COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 48 (2024) e70023

ISSN: 1551-6709 online
DOI: 10.1111/cogs.70023

Task and Timing Effects in Argument Role Sensitivity: Evidence From Production, EEG, and Computational Modeling

Masato Nakamura,[a] Shota Momma,[b] Hiromu Sakai,[c] Colin Phillips[d,e]

[a] *Department of Language Science and Technology, Saarland University*
[b] *Department of Linguistics, University of Massachusetts, Amherst*
[c] *Faculty of Science and Engineering, Waseda University*
[d] *Faculty of Linguistics, Philology, and Phonetics, University of Oxford*
[e] *Department of Linguistics, University of Maryland*

Received 19 February 2024; received in revised form 22 September 2024; accepted 11 November 2024

Abstract

Comprehenders generate expectations about upcoming lexical items in language processing using various types of contextual information. However, a number of studies have shown that argument roles do not impact neural and behavioral prediction measures. Despite these robust findings, some prior studies have suggested that lexical prediction might be sensitive to argument roles in production tasks such as the cloze task or in comprehension tasks when additional time is available for prediction. This study demonstrates that both the task and additional time for prediction independently influence lexical prediction using argument roles, via evidence from closely matched electroencephalogram (EEG) and speeded cloze experiments. In order to investigate the timing effect, our EEG experiment used maximally simple Japanese stimuli such as *Bee-nom/acc sting*, and it manipulated the time for prediction by changing the temporal interval between the context noun and the target verb without adding any further linguistic content. In order to investigate the task effect, we conducted a speeded cloze study that was matched with our EEG study both in terms of stimuli and the time available for prediction. We found that both the EEG study with additional time for prediction and the speeded cloze study with matched timing showed clear sensitivity to argument roles, while the EEG conditions with less time

Correspondence should be sent to Masato Nakamura, Department of Language Science and Technology, Saarland University, Biulding C7.1, 66123, Saarbrücken, Germany. E-mail: nakamura@lst.uni-saarland.de

for prediction replicated the standard pattern of argument role insensitivity. Based on these findings, we propose that lexical prediction is initially insensitive to argument roles but a monitoring mechanism serially inhibits role-inappropriate candidates. This monitoring process operates quickly in production tasks, where it is important to quickly select a single candidate to produce, whereas it may operate more slowly in comprehension tasks, where multiple candidates can be maintained until a continuation is perceived. Computational simulations demonstrate that this mechanism can successfully explain the task and timing effects observed in our experiments.

Keywords: Language comprehension; Language production; Prediction; N400; Cloze task

1. Introduction

Lexical items are not comprehended in isolation. Their processing is shaped by their preceding contexts. More predictable lexical items are easier to process than less predictable ones, as indexed by shorter reading times or reduced neural activity (e.g., Ehrlich & Rayner, 1981; Kutas & Hillyard, 1984). Such facilitation is often considered to reflect context-driven activation of stored representations of upcoming lexical items or their meanings (Federmeier, 2007). Many researchers have argued that people can use various types of contextual information to appropriately activate upcoming lexical items (verbs' selectional restrictions: Altmann & Kamide, 1999; negation: Nieuwland & Kuperberg, 2008; event knowledge: Kamide, Altmann, & Haywood, 2003; Bicknell et al., 2010; Rabs, Delogu, Drenhaus, & Crocker, 2022; the relationship between multiple events mediated by discourse connectives: Xiang & Kuperberg, 2015; discourse-specific characteristics of instances: Nieuwland & van Berkum, 2006). Context-driven activation of upcoming items is thought to support efficient and robust language comprehension in our daily lives, where linguistic information must be processed quickly and often in noisy settings. The abundant evidence for activation of appropriate continuations has led some researchers to argue that people immediately use all available information to generate highly accurate expectations for upcoming lexical items (e.g., Altmann & Mirković, 2009; Kuperberg & Jaeger, 2016). In this paper, we refer to context-driven lexical activation as (lexical) prediction, in line with the prior literature, even though the two concepts might not be identical.

Even though humans generally make appropriate predictions using different types of contextual information, there is increasing evidence of inappropriate predictions in specific circumstances. For example, multiple sources of evidence suggest that comprehenders activate verbs that are inappropriate continuations given the argument roles of prior noun phrases. Concretely, the verb *served* is a likely continuation given the preceding context in (1a), but it is an unlikely continuation in (1b), due to the reversed arguments. We henceforth refer to this type of anomalous sentence as an *argument role reversal*. Many studies in different languages have reported that the N400 component, which is often considered as an electrophysiological measure of prediction, is matched in amplitude for continuations such as *served* in the two contexts, despite the clear contrast in predictability (English: Chow, Smith, Lau, & Phillips, 2016; Kim & Osterhout, 2005; Kuperberg, Sitnikova, Caplan, & Holcomb, 2003; Dutch: Hoeks, Stowe, & Doedens, 2004; Kolk, Chwilla, van Herten, & Oor, 2003; Mandarin: Chow, Lau, Wang, & Phillips, 2018; Liao, Lau, & Chow, 2022; Japanese: Oishi & Sakamoto, 2010; German: Stone & Rabovsky, 2024. However, see Bornkessel-Schlesewsky et al., 2011 for exceptions). There are also consistent findings from eye-tracking studies (Visual world paradigm: Kukona, Fang, Aicher, Chen, & Magnuson, 2011; eye-tracking while reading: Burnsky, 2022). These findings have been taken as evidence for role-insensitive activation of upcoming lexical items (e.g., Brouwer, Crocker, Venhuizen, & Hoeks, 2017; Chow, Smith, et al., 2016; Kuperberg, 2016). This is surprising, considering that people make successful predictions using a variety of contextual information, but they fail to use argument roles, which provide crucial information about sentence meaning. Furthermore, this raises questions about people's ability to use available contextual information to generate predictions that closely mirror the statistical patterns in the past experience, which is often assumed to be excellent in the literature on prediction in language processing (e.g., Kuperberg & Jaeger, 2016) and in cognitive science more generally (e.g., Clark, 2013).

(1) a. The restaurant owner forgot which customer the waitress had served.
b. The restaurant owner forgot which waitress the customer had served.

Importantly, we take these studies to inform us about the predictions that people generate using argument roles, rather than about misinterpretation. That is, we do not take them to show that people systematically misinterpret the argument role reversal sentences such as (1b) in the same way as appropriate sentences such as (1a). In fact, in most of the studies above, participants reliably report that argument role reversals such as (1b) are implausible in the comprehension questions. Therefore, even when people obtain accurate interpretations of the sentences, the N400 component still shows no sensitivity to argument roles. Thus, argument role reversals do not cause a long-lasting misinterpretation, but this does not exclude the possibility that a transient misinterpretation is quickly revised and is not captured by such questions. However, this variant is still unlikely given the findings from the current experiments and we will discuss it further in the General Discussion section.

On the other hand, some studies have pointed out potential exceptions to this robust pattern. Chow et al. (2018) argue that people expect role-appropriate verbs more than role-inappropriate verbs when the verb is further away from its arguments such that there is additional time for context-driven lexical activation prior to the verb. There is also suggestive evidence that people show sensitivity to argument roles when they are explicitly asked to produce the continuations in a speeded cloze task (Chow, Kurenkov, Kraut, & Phillips, 2015). When speakers were prompted to provide a continuation to contexts such as (1), a small number of the responses were role-inappropriate verbs, but these were much less common than role-appropriate verbs. This suggests that prediction can show sensitivity to argument roles when it is measured through specific experimental paradigms such as production tasks.

In this study, we ran matched Japanese EEG and cloze experiments to investigate the contributions of timing and task differences to argument role sensitivity in context-driven activation. Using maximally simple stimuli such as (2), the EEG study compared N400 amplitudes at different moments by controlling the timing of the presentation of the target verbs. The cloze study used the same set of stimuli and tested the contrast between production and comprehension by inducing responses within specific time windows that matched the EEG study. After demonstrating under what circumstances people can show sensitivity to argument roles, we propose the hypothesis that role-sensitivity is the result of a serial monitoring process that checks for the suitability of highly activated lexical items, and we demonstrate through a computational simulation that it can simultaneously address the timing and task effects in lexical activation in argument role reversals.

(2) a. Hati-ga sas-u.
Bee-nom sting-pres.
"A/the bee stings something."
b. Hati-o sas-u.
Bee-acc sting-pres.
"Something stings a/the bee."

1.1. Background

Many studies suggest that predictions are insensitive to argument roles. Many of these are EEG studies using the N400 component of the event-related potential (ERP) (Hoeks et al., 2004; Kim & Osterhout, 2005; Kolk et al., 2003; Kuperberg et al., 2003; Chow, Smith, et al., 2016; Chow et al., 2018; Liao et al., 2022; Oishi & Sakamoto, 2010). The N400 component is a negative-going deflection of the ERP that peaks around 400 ms following stimulus onset. It is known that the N400 components elicited by more predictable lexical items are smaller (i.e., more positive) compared to those of less predictable lexical items (e.g., Kutas & Hillyard, 1984). This contrast in N400 amplitudes has been referred to as the *N400 Effect*. Although there are alternative interpretations for the N400 component (e.g., Nieuwland et al., 2020; Rabovsky, Hansen, & McClelland, 2018), it is often considered to reflect retrieval of stored lexical representations or the semantic representations tied to them (Lau, Phillips, & Poeppel, 2008). Therefore, the N400 effects caused by predictability manipulations have been taken to reflect facilitated lexical access due to predictive lexical activation driven by contextual information prior to the presentation of the lexical item. Given these linking assumptions for N400 amplitudes, the repeatedly found absence of N400 contrasts between role-appropriate and inappropriate continuations such as *served* in (1a) and (1b) have been taken to show that people at least initially make predictions about upcoming lexical items ignoring argument roles.

While many studies have repeatedly found insensitivity to argument role reversals in N400 amplitudes, timing and task manipulations seem to make lexical prediction sensitive to argument roles. Chow et al. (2018) argued that people activate upcoming lexical items in a role-appropriate manner if additional time is provided for prediction. They contrasted Mandarin SOV sentence pairs such as (3) and (4) including argument role reversals (3b and 4b). Both types of sentences included temporal adverbial phrases such as *zai shangxingqi*, which consists of a preposition *zai* and an adverbial time expression *shangxingqi* "last week." Whereas the sentence pairs in (3) had the adverbial phrase in sentence initial position, the pairs in (4) had the adverbial phrase between the second argument and the verb. This provided additional time for verb prediction in (4) compared to (3). Chow et al. found matched N400 amplitudes for the verb *zhuale* "arrested" in (3a) and (3b), consistent with previous studies. However, they found larger N400 amplitudes for (4a) than (4b). While there is some concern that (4) may sound slightly awkward to native Mandarin speakers, the study suggests that insensitivity to argument roles may be conditioned on timing, and that the role-sensitivity of prediction might change dynamically after the presentation of the context arguments but before the presentation of the target verb.

(3) a. Zai shangxingqi jingcha ba xiaotou zhua-le ...
ZAI last.week cop BA thief arrest...
'Last week the cop arrested the thief ...'
b. Zai shangxingqi xiaotou ba jingcha zhua-le ...
ZAI last.week thief BA cop arrest...
'Last week the thief arrested the cop...'
(4) a. Jingcha ba xiaotou zai shangxingqi zhua-le ...
Cop BA thief ZAI last.week arrest...
'Last week the cop arrested the thief ...'
b. Xiaotou ba jingcha zai shangxingqi zhua-le ...
Thief BA cop ZAI last.week arrest...
'Last week the thief arrested the cop ...'

Three important points should be highlighted here. First, if there is an initial stage when predictions are not fully accurate, then timing must matter a great deal, since it is clear that untimed, offline predictions are role sensitive, as reflected in measures like offline cloze tasks. In that regard, the main contribution of Chow et al. (2018) is to show that relatively little additional time is needed for role sensitivity to emerge. In practice, that amount of time is likely often available in naturalistic language comprehension. Second, the timing of the N400 effect relative to the presentation of the critical verb itself did not change in Chow et al.'s study. The timing manipulation was confined to the time between the context arguments and the critical verb. Hence, any differences in the ERP response to the verb must reflect differences in what happened prior to the presentation of the verb. Third, Chow et al. only found the effect of timing on N400 amplitudes for highly predictable verbs but not for somewhat predictable verbs. The two types of verbs were matched in their plausibility. This provides further evidence that it is specifically prediction processes that are impacted by timing rather than later processes that affect plausibility.

While Chow et al. argued that additional time can result in role-sensitive predictions, lexical predictions can also show sensitivity to argument roles when different measures are used. Context-driven lexical activation can be also measured by explicitly asking people to provide continuations to given contexts in the cloze task (Taylor, 1953). The proportion of each response given a context is called the cloze probability and it is widely used as a measure of prediction. It is known that cloze probability is sensitive to argument roles, and verbs have higher cloze probability in the contexts where they are appropriate regarding the roles (e.g., *served* in (1a)) than when they are inappropriate (e.g., *served* in (1b)). However, because most cloze studies are untimed, researchers have typically assumed that cloze probabilities do not reflect online sentence processing, but have rather taken them as reflecting the upper bound of people's usage of contexts, given unlimited or at least a lot of time and cognitive resources.

An unpublished cloze study by Chow et al. (2015) suggests that the online-offline contrast may not explain the difference between the cloze task and online measures of context-driven activation such as N400. They conducted a cloze task, but imposed deadlines for production by asking participants to produce a continuation within certain time limits. Their results showed that people were far more likely to produce role-appropriate continuations than role-inappropriate continuations, even in a cloze task under time pressure. This suggests that cloze data show sensitivity to argument roles not because they are offline measures. It also suggests that people can show early sensitivity to argument roles in at least some experimental tasks. However, we can only draw limited conclusions from this study because they did not match the language or timing with prior EEG studies that showed timing effects (Chow et al., 2018). Furthermore, Chow et al. focused more on the small number of role-inappropriate responses and the parallelism between the cloze and the EEG studies than on the dominant pattern of role-sensitivity.

The contrast between the EEG and the cloze studies is particularly interesting given prior claims about the linking hypotheses for N400s and cloze. Staub, Grant, Astheimer, and Cohen (2015) proposed a Race model for the speeded cloze task. In this model, the cloze process is understood as the parallel activation of lexical candidates, and a cloze response is considered as the first lexical candidate whose activation reaches a threshold for lexical access. This winner-take-all process reflects a fundamental characteristic of production that yields a single output. This is different from comprehension, where multiple possibilities may be maintained simultaneously. Their model successfully captures two patterns in their human cloze data. First, it successfully predicts that cloze probabilities and cloze reaction times (RTs) negatively correlate with each other. This is because fast activation is associated with both higher cloze probabilities and shorter cloze RTs. Second, the model predicts that when the cloze probabilities of two lexical candidates are matched, the one that competes with a stronger competitor should have shorter RTs. This is because it has to be fast to beat faster competitors, and hence this is a key diagnostic of a process in which multiple candidates compete for production. This exact pattern was observed in Staub et al.'s cloze experiments. Given the literature on the speeded cloze measures and N400 amplitudes, both types of measures are considered to reflect, albeit indirectly, the activation of lexical items. If role-insensitivity is only observed in N400 amplitudes but not in cloze probabilities, even when the timing is matched, the difference in the role-sensitivity of the prediction measures cannot be reduced to the online-offline contrast.

Argument role reversals are an interesting phenomenon in that they seem to be exceptions to the generally accurate predictions that people make, but these potential task and timing effects make them even more useful test cases. Prior models of prediction in language processing (e.g., Rabovsky et al., 2018; Brouwer, Fitz, & Hoeks, 2012, 2017; Kuperberg & Jaeger, 2016) claim that predictions are updated as each new piece of information is received and they typically focus on EEG data. If people's expectations are initially insensitive to argument roles but become sensitive to argument roles later without any new input, it suggests that there are fine-grained temporal dynamics in what happens in between inputs in updating predictions. Additionally, if we find qualitatively different patterns between the EEG study and the cloze study that cannot be reduced to the online-offline contrast, then models of prediction must account for the differences between the tasks.

1.2. The current study

This study aims to address the timing and task effects in argument role reversals. We conducted closely matched EEG and speeded cloze studies in order to control the timing, with minimal changes otherwise, and to compare different measures at the same timing. The characteristics of Japanese allowed us to create maximally simple stimuli such as (2) for these purposes. The sentences consisted of one context noun with a case marker and a target verb. Since Japanese is a verb-final language that allows free dropping of arguments, both (2a) and (2b) are grammatical. The role-appropriate (2a) and inappropriate (2b) differ only in the case markers on the context nouns. Participants had relatively easy access to argument role information because the explicitly marked cases strongly correlate with argument roles. In active voice constructions in Japanese, agents appear as subjects with the nominative case marker *-ga*, and patients appear as objects with the accusative case marker *-o*. Of course, case markers and argument roles are not perfectly correlated in Japanese. For example, patients in passive constructions are marked with the nominative case marker. However, there is a strong probabilistic association between case markers and argument roles in Japanese. Additionally, in our experiments, all stimuli were in active voice and arguments with the nominative case marker were always agents, and arguments with the accusative case marker were always patients. Importantly, the simplicity of our materials allowed our EEG study to manipulate timing by varying the interval between the context noun and the target verb, without changing any other aspects of the stimuli. It is difficult to create such simple stimuli in languages that encode argument roles using word order, and that do not allow free dropping of arguments, or that require complex structures for verb-final sentences. These Japanese stimuli thus can provide a stronger test of the timing effects than Chow et al. (2018), where the differences in the positions of the adverbial phrases in (3) and (4) could, in principle, have resulted in slightly different predictions.

In our EEG study, participants were presented with role-appropriate and inappropriate stimuli such as (2). The stimulus onset asynchrony (SOA) between the context noun and the target verb was 800 ms in the short condition and 1200 ms in the long condition (Fig. 1). We compared the N400 contrasts between the target verbs as role-appropriate continuations, such as (2a), or as role-inappropriate continuations, such as (2b), with different delays following the context nouns.

The design of the current EEG study is similar to EEG studies by Yano and Sakamoto (2016) and Yano (2018). Those studies used maximally simple Japanese sentences that consisted of a single noun and a single verb. The stimulus set included plausible sentences such as (5a) and corresponding argument role reversal sentences such as (5b). There was also a timing manipulation across the two experiments and the same stimuli were presented with an SOA of 1300 ms in Yano and Sakamoto (2016) and with an SOA of 600 ms in Yano

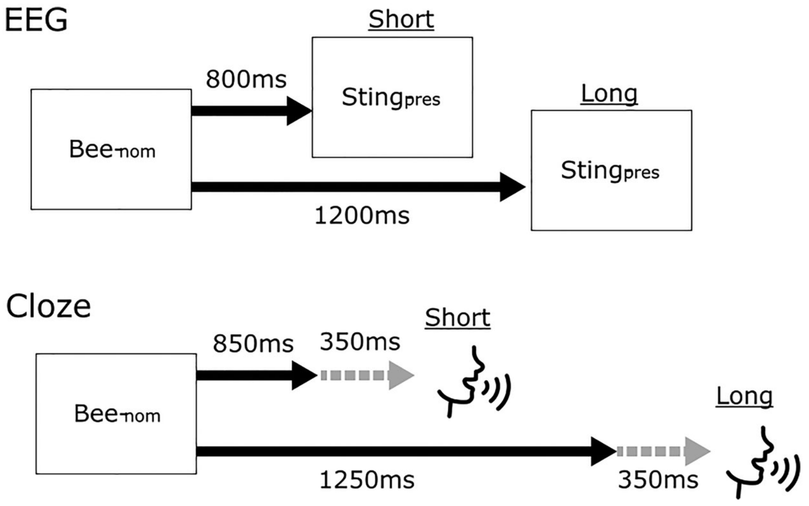


Fig. 1. Timing manipulation in the EEG and the speeded cloze studies. The SOA was 800 ms in the short condition and 1200 ms in the long condition in the EEG study. In the cloze task, if a continuation is produced 1200 ms after the onset of the context (short condition), there should have been less than 850 ms available for activation. If a continuation was produced at least 1600 ms following the context, there should have been more than 1250 ms for activation.

(2018). However, the current study is crucially different from Yano and Sakamoto's studies in its aim and the stimuli. All the nouns in Yano and Sakamoto's stimuli were inanimate since the studies were concerned with the structural predictions elicited by inanimate subjects. On the other hand, since we are primary concerned with semantic prediction of verbs using arguments, all the nouns in the current study were animate so that there was no animacy violation in addition to argument role reversal anomalies in the implausible sentences. In fact, Yano argued that the inanimate subject in (5b) elicited a left-anterior negativity (LAN effect) at the verb. It was important for the current experiment to use animate arguments in order to avoid eliciting such potentially confounding components.

(5) a. Mado-o simer-u.
Window-acc close-pres.
"Someone/something closes a/the window."
b. Mado-ga simer-u.
Window-nom close-pres.
"A/the window closes someone/something."

In the cloze study, we presented participants with the same contexts as the EEG study, but without the target verbs. Instead, participants were asked to produce a natural verb to continue each context noun within certain time windows that matched the EEG study. The time windows were 1200 ms following the context noun in the short condition and between 1600 and 2800 ms in the long condition.

Crucially, the SOAs in the EEG study and the response time windows in the cloze study were chosen so that the times available for lexical activation were matched, despite the different post-lexical processes involved in the two paradigms. While the N400 component is directly elicited by lexical access under the assumed linking hypothesis, the cloze response times also reflect post-lexical processes such as articulatory planning. Given that it is estimated to take about 350 ms to initiate articulation after completing lexical access (Indefrey & Levelt, 2004), we estimated that participants had up to 850 ms for lexical activation in the short condition and 1250–2450 ms in the long condition, which aligns with the EEG timing manipulation (Fig. 1). This timing manipulation may not have perfectly matched the EEG and the cloze studies, because the SOA in the EEG study might have underestimated the time available for context-driven lexical activation. Specifically, participants might continue activating possible upcoming verbs based on the contextual information even after the verb presentation until bottom-up processing of the verb is completed. However, this possibility has minimal influence on the interpretation of our data because this additional time should make N400 amplitude more sensitive to argument roles, if there is any effect. If we find that people are sensitive to argument roles in the cloze task but not in the EEG experiment, as we expect, the concern about timing alignment would be moot. Thus, even though the two paradigms do not share the same post-lexical processes, we matched the time available up to lexical access, taking the time for different post-lexical processes into account. Therefore, our design allowed us to compare people's context-driven lexical activation at different moments in the same EEG study, and at the same moment in different paradigms.

2. EEG

2.1. Methods[1]

2.1.1. Participants

Twenty-seven participants were recruited at Hiroshima University. We excluded three participants according to a data quality criterion described below, leaving 24 (12 males and 12 females; average age = 21.2). All participants were right-handed, and were native speakers of Japanese, without any significant early exposure to other languages. Prior written consent was obtained for each participant.

2.1.2. Materials

We created 160 plausible sentences with one case-marked context noun and one target verb. Eighty of them had the nominative (NOM) case marker *-ga*, such as (6a), and the other 80 had the accusative (ACC) case marker *-o*, such as (7a). These sentences were the plausible condition in the experiment. Then, we created the implausible condition by switching the cases of those context nouns ((6b) and (7b)) in the plausible condition. All the context nouns were animate, and all the target verbs were transitive verbs that can have multiple arguments. We also manipulated the timing of the presentation of the target verb. In the short condition, the target verb was presented 800 ms after the onset of the context noun. The interval was

1200 ms in the long condition. These two manipulations resulted in four conditions in total: plausible-short, plausible-long, implausible-short, and implausible-long. The 160 experimental items in the four conditions were distributed across four lists using a Latin Square design, so that the same participant saw each experimental item just once. In addition to the experimental items, we also created 80 plausible and 80 implausible filler sentences, which were also presented with an SOA of either 800 or 1200 ms.

(6) a. Hati-ga sas-u.
Bee-nom sting-pres.
"A/the bee stings something."
b. Hati-o sas-u.
Bee-acc sting-pres.
"Something stings a/the bee."

(7) a. Hae-o tatak-u.
Fly-acc swat-pres.
"Something swats a/the fly."
b. Hae-ga tatak-u.
Fly-nom swat-pres.
"A/the fly swats something."

One item was incorrectly labeled in the stimuli. This resulted in some participants seeing one more or fewer plausible items than implausible items. This is unlikely to have affected the results because we relabeled these stimuli for the analyses, and it is unlikely that this small shift in the ratio of plausible and implausible stimuli would result in observable effects.

2.1.3. Procedure

The 320 sentences were divided into four blocks, divided by short breaks. Each sentence was visually presented to participants word-by-word using PsychoPy (Peirce, 2007), while brain activity was recorded via EEG. In each trial, a fixation cross was presented at the center of the screen for 500 ms, and a blank screen was presented for the next 400 ms. The context noun and the target verb were presented word-by-word separated by a blank screen. Each word remained on the screen for 400 ms. The interval of the blank screen was 400 ms in the short condition and 800 ms in the long condition. Therefore, combined with the time the context noun was on the screen, the SOA was 800 ms in the short condition and 1200 ms in the long condition. Participants were asked about the plausibility of each sentence 1000 ms after the offset of the target verb. They provided answers by pressing the keys.

2.1.4. Electrophysiological recording

Thirty-two sintered Ag/AgCl passive electrodes were placed on participants' scalp in an EEG recording cap (Brain Products GmbH Easy Cap 40 Asian cut), according to the 10–20 system. One of the 32 electrodes was placed below the right eye to monitor eye movements and blinking. One electrode was used as ground, and another was used as a reference electrode and was attached to the nose. Impedances were kept below 5 kOhms for all scalp electrode sites. The EEG was sampled at 250 Hz.

2.2. Data analysis

The recorded EEG signals were preprocessed and averaged ERPs were obtained using EEGLAB (Delorme & Makeig, 2004) and ERPLAB (Lopez-Calderon & Luck, 2014) following Luck (2021)'s pipeline. The EEG signal was bandpass-filtered between 0.1 and 30 Hz. Eye blinks and movements were removed by independent component analysis. Nonresponding channels were excluded from this artifact correction procedure and were interpolated from neighboring channels using spherical splines after correction. Then, the EEG was segmented into 1000 ms intervals, starting 200 ms before the target verb onset and ending 800 ms after the verb onset. A baseline correction was applied to the 200 ms epoch prior to the target verb onset. Trials with EEG above 150 μV or below −150 μV were marked for rejection. Additionally, artifacts were detected and rejected using a moving window peak-to-peak algorithm with a window size of 200 ms, a window step of 100 ms, and a threshold of 125 μV. Three participants who had a 20% or greater rejection rate were excluded from further analyses. The rejection rate among the remaining participants was 1.1% on average.

We first independently compared the plausible and implausible conditions for each SOA (short vs. long). The EEG data was analyzed with nonparametric cluster-based permutation tests using the FieldTrip toolbox (Oostenveld, Fries, Maris, & Schoffelen, 2011). We obtained individual participant averages at each electrode for each time point from 300 to 800 ms after the onset of the target verb, and computed t-values for each electrode and time. We then selected spatially and temporally adjacent samples based on the t-values to form clusters ($\alpha = 0.05$). The summed t-values within each cluster were used as the cluster-level test statistic. Then, we randomly permuted the condition labels 1000 times and obtained the maximum of the cluster-level test statistics for each iteration. The test statistics of our data were compared against this randomly generated distribution, and the p-value of each cluster is the proportion of the generated test statistics greater than that of the cluster. We used an α-level of 0.025 for two-tailed tests. Within the 300−800 ms time window that we analyzed, we identified N400 effects as significant negative clusters in the 350−450 ms time window, which is the time window commonly used for N400 analyses. In addition to the analyses within each SOA, we also analyzed the interaction between plausibility and SOA by comparing the difference between the plausible and implausible conditions for each SOA and performing the same analysis as above (i.e., the difference between implausible-long and plausible-long compared against the difference between implausible-short and implausible-short).

We did not directly compare the two plausible conditions (plausible-long vs. plausible-short) or the two implausible conditions (implausible-long vs. implausible-short) because these comparisons could be confounded by the influence of the context nouns. The ERPs elicited by the target verbs are inevitably affected by the ERPs elicited by the preceding context nouns. Since the target verbs were presented with different latencies from the presentation of the context nouns in the long and the short conditions, the ERPs elicited by the context nouns might affect the ERPs of the target verbs differently in the long and the short conditions.

2.3. Results

2.3.1. Behavioral

The overall accuracy of the plausibility judgment task was 93.4%, including both experimental and filler items. More specifically, the accuracy was 96.4% for the plausible-short condition, 90.5% for the implausible-short condition, 96.8% for the plausible-long condition, and 89.8% for the implausible-long condition (all averages including filler items). A binomial generalized linear mixed-effects model with maximal random effects structures that converged revealed a significant effect of plausibility ($\beta = 1.34$, $p < .001$). This shows that participants were more likely to accept implausible items (e.g., *Bee-acc sting.*) than to reject plausible items (e.g., *Bee-nom sting.*). However, the accuracy was still overall high across all conditions. On the other hand, there was no statistically significant effect of SOA ($\beta = 0.12$, $p = .27$) or interaction between plausibility and SOA ($\beta = -0.25$, $p = .24$).

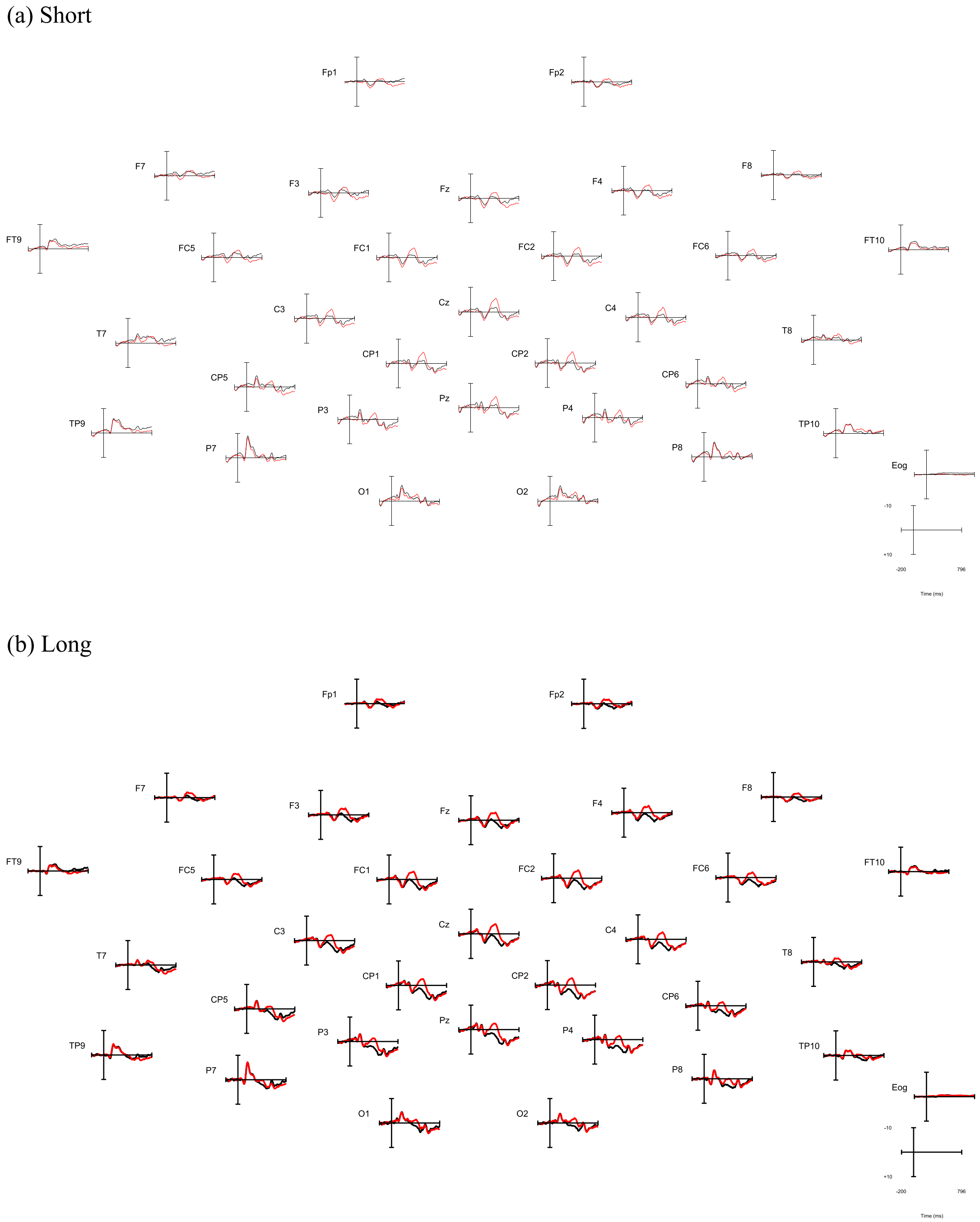
2.3.2. EEG

2.3.2.1. Filler items: Sixty trials were excluded from the analyses because they were erroneously shown to participants multiple times. This left 3783 filler trials to be analyzed. Fig. 2 shows the ERPs for the filler stimuli. As illustrated by the figures, classic N400 effects were observed in the comparison between the plausible and implausible conditions, both for the short and long conditions. Fig. 3 shows topographic maps of ERPs at 300–750 ms after the target verb onset, where each map is based on the average of a 50 ms window. Significant clusters are indicated with stars. As illustrated in the figure, cluster-based permutation tests found significant negative clusters in the N400 time window (350−450 ms) both in the short condition ($p = .010$) and the long condition ($p < .001$). That is, the ERPs in the implausible conditions were more negative than the plausible conditions in this cluster. We also unexpectedly found positive clusters in the late time window (570−800 ms) in the short condition ($p = .016$). No positive or negative clusters were found for the interaction analysis.

2.3.2.2. Experimental items: Fig. 4 shows the ERP waveforms for the experimental items. As Fig. 5 illustrates, there was no statistically significant positive or negative cluster in the short condition. On the other hand, there was a significant negative cluster in the N400 time window in the long condition ($p = .007$). Additionally, there was a significant negative cluster in the N400 time window in the interaction analysis ($p = .013$). These show that the long-implausible condition showed more negativity than the long-plausible condition in the cluster, but there was no evidence for such a contrast in the short conditions.

2.4. Discussion

Our EEG study showed an N400 contrast between the plausible-long condition and the implausible-long condition, but not between the plausible-short condition and the implausible-short condition. The lack of an N400 effect in the short conditions is consistent



(a) Short

(b) Long

Fig. 2. ERPs of (a) the short condition and (b) the long condition of the filler items.

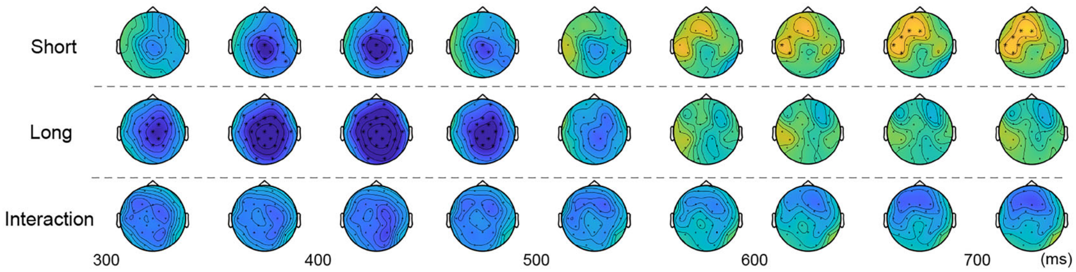


Fig. 3. Topographic maps of ERPs for the filler items. The top two rows indicate the difference between the plausible and the implausible conditions (implausible minus plausible) for each SOA. The bottom row shows the difference between the top two rows (long minus short). Stars (*) indicate members of significant clusters.

with many previous EEG studies of role reversals (e.g., Kolk et al., 2003; Kim & Osterhout, 2005; Chow, Smith, et al., 2016) and it cannot be reduced to a general absence of context effects in the short condition, as we find N400 effects in the filler items with the same SOAs. The presence of an N400 effect in the long condition replicates Chow et al. (2018). Thus, N400 amplitude at the verb is insensitive to argument roles in a setting where the interval between the context and the target verb is short, but it shows sensitivity to argument roles with an additional 400 ms delay before verb presentation. Note that this additional time (400 ms) is much shorter than the additional time added by moving a prepositional phrase (1200 ms) in Chow et al. (2018).

Previous EEG studies of argument role reversals have reported a late positivity for reversed sentences compared to canonical sentences (P600 effects). This was also the case for Chow et al. (2018), where they observed P600 effects both in the short and the long conditions. Some researchers have argued that P600 effects reflect the effort of semantic integration for a target lexical item that makes the sentence implausible (e.g., Brouwer et al., 2017). Others have argued that the P600 effects reflect syntactic reanalysis to obtain nonliteral but plausible meanings (e.g., van Herten, Kolk, & Chwilla, 2005). For example, comprehenders might reanalyze *bee-acc sting* as *bee-nom sting*. Interestingly, we did not find significant P600 effects in our current study, either in the short condition or the long condition. Although we do not have a clear account for the absence of P600 effects in our study, we suspect that this was due to the simplicity of our stimuli, which might have resulted in a relatively small integration cost or revision effort. Moreover, there might have been no syntactic reanalysis at all, because there was little reason for participants in our study to have any uncertainty over whether they had correctly understood the context.

The current findings regarding the timing of verb presentation are potentially closely related to Yano and Sakamoto (2016) and Yano (2018). Those studies used simple Japanese stimuli with argument role reversals that were very similar to the current experiment, such as (8). However, they used inanimate arguments in order to investigate the syntactic expectations for intransitive verbs elicited by inanimate subjects. Just like the current study, Yano and Sakamoto (2016) found a negativity in the N400 time window for (8b) compared to (8a) when the SOA was 1300 ms, but Yano (2018) did not find a negativity when the SOA

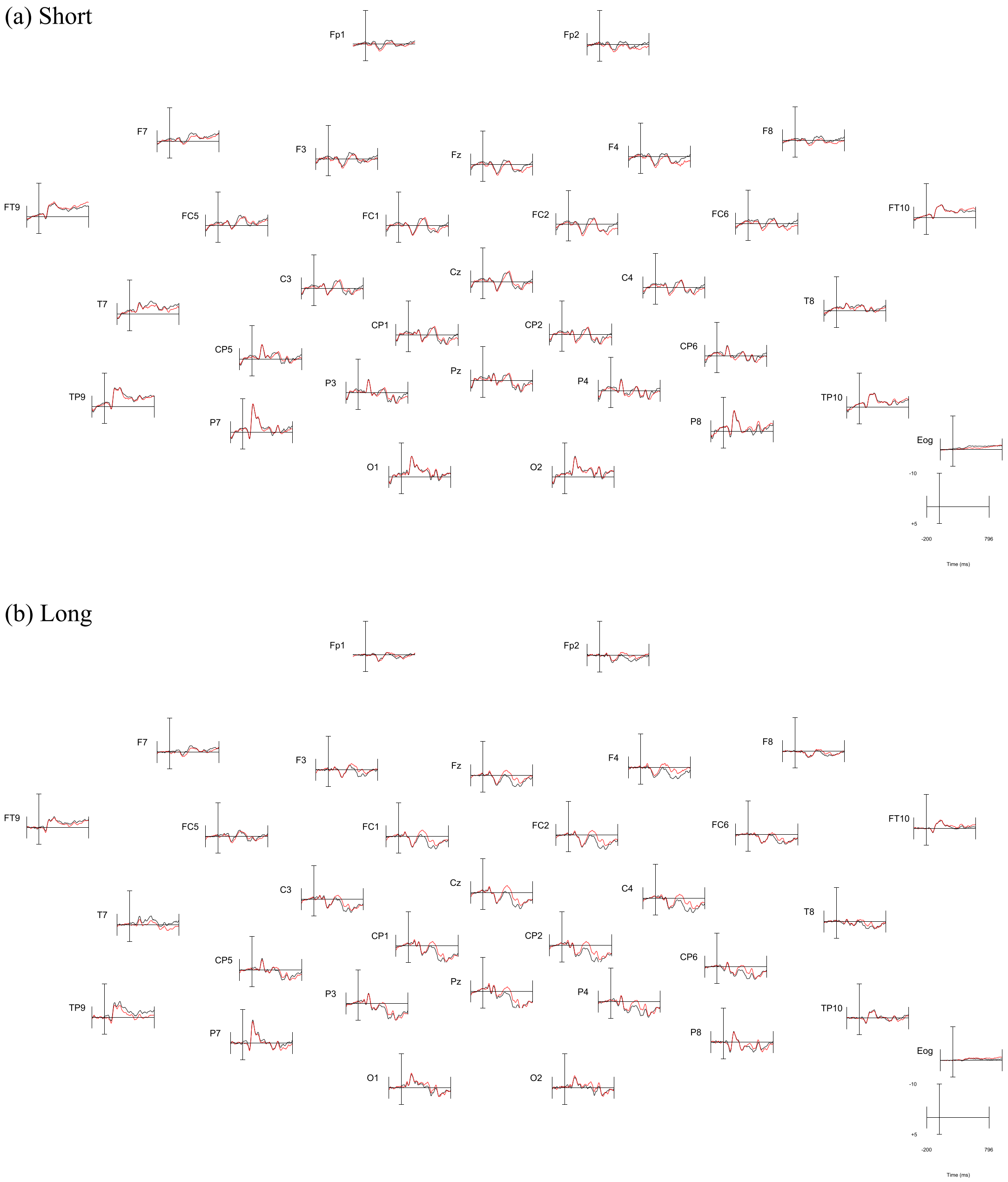


Fig. 4. ERPs of (a) the short condition and (b) the long condition of the experimental items.

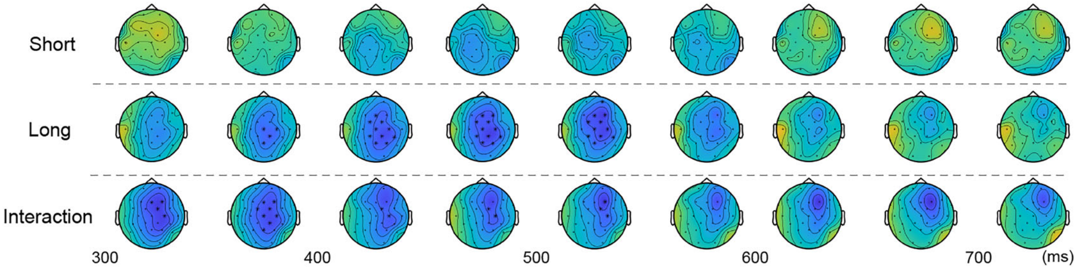


Fig. 5. Topographic maps of ERPs for the experimental items. The top two rows indicate the difference between the plausible and the implausible conditions (implausible minus plausible) for each SOA. The bottom row shows the difference between the top two rows (long minus short). Stars (*) indicate members of significant clusters.

was 600 ms. Since the negativity was frontally distributed, Yano and Sakamoto interpreted it as a LAN effect and argued that it reflected syntactic expectations for intransitive verbs in (8b) that were violated by the transitive verb *simeru* "close." However, the current study offers an alternative interpretation for Yano and Sakamoto's findings. The negativity they observed might be an N400 effect rather than a LAN effect. Although the negativity observed by Yano and Sakamoto was frontally distributed, it was immediately followed by a posterior positivity (P600). Therefore, the negativity in the posterior electrodes might have been canceled by the temporally and spatially overlapping positive component. In fact, we did not observe a P600 effect in our study and the negativity was widely distributed. Thus, Yano and Sakamoto's studies and the current experiment might be consistently showing that additional time prior to verb presentation elicits N400 effects in argument role reversals, an effect that was more clearly observed without an overlapping positive component in our experiment.

(8) a. Mado-o simer-u.
Window-acc close-pres.
"Someone/something closes a/the window."
b. Mado-ga simer-u.
Window-nom close-pres.
"A/the window closes someone/something."

One potential concern is that reversal anomalies with nominative context nouns might, in principle, be able to be repaired by passives or causatives. For example, (9b) is an argument role reversal sentence, but (9c) is a plausible sentence, where the nominative noun is a patient. Therefore, the lack of N400 effects in the short condition might be mainly caused by sentence pairs such as (9a) and (9b), where the same verb could be used in a plausible sentence. On the other hand, it is impossible to turn the argument role reversal sentence (10b) into a plausible sentence through further suffixation on the verb.

(9) a. Hae-o tatak-u.
Fly-acc swat-pres.
"Something swats a/the fly."
b. Hae-ga tatak-u.
Fly-nom swat-pres.
"A/the fly swats something."
c. Hae-ga tatak-arer-u.
Fly-nom swat-passive-pres.
"A/the fly was swatted by something."

(10) a. Hati-ga sas-u.
Bee-nom sting-pres.
"A/the bee stings something."
b. Hati-o sas-u.
Bee-acc sting-pres.
"Something stings a/the bee."

In order to address this concern, we also conducted the same analyses as above, but separately for accusative-plausible stimuli such as (9) and nominative-plausible stimuli such as (10). There were no statistically significant positive or negative clusters for any analyses, presumably due to the reduced data points, but we did find a visually similar pattern in this partial data analysis (see Figs. A1 and A2 in Appendix A). We take this to be consistent with the idea that the absence of an N400 effect in the short condition and its presence in the long condition was not caused specifically by one of the two types of stimuli. This suggests that the possibility of sentences such as (9c) did not contribute significantly to our results.

3. Speeded cloze

3.1. Methods

3.1.1. Participants

Eighty native Japanese speakers were recruited online on the crowdsourcing platform CrowdWorks. Participants received monetary compensation. Seven participants' data were not analyzed because the recordings were missing or unclear, or because they did not perform the task. Additionally, 19 participants were excluded from the data due to unsatisfactory performance as described in the Exclusion section. This left 54 participants for analysis. (39 females and 14 males, 1 did not provide gender information; average age = 41.1, 2 did not provide age information.) Prior consent was obtained for each participant.

3.1.2. Materials

We used 159 case-marked context nouns from the 160 experimental item pairs in our EEG experiment. Unlike the EEG experiment, each context noun did not appear with a target verb, and participants instead provided a verb that was a natural continuation of the presented context. The nouns were presented with either nominative or accusative case. One stimulus was replaced with another item, because the same context noun appeared twice in the EEG experiment. This context noun caused minimal problems for the EEG experiment because it appeared with two different verbs, but these two experimental items would have been identical in the cloze task, where verbs are not presented. No fillers were used for this experiment because we did not directly present reversal anomalies in the cloze task and, therefore, there was minimal risk for participants to realize the aim of the experiment. Participants were asked to start articulating a continuation before 1200 ms after the presentation of the context noun in the short condition and between 1600−2800 ms following the context noun in the long condition. Therefore, the same item appeared in four different conditions (nominative-short, nominative-long, accusative-short, and accusative-long). The 160 stimuli in the four conditions were distributed across four lists and assigned to participants using the Latin Square method.

3.1.3. Procedure

The experiment was conducted via the online experiment platform PCIbex (Zehr & Schwarz, 2018). The experiment consisted of two blocks, and the long condition trials were presented in the first block and the short condition trials were presented in the second block. We separated the long condition trials and the short condition trials in different blocks, rather than randomly interleaving them, because it was unlikely that participants could adjust to production deadlines randomly from one trial to the next. The long condition block always preceded the short condition block because piloting showed that the short condition was very challenging for most participants and it was easier for them to be familiarized with the speeded cloze task through the long condition, in order to be able to meet the strict production deadline in the short condition.

In each trial, a fixation cross was presented on the center of the screen for 800 ms, and then it was replaced with a context noun with a case marker. Participants were asked to produce a natural verb as a continuation of the presented context noun within the response time window. The response time windows were indicated in different ways in the two conditions, so that we could obtain responses in specific intervals. In the short condition, context nouns were presented for 1200 ms and participants were asked to provide responses while the stimuli were being presented. In the long condition, context nouns were presented in a black font color for 1600 ms, then they turned red and remained on the screen for the next 1200 ms, and then disappeared. Participants were asked to produce responses while the font color was red. Their responses were recorded by PCIbex and automatically sent to our server for data collection.

3.1.4. Data processing

The recordings were manually transcribed and coded for onsets by one of the authors using Praat (Boersma & Weenink, 2023), through both listening to the audio and visually inspecting the spectrograms. The coder had access to the context nouns but without case markers, in order to avoid bias in coding. After this process, each response was coded for

Table 1
Average cloze response latencies (ms)

Condition	Average response latency	Standard deviation
Long	2358	199
Short	1175	210
Short < 1200 ms	1021	124

role-appropriateness. Each pair of a context noun and a produced verb was classified into one of the three categories by the same author: (1) clearly more plausible with the nominative case (e.g., *bee* and *sting*); (2) clearly more plausible with the accusative case (e.g., *fly* and *swat*); or (3) no clear contrast between different cases (e.g., *cop* and *see*). Subsequently, each response was automatically coded according to the three categories. Responses produced with the preferred case were coded as role-appropriate, other responses produced with the dispreferred case were coded as role-inappropriate, and the rest were coded as neutral.

3.1.5. *Exclusion*

First, we identified participants who failed to follow the instructions by inspecting their performance in the same 40 experimental items. We excluded participants who produced short condition responses in the long condition time window or vice versa in more than 50% of trials. Because participants can, in principle, complete the experiment by producing the same verb in different trials regardless of the context, we explicitly instructed them to avoid producing the same verb in many trials. In order to detect participants who did not follow this instruction, we counted the number of each verb produced by each participant. We excluded participants who produced the same three verbs in more than 50% of trials. Nineteen participants were excluded by these two criteria, leaving 54 participants for further analysis.

Next, we excluded trials where nonwords, nonverbs, or unrelated words were produced, or no responses were recorded at all. We also excluded trials where participants produced responses in the wrong time window (e.g., producing short condition responses in the 1600–2800 ms long condition time window). We included short condition responses that were produced before the beginning of the long condition time window, because many participants struggled to meet the deadline of the short condition. However, the short condition responses were on average still produced in the expected time window, and the results did not differ if we focus on only the responses produced before the 1200 ms deadline. (See below for more details.) These exclusion processes left 6612 responses, which corresponds to 76.5% of all responses from the 54 participants.

3.2. *Results*

3.2.1. *Response latencies*

Response latencies (speech onset times) are summarized in Table 1. We confirmed that short condition responses including those produced after the deadline (1200 ms) but before the long condition time window (1600 ms) were on average comparable in timing to the short

Table 2
Proportion of role-appropriate, neutral, and role-inappropriate responses

	Long	Short	Short < 1200
Appropriate	74.1%	67.0%	70.8%
Neutral	22.2%	27.0%	24.0%
Inappropriate	3.6%	6.0%	5.2%

Note. $N = 2359$, 2195, 1240 for each column.

condition of our EEG study. Since it is estimated that it takes about 350 ms to start articulating a verb after completing lexical processing, on average the responses reflect activation up to 825 ms after the presentation of the context noun. This is comparable to the short condition of our EEG experiment where participants had an 800 ms interval between the onset of the context noun presentation and the target verb presentation.

3.2.2. Response categories

The proportion of responses in each category is summarized in Table 2. This excludes syntactically irreversible responses, such as intransitive verbs (28.9% of all responses). We made this decision so that we would not underestimate the proportion of role-inappropriate responses. People might avoid producing syntactically irreversible verbs after inappropriate contexts due to syntactic constraints rather than thematic constraints, although we are only concerned with thematic violations. We also excluded passives and causatives, because these types of responses require atypical case markings (2.3% of all responses). As indicated in the table, more than two-thirds of the responses had a clear bias for the roles of the context nouns and they were produced following the more appropriate case marker (e.g., *bee-nom sting-pres*). Combined with the neutral responses, plausible responses accounted for 94% of the responses, and role-inappropriate responses accounted for only 6% of the responses. Importantly, this pattern was consistent across both the short and the long conditions, as well as for the short condition responses that met the 1200 ms deadline. In order to demonstrate that role-appropriate responses are significantly more frequent than role-inappropriate conditions in both the long and the short conditions, we conducted a Pearson's chi-squared test for each condition, excluding the neutral responses. We used the α level of .025 to correct for multiple comparisons. The tests showed that the frequency of role-appropriate and inappropriate responses were significantly different within each condition ($p < .001$). Furthermore, we conducted an analysis using a binomial generalized linear mixed-effects model to compare the proportion of role-inappropriate responses between the long and the short conditions. The model revealed a statistically significant effect of condition ($\beta = 1.04$, $p = .002$). This shows that there were significantly more role-inappropriate responses in the short condition, though the percentages were low in both conditions.

We also computed cloze probabilities for the specific target verbs used in our EEG study with different case markers. This analysis included syntactically irreversible responses, following the ordinary way to obtain cloze probabilities (e.g., Taylor, 1953). As shown in Table 3, the target verbs were mostly produced in the appropriate context, and they were very rarely

Table 3
Cloze probability of the target verbs used in the EEG study

	Long	Short	Short < 1200
Appropriate	19.0%	13.2%	15.6%
Inappropriate	0.6%	1.7%	1.8%

Note. $N = 3430$, 3182, 1772 for each column.

produced as reversals. A binomial generalized linear mixed-effects model revealed a significant effect of appropriateness ($\beta = 4.53$, $p < .001$). It revealed a significant interaction between SOA and appropriateness ($\beta = 1.89$, $p < .001$), showing that the difference between the target verb production in the appropriate and the inappropriate contexts was greater in the long condition.

3.3. Discussion

Our cloze study showed that role-inappropriate responses are very rare both in the short condition and in the long condition. If people solely rely on role-insensitive prediction or mere lexical association in the speeded cloze task, there should be as many role-inappropriate responses as role-appropriate responses. The short condition is particularly important in that it shows that people can use argument roles in the cloze task, even in a time window where we do not observe N400 effects for the same stimuli. Thus, the role-sensitivity in the cloze task cannot be reduced to the additional time allowed in an offline cloze task.

Since participants were not allowed to produce responses before the response time window in the long condition, it is possible that these responses reflect some additional unnatural processes. In fact, there was a small but significant difference between the ratio of role-inappropriate responses in the long condition and the short condition. However, such a potential problem in the long condition does not undermine the main findings of our experiment. First, the most important data point in our cloze experiment is the short condition, where we found role-sensitivity in cloze responses, even though we did not observe N400 effects in the time-matched short condition in our EEG study. Second, the difference between the short and long conditions in the cloze task might be explained by the same mechanism that we propose to account for the role-sensitivity in the short condition of the cloze study or the long condition of the EEG study. There might have been fewer role-inappropriate responses in the long condition in the cloze study because the additional time in the long condition allowed participants to stop themselves from producing role-inappropriate responses before production and produce role-appropriate responses instead. We argue that such prevention of role-inappropriate responses is supported by a monitoring mechanism, and that it is also responsible for the role-sensitivity observed in the short condition of the cloze experiment or the long condition of the EEG study. We discuss this mechanism in more detail in the modeling section below.

It is important to note that 22.1% of the short condition trials had no responses or responses produced later than 1600 ms, and they were excluded from the analyses above. While this is a relatively large proportion, it is not sufficient to explain the large gap between role-appropriate responses and inappropriate responses.

4. Interim discussion

In this section, we discuss the results of the experiments before we proceed to computational modeling. In the EEG study, we replicated the well-known absence of N400 effects in the short condition. Even though *sting* should be expected after *Bee-nom* but not after *Bee-acc*, there was no evidence for an N400 contrast at the verb. More importantly, we found N400 contrasts for the same pairs of stimuli in the long condition. This not only replicates Chow et al. (2018)'s findings but also extends them. Chow et al. (2018) manipulated timing by inserting a temporal adverbial phrase between the arguments and the verb, which could have made the sentences somewhat unnatural to native speakers of Mandarin Chinese. On the other hand, the stimuli were maximally simple in our experiment and the only difference between the short and the long conditions was the additional 400 ms added to the interval between the context noun and the target verb. This provides stronger evidence that the absence of N400 effects in argument role reversals is specific to an earlier stage of processing, prior to presentation of the target verb, and hence that N400 effects change dynamically as a function of time after the presentation of the context, but before the presentation of the target verb.

While we confirmed that N400 amplitude can be insensitive to argument roles at least in the short condition time window, in the speeded cloze task, people produced role-appropriate verbs in almost all cases, both in the short condition and the long condition. The large contrast in cloze probabilities of role-appropriate and inappropriate responses persisted when we limited the analysis to only the target verbs used in the EEG experiment. Importantly, the timing of the short condition was carefully matched between the cloze experiment and the EEG experiment. People had 800 ms to process the context and activate likely continuations in the EEG experiment, while they took about 830 ms on average in the cloze experiment, considering the time required for post-lexical processes such as motor planning (350 ms; Indefrey & Levelt, 2004). Thus, the difference between N400 and cloze measures is not easily reduced to a contrast between online and offline comparisons, as has often been assumed. This puzzling mismatch between measures of context-driven lexical activation raises two questions: (1) Why do we find contrasts between different measures of prediction, even though the timing and the stimuli were carefully matched? (2) Why do earlier and later presentation of continuations result in the presence/absence of N400 effects?

Prior studies and the current study offer clues to answers for these questions. First, prior studies have shown that there are many parallels between N400 amplitude and cloze measures. It has been shown that the measures correlate with each other in most cases (e.g., Kutas & Hillyard, 1984; DeLong, Urbach, & Kutas, 2005; Nieuwland et al., 2020). Furthermore, studies on the linking hypotheses for each measure suggest that they both reflect context-driven lexical activation. Federmeier and Kutas (1999) showed that the N400 component is relatively small for lexical items that are unpredictable and implausible but that are semantically related to predictable lexical items. For example, the N400 amplitude of *pines* is smaller in (11) compared with *tulips*, due to the closer semantic relationship with highly expected *palms*. This is explained if the N400 component reflects the activation of lexical items or semantic representations tied to them, and if that activation can spread to related items. Meanwhile, Staub et al. (2015) showed that the patterns in human cloze probabilities and RTs can be explained by a model where cloze responses reflect parallel activation of lexical items. Second, while it is tempting to assume that both of these measures are transparent reflections of the shared lexical activation, the current findings challenge this assumption. Even though we closely matched the time course and used the same stimuli in the EEG study and the cloze study, we observed clear contrasts between the two studies.

(11) They wanted to make the hotel look more like a tropical resort. So along the driveway, they planted rows of palms/pines/tulips.

Based on these findings, we offer two hypotheses to account for the timing-based and task-based contrasts in lexical prediction using argument role information. The *faulty prediction hypothesis* attributes role-insensitivity to the context-driven lexical activation of verbs. Initially, role-inappropriate verbs are activated to a similar extent as role-appropriate verbs. However, they are successfully inhibited in production tasks, or in comprehension tasks with sufficient time, through a mechanism that monitors the plausibility of the utterance. The *faulty integration hypothesis* (e.g., van Herten et al., 2005; Bornkessel-Schlesewsky & Schlesewsky, 2008) attributes role-insensitivity to semantic integration after accessing the lexical representations. Role-inappropriate continuations are not (pre-)activated and hence they are difficult to access, but it is just as easy to incorporate them into the combinatorial semantic representations as role-appropriate verbs, and N400 amplitudes reflect such role-insensitive integration processes rather than lexical access.

Our experimental results are more consistent with the faulty prediction hypothesis. In our EEG experiment, more time was provided before the presentation of the target verb in the long condition rather than after it. That is, the amount of time allowed for processes after the presentation of verbs was identical in the short and the long conditions. Since the faulty integration hypothesis attributes role-insensitivity to processes that occur after the presentation of verbs, these two hypotheses have to explain why the additional time to process the very simple contexts resulted in role-sensitive processing of the role-incompatible verbs. On the other hand, the faulty prediction hypothesis attributes the role-insensitivity to context-driven lexical activation, and the long condition indeed provides more time for this process. Therefore, the timing contrast can be more straightforwardly explained by this account.

Thus, the faulty prediction hypothesis seems to be more consistent with our data, and in fact there are relevant prior computational models of predictive language processing that are informed by argument role reversals (Sentence Gestalt model: Rabovsky et al., 2018[2]; retrieval-integration account: Brouwer et al., 2012, 2017; multirepresentational hierarchical actively generative architecture: Kuperberg & Jaeger, 2016 and Kuperberg, 2016; noisy channel approach: Li & Ettinger, 2023). These models focus on the widely found absence of N400 effects in argument role reversals, and they assume that predictive processes are dominated by role-insensitive components or by heuristic components. However, these models tend to regard argument role reversals as a single data point to be captured, while our study demonstrates that there are richer generalizations to be captured, namely, the timing and task effects. The prior models update their predictions in a discrete step following each new input and, therefore, they cannot easily handle the timing contrast we observed between the short and long conditions in the EEG study, where predictions seem to be updated without any obvious new input. Additionally, these models focus on the ERP data and do not account for basic findings in the speeded cloze task, such as the effect of competitors found by Staub et al. (2015).

While the models introduced above do not straightforwardly handle our task and timing effects, the Race model by Staub et al. (2015) has attractive features, since it is designed to account for the speeded cloze data, and since it can model the dynamic accumulation of activation after the presentation of the context. Although the original Race model does not explain the new contrasts we obtained, we propose that it can account for the task and timing effects if a monitoring mechanism is added. In this version of the model, role-inappropriate candidates are activated as much as role-appropriate candidates by the context. Once a candidate's activation reaches a predetermined threshold, a monitoring mechanism that checks for appropriateness is triggered, and role-inappropriate candidates are then detected and inhibited. This account links sensitivity to argument roles to how often role-incompatible candidates reach the threshold within a specific time limit, thereby triggering the monitoring mechanism, leading to successful detection and inhibition. Either additional time for activation or a lowered threshold to allow rapid production of a continuation would allow role-inappropriate candidates to reach the threshold in higher proportions. We suggest that this monitoring mechanism is related to the monitoring process widely assumed in language production. Levelt (1983) claimed that people monitor their own speech in speech production and repair their utterance when an anomaly is detected. This repair can be done not only on overt speech but also on internal speech that is planned but not yet articulated. The monitoring mechanism in lexical prediction can be considered as one form of such covert monitoring of speech in language production, where the plausibility of the planned utterance is checked.

5. Modeling

5.1. Model descriptions

We developed a computational model of context-based activation of items, building on the Race model for the speeded cloze task (Staub et al., 2015). Given a context, lexical items are activated in parallel. The amount of activation each lexical item accumulates is determined by some function of the goodness of fit of the lexical item given the context, plus some noise. The candidate lexical item that reaches a threshold first (i.e., the "winner" of the race) is produced in the cloze task. Because there is some amount of random variation in the activation, the most likely continuations are produced often but not all the time.

We take the basic race model and generalize and extend it so that it can account for our findings in argument role reversals. First, we generalize the model to also account for the

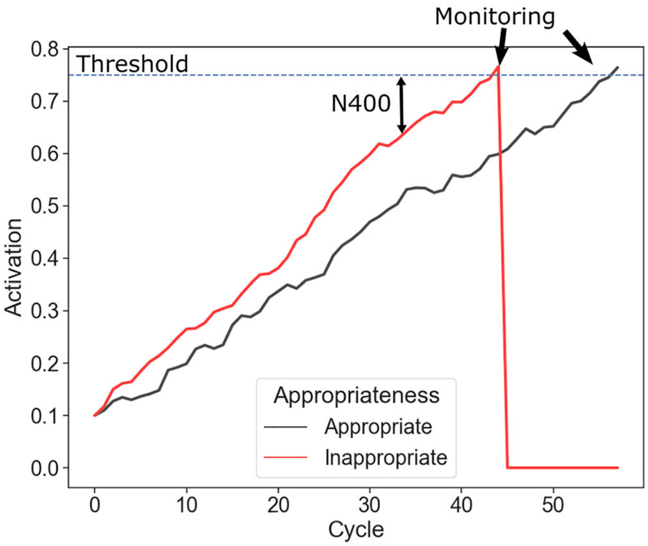


Fig. 6. An illustration of a simulation of the model. Candidate lexical items are activated in parallel and are monitored serially as their activation levels reach the threshold. We simulated step-by-step accumulation of activation. There is a normal distribution corresponding to each candidate's step-by-step activation, and the amount of activation each candidate accumulates in each step is determined by sampling from the normal distribution. Lexical items that fit to the context well have large values for the mu parameter (mean) of the normal distributions. Once a candidate reaches the threshold, a monitoring process is triggered and the candidate is inhibited if it is inappropriate. N400 amplitudes are operationalized as the distance between the threshold and the amount of activation at the point of presentation of the continuation.

EEG studies with comprehension tasks. In comprehension tasks, lexical items are activated in the same way as in the cloze task, until a continuation is presented. N400 amplitudes are operationalized as the distance between the threshold and the amount of activation when the continuation is presented (Fig. 6).

We also extend the model to capture step-by-step activation and inhibition dynamics. The original model by Staub et al. focused less on the details of the activation process and more on the time needed for a lexical item to reach the threshold for production. For this reason, Staub et al. used finishing time distributions as a proxy for the outcome of the activation process. Our approach allows us to model the temporal dynamics of the activation process itself, incorporating the monitoring mechanism and accounting for our empirical findings. More importantly, our model has a monitoring mechanism that inhibits the activation of inappropriate candidates that have reached the threshold. The candidates are activated in parallel, but then a monitoring mechanism evaluates each candidate in a serial manner, as soon as it reaches the activation threshold.

In the model, each candidate's activation is updated in each 25 ms cycle. The amount of activation each candidate accumulates is randomly determined by sampling from normal distributions. While the sigma parameter (i.e., variability) of the normal distributions is kept constant, each candidate has different values for the mu parameters (i.e., mean). Stronger

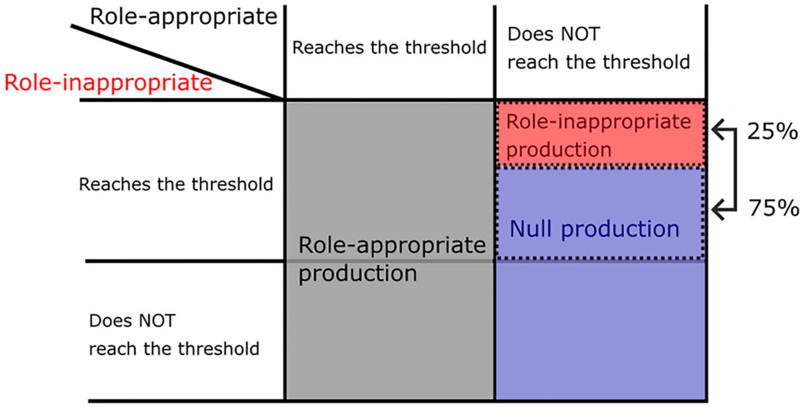

Role-appropriate
Role-inappropriate
Reaches the threshold
Does NOT reach the threshold
Reaches the threshold
Does NOT reach the threshold
Role-appropriate production
Role-inappropriate production
Null production
25%
75%

Fig. 7. Illustration of how the model produces role-appropriate, inappropriate, and null responses. Role-appropriate candidates are produced when they are the first to reach the threshold (the bottom-left cell), or when they reach the threshold after role-inappropriate candidates reach the threshold and are rejected (top-left cell). When no role-appropriate candidate reaches the threshold but a role-inappropriate candidate does reach the threshold (top-right cell), role-appropriate candidates are produced with 25% probability (red cell), and no responses are produced in the remaining 75% of trials (blue cell). There would also be no response if neither role-appropriate nor inappropriate candidates reach the threshold (bottom-right cell), but there was no such case in our simulation.

candidates have greater values for the mu parameters, so that they are likely to accumulate more activation in each step.

Once a candidate reaches the threshold, a monitoring process checks for its appropriateness in the context. When the winner is role-inappropriate, its activation is reduced to 0.[3] In comprehension tasks, this process of activation and monitoring continues until a continuation is externally presented. On the other hand, in a cloze task, the process ceases once a role-appropriate candidate reaches the threshold and passes the monitoring process. That candidate is produced as the cloze response.

In this model, the rare role-inappropriate responses in the speeded cloze task are considered as the last resort when no role-appropriate candidate reaches the threshold within the available response time window. In such cases, if there is a role-inappropriate candidate that has reached the threshold but is rejected through the monitoring process, the model may produce the role-inappropriate candidate rather than not producing anything. The detailed processes are described below and in Fig. 7.

While the model has identical activation processes for the speeded cloze task and for EEG studies with comprehension tasks, the difference between the tasks is operationalized via the activation threshold for monitoring. We propose that people set a lower threshold for the speeded cloze task compared to the comprehension task in the EEG studies, due to the task demands in the speeded cloze task. In our study, since participants have to decide on a single response within a specific time window in the speeded cloze task, they aim to have a lexical item that reaches the threshold before the end of the response time window. If they lower the threshold, there is more chance that they have a lexical item available to produce within the response time window. On the other hand, there is no reason to lower the threshold in the comprehension task because there is no such pressure to quickly decide on a single candidate, and participants can maintain multiple candidates until they see a continuation. The Race model is a variant of Drift Diffusion models (Ratcliff, 1979), but it is known that the task demands for quick decisions can be captured by lowered thresholds in Drift Diffusion models. Ratcliff and McKoon (2008) showed that response latencies and accuracy in decision tasks under different demands for response speed can be explained by changing the threshold of the model.

In our simulation, we focused on the contrast between the short condition of our EEG study and (1) the long condition of the same study, and (2) the short condition of our cloze study. This is because the absence of N400 effects in the short condition is consistent with many preceding studies, and the other two cases diverge from the general pattern and represent timing effects and task effects. We simulated 10,000 races for five pairings of role-appropriate and inappropriate candidates with the same mu parameters (mu = 0.017, 0.0145, 0.012, 0.0095, 0.007) and the same sigma parameter (sigma = 0.01). We set different thresholds for the comprehension task (0.75) and the production task (0.6).

In the comprehension task, the model was presented with the strongest role-appropriate or inappropriate candidate (i.e., the candidate with the strongest mu parameters of 0.017) at either 800 ms (i.e., 32nd cycle) or 1200 ms (i.e., 48th cycle) after the beginning of the race, which corresponds to the short and long conditions of our EEG experiment. The strongest role-appropriate candidates correspond to the modal cloze responses under the current model, given a sufficient number of trials, and the strongest role-inappropriate candidates (i.e., the candidate with the greatest mu parameter) can be considered as the most tempting lures. The N400 amplitude was modeled by subtracting the activation of the presented candidate from the threshold (0.75) at either timing. In the production task, if no role-appropriate candidate has reached the threshold by the deadline but if there is a role-inappropriate candidate that has reached the threshold, the model produces the role-inappropriate candidate 25% of the time but does not produce anything in the remaining 75% of the cases (Fig. 7). We made this assumption to account for the trials in which participants did not provide any responses in our speeded cloze experiment. The model's response (or nonresponse) was recorded for each trial, and we computed the proportion of role-appropriate and role-inappropriate responses. Since our aim was to propose a proof-of-concept model that can account for the novel data patterns we obtained, and our aim was not to perform quantitative comparisons with other models, we did not search for an optimal set of parameters that best explains our data. Instead, we picked a set of parameters (the mu and sigma of each candidate, the threshold for each task, and the rate of producing no responses when no role-appropriate candidate reaches the threshold) that generates a data set that is reasonably similar to our experimental data in terms of the proportion of role-appropriate and inappropriate responses we report below and the proportion of no-response/late-response trials in the short condition of the cloze task (human data: 22.1%; simulated data: 22.6%).

The model predicts that either additional time or a lowered threshold in the production task should result in more successful inhibition of role-inappropriate candidates. As illustrated in Fig. 8a, additional time for activation allows role-incompatible candidates to reach the threshold and be detected and inhibited by the monitoring process. Similarly, Fig. 8b shows

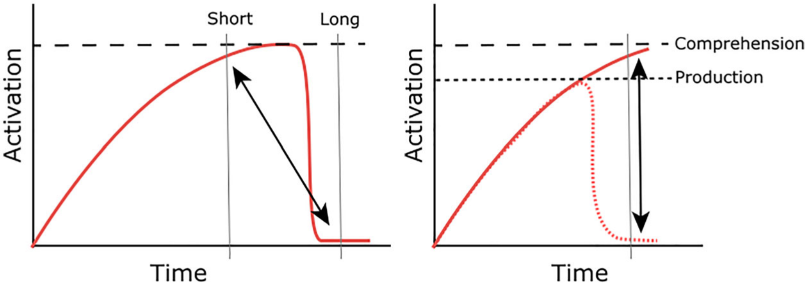


Fig. 8. Activation pattern of role-inappropriate candidates with (a) shorter versus longer time for activation, and (b) higher versus lower threshold for monitoring.

that a lowered threshold in the production task also allows more successful inhibition of role-inappropriate candidates given the same amount of time.

In this version of the model, competing items accumulate activation independent of one another, with no lateral inhibition between them, as also assumed by Staub et al. (2015). This contrasts with evidence for lateral inhibition mechanisms in models of lexical access in comprehension and production (Dahan, Magnuson, Tanenhaus, & Hogan, 2001; Chen & Mirman, 2012). Some recent studies have argued that a lateral inhibition mechanism is also needed to account for speeded cloze findings (Ness & Meltzer-Asscher, 2021b; Nakamura & Phillips, 2024). The current model leaves aside this mechanism in the interest of simplicity of exposition. However, we also conducted simulations that incorporated a lateral inhibition mechanism, yielding comparable results (see Appendix B).

5.2. *Simulation results*

We report the model-simulated N400 effects and the proportion of role-appropriate cloze responses. Fig. 9 shows the distance between the threshold and the activation of the presented candidates, which is assumed to be reflected in N400 amplitudes. The contrast between role-appropriate and inappropriate candidates was smaller for the short condition and greater for the long condition, successfully capturing the N400 patterns in our EEG experiment.

The model prediction might seem to be different from our EEG data because the simulated N400 amplitude for the implausible-long condition is much greater than the implausible-short conditions, although they seem to have matched N400 amplitudes in our EEG experiment. However, as we discussed in the Data analysis section of the EEG experiment, direct comparisons of the N400 amplitudes of the implausible-long and implausible-short conditions or between the plausible-long and plausible-short conditions can be misleading because the ERPs elicited by the context nouns might affect the long conditions and the short conditions differently. Therefore, the matched N400 amplitudes observed in the implausible-long and implausible-short conditions do not necessarily indicate that the target verbs were activated similarly in these two conditions. Since the simulated N400 amplitudes do not consider the effects from the context nouns, we argue that the simulation does not stand in contrast with our empirical data.

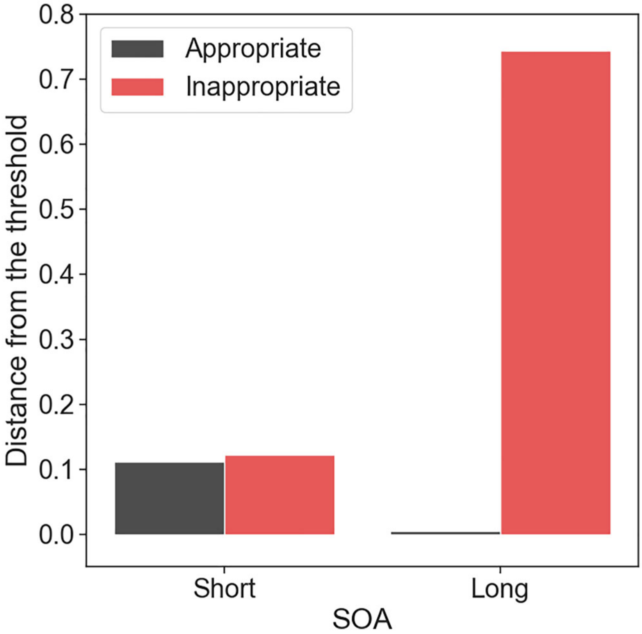


Fig. 9. Simulated N400 amplitudes with shorter (left) versus longer (right) time for activation (left).

In the production task, the responses generated by the model were overwhelmingly role-appropriate (91.7%) and role-inappropriate responses were rare (8.3%). This is comparable to our speeded cloze data, where role-appropriate responses and neutral responses accounted for 94% of the responses and role-inappropriate responses accounted for 6% of the responses.

6. General discussion

The absence of an N400 effect in argument role reversals has been taken to show that people do not use argument role information to activate upcoming lexical items. Therefore, the lack of N400 effects in argument role reversals has been considered as an exception to the generally successful use of all available contextual information to constrain lexical expectations (e.g., Altmann & Mirković, 2009).

We focused on two cases where measures of lexical prediction may show sensitivity to argument roles, despite the consistently reported lack of role-sensitivity in N400 and some eye-tracking measures. That is, additional time for context-driven activation in a comprehension task, and a production task (the speeded cloze task) both lead to role-sensitivity in prediction measures. We examined these cases by making use of simple Japanese stimuli and by running closely matched EEG and cloze experiments. The experiments replicated the widespread finding that role reversals fail to elicit an N400 effect, when the verb appears shortly after the context argument(s). They also corroborated the finding that role reversals elicit an N400 effect when there is additional time between the context argument(s) and the verb. We also found clear evidence for role sensitivity, even without additional time, when it was measured via a speeded cloze task. Thus, there are two closely related contrasts in argument role (in)sensitivity to be explained. One is about timing, and the other is about measures.

We demonstrated that the two apparently puzzling findings can be simultaneously addressed by incorporating a monitoring mechanism in the context-driven lexical activation process, a mechanism that is independently motivated and is a common component of language production models. We argued that context-driven activation is initially insensitive to argument roles, but that role-sensitivity arises when role-inappropriate candidates reach a threshold and are detected by a monitoring mechanism and successfully inhibited. Under this assumption, the timing and task effects can be explained by a single, independently motivated parameter adjustment. Role-inappropriate candidates are more likely to be detected and inhibited by the monitoring mechanism when additional time is available for activation to accumulate toward a threshold in a comprehension task, or when the threshold for monitoring is lowered due to the task demands of a speeded cloze paradigm to quickly produce a continuation.

The task and timing effects we found require models of predictive language comprehension to be refined in two ways. First, lexical prediction should be updated with a finer temporal grain size than once per word, not necessarily requiring new inputs for each update. This is supported by our EEG experiment, where N400 amplitudes showed a clearly different pattern when an additional temporal interval was added without any obvious new input. This challenges an assumption shared by prior models of predictive processing, whereby predictions are updated as new inputs are encountered. Second, a complete model of language prediction must take different task demands into account, as we have attempted to do via differences in thresholds in our model. Prior models have tended to focus on ERPs or comprehension measures such as reading times, and they have paid less attention to production measures such as speeded cloze. For those models, the drastic differences that we found between those tasks are striking, given the independently developed linking hypotheses of the speeded cloze measures (Staub et al., 2015) and the N400 component (e.g., Lau et al., 2008). The contrasts may be even more surprising considering a line of literature that highlighted the similarities between prediction and production (e.g., Pickering & Garrod, 2007).

6.1. Relationship to existing work: Experiments

We found an absence of N400 effects at verbs in Japanese argument role reversal sentences, consistent with earlier findings by Oishi and Sakamoto (2010). Japanese is a strictly verb final language and Japanese speakers have life-long experiences of hearing/reading verbs in sentence final position. Therefore, it is surprising that cases/roles of arguments still do not seem to affect their predictions about upcoming verbs, unless the verb appears after a longer delay.

Our study strengthens the evidence from Chow et al. (2018) that additional time yields role sensitivity in EEG paradigms. An advantage of the simple Japanese sentences used in our study is that they made it possible to implement a "pure" timing manipulation, in which the short and long conditions differed only in the length of the time interval between the context noun and the target verb, without the need to make any other word order adjustments. These effects of the timing manipulation challenge the view that the lack of N400 effects in argument role reversals reflect misinterpretation of the sentence (e.g., Kim & Osterhout, 2005) or uncertainty in the interpretation (Rabovsky et al., 2018) caused by attractive nonliteral interpretations (e.g., "A bee stings something" for *Bee-ACC sting*.). As mentioned earlier, there cannot be long-lasting misinterpretation or uncertainty, given the high accuracy in plausibility judgments or comprehension questions following argument role reversals. But even transient misinterpretation or uncertainty is unlikely given the timing effects. The accounts that assume transient misinterpretation attribute the lack of N400 effects to misinterpretation or uncertainty *after* verb presentation, but then it is not clear why additional time *before* verb presentation should affect the misinterpretation. These accounts might explain the timing effect by arguing that comprehenders have already misinterpreted the context or had uncertainty about the correct role assignment prior to verb presentation and that additional time resolves those issues. For example, *bee-ACC* might be misparsed as if *bee* is the subject of the clause. However, this seems unlikely since our stimuli were extremely simple and should have been processed easily, even in the short condition, where participants had 800 ms between the context noun and the verb. Alternatively, comprehenders might be uncertain about the argument role of *bee* given *bee-ACC* due to a bias toward bees being agents (Stone & Rabovsky, 2024). However, it is unlikely that the single argument contexts used in our experiment, as opposed to a combination of arguments such as *bee* and *boy*, create a strong bias toward the bee being an agent that overrides the clear structural cue provided by the case marker. Moreover, it is not clear why additional time should reduce such anomalous role assignments or promote the structural cues.

Meanwhile, our finding of role sensitivity in speeded cloze responses is also primarily a finding about timing. It is well known that people can give appropriate, role sensitive completions in standard untimed cloze tasks. These have been used as a source of norming data for many prior EEG studies. What is novel in our findings is that the same role sensitivity is also found when speakers have much less time to respond, even when it is the same amount of time as in the EEG study.

Our findings are relevant to existing accounts of the lack of N400 effects in role-reversal sentences. One approach attributes the early role insensitivity in the N400 component to a failure to parse the context fast enough to identify the argument roles of the items in the context. Chow, Momma, et al. (2016) questioned this approach because P600 effects are usually found in argument role reversals. Whether the P600 effects reflect increased semantic integration cost (e.g., Brouwer et al., 2017) or syntactic reanalysis (e.g., van Herten et al., 2005), they suggest that the processing mechanisms do successfully detect the semantic anomaly in argument role reversals, which requires identification of argument roles. Consistent with Chow and colleagues' proposal, our findings cast further doubt upon this approach because we find the same role insensitivity in the N400 component following exceedingly simple contexts, which speakers are unlikely to struggle to parse. In fact, participants successfully judged the argument role reversal sentences to be implausible in our experiment. Furthermore, we find role sensitivity in responses to the same context, on the same time scale, when the task involves a speeded cloze response. These findings show that comprehenders are able to parse argument role reversal sentences and identify the argument roles quickly. Another family of approaches attributes role insensitivity in the N400 component to the consideration of unsupported but highly attractive verb-argument pairings. A variant of this approach, proposed by Kuperberg (2016), proposes that argument role information in the context is selectively ignored or down-weighted in prediction processes, specifically in cases where ignoring the role information more readily yields a continuation than does taking the role information into account. For example, just the set (or a sequence) of {waitress, customer} may yield *serve* as an accessible candidate for the continuation, because it is very likely for a waitress to serve a customer. On the other hand, the full set of cues {waitress-patient, customer-agent} may not yield accessible candidates as continuations because there may not be very likely events involving customer-as-agent and waitress-as-patient. In that case, people may rely more on the mere set or sequence of the arguments but not with their roles because there would be likely candidates. This strikes us as an unlikely account of the role insensitivity observed in our Japanese materials, where each context consists of a single noun and a case marker. In such a simple context, only the identity of a single argument is available as a cue for prediction if comprehenders ignore the argument roles. It seems unlikely that people can generate more accessible continuations given a single argument without a role such as *bee*, than they can generate when given the combination of an argument and its role, such as *bee-as-patient*.

6.2. *Relationship to existing work: Modeling*

6.2.1. *Accounts of the absence of N400 effects*

We highlighted two approaches to the contrasting profiles of role (in)sensitivity in our studies. We used an implemented computational model to simulate one of these approaches, in which initial activations are role insensitive, and where role sensitivity is the result of a serial monitoring process. This simulation not only successfully explains why people show sensitivity to argument roles in (speeded) cloze studies and in EEG studies with additional time prior to verb presentation, it is also consistent with existing accounts that assume initial role-insensitive activation of lexical candidates.

Brouwer et al. (2012, 2017) attribute insensitivity to lexical activation caused by simple association and scenario-based world knowledge. Under their account, *sting* is activated by *bee-acc* because (1) *bee* and *sting* are associated, and (2) *sting* is likely to appear in a scenario where a *bee* is involved. The role-sensitive monitoring mechanism in our model can be understood as an additional process that constrains this association-based activation.

A related but slightly different account by Chow, Momma, et al. (2016) attributes the role-insensitivity of N400 to event memory search. They suggest that lexical prediction requires retrieval of likely events from long-term memory, and that argument roles might not be able to be used for retrieval of event memories. As one possible mechanism for this, they suggest that thematic roles might be encoded in event-specific manners such as *stinger* or *stingee* rather than in terms of abstract thematic roles such as agent or patient. However, when comprehenders predict verbs using arguments, they do not have access to such event-specific roles but only to event-general roles such as agent or patient, because verbs are absent. Thus, there is a mismatch between the thematic role representations available to comprehenders in verb prediction and the thematic role representations in the memory. Such a mismatch would only allow memory search using the arguments without roles as cues, and hence both role-appropriate and inappropriate verbs are generated as candidates. Chow, Momma, et al. argued that events retrieved by such a parallel role-insensitive memory search can be rejected by serially evaluating each of them, which requires additional time. Our model can be taken as an extension of their suggestion that provides a concrete mechanism for this rejection process and simultaneously accounts for the cloze data.

6.2.2. The monitoring model and preupdating

Our assumption of a serial monitoring process that leads role-inappropriate candidates to be inhibited looks like a departure from other existing models, but we see a number of precedents. Although accounts of prediction in language processing tend to focus on the parallel activation of multiple promising candidates, this cannot be the whole story. As we showed through computational modeling, apparently surprising contrasts involving task and timing can be explained by an additional step that applies to specific individual candidates following parallel activation. In production, Staub et al. (2015) Race model assumes the notion of an activation threshold, corresponding to the point at which a lexical candidate has accumulated sufficient activation for it to be ready for phonological processing. In this approach, the first item to reach the threshold is the one that gets articulated. The only addition in our account here is the suggestion that reaching this threshold also triggers a monitoring process, which is a standard component of production models (e.g., Levelt, 1983).

Similarly, in comprehension, simply preactivating multiple lexical continuations cannot be enough. There must eventually be a point at which an individual item is integrated into the representation of the context. This could happen at the point when the actual incoming word is presented. But there is also interesting evidence that this process sometimes occurs before the presentation of continuations, in situations in which the comprehender has a high degree of confidence in an individual predicted continuation. Such early integration is sometimes called *preupdating* (e.g., Lau, Holcomb, & Kuperberg, 2013; Ness & Meltzer-Asscher, 2018, 2021a). Some prior studies have found P600 effects at words preceding highly predicted continuations. Referring to prior studies that observed P600 effects when wh-dependencies are resolved (e.g., Felser, Clahsen, & Münte, 2003), Ness and Meltzer-Asscher argued that the P600 component reflects the integration of words and that the P600 effects at words preceding highly predicted continuations reflect integration of the predicted continuations prior to their presentation. Our model is consistent with this literature, in that a high degree of activation for predicted candidates triggers subsequent processes. The monitoring process can be taken as a part of this preupdating process, where the lexical item is not just integrated into the semantic representation but is also evaluated for the plausibility.

The parallelism between our model and preupdating makes the empirical prediction that there should be a late positivity following the context noun of the plausible-long condition of the EEG study compared to that of the plausible-short condition, because additional time should allow role-appropriate candidates to reach the threshold and to be integrated into the sentential semantic representations. However, our current data set cannot assess this prediction because if additional time allows more preupdating, that should mostly occur 800–1200 ms after the presentation of the context noun. We do not have EEG data for this time window for the short condition, because participants had already been presented with the target verb by the time. Therefore, the examination of this empirical prediction is left for future studies.

There is, in fact, a recent finding that supports the existence of a shared monitoring mechanism in speeded cloze tasks and in the comprehension tasks used in EEG studies. Lee and Phillips (2023) observed N400 effects for argument role reversals in an EEG experiment that interleaved comprehension trials with speeded cloze trials. Two points should be highlighted about their experiment. First, participants did not know if they were in a cloze trial or a comprehension trial until they were asked to provide a continuation or were shown a sentence completion. Second, Lee and Phillips observed these N400 effects in comprehension trials using the same stimuli as a prior EEG study on argument role reversals. This is consistent with our model, which attributes the differences between the cloze studies and the EEG studies to differences in the threshold for the monitoring process. The interleaved cloze trials would generally lower the threshold because participants did not know which type of trials they were in. Hence, role-inappropriate candidates in the comprehension trials, as well as the cloze trials, would be detected and inhibited by the monitoring mechanism.

We proposed the monitoring process in order to account for our findings in argument role reversals, but our model makes testable predictions for other cases of N400 neutralization. For example, words heavily repeated in the discourse context elicit a similarly small N400 as predictable continuations even when they appear as unpredictable continuations, such as *suitcase* compared with predictable *tourist* in (12) (Nieuwland & van Berkum, 2005; Aurnhammer, Delogu, Brouwer, & Crocker, 2023). In another case, things that are likely to be involved in the situation elicit a small N400 even if they are unpredictable in the context. For example, Metusalem et al. (2012) found that the N400 amplitude for *jacket* in (13) was larger than *snowman* but was smaller than *towel*. In these cases, unpredictable continuations seem to be activated due to repetition or involvement in the event. Since these lexical items are both implausible in the contexts, they should be detected and inhibited by the monitoring mechanism. Therefore, our model predicts that these lexical items should be (1) very rarely produced in a speeded cloze task and (2) elicit greater N400 effects when additional time is provided before the target word in an EEG study.

(12) Next, the woman told the tourist/suitcase ...

(13) A huge blizzard ripped through town last night. My kids ended up getting the day off from school. They spent the whole day outside building a big snowman/jacket/towel in the front yard.

7. Conclusion

In the current study, we examined how argument roles affect the activation of upcoming lexical items. Our closely matched EEG and cloze studies replicated the N400 amplitudes' insensitivity to argument roles in the short condition of the EEG study, but also showed that people show sensitivity to argument roles (1) with additional time in the same EEG study or (2) at the same moment in a speeded cloze paradigm. This raises the question of why different measures of context-driven activation or different timing result in different sensitivity to argument roles. These findings require models of language prediction that have fine-grained temporal dynamics and that implement the task demands of comprehension and production tasks, features that prior models lack. Our model provides one potential solution to the puzzle that explains both timing and task effects simultaneously, where both additional time or task demands in the speeded cloze study enable successful detection and inhibition of the role-inappropriate candidates.

Acknowledgments

This research was supported by the National Science Foundation grants DGE-0801465, DGE-1449815, BCS-1530332, and BCS-2240434, and Office of Naval Research grant N00014-18-1-2670. Open access was funded by Saarland University based on DEAL.

We are grateful to Wing Yee Chow for sharing the data of a prior study that triggered the current research. We also thank Eun-Kyoung Rosa Lee for the discussions that helped us develop our model.

Open access funding enabled and organized by Projekt DEAL.

Notes

[1] Materials, data, and codes for all experiments and simulations are available at https://osf.io/b4h6q/?view_only=47a44aa6f4d947f0af97482985e3bb15.

[2] Sentence Gestalt model proposed by Rabovsky, Hansen, and McClelland (2018) could be seen as a version of integration hypotheses in that the N400 amplitudes reflect updates in the combinatorial semantic representations and role-inappropriate verbs induce small effort to update the representations. However, in this paper, we treat it as a version of faulty prediction hypothesis in that it makes role-insensitive expectations about the sentence meaning and upcoming verbs in argument role reversals.

[3] We made this decision to make predictions about the N400 amplitudes. An alternative approach would be to completely throw out the rejected candidates rather than to modify the activation. Although it is a theoretical possibility, we did not adopt this approach because it would not make any predictions about the N400 amplitudes under our linking hypothesis. On the other hand, as long as the activation is reduced, the reducing the activation to 0 or other small values should not change the results of our simulations significantly.

References

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264. https://doi.org/10.1016/S0010-0277(99)00059-1

Altmann, G. T. M., & Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science*, *33*(4), 583–609. https://doi.org/10.1111/j.1551-6709.2009.01022.x

Aurnhammer, C., Delogu, F., Brouwer, H., & Crocker, M. W. (2023). The P600 as a continuous index of integration effort. *Psychophysiology*, *60*(9), e14302. https://doi.org/10.1111/psyp.14302

Bicknell, K., Elman, J. L., Hare, M., McRae, K., & Kutas, M. (2010). Effects of event knowledge in processing verbal arguments. *Journal of Memory and Language*, *63*(4), 489–505. https://doi.org/10.1016/j.jml.2010.08.004

Boersma, P., & Weenink, D. (2023). *Praat: Doing Phonetics by Computer* (6.1.55). Retrieved from http://www.praat.org/

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

Brouwer, H., Crocker, M. W., Venhuizen, N. J., & Hoeks, 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

Brouwer, H., Fitz, H., & Hoeks, J. (2012). Getting real about semantic illusions: Rethinking the functional role of the P600 in language comprehension. *Brain Research*, *1446*, 127–143. https://doi.org/10.1016/j.brainres.2012.01.055

Burnsky, J. (2022). *What did you expect? An investigation of lexical preactivation in sentence processing*. University of Massachusetts Amherst.

Chen, Q., & Mirman, D. (2012). Competition and cooperation among similar representations: Toward a unified account of facilitative and inhibitory effects of lexical neighbors. *Psychological Review*, *119*(2), 417–430. https://doi.org/10.1037/a0027175

Chow, W.-Y., Kurenkov, I., Kraut, B., & Phillips, C. (2015). How predictions change over time: Evidence from an online cloze paradigm. The 28th Annual CUNY Conference on Human Sentence Processing. Los Angeles, CA. https://doi.org/10.17605/OSF.IO/SCB5M

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*, *33*(7), 803–828. 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 and Neuroscience*, *31*(5), 617–627. https://doi.org/10.1080/23273798.2016.1160135

Chow, W.-Y., Smith, C., Lau, E., & Phillips, C. (2016). A "bag-of-arguments" mechanism for initial verb predictions. *Language, Cognition and Neuroscience*, *31*(5), 577–596. https://doi.org/10.1080/23273798.2015.1066832

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525x12000477

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. M. (2001). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes*, *16*(5–6), 507–534. https://doi.org/10.1080/01690960143000074

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117–1121. https://doi.org/10.1038/nn1504

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, *20*(6), 641–655. https://doi.org/10.1016/S0022-5371(81)90220-6

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, *44*(4), 491–505. https://doi.org/10.1111/j.1469-8986.2007.00531.x

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, *41*(4), 469–495. https://doi.org/10.1006/jmla.1999.2660

Felser, C., Clahsen, H., & Münte, T. F. (2003). Storage and integration in the processing of filler-gap dependencies: An ERP study of topicalization and wh-movement in German. *Brain and Language*, *87*(3), 345–354. https://doi.org/10.1016/S0093-934X(03)00135-4

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

Indefrey, P., & Levelt, W. J. M. (2004). The spatial and temporal signatures of word production components. *Cognition*, *92*(1), 101–144. https://doi.org/10.1016/j.cognition.2002.06.001

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, *49*(1), 133–156. https://doi.org/10.1016/S0749-596X(03)00023-8

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

Kukona, A., Fang, S.-Y., Aicher, K. A., Chen, H., & Magnuson, J. S. (2011). The time course of anticipatory constraint integration. *Cognition*, *119*(1), 23–42. https://doi.org/10.1016/j.cognition.2010.12.002

Kuperberg, G. R. (2016). Separate streams or probabilistic inference? What the N400 can tell us about the comprehension of events. *Language, Cognition and Neuroscience*, *31*(5), 602–616. https://doi.org/10.1080/23273798.2015.1130233

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32–59. https://doi.org/10.1080/23273798.2015.1102299

Kuperberg, G. R., 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., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161–163. https://doi.org/10.1038/307161a0

Lau, E. F., Holcomb, P. J., & Kuperberg, G. R. (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

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (De)constructing the N400. *Nature Reviews Neuroscience*, *9*(12), 920–933. https://doi.org/10.1038/nrn2532

Lee, E.-K., & Phillips, C. (2023). Top-down goals modulate the use of argument roles in prediction: Evidence from ERPs. In 36th Annual Conference on Human Sentence Processing. Pittsburgh, PA.

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition*, *14*(1), 41–104. https://doi.org/10.1016/0010-0277(83)90026-4

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

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. Frontiers in Human Neuroscience, *8*, 213. Retrieved from https://www.frontiersin.org/articles/10.3389/fnhum.2014.00213

Luck, S. J. (2021). *Applied event-related potential data analysis*. LibreTexts. https://doi.org/10.18115/D5QG92

Metusalem, R., Kutas, M., Urbach, T. P., Hare, M., McRae, K., & Elman, J. L. (2012). Generalized event knowledge activation during online sentence comprehension. *Journal of Memory and Language*, *66*(4), 545–567. https://doi.org/10.1016/j.jml.2012.01.001

Nakamura, M., & Phillips, C. (2024). *Lateral inhibition accounts for quantitative patterns in speeded cloze data*. Manuscript in preparation.

Ness, T., & Meltzer-Asscher, A. (2018). Predictive pre-updating and working memory capacity: Evidence from event-related potentials. *Journal of Cognitive Neuroscience*, *30*(12), 1916–1938. https://doi.org/10.1162/jocn_a_01322

Ness, T., & Meltzer-Asscher, A. (2021a). From pre-activation to pre-updating: A threshold mechanism for commitment to strong predictions. *Psychophysiology*, *58*(5), e13797. https://doi.org/10.1111/psyp.13797

Ness, T., & Meltzer-Asscher, A. (2021b). Love thy neighbor: Facilitation and inhibition in the competition between parallel predictions. *Cognition*, *207*, 104509. https://doi.org/10.1016/j.cognition.2020.104509

Nieuwland, M. S., Barr, D. J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D. I., Ferguson, H. J., Fu, X., Heyselaar, E., Huettig, F., Matthew Husband, E., Ito, A., Kazanina, N., Kogan, V., Kohút, Z., Kulakova, E., Mézière, D., Politzer-Ahles, S., Rousselet, G., Shirley-Ann, Rueschemeyer, Katrien, Segaert, Jyrki, Tuomainen, & Von Grebmer Zu Wolfsthurn, S. (2020). Dissociable effects of prediction and integration during language comprehension: Evidence from a large-scale study using brain potentials. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1791), 20180522. https://doi.org/10.1098/rstb.2018.0522

Nieuwland, M. S., & Kuperberg, G. R. (2008). When the truth is not too hard to handle: An event-related potential study on the pragmatics of negation. *Psychological Science*, *19*(12), 1213–1218. https://doi.org/10.1111/j.1467-9280.2008.02226.x

Nieuwland, M. S., & Van Berkum, J. J. A. (2005). Testing the limits of the semantic illusion phenomenon: ERPs reveal temporary semantic change deafness in discourse comprehension. *Cognitive Brain Research*, *24*(3), 691–701. https://doi.org/10.1016/j.cogbrainres.2005.04.003

Nieuwland, M. S., & Van Berkum, J. J. A. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience*, *18*(7), 1098–1111. https://doi.org/10.1162/jocn.2006.18.7.1098

Oishi, H., & Sakamoto, T. (2010). *The integration of outputs from syntactic and semantic/thematic processing: "Semantic P600" effects in Japanese sentence processing*. Unpublished manuscript.

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, *2011*, 1:1–1:9. https://doi.org/10.1155/2011/156869

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

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

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

Rabs, E., Delogu, F., Drenhaus, H., & Crocker, M. W. (2022). Situational expectancy or association? The influence of event knowledge on the N400. *Language, Cognition and Neuroscience*, *37*(6), 766–784. https://doi.org/10.1080/23273798.2021.2022171

Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin*, *86*, 446–461. https://doi.org/10.1037/0033-2909.86.3.446

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Staub, A., Grant, M., Astheimer, L., & Cohen, A. (2015). The influence of cloze probability and item constraint on cloze task response time. *Journal of Memory and Language*, *82*, 1–17. https://doi.org/10.1016/j.jml.2015.02.004

Stone, K., & Rabovsky, M. (2024). *The role of syntactic and semantic cues in resolving illusions of plausibility*. Manuscript submitted for publication.

Taylor, W. L. (1953). “Cloze procedure”: A new tool for measuring readability. *Journalism Bulletin*, *30*(4), 415–433. https://doi.org/10.1177/107769905303000401

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

Xiang, M., & Kuperberg, G. (2015). Reversing expectations during discourse comprehension. *Language, Cognition and Neuroscience*, *30*(6), 648–672. https://doi.org/10.1080/23273798.2014.995679

Yano, M. (2018). Predictive processing of syntactic information: Evidence from event-related brain potentials. *Language, Cognition and Neuroscience*, *33*(8), 1017–1031. https://doi.org/10.1080/23273798.2018.1444185

Yano, M., & Sakamoto, T. (2016). The interaction of morphosyntactic and semantic processing in Japanese sentence comprehension: Evidence from event-related brain potentials. *Gengo Kenkyu [Journal of the Linguistic Society of Japan]*, *149*, 43–59.

Zehr, J., & Schwarz, F. (2018). *PennController for Internet Based Experiments (IBEX)*. https://doi.org/10.17605/OSF.IO/MD832

APPENDIX A

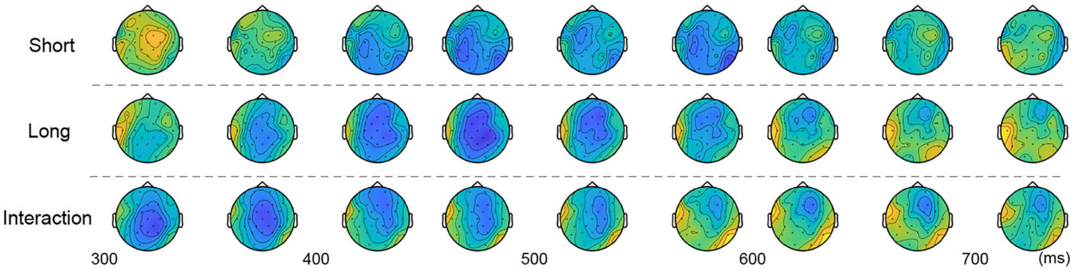


Fig. A1. Topographic maps of ERPs of the accusative-plausible stimuli. There were no significant clusters, but the visual pattern was similar as nominative-plausible stimuli.

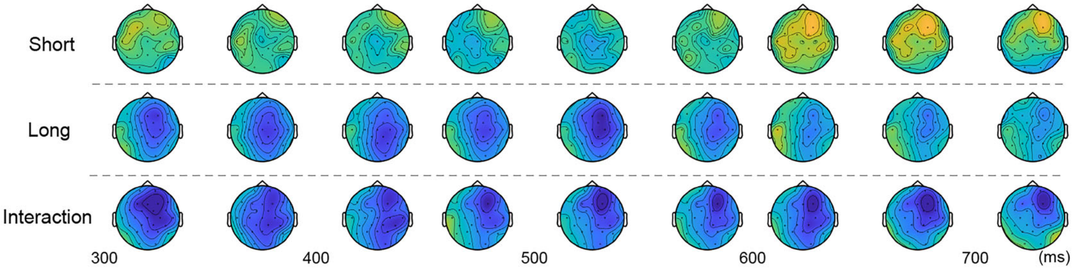


Fig. A2. Topographic maps of ERPs of the nominative-plausible stimuli. There were no significant clusters, but the visual pattern was similar as accusative-plausible stimuli.

APPENDIX B

We also conducted simulations with lateral inhibition assumptions. In this simulation, each candidate does not only accumulate activation but also inhibits the other candidates based on its activation (Nakamura & Phillips, 2024). The amount of inhibition is determined by a sigmoid function, based on Chen and Mirman (2012)'s model that accounted for the neighborhood effects of lexical access in production and comprehension tasks by the sigmoid function.

We used a different set of parameters than the original simulation, which is listed in Table B1. Specifically, greater values for mus were necessary to compensate the activation lost by lateral inhibition. Additionally, the probability of not producing any response when role-appropriate candidates did not reach the threshold was adjusted to 65% so that the proportion of no-response trials (18.1%) matches the proportion in the human data (22.1%).

This simulation with the lateral inhibition assumption showed the same pattern as the original simulation reported above. The simulated N400 amplitudes showed a small contrast in the short condition but a larger contrast in the long condition (Fig. B1).

Table B1
List of parameters in the simulation with lateral inhibition

mu	0.055, 0.0525, 0.05, 0.0475, 0.045
sigma	0.03

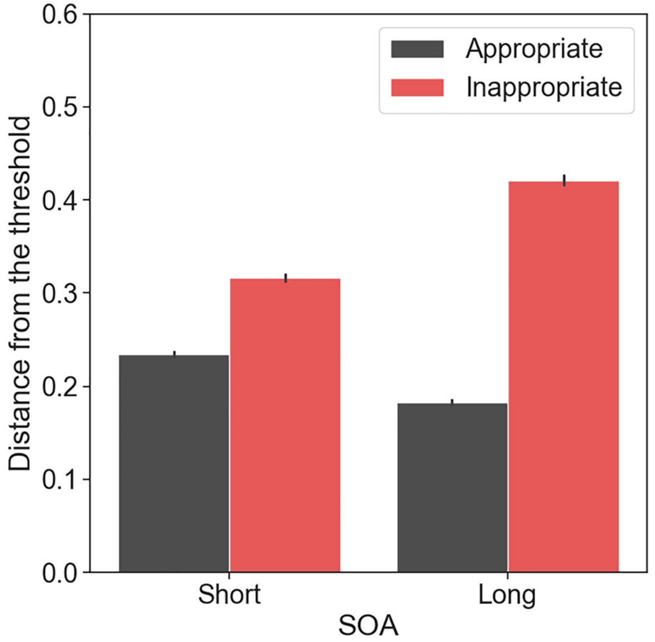


Fig. B1. The simulated N400 effects with the lateral inhibition assumption.

In the production simulation, most responses were role-appropriate (91.5%), while role-inappropriate responses were rare (8.5%). Thus, this simulation showed the same pattern as the simulation without lateral inhibition assumptions.