

When components collide: Spatiotemporal overlap of the N400 and P600 in language comprehension

Francesca Delogu ^{a,*}, Harm Brouwer ^a, Matthew W. Crocker ^a

^a Department of Language Science and Technology, Saarland University, Building C7.1, 66123 Saarbrücken, Germany

ARTICLE INFO

Keywords:
Event-Related brain Potentials (ERPs)
N400
P600
Spatiotemporal component overlap
Language comprehension

ABSTRACT

The problem of spatiotemporal overlap between event-related potential (ERP) components is generally acknowledged in language research. However, its implications for the interpretation of experimental results are often overlooked. In a previous experiment on the functional interpretation of the N400 and P600, it was argued that a P600 effect to implausible words was largely obscured – in one of two implausible conditions – by an overlapping N400 effect of semantic association. In the present ERP study, we show that the P600 effect of implausibility is uncovered when the critical condition is tested against a proper baseline condition which elicits a similar N400 amplitude, while it is obscured when tested against a baseline condition producing an N400 effect. Our findings reveal that component overlap can result in the apparent absence or presence of an effect in the surface signal and should therefore be carefully considered when interpreting ERP patterns. Importantly, we show that, by factoring in the effects of spatiotemporal overlap between the N400 and P600 on the surface signal, which we reveal using rERP analysis, apparent inconsistencies in previous findings are easily reconciled, enabling us to draw unambiguous conclusions about the functional interpretation of the N400 and P600 components. Overall, our results provide compelling evidence that the N400 reflects lexical retrieval processes, while the P600 indexes compositional integration of word meaning into the unfolding utterance interpretation.

1. Introduction

In electrophysiological research, the Event-Related Potential (ERP) method is widely used to study the cognitive processes underlying online language comprehension. The advantage of this technique over other methods lies in its high temporal resolution and the multidimensional nature of the signal, which offers insights into not just quantitative, but also qualitative differences in the underlying cognitive processes. Yet, there are some fundamental and methodological issues that – despite being long recognized in the field – remain unsolved.

First, there remain important differences regarding the functional interpretation of the two most salient language-related ERP components, the N400 and the P600 (see, for a discussion, [Delogu et al., 2019](#)). The N400 has been linked to lexical access/retrieval processes ([Brouwer et al., 2012](#); [Kutas and Federmeier, 2000](#); [Kutas and Federmeier, 2011](#); [Lau et al., 2009](#); [Lau et al., 2008](#); [van Berkum et al., 2009](#); [van Berkum, 2010](#)), semantic integration ([Brown and Hagoort, 1993](#); [Brown and Hagoort, 1999](#); [Hagoort et al., 2004](#)), or both ([Baggio and Hagoort, 2011](#); [Lau et al., 2016](#); [Nieuwland et al., 2019](#)). The P600, on the other

hand, has been initially associated with syntactic re-analysis/repair (e.g., [Osterhout and Holcomb, 1992](#); [Friederici, 1995](#); [Hagoort et al., 1999](#)), syntactic integration difficulty (e.g., [Kaan et al., 2000](#); [Kaan and Swaab, 2000](#)), and then with more general integration processes ([Brouwer et al., 2012](#); [Brouwer et al., 2017](#); [Burkhardt, 2006](#); [Burkhardt, 2007](#); [Regel et al., 2011](#); [Delogu et al., 2018](#); [Hoeks et al., 2013](#)). Clearly, in order to fully exploit ERPs in the study of language understanding, it is essential to clarify what cognitive processes are indexed by these two components. Very different – if not inconsistent – conclusions can be drawn about the processes underlying an individual effect depending on which interpretation of the relevant component is assumed.

An important related issue, which has long been acknowledged but is rarely explicitly addressed in the psycholinguistic literature, is the spatiotemporal overlap between ERP components ([Luck, 2005](#); [Donchin et al., 1978](#); [Näätänen and Picton, 1987](#)). The scalp-recorded ERP waveform reflects the contribution of simultaneously present latent ERP components, and when these have opposite polarity, such as the N400 and the P600, they cancel each other out in the surface signal. This may result in apparent absence (or presence) of an effect in one component,

* Corresponding author.

E-mail address: delogu@coli.uni-saarland.de (F. Delogu).

potentially leading to misinterpreting the nature of the underlying cognitive processes. Indeed, Brouwer and Crocker (2017) argue that component overlap may account for several inconsistencies in the data reported in the literature, both within (see, for example, Kim and Osterhout, 2005; Kolk et al., 2003; Kuperberg et al., 2007; Delogu et al., 2019) as well as across studies (see, for reviews, Bornkessel-Schlesewsky and Schlesewsky, 2008; Brouwer et al., 2012; van Petten and Luka, 2012; Kuperberg, 2007).

These two issues – the functional interpretation of ERP components and component overlap – are closely related to each other and should be addressed simultaneously: Determining if and how the N400 and P600 interact in the ERP signal is in fact critical to elucidating their functional interpretation, to the extent that component overlap affects the empirical evidence on which the various theoretical accounts are based. For example, the Retrieval-Integration (RI) account (Brouwer et al., 2012; Brouwer et al., 2017) links the N400 to lexical retrieval processes. More specifically, the amplitude of the N400 is argued to reflect the effort involved in retrieving from long-term memory the conceptual knowledge associated with the eliciting word, which is sensitive to the degree to which this knowledge is cued by the context (see also, e.g., Kutas and Federmeier, 2000; Lau et al., 2008). The P600, on the other hand, is assumed to be a family of late positivities reflecting the different subprocesses underlying the integration of word meaning into the unfolding utterance interpretation (e.g., Brouwer et al., 2012; Burkhardt, 2006; Burkhardt, 2007). The RI account predicts biphasic N400/P600 effects whenever both lexical retrieval and semantic integration are experimentally manipulated. However, as argued by the authors, retrieval and integration processes likely overlap in time, resulting in a corresponding overlap between the latent components (Brouwer and Crocker, 2017; Brouwer and Hoeks, 2013). As a consequence, the scalp-recorded ERP waveforms may reveal monophasic rather than biphasic N400/P600 patterns, due to the effect of component overlap on the morphology of the surface signal.

Quantifying the independent contribution of the latent components from the surface signal is extremely challenging (as noted by Kappenman et al., 2012; Luck, 2005; Fabiani et al., 2012; Brouwer and Crocker, 2017, among others). Understandably, this may be one of the reasons why researchers in the psycholinguistic community are reluctant to consider component overlap in the interpretation of their data (but see, Hagoort, 2003; Tanner and van Hell, 2014). However, one way to approach the problem, which we consider in this paper, is to act at the experimental design level, by incorporating a means of isolating the component of interest by controlling for component overlap. Traditional ERP analysis techniques have been recently extended with the regression-based ERP (rERP) framework proposed by Smith and Kutas (2015a, b). This analysis technique allows for decomposing the scalp-recorded waveform into the underlying latent components, by isolating the relative contribution made by each experimentally-manipulated factor (for a step-by-step derivation of an rERP analysis, see, Brouwer et al., 2021a).

In the present ERP study, we employ both approaches to examine an apparent inconsistency in the findings reported by Delogu et al. (2019), henceforth DBC, which they discuss as resulting from component overlap. DBC investigated the functional interpretation of the N400 and P600 with a single design manipulating semantic association (tapping into lexical retrieval processes) and plausibility (tapping into integrative processes). They measured ERPs to target words (e.g., *menu*) in mini-discourses such as (1):

(1) a. *Baseline*

John entered the restaurant. Before long, he opened the menu and...

b. *Event related violation condition*

John left the restaurant. Before long, he opened the menu and...

c. *Event unrelated violation condition*

John entered the apartment. Before long, he opened the menu and...

In the baseline condition (1a), the target word was semantically related to the context (via the word *restaurant*) and plausible according to world knowledge. In (1b), the target was still related to the context – resulting in similar lexical retrieval effort as in the baseline – but implausible, and therefore more difficult to integrate into the preceding context. Finally, in (1c), the target was both unrelated and implausible, leading to more difficult retrieval and integration processes.

DBC found clear evidence that the N400 indexes retrieval and not integration processes, as evidenced by the absence of an N400 effect for the event related violation condition (1b) relative to the baseline (1a), while a large and broadly distributed N400 effect was elicited by the event unrelated violation condition (1c). That is, N400 effects patterned with association manipulation and not plausibility. The P600, on the other hand, was sensitive to plausibility, as evidenced by the P600 effect elicited by the targets in the event related violation condition compared to the baseline. Somewhat unexpectedly, however, no P600 effect was observed for event unrelated violating targets, but rather a sustained negativity in fronto-central electrodes becoming gradually more positive over posterior sites, and producing a significant effect only over occipital electrodes. As noted by the authors, no existing account of the P600 predicts the observed pattern of effects. Syntactic accounts would predict absence of a P600 effect in both violating conditions, since both are syntactically well-formed and unambiguous. Integration accounts would predict presence of a P600 effect in both violating conditions, as both conditions describe an implausible situation. As DBC pointed out, multi-stream models (e.g., Kim and Osterhout, 2005; Kuperberg, 2007; van Herten et al., 2006; see Brouwer et al., 2012 for an overview) are also unable to explain the results. The observed pattern of effects, therefore, appears to be internally inconsistent: Is the P600 sensitive to plausibility? If not, what drives the P600 effect in the event related violation condition?

DBC argue that this apparent inconsistency can be reconciled once the effects of component overlap are taken into account. They point out that the positive-going deflection of the waveforms over posterior sites observed for the event unrelated condition likely results from an overlap between the broadly distributed negativity starting in the N400 time window and a positivity more pronounced over posterior electrodes. The negativity might have attenuated the overlapping positivity, thereby masking the expected P600 effect of plausibility. Under this interpretation, the results of the study provide unambiguous support for the RI account, according to which the N400 reflects lexical retrieval processes and the P600 indexes integration.

If the component overlap interpretation is correct, it makes a unique prediction: a P600 effect of plausibility to *unrelated* targets should be visible when not preceded by an N400 effect, that is, relative to a plausible and *unrelated* baseline condition, against which no N400 effect of semantic association is expected. By contrast, no P600 effect should be observed relative to a plausible and *related* baseline, precisely because this contrast should result in an N400 effect of association overlapping – and therefore masking – the P600 effect of plausibility (as in DBC's experiment).

We tested this hypothesis in an ERP study using a design similar to the one employed by DBC (i.e., manipulating association and plausibility), but in which the three conditions were modified as in (2):

(2) a. *Related-Plausible Baseline*

John went out in the rain. Before long, he opened the umbrella and...

b. *Unrelated-Plausible Baseline*

John left the restaurant. Before long, he opened the umbrella and...

c. Unrelated-Implausible condition

John entered the restaurant. Before long, he opened the umbrella and...

The target word *umbrella* is semantically related to the context in (2a) – via semantic association with the word *rain* – and unrelated in (2b) and (2c). We should therefore observe a similar N400 effect of association for both (2b) and (2c) relative to (2a). Crucially, (2b) and (2c) differ in plausibility (opening an umbrella is more plausible after leaving than after entering a building). Thus, (2c) should elicit a P600 effect of plausibility relative to its proper baseline (2b), but not relative to (2a), due to overlap with the N400 effect elicited by this contrast. Relative to (2a), we might rather observe a sustained negativity in frontal electrodes, as reported in DBC's study. To elucidate this prediction, we not only provide the canonical analysis of ERPs by considering mean amplitudes in the N400 and P600 time windows, but also present an rERP analysis. This provides a clear window into examining how the manipulated factors of association and plausibility, as well as the cloze probability of the target word – which has been shown to be inversely correlated with N400 amplitude (see Kutas and Federmeier, 2011) – reflect modulations in the underlying generators of the N400 and P600 components, to tease apart their contribution to the scalp-recorded activity.

Overall, the predicted pattern of effects would provide further support for the RI account regarding the functional interpretation of the N400 and P600. More importantly, it would also directly confirm the prediction that biphasic N400/P600 effects of retrieval and integration processes may be obscured in the surface signal by the effects of component overlap in the latent structure.

2. Results

2.1. Plausibility judgments

In the ERP stimuli, several words were added after the critical target word (which was the final word in the offline plausibility rating study, see Section 4.2) to avoid wrap-up effects. As the plausibility judgments provided in the ERP experiment (see Section 4.3) by the first 12 participants revealed that 25 out of 90 items did not show the intended plausibility ratings for the Unrelated-Plausible condition (percentage rated as plausible: Related-Plausible: 94%, Unrelated-Plausible: 70%; Unrelated-Implausible: 13%), we revised the post target region in these items. Notice that these small changes could not affect the ERPs elicited up to and including the critical word, as this was never altered. On average, participants judged the conditions to be plausible as follows: Related-Plausible, 95% (SD = 21); Unrelated-Plausible, 75% (SD = 44);

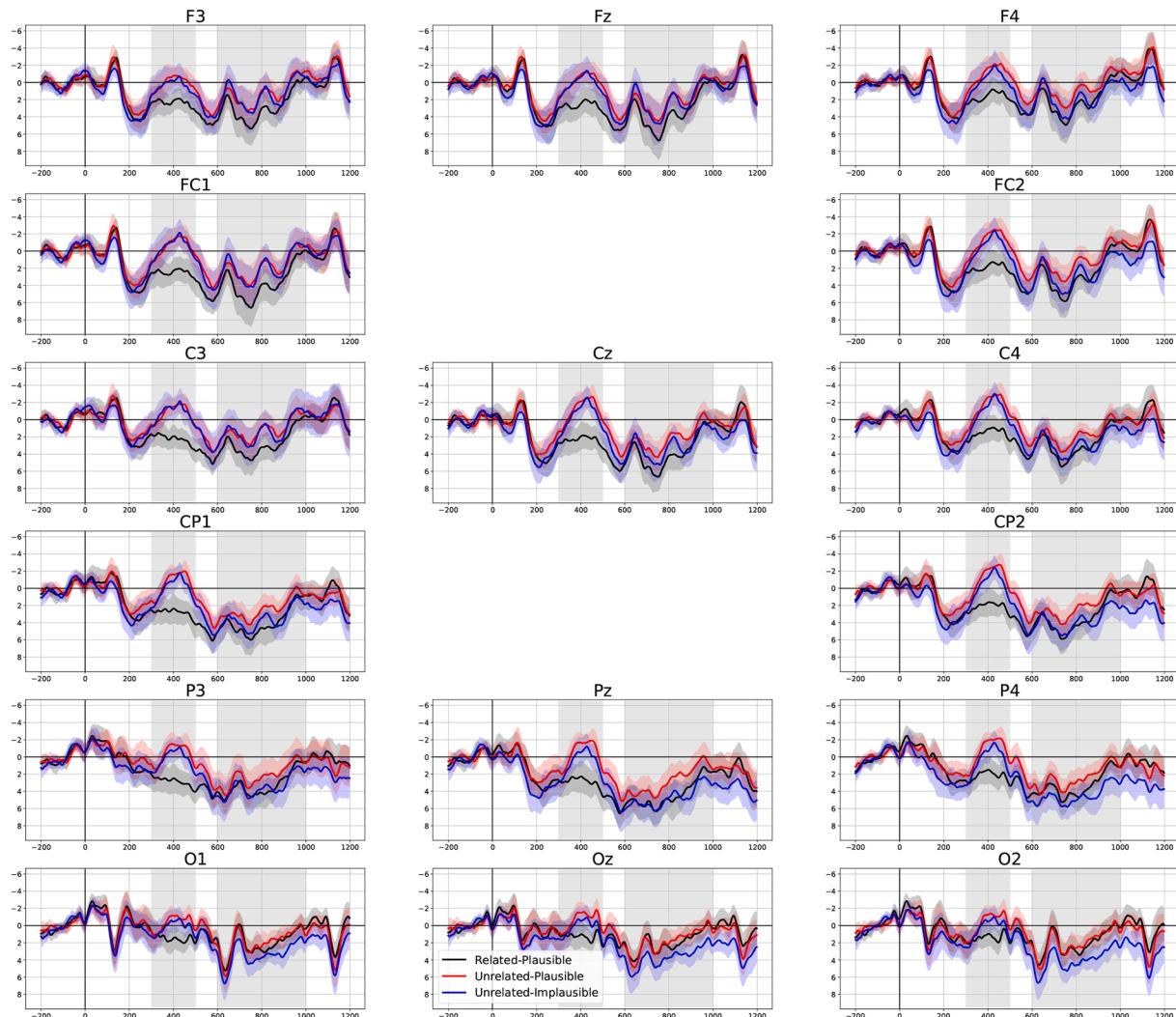


Fig. 1. Grand-average ERP waveforms time-locked to the target nouns in the Related-Plausible condition (black line), Unrelated-Plausible condition (red line), and Unrelated-Implausible condition (blue line), for a subset of electrodes. Shaded regions around the waveforms show mean voltage ± 2 SE. Negative voltages are plotted upwards.

Unrelated-Implausible, 13% ($SD = 34$). Pairwise comparisons with Bonferroni correction revealed a significant difference between all conditions. These results qualitatively mirror the plausibility ratings on a 1–7 Likert scale observed in the offline norming study. Plausibility judgements for the group of 9 participants who saw the revised version of the items were: Related-Plausible: 96%, Unrelated-Plausible: 80%; Unrelated-Implausible: 15%. No trials were rejected based on the behavioral results.

2.2. ERP analysis

Grand-average ERP waveforms to the target nouns in all three conditions are shown in Fig. 1. Table 1 reports the ANOVAs testing all pairwise comparisons in both the N400 and P600 time windows. Fig. 2 and Fig. 3 show the topographic maps for all pairwise comparisons in the N400 and the P600 time window, respectively. We note that the ERPs in the centro-posterior electrodes show an increased positivity around 200 ms post stimulus onset in response to implausible targets. A similar pattern can be observed also in DBC waveforms, although more visible on central electrodes (see Delogu et al., 2019, Fig. 1). To further investigate this effect, we performed ANOVAs in the time window between 100–300 ms and report statistically significant results. Based on visual inspection of the ERP waveforms, this effect could be attributed either to an early onset of the N400 for Unrelated-Implausible targets, which is slightly less negative than the N400 to Unrelated-Plausible

Table 1
ANOVAs on ERPs to target nouns across the N400 time window (300–500 ms) and the P600 time window (600–1000 ms).

		300–500 ms			600–1000 ms			
		df	F	p	$\eta^2 G$	F	p	$\eta^2 G$
<i>Unrelated-Plausible vs. Related-Plausible</i>								
Midline	Cond	(1, 20)	21.7	< .001	.23	7.48	.01	.06
	Cond × AP	(2, 40)	1.50	.24	.002	< 1	.90	< .001
ROIs	Cond	(1, 20)	24.0	< .001	.25	5.31	.03	.05
	Cond × AP	(1, 20)	1.03	.32	.001	2.49	.13	.002
	Cond × H	(1, 20)	< 1	.52	< .001	< 1	.58	< .001
<i>Unrelated-Implausible vs. Related-Plausible</i>								
Midline	Cond	(1, 20)	22.7	< .001	.14	< 1	.43	.007
	Cond × AP	(2, 40)	1.09	.33	< .01	3.58	.05	.01
ROIs	Cond	(1, 20)	25.6	< .001	.15	< 1	.62	.002
	Cond × AP	(1, 20)	< 1	.97	< .001	12.8	< .01	
	Cond × H	(1, 20)	3.15	.09	.001	16.2	< .001	.008
<i>Unrelated-Implausible vs. Unrelated-Plausible</i>								
Midline	Cond	(1, 20)	< 1	.33	.01	4.50	< .05	.02
	Cond × AP	(2, 40)	1.37	.26	< .01	7.07	< .01	
ROIs	Cond	(1, 20)	1.08	.31	.01	3.96	.06	.03
	Cond × AP	(1, 20)	< 1	.37	< .001	7.37	.01	.005
	Cond × H	(1, 20)	8.14	< .01	.002	14.31	< .01	.007

Notes. Cond × AP = Condition × Anterior–Posterior distribution; Cond × H = Condition × Hemisphere.

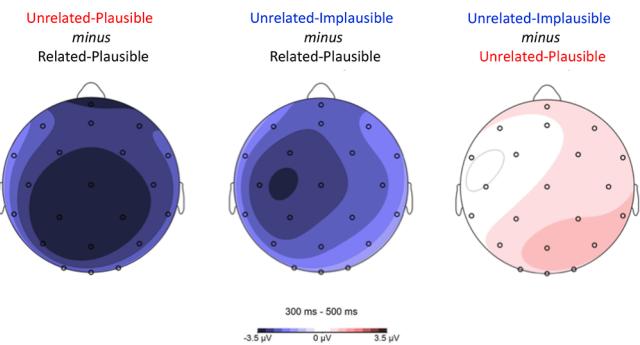


Fig. 2. Topographic maps of the simple effects in the N400 time window (300–500 ms).

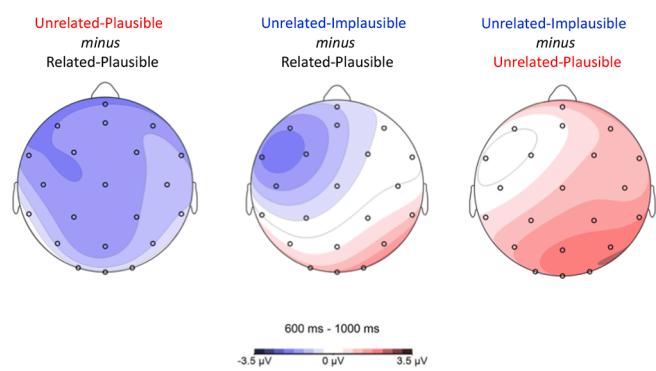


Fig. 3. Topographic maps of the simple effects in the P600 time window (600–1000 ms).

targets, or to an increased P200 in the implausible condition. We further examine this issue in the rERP analysis presented in Section 4.2.

2.3. P200 time window (100–300 ms)

In the midline electrodes there was a significant effect of Condition, $F(2, 40) = 3.87$, $p = .03$, $\eta^2 G = .03$, and no interaction with AP distribution, $F(4, 80) = 1.23$, $p = .31$, $\eta^2 G = .001$. Pairwise comparisons revealed a significant difference only between the Unrelated-Implausible ($M = 2.85 \mu V$, $SD = 3.02$) and the Unrelated-Plausible conditions ($M = 1.64 \mu V$, $SD = 2.66$), $F(1, 20) = 7.28$, $p = .014$, $\eta^2 G = .037$. The analysis over lateral electrodes revealed a significant effect of Condition, $F(2, 40) = 4.16$, $p = .03$, $\eta^2 G = .03$, and an interaction with Hemisphere, $F(2, 40) = 6.03$, $p = .006$, $\eta^2 G = .002$. Pairwise comparisons revealed that the Unrelated-Implausible condition ($M = 1.90 \mu V$, $SD = 2.49$) was more positive than the Related-Plausible condition ($M = 1.22 \mu V$, $SD = 1.85$), $F(1, 20) = 4.48$, $p = .047$, $\eta^2 G = .02$. The effect was more pronounced over right ($M_{Difference} = .97$) than left electrodes ($M_{Difference} = .39$), as revealed by a significant Condition × Hemisphere interaction, $F(1, 20) = 9.43$, $p = .006$, $\eta^2 G = .003$. The same pattern was observed in the comparison between the Unrelated-Implausible and the Unrelated-Plausible conditions ($M = 0.86 \mu V$, $SD = 2.18$). There was an effect of Condition, $F(1, 20) = 6.50$, $p = .02$, $\eta^2 G = .04$, and a Condition × Hemisphere interaction, $F(1, 20) = 5.76$, $p = .03$, $\eta^2 G = .002$, such that the difference was larger over right ($M_{Difference} = 1.26$) than left electrodes ($M_{Difference} = .83$). In sum, the analyses revealed a larger positivity for the Unrelated-Implausible condition, more pronounced over the right hemisphere. While increased P200s have been observed in response to irony (Regel et al., 2011) and expected or unexpected words completing strongly constraining sentence contexts (Wlotko and Federmeier, 2007), the functional interpretation of this component in sentential and discourse contexts is still poorly understood. We further

discuss these results in Section 3.

2.3.1. N400 time window (300–500 ms)

The ANOVA on midline electrodes revealed a significant effect of Condition, $F(2, 40) = 14.95$, $p < .0001$, $\eta^2G = .17$, and no interaction with AP distribution, $F(4, 80) = 1.32$, $p = .28$, $\eta^2G = .003$. Pairwise comparisons (see Table 1 and Fig. 2) showed a larger negativity for both the Unrelated-Plausible ($M = -0.59 \mu\text{V}$, $SD = 2.44$) and the Unrelated-Implausible conditions ($M = 0.03 \mu\text{V}$, $SD = 2.98$) compared to the Related-Plausible condition ($M = 2.60 \mu\text{V}$, $SD = 2.9$). The two unrelated conditions did not differ from each other.

The ANOVA over lateral sites revealed an effect of Condition, $F(2, 40) = 16.001$, $p < .0001$, $\eta^2G = .18$, and an interaction with Hemisphere, $F(2, 40) = 3.59$, $p < .04$, $\eta^2G = .002$, but not with AP distribution ($Fs < 1$). Pairwise comparisons showed again a larger negativity for both the Unrelated-Plausible ($M = -0.70 \mu\text{V}$, $SD = 2.09$) and the Unrelated-Implausible conditions ($M = -0.14 \mu\text{V}$, $SD = 2.48$) compared to the Related-Plausible condition ($M = 2.06 \mu\text{V}$, $SD = 2.2$). The comparison between the two unrelated conditions revealed an interaction with Hemisphere such that the implausible condition was more positive-going than the plausible condition over the right hemisphere (see Fig. 2). Notice that this effect cannot be explained in terms of cloze probability, since the cloze probability of implausible targets was lower than that of plausible targets, and should therefore have resulted in a larger negativity (Kutas and Hillyard, 1984; DeLong et al., 2005). We interpret this positivity as the start of the P600 effect for the implausible condition observed in the later time window (see rERP analysis in Section 2.4).

To summarize, the pattern of effects in the N400 time window largely replicates DBC's results. We found that the N400 was clearly sensitive to semantic association, as weakly associated targets elicited a larger N400 than more strongly associated targets. There was no evidence that implausible, and therefore more difficult to integrate, words were associated with larger N400 amplitudes, as would be predicted by the integration account of the N400. On the contrary, in this time window plausible targets were actually more negative than implausible targets.

2.3.2. P600 time window (600–1000 ms)

The ANOVA over midline sites revealed an effect of condition, $F(2, 40) = 3.57$, $p < .05$, $\eta^2G = .04$, and an interaction with AP distribution $F(4, 80) = 3.05$, $p < .04$, $\eta^2G = .01$. Pairwise comparisons showed a larger negativity for the Unrelated-Plausible condition ($M = 2.30 \mu\text{V}$, $SD = 2.68$) compared to the Related-Plausible condition ($M = 3.82 \mu\text{V}$, $SD = 3.07$). The Unrelated-Implausible condition ($M = 3.27 \mu\text{V}$, $SD = 3.21$) was more negative than the Related-Plausible condition over more anterior sites (see Table 1 and Fig. 3). Crucially, the comparison between the two unrelated conditions revealed a larger positivity for the implausible condition compared to the plausible one, which was more pronounced over posterior electrodes.

The ANOVA over lateral sites revealed a marginal effect of Condition, $F(2, 40) = 2.92$, $p = .07$, $\eta^2G = .03$, and an interaction with AP distribution, $F(2, 40) = 8.06$, $p < .01$, $\eta^2G = .01$, and with Hemisphere, $F(2, 40) = 10.0$, $p < .01$, $\eta^2G = .01$. Pairwise comparisons showed again a larger negativity for the Unrelated-Plausible condition ($M = 1.67 \mu\text{V}$, $SD = 2.38$) compared to the Related-Plausible condition ($2.95 \mu\text{V}$, $SD = 2.68$). The Unrelated-Implausible condition ($M = 2.64 \mu\text{V}$, $SD = 3.09$) was more negative than the Related-Plausible condition over left-anterior sites, while it was more positive than the Unrelated-Plausible condition over right-posterior sites.

In sum, there are two main outcomes in the P600 time window. Firstly, we replicated the rather puzzling findings observed in DBC's experiment, showing no visible P600 effect for the event unrelated violation condition (1c) relative to the (related and plausible) baseline (1a). As in their study, we rather observed a sustained anterior negativity for the Unrelated-Implausible condition relative to the Related-

Plausible baseline (although in the present study the effect was more pronounced over the left hemisphere), becoming gradually more positive over posterior and occipital sites (see Figs. 1 and 3). Following DBC, we performed statistical analyses over occipital electrodes, to assess the reliability of the effect. The ANOVA with Condition (Related-Plausible, Unrelated-Plausible, Unrelated-Implausible) and Electrode (O1, Oz, O2) as within-subject factors revealed an effect of Condition, $F(2, 40) = 8.81$, $p < .01$, $\eta^2G = .10$, and no interaction with Electrode ($F < 1.6$). Crucially, and replicating DBC's findings, the Unrelated-Implausible condition ($M = 3.67 \mu\text{V}$, $SD = 2.57$) elicited a larger positivity than the Related-Plausible condition ($2.02 \mu\text{V}$, $SD = 2.19$), $F(1, 20) = 10.26$, $p < .01$, $\eta^2G = .10$.

Secondly, we validated DBC's interpretation of this pattern as resulting from spatiotemporal overlap between a P600 effect of plausibility and the long-lasting N400 effect of association elicited by the same contrast. As predicted by the component overlap hypothesis, a P600 effect of plausibility was clearly observed when the Unrelated-Implausible condition was contrasted against its proper baseline, i.e., the Unrelated-Plausible condition. The latter functions as a proper baseline precisely because it results in an N400 effect similar to the one elicited by the Unrelated-Implausible condition, thereby not obscuring the P600 effect to implausible targets. Although the P600 effect was more pronounced over the right hemisphere than it was in DBC's study, right-lateralised P600 effects are attested in the literature, and might result from differences in the stimuli or individual differences between participant groups (e.g., Palolahti et al., 2005; Tanner and van Hell, 2014; Zheng and Lemhöfer, 2019). Importantly, this P600 effect starts already in the N400 time window, as revealed by the similarly distributed positivity to implausible targets observed in the 300–500 ms time window, and by the rERP analysis reported below, further corroborating the hypothesis that the P600 and the N400 overlap in space and time.

Finally, the contrast between the two plausible conditions revealed a long-lasting negativity and no emerging positivity, providing further evidence that the P600 component is only sensitive to plausibility. Thus, the overall pattern of effects in the P600 time window, including the surface waveforms as well as what can be inferred about the underlying generators, provide strong support for the integration account of the P600 component (e.g., Brouwer et al., 2012).

2.4. rERP analysis

We supplement the traditional analysis of the ERP components with an rERP analysis (Brouwer et al., 2021a; Smith and Kutas, 2015a, b) which has been shown to provide more insights into how the experimentally manipulated variables contribute to the observed ERP signal (see Section 4.5).¹ Fig. 4 shows the residuals and coefficients (anchored to the intercept) on electrode Pz from Model 1 ($y = \beta_0 + \beta_1 \times \text{plausibility} + \beta_2 \times \text{association} + \epsilon$) and Model 2 ($y = \beta_0 + \beta_1 \times \text{plausibility} + \beta_2 \times \text{cloze} + \epsilon$). The residuals show that Model 1 offers a better fit to the observed data than Model 2, since in Model 1 the difference between observed and estimated voltages is closer to 0 in the entire epoch, including the N400 time window (where cloze might be expected to be a better predictor, as shown, for example, by Frank et al., 2013). Thus, in our further analyses we only consider Model 1, with association and plausibility as continuous predictors.

In the ERP data we observed an effect of association in the N400 time window and an effect of plausibility in the P600 time window. As expected under the hypothesis of spatiotemporal overlap between the N400 and the P600, the positive-going effect of plausibility was observed between conditions producing a similar N400 effect, while it did not emerge relative to a condition eliciting a less pronounced N400. Inspection of the regression coefficients from the rERPs (Model 1 as

¹ Standalone code and the data required to replicate this analysis is available at <https://github.com/hbrouwerdbc2021rerp>.

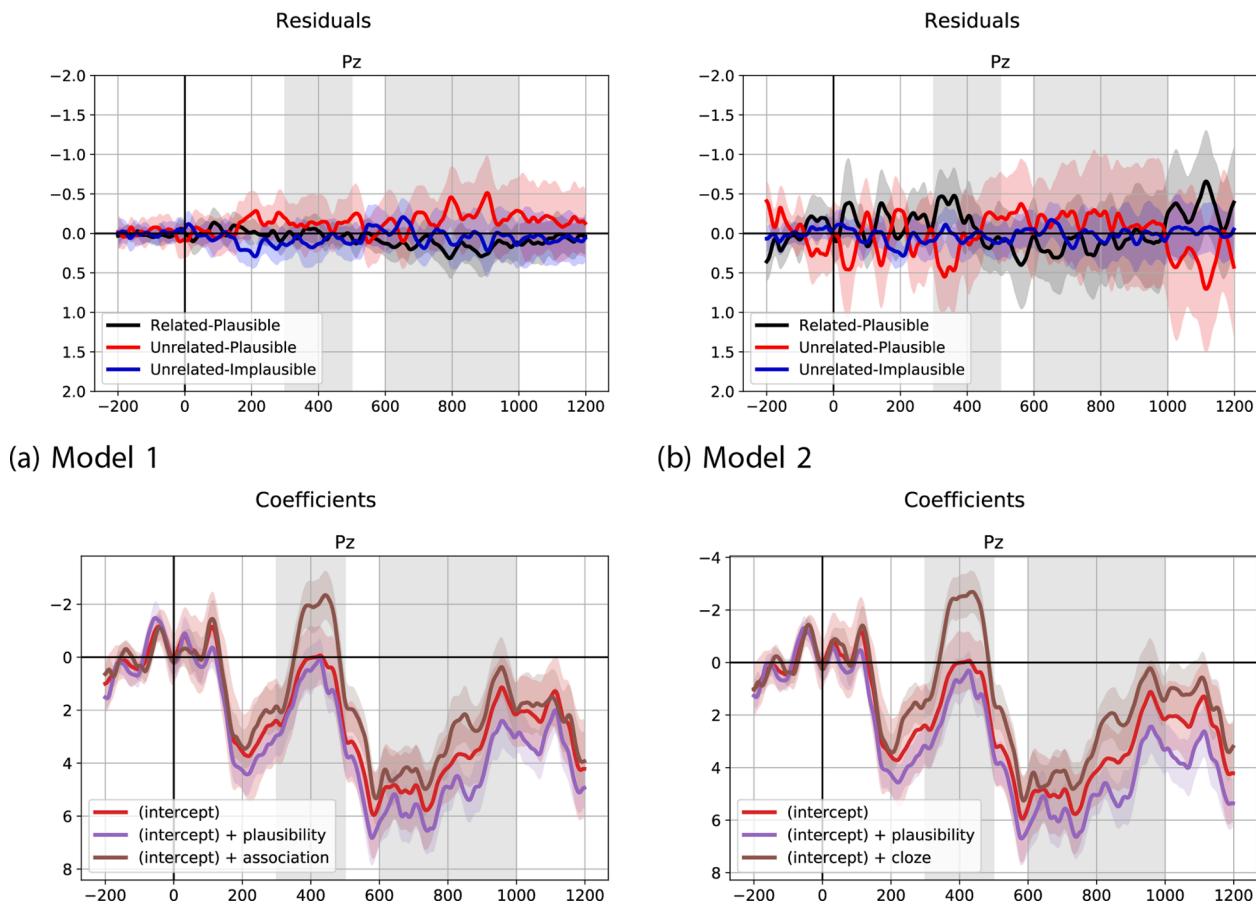


Fig. 4. Residuals and regression coefficients from Model 1 (a, left column) and Model 2 (b, right column). See text for more details. Slopes of predictors are anchored to the intercept. Negative is plotted upwards. Shaded regions around the waveforms show mean ± 2 SE.

shown in Fig. 5) indicates that association accounts for negative-going voltages in the N400 time window and, for most electrodes, in the P600 time window as well (see, for example, Pz compared to Oz). Plausibility, on the other hand, predicts positive-going voltages in the P600 time window (and also in the P200 window). From these observations it follows that the P600 effect to Unrelated-Implausible targets relative to the Related-Plausible baseline should be less likely to survive on electrodes in which the P600 time window shows an overlapping negativity (such as Pz) and more likely to be observed when no such overlapping negativity is produced, as in Oz. This was confirmed by the ERP analysis, showing a significant effect between these conditions only over occipital sites (see Section 2.3.2). Following Brouwer et al. (2021a), we also re-estimated voltages from two regression models isolating the contribution of association and plausibility, respectively (Fig. 6).

The rERP waveforms show that association alone predicts an N400 effect for the two unrelated conditions followed by a long lasting negativity, while plausibility predicts a P600 effect for the Unrelated-Implausible condition relative to both plausible conditions, already emerging in the N400 time window (thus replicating Brouwer et al., 2021a) and preceded by a positivity in the P200 time window (or, alternatively, emerging already at 200 ms and persisting along the entire epoch). This provides compelling evidence that a P600 effect of plausibility between the Unrelated-Implausible and the Related-Plausible conditions in the ERP waveforms (Fig. 1) is obscured by an overlapping negative-going effect of association. Concerning the P200 effect observed in the ERP waveforms, both the model coefficients from Model 1 (Fig. 5) and the estimates in Fig. 6 suggest a role of plausibility rather than association in explaining the effect. That is, the P200 seems to reflect a larger positivity for the implausible condition rather than a larger negativity for the unrelated conditions (i.e., the onset of the N400

effect).

To summarize, the rERP analysis confirmed and further qualified the effects in the N400 and P600 time windows observed with ERPs. The N400 effect appears to be driven by semantic association, as predicted by the retrieval account of the N400 (e.g., Brouwer et al., 2012; Kutas and Federmeier, 2000), while the late positivity is modulated by plausibility, consistent with the integration account of the P600 (e.g., Brouwer et al., 2012; Burkhardt, 2007). The two components overlap, resulting in the pattern of effects observed in the ERP waveforms. Additionally, the rERP analysis shows that association is a better predictor of N400 amplitudes than cloze, as shown by the residuals in the N400 time window. This is not surprising as the less predictable, implausible condition resulted in a more positive-going waveform. In order to account for this effect in Model 2, the coefficients for plausibility in the N400 time window are slightly more positive compared to Model 1, to compensate for the slight difference between Unrelated-Plausible and Unrelated-Implausible targets, for which the cloze ratings go in the opposite direction of observed N400 amplitude (see Fig. 4). Indeed, if one were to isolate the effect of cloze by setting plausibility to its mean (0) across trials, Unrelated-Implausible targets would actually produce a larger negativity ($M = -0.65 \mu V$) than the Unrelated-Plausible targets ($M = -0.59 \mu V$), reflecting the small difference in cloze probability ($=.014$) between the two conditions. Crucially, regardless of whether cloze or association is entered as a predictor into the model along with plausibility, it remains necessary to invoke spatiotemporal component overlap in order to account for the observed modulations of both the N400 and the P600.

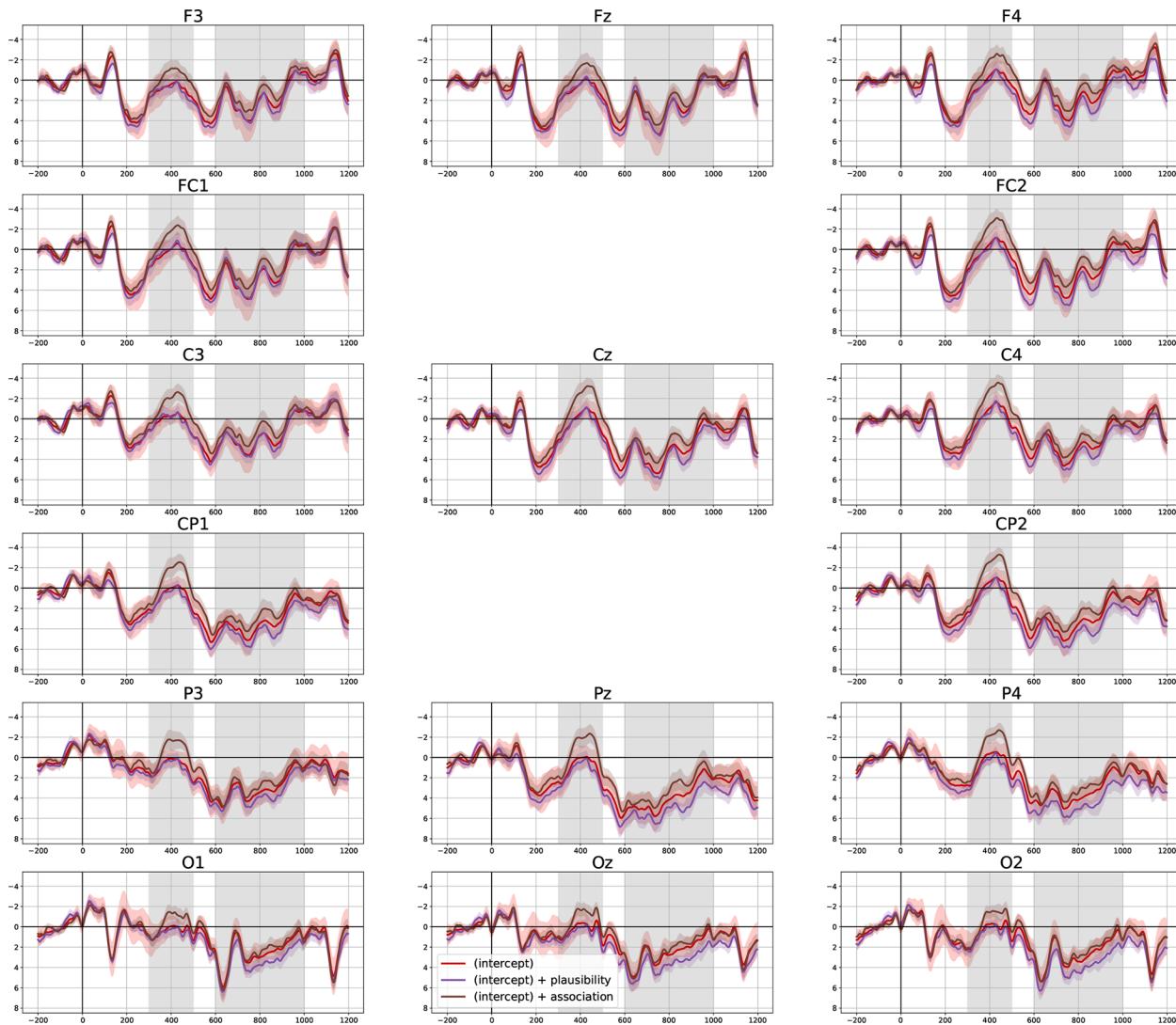
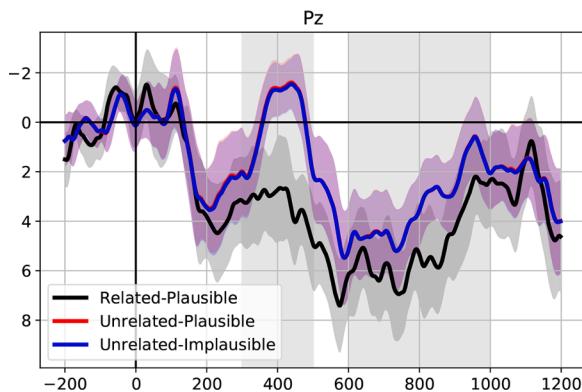


Fig. 5. Regression coefficients for the effects of plausibility and association (Model 1) in each time sample from -200 to 1200 ms relative to word onset. Slopes of predictors are anchored to the intercept. Negative is plotted upwards. Shaded areas around the waveforms indicate mean ± 2 SE.

rERPs: Effect of Association



rERPs: Effect of Plausibility

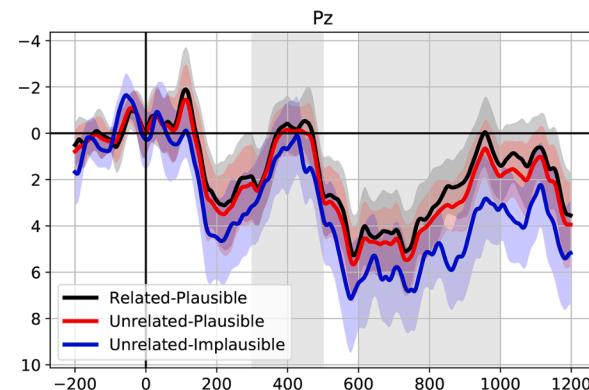


Fig. 6. Grand-average rERPs resulting from the intercept plus plausibility plus association model when plausibility (left) or association (right) is set to its mean rating (0) for all trials. Negative is plotted upwards. Shaded regions around the waveforms show mean estimated voltage ± 2 SE across subjects.

3. Discussion

The goal of the present study was to elucidate the impact of

spatiotemporal overlap between the N400 and P600 on the ERP signal. To this end, we implemented a design inspired by a previous study on the functional interpretation of the N400 and P600 (Delogu et al., 2019,

DBC), in which semantic association and plausibility were manipulated to tap into lexical retrieval and integration processes. While DBC found clear evidence that the N400 was sensitive to lexical retrieval processes, and not to semantic integration, the link of the P600 to integration processes was less clearly established, as a P600 effect of plausibility was observed only when the target word was semantically related to the context, that is, in absence of a preceding N400 effect. When the implausible target was unrelated to the context, resulting in an N400 effect, the expected biphasic N400/P600 effect of association and plausibility was observed only in occipital electrodes, while frontal electrodes showed a long lasting negativity.

DBC interpreted these rather puzzling findings as resulting from the effects of spatiotemporal overlap between the N400 and the P600. We investigated the component overlap hypothesis by testing a novel control condition, in which the critical word was a plausible continuation of the mini-discourse, but was semantically unrelated to the context. This condition should provide a proper baseline to reveal the presence of a P600 effect to Unrelated-Implausible targets, because the equally unrelated conditions should elicit similar N400 effects of association (relative to a related and plausible baseline), thereby not obscuring an emerging P600 effect of plausibility through spatiotemporal overlap between components. We therefore tested three conditions in which an Unrelated-Implausible target (2c) was contrasted against two baseline conditions: a *related* plausible condition (2a, as in DBC's study) and an *unrelated* plausible condition (2b). The results in the N400 time window replicated DBC's findings. We observed a clear sensitivity of the N400 to semantic association, with larger N400 amplitudes to unrelated targets (2b-c) than to related ones (2a). We also replicated DBC's surprising results in the P600 time window. Relative to a related and plausible baseline condition (2a), unrelated and implausible target words (2c) elicited a larger negativity over more anterior electrodes (although in the present experiment the effect was more lateralised than in DBC's study), becoming gradually more positive over posterior and occipital sites. Crucially, in the absence of an N400 effect of association between the two unrelated conditions, the P600 effect of plausibility was no longer obscured, as evidenced by the larger positivity produced by Unrelated-Implausible targets when compared to the novel baseline condition (2b). Thus, by introducing this novel contrast, we were able to show that spatiotemporal overlap between components can conceal an effect that is actually present in the underlying generators.

The presence of component overlap in the ERP signal was further confirmed by an rERP analysis of the data (Smith and Kutas, 2015a, b). This technique is particularly well suited to investigate component overlap, in that it allows for modelling the latent components underlying the surface signal by assessing the combined contribution of relevant, manipulated factors to predicting the observed signal modulations (Brouwer et al., 2021a). Our rERP analysis shows that, over the entire epoch, lexical association predicts increased negative voltages, while plausibility results in more positive-going waveforms. To the extent that these two predictors reflect the underlying components, their overlap in time pulls the waveforms in opposite directions, determining the absence or presence of an effect in the surface signal. Interestingly, both our ERP and rERP analyses revealed that the effect of plausibility has an early onset, already in the P200 time window. Although we did not have any specific predictions regarding the P200 component, P200 effects in language studies have been reported for a variety of semantic factors, at the lexical level (e.g., Dambacher et al., 2006; Coulson et al., 2005; Feng et al., 2019), but also in response to higher level language comprehension, such as irony or metaphors (Regel et al., 2011; Spotorno et al., 2013; Schneider et al., 2014), contextual constraint (Wlotko and Federmeier, 2007), and semantic or world knowledge violations (e.g., Wang et al., 2012; Leuthold et al., 2015). While the functional role of this component in language processing is still under debate, what seems relevant to the present study is that it has been observed in discourse contexts (e.g., Dambacher et al., 2006; Regel et al., 2011) and often in presence of a P600 effect, giving rise to a P200/P600 complex (e.g.,

Regel et al., 2011; Regel and Gunter, 2017; Domaneschi et al., 2018, see also Fritz and Baggio (2020)). Our results, however, do not allow us to establish whether plausibility results in a P200 followed by a P600 or, alternatively, in a sustained positivity with an early onset (masked by an intervening N400 in specific electrodes).

Recently, Nieuwland et al. (2019) similarly used an rERP approach to model ERPs. In particular, they explored effects of predictability, plausibility and semantic similarity. While they found more implausible words to elicit more negative voltages in the N400 time window (patterning with predictability), no significant effects in a later time window (from about 650 ms) were observed. Their design, however, was not intended to investigate plausibility effects, as shown by the reported plausibility ratings. While high predictability targets were rated as plausible, low predictability targets were neither plausible nor implausible. This might have attenuated any plausibility effect in the P600 time window. Our rERP analysis has instead shown that, over the entire epoch, the increased negativity is driven by association, while the increased positivity is predicted by plausibility, and that the two components overlap in time, affecting the surface signal.

Hence, the rERP analysis allows us to make inferences about the underlying cognitive processes, while taking into account the effects of component overlap. As we pointed out in the introduction, this is particularly important when it comes to theorising about the functional interpretation of ERP components, because component overlap affects the empirical evidence on which the various accounts are based. Without taking into account these effects, many results in the literature, both within and across studies, appear to be inconsistent. For example, in a systematic review on semantic incongruity effects, van Petten and Luka (2012) found that around 30% of the reviewed studies (which were selected to be as homogeneous as possible in terms of type of manipulation, task, etc.) reported biphasic N400/P600 effects, whereas the remaining studies reported N400 effects only. Obviously, this inconsistency makes it difficult to formulate hypotheses about the functional interpretation of the P600. Similarly, in DBC's study, the finding that the presence of a P600 effect depended on whether or not the target was semantically related to the context was puzzling, weakening the theoretical account of the P600 that was proposed by the authors. However, in both cases, the picture becomes more coherent once the effects of component overlap are factored in (see Brouwer and Crocker, 2017 for a discussion of van Petten and Luka, 2012's findings). Thus, future research should take into account how component overlap may shape the data, possibly masking effects in the surface waveforms, and use appropriate techniques to tackle this issue.

Beyond reconciling apparent inconsistencies within and across studies, component overlap might be invoked to account for differences in topographic distributions, such as the frontal positivity elicited by plausible but highly unexpected words occurring in highly constraining sentences (e.g., Brothers et al., 2020; Federmeier et al., 2007; Brothers et al., 2015; van Petten and Luka, 2012; Thornhill and van Petten, 2012; DeLong et al., 2014). For example, Thornhill and van Petten (2012) report a frontal positivity with an early onset in more anterior electrodes, co-occurring with a centro-parietal N400 effect, in response to unrelated low-cloze completions in high-constraint sentences. As their Fig. 1 shows (Thornhill and van Petten, 2012, p. 386), the early part of the positivity is overtaken by the N400, while the later part becomes gradually closer to the baseline the higher the N400 amplitude is. This pattern may result from a broadly distributed P600 effect that starts in the N400 time window and overlaps with a centro-parietal N400, which cancels out the posterior part of the P600. While we agree that this is only speculative, it is possible that particular distributions of ERP effects result from spatiotemporal overlap between components.

The present results show that, when examining the underlying generators, the data provide unambiguous evidence that the N400 reflects lexical retrieval processes and not semantic integration (e.g., Brouwer et al., 2012; Kutas and Federmeier, 2000; Kutas and Federmeier, 2011; Lau et al., 2009; Lau et al., 2008), while the P600 indexes compositional

integration processes, not limited to syntactic aspects alone (e.g., Brouwer et al., 2012; Brouwer et al., 2017; Burkhardt, 2007). That is, syntactic/reanalysis accounts of the P600 are subsumed by the semantic integration account, in that syntactically ill-formed or complex sentences necessary result in more effortful utterance meaning construction (see Brouwer et al., 2012, for further discussion). On this account, the P600 is assumed to be a family of late positivities – varying in amplitude, scalp distribution, latency, and duration – reflecting different subprocesses (e.g., referential processing, thematic role revisions, pragmatic inferences) involved during the incremental construction or updating of utterance meaning (Brouwer et al., 2012).

Overall, our results are consistent with the Retrieval-Integration (RI) account of the functional interpretations of the N400 and the P600 (Brouwer et al., 2012; Brouwer et al., 2017), in which the two components reflect the retrieval and integration processes that are routinely performed during comprehension. Recently, the RI account has been instantiated within a neurocomputational model of incremental language comprehension (Brouwer et al., 2021b) that links retrieval and integration processes to a comprehension-centric notion of surprisal (Venhuizen et al., 2019). This surprisal metric, which incorporates both linguistic experience and world knowledge, reflects the likelihood of a change in interpretation resulting from integrating the meaning of an incoming word into the unfolding utterance. The model predicts that the P600 component, which reflects integration difficulty, indexes comprehension-centric surprisal, i.e., the degree to which a word is (syntactically, semantically, and pragmatically) expected. The N400 indexing retrieval processes, on the other hand, is predicted to be sensitive to both lexical association and expectancy via lexical and contextual priming. In particular, contextual priming explains the well-documented expectancy and discourse-level effects on the N400 (e.g., Frank et al., 2015; Delogu et al., 2017; Kutas and Federmeier, 2011; Nieuwland and van Berkum, 2006). Importantly, the RI hypothesis does not only account for processing effort in implausible linguistic input, but also in more naturalistic language, with implausibility being an extreme case of totally unexpected input. In future work, we plan to validate this hypothesis using more naturalistic stimulus materials. An interesting open question, for instance, is if there is any modulation of the P600 component in continuous EEG recording using naturalistic stimuli as in Frank et al. (2015) or Brennan and Hale (2019). A treatment of spatio-temporal component overlap as potentially offered by the rERP approach would likely be essential to answer this question.

Importantly, under the RI view, integration processes may start while lexical information is not yet completely retrieved, predicting a temporal overlap between the (generators of) the N400 and the P600 (Brouwer and Hoeks, 2013). That is, upon encountering a word, the generator of the P600 may continuously attempt to integrate even incomplete lexical information, while additional information is still being retrieved. The theory predicts biphasic N400/P600 effects in the latent components whenever retrieval and integration processes are manipulated. Biphasic patterns, however, will be detected in the surface signal only if the P600 effect survives component overlap, which often occurs in presence of strong semantic anomalies (e.g., Hoeks et al., 2004; van Herten et al., 2006). In case of less strong violations, such as the kinds of world-knowledge violations investigated in DBC's as well as in the present study, the integration process indexed by the P600 is predicted to be less demanding, and therefore less likely to emerge in the surface signal when overlapping with a strong N400 effect. As a consequence, the absence of an expected P600 effect in presence of an N400 effect should be treated with caution. In such cases, the rERP technique may be used to further investigate the observed ERP responses and identify potentially hidden effects. This technique, as already pointed out, is also particularly suited to investigate the electrophysiological correlates of language comprehension in more naturalistic settings, such as reading a novel, where the effects of component overlap cannot be detected via experimental design.

3.1. Conclusions

In conclusion, the present study provides concrete support for our hypothesis that component overlap between the N400 and P600 can affect the visibility of effects found in the surface morphology of the ERP signal. As a consequence, our results in conjunction with those by DBC provide robust evidence that the underlying generators of the observed signal are sensitive to association and plausibility, as indexed by the N400 and P600 respectively. We interpret this in the context of the Retrieval-Integration account, in which the N400 indexes lexical retrieval and the P600 reflects compositional semantic integration processes, which are assumed to dynamically interact in time. Generally, our results underline the need to establish methodologies that deal with component overlap, both at the experimental design and data analysis level, in order to accurately inform theoretical accounts of the underlying cognitive processes. Equally, it may be necessary to consider component overlap when assessing the evidence reported by existing studies, and their theoretical consequences.

4. Methods and materials

4.1. Participants

Twenty-three participants from Saarland University took part in the experiment. All were right-handed, native speakers of German, and had a normal or corrected-to-normal vision. All participants gave written informed consent and were paid for taking part in the experiment. The reported studies were conducted in accordance with the ethics approval granted by the Deutsche Gesellschaft für Sprache (DGfS).

4.2. Materials

We created ninety triplets in German, as illustrated in (2) (see [Supplementary Materials](#) for the original versions in German). Following DBC, we pre-tested the materials by collecting association and plausibility ratings, and by estimating the cloze probability of the target words. All participants in the pre-tests were recruited via Prolific Academic (www.prolific.co).

To estimate the cloze probability of the target words, three lists counterbalancing items and conditions were created with the sentence pairs presented up to and including the determiner/possessive preceding the target word (e.g., “Peter ging raus in den Regen. Sofort öffnete er seinen ...”). Each list was assigned to 10 participants. The mean cloze probability of the target word was higher in the Related-Plausible baseline ($M = .61$, $SD = .27$) than in both the Unrelated-Plausible ($M = .02$, $SD = .04$) and the Unrelated-Implausible conditions ($M = .004$, $SD = .03$). Pairwise comparisons with Bonferroni correction showed a significant difference between all conditions (all $p < .01$). Thus, while the target word was highly unexpected in both the unrelated conditions, it was even more so in the Unrelated-Implausible condition.

The semantic association between the target nouns (e.g., *umbrella*) and the related vs. unrelated nouns in the context sentence (*rain* vs. *restaurant*), was rated on 7-points Likert scale (1 = not at all related, 7 = strongly related) by 20 participants who did not take part in the cloze study. Mean ratings for the related condition (*rain – umbrella*) was 6.66 ($SD = .47$), while for the unrelated conditions (*restaurant – umbrella*) was 1.46 ($SD = .56$). Thus, as expected, the target word was highly associated with the context in condition (2a) and weakly associated in conditions (2b-c).

Finally, we collected plausibility ratings on a 7-points Likert scale (1 = highly implausible, 7 = highly plausible) for all the experimental items presented up to the target noun, to avoid judgments to be affected by sentential materials appearing after it. Three counterbalanced lists were created, each one presented to 10 participants who did not take part into the cloze or the association rating studies. The Unrelated-Implausible condition was judged to be less plausible ($M = 1.61$, SD

= 0.49) than both the Related-Plausible ($M = 6.65$, $SD = 0.45$) and the Unrelated-Plausible ($M = 5.45$, $SD = 0.79$) conditions. Pairwise comparisons with Bonferroni correction showed significant differences between all conditions (all $p < .01$).

Three counterbalanced lists were created so that each item appeared in each list in a different condition. The experimental items were intermixed with 102 filler passages created to counterbalance the proportion of related vs. unrelated and plausible vs. implausible items throughout the experiment, as well as the frequency with which the verb in the context sentence appeared in a certain condition, so that participants could not use the verb in the context to predict the plausibility of the item. Fillers also varied in structure and length (for example, by incorporating subordinate or relative clauses) to introduce some variability in the stimuli.

4.3. Procedure

Participants were seated in a dimly lit sound-proof, electro-magnetically shielded booth, in front of a 24 inch computer screen. Stimuli were presented with the E-prime software (Psychology Software Tools, Inc.) in white font on a black background. Each trial began with a screen in which participants were asked to press a button to start reading the passages. The context sentence appeared as a whole until participants pressed a button to proceed. Then a fixation cross appeared for 750 ms, after which the target sentence was displayed word-by-word in the center of the screen, for 350 ms preceded by a 150 ms inter-stimulus interval. After each trial, participants judged whether the passage was plausible by pressing one of two buttons (plausible – implausible) on a response box. Items were presented in pseudo-randomized order in 3 blocks, with breaks after each block. Before the experiment, participants performed a short training session to familiarize with the task. The experiment session lasted approximately 1 hour.

4.4. Electrophysiological recording and processing

The EEG was recorded using 26 active scalp electrodes placed according to the 10–20 system. The horizontal electro-oculogram (EOG) was monitored with two electrodes placed at the outer canthi of each eye and the vertical EOG with two electrodes above and below the left eye. Electrode impedance was kept below 5 k Ω for all scalp electrode sites, and below 10 k Ω for the EOG electrodes. The signal was digitized at a sampling rate of 500 Hz. During recording no online filters were used. The EEG signal was band-pass filtered offline at 0.1–30 Hz and re-referenced to the average of the left and right mastoid electrodes. The EEG was segmented into epochs time-locked to the onset of the target nouns (-200 ms to 1200 ms). Epochs containing ocular and muscular artifacts were rejected prior to analyses. Two participants showing excessive artifacts were discarded. EEG data were averaged for each subject and condition using 200 ms pre-stimulus baseline.

4.5. Analyses

Following DBC, we computed mean amplitudes in the 300–500 ms (N400) and in the 600–1000 ms (P600) time windows. Separate analyses were performed on midline sites (Fz, Cz, Pz) and lateral sites (F3, FC1, FC5, P3, CP1, CP5, F4, FC2, FC6, P4, CP2, CP6). Data from lateral sites were grouped into four regions of interest (ROIs): left anterior (F3, FC1, FC5), right anterior (F4, FC2, FC6) left posterior (P3, CP1, CP5) right posterior (P4, CP2, CP6). Within each time window, ANOVAs on midline sites were carried out with Condition (Related-Plausible, Unrelated-Plausible, Unrelated-Implausible) and anterior-posterior (AP) distribution (anterior, central, posterior) as within-subject factors. The ANOVAs over later sites included Condition, AP distribution (anterior, posterior), and Hemisphere (left, right) as within-subject factors. The ANOVAs were followed by pairwise comparisons of interest. The Greenhouse-Geisser correction was applied to all ANOVAs with greater than one degree of

freedom in the numerator. In such cases, the corrected p-value is reported. Generalized eta-squared (η^2_G) is reported as a measure of effect size.

We also performed an rERP analysis (Smith and Kutas, 2015a, b) to decompose how plausibility, association, and cloze probability contribute to the scalp-recorded voltage and to investigate the effects of component overlap on the surface signal (Brouwer et al., 2021a).

Multiple linear regression models are fitted for each electrode, subject, and 2 ms time slice (corresponding to a 500 Hz sampling rate over a 1400 ms epoch) separately. In order to aid interpretation of the slopes, all predictors are inverted by subtracting each rating from the maximum possible rating (7 for plausibility and association; 1.0 for cloze probability), and subsequently z-standardized. After fitting the models by subjects across trials, trials are regrouped by condition, and model fit is qualitatively assessed by looking at mean residuals (mean ϵ) per electrode and time-slice within each condition.

Our norming ratings reveal that while cloze and association are strongly correlated in that both unrelated conditions (Unrelated-Plausible and Unrelated-Implausible) yield lower cloze probabilities and association ratings than the Related-Plausible condition, plausibility reveals a different pattern in that both plausible conditions (Related-Plausible and Unrelated-Plausible) are rated as more plausible than the Related-Implausible condition. We therefore consider two rERP analyses of our data: A model including plausibility and association as predictors Model 1:

$y = \beta_0 + \beta_1 \times \text{plausibility} + \beta_2 \times \text{association} + \epsilon$, and a model including plausibility and cloze probability Model 2:

$y = \beta_0 + \beta_1 \times \text{plausibility} + \beta_2 \times \text{cloze} + \epsilon$. Given that association and cloze are highly correlated, a model including all three predictors is not expected to offer a better explanation of the data and hence we do not consider it. It is important to note that, unlike in previous studies employing the rERP technique for inferential statistics (e.g., Nieuwland et al., 2019), the aim of this analysis is not to determine the significance of the predictors in a given model, but rather in how well the models that include (or exclude) those predictors fit the observed voltages. A predictor could in fact be highly significant in a model that poorly fits the observed data, potentially leading to wrong conclusions about the importance of that predictor (see Brouwer et al., 2021a, for further discussion). The coefficients for the model that best fits the data are then examined to see how the predictors combine in producing the scalp-recorded voltages. To visualise the influence of each individual predictor independent of the other, rERP waveforms are re-estimated from the fitted regression models while setting one of the predictors to its mean (0) for all trials (e.g., effects of plausibility are investigated by estimating $y = \beta_0 + \beta_1 \times \text{plausibility} + \beta_2 \times 0 + \epsilon$).

As we use the analyses solely for illustrative means, no further inferential statistics will be conducted on the rERP results.

CRediT authorship contribution statement

Francesca Delogu: Conceptualization, Methodology, Formal analysis, Investigation, Writing - original draft, Writing - review & editing. **Harm Brouwer:** Conceptualization, Methodology, Formal analysis, Writing - review & editing, Funding acquisition. **Matthew W. Crocker:** Conceptualization, Methodology, Writing - review & editing, Funding acquisition.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102. The authors would like to thank Miriam Schulz for help creating the stimuli and carrying out the experiment.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <https://doi.org/10.1016/j.brainres.2021.147514>.

References

- Baggio, G., Hagoort, P., 2011. The balance between memory and unification in semantics: a dynamic account of the N400. *Language Cogn. Processes* 26 (9), 1338–1367.
- Bornkessel-Schleesewsky, I., Schleesewsky, M., 2008. An alternative perspective on semantic P600 effects in language comprehension. *Brain Res. Rev.* 59, 55–73.
- Brennan, J.R., Hale, J.T., 2019. Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PLoS ONE* 14 (1), e0207741.
- Brothers, T., Swaab, T.Y., Traxler, M.J., 2015. Effects of prediction and contextual support on lexical processing: prediction takes precedence. *Cognition* 236, 135–149.
- Brothers, T., Wlotko, E.W., Warnke, L., Kuperberg, G.R., 2020. Going the extra mile: effects of discourse context on two late positivities during language comprehension. *Neurobiol. Language* 1 (1), 135–160.
- Brouwer, H., Crocker, M.W., 2017. On the proper treatment of the N400 and the P600 in language comprehension. *Front. Psychol.* 8, 1327.
- Brouwer, H., Crocker, M.W., Venhuizen, N.J., Hoeks, J.C.J., 2017. A neurocomputational model of the N400 and the P600 in language processing. *Cogn. Sci.* 41, 75–83.
- Brouwer, H., Delogu, F., Crocker, M.W., 2021a. Splitting event-related potentials: modeling latent components using regression-based waveform estimation. *Eur. J. Neurosci.* 53 (4), 974–995.
- Brouwer, H., Delogu, F., Venhuizen, N.J., Crocker, M.W., 2021b. Neurobehavioral correlates of Surprise in language comprehension: a neurocomputational model. *Front. Psychol.* 12, 615538.
- Brouwer, H., Fitz, H., Hoeks, J., 2012. Getting real about semantic illusions: rethinking the functional role of the P600 in language comprehension. *Brain Res.* 1446, 127–143.
- Brouwer, H., Hoeks, J.C.J., 2013. A time and place for language comprehension: mapping the N400 and the P600 to a minimal cortical network. *Front. Human Neurosci.* 7, 758.
- Brown, C.M., Hagoort, P., 1993. The processing nature of the N400: evidence from masked priming. *J. Cogn. Neurosci.* 5, 34–44.
- Brown, C.M., Hagoort, P., 1999. On the electrophysiology of language comprehension: Implications for the human language system. In: Crocker, M., Pickering, P., Clifton, C. (Eds.), *Architectures and mechanisms for language processing*. Cambridge University Press, pp. 213–237.
- Burkhardt, P., 2006. Inferential bridging relations reveal distinct neural mechanisms: Evidence from event-related brain potentials. *Brain Lang.* 98 (2), 159–168.
- Burkhardt, P., 2007. The P600 reflects cost of new information in discourse memory. *NeuroReport* 18 (17), 1851–1854.
- Coulson, S., Federmeier, K.D., Petten, C.V., Kutas, M., 2005. Right hemisphere sensitivity to word- and sentence-level context: Evidence from event-related brain potentials. *J. Exp. Psychol.: Learn. Memory Cogn.* 31, 129–147.
- Dambacher, M., Kliegl, R., Hofmann, M., Jacobs, A.M., 2006. Frequency and predictability effects on event-related potentials during reading. *Brain Res.* 1084 (1), 89–103.
- Delogu, F., Brouwer, H., Crocker, M.W., 2019. Event-related potentials index lexical retrieval (N400) and integration (P600) during language comprehension. *Brain Cogn.* 135, 103569.
- Delogu, F., Crocker, M.W., Drenhaus, H., 2017. Teasing apart coercion and surprise: evidence from eye-movements and ERPs. *Cognition* 161, 46–59.
- Delogu, F., Drenhaus, H., Crocker, M.W., 2018. On the predictability of event boundaries in discourse: an ERP investigation. *Memory Cogn.* 46 (2), 315–325.
- DeLong, K.A., Quant, L., Kutas, M., 2014. Predictability, plausibility, and two late ERP positivities during written sentence comprehension. *Neuropsychologia* 61, 150–162.
- DeLong, K.A., Urbach, T.P., Kutas, M., 2005. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nat. Neurosci.* 8 (89), 1117–1121.
- Domaneschi, F., Canal, P., Masia, V., Vallauri, E.L., Bambini, V., 2018. N400 and P600 modulation in presupposition accommodation: the effect of different trigger types. *J. Neurlinguistics* 45, 13–35.
- Donchin, E., Ritter, W., McCallum, W.C., 1978. Cognitive psychophysiology: the endogenous components of the ERP. In: Callaway, E., Tueting, P., Koslow, S.H. (Eds.), *Event-related brain potentials in man*. Academic Press, pp. 349–411. <https://doi.org/10.1016/B978-0-12-155150-6.50019-5>.
- Fabiani, M., Gratton, G., Federmeier, K.D., 2012. Event-related brain potentials: methods, theory, and applications. In: Cacioppo, J.T., Tassinary, L.G., Bernston, G.G. (Eds.), *Handbook of psychophysiology*. Oxford University Press, pp. 85–119.
- Federmeier, K.D., Wlotko, E.W., Ochoa-Dewald, E.D., Kutas, M., 2007. Multiple effects of sentential constraint on word processing. *Brain Res.* 1146, 75–84.
- Feng, J., Gong, T., Shuai, L., Wu, Y., 2019. No morphological markers, no problem: ERP study reveals semantic contribution to distinct neural substrates between noun and verb processing in online sentence comprehension. *Front. Neurosci.* 13, 957.
- Frank, S.L., Otten, L.J., Galli, G., Vigliocco, G., 2013. Word surprisal predicts N400 amplitude during reading. In: Proceedings of the 51st annual meeting of the association for computational linguistics. ACL, pp. 878–883.
- Frank, S.L., Otten, L.J., Galli, G., Vigliocco, G., 2015. The ERP response to the amount of information conveyed by words in sentences. *Brain Lang.* 140, 1–11.
- Friederici, A.D., 1995. The time course of syntactic activation during language processing: a model based on neurological and neurophysiological data. *Brain Lang.* 50, 259–281.
- Fritz, J., Baggio, G., 2020. Meaning composition in minimal phrasal contexts: distinct ERP effect of intensionality and denotation. *Lang. Cogn. Neurosci.* 35 (10), 1295–1313.
- Hagoort, P., 2003. Interplay between syntax and semantics during sentence comprehension: ERP effects of combining syntactic and semantic violations. *J. Cogn. Neurosci.* 15 (6), 883–899.
- Hagoort, P., Brown, C., Osterhout, L., 1999. The neurocognition of syntactic processing. In: Brown, C., Hagoort, P. (Eds.), *Neurocognition of language*. Oxford University Press, Oxford, UK, pp. 273–316.
- Hagoort, P., Hald, L., Bastiaansen, M., Petersson, K.M., 2004. Integration of word meaning and world knowledge in language comprehension. *Science* 304 (5669), 438–441.
- Hoeks, J.C.J., Stowe, L.A., Doedens, G., 2004. Seeing words in context: the interaction of lexical and sentence level information during reading. *Cogn. Brain Res.* 19 (1), 59–73.
- Hoeks, J.C.J., Stowe, L.A., Hendriks, P., Brouwer, H., 2013. Questions left unanswered: how the brain responds to missing information. *PLoS ONE* 8 (10), e7359.
- Kaan, E., Harris, A., Gibson, E., Holcomb, P., 2000. The P600 as an index of syntactic integration difficulty. *Lang. Cogn. Processes* 15 (2), 159–201.
- Kaan, E., Swaab, T., 2000. Repair, revision, and complexity in syntactic analysis: an electrophysiological differentiation. *J. Cogn. Neurosci.* 15 (1), 98–110.
- Kappenman, E., Luck, S.J., 2012. The ups and down of brainwave recordings. In: Luck, S.J., Kappenman, E.S. (Eds.), *The Oxford handbook of event-related potential components*. Oxford University Press, pp. 3–30.
- Kim, A., Osterhout, L., 2005. The independence of combinatory semantic processing: evidence from event-related potentials. *J. Mem. Lang.* 52, 205–225.
- Kolk, H.H.J., Chwilla, D.J., van Herten, M., Oor, P.J., 2003. Structure and limited capacity in verbal working memory: a study with event-related potentials. *Brain Lang.* 85, 1–36.
- Kuperberg, G.R., 2007. Neural mechanisms of language comprehension: challenges to syntax. *Brain Res.* 1146, 23–49.
- Kuperberg, G.R., Kreher, D.A., Sitnikova, T., Caplan, D., Holcomb, P.J., 2007. The role of animacy and thematic relationships in processing active English sentences: evidence from event-related potentials. *Brain Lang.* 100, 223–238.
- Kutas, M., Federmeier, K.D., 2000. Electrophysiology reveals semantic memory use in language comprehension. *Trends Cogn. 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.
- Kutas, M., Hillyard, S.A., 1984. Brain potentials during reading reflect word expectancy and semantic association. *Nature* 307, 161–163.
- Lau, E.F., Almeida, D., Hines, P.C., Poeppel, D., 2009. A lexical basis for N400 context effects: evidence from MEG. *Brain Lang.* 111, 161–172.
- Lau, E.F., Namyst, A., Fogel, A., Delgado, T., 2016. A direct comparison of N400 effects of predictability and incongruity in adjective-noun combination. *Collabora: Psychol.* 2 (1)(13), 1–19.
- Lau, E.F., Phillips, C., Poeppel, D., 2008. A cortical network for semantics: (de)constructing the N400. *Nat. Rev. Neurosci.* 9, 920–933.
- Leuthold, H., Kunkel, A., Mackenzie, I.G., Filik, R., 2015. Online processing for moral transgressions: ERP evidence for spontaneous evaluation. *Soc. Cogn. Affect. Neurosci.* 10 (8), 1021–1029.
- Luck, S., 2005. An introduction to the event-related potential technique. MIT Press, Cambridge, MA.
- Näätänen, R., Picton, T., 1987. The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology* 24 (4), 375–425.
- Nieuwland, M.S., Barr, D.J., Bartolazzi, F., Busch-Moreno, S., Darley, E., Ferguson, H.J., Wolfsturn, S.V.G.Z., 2019. Dissociable effects of prediction and integration during language comprehension: evidence from a large-scale study using brain potentials. *Philos. Trans. R. Soc. London B* 375, 20180522.
- Nieuwland, M.S., van Berkum, J.J.A., 2006. When peanuts fall in love: N400 evidence for the power of discourse. *J. Cogn. Neurosci.* 18 (7), 1098–1111.
- Osterhout, L., Holcomb, P.J., 1992. Event-related brain potentials elicited by syntactic anomaly. *J. Mem. Lang.* 31, 785–806.
- Palolahti, M., Leino, S., Jokela, M., Kopra, K., Paavilainen, P., 2005. Event-related potentials suggest early interaction between syntax and semantics during on-line sentence comprehension. *Neurosci. Lett.* 384, 222–227.
- Regel, S., Gunter, T.C., 2017. Don't get me wrong: ERP evidence from cueing communicative intentions. *Front. Psychol.* 8, 1465.
- Regel, S., Gunter, T.C., Friederici, A.D., 2011. Isn't it ironic? An electrophysiological exploration of figurative language processing. *J. Cogn. Neurosci.* 23, 277–293.
- Schneider, S., Rapp, A.M., Heußinger, F.B., Ernst, L.H., Hamm, F., Fallgatter, A.J., Ehli, A.-C., 2014. Beyond the N400: complementary access to early neural correlates of novel metaphor comprehension using combined electrophysiological and haemodynamic measurements. *Cortex* 53, 45–59.
- Smith, N.J., Kutas, M., 2015a. Regression-based estimation of ERP waveforms: I. The ERP framework. *Psychophysiology* 52, 157–168.
- Smith, N.J., Kutas, M., 2015b. Regression-based estimation of ERP waveforms: II. Nonlinear effects, overlap correction, and practical considerations. *Psychophysiology* 52, 169–181.
- Spotorio, N., Cheylus, A., Henst, J.B.V.D., Noveck, I.A., 2013. What's behind a P600? Integration operations during irony processing. *PLoS ONE* 8 (6), e66839.

- Tanner, D., van Hell, J.G., 2014. ERPs reveal individual differences in morphosyntactic processing. *Neuropsychologia* 56, 289–301.
- Thornhill, F.E., van Petten, C., 2012. Lexical versus conceptual anticipation during sentence processing: frontal and N400 components. *Int. J. Psychophysiol.* 83, 382–392.
- van Berkum, J.J.A., 2009. The neuropragmatics of ‘simple’ utterance comprehension: an ERP review. In: Sauerland, U., Yatsushiro, K. (Eds.), Semantics and pragmatics: from experiment to theory. Palgrave Macmillan, pp. 276–316.
- van Berkum, J.J.A., 2010. The brain is a prediction machine that cares about good and bad – any implications for neuropragmatics? *Ital. J. Linguist.* 22, 181–208.
- van Herten, M., Chwilla, D.J., Kolk, H.H.J., 2006. When heuristics clash with parsing routines: ERP evidence for conflict monitoring in sentence perception. *J. Cogn. Neurosci.* 18 (7), 1181–1197.
- van Petten, C., Luka, B.J., 2012. Prediction during language comprehension: Benefits, costs, and ERP components. *Int. J. Psychophysiol.* 83 (1), 176–190.
- Venuhuizen, N.J., Crocker, M.W., Brouwer, H., 2019. Expectation-based comprehension: modeling the interaction of world knowledge and linguistic experience. *Discourse Processes* 56, 229–255.
- Wang, L., Zhu, Z., Bastiaansen, M., 2012. Integration or predictability? A further specification of the functional role of gamma oscillations in language comprehension. *Front. Psychol.* 3, 187.
- Wlotko, E.W., Federmeier, K.D., 2007. Finding the right word: hemispheric asymmetries in the use of sentence context information. *Neuropsychologia* 45, 3001–3014.
- Zheng, X., Lemhöfer, K., 2019. The semantic P600 in second language processing: When syntax conflicts with semantics. *Neuropsychologia* 127, 131–147.