.

Brothers, T., Swaab, T.Y., Traxler, M.J., 2015. Effects of prediction and contextual support on lexical processing: prediction takes precedence. Cognition 136, 135–149.

Brothers, T., Swaab, T.Y., Traxler, M.J., 2017. Goals and strategies influence lexical prediction during sentence comprehension. J. Mem. Lang. 93, 203–216. https://doi.org/10.1016/j.jml.2016.10.002.

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

Brunelliere, A., Soto-Faraco, S., 2013. The speakers' accent shapes the listeners' phonological predictions during speech perception. Brain Lang. 125 (1), 82–93. https://doi.org/10.1016/j.bandl.2013.01.007.

Bubic, A., von Cramon, D.Y., Schubotz, R.I., 2010. Prediction, cognition and the brain. Front. Hum. Neurosci. 4, 25. https://doi.org/10.3389/fnhum.2010.00025.

Bulkes, N., Christianson, K., Tanner, D., 2018. Semantic Constraint, Reading Control, and the Granularity of Form-based Expectations During Sentence Processing: Evidence From ERPs. Article in Preparation. .

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

Camblin, C.C., Ledoux, K., Boudewyn, M., Gordon, P.C., Swaab, T.Y., 2007. Processing new and repeated names: effects of coreference on repetition priming with speech and fast RSVP. Brain Res. 1146, 172–184. https://doi.org/10.1016/j.brainres.2006.07.033.

Carreiras, M., Armstrong, B.C., Perea, M., Frost, R., 2014. The what, when, where, and how of visual word recognition. Trends Cogn. Sci. 18 (2), 90–98. https://doi.org/10.1016/j.tics.2013.11.005.

Chambers, C., 2017. The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice. Princeton University Press.

Chauncey, K., Holcomb, P.J., Grainger, J., 2008. Effects of stimulus font and size on masked repetition priming: an event-related potentials (ERP) investigation. Lang. Cogn. Process. 23 (1), 183–200. https://doi.org/10.1080/01690960701579839.

Christiansen, M.H., Chater, N., 2016. The Now-or-Never bottleneck: a fundamental constraint on language. Behav. Brain Sci. 39, e62. https://doi.org/10.1017/S0140525X1500031X.

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

Coffman, B.A., Kodituwakku, P., Kodituwakku, E.L., Romero, L., Sharadamma, N.M., Stone, D., Stephen, J.M., 2013. Primary visual response (M100) delays in adolescents with FASD as measured with MEG. Hum. Brain Mapp. 34 (11), 2852–2862. https://doi.org/10.1002/hbm.22110.

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

Connolly, J.F., Phillips, N.A., Stewart, S.H., Brake, W.G., 1992. Event-related potential sensitivity to acoustic and semantic properties of terminal words in sentences. Brain Lang. 43 (1), 1–18. https://doi.org/10.1016/0093-934x(92)90018-A.

Corley, M., MacGregor, L.J., Donaldson, D.I., 2007. It's the way that you, er, say it: hesitations in speech affect language comprehension. Cognition 105 (3), 658–668. https://doi.org/10.1016/j.cognition.2006.10.010.

Coulson, S., King, J.W., Kutas, M., 1998. Expect the unexpected: event-related brain response to morphosyntactic violations. Lang. Cogn. Process. 13 (1), 21–58. https://doi.org/10.1080/016909698386582.

Curran, T., Tucker, D.M., Kutas, M., Posner, M.I., 1993. Topography of the N400: brain electrical activity reflecting semantic expectancy. Electroencephalogr. Clin. Neurophysiol. 88 (3), 188–209.

Dahan, D., Magnuson, J.S., 2006. Spoken word recognition. Handbook of Psycholinguistics, second edition. pp. 249–283.

Dambacher, M., Rolfs, M., Gollner, K., Kliegl, R., Jacobs, A.M., 2009. Event-related potentials reveal rapid verification of predicted visual input. PLoS One 4 (3), e5047. https://doi.org/10.1371/journal.pone.0005047.

Dambacher, M., Dimigen, O., Braun, M., Wille, K., Jacobs, A.M., Kliegl, R., 2012. Stimulus onset asynchrony and the timeline of word recognition: event-related potentials during sentence reading. Neuropsychologia 50 (8), 1852–1870. https://doi.org/10.1016/j.neuropsychologia.2012.04.011.

Dave, S., Brothers, T.A., Swaab, T.Y., 2018. 1/f neural noise and electrophysiological indices of contextual prediction in aging. Brain Res. https://doi.org/10.1016/j.brainres.2018.04.007.

Dehaene, S., Cohen, L., Sigman, M., Vinckier, F., 2005. The neural code for written words: a proposal. Trends Cogn Sci 9 (7), 335–341. https://doi.org/10.1016/j.tics.2005.05.004.

Dell, G.S., Chang, F., 2014. The P-chain: Relating sentence production and its disorders to comprehension and acquisition. Phil. Trans. R. Soc. B 369 (1634), 20120394.

Den Ouden, H.E., Kok, P., De Lange, F.P., 2012. How prediction errors shape perception, attention, and motivation. Front. Psychol. 3, 548.

Diaz, M.T., Swaab, T.Y., 2007. Electrophysiological differentiation of phonological and semantic integration in word and sentence contexts. Brain Res. 1146, 85–100. https://doi.org/10.1016/j.brainres.2006.07.034.

Dien, J., 1998. Issues in the application of the average reference: review, critiques, and recommendations. Behav. Res. Methods Instrum. Comp. 30 (1), 34–43. https://doi.org/10.3758/Bf03209414.

Dikker, S., Pylkkanen, L., 2011. Before the N400: effects of lexical-semantic violations in visual cortex. Brain Lang. 118 (1-2), 23–28. https://doi.org/10.1016/j.bandl.2011.02.006.

Dikker, S., Rabagliati, H., Pylkkanen, L., 2009. Sensitivity to syntax in visual cortex. Cognition 110 (3), 293–321. https://doi.org/10.1016/j.cognition.2008.09.008.

Dikker, S., Rabagliati, H., Farmer, T.A., Pylkkanen, L., 2010. Early occipital sensitivity to syntactic category is based on form typicality. Psychol. Sci. 21 (5), 629–634. https://doi.org/10.1177/0956797610367751.

Ettinger, A., Linzen, T., Marantz, A., 2014. The role of morphology in phoneme prediction: evidence from MEG. Brain Lang. 129, 14–23. https://doi.org/10.1016/j.bandl.2013.11.004.

Farmer, T.A., Christiansen, M.H., Monaghan, P., 2006. Phonological typicality influences on-line sentence comprehension. Proc. Natl. Acad. Sci. 103 (32), 12203–12208.

Federmeier, K.D., Kutas, M., 1999. A rose by any other name: long-term memory structure and sentence processing. J. Mem. Lang. 41 (4), 469–495. https://doi.org/10.1006/jmla.1999.2660.

Federmeier, K.D., McLennan, D.B., De Ochoa, E., Kutas, M., 2002. The impact of semantic memory organization and sentence context information on spoken language processing by younger and older adults: an ERP study. Psychophysiology 39 (2), 133–146. https://doi.org/10.1111/1469-8986.3920133.

Fine, A.B., Jaeger, T.F., Farmer, T.A., Qian, T., 2013. Rapid expectation adaptation during syntactic comprehension. PLoS One 8 (10) doi: ARTN e7766110.1371/journal.-pone.0077661.

Folstein, J.R., Van Petten, C., 2008. Influence of cognitive control and mismatch on the N2 component of the ERP: a review. Psychophysiology 45 (1), 152–170. https://doi.org/10.1111/j.1469-8986.2007.00602.x.

Folstein, J.R., Van Petten, C., 2011. After the P3: late executive processes in stimulus categorization. Psychophysiology 48 (6), 825–841. https://doi.org/10.1111/j.1469-8986.2010.01146.x.

Forstmeier, W., Wagenmakers, E.J., Parker, T.H., 2017. Detecting and avoiding likely false-positive findings - a practical guide. Biol. Rev. 92 (4), 1941–1968. https://doi.org/10.1111/brv.12315.

Freunberger, D., Roehm, D., 2016. Semantic prediction in language comprehension: evidence from brain potentials. Lang. Cogn. Neurosci. 31 (9), 1193–1205. https://doi.org/10.1080/23273798.2016.1205202.

Friederici, A.D., 2002. Towards a neural basis of auditory sentence processing. Trends Cogn Sci 6 (2), 78–84.

Friederici, A.D., Pfeifer, E., Hahne, A., 1993. Event-related brain potentials during natural speech processing - effects of semantic, morphological and syntactic violations. Cogn. Brain Res. 1 (3), 183–192. https://doi.org/10.1016/0926-6410(93)90026-2.

Friston, K., 2005. A theory of cortical responses. Philos. Trans. R. Soc. Lond. B Biol. Sci. 360 (1456), 815–836. https://doi.org/10.1098/rstb.2005.1622.

Friston, K., 2008. Hierarchical models in the brain. PLoS Comput. Biol. 4 (11), e1000211.

Friston, K., 2010. The free-energy principle: a unified brain theory? Nat. Rev. Neurosci. 11 (2), 127–138. https://doi.org/10.1038/nrn2787.

Friston, K., Kiebel, S., 2009. Predictive coding under the free-energy principle. Philos. Trans. R. Soc. Lond. B Biol. Sci. 364 (1521), 1211–1221. https://doi.org/10.1098/rstb.2008.0300.

Fruchter, J., Linzen, T., Westerlund, M., Marantz, A., 2015. Lexical preactivation in basic linguistic phrases. J. Cogn. Neurosci. 27 (10), 1912–1935. https://doi.org/10.1162/jocn_a_00822.

Gagnepain, P., Henson, R.N., Davis, M.H., 2012. Temporal predictive codes for spoken words in auditory cortex. Curr. Biol. 22 (7), 615–621. https://doi.org/10.1016/j.cub.2012.02.015.

Gelman, A., Loken, E., 2014. The statistical crisis in science. Am. Sci. 102 (6), 460–465. https://doi.org/10.1511/2014.111.460.

Gelman, A., Stern, H., 2006. The difference between "significant" and "not significant" is not itself statistically significant. Am. Stat. 60 (4), 328–331. https://doi.org/10.1198/000313006X152649.

Grainger, J., Holcomb, P.J., 2009. An ERP investigation of orthographic priming with relative-position and absolute-position primes. Brain Res. 1270, 45–53. https://doi.org/10.1016/j.brainres.2009.02.080.

Grainger, J., Jacobs, A.M., 1996. Orthographic processing in visual word recognition: a multiple read-out model. Psychol. Rev. 103 (3), 518.

Grainger, J., Kiyonaga, K., Holcomb, P.J., 2006. The time course of orthographic and phonological code activation. Psychol. Sci. 17 (12), 1021–1026. https://doi.org/10.1111/j.1467-9280.2006.01821.x.

Groppe, D.M., Choi, M., Huang, T., Schilz, J., Topkins, B., Urbach, T.P., Kutas, M., 2010. The phonemic restoration effect reveals pre-N400 effect of supportive sentence context in speech perception. Brain Res. 1361, 54–66. https://doi.org/10.1016/j.brainres.2010.09.003.

Groppe, D.M., Urbach, T.P., Kutas, M., 2011. Mass univariate analysis of event-related brain potentials/fields I: a critical tutorial review. Psychophysiology 48 (12), 1711–1725. https://doi.org/10.1111/j.1469-8986.2011.01273.x.

Hagoort, P., 2017. The core and beyond in the language-ready brain. Neurosci. Biobehav. Rev. https://doi.org/10.1016/j.neubiorev.2017.01.048.

Hagoort, P., Brown, C.M., 2000. ERP effects of listening to speech compared to reading: the P600/SPS to syntactic violations in spoken sentences and rapid serial visual presentation. Neuropsychologia 38 (11), 1531–1549.

Hahne, A., Friederici, A.D., 1999. Electrophysiological evidence for two steps in syntactic analysis. Early automatic and late controlled processes. J. Cogn. Neurosci. 11 (2), 194–205.

Hauk, O., Patterson, K., Woollams, A., Watling, L., Pulvermüller, F., Rogers, T.T., 2006. [Q:] when would you prefer a SOSSAGE to a SAUSAGE? [A:] At about 100 msec. ERP correlates of orthographic typicality and lexicality in written word recognition. J. Cogn. Neurosci. 18 (5), 818–832. https://doi.org/10.1162/jocn.2006.18.5.818.

Hickok, G., 2012. The cortical organization of speech processing: feedback control and predictive coding the context of a dual-stream model. J. Commun. Disord. 45 (6), 393–402. https://doi.org/10.1016/j.jcomdis.2012.06.004.

Hillyard, S.A., Anllo-Vento, L., 1998. Event-related brain potentials in the study of visual selective attention. Proc. Natl. Acad. Sci. U.S.A. 95 (3), 781–787. https://doi.org/10.1073/pnas.95.3.781.

Hillyard, S.A., Teder-Salejarvi, W.A., Munte, T.F., 1998. Temporal dynamics of early perceptual processing. Curr. Opin. Neurobiol. 8 (2), 202–210. https://doi.org/10.1016/S0959-4388(98)80141-4.

Holcomb, P.J., Grainger, J., 2006. On the time course of visual word recognition: an event-related potential investigation using masked repetition priming. J. Cogn. Neurosci. 18 (10), 1631–1643. https://doi.org/10.1162/jocn.2006.18.10.1631.

Ito, A., Corley, M., Pickering, M.J., Martin, A.E., Nieuwland, M.S., 2016. Predicting form and meaning: evidence from brain potentials. J. Mem. Lang. 86, 157–171. https://doi.org/10.1016/j.jml.2015.10.007.

Ito, A., Martin, A.E., Nieuwland, M.S., 2017. How robust are prediction effects in language comprehension? Failure to replicate article-elicited N400 effects. Lang. Cogn. Neurosci. 32 (8), 954–965. https://doi.org/10.1080/23273798.2016.1242761.

Johannes, S., Munte, T.F., Heinze, H.J., Mangun, G.R., 1995. Luminance and spatial attention effects on early visual processing. Brain Res. Cogn. Brain Res. 2 (3), 189–205.

Junghofer, M., Elbert, T., Tucker, D.M., Braun, C., 1999. The polar average reference effect: a bias in estimating the head surface integral in EEG recording. Clin. Neurophysiol. 110 (6), 1149–1155.

Kaan, E., 2002. Investigating the effects of distance and number interference in processing subject-verb dependencies: an ERP study. J. Psycholinguist. Res. 31 (2), 165–193.

Kaan, E., Kirkham, J., Wijnen, F., 2016. Prediction and integration in native and second-language processing of elliptical structures. Bilingualism-Language and Cognition 19 (1), 1–18. https://doi.org/10.1017/S1366728914000844.

Kazanina, N., Bowers, J.S., Idsardi, W., 2018. Phonemes: lexical access and beyond. Psychon. Bull. Rev. 25 (2), 560–585. https://doi.org/10.3758/s13423-017-1362-0.

Kehler, A., Rohde, H., 2013. A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation. Theor. Linguist. 39 (1-2), 1–37. https://doi.org/10.1515/tl-2013-0001.

Kilner, J.M., Friston, K.J., Frith, C.D., 2007. Predictive coding: an account of the mirror neuron system. Cognit. Process. 8 (3), 159–166.

Kim, A.E., Gilley, P.M., 2013. Neural mechanisms of rapid sensitivity to syntactic anomaly. Front. Psychol. 4, 45. https://doi.org/10.3389/fpsyg.2013.00045.

Kim, A., Lai, V., 2012. Rapid interactions between lexical semantic and word form analysis during word recognition in context: evidence from ERPs. J. Cogn. Neurosci. 24 (5), 1104–1112. https://doi.org/10.1162/jocn_a_00148.

Kiyonaga, K., Grainger, J., Midgley, K., Holcomb, P.J., 2007. Masked cross-modal repetition priming: an event-related potential investigation. Lang. Cogn. Process. 22

(3), 337–376. https://doi.org/10.1080/01690960600652471.

Kok, P., de Lange, F.P., 2014. Shape perception simultaneously up- and downregulates neural activity in the primary visual cortex. Curr. Biol. 24 (13), 1531–1535. https://doi.org/10.1016/j.cub.2014.05.042.

Kok, P., Jehee, J.F., de Lange, F.P., 2012. Less is more: expectation sharpens representations in the primary visual cortex. Neuron 75 (2), 265–270. https://doi.org/10.1016/j.neuron.2012.04.034.

Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S.F., Baker, C.I., 2009. Circular analysis in systems neuroscience: the dangers of double dipping. Nat. Neurosci. 12 (5), 535–540. https://doi.org/10.1038/nn.2303.

Kuperberg, G.R., Jaeger, T.F., 2016. What do we mean by prediction in language comprehension? Lang. Cogn. Neurosci. 31 (1), 32–59. https://doi.org/10.1080/23273798.2015.1102299.

Kutas, M., Federmeier, K.D., 2000. Electrophysiology reveals semantic memory use in language comprehension. Trends Cognit. Sci. 4 (12), 463–470.

Kutas, M., Federmeier, K.D., 2011. Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). Annu. Rev. Psychol. 62, 621–647. https://doi.org/10.1146/annurev.psych.093008.131123.

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

Kutas, M., Hillyard, S.A., 1984. Brain potentials during reading reflect word expectancy and semantic association. Nature 307 (5947), 161–163.

Kutas, M., McCarthy, G., Donchin, E., 1977. Augmenting mental chronometry: the P300 as a measure of stimulus evaluation time. Science 197 (4305), 792–795.

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

Laszlo, S., Federmeier, K.D., 2009. A beautiful day in the neighborhood: an event-related potential study of lexical relationships and prediction in context. J. Mem. Lang. 61 (3), 326–338. https://doi.org/10.1016/j.jml.2009.06.004.

Lau, E., Stroud, C., Plesch, S., Phillips, C., 2006. The role of structural prediction in rapid syntactic analysis. Brain Lang. 98 (1), 74–88.

Lau, E., Phillips, C., Poeppel, D., 2008. A cortical network for semantics: (de)constructing the N400. Nat. Rev. Neurosci. 9 (12), 920–933. https://doi.org/10.1038/nrn2532.

Lau, E., Almeida, D., Hines, P.C., Poeppel, D., 2009. A lexical basis for N400 context effects: evidence from MEG. Brain Lang. 111 (3), 161–172. https://doi.org/10.1016/j.bandl.2009.08.007.

Lau, E., Namyst, A., Fogel, A., Delgado, T., 2016. A direct comparison of N400 effects of predictability and incongruity in adjective-noun combination. Collabra Psychol. 2 (1).

Lee, C.Y., Liu, Y.N., Tsai, J.L., 2012. The time course of contextual effects on visual word recognition. Front. Psychol. 3 doi: Artn 28510.3389/Fpsyg.2012.00285.

Leopold, D.A., Logothetis, N.K., 1998. Microsaccades differentially modulate neural activity in the striate and extrastriate visual cortex. Exp. Brain Res. 123 (3), 341–345. https://doi.org/10.1007/s002210050577.

Loken, E., Gelman, A., 2017. Measurement error and the replication crisis. Science 355 (6325), 584–585. https://doi.org/10.1126/science.aal3618.

Luck, S.J., 2014. Introduction to the event-related potential technique. Introduction to the Event-Related Potential Technique, 2nd edition. pp. 1–406.

Luck, S.J., Gaspelin, N., 2017. How to get statistically significant effects in any ERP experiment (and why you shouldn't). Psychophysiology 54 (1), 146–157. https://doi.org/10.1111/psyp.12639.

Luck, S.J., Hillyard, S.A., 1994. Electrophysiological correlates of feature analysis during visual-search. Psychophysiology 31 (3), 291–308. https://doi.org/10.1111/j.1469-8986.1994.tb02218.x.

Luke, S.G., Christianson, K., 2016. Limits on lexical prediction during reading. Cogn. Psychol. 88, 22–60. https://doi.org/10.1016/j.cogpsych.2016.06.002.

Lupyan, G., 2015. Cognitive penetrability of perception in the age of prediction: predictive systems are penetrable systems. Rev. Philos. Psychol. 6 (4), 547–569.

MacGregor, L.J., Corley, M., Donaldson, D.I., 2009. Not all disfluencies are are equal: the effects of disfluent repetitions on language comprehension. Brain Lang. 111 (1), 36–45. https://doi.org/10.1016/j.bandl.2009.07.003.

MacGregor, L.J., Corley, M., Donaldson, D.I., 2010. Listening to the sound of silence: disfluent silent pauses in speech have consequences for listeners. Neuropsychologia 48 (14), 3982–3992. https://doi.org/10.1016/j.neuropsychologia.2010.09.024.

MacGregor, L.J., Pulvermüller, F., van Casteren, M., Shtyrov, Y., 2012. Ultra-rapid access to words in the brain. Nat. Commun. 3, 711. https://doi.org/10.1038/ncomms1715.

Maess, B., Mamashli, F., Obleser, J., Helle, L., Friederici, A.D., 2016. Prediction signatures in the brain: semantic pre-activation during language comprehension. Front. Hum. Neurosci. 10 doi: Artn 59110.3389/Fnhum.2016.00591.

Mangun, G.R., Buonocore, M.H., Girelli, M., Jha, A.P., 1998. ERP and fMRI measures of visual spatial selective attention. Hum. Brain Mapp. 6 (5-6), 383–389.

Marslen-Wilson, W.D., 1987. Functional parallelism in spoken word-recognition. Cognition 25 (1-2), 71–102.

Marslen-Wilson, W., Tyler, L.K., 1980. The temporal structure of spoken language understanding. Cognition 8 (1), 1–71.

Martin, A.E., 2016. Language processing as cue integration: grounding the psychology of language in perception and neurophysiology. Front. Psychol. 7 doi: Artn 12010.3389/Fpsyg.2016.00120.

McCarthy, G., Donchin, E., 1981. A metric for thought: a comparison of P300 latency and reaction time. Science 211 (4477), 77–80.

McClelland, J.L., Rumelhart, D.E., 1981. An interactive activation model of context effects in letter perception .1. An account of basic findings. Psychol. Rev. 88 (5), 375–407. https://doi.org/10.1037/0033-295x.88.5.375.

McDowell, J.E., Dyckman, K.A., Austin, B.P., Clementz, B.A., 2008. Neurophysiology and neuroanatomy of reflexive and volitional saccades: evidence from studies of humans. Brain Cogn. 68 (3), 255–270. https://doi.org/10.1016/j.bandc.2008.08.016.

Meyberg, S., Werkle-Bergner, M., Sommer, W., Dimigen, O., 2015. Microsaccade-related brain potentials signal the focus of visuospatial attention. Neuroimage 104, 79–88. https://doi.org/10.1016/j.neuroimage.2014.09.065.

Molinaro, N., Conrad, M., Barber, H.A., Carreiras, M., 2010. On the functional nature of the N400: contrasting effects related to visual word recognition and contextual semantic integration. Cogn. Neurosci. 1 (1), 1–7. https://doi.org/10.1080/17588920903373952.

Monahan, P.J., 2018. Phonological knowledge and speech comprehension. Annu. Rev. Linguist. 4, 21–47.

Morton, J., 1979. Facilitation in word recognition: experiments causing change in the logogen model. Processing of Visible Language. Springer, Boston, MA, pp. 259–268.

Munafo, M.R., Nosek, B.A., Bishop, D.V.M., Button, K.S., Chambers, C.D., du Sert, N.P., et al., 2017. A manifesto for reproducible science. Nat. Hum. Behav. 1 (1) doi: Unsp 002110.1038/S41562-016-0021.

Naatanen, R., Paavilainen, P., Rinne, T., Alho, K., 2007. The mismatch negativity (MMN) in basic research of central auditory processing: a review. Clin. Neurophysiol. 118 (12), 2544–2590. https://doi.org/10.1016/j.clinph.2007.04.026.

Neely, J.H., 1977. Semantic priming and retrieval from lexical memory - roles of inhibition-less spreading activation and limited-capacity attention. J. Exp. Psychol. Gen. 106 (3), 226–254. https://doi.org/10.1037//0096-3445.106.3.226.

Neville, H., Nicol, J.L., Barss, A., Forster, K.I., Garrett, M.F., 1991. Syntactically based sentence processing classes - evidence from event-related brain potentials. J. Cogn. Neurosci. 3 (2), 151–165. https://doi.org/10.1162/jocn.1991.3.2.151.

Nieuwenhuis, S., Forstmann, B.U., Wagenmakers, E.J., 2011. Erroneous analyses of interactions in neuroscience: a problem of significance. Nat. Neurosci. 14 (9), 1105–1107. https://doi.org/10.1038/nn.2886.

Nieuwland, M.S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., et al., 2018a. Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. eLife 7, e33468.

Nieuwland, M.S., Barr, D.J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D.I., et al., 2018b. Dissociable effects of prediction and integration during language comprehension: evidence from a large-scale study using brain potentials. bioRxiv, 267815.

Norris, D., McQueen, J.M., Cutler, A., 2000. Merging information in speech recognition: feedback is never necessary. Behav. Brain Sci. 23 (3), 299–325 discussion 325-270.

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

Osterhout, L., Holcomb, P.J., Swinney, D.A., 1994. Brain potentials elicited by garden-path sentences: evidence of the application of verb information during parsing. J. Exp. Psychol. Learn. Mem. Cogn. 20 (4), 786–803.

Osterhout, L., McLaughlin, J., Kim, A., Greewald, R., Inoue, K., 2004. Sentences in the brain: event-related potentials as real-time reflections of sentence comprehension and language learning. In: Carreiras, M., Clifton, C. (Eds.), The On-Line Study of Sentence Comprehension: Eyetracking-ERPs, and Beyond. Psychology Press, New York, pp. 271–308.

Otten, Van Berkum, J., 2008. Discourse-Based Word Anticipation During Language Processing: Prediction or Priming? Discourse Process. 45 (6), 464–496. https://doi.org/10.1080/01638530802356463. doi: Pii 905987649.

Otten, M., Van Berkum, J.J., 2009. Does working memory capacity affect the ability to predict upcoming words in discourse? Brain Res. 1291, 92–101.

Otten, M., Nieuwland, M.S., van Berkum, J.J.A., 2007. Great expectations: specific lexical anticipation influences the processing of spoken language. BMC Neurosci. 8https://doi.org/10.1186/1471-2202-8-89. doi: Artn 89.

Patel, S.H., Azzam, P.N., 2005. Characterization of N200 and P300: selected studies of the event-related potential. Int. J. Med. Sci. 2 (4), 147–154.

Penolazzi, B., Hauk, O., Pulvermüller, F., 2007. Early semantic context integration and lexical access as revealed by event-related brain potentials. Biol. Psychol. 74 (3), 374–388. https://doi.org/10.1016/j.biopsycho.2006.09.008.

Pickering, M.J., Clark, A., 2014. Getting ahead: forward models and their place in cognitive architecture. Trends Cognit. Sci. 18 (9), 451–456.

Pickering, M.J., Gambi, C., 2018. Predicting while comprehending language: a theory and review. Article in press at. Psychol. Bull.

Pickering, M.J., Garrod, S., 2007. Do people use language production to make predictions during comprehension? Trends Cogn. Sci. 11 (3), 105–110. https://doi.org/10.1016/j.tics.2006.12.002.

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

Pitcher, D., Walsh, V., Duchaine, B., 2011. The role of the occipital face area in the cortical face perception network. Exp. Brain Res. 209 (4), 481–493. https://doi.org/10.1007/s00221-011-2579-1.

Polich, J., Kok, A., 1995. Cognitive and biological determinants of P300: an integrative review. Biol. Psychol. 41 (2), 103–146.

Proverbio, A.M., Adorni, R., 2009. C1 and P1 visual responses to words are enhanced by attention to orthographic vs. Lexical properties. Neurosci. Lett. 463 (3), 228–233. https://doi.org/10.1016/j.neulet.2009.08.001.

Pylkkanen, L., Marantz, A., 2003. Tracking the time course of word recognition with MEG. Trends Cogn. Sci. 7 (5), 187–189. https://doi.org/10.1016/S1364-6613(03)00092-5.

Radach, R., Inhoff, A., Heller, D., 2004. Orthographic regularity gradually modulates saccade amplitudes in reading. Eur. J. Cogn. Psychol. 16 (1-2), 27–51. https://doi.org/10.1080/09541440340000222.

Rao, R.P., Ballard, D.H., 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci. 2 (1), 79–87. https://doi.org/10.1038/4580.

Rastle, K., 2007. Visual word recognition (2012) In: Gaskell, M.G. (Ed.), The Oxford Handbook of Psycholinguistics. Oxford University Press, Oxford, pp. 71–87.

Rauss, K., Schwartz, S., Pourtois, G., 2011. Top-down effects on early visual processing in humans: a predictive coding framework. Neurosci. Biobehav. Rev. 35 (5), 1237–1253.

Rayner, K., 1998. Eye movements in reading and information processing: 20 years of research. Psychol. Bull. 124 (3), 372–422.

Roehm, D., Bornkessel-Schlesewsky, I., Rösler, F., Schlesewsky, M., 2007. To predict or not to predict: influences of task and strategy on the processing of semantic relations. J. Cognit. Neurosci. 19 (8), 1259–1274.

Roelofs, A., Meyer, A.S., Levelt, W.J., 1998. A case for the lemma/lexeme distinction in models of speaking: comment on Caramazza and Miozzo (1997). Cognition 69 (2), 219–230.

Rommers, J., Meyer, A.S., Praamstra, P., Huettig, F., 2013. The contents of predictions in sentence comprehension: activation of the shape of objects before they are referred to. Neuropsychologia 51 (3), 437–447. https://doi.org/10.1016/j.neuropsychologia.2012.12.002.

Rousselet, G.A., 2012. Does filtering preclude us from studying ERP time-courses? Front. Psychol. 3https://doi.org/10.3389/Fpsyg.2012.00131. doi: Artn 131.

Rousselet, G.A., 2016. How to Chase ERP Monsters Hiding Behind Bars. Retrieved from. https://garstats.wordpress.com/2016/06/03/erp-monsters.

Rousselet, G.A., Pernet, C.R., 2011. Quantifying the time course of visual object processing using ERPs: it's time to up the game. Front. Psychol. 2https://doi.org/10.3389/Fpsyg.2011.0010. doi: Artn 107.

Rousselet, G.A., Foxe, J.J., Bolam, J.P., 2016. A few simple steps to improve the description of group results in neuroscience. Eur. J. Neurosci. 44 (9), 2647–2651. https://doi.org/10.1111/ejn.13400.

Salmelin, R., 2007. Clinical neurophysiology of language: the MEG approach. Clin. Neurophysiol. 118 (2), 237–254. https://doi.org/10.1016/j.clinph.2006.07.316.

Shtyrov, Y., Pulvermüller, F., 2007. Early MEG activation dynamics in the left temporal and inferior frontal cortex reflect semantic context integration. J. Cogn. Neurosci. 19 (10), 1633–1642.

Siegelman, M., Mineroff, Z., Blank, I., Fedorenko, E., 2017. An attempt to replicate a dissociation between syntax and semantics during sentence comprehension reported by Dapretto & Bookheimer (1999, Neuron). bioRxiv 110791.

Sohoglu, E., Peelle, J.E., Carlyon, R.P., Davis, M.H., 2012. Predictive top-down integration of prior knowledge during speech perception. J. Neurosci. 32 (25), 8443–8453. https://doi.org/10.1523/Jneurosci.5069-11.2012.

Solomyak, O., Marantz, A., 2009. Lexical access in early stages of visual word processing: a single-trial correlational MEG study of heteronym recognition. Brain Lang. 108 (3), 191–196. https://doi.org/10.1016/j.bandl.2008.09.004.

Stack, C.M.H., James, A.N., Watson, D.G., 2018. A failure to replicate rapid syntactic adaptation in comprehension. Mem. Cognit. 1–14.

Staub, A., 2015. The effect of lexical predictability on eye movements in reading: critical review and theoretical interpretation. Lang. Linguist. Compass 9 (8), 311–327. https://doi.org/10.1111/lnc3.12151.

Staub, A., Clifton, C., 2006. Syntactic prediction in language comprehension: evidence from either … or. Journal of Exp. Psychol. Learn. Mem. Cogn. 32 (2), 425–436. https://doi.org/10.1037/0278-7393.32.2.425.

Staub, A., Grant, M., Clifton Jr, C., Rayner, K., 2009. Phonological typicality does not influence fixation durations in normal reading. J. Exp. Psychol. Learn. Mem. Cogn. 35 (3), 806.

Staub, A., Grant, M., Clifton, C., Rayner, K., 2011. Still no phonological typicality effect on word reading time (and no good explanation of one, either): a rejoinder to Farmer, Monaghan, Misyak, and Christiansen. J. Exp. Psychol. Learn. Mem. Cogn. 37 (5), 1326–1328.

Steinhauer, K., Drury, J.E., 2012. On the early left-anterior negativity (ELAN) in syntax studies. Brain Lang. 120 (2), 135–162. https://doi.org/10.1016/j.bandl.2011.07.001.

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

Summerfield, C., Egner, T., Greene, M., Koechlin, E., Mangels, J., Hirsch, J., 2006. Predictive codes for forthcoming perception in the frontal cortex. Science 314 (5803), 1311–1314. https://doi.org/10.1126/science.1132028.

Tanner, D., 2015. On the left anterior negativity (LAN) in electrophysiological studies of morphosyntactic agreement: a commentary on "grammatical agreement processing in reading: ERP findings and future directions" by Molinaro et al., 2014. Cortex 66, 149–155. https://doi.org/10.1016/j.cortex.2014.04.007.

Tanner, D., McLaughlin, J., Herschensohn, J., Osterhout, L., 2013. Individual differences reveal stages of L2 grammatical acquisition: ERP evidence. Bilingualism: Lang. Cognit. 16 (2), 367–382.

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

Tanner, D., Grey, S., van Hell, J.G., 2017. Dissociating retrieval interference and reanalysis in the P600 during sentence comprehension. Psychophysiology 54 (2), 248–259.

Tarkiainen, A., Helenius, P., Hansen, P.C., Cornelissen, P.L., Salmelin, R., 1999. Dynamics of letter string perception in the human occipitotemporal cortex. Brain 122, 2119–2131. https://doi.org/10.1093/brain/122.11.2119.

Tarkiainen, A., Cornelissen, P.L., Salmelin, R., 2002. Dynamics of visual feature analysis and ob object-level processing in face versus letter-string perception. Brain 125, 1125–1136. https://doi.org/10.1093/Brain/Awf112.

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

van Berkum, J.J.A., Zwitserlood, P., Hagoort, P., Brown, C.M., 2003. When and how do listeners relate a sentence to the wider discourse? Evidence from the N400 effect. Cogn. Brain Res. 17 (3), 701–718. https://doi.org/10.1016/S0926-6410(03)00196-4.

Van Berkum, J.J.A., 2009. The neuropragmatics of 'simple' utterance comprehension: an ERP review. Semantics and Pragmatics: From Experiment to Theory. Palgrave Macmillan, pp. 276–316.

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

Van Berkum, J.J.A., van den Brink, D., Tesink, C.M.J.Y., Kos, M., Hagoort, P., 2008. The neural integration of speaker and message. J. Cogn. Neurosci. 20 (4), 580–591. https://doi.org/10.1162/jocn.2008.20054.

van den Brink, D., Hagoort, P., 2004. The influence of semantic and syntactic context constraints on lexical selection and integration in spoken-word comprehension as revealed by ERPs. J. Cogn. Neurosci. 16 (6), 1068–1084. https://doi.org/10.1162/0898929041502670.

van den Brink, D., Brown, C.M., Hagoort, P., 2001. Electrophysiological evidence for early contextual influences during spoken-word recognition: N200 versus N400 effects. J. Cogn. Neurosci. 13 (7), 967–985. https://doi.org/10.1162/089892901753165872.

van den Brink, D., Brown, C.M., Hagoort, P., 2006. The cascaded nature of lexical selection and integration in auditory sentence processing. J. Exp. Psychol. Learn. Mem. Cogn. 32 (2), 364–372. https://doi.org/10.1037/0278-7393.32.3.364.

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

Van Petten, C., Coulson, S., Rubin, S., Plante, E., Parks, M., 1999. Time course of word identification and semantic integration in spoken language. J. Exp. Psychol. Learn. Mem. Cogn. 25 (2), 394–417. https://doi.org/10.1037//0278-7393.25.2.394.

Vasishth, S., Mertzen, D., Jäger, L.A., Gelman, A., 2018. The statistical significance filter leads to overoptimistic expectations of replicability. J. Mem. Lang. 103, 151–175.

Vega-Mendoza, M., Pickering, M.J., Nieuwland, M.S., 2017. Concurrent Use of Animacy and Event-knowledge During Comprehension: Evidence From Event-related Potentials. PsyArxiv.

Vespignani, F., Canal, P., Molinaro, N., Fonda, S., Cacciari, C., 2010. Predictive mechanisms in idiom comprehension. J. Cogn. Neurosci. 22 (8), 1682–1700. https://doi.org/10.1162/jocn.2009.21293.

Vissers, C.T.W.M., Chwilla, D.J., Kolk, H.H.J., 2006. Monitoring in language perception: the effect of misspellings of words in highly constrained sentences. Brain Res. 1106, 150–163. https://doi.org/10.1016/j.brainres.2006.05.012.

Vul, E., Pashler, H., 2012. Voodoo and circularity errors. Neuroimage 62 (2), 945–948. https://doi.org/10.1016/j.neuroimage.2012.01.027.

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

Wicherts, J.M., Bakker, M., Molenaar, D., 2011. Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. PLoS One 6 (11). https://doi.org/10.1371/journal.pone.0026828. ARTN e26828.

Willems, R.M. (Ed.), 2015. Cognitive Neuroscience of Natural Language Use. Cambridge University Press.

Wittenberg, B., 2012. The Effects of Context and Presentation Rate on the First 200 msec of Word Recognition. Undergraduate Honors Theses. pp. 1272. https://scholar.colorado.edu/honr_theses/1272.

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. https://doi.org/10.1016/j.cortex.2015.03.014.

Zwaan, R.A., Etz, A., Lucas, R.E., Donnellan, M.B., 2018. Making replication mainstream. Behav. Brain Sci. 41.

Zwitserlood, P., 2004. Sublexical and morphological information in speech processing. Brain Lang. 90 (1-3), 368–377. https://doi.org/10.1016/S0093-934X(03)00448-6.