Accepted manuscript. Please cite as Stone, K. & Rabovsky, M. (2025). The role of syntactic and semantic cues in preventing temporary illusions of plausibility. Journal of Cognitive Neuroscience, 1-27, https://doi.org/10.1162/jocn_a_02320.

The role of syntactic and semantic cues in preventing temporary illusions of plausibility

Kate Stone[1,2]* and Milena Rabovsky[2]

[1]University of Hull, [2]University of Potsdam

Abstract

Unexpected words within a context elicit large N400 brain potentials. However, sometimes the N400 at an unexpected word is small when stereotypical agent and patient roles are reversed, such as at *'arrested'* in '*the cop that the thief arrested*'. In a study of 74 native German speakers, we demonstrate evidence that readers can avoid this so-called "N400 semantic illusion" if the verb is delayed with neutral information such as '*that evening*', but are less able to do so if the delay contains cues that could further strengthen the canonical interpretation, such as '*with handcuffs*'. In doing so, we provide a conceptual replication of a relatively new finding, and extend previous research by showing that the semantic content of the delay is important. Moreover, we demonstrate evidence that the effect of only the neutral delay increases as the experiment progresses. We propose an interpretation of these findings with reference to the Sentence Gestalt model (Rabovsky et al., 2018), which accounts for the initial illusion as resulting from uncertainty and an erroneous interpretation based on a strong semantic attractor. Two additional, novel contributions of the work are a demonstration that the illusion can be elicited in German, despite its explicit subject-object case marking, and an exploration of illusion effect among individual readers.

Introduction

The N400 is a well-established index of meaning processing (Kutas & Federmeier, 2011). Amplitude of the N400 is usually large at semantically implausible words in sentences such as *I take coffee with cream and dog*, suggesting that readers have comprehended the meaning of the preceding context such that they know what new words are plausible (Kutas & Hillyard, 1980). However, it is sometimes observed that an implausible word triggers a small-amplitude N400. For example, in the sentence "*For breakfast the eggs would eat…*", the usual thematic role of *eggs* has been reversed such that the verb *eat* is semantically implausible, yet it elicits only a small N400 similar in size to the corresponding congruent sentence (Kuperberg et al., 2003). This is despite all the syntactic and semantic cues necessary for interpretation being available by the time the verb is encountered. A similar effect has been observed in a number of languages (Chow et al., 2016, 2018; Hoeks et al., 2004; Kim & Osterhout, 2005; Kolk et al., 2003;

Kuperberg et al., 2003; Liao et al., 2022; van Herten et al., 2005, 2006), although it has not been tested as directly in German, as we discuss below.

Based on the lack of an N400 effect between canonical and reversed sentences, it is inferred by some that readers experience an "illusion" of semantic plausibility, although evidence from plausibility judgements suggests the illusion is temporary (Chow et al., 2018; Ehrenhofer et al., submitted; Hoeks et al., 2004; Kim & Osterhout, 2005; Kolk et al., 2003; Kuperberg et al., 2003). In this paper, we will refer to this phenomenon as a "temporary semantic illusion" and sometimes a "semantic illusion" as a short-hand, since we focus only on the time period of this temporary illusion (i.e. the N400 time window) while remaining agnostic as to whether the illusion persists.

The semantic illusion can be prevented, however: Chow et al. (2018) showed that the verb *arrest* elicits a small N400 in (1a), but when it is delayed in (1b), N400 amplitude relative to a non-role-reversed control sentence increases (translated from Mandarin):

(1) a. Thief.$_{\text{AGENT}}$ cop arrest… (role reversal, implausible)

b. Thief.$_{\text{AGENT}}$ cop that evening arrest… (role reversal, implausible)

This delay effect challenges existing accounts of the illusion as we will discuss shortly, and thus has the potential to distinguish between accounts, refining our understanding of argument-verb computation (Liao et al., 2022). The effect also raises two questions: what is it about the delay that prevents the illusion, and what does this tell us about the nature of the initial illusion? In this paper we present an experiment where we manipulated the content of the delaying sentence fragment and show that only a neutral delay with little semantic association with the context serves to prevent the illusion. In contrast, more semantically associated delays may strengthen the canonical interpretation and appear to sustain the illusion. We propose that this suggests continuous, ongoing conflict between syntactic and semantic cues. This conflict can be resolved by a semantically neutral delay because it reinforces the syntactic structure of the sentence or provides additional time to resolve the conflict (although we do not directly test a time-based explanation in the current experiment, it will be the subject of forthcoming work; Stone & Rabovsky, in preparation). In contrast, the conflict persists when a semantically associated delay reinforces the syntactic structure and/or provides additional time for conflict resolution, but additionally reinforces the canonical interpretation.

How does the N400 semantic illusion arise?

The original report of a semantic illusion caused by "*For breakfast the eggs would eat…"* (Kuperberg et al., 2003), has provided fertile ground for debate about how meaning is computed during sentence processing, particularly about the role of competing semantic and syntactic cues. Kuperberg et al. (2003) proposed that because of the typical word order of agents and patients in English sentences (agent first), readers initially assign *eggs* the thematic role of agent (the "doer" in the event). At the verb *eat*, this assignment is violated because *eat* requires an animate agent. The animacy violation triggers a syntactic reanalysis process associated with a late ERP positivity (P600) and not the N400.

However, subsequent research showed that animacy violations are not sufficient to produce the illusion if there is no semantic association between the target verb and the context (Kim & Osterhout, 2005). In that study, "*...the dusty tabletops were devouring*" did not elicit the illusion (there was an N400 for *devouring* relative to a plausible control), but "*...the hearty meal was devouring*" did. The authors concluded

that semantic association of the verb and its context drove an illusion of plausibility by inducing a theme-first analysis of the verbal argument, despite the agent-first interpretation indicated unambiguously by syntactic cues. A misinterpretation or semantic illusion thus arises, similar in nature to that of a "good enough" parse (Ferreira et al., 2002). A number of subsequent accounts have proposed that such illusions are possible because of the temporary separation or even blocking of one or more syntactic, semantic, and/or thematic processing streams (Bornkessel-Schlesewsky & Schlesewsky, 2008; Kolk et al., 2003; Kuperberg, 2007; van Herten et al., 2005, 2006). Each of these accounts focuses on processing issues that are triggered once the verb is encountered.

Other accounts focus on processing occurring before presentation of the verb. One of these is a time-based account of argument-verb computation (Chow et al., 2016, 2018; Liao et al., 2022). This three-stage, sequential theory of argument-verb computation proposes that verb prediction is initially based on lexical association, then on recognition of nouns as potential verbal arguments without thematic role assignment, and finally on full thematic role assignment. Verb predictions are made at each stage of this process but are not fully constrained until the final step. The small or absent N400 at the verb thus occurs because the verb in experimental sentences is presented before processing has reached the final step, so the reader makes the same verb prediction in both role-reversed and canonical sentences, which is consistent with the verb they see. The Retrieval-Integration (RI) model (Brouwer et al., 2017) proposes a similar forward influence of lexical priming, although this influences retrieval of the verb rather than prediction and does not incorporate a temporal ordering of processing steps prior to the N400.

Other accounts in this category focus on the probabilistic representation created by the global context rather than specific lexical items, accounting for the illusion—via different mechanisms—as resulting from the strong influence of prior world knowledge (Bornkessel-Schlesewsky & Schlesewsky, 2019; Li & Ettinger, 2023; Rabovsky et al., 2018). In this paper we focus on one of these: the Sentence Gestalt (SG) model (Rabovsky et al., 2018). The SG model proposes that readers generate a predictive representation of an event described by a presented sentence in which certain concepts and their roles in the event become more or less activated depending on the input and the reader's experience. N400 amplitude is driven by the degree of update that each new word triggers across this predictive representation. Under the SG model, the semantic illusion as reflected by small N400 amplitude arises because conflict between knowledge about canonical word order and stereotypical agent-patient roles creates uncertainty about the event. This uncertainty is not resolved by the time the word *eat* is seen in *the eggs would eat*, and therefore *eat* triggers only a small update of the reader's meaning representation. More explicitly, model simulations show that after *At breakfast*…, activation levels indicate that the most likely action is *eat*, that the most likely agents are animate nouns, and that the most likely patients are foods (particularly *eggs;* Rabovsky et al., 2018). After *the eggs*, the model's activation for *eggs* as the patient becomes even stronger. Even after *eat* is encountered, activation for *eggs* as the patient is still strong. The change in overall state of activation from *eggs* to *eat* is therefore very small, resulting in only a small N400 correlate relative to other semantically incongruent sentences without role reversals. The illusion in this account thus arises from conflict between the interpretation suggested by the syntactic structure of the sentence input and the familiar roles of events involving eggs, breakfast, and eating. The initial understanding of the agent and patient is therefore incorrect at the time of verb presentation.

The delay effect (Chow et al., 2018) was observed subsequent to publication of most of these accounts (with the exception of Liao et al.'s three-stage processing theory) so that none has explicitly addressed the delay effect. However, we can speculate whether and how each of the accounts might accommodate the effect.

How might a delay prevent the illusion?

The delay effect presents a challenge to existing accounts of the illusion in three ways: The first is that if the delay effect does result simply from the additional time provided by the delaying sentence fragment, models that treat comprehension as a discrete process which is updated at each new word in a sentence may not be able to accommodate it. This is, however, more an implementation rather than a theoretical challenge as most would probably agree that comprehension is a continuous process and discrete models are only designed as such for the purpose of simplicity. The second challenge is to accounts in which the illusion results from processing triggered by the verb (Kim & Osterhout, 2005; Kolk et al., 2003; Kuperberg, 2007; van Herten et al., 2005, 2006), since for the delay to have any effect on the N400 it must affect processing that occurs before the verb is seen. The third, more difficult challenge to overcome for any account is why the additional sentence fragment in Chow et al. (2018) would have any effect: "*that evening*" in (1) does not contain any information that would facilitate lexical retrieval or increase the probability of a different verb and in any case provides an identical amount of time and information in both the canonical (1a) and reversed (1b) conditions.

Liao and colleagues' proposal is that time alone is required to allow role assignment to complete and constrain an initial, lexical association-based verb prediction (Liao et al., 2022). Thus in the role-reversed sentence (1b), delaying *arrest* with *that evening* allows sufficient time for *thief* to be assigned the agent of the sentence and a congruent verb to be preactivated (e.g. *escaped*). An N400 is then elicited when the input *arrested* falsifies this prediction. The possibility that time plays a role in preventing the illusion is corroborated by evidence from similar constructions in Japanese, where simply slowing the word presentation rate was sufficient to prevent the semantic illusion, as indicated by a larger N400 in the role reversed condition (Nakamura et al., 2024).

As a tentative hypothesis, we propose that the current SG model (Rabovsky et al., 2018) could accommodate the delay finding via the following account: In the case of role-reversed sentences, the initial interpretation of the sentence is highly uncertain or even incorrect due to the stronger influence of the semantic "attractor" (a state toward which the model settles across e.g. a sentence). The role of new input might therefore be to provide new information that changes the model's initial representation. In (1) the new words (e.g., *that evening*) are consistent with the structure of the sentence so far but have no particular semantic relationship with thief, cop, or arrest. They thus corroborate the structural interpretation ("the current syntactic parse is still viable"), which contributes slightly more information than the semantic content ("null"). This could result in strengthening of the syntactic attractor—even if only weakly—which increases its influence over the sentence interpretation in both role-reversed and canonical conditions. The literal interpretation of the role-reversed sentence may therefore be more likely to emerge and the verb detected as semantically odd, triggering a larger N400.

An alternative possibility is that it is time alone that resolves the delay, although this is not consistent with the current SG model's implementation as it models N400 amplitude as discrete per-word updates. However, a temporal contribution is completely consistent with the underlying theory: If updating continues between words, additional time should help the model to resolve the conflict and thus overcome the illusion, in part because the influence of the syntactic attractor over the sentence representation will continue to grow. This generates a novel prediction: If the delay contained words that were not only consistent with the structure of the sentence but additionally strengthened the semantic association that led to the initial illusion, the benefit of the delay observed in Chow et al. (2018) might be reduced and the

illusion sustained. Thus under this hypothesis, the content of the delay may be crucial to whether the illusion is avoided or not. We further develop this tentative hypothesis in light of some of the experimental results in the Discussion.

Incremental role assignment and reversal anomalies in German

The current experiment tests the role reversal phenomenon in German. In the only previous study of German with a comparable design comparable, Bornkessel-Schlesewsky et al. (2011) cite evidence showing that role reversed sentences such as 2b elicited larger N400 amplitude relative to their canonical counterparts in 2a, suggesting that German readers may be immune to the illusion and attribute this to German's explicit case marking (Schlesewsky & Bornkessel-Schlesewsky, 2009, Birgit Rausing Language Conference Program in Linguistics).

(2) a. …dass den.$_{ACC}$ Schalter der.$_{NOM}$ Techniker bedient (canonical)
…that the.$_{ACC}$ switch the.$_{NOM}$ technician operates
b. …dass der.$_{NOM}$ Schalter den.$_{ACC}$ Techniker bedient (reversed)
...that the.$_{NOM}$ switch the.$_{ACC}$ technician operates
c. …dass Techniker Schalter bedienen (canonical, no case marking)
...that technicians operate switches
d. …dass Schalter Techniker bedienen (reversed, no case marking)
…that switches operate technicians

Bornkessel-Schlesewsky et al. (2011) propose that case marking provides a cue to thematic role, which in 2b conflicts with the general preference to assign agency to animate nouns. Under their own extended Argument Dependency Model (eADM) account, this violates the "compute linking" step reflected by the N400 and explains why the N400 is elicited in these German stimuli but not in other role reversal stimuli in languages with animacy cues but without case marking (e.g. the English "*the eggs would eat")*. In languages without case marking, an inanimate noun can be assigned a patient role and pass the compute linking step even in the role reversed version because of the semantic plausibility of the verb and arguments; no N400 is elicited. A conflict is only detected at the "generalised mapping" step reflected by the P600 when syntactic cues indicate an agent role for the inanimate noun. This could suggest that it is the combination of case marking and animacy cues in German that prevent the illusion and that sentences without case marking such as 2d should yield an illusion; however a larger N400 was still observed in 2d relative to 2c. Bornkessel-Schlesewsky et al. (2011) argue that the word order preference for *switches* as the agent in 2d conflicts with the animacy-based expectation for *technicians* as agent, eliciting the N400; however it is not clear how this finding can be reconciled with findings demonstrating the illusion in other case-less languages with word order and animacy expectations (Kolk et al., 2003; Kuperberg, 2007; Kuperberg et al., 2003; van Herten et al., 2005, 2006). In sum, case marking could mean German readers are more robust than speakers of other languages to the role reversal illusion in sentences with one inanimate and one animate argument. However, the available results leave open the question of whether semantically implausible sentences where both arguments are animate may yield an illusion in German.

In addition to the benefit of thematic role cues from case marking, the single-word presentation paradigms used for ERP studies mean that this thematic role information is presented earlier to participants in German than in previous languages attesting the N400 semantic illusion, and is seen before each noun. In Mandarin, subjecthood is marked by a particle that is seen only after the noun (Chow et al., 2015, 2018;

cf. Bornkessel-Schlesewsky et al., 2011, but only in animate-inanimate argument pairs). Japanese similarly uses post-nominal particles marking subject and object (Nakamura et al., 2024). Other tested languages use no case marking at all (English, Kim & Osterhout, 2005; Kuperberg, 2007; Kuperberg et al., 2003; Dutch, Hoeks et al., 2004; Kolk et al., 2003; van Herten et al., 2005, 2006; cf. Turkish and Icelandic in Bornkessel-Schlesewsky et al., 2011, but using a different study design). Although subjecthood (and thus likely agency) of the first noun phrase can be strongly inferred from word order in these languages, there remains a non-zero chance that the noun phrase is not the subject and thus the strategy is not as fail-proof as explicit case marking. Considering that time may be important in eliciting the illusion, earlier access to role cues in the rapid serial visual presentation (RSVP) paradigm may be an additional reason why German readers are less susceptible.

The current experiment

The current experiment had two aims: (i) to determine whether the N400 semantic illusion could be demonstrated in German and (ii) to further interrogate the delay effect by determining whether the content of the delay would affect readers' ability to avoid the N400 semantic illusion. Whereas previous studies of role reversals have also examined the P600, we chose to focus only on the N400, firstly because it was the locus of the delay effect, and secondly because the theoretical model with which we align our interpretation (the SG model; Rabovsky et al., 2018) currently only models the N400 and not the P600. We do however discuss the P600 window in the *Discussion*.

With respect to aim (ii), our first goal was to replicate the delay effect observed in Chow et al. (2018). The resolution of the N400 semantic illusion has—at the time of writing—only been observed in Mandarin (Chow et al., 2018) and Japanese (Nakamura et al., 2024). We therefore first aimed to replicate Chow et al.'s (2018) Experiment 3 by showing that a neutral delay such as *that evening* would prevent the illusion. The second goal was to extend Chow et al.'s work by investigating the effect of additional semantic cues in the delay to test whether the delay would be sufficient to generate a robust sentence representation or whether this representation would be subject to ongoing interference from semantic cues.

The experiment had a 2×3 design with the factors S(ubject)-O(bject) order (canonical/reversed) and delay type (none/neutral/associated; **Table 1**). In the neutral delay conditions (c/d), the verb in the sentence "*The viewers want to know which guest the moderator invited"* was delayed by adding a neutral sentence fragment (e.g. *additionally*) between the verbal arguments and the verb. While *additionally* does not provide specific cues disambiguating argument roles or predicting the verb, it does slightly reinforce the syntactic structure of the preceding sentence fragment (in the sense that it is a cue that the sentence is still compatible with the current parse). In the associated delay conditions (e/f), the verb was delayed with a sentence fragment that similarly reinforced structure, but was also lexico-semantically associated with the context in the canonical sentence (e.g. *to the panel show*) and was meant to reinforce the canonical interpretation. Note that we call this the "associated" delay condition, but the way in which the delaying sentence fragment was linked to the target verb varied across the experimental materials. The important common point from our perspective was that the fragment biased the interpretation of the sentence towards the canonical interpretation, through lexical association and/or its propositional content, which is crucial for the SG model's account of the effect.

We placed the critical sentence fragment within a wh- clause as in German, this allowed us to place the verb in sentence-final position and hold the position of the verbal arguments constant while switching

only their subject and object affixes (*-er* for subject and *-en* for object). The ordering of subject and object in wh-clauses is relatively free in German, meaning that although there is a preference for subject-first order if the grammatical case marking of the first noun is ambiguous (Schlesewsky et al., 2000), either order is considered well-formed by native speakers if syntactically unambiguous, as in the current study.

Table 1

Example experimental item

Condition	S-O order	Delay type	Critical sentence fragment
Die Zuschauer wollen erfahren, … *The viewers want to know, ...*			
(a)	Canonical	None	…welchen Gast der Moderator **<u>eingeladen</u>** hat. *...which.$_{ACC}$ guest the.$_{NOM}$ moderator **<u>invited</u>** has.*
(b)	Reversal	None	…welcher Gast den Moderator **<u>eingeladen</u>** hat. *...which.$_{NOM}$ guest the.$_{ACC}$ moderator **<u>invited</u>** has.*
(c)	Canonical	Neutral	…welchen Gast der Moderator auch noch **<u>eingeladen</u>** hat. *...which.$_{ACC}$ guest the.$_{NOM}$ moderator additionally **<u>invited</u>** has.*
(d)	Reversal	Neutral	…welcher Gast den Moderator auch noch **<u>eingeladen</u>** hat. *...which.$_{NOM}$ guest the.$_{ACC}$ moderator additionally **<u>invited</u>** has.*
(e)	Canonical	Associated	…welchen Gast den Moderator zur Diskussionsrunde **<u>eingeladen</u>** hat. *...which.$_{ACC}$ guest the.$_{NOM}$ moderator to the panel show **<u>invited</u>** has.*
(f)	Reversal	Associated	…welcher Gast den Moderator zur Diskussionsrunde **<u>eingeladen</u>** hat. *...which.$_{NOM}$ guest the.$_{ACC}$ moderator to the panel show **<u>invited</u>** has.*
Translation: The viewers want to know which.$_{ACC/NOM}$ guest the.$_{NOM/ACC}$ moderator has (additionally) invited (to the panel show).			

Note. The target verb is bolded and underlined. Note that in the example, the delay fragment in the associated condition is syntactically more complex than the fragment in the neutral condition. This was not always the case across items - indeed often the reverse was true. We provide further summary statistics in **Table 2**.

Manipulating the sentence before the verb did of course mean that the pre-verbal region differed between delay conditions. This would have meant that direct comparisons of the verb between, for example, the canonical sentences between the neutral and associated delay conditions would have been confounded

by a differing baseline. However, we were only interested in the canonical vs. reversed contrast within each delay condition (nested effects) and whether these nested effects differed between associated, neutral, and no delay conditions (interaction). This meant that any statistical comparisons were conducted either on verbs with identical baseline regions (nested effects) or on the difference of amplitude differences (interaction) where the baseline was no longer of concern.

In the example given in **Table 1**, the delaying fragment in the associated delay condition is syntactically more complex than that in the neutral condition. We note that sometimes the number of words also differed between these conditions across items. **Table 2** summarises some features of the delay fragments by condition. Delay fragments in the associated conditions were, on average, 0.38 words longer ($p < 0.001$) and associated delay fragments were 36% more likely to have one or more noun phrases than neutral delay fragments ($p < 0.001$). The length of the delay fragment in number of words and the number of noun phrases were therefore used as predictors in the statistical models where the position of the target verb was not matched.

Table 2

Summary statistics of features of the experimental items

Condition	Verb cloze probability (SD)	Mean number verb-delaying words (SD)	% noun phrases (NPs) in delay fragment		
			0 NPs	1 NP	2 NPs
(a)	0.32 (0.22)	-	-	-	-
(b)	0.07 (0.09)	-	-	-	-
(c)	0.34 (0.21)	2.28 (0.61)	52	47	2
(d)	0.11 (0.11)	2.28 (0.61)	52	47	2
(e)	0.44 (0.26)	2.66 (0.75)	16	81	3
(f)	0.18 (0.17)	2.66 (0.75)	16	81	3

Note. The difference in cloze probabilities between canonical and reversed conditions was 0.25 (no delay condition), 0.23 (neutral delay condition), and 0.26 (associated delay condition).

Hypothesis 1: Replicating Chow et al (2018)

We predicted an interaction effect between subject-object order (canonical/reversed) and delay (none/neutral delay): The difference in mean amplitude in the N400 window should be larger d vs. c than b vs. a (see Table 1).

Hypothesis 2: Examining semantic cues in the delay

The additional semantic cues could affect processing in different ways, visualised in **Figure 1**. Panel A represents a situation where additional semantic cues contribute to resolution of the semantic illusion over

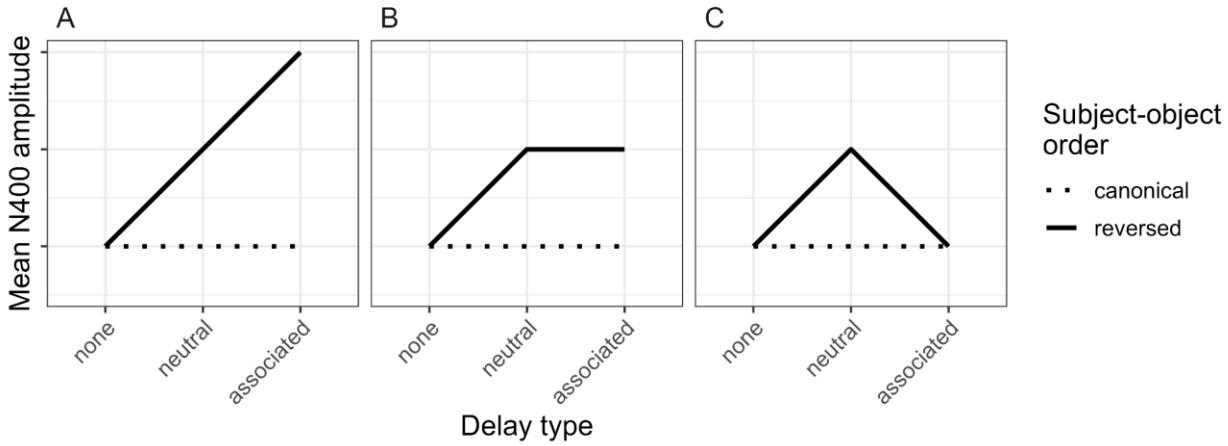
and above a neutral delay, such that the difference in N400 amplitude between canonical and reversed sentences increases from the neutral to the associated condition. This would suggest an additive contribution of increased syntactic constraint/time and semantic cues to disambiguating argument roles. Panel B represents a situation where semantic cues do not provide an additional boost over the neutral delay and so the N400 effect is of a similar magnitude for both delay types. This would suggest that the key to preventing the semantic illusion is the time or the additional syntactic constraint and that new semantic cues do not further facilitate disambiguation of argument roles or interfere with the now-correct representation.

Panel C represents the most likely scenario given the importance of the semantic attractor in the SG model. The associated delay in our stimuli was associated with the event in the canonical sentences only, e.g. a moderator inviting a guest *to the panel show*. In the reversed sentences, it did not strengthen the plausibility of the guest as the agent. This was evident from the relatively high cloze probability in the associated, reversed condition (e); some cloze test participants insisted that "invited" was still the most likely continuation even when given a second chance to review the sentence. In this scenario, the strengthened semantic cues outweigh the syntactic constraints and sustain the illusion.

The experimental design, hypotheses, and analysis plan were pre-registered at https://osf.io/b65vu/. Minor deviations were made from the pre-registration: i) we used different contrast coding with equivalent interpretation (-0.5/0.5 instead of -1/1); ii) additionally reported the proportion of the posterior distribution that was in the same direction as the effect for each analysis; iii) reported the Bayes factor of a less informative prior to simplify the Methods section by reducing the number of different priors used for different analyses (the conclusion was identical to the pre-registered prior which was included in the sensitivity analysis); iv) proposed two additional hypotheses (Fig. 1B and 1C) as 1C in particular was felt to be more consistent with the SG model and the results of the cloze test. Data, code, and experimental materials can be found at https://osf.io/qtek9/.

Figure 1.

Graphical predictions for the different possible effects of delay type on role reversals.



Note. **A.** N400 amplitude is the same for canonical and reversed sentences in the no-delay condition, but becomes larger in the reversed condition with increasing informativity of the delay. **B.** Both delays prevent the illusion to a similar extent. **C.** Only the neutral delay prevents the illusion while the associated delay does not.

Methods

Participants

Participants were 74 healthy, right-handed, German native-speaking adults recruited via the University of Potsdam's online participant pool (mean age 25 years, range: 18-37 years, SD = 5 years). All participants had normal or corrected-to-normal vision, reported no history of developmental or current language, neurological, or psychiatric disorder, and had not participated in the cloze test. Eight additional participants were excluded: two due to sensitivity to the electrode gel which prevented obtaining good impedances, one for answering comprehension questions with less than 70% accuracy, two due to data loss, and three due to irreparable artefact in over 75% of target EEG segments. Mean accuracy in the final sample of 74 participants was 93% (SD = 0.04). In line with university policy, participants were reimbursed for their time either financially or in the form of credit points toward their studies. All participants provided written consent to participation in the study. Ethical approval for the study was granted by the University of Potsdam Ethical Review Committee (27/2021).

The sample size was originally pre-registered to be determined by continuing recruitment until we reached a Bayes factor of 6 for Hypothesis 1 or our sample size cap of 100, whichever came first (Schönbrodt & Wagenmakers, 2018). A poorer than expected recruitment rate after the lifting of Covid-19 lockdowns meant that the pre-registered plan became infeasible within time constraints. A design analysis using data simulated from the final sample indicated that, assuming the data reported below were a good representation of true values, even with a sample size of 200 participants we would not have reached our cut-off criterion *(Appendix A3)*.

Materials

The experimental stimuli were 186 items consisting of three pairs of canonical and role-reversed sentences. For the first pair, there was no delay between the verbal arguments and the target verb. In the second pair, there was a neutral sentence fragment delaying the verb and in the third pair, an associated fragment. An example stimulus is in **Table 1**. For the EEG experiment, the stimuli were split into six lists in a Latin square design and presented in pseudo-randomised order such that the same condition was never presented more than two times in a row. The critical sentences were interspersed with filler sentences in a 1:1 ratio to disguise the purpose of the experiment (186 fillers in total). The fillers were plausible sentences with a variety of structures without wh- relative clauses, such as "*When Dieter left the house, he was secretly photographed by a journalist*.". After 50% of the fillers, a yes/no question appeared, e.g. "*Did the photographer photograph Dieter with his permission?*". While Chow et al. (2018) used plausibility judgements of each sentence, experimental or filler, we chose not to ask questions of the experimental sentences in order to avoid task-related effects. Asking questions of only a proportion of the filler sentences meant that participants needed to remain attentive as it was unpredictable when a question might appear. It also served to minimise the length of an already long experiment. No participants reported noticing that questions only appeared with certain types of sentences.

Cloze test

Prior to the EEG experiment, a cloze test was conducted to determine the probability with which readers expected the verb in each sentence, as well as to ensure the stimuli were well-constructed. The experimental

sentences were truncated before the critical verb. The truncated sentences were split into six lists in a Latin square design and presented in fully randomised order. We obtained 30 completions per sentence from participants via the online platform Prolific.co. Participants gave explicit consent to participate in the study and received a financial reimbursement for their time.

Procedure

Participants sat in a shielded EEG cabin approximately 65 cm from a 30 x 54 cm presentation screen. The experimental paradigm was built and presented using Open Sesame (Mathôt et al., 2012). Each experimental session began with instructions advising participants that they would read sentences presented word-by-word and that after some sentences, they would answer a yes/no question using the keyboard. Each experimental session began with five practice trials. Each trial in the experiment began with a 500 ms fixation point in the centre of the screen followed by a blank screen jittered with a mean of 1000 ms and standard deviation 250 ms.

Each sentence was presented word-by-word for a duration of 190 ms per word plus 20 ms for each letter, with a minimum duration of 250 ms for any word. The variable word presentation rate was used to accommodate long German words as a set presentation rate would have resulted in a noticeably brief presentation duration for these words. All word presentation durations were matched within items. The target word was always presented for 700 ms regardless of length. The inter-word interval was 300 ms. The longer presentation duration of the target relative to other non-target words could potentially have strengthened any N400 effect (Wlotko & Federmeier, 2015); however, since the duration was identical across all target words, we reasoned that the possibility of such effects was outweighed by the benefit of having a 1000 ms analysis window before the next word onset. Moreover, we note that the pattern of N400 results mirrors previous studies using the same manipulation. The comprehension questions were answered via the keyboard, which triggered the next trial. Breaks were offered after every 28 sentences. The testing session including EEG setup lasted approximately three hours.

EEG recording and preprocessing

The EEG recordings were made in the Department of Psychology at the University of Potsdam, Germany, during 2022 using BrainVision Recorder (Version 1.23.001, Brain Products, 2020). EEG was recorded in an electro-magnetically shielded EEG cabin using a 32-lead actiChamp Plus system (Brain Products, 2020a) and electrodes arranged on the head based on the international 10-20 system (Jasper, 1958). Electrode impedances were kept below 5 kOhm throughout the experiment. EEG was recorded at a sampling rate of 500 Hz and online filtered with a low-pass filter of 140 Hz, using the right mastoid as a reference. Raw EEG recordings were preprocessed using the *R* package *eeguana* (Nicenboim, 2018). Butterworth FIR filters were applied with a high-pass cut-off at 0.01 Hz (order of 4, transition band width 0.01 Hz) and a low-pass cutoff at 30 Hz (order of 4, transition band width 7.50 Hz). The EEG was then re-referenced to the average of the two mastoids.

The recording was then segmented into epochs from sentence onset to sentence end. Blinks were corrected using automatic independent component analysis (ICA; Jung et al., 2001) with the Fast ICA algorithm (Hyvärinen et al., 2001). ICA components were manually inspected for each participant and removed if they strongly correlated with the ocular channels. The target words for the critical sentences were then extracted and segmented into 1200 ms epochs representing 200 ms pre-stimulus baseline and 1000 ms post-stimulus window. EEG channels with muscle artefact or irreparable eye-blink or movement

artefact were automatically rejected, defined as voltage steps exceeding a maximum 50 μV, or a maximum voltage peak-to-peak difference of more than 100 μV in a 150 ms window. Overall, this resulted in the exclusion of 7.47% of the 13,764 target trials. A further 0.01% were not recorded due to a technical issue or experimenter error, leaving approximately 2100 trials per condition for the statistical analysis. Each trial epoch was baseline-corrected relative to the 200 ms pre-stimulus interval.

Planned analysis

Linear mixed effects models were fit in *brms* (Buerkner, 2018) in *R* (R Core Team, 2020) with maximal random effects structure for subjects and items. The dependent variable was mean N400 amplitude in the window 300-500 ms across electrodes Cz, CP1, CPz, CP2, P3, Pz, P4, PO3, POz, PO4. This pre-registered time window and electrode selection differed slightly from Chow et al. (2018; 250-450 ms over electrodes P3, PZ, P4, O1, OZ, and O2). The N400 is known to be broad and its distribution may differ slightly for reasons not related to the experiment (e.g. the effect of skull shape on dipoles, differences in lab procedures). Our pre-registration decision took into account previous research in our lab and student population which had yielded a more centro-posterior distribution than that reported in Chow et al. (2018), with maximal effects in a 300-500 ms window. Both of these parameters are consistent with the typical spatio-temporal properties of the N400 reported in the literature (Kutas & Federmeier, 2011).

The first model we fit was a model of the full 2 × 3 design using simple difference contrasts. The contrasts reflected the hypotheses that the neutral delay would prevent the illusion in reversed vs. canonical sentences in the delay conditions, but that the resolution—if any—would be smaller in the associated delay condition relative to the neutral delay condition. This model used non-directional priors and was evaluated using the 95% credible interval (CrI) of the posterior probability distributions, as well as the proportion of the posterior samples that were greater or less than zero, depending on the direction of the respective effect. The proportion of posterior samples was taken as an indication of how much probability mass each posterior had that supported a non-zero effect. This was especially useful when the 95% credible interval contained zero. However, this was not intended as a metric that conclusively excluded a null effect; rather, we used it as an additional descriptive measure. We considered proportions of at least 95% to be consistent with an effect and proportions less than 95% to be inconsistent with an effect, since in the latter case more than 5% of the posterior estimates would be zero or in the opposite direction. This proportion reflects the significance threshold at an alpha of 0.05 in the frequentist framework (but see also McElreath, 2015, p58).

For the specific pre-registered hypothesis about an interaction of role order and delay type where we wanted a more conclusive answer, we computed Bayes factors using 2 × 2 subsets of the data and a range of priors on the interaction. We interpreted evidence favouring the alternative over the null hypothesis in line with Lee & Wagenmakers (2014) and Jeffreys (1939), where a ratio of at least 3:1 (Bayes factor of 3) is considered moderate evidence in favour of the alternative hypothesis and at most 1:3 (Bayes factor of 0.3) as moderate evidence in favour of the null hypothesis. For each hypothesis test, a range of priors was used in order to examine how these may have changed the Bayes factor and our conclusions (Schad et al., 2020). A description of each model's parameters follows.

Main model

The predictors were subject-object order with sum contrast coding (canonical -0.5, reversed 0.5) and delay type with simple difference contrast coding (no delay -0.5, neutral 0.5, associated 0 for the comparison neutral vs. no delay; and no delay 0, neutral 0.5, associated -0.5 for the comparison associated vs. neutral

delay). We were primarily interested in the interaction, as well as in nested effects: This meant that we were comparing differences of differences, or differences within each delay condition where the target word position, the number of words in the delaying fragment, and the number of noun phrases in the delaying fragment were matched. For that reason, we did not include these features of the delay fragment as predictors in the model.[1] Maximal random effects structure was used to model the nested effects of individual subjects and experimental items. The model specification was:

$$N400\_amplitude \sim role\_order * delay\_type + random\ effects$$

Prior distributions were used to encode our expectations about the range of plausible effect sizes for each model estimate. These were informed by the effects observed in previous Bayesian ERP analyses (Nicenboim et al., 2020; Stone et al., 2022), but were not made strictly informative in order to account for the different experimental design. The priors were:

$$intercept \sim Normal(0,5)$$
$$\beta_{main,interaction} \sim Normal(0,1)$$
$$\sigma_{subject,item} \sim Normal_{+}(0,0.5)$$
$$\sigma_{residual} \sim Normal_{+}(8,2)$$
$$\rho_{random\ effect\ correlations} \sim LKJ(2)$$

Bayes factor analyses

The data were subset into three 2 × 2 datasets to test the interaction with role order of: i) neutral vs. no delay (testing the replication of Chow et al., 2018), ii) neutral vs. associated delay (testing whether the neutral and associated delays influenced the role order effect differently), and iii) associated vs. no delay (testing whether the associated delay prevented the illusion). The predictors were role order with sum contrast coding (canonical -0.5, reversed 0.5) and delay type with sum contrast coding (i. no delay -0.5, neutral 0.5; ii. associated -0.5, neutral 0.5; iii. no delay -0.5, associated 0.5). The key model estimate was the interaction of subject-object order and delay type. The model specification and priors were as above.

For each contrast (i, ii, iii), we conducted a sensitivity analysis to see how different priors would have affected our conclusions (Schad et al., 2020; *Appendix A1*). We used a range of priors that constrained the model to test for plausible interaction effect sizes based on the literature: these ranged from constraining priors which assumed only a narrow range of effects (mean of zero and standard deviation of 0.1 μV, meaning that we expected the effect to fall with 95% probability between -0.2 and $0.2\ \mu V$), to less constraining priors that assumed a broader range of effects (mean of zero and standard deviation of 2 μV, meaning that the effect could fall with 95% probability anywhere between -4 and $4\ \mu V$).

Results

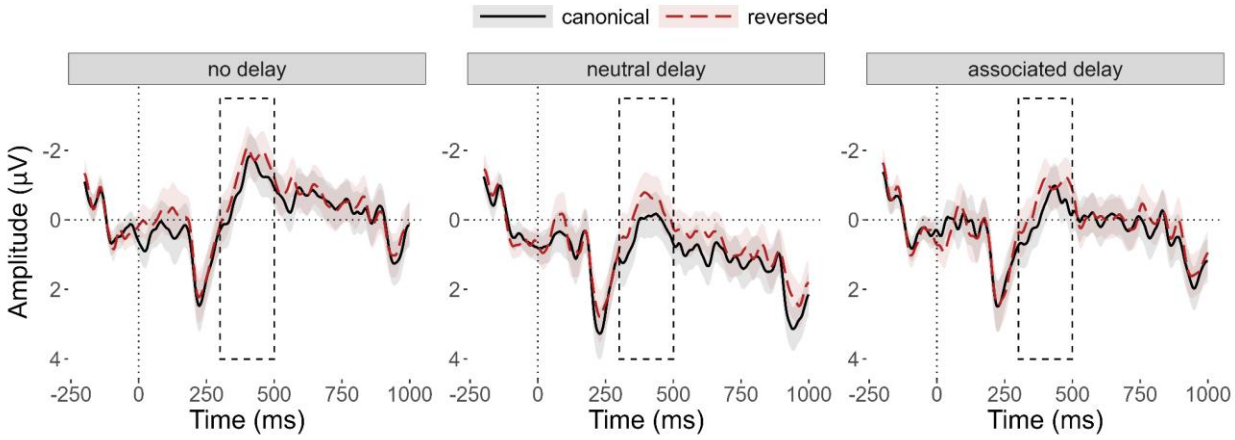
Mean amplitude of the ERP at the target word is plotted in Figure 2. Visual inspection suggested a semantic illusion (no difference in N400 amplitude between canonical and reversed sentences) in the no-delay condition. In the neutral delay condition, there was a small increase in N400 amplitude for role-reversed

[1] We also tested models using these predictors: The interaction and nested effect estimates were nearly identical and so for computational efficiency, we report only the models without the additional predictors.

sentences, suggesting the illusion may have been prevented. In the associated delay condition, amplitude was similar between the canonical and reversed sentences, but the N400 in the reversed condition was wider, possibly suggesting more latency variability between participants than in the canonical condition.

Figure 2

ERPs at the target verb



Note. Mean amplitude across electrodes Cz, CP1, CPz, CP2, P3, Pz, P4, PO3, POz, PO4 is plotted in the no-delay (e.g. *which guest the moderator* ***invited***), neutral delay (e.g. *which guest the moderator* additionally ***invited***), and associated delay conditions (e.g. *which guest the moderator* to the panel show ***invited***). The dashed box indicates the analysis time window.

Main model

The model was not consistent with an interaction of role order and delay in the neutral versus no delay comparison, $\hat{\beta} = -0.05\,\mu V, 95\%\,CrI\ [-0.39, 0.30]\,\mu V$. The proportion of posterior probability mass below zero for this comparison was only 61%, $P(\beta < 0) = 0.61$. The model was also not consistent with an interaction of role order and delay in the associated versus neutral delay comparison, $\hat{\beta} = -0.09\,\mu V, 95\%\,CrI\ [-0.44, 0.25]\,\mu V$, $P(\beta < 0) = 0.70$. Thus, the model was not consistent with readers having prevented the illusion in the neutral delay condition, from which the associated condition did not appear to differ. However, there was a clear main effect of role order with amplitude more negative for role reversed than canonical sentences across all delay conditions, $\hat{\beta} = -0.45\,\mu V, 95\%\,CrI = [-0.73, -0.17]\,\mu V$, $P(\beta < 0) = 1.00$. To determine which—if any—delay condition might have been driving this effect, we conducted nested comparisons.

Nested comparisons

To determine which of the delay conditions may have been driving the main effect of role order, we fit a model with simple difference contrast coding that estimated the effect of role order nested within each of the three delay conditions. The model was consistent with a larger N400 for reversed sentences within the neutral delay condition, $\hat{\beta} = -0.53\,\mu V, 95\%\,CrI = [-1.00, -0.05]\,\mu V$, $P(\beta < 0) = 0.99$, and to a lesser extent within the associated delay condition, $\hat{\beta} = -0.43\,\mu V$, $95\%\,CrI = [-0.93, 0.05]\,\mu V$,

$P(\beta < 0) = 0.96$. The model was not consistent with a larger N400 for reversed sentences in the no delay condition, although the majority of the posterior probability mass was compatible with an effect in this direction, $\hat{\beta} = -0.35\,\mu V, 95\%\,CrI = [-0.82, 0.13]\,\mu V, P(\beta < 0) = 0.93$. Thus it appeared that the main effect of role order was being driven by both delay conditions, despite the similarity in cloze probability differences across all three delay conditions.

Bayes factor analyses

In the subset of the data containing only the neutral and no delay conditions, the model was not consistent with an interaction of role order and delay type, $\hat{\beta} = -0.17\,\mu V, 95\%\,CrI = [-0.83, 0.49]\,\mu V$. The Bayes factor was 0.46, indicating no evidence either for or against the effect and that the data were insufficient to distinguish between hypotheses given the model and priors. The sensitivity analysis indicated different prior choices would not have yielded more conclusive evidence unless we had assumed that a broader range of effect sizes were also plausible—including relatively large effect sizes—in which case there would have been moderate evidence against the interaction (*Appendix A1*).

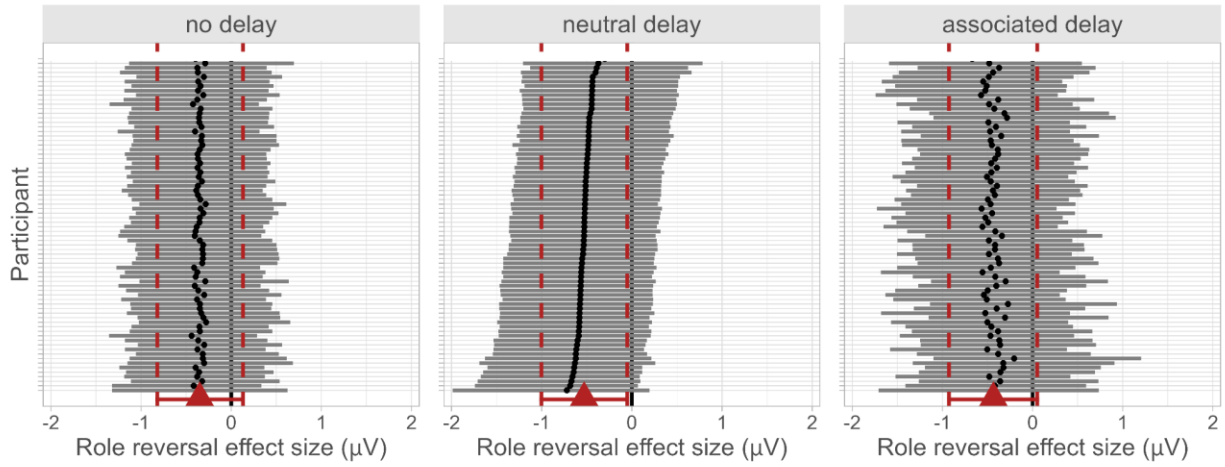
In the subset of the data containing only the associated and neutral delay conditions, the model was not consistent with an interaction of delay with role order, $\hat{\beta} = 0.09\,\mu V, 95\%\,CrI\ [-0.57, 0.73]\,\mu V$, $BF_{10} = 0.36$, nor was it in the subset of the data containing only the associated and no delay conditions, $\hat{\beta} = -0.09\,\mu V, 95\%\,CrI\ [-0.74, 0.54]\,\mu V$, $BF_{10} = 0.31$. The sensitivity analyses indicated inconclusive evidence if we had assumed a priori small effect sizes and moderate evidence against the interaction assuming a broader range of effect sizes (*Appendix A1*). Altogether, there was inconclusive evidence about whether the delay prevented the illusion and whether the neutral and associated delays differed in their ability to do so.

Individual differences

The inconclusiveness of the Bayes factors could suggest that small interaction effects may have been present that were obscured by a high degree of variability between participants. We therefore examined the distribution of effect sizes among individual participants by extracting the participant random effect estimates from the nested comparisons model (Figure 3). To examine whether there was any visual pattern in the by-participant estimates, we plotted the effects in all delay conditions sorted by effect size in the neutral condition, since this was the condition that was most consistent with a role reversal effect in the nested analysis: this allowed us to inspect whether participants who had larger role reversal effects in the neutral delay condition tended to show larger or smaller effects in the other conditions. This was not the case. Interestingly however, while the role order effect in the no delay condition was not statistically supported at the group level (red 95% CrI contained zero), all participants showed a small but relatively uniform increase in N400 amplitude for reversed sentences, irrespective of their effect size in the neutral delay condition. The group-level effect in the associated delay condition showed a similar pattern: all participants had a small increase in amplitude for reversed sentences. There was no clear visual relationship between effect sizes in the neutral and associated conditions, other than possibly more variability among effect sizes in the associated condition for participants who had larger effects in the neutral condition.

Figure 3

By-participant estimates of the main effect of role order within each delay condition



Note. The black points and error bars show each participant's mean estimated N400 difference between reversed and canonical sentences and 95% credible interval. A negative value indicates that the N400 was more negative in the reversed condition. The red triangles indicate the group estimate for each contrast.

Others have raised the importance of different features of the experimental stimuli in eliciting the illusion (Ehrenhofer et al., submitted; Li & Ettinger, 2023) and so we also examined the by-item effects. Since these were not directly relevant to the question in the main manuscript, we present them in *Appendix A3*. In sum, the pattern of effects was similar to that of the by-participant effects. In addition, since the small but consistent effect in our no delay condition was surprising, we confirmed that a similar pattern was observed in Chow et al. Experiment 3 (2018; *Appendix A4*).

Exploring features of the delaying fragment

Our "neutral" and "associated" delay conditions were planned as categorical predictors. What was considered neutral and associated was largely determined by cloze probability, with higher probability taken as indicating that the delaying fragment created stronger association with the target verb. However, we were agnostic about how the fragment induced this stronger association and individual stimuli differed in this respect. In this section we report exploratory analyses probing different features of the delaying fragment and their effect on whether or not an illusion was observed. First, we examine lexico-semantic similarity of the fragment with the context, independently of the fragment's propositional content. We then examine how much information the added delay fragment contributed over and above the no delay condition according to cloze probabilities—this measure did take into account the fragment's propositional content. Finally, we examine whether stronger contextual constraint in the canonical condition was more predictive of a canonical interpretation in the reversed condition (i.e., more likely to elicit an illusion), which also tested Chow et al.'s (2018) claim that any benefit of a delay should be restricted to highly predictable contexts.

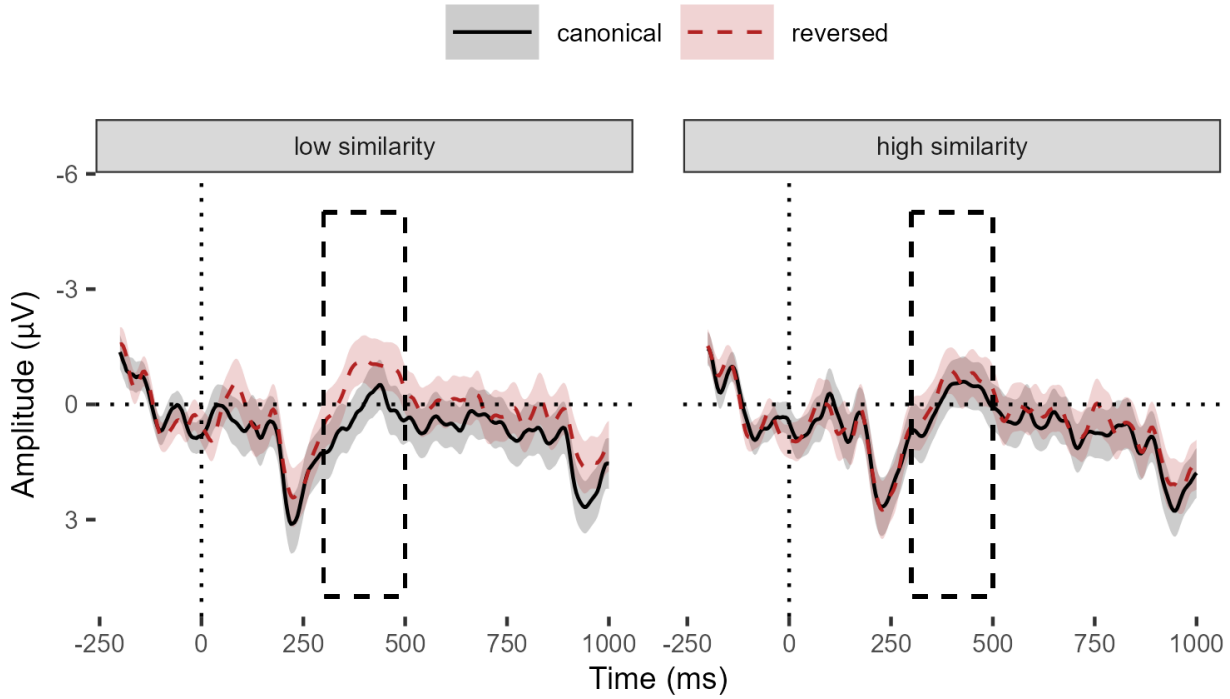
Semantic associatedness of fragment and context

Cosine similarity is a numerical value reflecting the degree of semantic similarity of two words based on how often they occur in the same contexts. We computed cosine similarities using the R package *LSAfun* (Günther et al., 2015) with a German latent semantic analysis (LSA) space of 300 dimensions (Günther, 2022) trained on the 1.7 billion-word deWaC corpus (Baroni et al., 2009). We computed cosine similarity between each delay fragment and other regions of the sentence, including one or both of the verbal arguments, the context, the verb, and various combinations of these elements. Then, in separate models, we replaced the categorical "delay type" predictor with continuous cosine similarity. Additional predictors were added to account for the effects of cloze probability and—since these were now not matched—target word position, the number of words in the delaying fragment, and the number of noun phrases in the delaying fragment. We used the same priors as for the main analysis. The posteriors for the interaction effect from each model are presented in **Figure A5** in *Appendix 5***.** Summary statistics of the cosine values for each combination of sentence regions can be found in *Appendix 6*.

The cosine similarity that best predicted a larger N400 in the reversed versus canonical condition was the similarity between the first noun phrase after *which* and the delay fragment ("delay-NP1"); in the example in **Table 1**, this would mean the similarity between *to the panel show* or *additionally* and *guest*. That is, the more semantically similar the delay fragment was with the linearly first verbal argument (the stereotypical patient of the verb), the more likely it was that the illusion occurred, $\hat{\beta} = 0.39, 95\%\ CrI = [0.02, 0.78], P(\beta > 0) = 0.98$. All posteriors were suggestive of an effect in the same direction. Taking the similarity of the first noun phrase and the delay fragment as the best predictor, we then re-plotted the grand average ERPs in **Figure 4**, categorising the cosine values into low and high categories via median split. The N400 in the reversed condition is now clearly absent when cosine similarity is high—comparable to the previous "associated" delay condition—but present when similarity is low—comparable to the previous "neutral" delay condition.

Figure 4

Grand average ERPs averaged by median-split cosine similarity of the delay and the first verbal argument



Note. The grand average was computed across electrodes Cz, CP1, CPz, CP2, P3, Pz, P4, PO3, POz, PO4. Low indicates low cosine similarity of the delay fragment and the first verbal argument, comparable with the "neutral" category in the main analysis. High indicates high similarity, comparable with the "associated" category. The dashed box indicates the analysis time window.

The additional verb biasing effect of the delaying fragment

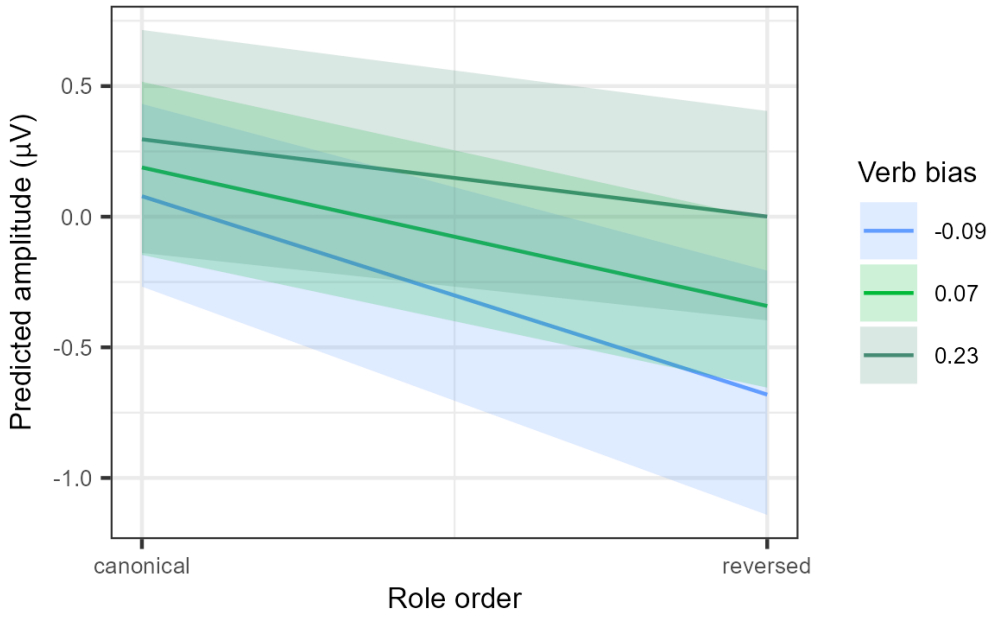
For this analysis, we used cloze probabilities to compute how much more likely cloze test participants were to give the target verb in the neutral vs. no delay condition and associated vs. no delay conditions. A "bias" value was computed by subtracting the cloze probability of, e.g., the neutral delay canonical condition from the no delay canonical condition and the neutral delay reversed condition from the no delay reversed condition. The same was computed for the associated vs. no delay conditions. The resulting value gave us the biasing effect of the fragments alone, all else being equal. Weak bias indicated that the fragment did not change readers' expectations about the target verb relative to the no delay condition; strong bias indicated that it did. The neutral fragment increased the probability of the target verb by a mean of 0.02 (SD = 0.15) in the canonical condition and 0.03 (SD = 0.08) in the reversed condition. The associated fragment increased the probability of the target verb by 0.12 (SD = 0.21) in the canonical condition and 0.11 (SD = 0.15) in the reversed condition.

As for the cosine similarity analysis, we replaced the categorical delay type predictor with the continuous bias predictor. All other model aspects were the same as the cosine similarity model. While the interaction of role order and bias did not meet our criteria for being statistically consistent with an effect, there was a numerical trend indicating that strongly biasing fragments had less of an effect on the N400 in the reversed condition than weakly biasing fragments, $\hat{\beta} = 0.26, 95\%\ CrI = [-0.15, 0.67], P(\beta > 0) =$

0.89. In other words, as can be seen in **Figure 5**, fragments that biased more toward the target verb (and thus toward the canonical interpretation) reduced the difference in N400 amplitude between canonical and reversed conditions.

Figure 5

Marginal effects of the model of verb bias relative to the no delay condition



Note. Verb bias is split into three colours, where values of around zero indicate weak bias and larger positive values indicate stronger bias.

The delay effect and strength of bias toward the canonical interpretation

Chow et al. (2018) claim that the benefit of delaying the verb is only apparent for role reversals where the verb in the canonical condition is highly predictable. This claim can be further distilled into the effects of predictability of the specific target verb and of how strongly the context constrains toward any one word, not necessarily the target. We therefore analysed the effect of cloze probability (predictability of target verb) and computed entropy among the responses given in the cloze test (contextual constraint). Entropy is a measure of uncertainty that quantifies the distribution of responses given by cloze test participants at the target word position. Values closer to zero indicate lower uncertainty and thus a more constraining context, independently of the identity of the target verb. The higher the value above zero, the more uncertainty and thus weaker constraint. Summary statistics of the entropy values by condition are in **Table 3**.

Table 3

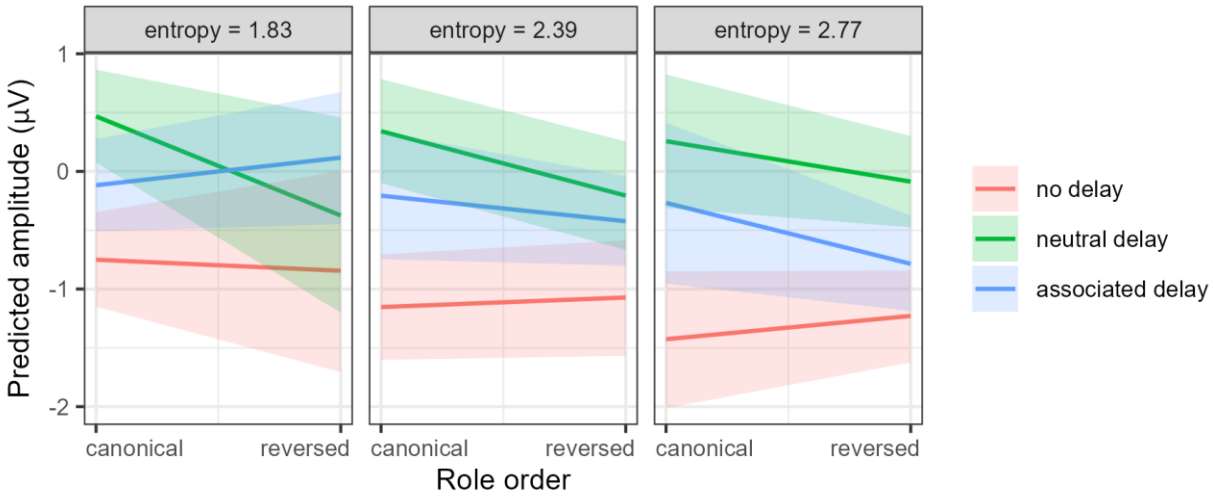
Summary entropy statistics

	Mean (bits)	SD
(a) No delay, canonical	1.98	0.55
(b) No delay, reversed	2.77	0.36
(c) Neutral delay, canonical	1.95	0.58
(d) Neutral delay, reversed	2.70	0.34
(e) Associated delay, canonical	1.58	0.61
(f) Associated delay, reversed	2.52	0.48

Entropy was added to the main model used in the planned analysis to estimate the 3-way interaction of role order, delay type, and constraint. All other model specifications remained the same. We used contrast coding to test the effect of role order in between the neutral and no delay conditions, between the associated and no delay conditions, and between the associated and neutral conditions. The model was consistent with a 3-way interaction in the neutral vs. no delay comparison, $\hat{\beta} = 0.44, 95\%\, CrI = [-0.05, 0.92], P(\beta > 0) = 0.96$. The 3-way interaction in the associated vs. no delay comparison did not meet our criteria for being consistent with an effect, although the bulk of the posterior favoured a positive effect, $\hat{\beta} = 0.37, 95\%\, CrI = [-0.12, 0.86], P(\beta > 0) = 0.93$. The 3-way interaction in the associated vs. neutral comparison was not consistent with an effect, $\hat{\beta} = 0.05, 95\%\, CrI = [-0.49. 0.58], P(\beta > 0) = 0.56$. **Figure 6** shows the interactions in graphical form: The reversed vs. canonical effect does not appear to differ with increasing constraint in the no delay condition (pink). The neutral delay (green) appears to benefit readers most in more constraining sentences, with the N400 most negative in the reversed condition at the lower quartile of entropy values. The effect of the associated delay (blue) appears to benefit readers in less constraining contexts (upper quartile of entropy values).

Figure 6

Marginal effects of the 3-way interaction of role order, delay type, and entropy

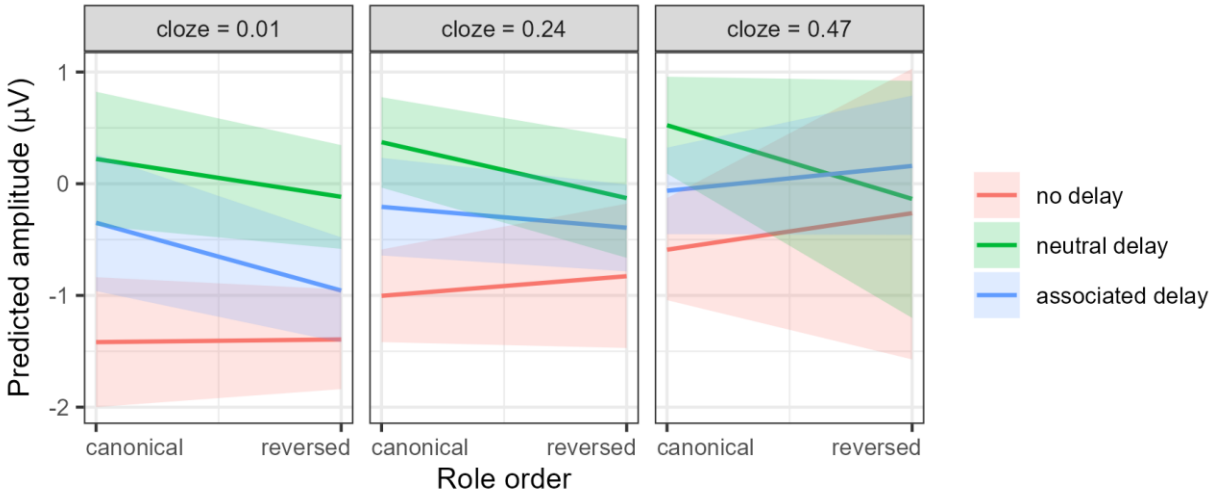


Note. Entropy is split into three panels reflecting the lower quartile (left), median (middle) and upper quartile (right).

The same model was then fit using cloze probability of the target word instead of entropy. Visually, the neutral delay appeared to benefit readers more in the highest predictability sentences (**Figure 7**), although the effect was not statistically convincing, $\hat{\beta} = -0.29, 95\%\, CrI = [-0.77. 0.19], P(\beta < 0) = 0.88$. Cloze probability did not appear to affect the interaction of role order in the associated vs. no delay comparison, $\hat{\beta} = -0.13, 95\%\, CrI = [-0.66, 0.39], P(\beta < 0) = 0.69$. The 3-way interaction in the associated vs. neutral comparison was also not consistent with an effect, $\hat{\beta} = -0.11, 95\%\, CrI = [-0.71, 0.47], P(\beta < 0) = 0.65$.

Figure 7

Marginal effects of the 3-way interaction of role order, delay type, and cloze probability



Note. Cloze probability is split into three panels reflecting the lower quartile (left), median (middle) and upper quartile (right).

Meta-analysis of delay effects

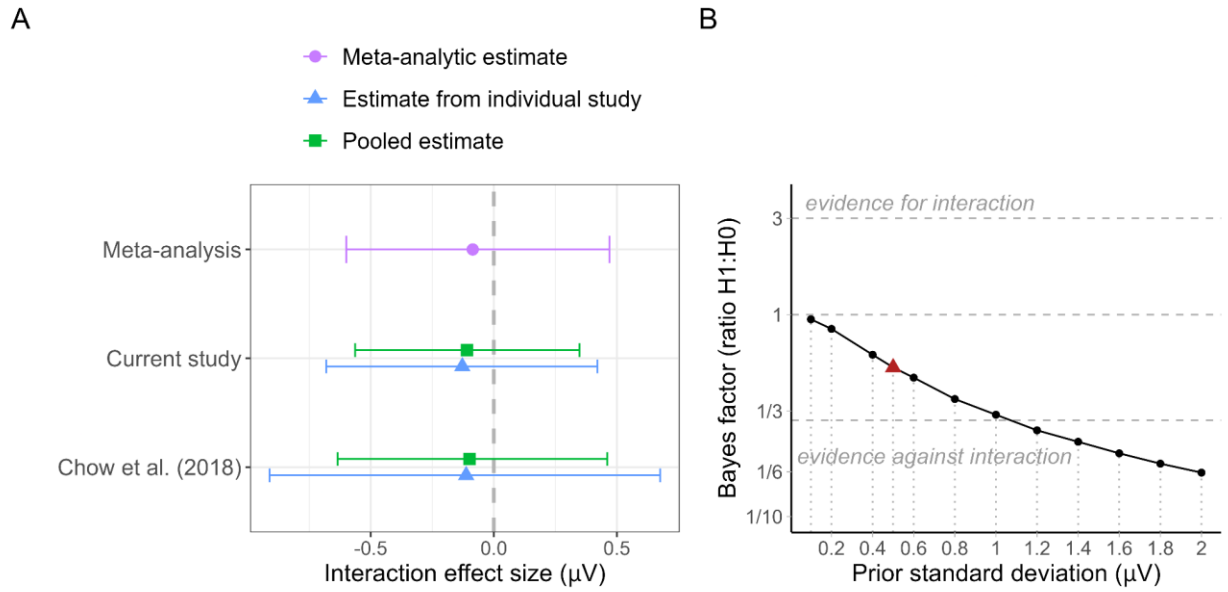
Although the planned analysis yielded inconclusive support for an interaction of role order and the delay, the small estimated effect size was at least consistent in sign with the previous study in which an effect was observed (Experiment 3, Chow et al., 2018). That experiment included 24 subjects and 120 experimental items containing four conditions, meaning 30 sentences per condition per subject. We therefore quantified whether the pooled data might yield more conclusive evidence for the effect. We undertook the meta-analytic approach outlined in Nicenboim et al. (2020; see also Jäger et al., 2017; Vasishth et al., 2013), in which separate models are fit to each study's data and a meta-analytic estimate is obtained by modelling the individual estimates in an intercepts-only model with study identifier as a random effect. The meta-analytic estimate of the interaction effect therefore took into account the differing variances of the two studies.

As Chow et al.'s study only had the equivalent of the neutral delay condition, only the neutral delay condition from our study was considered in the meta-analysis. We note also that to keep the analyses consistent between both studies, we reanalysed Chow et al.'s data with a method different to that used in their paper, using only the posterior region of electrodes (to be consistent with the region of the effect identified as significant in their analysis), and so the estimates we report here are different to (but consistent with) their paper. We used a prior of $\boldsymbol{Normal(0, 0.5)}$ to reflect the range of effect sizes observed in the current study ($\boldsymbol{-0.05\ \mu V}$) and Chow et al. ($\boldsymbol{< 1\ \mu V}$). While these effect signs were all negative, they stem from only two datasets and so the non-truncated prior encoded the possibility that the true effect could also be positive. The results of the meta-analysis are in Figure 8A. A sensitivity analysis was additionally conducted to see whether other prior choices would have changed our conclusion (Figure 8B).

Given that only data for two studies was available, it is perhaps not surprising that the meta-analytic estimate still did not yield conclusive evidence in favour the interaction, $\widehat{\boldsymbol{\beta}} = -\mathbf{0.09}\ \boldsymbol{\mu V}, \mathbf{95\%}\ \boldsymbol{CrI}\ [-\mathbf{0.60}, \mathbf{0.47}]\ \boldsymbol{\mu V}, \boldsymbol{BF_{10}} = \mathbf{0.55}$. With only two studies, publication bias is an issue with this meta-analysis, however we present the analysis to demonstrate the current state of evidence about the interaction effect and provide the code so that future researchers may contribute.

Figure 8

Results of the mini meta-analysis of the role reversal vs. delay type interaction



Note. **A.** The purple circle represents the meta-analytic estimate considering both the current study and Chow et al. (2018). The blue triangles represent estimates based on the original study data. The green squares represent estimates assuming shrinkage to the meta-analytic mean. **B.** Bayes factors for the ratio of evidence of a model with the meta-analytic estimate (H1) vs. a model without (H0). The red triangle indicates the prior used for the meta-analysis. The plot demonstrates that if we had assumed a priori that the interaction effect size was larger than 1 μV, the combined data would have yielded moderate evidence against the effect. For smaller effect size assumptions, evidence was inconclusive.

Discussion

We examined how delaying the verb in sentences such as "*which cop the thief arrested*" might help readers to avoid the N400 semantic illusion that "*arrested*" is a plausible verb even when "*thief*" is the agent. We tested two types of delay: a neutral delay where the verb was delayed with information that was congruent with the preceding sentence but did not add any additional semantic cues about the agent/patient roles, e.g., "*that evening*", and an associated delay which contained words that biased the interpretation towards the event described in the canonical sentence, e.g., "*with handcuffs*". The findings suggest that the neutral delay enabled participants to avoid the illusion, consistent with Chow et al. (2018). That is, when the verb appeared immediately, there was an N400 semantic illusion; when it was delayed, the results were

consistent with a larger N400 for reversed sentences, suggesting the delay—particularly the neutral delay—contributed to readers avoiding the illusion. We were also able to demonstrate in the no delay condition that the semantic illusion can be elicited in German, despite the fact that in the single word presentation paradigm, readers had access to unambiguous thematic role information via case marking on the determiner before seeing the verbal arguments and the verb. Moreover, the differing effect of the role reversal across delay conditions was present despite the difference in cloze probability of the verb being similar between canonical and role reversed sentences in each delay condition.

Our conclusions about the ability of the delay to prevent an illusion take into consideration the results of the nested comparisons and the similarity of the effect size and direction in the meta-analysis with Chow et al.'s (2018) data. Our conclusion that the neutral delay was better able than the associated delay to prevent the illusion is based on the nested comparisons as well as the cosine similarity analysis, which showed that higher semantic association between the delay fragment and the first verbal argument was more likely to sustain the illusion despite the delay. Similarly, fragments that biased more strongly toward the canonical interpretation and thus the presented target verb also appeared more likely to sustain the illusion, although statistical support was weak. Given the inconclusive result at the level of the interactions, we draw these conclusions tentatively and hypothesise that the interaction effect size is likely too small to yield conclusive evidence for an interaction at the current sample size (Gelman, 2020; see also the prospective power analysis in *Appendix 3*). Based on our conclusions, we will propose an account of thematic role assignment as a continuous process that can be swayed by lexico-semantic cues.

One surprising finding was that of the small, consistent increase in N400 amplitude in the reversed condition for every participant within the no delay condition. We conclude that there was an illusion at the group level despite this finding since the same pattern of individual results was also observed in a previous study observing an illusion (Chow et al., 2018; see *Appendix 4*). We do not know if this pattern across individual participants may actually be present in all previous studies also observing the semantic illusion (Hoeks et al., 2004; Kim & Osterhout, 2005; Kolk et al., 2003; Kuperberg et al., 2003; van Herten et al., 2005, 2006). Thus, we see our data as consistent with the illusion as it has been reported in previous studies. At the same time, the by-participant pattern could be seen as calling into question the completeness of the illusion effect or at least as raising the question of how big an N400 effect has to be before it is considered cognitively meaningful. Previous studies have not reported on individual differences, thus it would be interesting to see whether future studies (or re-analyses of previous studies) demonstrate the same pattern or whether it is unique to the two studies using the delaying sentence fragment (the current study and Chow et al., 2018).

Implications for how the N400 semantic illusion arises and how the delay prevents it

The difference between the neutral and associated delay effects in the current experiment presents an interesting test of theoretical accounts about the cognitive origins of the N400 semantic illusion. We discuss each of these accounts now. First, in the *Introduction* we tentatively proposed that the SG model (Rabovsky et al., 2018) may be able to accommodate the delay effect and the difference between the neutral and associated delay effects in the following way: Under the SG model, the initial illusion arises due to uncertainty between semantic cues based on real-world event probabilities and syntactic cues like word order. The event is misinterpreted based on semantic relationships and *'arrested'* perceived as plausible. The illusion may therefore be avoided because the neutral delay corroborates the syntactic structure,

strengthening the syntactic attractor which then has more influence in constraining the event representation toward the literal interpretation. In contrast, the associated delay corroborates the syntactic structure but additionally strengthens the semantic attractor that led to the illusion in the first place, increasing uncertainty and yielding greater variability in the observed N400 effect.

Such an account may be supported by an analysis reported in Appendix 7, which shows that the neutral but not the associated delay effect increased as the experiment progressed. It may therefore be that the repeated presentation of sentences with the same structure further increases attention to structural cues, which in the absence of any lexically associated material, strengthens the syntactic attractor relative to the semantic attractor. This could predict that manipulating the ratio of fillers to critical items in the experiment influences whether the neutral delay effect increases or not. Elsewhere, increasing the ratio of semantic cues has been observed to affect N400 amplitude (Brothers et al., 2015, 2017; Lau et al., 2013), so it is conceivable that a structural cue that facilitates interpretation could have a similar effect. However, strengthening of the syntactic attractor is a tentative hypothesis.

An alternative possibility, as Liao et al.'s (2022) account proposes and which is equally compatible with the core ideas of the SG model, is that the additional time provided by the neutral delay is the critical factor, not the additional information. Indeed, in forthcoming work (Stone & Rabovsky, in preparation), we test different methods of temporally delaying the verb and show that this is sufficient to prevent the illusion without any additional sentence material, replicating and extending Nakamura et al.'s (2024) findings. Under the SG account, time may help the model to move towards resolving the conflict between semantic and syntactic cues (explaining the N400 effect in the neutral delay condition), while the associated delay, which biases towards the canonical interpretation, enhances the conflict (explaining the lack of an N400 effect in the associated delay condition). The effect of the associated delay would thus be to counteract the influence of additional time by reinforcing the canonical plausibility-based interpretation via strengthening the semantic attractor. As noted earlier, even though the current version of the SG model implements discrete per word updates, this is just a simplification, and the idea that updates are continuous and time thus has an impact on SG representations is fully compatible with the ideas underlying the model. Work to adapt the SG model to capture these effects of time alone is currently underway.

Liao et al.'s (2022) three-stage account could also account for the difference between the two delay conditions. In their account, a delay provides the time for role assignment to be completed and constrain verb prediction, which at first blush would seem to predict that both the neutral and the associated delays in the current study should have prevented the illusion. However, in their discussion, Liao et al. do hypothesise that additional cues in the context or discourse that weight the event toward the canonical interpretation could prolong the implementation of the three processing steps, although they did not explicitly test this hypothesis. Our results in the associated delay condition—as well as the cosine similarity and bias analyses suggesting that anything about the fragment that strengthened the canonical interpretation sustained the illusion in the reversed condition—would therefore appear to provide a successful test of their hypothesis. Our exploratory analyses also provide support for the claim by Chow et al. (2018) that the neutral delay only benefits readers in high predictability contexts, although we qualify this further to show that the delay offers the most benefit in highly constraining contexts and that predictability of a specific target word may be less crucial.

With respect to other existing models of the role reversal effect, our successful replication of Chow et al.'s (2018) delay effect, as well as the exploratory analyses underlining the important role of context, confirm that the processes that lead to the illusion at the verb are already underway before it is presented. This potentially rules out a number of accounts in which the illusion results from processing that is triggered

by presentation of the verb (Bornkessel-Schlesewsky & Schlesewsky, 2008; Kolk et al., 2003; Kuperberg, 2007; van Herten et al., 2005, 2006). Accounts either already incorporating a pre-verbal mechanism or that could reasonably be adjusted to do so will be challenged to explain how the delaying sentence fragment—identical across both canonical and reversed conditions—could change the N400 in the reversed condition only (Bornkessel-Schlesewsky & Schlesewsky, 2008, 2019; Brouwer et al., 2017; Kim & Osterhout, 2005; Li & Ettinger, 2023).

Specifically, the delay finding seems difficult to explain if the reduced N400 in the reversed condition is the result of priming at the word level as assumed by Brouwer et al. (2017) rather than the result of a transient illusion at the level of sentence meaning as assumed by the SG model (Rabovsky et al., 2018). This is because the priming effect should be equally influenced by the delaying fragments in both the canonical and the corresponding reversed conditions, i.e., in both canonical and reversed conditions the neutral delay should reduce the priming while the associated delay should restore the priming effect. Thus under a priming account the difference in N400 amplitude between the canonical and reversed conditions should stay the same across the no delay, neutral delay and associated delay conditions. On the other hand, explaining the reduced N400 as resulting from a temporary illusion at the level of sentence meaning can explain how the neutral delay might help to resolve the illusion and thus increase the difference in interpretation between the canonical and reversed conditions, increasing the difference in N400 amplitude, while the associated delay, which provides a bias towards the canonical interpretation, should reduce the difference in interpretations between the canonical and reversed conditions and reduce the difference in N400 amplitudes. While this provides a tentative hypothesis for how the SG model (Rabovsky et al., 2018) might explain our findings, this is currently only a verbally descriptive account; a computational implementation remains the subject of ongoing work.

The finding that the semantically associated delay appeared more likely among individual participants to sustain the illusion than the neutral delay is reminiscent of findings in anomaly detection research.[2] Participants have been found to be less likely to notice a semantic anomaly if it is found within information that is relevant for completing a task (Barton & Sanford, 1993) or semantically related to the global context (Sanford et al., 2011). Sanford & Garrod (1998) propose that this is because early processing in comprehension attempts to relate new information with the existing discourse ("scenario mapping"); if mapping is possible, deeper processing will not occur. Semantic association of the target word with the context provides the illusion of mapping, even if implausible, and so deeper processing does not occur. This idea is echoed by the first 'bag of words' stage of processing in Liao et al.'s (2022) three-stage argument-verb processing theory, although deeper processing in their account is prevented by time constraints rather than an illusion of successful mapping. However, even without time constraints, participants in anomaly detection experiments fail to detect anomalies in the presence of semantic information, echoing our findings and highlighting the strong, ongoing influence of lexico-semantic cues on processing language.

One difference between the anomaly detection literature and the role reversal paradigm is that the majority of participants in role reversal experiments report having noticed the anomaly (Chow et al., 2018; Ehrenhofer et al., submitted; Hoeks et al., 2004; Kim & Osterhout, 2005; Kolk et al., 2003; Kuperberg et al., 2003). Detection of the anomaly is further supported by the presence of a P600 in those experiments, which suggests that any "illusion" in the N400 time window is resolved in the P600 window such that the literal interpretation is derived. This could suggest that an account where shallow processing extends beyond the verb is not a good explanation of the current results. We note that we did not include plausibility

[2] We thank an anonymous reviewer for highlighting this line of work.

judgements in our design and previous role reversal studies did not contain a semantically associated delay condition, so it is impossible to know whether the associated condition affected the proportion of readers who detected the anomaly. However, an increased probability of giving the canonical verb in the reversed condition—where it was implausible—was apparent among individual items from the cloze testing of the current study. On the other hand, we did not observe a P600, which we attribute to our in-experiment task: Our yes/no comprehension questions perhaps did not draw participants' attention to the semantic plausibility of the sentences as strongly as previous studies, which may be a necessary condition for eliciting a P600 effect (Li & Ettinger, 2023; cf. Van Herten 2005; 2006 who observed the effect with content questions only). A future study could address this limitation by including plausibility judgements to quantify whether the associated delay affected readers' ability to detect the anomaly.

Overall, the delay effect and the differences between delay content observed in the current experiment thus allow us to constrain our hypotheses about the cognitive processes underlying the N400 semantic illusion. The accounts that seem to best accommodate the delay effect suggest that either the illusion results from uncertainty about the correct interpretation due to conflicting syntactic and semantic information (Rabovsky et al., 2018), or from thematic roles being slower to constrain verb predictions than semantic associations (Liao et al., 2022).

The initial illusion occurs even with early structural cues to role information

The finding of a role reversal illusion in the no delay condition conceptually replicates the illusion seen in previous studies (Chow et al., 2015, 2018; Hoeks et al., 2004; Kim & Osterhout, 2005; Kolk et al., 2003; Kuperberg, 2007; Kuperberg et al., 2003; Liao et al., 2022; Nakamura et al., 2024; van Herten et al., 2005, 2006). This is—to our knowledge—the first demonstration of the role reversal semantic illusion in German, which is notable because the pre-nominal morphological marking of case in German means readers have earlier access to thematic role information than readers in previously tested languages with post-nominal or no case marking. This fact becomes particularly relevant in the single-word presentation mode of ERP experiments and especially in light of evidence that thematic roles are assigned incrementally by German readers (Bornkessel et al., 2003; Friederici et al., 1998; Frisch & Schlesewsky, 2001; Haupt et al., 2008; Schlesewsky & Bornkessel, 2006). Neither of these factors appeared to make German readers immune to the illusion however—even as the experiment progressed (Figure A7)—highlighting the strong influence of semantic expectation on processing even in the presence of early, unambiguous syntactic cues.

The incremental processing of these structural role cues in German may relate to why it was particularly the cosine similarity of the first verbal argument to the delay fragment that sustained the illusion. The unambiguous case marking of this argument would already constrain predictions about both the role of any upcoming noun, as well as lexico-semantic predictions about the plausible identity of the agent and verb. Cosine similarities including other regions of the sentence were less predictive of this effect (although all tended in the same direction), possibly suggesting that the first verbal argument was the strongest driver of predictions. Even the unambiguous case marking may have been overridden by the strong bias toward the canonical verb if there was a strong link between this argument and the associated delay fragment. We also note that cosine similarity does not purposefully reflect whether *thief* is often found as an agent or patient in events that include words in the delay fragment, only the frequency of their co-occurrence, although this information may be implicitly encoded via the frequency of *thief* as an agent or patient in general. Ehrenhofer et al. (submitted) have indeed suggested that stereotypical agent- or patienthood of the verbal arguments is important to eliciting the N400 semantic illusion. To test the relative contributions of case marking or lexical identity of the first verbal argument, a future study could swap the

linear position of the agent and patient rather than the case marking, or test the effect in languages with less case marking but stronger word order cues.

Do individual readers even experience an illusion?

One striking feature of the current data is that although the posterior estimate for the no delay condition did not meet the threshold for being considered consistent with an N400 effect, the posterior probability distribution was shifted in the direction of a small increase in amplitude for role reversed sentences; 93% of the posterior samples were less than zero. Moreover, every participant showed a small effect in the same direction. This may suggest that readers were able to use the syntactic cues to constrain interpretation of the sentence to some extent, but perhaps not (yet) sufficiently by the time the verb was presented to fully notice its incompatibility with the context. We examined individual effects in the data from Chow et al. (2018) and noticed a similar pattern: In the no delay condition, most participants showed a small role order effect but two showed null effects and one showed a small effect in the opposite direction; in the neutral delay condition, all participants showed a small role order effect (see **Figure A4** in the *Appendices*). This may suggest that the by-participant pattern in our data is not specific to German, but rather a common feature of role reversal manipulations, which would also be in line with the observation that the N400 amplitude in previous studies is often numerically somewhat larger in the role reversed as compared to the canonical condition, even though the difference does not reach significance (e.g., Chow et al., 2018; Hoeks et al., 2004; Kuperberg et al., 2003). In line with the temporal processing accounts discussed in this paper, the small effect likely reflects the beginning of a continuous process of settling into the literal interpretation, but this raises interesting questions for future research: Is the small N400 effect cognitively meaningful? How large does an N400 effect need to be to be meaningful? Do we need to reinterpret the initial illusion?

Conclusions

A temporary semantic illusion has been demonstrated in role reversal sentences, in which an implausible verb does not increase N400 amplitude. We sought to demonstrate that German readers—despite receiving early morphosyntactic cues to thematic roles in a word-by-word presentation paradigm—are still susceptible to the semantic illusion, and more importantly, to replicate a previous finding that delaying the verb can prevent the illusion. We demonstrate that the illusion could be elicited and that it could be prevented if the verb was delayed by neutral information, but that it was less likely to be prevented by information that was semantically related to the verb and/or its context, because it induced a bias towards the canonical interpretation. We interpret the findings with respect to the Sentence Gestalt model, which suggests that the initial illusion arises because readers make quick but uncertain and possibly incorrect interpretations based on semantic association and event probability, conflicting with syntactic cues. The additional pre-verbal information may strengthen the syntactic interpretation due to its compatibility with the previous syntactic cues or via providing additional time to resolve the conflict, as long as it contains no further semantic cues. The findings provide confirmatory evidence for a relatively new finding, which itself provides a challenge to some existing models of the N400 semantic illusion.

Data availability statement

All materials, data, and analysis scripts for the analyses reported in the manuscript can be found in the paper's Open Science Framework (OSF) repository: https://osf.io/qtek9/.

Author contribution statement

Conceptualisation: MR and KS; data curation: KS; formal analysis: KS; funding acquisition: MR; investigation: KS; methodology: MR and KS; project administration: KS; resources: MR; supervision: MR; visualisation: KS; writing - original draft: KS; writing - review and editing: MR and KS.

Acknowledgments

This work was supported by an Emmy Noether grant (RA 2715/2-1) awarded to Milena Rabovsky and by project B03 of the Collaborative Research Centre (SFB) 1287. We thank Jasmin Barthelmeß and Melissa Höger for help with data collection. We also thank Wing Yee Chow for sharing data and sentence stimuli, Yana Arkhipova for sharing German sentence stimuli, and Lioba Berndt and Antonia Heinrich for creating additional sentence stimuli. Last but definitely not least, we thank Colin Phillips and two anonymous reviewers for their detailed and insightful reviews of various versions of the manuscript.

References

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, *43*(3), 209–226. https://doi.org/10.1007/s10579-009-9081-4

Barton, S. B., & Sanford, A. J. (1993). A case study of anomaly detection: Shallow semantic processing and cohesion establishment. *Memory & Cognition*, *21*(4), 477–487. https://doi.org/10.3758/BF03197179

Bornkessel, I., Schlesewsky, M., & Friederici, A. D. (2003). Eliciting thematic reanalysis effects: The role of syntax-independent information during parsing. *Language and Cognitive Processes*, *18*(3), 269–298. https://doi.org/10.1080/01690960244000018

Bornkessel-Schlesewsky, I., Kretzschmar, F., Tune, S., Wang, L., Genç, S., Philipp, M., Roehm, D., & Schlesewsky, M. (2011). Think globally: Cross-linguistic variation in electrophysiological activity during sentence comprehension. *Brain and Language*, *117*(3), 133–152. https://doi.org/10.1016/j.bandl.2010.09.010

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2008). An alternative perspective on "semantic P600" effects in language comprehension. *Brain Research Reviews*, *59*(1), 55–73. https://doi.org/10.1016/J.BRAINRESREV.2008.05.003

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2019). Toward a Neurobiologically Plausible Model of Language-Related, Negative Event-Related Potentials. *Frontiers in Psychology*, *10*, 298.

Brain Products. (2020a). *actiChamp Plus* [Computer software]. Brain Products GmbH.

Brain Products. (2020b). *BrainVision Recorder* (Version 1.23.001) [Computer software]. Brain Products GmbH.

Brothers, T., Swaab, T. Y., & Traxler, M. J. (2015). Effects of prediction and contextual support on lexical processing: Prediction takes precedence. *Cognition*, *136*. https://doi.org/10.1016/j.cognition.2014.10.017

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

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. *Cognitive Science*, *41*(S6), 1318–1352. https://doi.org/10.1111/cogs.12461

Buerkner, P.-C. (2018). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Chow, W.-Y., Lau, E., Wang, S., & Phillips, C. (2018). Wait a second! Delayed impact of argument roles on on-line verb prediction. *Language, Cognition and Neuroscience*, *0*(0), 1–26. https://doi.org/10.1080/23273798.2018.1427878

Chow, W.-Y., Momma, S., Smith, C., Lau, E., & Phillips, C. (2016). Prediction as Memory Retrieval: Timing and Mechanisms. *Language Cognition & Neuroscience*, *44*(5). https://doi.org/10.1080/23273798.2016.1160135

Chow, W.-Y., Smith, C., Lau, E., & Phillips, C. (2015). A ‘ bag-of-arguments ’ mechanism for initial verb predictions. *Language, Cognition and Neuroscience*, *44*(5). https://doi.org/10.1080/23273798.2015.1066832

Ehrenhofer, L., Lau, E., & Colin Phillips. (submitted). *A possible cure for ‘N400 blindness’ to role reversal anomalies in sentence comprehension*.

Ferreira, F., Bailey, K. G. D., & Ferraro, V. (2002). Good-Enough Representations in Language Comprehension. *Current Directions in Psychological Science*, *11*(1), 11–15. https://doi.org/10.1111/1467-8721.00158

Friederici, A. D., Steinhauer, K., Mecklinger, A., & Meyer, M. (1998). Working memory constraints on syntactic ambiguity resolution as revealed by electrical brain responses. *Biological Psychology*, *47*(3), 193–221. https://doi.org/10.1016/S0301-0511(97)00033-1

Frisch, S., & Schlesewsky, M. (2001). The N400 reflects problems of thematic hierarchizing. *NeuroReport*, *12*(15), 3391–3394. https://journals.lww.com/neuroreport/abstract/2001/10290/the_n400_reflects_problems_of_thematic.48.aspx

Gelman, A. (2020). Linear or logistic regression with binary outcomes. *Statistical Modeling, Causal Inference, and Social Science*. https://statmodeling.stat.columbia.edu/2020/01/10/linear-or-logistic-regression-with-binary-outcomes/

Günther, F. (2022). *Homepage of Fritz Günther—Semantic Spaces*. https://sites.google.com/site/fritzgntr/software-resources/semantic_spaces

Günther, F., Dudschig, C., & Kaup, B. (2015). LSAfun—An R package for computations based on Latent Semantic Analysis. *Behavior Research Methods*, *47*(4), 930–944. https://doi.org/10.3758/s13428-014-0529-0

Haupt, F. S., Schlesewsky, M., Roehm, D., Friederici, A. D., & Bornkessel-Schlesewsky, I. (2008). The status of subject–object reanalyses in the language comprehension architecture. *Journal of Memory and Language*, *59*(1), 54–96. https://doi.org/10.1016/j.jml.2008.02.003

Hoeks, J. C. J., Stowe, L. A., & Doedens, G. (2004). Seeing words in context: The interaction of lexical and sentence level information during reading. *Cognitive Brain Research*, *19*(1), 59–73. https://doi.org/10.1016/j.cogbrainres.2003.10.022

Hyvärinen, A., Karhunen, J., & Oja, E. (2001). Adaptive and learning systems for signal processing, communications, and control. In *Independent component analysis* (Vol. 1, pp. 11–14). John Wiley & Sons, Inc. https://doi.org/10.1002/0471221317

Jäger, L. A., Engelmann, F., & Vasishth, S. (2017). Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis. *Journal of Memory and Language*, *94*, 316–339. https://doi.org/10.1016/j.jml.2017.01.004

Jasper, H. (1958). Report of the committee on methods of clinical examination in electroencephalography: 1957. *EEG and Clinical Neurophysiology*, *10*(2), 370–375.

Jeffreys, H. (1939). *Theory of Probability*. Oxford University Press.

Jung, T., Makeig, S., Westerfield, M., Townsend, J., Courchesne, E., & Sejnowski, T. J. (2001). Analyzing and Visualizing Single-Trial Event-Related Potentials. *Human Brain Mapping*, *185*(March), 166–185. https://doi.org/10.1002/hbm.1050

Kim, A., & Osterhout, L. (2005). The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of Memory and Language*, *52*(2), 205–225. https://doi.org/10.1016/j.jml.2004.10.002

Kolk, H. H. J., Chwilla, D. J., van Herten, M., & Oor, P. J. W. (2003). Structure and limited capacity in verbal working memory: A study with event-related potentials. *Brain and Language*, *85*(1), 1–36. https://doi.org/10.1016/S0093-934X(02)00548-5

Kuperberg, G. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research, 1146*, 23–49. https://doi.org/10.1016/j.brainres.2006.12.063

Kuperberg, G., Sitnikova, T., Caplan, D., & Holcomb, P. J. (2003). Electrophysiological distinctions in processing conceptual relationships within simple sentences. *Cognitive Brain Research*, *17*(1), 117–129. https://doi.org/10.1016/S0926-6410(03)00086-7

Kutas, M., & Federmeier, K. D. (2011). Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology*, *62*(1), 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. https://doi.org/10.1126/science.7350657

Lau, E., Holcomb, P. J., & Kuperberg, G. (2013). Dissociating N400 effects of prediction from association in single-word contexts. *Journal of Cognitive Neuroscience*, *25*(3), 484–502. https://doi.org/10.1162/jocn_a_00328

Lee, M., & Wagenmakers, E.-J. (2014). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press. https://doi.org/10.1017/CBO9781139087759

Li, J., & Ettinger, A. (2023). Heuristic interpretation as rational inference: A computational model of the N400 and P600 in language processing. *Cognition*, *233*, 105359. https://doi.org/10.1016/j.cognition.2022.105359

Liao, C.-H., Lau, E., & Chow, W.-Y. (2022). Towards a processing model for argument-verb computations in online sentence comprehension. *Journal of Memory and Language*, *126*, 104350. https://doi.org/10.1016/j.jml.2022.104350

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324. https://doi.org/10.3758/s13428-011-0168-7

McElreath, R. (2015). *Statistical Rethinking: A Bayesian Course with Examples in R and STAN*. CRC Press.

Nakamura, M., Momma, S., Sakai, H., & Phillips, C. (2024). Task and Timing Effects in Argument Role Sensitivity: Evidence From Production, EEG, and Computational Modeling. *Cognitive Science*, *48*(12), e70023. https://doi.org/10.1111/cogs.70023

Nicenboim, B. (2018). *eeguana: A package for manipulating EEG data in R* [Computer software]. https://github.com/bnicenboim/eeguana

Nicenboim, B., Vasishth, S., & Rösler, F. (2020). Are words pre-activated probabilistically during sentence comprehension? Evidence from new data and a Bayesian random-effects meta-analysis using publicly available data. *Neuropsychologia*, *142*, 107427. https://doi.org/10.1016/j.neuropsychologia.2020.107427

R Core Team. (2020). *R: A Language and Environment for Statistical Computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org

Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, *2*(9), 693. https://doi.org/10.1038/s41562-018-0406-4

Sanford, A. J., & Garrod, S. C. (1998). The role of scenario mapping in text comprehension. *Discourse Processes*, *26*(2–3), 159–190. https://doi.org/10.1080/01638539809545043

Sanford, A. J., Leuthold, H., Bohan, J., & Sanford, A. J. S. (2011). Anomalies at the Borderline of Awareness: An ERP Study. *Journal of Cognitive Neuroscience*, *23*(3), 514–523. https://doi.org/10.1162/jocn.2009.21370

Schad, D. J., Betancourt, M., & Vasishth, S. (2020). Toward a principled Bayesian workflow: A tutorial for cognitive science. *Psychological Methods*. https://doi.org/10.1037/met0000275

Schlesewsky, M., & Bornkessel, I. (2006). Context-sensitive neural responses to conflict resolution: Electrophysiological evidence from subject–object ambiguities in language comprehension. *Brain Research*, *1098*(1), 139–152. https://doi.org/10.1016/j.brainres.2006.04.080

Schlesewsky, M., & Bornkessel-Schlesewsky, I. D. (2009). When semantic P600s turn into N400s: On cross-linguistic differences in online verb-argument linking. In *Papers from Brain Talk. The 1st Birgit Rausing Language Program Conference in Linguistics* (pp. 75–97). Lund University, Media Tryck. https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp

Schlesewsky, M., Fanselow, G., Kliegl, R., & Krems, J. (2000). The Subject Preference in the Processing of Locally Ambiguous WH-Questions in German. In B. Hemforth & L. Konieczny (Eds.), *German Sentence Processing* (pp. 65–93). Springer Science & Business Media. DOI 10.1007/978-94-015-9618-3

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y

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. *Journal of Cognitive Neuroscience*, *18*(7), 1181–1197. https://doi.org/10.1162/jocn.2006.18.7.1181

van Herten, M., Kolk, H. H. J., & Chwilla, D. J. (2005). An ERP study of P600 effects elicited by semantic anomalies. *Cognitive Brain Research*, *22*(2), 241–255. https://doi.org/10.1016/j.cogbrainres.2004.09.002

Vasishth, S., Malsburg, T. von der, & Engelmann, F. (2013). What eye movements can tell us about sentence comprehension. *WIREs Cognitive Science*, *4*(2), 125–134. https://doi.org/10.1002/wcs.1209

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

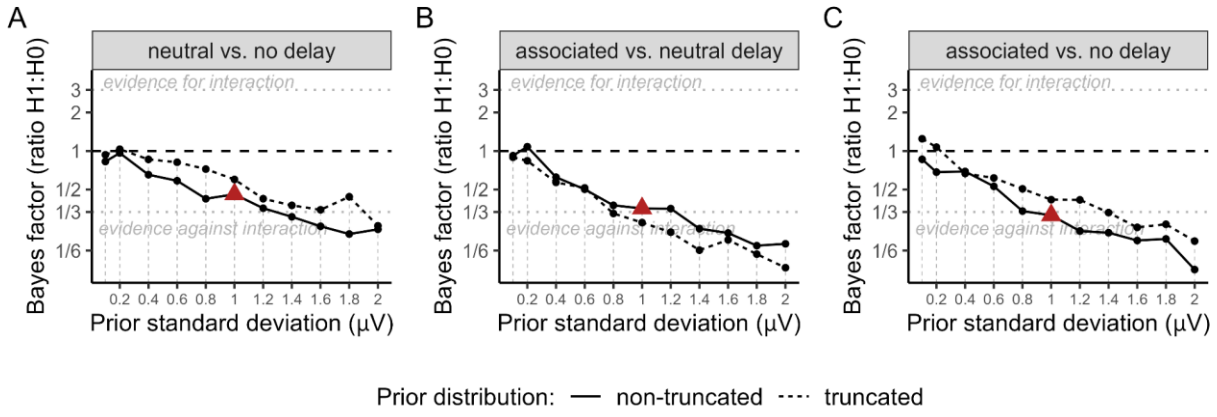
Appendices

A1. Bayes factor sensitivity analyses

The Bayes factors for the main analyses were based on a prior of $Normal(0, 1)$, which assumed that the interaction effect could have either a positive or negative sign and would lie with 95% probability between -4 and 4 μV. Since the Bayes factor is sensitive to the choice of prior, we conducted a sensitivity analysis with a range of priors (Schad et al., 2020). These included truncated priors, which assumed the interaction effect would have a negative sign as suggested by previous research (Chow et al., 2018). Figure A1 shows that unless we had a priori assumed interaction effect sizes of $1\mu V$ or more, the Bayes factor would still have indicated inconclusive evidence for an interaction of delay type and role order. For larger effect sizes, the Bayes factor would have indicated moderate evidence against an interaction.

Figure A1

Bayes factor sensitivity analyses for the interaction of delay type and role order



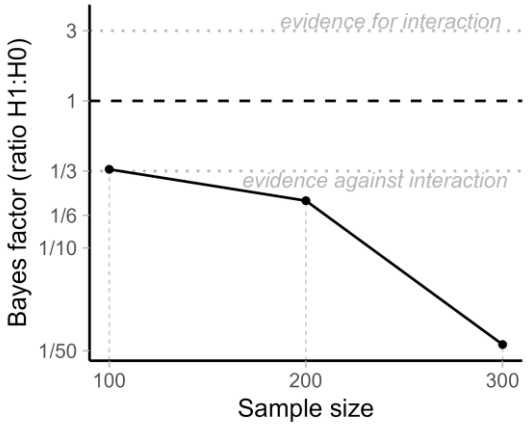
Note. **A.** Interaction of role order with delay type (neutral vs. none) in the subset of the data containing only the neutral and no delay conditions. Priors with increasing standard deviations are shown on the x-axis and the corresponding Bayes factor (BF) for a model with versus without the interaction is shown on the y-axis. The red triangle indicates the prior used in the main analysis. Grey horizontal dotted lines indicate what would be considered at least moderate evidence for (BF ≥ 3) or against (BF ≤ ⅓) the interaction according to Lee & Wagenmaker's (2014) adaptation of Jeffreys' (1939) scale. **B.** Interaction of the role order with delay type (associated vs. neutral) in the subset of the data containing only the associated and neutral delay conditions. **C.** Interaction of the role order with delay type (associated vs. none) in the subset of the data containing only the associated and no delay conditions.

A2. Design analysis

Since we had to abandon our pre-registered sample size criterion of continuing recruitment until the Bayes factor for the role order × delay type interaction reached 6 or we reached 100 participants, we used the data we had collected for 74 participants to simulate hypothetical experiments of 100, 200 and 300 participants (Figure A2). Assuming the current data were a good representation of true values, not even 200 participants would have reached our pre-registered evidence cut-off (evidence 6:1 either for or against the interaction). We focused on the interaction or role order and delay in the subset of the data containing only the neutral and no delay conditions, since this was the effect for which there was previous support from the literature.

Figure A2

Bayes factors for simulated sample sizes



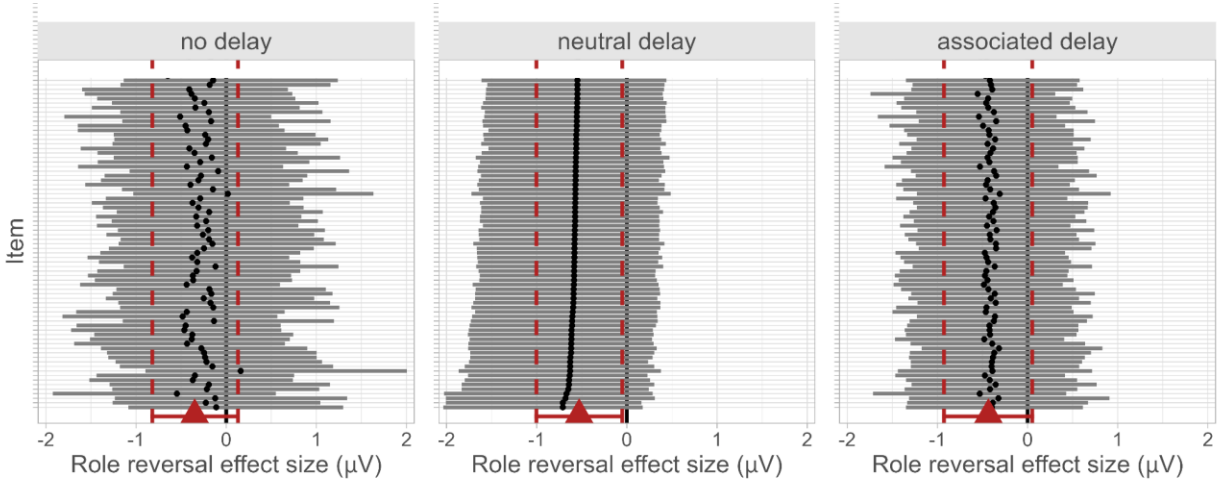
Note. Selected sample sizes are shown on the x-axis and the corresponding Bayes factor (BF) for a model with versus without the interaction is shown on the y-axis. Grey horizontal dotted lines indicate what would be considered at least moderate evidence for (BF ≥ 3) or against (BF ≤ ⅓) the interaction according to Lee & Wagenmaker's (2014) adaptation of Jeffreys' (1939) scale.

A3. By-item differences in the delay effects

In the main text, we examined by-participant effects of canonical and reversed sentences within each delay condition. It has also been suggested that the type of items may also be relevant for inducing the illusion (Ehrenhofer et al., submitted; Li & Ettinger, 2022), so we additionally extracted by-item effects from the same nested model. Figure A3 shows a similar pattern to the by-participant effects in that all conditions showed a small effect of role reversals across items with a negative sign (N400 lager in the reversed condition). However, the pattern of variability of the estimates is reversed: whereas the no delay condition was the least variable in the by-participant estimates, it is the most variable in the by-item effects. According to Ehrenhofer et al. or Li and Ettinger, this could result from variability between items in the frequency or typicality of agent-hood or patient-hood of the verbal arguments within the verbal event, and/or from variability in the difference in plausibility between the canonical and reversed sentences. That these factors created strikingly more variability in the no delay condition than the delay conditions corroborates the finding that the delay helped readers better apply the syntactic constraints of the sentence.

Figure A3

By-item estimates of the main effect of role order within each delay condition



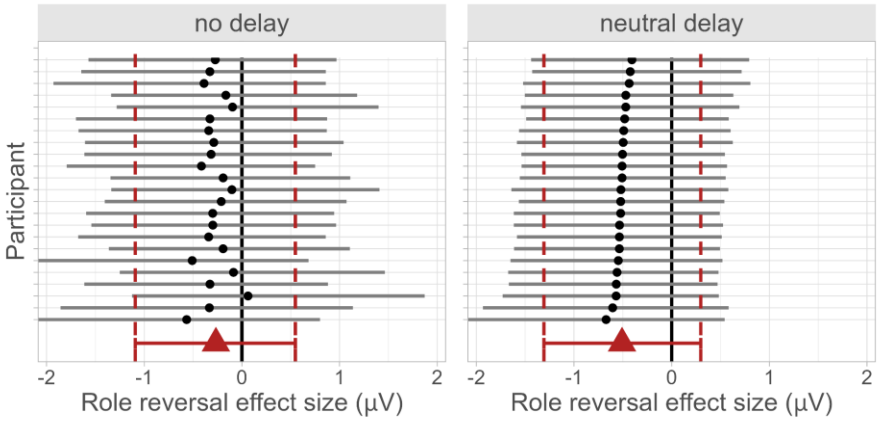
Note. The black points and error bars show each item's mean estimated N400 difference between reversed and canonical sentences and 95% credible interval. A negative sign indicates that the N400 was more negative in the reversed condition. The red triangles indicate the group estimate for each contrast.

A4. By-participant estimates from Chow et al. (2018)

In the main text, we noted that there was actually a small but consistent N400 effect in the no delay condition which was unexpected, since this condition reflects the role reversal construction in which multiple studies have observed an illusion. To check whether this was unique to our study, we also fit a nested model to the data from Chow et al. (2018). Figure A4 shows that a similar pattern was observed in their study. If this is common across all illusion studies, it may be that most readers are able to apply thematic role constraints even if the verb appears immediately, but that something about this process is incomplete or requires time for preactivation of a suitable verb to reach a level such that a discrepancy is detectable as an N400 in the reversed condition.

Figure A4

By-participant estimates from Chow et al. (2018) ordered by effect size in the neutral delay condition

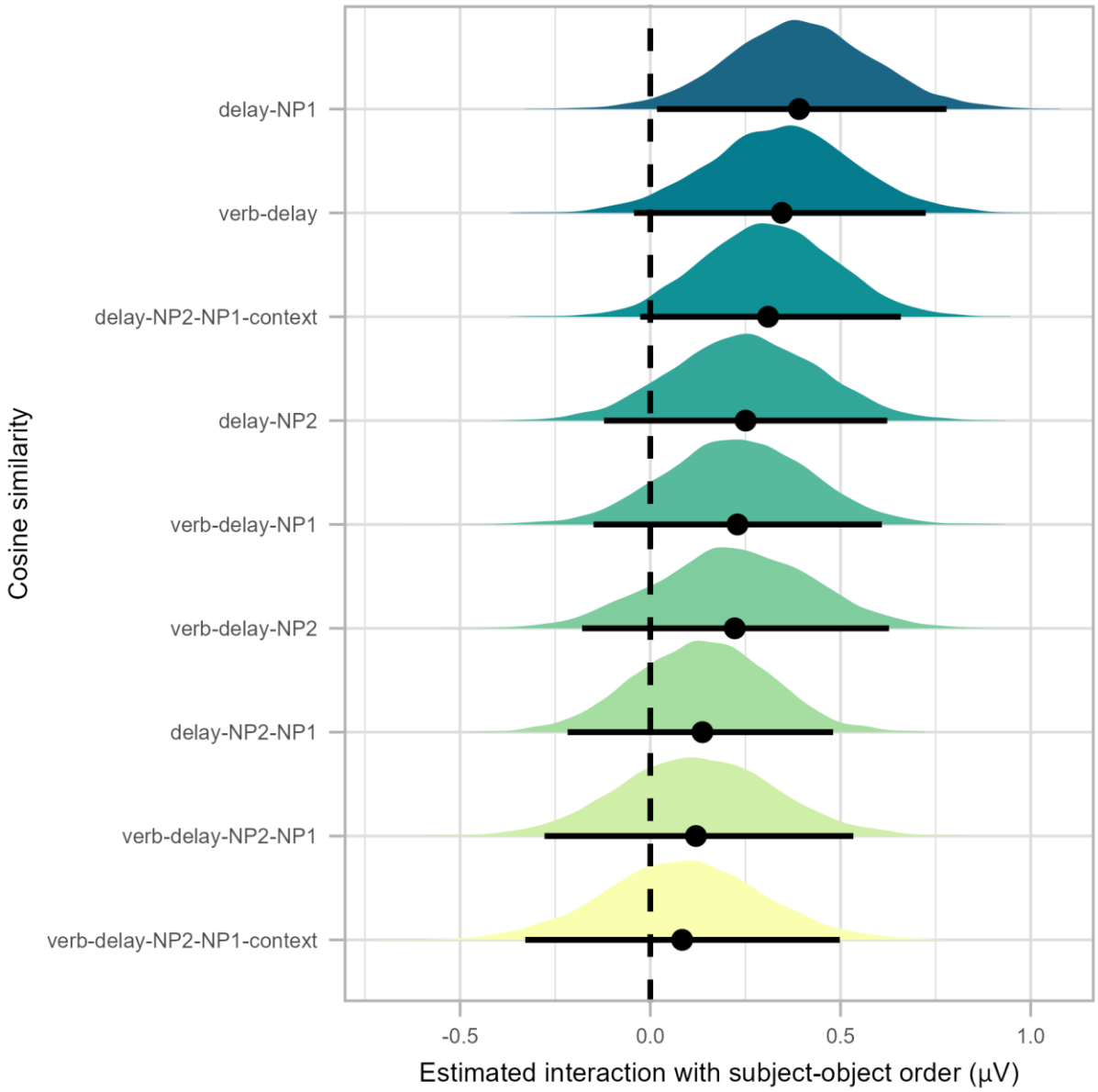


Note. The black points and error bars show each participant's mean estimated N400 difference between reversed and canonical sentences and 95% credible interval. A negative sign indicates that the N400 was more negative in the reversed condition. The red triangles indicate the group estimate for each contrast.

A5. Cosine similarity analysis model outputs

Figure A5

Posterior estimates for the interaction of role order and cosine similarities



Note. The interaction effect estimates are from separate models. A positive coefficient means more negative N400 amplitude in the role reversed versus the canonical condition.

A6. Cosine similarity summary statistics

Table A6 gives the mean cosine similarity for low- and high-similarity groups split by median cosine value. Cosine similarity was computed between the delay and various other sections of the corresponding sentence.

Table A6

Cosine similarity summary statistics

	Mean cosine by median split (SD)	
	low	**high**
Verb-Delay	0.19 (0.06)	0.40 (0.10)
Verb- Delay- NP1- NP2	0.26 (0.07)	0.50 (0.10)
Verb- Delay-NP1- NP2- Context	0.31 (0.08)	0.50 (0.07)
Delay-NP1- NP2	0.15 (0.05)	0.31 (0.08)
Delay-NP2	0.14 (0.05)	0.34 (0.11)
Delay-NP1	0.16 (0.05)	0.39 (0.13)
Delay-NP1- NP2- Context	0.27 (0.07)	0.46 (0.08)
Verb- Delay-NP1	0.24 (0.07)	0.47 (0.10)
Verb- Delay-NP2	0.23 (0.07)	0.46 (0.11)

Note. NP1 reflects the stereotypical patient of the verb and NP2 the stereotypical agent, in line with the linear ordering of the verbal arguments in the canonical sentences.

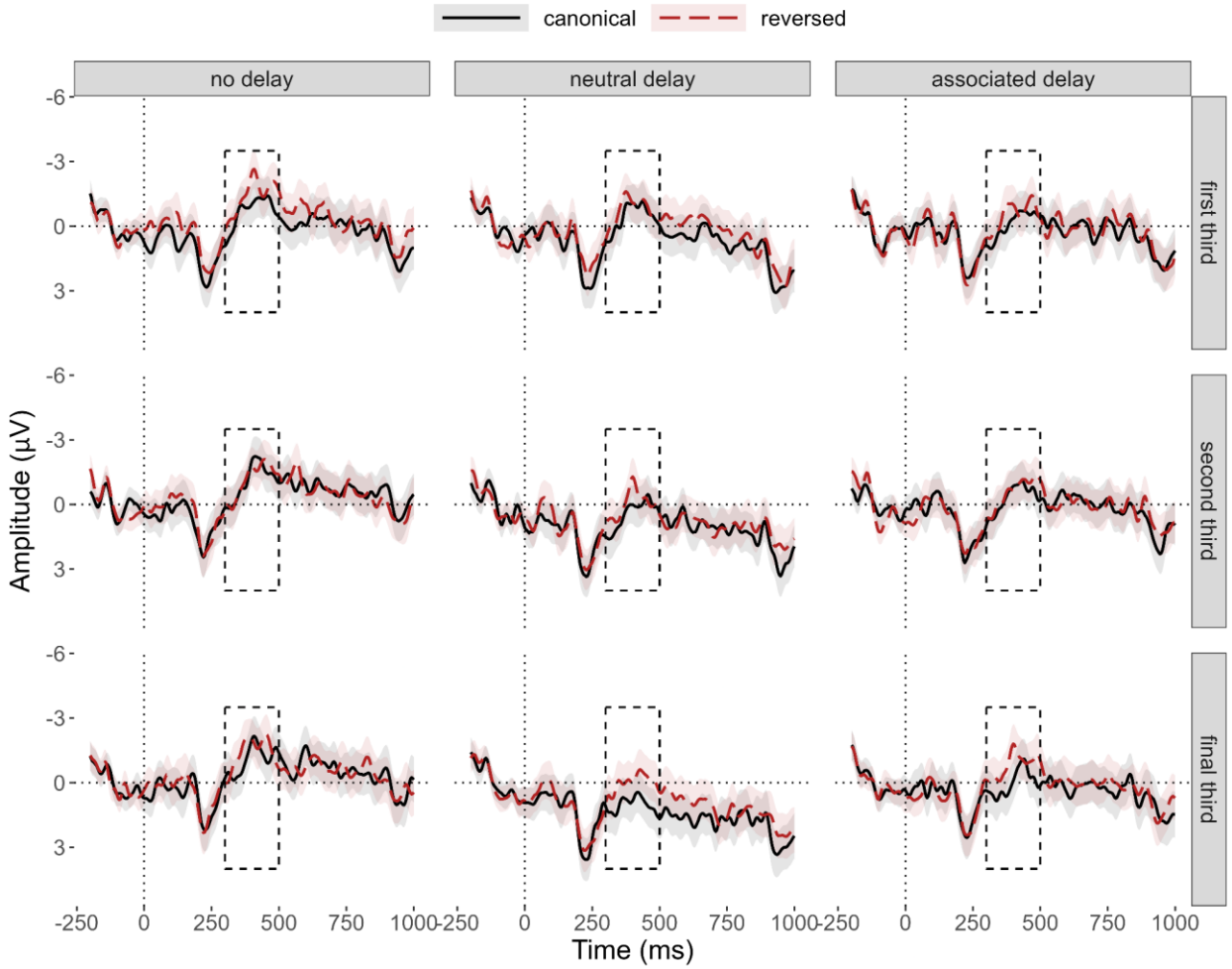
A7. The delay effect as the experiment progressed

Many participants reported that they noticed the role reversals, at first thinking it was a grammatical error and later realising it was likely part of the experiment. This could suggest a change in strategy as the experiment progressed; for example, switching from an analytical reading mode to a more passive mode, or vice versa. Figure 4 plots the averaged ERPs in the first, middle, and final thirds of the experiment, for each of the delay conditions. A large N400 effect was visually apparent in the neutral delay condition (and no other condition) in the final third of the experiment. When trial order was added as a scaled, centred predictor to the full 2 × 3 model using a weakly constraining, non-directional prior of $Normal(0,1)$, the posterior for the interaction of role order and delay in the neutral versus no delay comparison was not consistent with a change as trial order increased, although 89% of the posterior favoured an increase in the size of the interaction as the experiment progressed, $\hat{\beta} = -0.21\,\mu V, 95\%\,CrI = [-0.54, 0.12]\,\mu V$, $P(\beta < 0) = 0.89$. There was no indication that the interaction of role order and delay in the informative versus neutral delay comparison changed as trial order increased, $\hat{\beta} = 0.06\,\mu V, 95\%\,CrI = [-0.29, 0.40]\,\mu V, P(\beta < 0) = 0.63$. We thus conducted follow up comparisons within each of the delay

conditions within the final third of the experiment and found a main effect of role order in the neutral delay condition, $\hat{\beta} = -0.96\,\mu V, 95\%\,CrI = [-1.76, -0.16]\,\mu V$, $P(\beta < 0) = 1.00$, but not in the no delay condition, $\hat{\beta} = -0.37\,\mu V, 95\%\,CrI = [-1.19, 0.46]\,\mu V$, $P(\beta < 0) = 0.81$, or the informative delay condition, $\hat{\beta} = -0.60\,\mu V, 95\%\,CrI = [-1.39, 0.21]\,\mu V$, $P(\beta < 0) = 0.93$.

Figure A7

ERPs at the target verb in the first, middle, and last thirds of the experiment



Note. Mean amplitude across electrodes Cz, CP1, CPz, CP2, P3, Pz, P4, PO3, POz, and PO4 is plotted in the no-delay (e.g. *which guest the moderator* ***invited***), neutral delay (e.g. *which guest the moderator* additionally ***invited***), and informative delay conditions (e.g. *which guest the moderator* to the panel show ***invited***). The dashed box indicates the analysis time window.