Cognition 241 (2023) 105628

Contents lists available at ScienceDirect

Cognition

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

Context-specific effects of violated expectations: ERP evidence

Jiaxuan Li [a,1], Jinghua Ou [b], Ming Xiang [b,*]

[a] Department of Language Science, University of California, Irvine, USA
[b] Department of Linguistics, University of Chicago, USA

ARTICLE INFO

Keywords:
Sentence comprehension
Predictive processing
ERPs
N400
Post-N400 positivity
Neural oscillation
Chinese classifiers
Verb-argument structure

ABSTRACT

A complete understanding of the predictive processing effect in sentence comprehension needs to understand both the facilitation effect of successful prediction and the cost associated with disconfirmed predictions. The current study compares the predictive processing effect across two types of contexts in Mandarin Chinese: the classifier-noun vs. verb-noun phrases, when controlling for the degree of contextual constraints and cloze probability of the target nouns across the two contexts. The two contexts showed similar N400 patterns for expected target nouns, indicative of an identical facilitation effect of confirmed contextual expectation. But in the post-N400 time window, the processing cost associated with the unexpected words differed between the two contexts. Additional differences between the two contexts were also revealed by the neural oscillation patterns obtained prior to the target noun. The differences between the classifier vs. verb contexts shed new light on the revision mechanism that deals with disconfirmed expectations.

1. Introduction

Efficient language comprehension has been argued to involve active anticipation of the upcoming linguistic input. As a sentence unfolds in real time, comprehenders not only construct an interpretation of the existing sentence fragment, but also generate expectations about the upcoming inputs. Comprehenders can rely on a wide range of contextual information to generate forward expectations during incremental sentence processing, such as semantic and syntactic information, phonology, event knowledge, and inferences about discourse relations, etc (Berkum et al., 1999; Federmeier & Kutas, 1999; Kim et al., 2016; Kuperberg et al., 2011; Luke & Christianson, 2016; Van Berkum et al., 2005; Van Petten & Luka, 2012; Wicha et al., 2004; Xiang & Kuperberg, 2015). Across different experimental paradigms, the processing of expected words has been found to be facilitated compared to those that are less expected, as demonstrated by shorter reading time (see Staub, 2015 for a review), quicker eye-fixations towards the targets in visual world eye-tracking studies (Altmann & Kamide, 1999, 2007; Staub et al., 2012), and reduced N400 amplitude in ERP studies (Kutas & Federmeier, 2000; Van Berkum et al., 2005).

Although active anticipation could play an important role in language comprehension, it is also recognized that highly predictable words are rare in normal discourse and text. For example, Luke and Christianson (2016) found that highly predictable words are infrequent in texts, with only 5% of content words and 19% of function words being highly predictable (over .67 cloze probability). For words that are not highly predictable, the contexts they appeared in could be misleading, with 70% chances guiding comprehenders to generate a wrong prediction. Observations like these raise the problem that prediction failure could arise with a non-trivial frequency in regular reading and language comprehension. For language comprehension to successfully take place, there needs to be a recovery mechanism that can efficiently monitor and detect prediction errors and also quickly integrate the less expected input. A detailed understanding of this process is still lacking. The goal of the current study is to investigate how the brain responds to upcoming words that confirms or disconfirms prior anticipations, and how the neural responses are modulated by different types of contexts and contextual constraint. To this end, we will compare two types of contexts, classifier-noun v.s. verb-noun phrases in Mandarin Chinese. In both contexts, the target noun will either match or mismatch the contextual expectation, and we will measure the brain responses both on the critical target noun as well as during the time window prior to the target. As will be argued below, the similarities and differences on the brain responses in these two types of contexts provide interesting insights on how comprehenders deal with information that meets or disconfirms contextual expectations.

1.1. N400 and post-N400 positivities (PNPs)

In ERP studies, the facilitation effect of expected words has been associated with the attenuated N400 response, a negative-going waveform that peaks at around 300-500 ms after the onset of the stimuli.

* Correspondence to: 1115 E. 58th Street, Chicago, IL, 60637, USA.
E-mail addresses: jixuan.li@uci.edu (J. Li), mxiang@uchicago.edu (M. Xiang).
[1] SSPB, University of California Irvine, Irvine, CA, 92617, USA.

https://doi.org/10.1016/j.cognition.2023.105628
Received 21 December 2022; Received in revised form 21 September 2023; Accepted 22 September 2023
Available online 4 October 2023

Unexpected words (including those that are incongruent with context) elicit a larger N400 than expected ones (Ito et al., 2016, 2020; Kutas & Federmeier, 2000; Kutas & Hillyard, 1984; Van Berkum et al., 2005; Van Petten et al., 1999; Wicha et al., 2004; Wlotko & Federmeier, 2013). Prediction is generally believed to involve a broad and graded pre-activation of linguistic features instead of a categorical all-or-nothing activation (Federmeier & Kutas, 1999; Kutas & Federmeier, 2000; Luke & Christianson, 2016; Roland et al., 2012). In experimental work, cloze probability, i.e. how likely a comprehender is to continue the sentence fragment with a specific word in a sentence completion task, is often used to operationalize the degree of word predictability in context, and it has been found that the amplitude of N400 on a word in a given context is inversely correlated with the word's cloze probability (Kutas & Hillyard, 1984; Van Petten et al., 1999; Wlotko & Federmeier, 2012).

Whether there is a processing cost for disconfirmed predictions, however, is not entirely clear. Intuitively, if what is encountered turns out to be different from what is expected, comprehenders would need to update their current belief, which may involve inhibiting the incorrectly predicted representations and reviving the previously unexpected alternatives. One way to evaluate whether there is cost associated with this process is to investigate the effect of contextual constraint on unexpected words. A highly constraining context can lead to a stronger prediction than a weakly constraining context, and therefore an unexpected word in a strongly constraining context would violate the prediction to a greater extent than in a weakly constraining context. If there is an extra effort associated with processing disconfirmed predictions, there should be an increased processing cost for an unexpected word in the highly constraining context compared to one in the weakly constraint context. Some behavioral studies have found evidence for such a cost with lexical decision tasks (Fischler & Bloom, 1985), naming tasks (Stanovich & West, 1983) and self-paced reading tasks (Ng et al., 2017; Payne & Federmeier, 2017). For instance, Payne and Federmeier (2017) found that comprehenders read unexpected words in strongly constraining contexts more slowly than in weakly constraining contexts in a self-paced reading paradigm. The behavioral evidence is not totally conclusive though. Eye-tracking reading time appears to be insensitive to failed predictions (Ehrlich & Rayner, 1981; Frisson et al., 2017; Luke & Christianson, 2016; Staub, 2015). Results from ERP studies are a bit mixed, with majority of the studies found that N400 appeared to be insensitive to failed prediction–the N400 amplitude elicited by unexpected words in the strongly constraining context is usually as large as that in the weakly constraining context (Brothers et al., 2015; DeLong & Kutas, 2020; DeLong et al., 2014; Delong et al., 2011; Federmeier et al., 2007; Kuperberg et al., 2020; Kutas & Hillyard, 1984; Lau et al., 2009; Thornhill & Van Petten, 2012; Van Petten & Luka, 2012). But a small number of studies, such as Hoeks et al. (2004), reported a significant difference in N400 amplitude between words that are poor fit in strongly and weakly constraining context respectively. Husband and Bovolenta (2020) also found that a locally predictive context did not necessarily reduce N400 if the upcoming word meets the local expectation but is incongruent with the global context.

In contrast to the N400 response, a series of studies have reported a larger post-N400 frontal positivity response (frontal PNP) on unexpected but still congruent words in strongly constraining context relative to those in weakly constraining context (DeLong & Kutas, 2020; DeLong et al., 2014; Delong et al., 2011; Federmeier et al., 2007; Lai et al., 2021; Lau et al., 2009; Thornhill & Van Petten, 2012; Van Petten & Luka, 2012). A number of studies have also reported that unexpected but plausible words evoked larger frontal PNPs relative to semantically anomalous continuations (DeLong & Kutas, 2020; DeLong et al., 2014; Federmeier et al., 2007; Payne & Federmeier, 2017; Van Petten & Luka, 2012). These late frontal positivities therefore might index the cost of failed predictions. Some studies have even found frontal PNPs prior to the appearance of the target word when the target word stands in a systematic morpho-syntactic relation with a previous word (Hubbard et al., 2019; Van Berkum et al., 2005; Wicha et al., 2004). For example, in Dutch, the gender of adjectives must agree with the nouns they modify. Van Berkum et al. (2005) created two conditions where a contextually most expected noun was paired with an adjective whose gender feature either correctly or incorrectly matches the noun. If comprehenders actively predict the noun given the highly constraining context, they would also expect the matching gender feature on the modifying adjective. The results found a greater frontally distributed PNP on gender-mismatched adjectives than on gender-matched ones. It is important to distinguish the frontal PNPs from the more traditional P600 component, which is also a post-N400 positivity response but with a posterior scalp distribution. In a systematic review on late positivities, Van Petten and Luka (2012) found that whereas the frontal PNPs are elicited by unexpected words that were nonetheless congruent with the context, the majority of late positivities elicited by incongruent/anomalous words has a posterior-parietal scalp distribution.

Particularly relevant for the current purpose is a study by Kuperberg et al. (2020). In this study, comprehenders read three-sentence scenarios, which either provided a rich situation model that can lead to a strong prediction of the upcoming word (e.g. the prediction of the word *swimmers* in the *high constraint* context), or a simple situation model that is only weakly constraining the upcoming word (e.g. the prediction of an entity that is sentient and movable in the *low constraint* context below).

(1a) **Low constraint**: Eric and Grant received the news late in the day. They mulled over the information, and decided it was better to act sooner rather than later. Hence, they cautioned the...(trainees/drawer).

(1b) **High constraint**: The lifeguards received a report of sharks right near the beach. Their immediate concern was to prevent any incidents in the sea. Hence, they cautioned the...(swimmers/trainees/drawer).

Following each context, the critical word in the object position was either expected (*swimmers*), unexpected but still congruent to the context (*trainees*), or it violated the prediction (*drawer*). They found that there was an increased *frontally-distributed* positivity to unexpected continuations in high constraint context vs. low constraint context, whereas anomalous continuations in high constraint contexts elicited a greater *posterior* positivity than in low constraint contexts.

There are several hypotheses regarding what underlying processes the frontal PNP component may reflect. As pointed out by Husband and Bovolenta (2020), these late frontal positivities could simply reflect the detection of the failed prediction based on some sort of error signal (Van Petten & Luka, 2012), or an effort to inhibit the previously predicted words when the actual outcome mismatches the prediction (Kutas, 1993), or they may reflect a discourse update process such that the previously anticipated discourse relations and inferences can be revised when the unexpected linguistic input is ultimately integrated (Brothers et al., 2015; Kuperberg et al., 2020), or it is also possible they index the adaptation process to the overall environmental structure (Kuperberg & Jaeger, 2016). These hypotheses are not mutually exclusive. Brothers et al. (2015) found that unpredicted sentence completions, as defined by a self-report after reading, elicited a larger frontal PNP than predicted continuations. In addition, the frontal PNP to unexpected words is modulated by contextual constraint. These results support the hypothesis that the frontal PNP reflects not only the detection or inhibition of prediction errors, but also the resolution in the discourse update process.

1.2. Context-dependent effect of constraints

Prior work on contextual constraint normally approaches the question by manipulating the degree of semantic fit of the target word in a given context. But an understudied question is whether the contextual effect on predictive processing and any necessary later stage belief revision would vary depending on the nature of the context, even when the semantic fit, often operationalized by cloze probabilities, of the target words are controlled for. There are some preliminary evidence suggesting that the specific ways contextual constraint exerts its influence is indeed context sensitive. The relevant observations have been reported for both N400 (Chow et al., 2018; Liao & Lau, 2020; Momma et al., 2015) and the post-N400 late positivities (Brothers et al., 2020). For example, in a study on Mandarin Chinese, Liao and Lau (2020) compared the N400 responses on an object noun phrase following two different types of verbs: a resultative verb compound and a coordinated verb compound. Even though the object noun phrases following each verb type were both predictable and have very similar cloze probabilities, the resultative verb context did not produce the N400 effect when the target noun phrase was compared to its relevant low-cloze baseline condition, whereas the coordinated verb context did. Liao and Lau (2020) suggested that the resultative verb compound is more complex in its lexical semantic representations than the coordinated verb compound, and therefore requires more time to compute the relevant contextual information. Also highly relevant for the current purpose, Brothers et al. (2020) found that extended linguistic context is more effective at eliciting late positivities than simple lexical context. This study looked at verb-noun phrases of different semantic fit, e.g. *unlock the door/laptop/gardener*, and found that when the phrases were embedded within a simple sentence (e.g. *John unlocked the door*), different levels of semantic fit only modulated the N400 amplitude but did not trigger any post-N400 positivities. But when the phrases were embedded under a 3-sentence discourse context, the semantically anomalous noun continuation (e.g. *unlock the gardener*) elicited a P600; and the unexpected but still plausible continuation (e.g. *unlock the laptop*) triggered a late frontal positivity when the global discourse context was informationally rich and contributed to the construction of a situation model prior to the appearance of the verb-noun pairing.

Existent observations provided preliminary evidence for the possibility that the effect of contextual constraint could be carried out in different ways depending on the nature of the context. Previous studies, however, did not directly compare the effect of strong vs. weak constraints across different contexts. The current study addresses this question by taking advantages of some specific properties of Mandarin Chinese. In particular, we will compare the effect of constraints, by contrasting strongly vs. weakly constraining contexts, across two different constructions: the verb-noun and the classifier-noun constructions.

The semantic fit between a transitive verb and its object noun depends on what type of event is being described and how the relevant event relation is encoded by the verb and its object noun. Upon encountering a transitive verb, a comprehender could rapidly retrieve from their semantic and episodic memory about the event information signaled by the verb meaning, and generate probabilistic expectations about the upcoming noun. Some verbs are more constraining than others. For instance, the verb "hold (举办)" in Mandarin imposes fewer constraints on the type of nouns that can follow compared to a more strongly constraining verb such as "set aside (搁置)". In situations where an expectation of the upcoming noun is disconfirmed by the bottom-up input when the noun actually appears, a comprehender can again recruit information from their semantic and episodic memory to revise the event structure they have committed to earlier and integrate the noun into the updated event structure. It is important to note that event structure information encompassed by a verb is highly nuanced. To understand a verb's event structure entails the understanding of a host of information, including the argument structure preferences of a verb, the temporal properties of the event encoded by the verb, the type of sub-events that could be represented, the plausibility of involving a particular participant in an event, etc. A comprehender draw upon their rich world experience to understand what events are possible in what contexts, what/who are the possible participants in an event, what are the causes/consequences of an event, and many other details . When a participant is cued by a verb to retrieve the relevant event structure information, they are also activating the various relevant "schemas" about the world.

What is unique about Mandarin Chinese is that it allows us to carry out a relatively well-controlled comparison between the verb-noun context and a different type of context: the classifier-noun context. In Mandarin, when a noun is preceded by a numeral (e.g., one, two) or a demonstrative (e.g., this) or a quantifier (e.g., a few), a classifier is required. For example, *one book* in English would be translated into Mandarin as *one-classifier-book* (*yi-ben-shu*,一本书). There are some loose semantic relations governing which classifiers are compatible with which nouns, based on salient perceptual properties of the noun (Tai, 1994). For example, the classifier 'ben' (本) is used to pick out large-volume printed materials, such as books and magazines, whereas the classifier 'zhang' (张) targets objects with a flat appearance, such as papers, tables or thin pancakes. Importantly, however, although there is some semantic motivation for the correspondences between classifiers and nouns, the relationship also oftentimes seem arbitrary and unpredictable. For example, a horse can share the same classifier with a wolf (pi, 匹), but a cow and a tiger would share a different one (tou, 头), both are yet different from the classifiers used for other animals. For language learners therefore, mastering the classifier-noun mapping largely requires memorization of lexicalized collocations.[2]

Classifiers, like verbs, can constrain the space of possible upcoming nouns. Some classifiers are strongly constraining, since they are compatible with only a small set of nouns; whereas some other classifiers are weakly constraining and can be paired with a large set of nouns. Previous studies have also shown that in the classifier-noun context, violations of constraint-based expectations lead to standard N400 effects on nouns following classifiers (Chan, 2019; Chou et al., 2014; Frankowsky et al., 2022; Hsu et al., 2014; Kwon et al., 2017; Li et al., 2021; Qian & Garnsey, 2016; Zhou et al., 2010). What is important for the current study is that the nature of the constraints for the classifier-noun and the verb-noun contexts, however, is crucially different. As discussed above, the constraining effect of a verb arises from people's nuanced knowledge about the event structure information encoded on a verb. The constraining effect of a classifier, on the other hand, is only based on the learned classifier-noun mappings (to some extent idiosyncratic) between a classifier and a noun. Given the differences between the verb-noun and the classifier-noun relationships, it is possible that contextual constraints in these two contexts may exert their impact on the prediction generation and belief revision processes in different ways. The current study examines this question by directly comparing the contextual constraint/expectation effect between the verb-noun and classifier-noun contexts. Empirically, we will probe what happens on the critical noun following a verb or a classifier when the strength of contextual expectation/constraint is being manipulated. In addition, we will also probe what happens prior to the critical noun, asking the question whether we can capture the incremental built-up of the contextual constraint effect.

1.3. Tracking the built-up of contextual constraint prior to the target word

Observations about N400 and frontal PNPs provided some information about the neural basis for the contextually driven anticipatory and the subsequent belief revision processes. But since these observations are made on the target word instead of on the pre-target context directly, it is an open question how contextual information is accumulated to guide the subsequent processes. A small number of previous ERP studies have examined the constraint effect prior to the critical

[2] There is some debate as to whether the classifier-noun relationship should be analyzed as syntactic agreement (Chan, 2019). Since there is no clear consensus, we maintain a neutral position on this matter.

target, but the findings are mixed. Ness and Meltzer-Asscher (2018, 2021) observed an increased P600 amplitude in strongly constraining verb context prior to the appearance of the expected noun, but in a different study, strongly constraining verb context elicited a late left-anteriorly distributed negative deflection prior to the onset of the target (Li et al., 2017).

Some recent studies have also suggested that neural oscillatory activities might be informative in this regard (Lewis et al., 2015). Particularly, alpha (8–12 Hz) and beta (13–30 Hz) oscillatory activity have been shown to be involved in linguistic prediction prior to the appearance of a target word (Piai et al., 2018; Rommers et al., 2017; Terporten et al., 2019; Wang et al., 2018a). Alpha oscillations have been argued to reflect cortical idling and the engagement of task relevant neural network. Beta oscillations, on the other hand, have been proposed to be relevant for the top-down propagation of predictions to lower processing levels (Friston et al., 2015; Lewis & Bastiaansen, 2015). For instance, when participants were asked to name pictures in strongly or weakly constraining sentence contexts, they showed reduction in alpha and beta power in strongly constraining contexts before the naming, which might reflect working memory demands related to pre-selection and maintenance of lexical candidates (Piai et al., 2014, 2015). In addition to production, the alpha/beta suppression was also observed prior to target words for highly constraining relative to weakly constraining contexts during sentence comprehension (Rommers et al., 2017; Wang et al., 2018b). The power suppression has been suggested to indicate stronger engagement of the task-related language network when predictions can be made based on context. Results were however mixed when more than two levels of sentence context constraint were employed, with pre-stimulus alpha/beta power decreasing most strongly for intermediate constraints, followed by high and low constraints, raising some questions about whether the alpha/beta modulation actually indexes the degree of word predictability based on sentence context (Terporten et al., 2019).

In current study, we will measure ERPs on the target noun and also on the verb/classifier before the target noun. We will also examine oscillatory activities prior to the target words. Obtaining information on both the target word and the pre-target context can help us better understand how context facilitates linguistic prediction and belief revision.

1.4. The current study: comparing verb and classifier contexts

The present study investigates the neural processes in response to confirmed and violated expectations in both the classifier-noun and the verb-noun contexts, combining measures about the oscillatory activities prior to the target words and the ERP responses obtained on the target words. Under each type of context, we adopted an experimental design from Kuperberg et al. (2020), controlling for the degree of constraints based on the classifier or the verb information (high constraint vs. low constraint) and the expectancy of target nouns in their respective context (expected, unexpected but congruent, and anomalous). For data analyses, prior to the target noun, we will compare the ERP and the oscillatory activities elicited by the strongly vs. weakly constraining contexts, for verb-noun and classifier-noun structures separately. On the target noun, we will focus on the N400 and the post-N400 frontal positivities (PNPs) to evaluate the effects of contextual constraints across two different contexts.

2. Method

2.1. Participants

The experiment was conducted in Guangzhou, China. We recruited 24 (11 males) native Mandarin speakers, aging between 18 and 30 years old. All of them were university students, right-handed and with normal vision. The experiment was approved by the research ethics committee of the local institution, and consent was obtained from all participants before their participation.

2.2. Design and materials

Two sets of phrases were created, one set with the numeral -classifier-noun structure and the other set with the verb-noun structure. Within each structure type, following the experimental design from Kuperberg et al. (2020), there were five conditions (see an example in Table 1): high constraint expected (HC.Exp), high constraint unexpected (HC.Unexp), high constraint anomalous(HC.AN), low constraint unexpected (LC.Unexp), low constraint anomalous (LC.AN). We note that, different from Kuperberg et al. (2020), high and low constraints conditions have different target nouns. To control for lexical frequency and visual complexity effects, we included frequency and number of strokes in our regression analyses as predictors, as detailed below.

The five conditions were constructed as follows. First, we parsed the Chinese Treebank 9.0 (Xue et al., 2005), a Chinese language corpus consisting of two million words from a wide variety of texts including news, weblogs, transcribed phone conversations. We collected all possible quantifier-classifier-noun and verb-noun patterns within the corpus. We then calculated the entropy for each classifier and verb based on the probability distribution of the following nouns. The entropy was used as an approximation of the constraining effect of each classifier and verb on the following nouns. A high entropy value on a classifier or verb indicates high degree of uncertainty of the upcoming noun, and therefore the context is less constraining; whereas a low entropy value indicates the context is more constrained. An initial set of high and low-constraint verbs and classifiers were then selected based on their entropy values. The final set of stimuli were chosen after the cloze norming study detailed below. For the final set of stimuli, the average entropy in high constraint conditions for classifiers and verbs were 2.56 and 2.63, respectively, while the values in low constraint conditions for classifiers and verbs were 3.90 and 3.91.

Next, we conducted a cloze norming study. Seventy-two native Mandarin Chinese speakers recruited from Qualtrics were presented with the initial set of classifiers and verbs. They were asked to write down a single noun that was the most likely continuation after each classifier or verb. The responses from the subjects were used to construct the final set of stimuli that were controlled for two metrics — contextual constraint and cloze probability of the noun. The cloze probability of nouns was operationalized as the proportion of the target nouns being used to complete a given phrase, and the degree of constraint was defined as the cloze probability of the most frequent completion for a given classifier/verb. As shown in Table 1, cloze probabilities of the target nouns and contextual constraints were both matched between the verb and classifier groups. The average constraint values of the high constraint (HC) conditions for classifiers and verbs were 0.52 (Min: 0.32; Max: 0.86) and 0.51 (Min: 0.31; Max: 1) respectively, with no reliable statistical difference between the two ($p > 0.7$), while those of the low constraint (LC) conditions for classifiers and verbs were 0.18 (Min: 0.06; Max: 0.29) and 0.2 (Min: 0.05; Max: 0.31) ($p > 0.1$). The average cloze probabilities of the high constraint expected (HC.Exp) verb and classifier structures were 0.52 and 0.51, with all cloze values above 0.3. The cloze probabilities of the low cloze and anomalous conditions were close or equal to zero. Although matched in cloze probability, the unexpected conditions were different from the anomalous conditions in terms of plausibility — the target nouns in the unexpected conditions were unpredictable but still plausible, whereas the anomalous conditions are semantically implausible. The plausibility ratings presented in Table 1 came from the behavioral responses during the ERP recording session.

We also matched the noun frequencies between the classifier and verb contexts for all the low cloze and anomalous conditions (i.e. four conditions total). But it was impossible to match the noun frequencies between high constraint high cloze condition with the other conditions. Table 2 shows the frequency counts per million words for the target nouns. The frequency counts are extracted from Chinese Internet

Table 1
Examples of experimental conditions. The ratings of constraint, cloze and plausibity represent mean ratings for each condition.

Condition	Classifier (CL)				Verb (VB)			
	Example	Constraint	Cloze	Plausibility	Example	Constraint	Cloze	Plausibility
High Constraint Expected (HC.Exp)	一台电脑 one-CL computer	.52	.52	95.1%	搁置争议 set aside arguments	.51	.51	93.5%
High Constraint Unexpected (HC.Unexp)	一台手术 one-CL surgery	.52	.02	89.2%	搁置提案 set aside proposals	.51	.02	91.3%
High Constraint Anomalous (HC.AN)	一台学问 one-CL knowledge	.52	.0	8.1%	搁置云朵 set aside clouds	.51	.0	8.1%
Low Constraint Unexpected (LC.Unexp)	一份感情 one-CL relationship	.18	.02	89.1%	举办婚礼 celebrate wedding	.20	.01	90.8%
Low Constraint Anomalous (LC.AN)	一份墙 one-CL wall	.18	.0	7.4%	举办鞋子 celebrate shoes	.20	.0	9.5%

Table 2
Mean word frequency (counts per million words) and number of strokes for pre-noun contexts and target nouns, across classifier and verb conditions.

	Count	Position	High constraint			Low constraint	
			Exp	Unexp	AN	Unexp	AN
CL	Frequency	classifier		148		329	
		target noun	521	98	98	98	98
	Stroke	classifier		9		9	
		target noun	10	14	13	14	12
VB	Frequency	verb		77		71	
		target noun	239	98	98	98	98
	Stroke	verb		16		16	
		target noun	15	15	13	16	12

Corpus with 90 million words (Sharoff, 2006). We also computed the number of strokes for the target nouns, as presented in Table 2 as well. In addition, since the high vs. low constraint contexts involve different verbs/classifiers, we computed the frequency counts and the number of strokes for the pre-noun verbs/classifiers too, and presented them in Table 2. The data analyses below will take into account the word frequency and visual complexity (as indexed by the number of strokes).

In the final set of experimental material there are 90 high-constraint classifiers (90 unique classifiers with no repetition) and 60 low-constraint classifiers (51 unique ones and 9 repeated). There are also 90 high-constraint verbs (no repetition) and 60 low-constraint verbs (no repetition). Among the classifiers in the final set of stimuli, about 7% are so-called measure words, since they can be used to measure quantities, such as "mile" (里) in "a mile of road" (一里路). Each high constraint classifier was paired with three types of nouns (expected, unexpected, or anomalous target noun), yielding 3 classifier-noun sublists. Similarly, there were also 3 sublists for the high constraint verb-noun phrases. For the low constraint classifiers/verbs, each classifier/verb was paired with an unexpected or an anomalous noun, yielding 2 low-constraints classifier-noun lists and also 2 low-constraints verb-noun sublists. The high constraint and low constraint sub-lists were then combined to create 6 (3 × 2) experimental lists. Each experimental list consisted of 300 trials (150 for the classifier-noun phrases and 150 for the verb-noun phrases), with 30 trials in each condition (see an example in Table 1 for all the conditions).

2.3. Procedure

During the experimental session, participants were randomly assigned to one of the 6 stimuli lists. In each trial they read either a numeral-classifier-noun phrase or a verb-noun phrase, and judged the plausibility of the phrase. Each phrase was divided into two segments and presented sequentially on the screen. The first segment contained a numeral-classifier or a verb, and the second segment presented the critical target noun. At the beginning of each trial, participants saw a fixation at the center of the screen for 400 ms. Then each segment of a phrase appeared for 850 ms, followed by a 150 ms blank screen. After the offset of the second segment, which is also the critical word, participants were prompted to decide the plausibility of the phrase by pressing 'f' (for the plausible phrases) and 'j' (for the implausible phrases) on the keyboard. In each experimental session there were five practice trials and 300 experimental trials, divided into 4 blocks. The sequence of the trials was randomized within each block. Participants were seated in a comfortable chair in front of a computer, instructed to avoid any excessive eye or body movement.

2.4. EEG recording

The EEGs were recorded with a Brain Vision actiCHamp Plus System (Brain Products GmbH) from 32 active electrodes. Thirty electrodes were attached on an elastic cap (Fp1, Fp2, F7, F3, Fz, F4, F8, FT5, FC5, FC1, FC2, FC6, FT10, T7, C3, Cz, C4, T8, CP5, CP1, CP2, CP6, P7, P3, Pz, P4, P8, O1, Oz, O2). A separate vertical EOG (VEOG) electrode was attached below the left eye. A horizontal EOG (HEOG) electrode was placed at the outer cantus of left eye. The left (TP9) or the right mastoid (TP10) was used as online reference electrode (counterbalanced between participants). The impedance of all electrodes was kept below 10kω throughout the experiment. Continuous data were digitized using the BrainVision Recorder with a sampling rate of 1000 Hz without any online filter. The entire recording lasted for about 30 min.

2.5. Data analyses pre-processing

We performed two sets of analyses, one focused on the critical noun, and the other focused on the pre-noun contexts (i.e. the verbs and

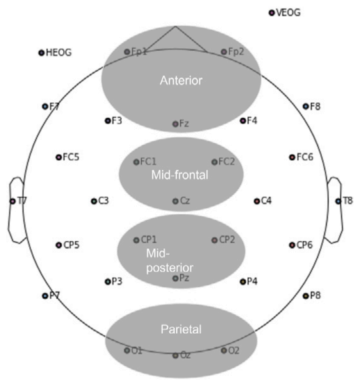

Fig. 1. Electrodes layout and the midline area channels included in the data analysis. The midline regions, shown in shaded circles, from front to the back: Anterior, mid-frontal, mid-posterior, parietal.

classifiers prior to the nouns). The analyses on the target nouns are traditional ERP analyses time locked to the onset of the noun. For the analysis on the pre-noun contexts, we not only analyzed the ERP data time locked to the onset of the classifier/verb, we also carried out a time-frequency analysis on the oscillatory activities prior to the noun. More details about the analyses on the pre-noun contexts will be reported in Section 4. In this section we focus on data/results from the target noun.

The raw EEG data was preprocessed with EEGLab (Delorme & Makeig, 2004) and ERPLab (Lopez-Calderon & Luck, 2014). The continuous EEG data were bandpass filtered from 0.01 to 30 Hz. The continuous data around the head noun were epoched with a 1200 time window — 200 ms before the onset of the target noun and another 1000 ms after the noun onset. Baseline correction was performed by subtracting the mean pre-stimulus voltage (−200 – 0 ms) from the epoched data. The data were then re-referenced to the average of left and right mastoid electrodes. We removed trials with peak-to-peak amplitude on eye electrodes above 100 μV using a moving-window peak-to-peak threshold method. Trials with vertical and horizontal eye movements were removed with a step function which detected sudden changes above a threshold of 75 μV within a step window of 50 ms. We also excluded trials with extreme voltages on any other electrodes greater than the ±75 μV threshold. The artifact rejection removed 13.4% of the total number of trials. We further excluded trials with incorrect behavioral responses, which removed another 7.3% of trials. Among the remaining trials, there are on average 22–26 trials per condition.

The data analyses focused on the midline area (see Fig. 1). The midline area was further divided into four regions of interests (ROIs): anterior, frontal, posterior and parietal regions (marked with shaded circles in Fig. 1). Following the previous study (Kuperberg et al., 2020), we extracted ERPs from an interval of 300-500 ms post critical word onset to examine the N400, and ERPs from 600-1000 ms post critical word onset to analyze the post-N400 positivities (PNP). For each time window, a series of linear mixed-effect models were fit, implemented with the lmer4 package in R using the Satterthwaite method.

For the statistical analysis, we set up two user-defined contrasts. In the first contrast, using a backward difference coding, we set target nouns of expected (high cloze), unexpected (low cloze) and anomalous (zero cloze) nouns as a 3-level contrast *contrast1*. The backward difference coding compares the mean of one level of a dependent variable to the mean of the prior level. Contrast1 therefore helps us examine two effects: the comparison between expected and unexpected (but plausible) nouns represents the *effect of Cloze probability*; and the comparison between unexpected (but plausible) and anomalous nouns represents the *effect of Plausibility*. The second contrast *contrast2* examines the *effect of Constraint* by comparing the low cloze nouns in the high constraint context with those in the low constraint context. Again with a user-defined coding, we set high constraint unexpected condition (HC.Unexp) as 1, low constraint unexpected condition (LC.Unexp) as −1, and the rest of conditions as 0. This second contrast *contrast2* only compared the unexpected target nouns under high vs. low constraining context, because this is the critical comparison in previous studies that addresses questions about the cost of violated expectations (see Section 1.1).

Since our primary interest is whether the abovementioned effects interact with verb vs. classifier context in different ways, as our first step of analysis, we specifically evaluated whether there are interactions between Context (verb vs. classifier) and the two user-defined contrasts above. When there are significant interactions, we then carried out further analyses for classifier-noun and verb-noun contexts separately. More details of the analyses will be reported in the next section.

3. Results

3.1. Behavioral results

The plausibility judgments during the ERP recording session are reported in Table 1. For the Classifier-Noun conditions, participants judged most trials to be plausible for the *high constraint expected* (98.0%), *high constraint unexpected* (84.9%) and *low constraint unexpected* (87.1%) conditions. Only a small number of trials in *high constraint anomalous* (6.3%) and *low constraint anomalous* (6.5%) conditions were judged as plausible. Similarly, for the Verb-Noun conditions, participants considered most trials plausible for the *high constraint expected* (97.1%), *high constraint unexpected* (84.9%) and *low constraint unexpected* (90.3%) conditions, and only a small number of trials in the *high constraint anomalous* (10.3%) and *low constraint anomalous* (9.5%) conditions were judged as plausible.

3.2. ERP results on the target noun

To examine whether there is an interaction between Contexts (verb vs. classifier) and any of our user-defined contrasts, we first constructed a full model m_0 in (1) below. This model included the following predictors: *contrast1* (high cloze vs. low cloze vs. anomalous), *contrast2* (HC.Unexp vs. LC.Unexp), *context* (classifier vs. verb), and *ROI* (anterior vs. mid-frontal vs. mid-posterior vs. parietal), frequency (log-transformed) and number of strokes of the noun. We apply this model separately to analyze data from the 300-500 ms and 600-1000 ms time windows.[3] We then constructed a series of models by removing three-way or two-way interactions between predictors from the full model m_0, and used nested model comparisons to estimate the effect of the respective interactions. For example, to estimate the effect of the 3-way interaction $ROI \times Context \times Contrast1$, we constructed a new model m_{01} that does not contain this interaction term but is otherwise identical to m_0. We then did a model comparison between m_0 and m_{01}. Similarly, to estimate the effect for the 3-way interaction $ROI \times Context \times Contrast2$, we constructed a model m_{02} that is again almost identical to m_0 but does not contain the relevant interaction, and we then did a model comparison between m_0 and m_{02}. The main effects for our key predictors are presented in Table 3. We present all the 2-way and 3-way interactions in Table 4. The results showed a significant $ROI \times contrast1$ interaction

[3] Since the nouns in the highly constraining conditions also tend to be more frequent than the nouns in other conditions (see Table 2), to ensure frequency is not correlated with other predictors we calculated the variance inflation factor (VIF) measure to detect multicollinearity between predictors using the *vif* function from the *car* package (Fox & Weisberg, 2019). VIFs between 5–10 indicate a moderate correlation between predictors. In our results, for both the 300-500 ms and 600-1000 ms time windows, the VIFs for frequency are very close to 1, suggesting that frequency is unlikely to be correlated with other predictors in the model.

Table 3
The effects of context, cloze probability, plausibility and constraint in the 300–500 ms and 600–1000 ms time windows.

Time window	Context			Cloze probability			Plausibility			Constraint		
	coef	*se*	*t*	*coef*	*se*	*t*	*coef*	*se*	*t*	*coef*	*se*	*t*
300–500 ms	−0.75	0.30	−2.50*	0.69	0.36	1.86^	1.56	0.30	5.18***	0.38	0.21	1.80
600–100 0ms	−0.09	0.31	−0.29	−1.12	0.41	−2.71**	1.90	0.35	5.50***	0.58	0.24	2.38*

Table 4
Interaction between regions of interests (ROIs), two user-defined contrasts and contexts.

Contrast	Interaction	300–500 ms window		600–1000 ms window	
		chi-square	*p*	*chi-square*	*p*
Contrast1	ROI:context:contrast1	1.13	0.98	2.94	0.82
	ROI:context	16.36	<0.001***	2.23	0.52
	context:contrast1	2.50	0.28	7.40	0.02*
	ROI:contrast1	63.15	<0.001***	31.20	<0.001***
Contrast2	ROI:context:contrast2	0.11	0.99	1.62	0.65
	ROI:context	0	0.99	0	0.99
	context:contrast2	0.72	0.39	4.92	0.03*
	ROI:contrast2	3.46	0.33	2.48	0.48

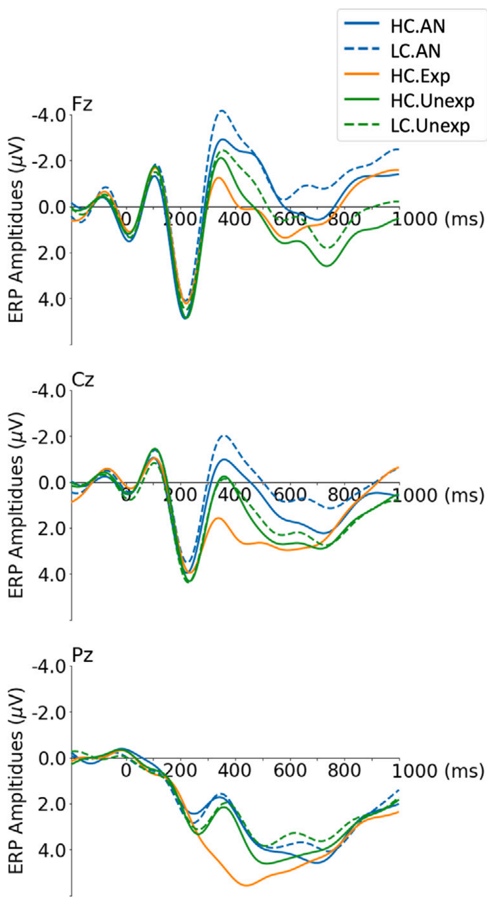

Fig. 2. The Classifier-Noun (CL-N) construction: Grand-averaged ERP waveforms to the critical nouns at electrode Fz (Top), Cz (Middle) and Pz (Bottom).

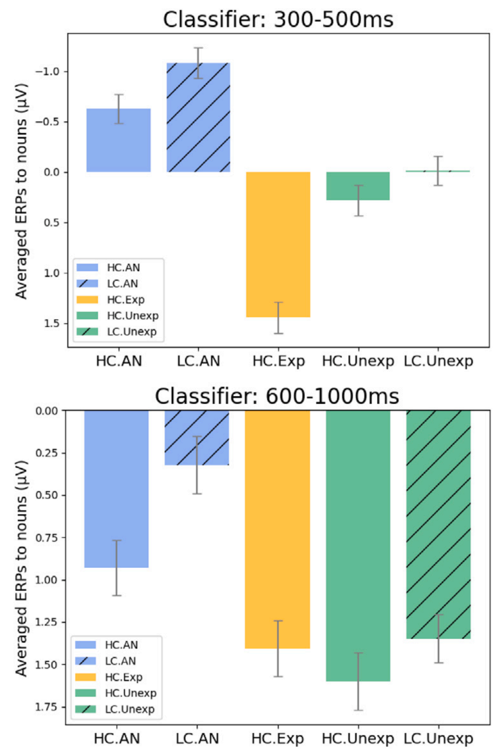

Fig. 3. The Classifier-Noun (CL-N) construction. Top: Averaged ERPs to the critical nouns across 300-500 ms time window from all midline region ROIs. Bottom: Averaged ERPs to the critical nouns across 600-1000 ms time window from all midline region ROIs.

in the 300-500 ms time window (chi-square = 63.15, p <.001) and in the 600-1000 ms time window (chi-square = 31.20, p <.001). There is also a $context \times contrast1$ interaction (chi-square = 7.40, p = 0.02) effect in the 600-1000 ms window. There are no 3-way interactions in either time windows. Since $contrast1$ encompasses two effects, cloze effect (high vs. low cloze) and plausibility effect (low cloze vs. anomalous), these findings suggest that cloze and plausibility had similar effects for both verbs and classifiers contexts in the 300-500 ms window, but contexts modulated these effects in different ways in the 600-1000 ms window. Regarding $contrast2$, which represents the effect of constraint (HC.unexpcted vs. LC unexpected), the only significant interaction is $context \times contrast2$ (chi-square = 4.92, p = 0.03). This indicates that the effect of constraint only emerged in the 600-1000 ms window, and was modulated by contexts. Additionally, in the 300-500 ms time window, there was a significant $ROI \times context$ interaction (chi-square = 16.36, p <.001). Since this is not an effect of our primary interest, we will not pursue it further.

$$\begin{aligned} m_0 : Amplitude \sim & ROI * context * contrast1 + ROI * context * contrast2 \\ & + frequency + stroke + (1 \mid item) \\ & + (1 + ROI * context * contrast1 \\ & + ROI * context * contrast2 \mid subject) \end{aligned} \tag{1}$$

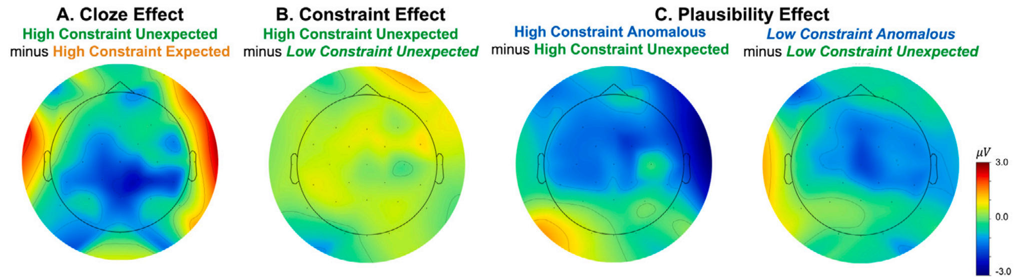

Fig. 4. The Classifier-Noun (CL-N) construction (300-500 ms): Topographic maps of ERP amplitude differences for the cloze effect (A), plausibility effect (B) and constraint effect (C). ERP amplitudes averaged across the 300-500 ms time window.

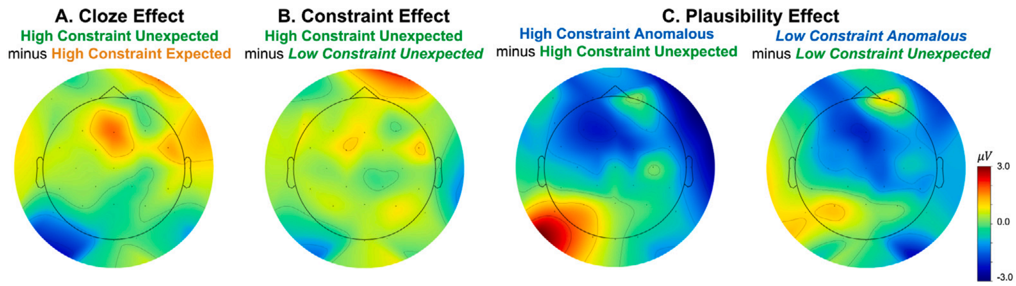

Fig. 5. The Classifier-Noun (CL-N) construction (600-1000 ms): Topographic maps of ERP amplitude differences for the cloze effect (A), plausibility effect (B) and constraint effect (C). ERP amplitudes averaged across the 600-1000 ms time window.

Since there are interactions between contexts and our effects of interests, next we analyzed the effects of cloze probability, plausibility and constraint for verbs and classifiers separately.

3.2.1. ERP results: the classifier-noun context

The ERP results for the Classifier-Noun trials were presented in the following figures. Fig. 2 presents the ERP waveforms elicited on the target nouns following a classifier at Fz, Cz and Pz. Fig. 3 presents the mean ERP amplitudes on the critical nouns during the N400 time window (300-500 ms post-noun-onset) and Post-N400 time window (600-1000 ms post-noun-onset), averaged over all the midline region ROIs. Figs. 4 and 5 present the topographic distribution of the relevant effects in the N400 (300-500 ms) and Post-N400 time window (600-1000 ms).

For the 300-500 ms and 600-1000 ms time window, we constructed a mixed effect model using the two user-defined contrasts, frequency and number of strokes as the fixed predictors and the maximal by-participant, by-item and by-ROI random effects, shown in (2).

$$\begin{aligned} m_1 : \ & Amplitude \sim contrast1 + contrast2 + frequency + stroke \\ & + (1 \mid item) + (1 + contrast1 + contrast2 \mid subject) \\ & + (1 + contrast1 + contrast2 \mid ROI) \end{aligned} \quad (2)$$

Since the three levels in $contrast1$ were backward difference coded, in the model outcome, the coefficients associated with $contrast1$ represent the two effects *cloze* and *plausibility*. The coefficient for $contrast2$ represents the effect of *constraint*. These effects were presented in Table 5. In the 300-500 ms time window, there was a significant effect of cloze probability (t = 3.19, p < 0.05), with the amplitudes of N400 evoked by the unexpected target nouns larger than the expected target nouns. There was also a significant effect of plausibility (t = 2.45, p < 0.05), due to the fact that the N400 elicited by anomalous nouns was larger than unexpected nouns that were congruent to contexts. The effect of contextual constraint, however, was not significant during the N400 window. In particular, there was no amplitude difference on the low cloze target nouns in a high constraint vs. a low constraint context. In the post-N400 time window (600-1000 ms), there was no significant effect (all ts < 1.5, all ps > 0.1).

As we showed in Table 4, there was not any 3-way interaction involving contexts, ROI and our effects of interests. But we still carried out a post-hoc analysis for each ROI separately. This was done to simply provide more details for readers that are interested, but we are not drawing statistical conclusions about any particular ROI based on this post-hoc analysis. The procedure and results of these additional analyses were presented in Appendix A.

3.2.2. ERP results: the verb-noun context

Fig. 6 displays the averaged ERP responses to the target nouns in verb-noun phrases across five conditions at the three midline electrodes. Fig. 7 displays mean ERP amplitudes on target nouns in the N400 and Post-N400 time windows, averaged over all the midline region ROIs.

Similar to the classifier context, we also conducted a mixed effect model specified in (2), and the results were presented in Table 5. In the N400 window (300-500 ms), there was a significant main effect of cloze probability on N400 amplitudes (t = 3.08, p < 0.01), with a larger N400 on unexpected but plausible nouns than expected nouns. There was also an effect of plausibility (t = 2.63, p < 0.05), with a larger N400 to anomalous nouns than unexpected but plausible nouns. Similar to the classifier-noun phrases, in the N400 window, there was no significant effect of constraint on unexpected nouns in high constraint vs. low constraint conditions. We present the topographic distributions for the N400 time window in Fig. 8. Also shown in Table 5 were the results from the post-N400 time window (600-1000 ms). There was a post-N400 positivity effect (PNP) that was sensitive to plausibility (t = 2.53, p < 0.05), with a larger PNP to unexpected but plausible vs. anomalous words. Most importantly, in contrast to the classifier-noun context, in the verb-noun context, the amplitudes of PNPs elicited by unexpected nouns in high constraint sentences were significantly greater than in the low constraint sentences (t = 2.34, p < 0.05). We present the topographic distributions in the post-N400 time window in Fig. 9. We also carried out the post-hoc by-ROI analyses for the 300-500 ms and 600-1000 ms windows, and the results were presented in Appendix A as well.

Table 5
The effects of cloze probability, plausibility and constraint in the N400 (300–500 ms) and post-N400 (600–1000 ms) time windows for Classifier-Noun (CL-N) and Verb-Noun (VB-N) constructions.

Construction	Time window	Cloze probability			Plausibility			Constraint		
		coef	*se*	*t*	*coef*	*se*	*t*	*coef*	*se*	*t*
CL-N	N400	1.37	0.43	3.19*	0.93	0.38	2.45*	0.26	0.21	1.22
	Post-N400	0.21	0.44	0.47	0.63	0.48	1.30	0.22	0.28	0.79
Verb-Noun	N400	1.61	0.52	3.08**	1.22	0.46	2.63*	0.26	0.17	1.48
	Post-N400	−0.27	0.61	−0.60	1.31	0.52	2.53*	0.55	0.24	2.34*

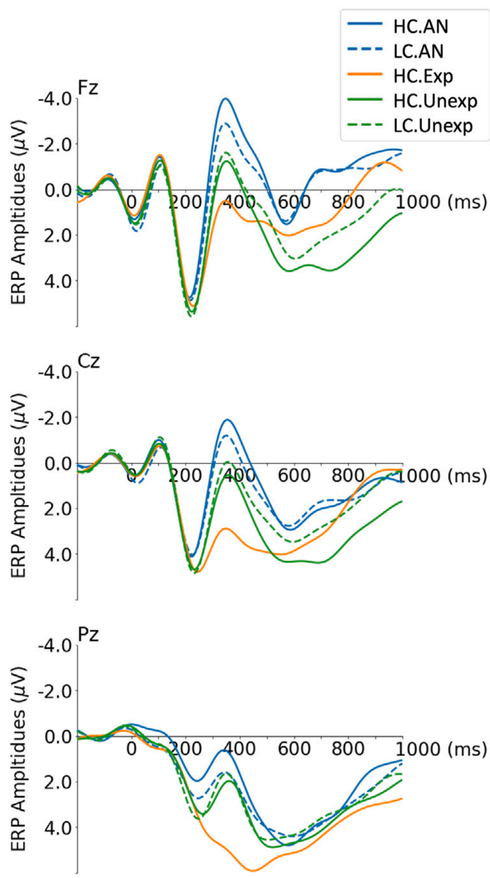

Fig. 6. The Verb-Noun (VB-N) construction: Grand-averaged ERP waveforms to the critical nouns at electrode Fz (Top), Cz (Middle) and Pz (Bottom).

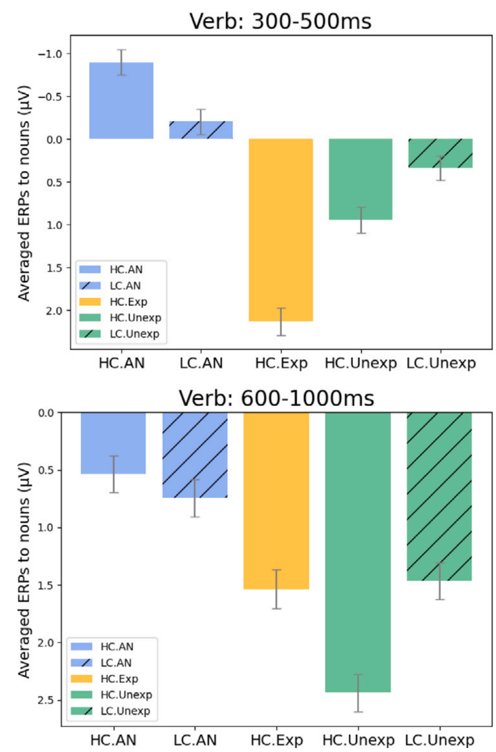

Fig. 7. The Verb-Noun (VB-N) construction: Top: Averaged ERPs to the critical nouns across 300-500 ms time window from all the midline region ROIs. Bottom: Averaged ERPs to the critical nouns across 600-1000 ms time window from all the midline region ROIs.

3.3. Summary about the comparisons between the classifier-noun and verb-noun contexts

To summarize, the classifier-noun and the verb-noun contexts showed very similar ERP effects in the N400 time window: both showed a gradient N400 effect modulated by cloze probability and plausibility (anomalous > unexpected > expected). But for the unexpected yet plausible target nouns, there is no difference between the high constraint vs. the low constraint sentences in the N400 window. During the post-N400 window, the verb-noun contexts showed the plausibility effect and the cloze probability effect. The classifier-noun context did not show reliable effects in the post-N400 window (Table 5)[4]. The most critical difference between the two contexts is that, only in the verb-noun context, we observed a larger PNP for the unexpected words in the high vs. low constraint sentences. This latter finding is consistent with previous findings that found larger PNP for words that violated contextual expectation to a greater degree (i.e. unexpected words in high constraint context) (DeLong & Kutas, 2020; DeLong et al., 2014; Delong et al., 2011; Federmeier et al., 2007; Kuperberg et al., 2020; Lai et al., 2021; Lau et al., 2009; Payne & Federmeier, 2017; Thornhill & Van Petten, 2012; Van Petten & Luka, 2012). It is also worth noting that previously discussed PNP effect often has a frontal distribution. We are not able to make strong conclusions about the topographic distribution of the PNP effect in the current study since we did not find a reliable 3-way interaction between contexts, ROIs and the constraint effect (see Table 4, $ROI \times context \times contrast2$, 600-1000 ms window). But the post-hoc tests presented in Appendix A (see Table A.2) are consistent with a central-frontal distribution of the PNP constraint effect. It is possible that the current study is not sufficiently powered to detect a 3-way interaction. This would be a question for future work.

4. Analyses on the pre-noun context

The previous section looked at the ERP results time locked to the onset of the target noun. In this section we report the results from two analyses on the time window prior to the target noun.

[4] Although we note that the post-hoc analyses in each ROI revealed signs of plausibility effects in anterior and mid-frontal regions (see Table A.2) in Appendix A.

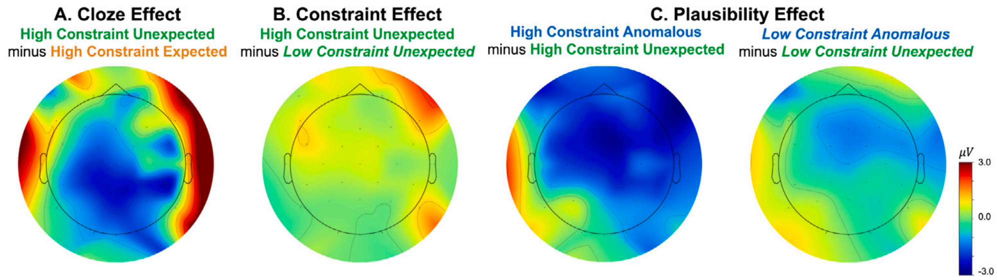

Fig. 8. The Verb-Noun (VB-N) construction (300-500 ms): Topographic maps of ERP amplitude differences for the cloze effect (A), plausibility effect (B) and constraint effect (C). ERP amplitudes averaged across the 300-500 ms time window.

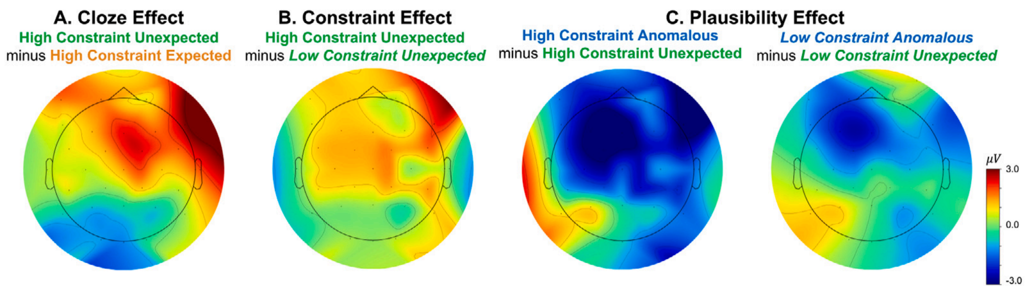

Fig. 9. The Verb-Noun (VB-N) construction (600-1000 ms): Topographic maps of ERP amplitude differences for the cloze effect (A), plausibility effect (B) and constraint effect (C). ERP amplitudes averaged across 600-1000 ms time window.

4.1. ERP analyses and results

For the ERP analysis, the pre-noun context data was segmented into epochs with a 1200 time window — 200 ms before the onset of the classifier/verb and 1000 ms after. This time window corresponds to the −1200 ms to 0 ms window relative to the onset of the critical noun. We collapsed all the conditions into *High constraint* and *Low constraint* conditions. But to balance the number of trials in the high vs. low constraint condition, we excluded data from the *expected* condition in Table 2, leaving 60 trials in each of the high and low constraining condition, for verbs and classifiers contexts respectively. The data preprocessing procedure was identical to the procedure in Section 2.5. The artifact rejection removed 21.4% of the total number of trials. After artifact rejection, the average remaining number of trials per participant per condition was: verb high constraint 47; verb low constraint 48; classifier high constraint 47; classifier low constraint 48.

Figs. 10 and 11 show the topographic maps for the classifier and verb contexts separately, with the difference between high vs. low constraining conditions. Since there was no strong consensus from previous studies regarding the expected ERP patterns during the pre-target noun window, our initial objective was to identify significant temporal-spatial clusters that would reveal the differences between high and low constraining contexts. To accomplish this, we conducted a cluster-mass based permutation test (Bullmore et al., 1999) for the classifier and verb constructions separately, using the Mass Univariate ERP Toolbox (Groppe et al., 2011) and Factorial Mass Univariate ERP Toolbox (Fields, 2017). But the permutation tests did not find any significant constraint effect for either classifier or verb constructions.

For the permutation test, the data was first downsampled from 1000 Hz to 250 Hz and then re-baselined using the −200 ms baseline. The permutation test was performed for the 0-1000 ms time window post classifier/verb onset over all electrodes, excluding reference channel (TP9 and TP10) and eye electrodes (VEOG and HEOG). ANOVA tests were performed for the original data, as well as 1000 random within-participant permutations of the original data. For each permutation, a F-value was computed to quantify the between-condition difference at each time point, ERP amplitude and channel combination. F-values that were above the significance threshold (a significance level of 0.05) and were adjacent spatially and temporally were grouped together to form clusters. A cluster mass statistic was then calculated by summing all the F-values in a cluster. The largest cluster mass value for each of the 1000 permutations was recorded, which together formed the null distribution. Using this distribution, p-values were derived for each cluster in the original data, allowing us to derive statistically significant clusters. No significant clusters was identified by this procedure.

4.2. Time-frequency analysis (TFA)

4.2.1. TFA procedure

To facilitate time-frequency analysis (TFA), the raw continuous EEG data were band-pass filtered between 0.01 Hz and 50 Hz, and then segmented into epochs from −200 prior to the verb/classifier onset and 2000 ms after the verb/classifier. This included 1000 ms after the onset of the target noun. Our primary interest here was only the constraint effect prior to the noun onset, but the initial epochs covered a longer time window in order to increase the frequency resolution in the TFA analysis. But we only report the results below for the 1000 ms time window prior to the noun onset (starting from the verb/classifier onset). Trials were grouped into *high* and *low constraint* respectively for the *verb* and *classifier* condition. Additionally, to balance the number of trials between high and low constraint condition, trials in the high constraint expected condition were not included in the TFA. As such, the *high constraint* and *low constraint* conditions each comprised of 60 trials. Epochs with eye movements, blinks, or other artifacts were removed in the same way as the ERP data pre-processing described in Section 2.5. After artifact rejection, the mean remaining number of trials per participant per condition was: verb high constraint 47; verb low constraint 48; classifier high constraint 47; classifier low constraint 47.

As the TFA can complements the ERP analysis and provide information on the neural oscillatory dynamics that cannot be observed in the ERP results, the current TFA was performed on non-phased-locked EEG activities. EEG not only contains synchronous neural activity phase-locked to a stimulus (i.e. ERP, also known as "evoked" responses), but also oscillatory activities that are not necessarily time- or phase-locked

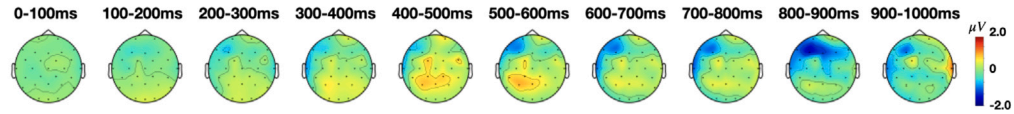

Fig. 10. The Classifier construction: Topographic maps of ERP amplitude differences between High Constraint and Low Constraint conditions. ERP amplitudes averaged over every 100 ms duration from the classifier onset to 1000 ms after the classifier onset.

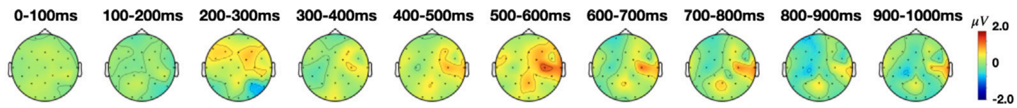

Fig. 11. The Verb construction: Topographic maps of ERP amplitude differences between High Constraint and Low Constraint conditions. ERP amplitudes averaged over every 100 ms duration from the verb onset to 1000 ms after the verb onset.

to the stimulus ("induced" responses; Bertrand et al. (2000)), which are canceled out through the averaging process across trials in the time-domain. To isolate the induced EEG, the ERP (generated from averaging the EEG across trials) was subtracted from each of the single-trial EEG (Crone et al., 2001; Kalcher & Pfurtscheller, 1995). The resulting single-trial induced EEG were then decomposed into time-frequency representations using a moving window Fast Fourier Transform (FFT) approach implemented in the MATLAB toolbox Fieldtrip (Oostenveld et al., 2011). Specifically, the single-trial data were convolved using a Hanning-tapered 500-ms window that moves in steps of 10 ms along the temporal dimension. In the spectral dimension, a set of sinusoidal wavelets with linearly increased cycles from 2 cycles for the lowest frequency (2 Hz) to 10 cycles for the highest (50 Hz) was used. No baseline correction was used, but the spectrograms from each condition were divided by the average power across all four conditions to yield relative power changes, then transformed into a dB value ($10*\log10$ of the signal).

Statistical evaluation between *high* and *low constraint* condition were performed using non-parametric cluster-based permutation tests (Maris & Oostenveld, 2007). First, we computed a dependent t-test to quantify the between-condition difference at each time point-, frequency- and channel-pair. These t-values were used to define the clusters for the non-parametric statistical testing: clusters comprised samples whose t-values were above threshold (a significance level of .05) and were adjacent spatially and temporally. Cluster-level statistics were the sum of all t-values within the cluster, and the cluster with the maximum sum was selected. The distribution of the cluster-level statistics under the null hypothesis was generated by random relabeling of the conditions 1000 times, and computing the cluster-level t-values for each randomization. On the original data, clusters whose test-statistics fell within the top 5th percentiles of the null distribution were considered significant. The differences between high and low constraint were separately examined within the verb and classifier conditions.

To further quantify the differences between *verb* and *classifier* with respect to the context constraint effect (high vs. low constraint), we extracted the time course of frequency bins emerged as significant in the above permutation test, and averaged the power across these bins for high and low constraint respectively. The time course of context constraint effect was computed by subtracting the power of low constraint from that of high constraint. The time courses of context constraint effect were subsequently compared between verb and classifier condition by contrasting average power estimates at each time point using a permutation test. Specifically, paired sample t-tests were conducted for each comparison using the original data (i.e. observed t-values) and 1000 random between-condition permutations of the data (i.e. permuted t-values). For each permutation, data points were randomly assigned to either the verb or classifier condition without replacement. For each set of permutation, the maximum t-value was recorded and used to estimate the distribution under the null hypothesis. The observed t-values were then compared against the null distribution, with values above the 95th percentile rendered as significant.

4.2.2. TFA results

The effects of context constraint on power modulations prior to the onset of the critical nouns (−1000 ms to 0 ms) in the *verb* condition are illustrated in Fig. 12A, which provided an initial overview of the between-condition differences across the induced EEG's spectrum. Time-frequency representations to low constraint seemed to show stronger synchronization in the frequency range of alpha and beta as compared with that to *high constraint* (Fig. 12B). Permutation test revealed significant clusters between 26 to 28 Hz from −291 to −21 ms, and between 13 to 17 Hz from around −433 ms to the noun onset ($p = .034$) with a posterior distribution, suggesting an alpha/beta power suppression in the *verb high constraint* context prior to the onset of critical nouns. The power modulations in *high* and *low constraint* prior to noun onset in the *classifier* condition are exhibited in Fig. 13A. We found a similar alpha/beta power decrease occurred in the *classifier high constraint* context, cluster $p = .005$ (Fig. 13B). Specifically, the cluster spanned frequencies from 19 to 26 Hz between −874 ms and −600 ms, and from 17 to 24 Hz between −397 ms to 0 ms on posterior channels. These findings suggest that an alpha/beta power suppression as modulated by context constraint was also observed in the *classifier* condition.

To compare the difference between verb and classifier condition with respect to the alpha/beta suppression occurred before noun onset, the time courses of power modulation to high and low constraint in the frequency range of 10 to 30 Hz on posterior channels (including Pz, P3, P4) were extracted and averaged respectively for verb and classifier (Fig. 14). This frequency range encompassed frequencies that emerge as significant when comparing *high* and *low constraint* in both *verb* and *classifier* condition. Furthermore, the constraint difference was computed by subtracting the alpha/beta power modulation in the *low constraint* from that in the *high constraint*, and the time courses of constraint difference were then compared between *verb* and *classifier* condition via a permutation test. Results reveal significant differences from −590 ms to −500 ms prior to the onset of the critical nouns ($p < .05$, $N = 1000$ permutations), with stronger alpha/beta suppression (in the high relative to the low constraint context) observed in the *classifier* condition than in the *verb* condition.

To summarize, prior to the noun onset, we did not find any contextual constraint effect reflected by the ERP results. This is different from some previous findings (Li et al., 2017; Ness & Meltzer-Asscher, 2018, 2021). However, based on the results from the time-frequency domain analyses, we observed neural oscillatory modulations in the alpha and beta frequency band prior to the onset of the target nouns, with stronger alpha/beta suppression for high relative to the low constraint context. These context-induced power modulations were exhibited in both verb and classifier conditions, but contextual constraint seemed to exert stronger alpha/beta suppression for the high constraint sentences in the classifier context than in the verb context.

One potential confound that might impact our interpretation of the TFA results is that, as shown in Table 2, classifiers in the high constraint condition were less frequent than those in the low constraint condition

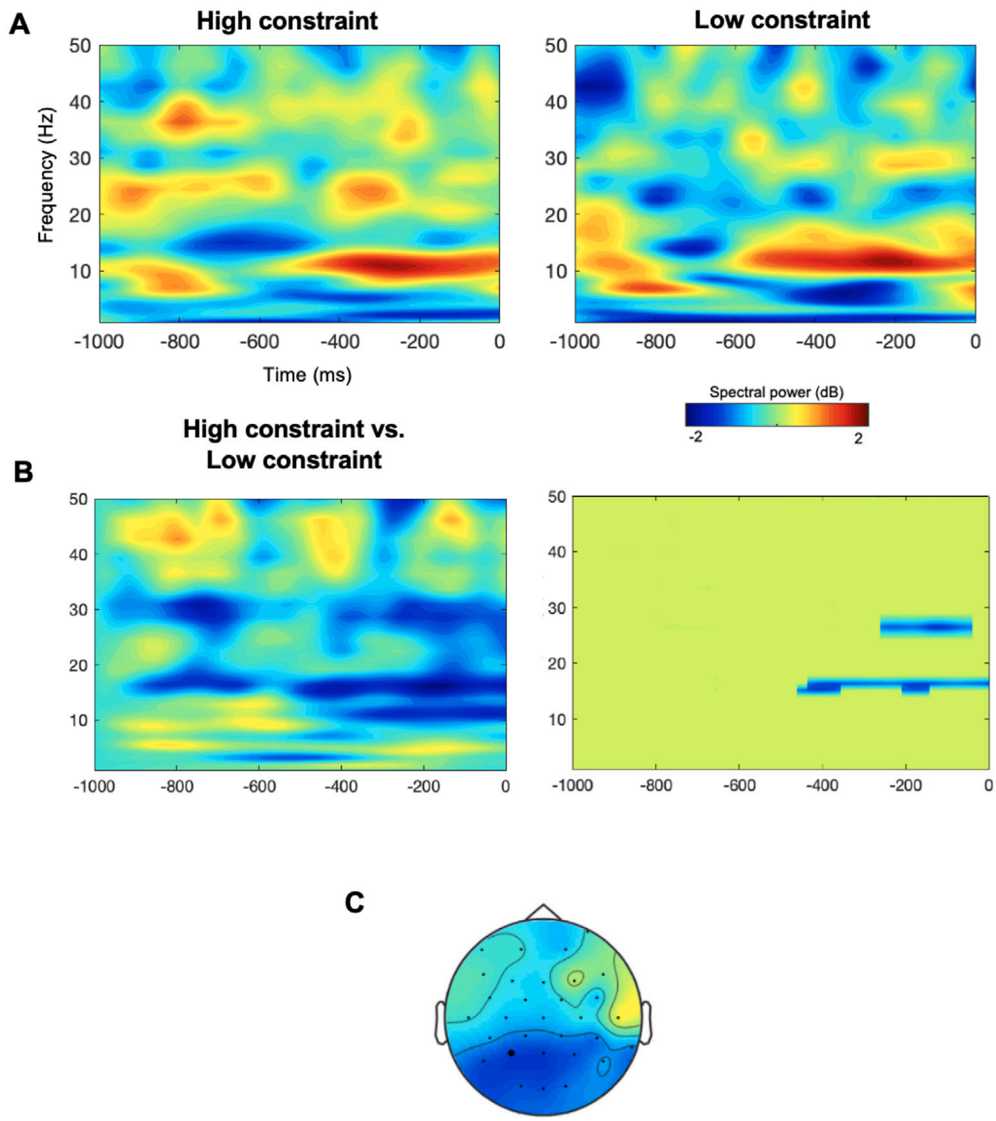

Fig. 12. A. Time-frequency plots of power changes in high constraint (left) and low constraint (right) in the verb condition at a parietal channel (P3, indicated with a black dot in the scalp map). Time zero indicates the onset of the critical noun; negative values indicate time prior to its onset. **B.** Contrast between the conditions in raw difference (left) and statistically significant difference (right), showing a stronger alpha/beta decrease in response to high constraint compared with low constraint. **C.** Difference scalp topography of the alpha/beta (13 to 17 Hz) decrease from −433 ms to 0 ms (noun onset).

($t = -3.89$, $p < 0.001$). It is therefore possible that the differences we observed between high vs. low constraint classifier conditions could be due to a lexical frequency effect in the time-frequency domain. This concern does not arise for the verb conditions, since the verb frequencies under the high and low constraining conditions were not different ($t = 0.33$, $p > 0.7$). To evaluate whether there is an independent effect of lexical frequency in the time-frequency domain, we performed a separate TFA analysis. The details of this additional analysis were presented in Appendix B. In summary, we categorized the classifiers used in our study into high and low frequency groups and found no significant difference between these groups in the new TFA analysis. This finding alleviates concerns that the constraint effect observed in the classifier context might be merely a side effect of lexical frequency. Interestingly, while lexical frequency of verbs did not confound our main objective, we did observe a lexical frequency effect for verbs (high vs. low frequency groups) in the new TFA analysis. The contrasting lexical frequency effects between classifiers and verbs in the time-frequency domain present an intriguing topic for future investigation. Another potential concern arises from the general classification of classifiers as close-class words and verbs as open-class words. It remains uncertain whether the close-class versus open-class classification could influence the constraint effect in distinct ways, guided by independent principles. The current study was unable to investigate this matter in depth, as the close-class vs. open-class distinction completely mirrors the classifier vs. verb distinction. We leave this question for future work.

5. General discussions

The current study compared the contextual effect on word processing in two different types of contexts: the classifier-noun and the verb-noun contexts. In this section, we compare the similarities and differences between these two contexts. Our discussion will focus on two different time windows, the ERP and neural oscillatory activities prior to the critical target noun, and the N400 and the post-N400 frontal positivities measured on the target noun.

5.1. Similarities between the two contexts: the N400 window

The N400 patterns on the target noun across the two contexts are very similar to each other. Under both contexts, the amplitudes of

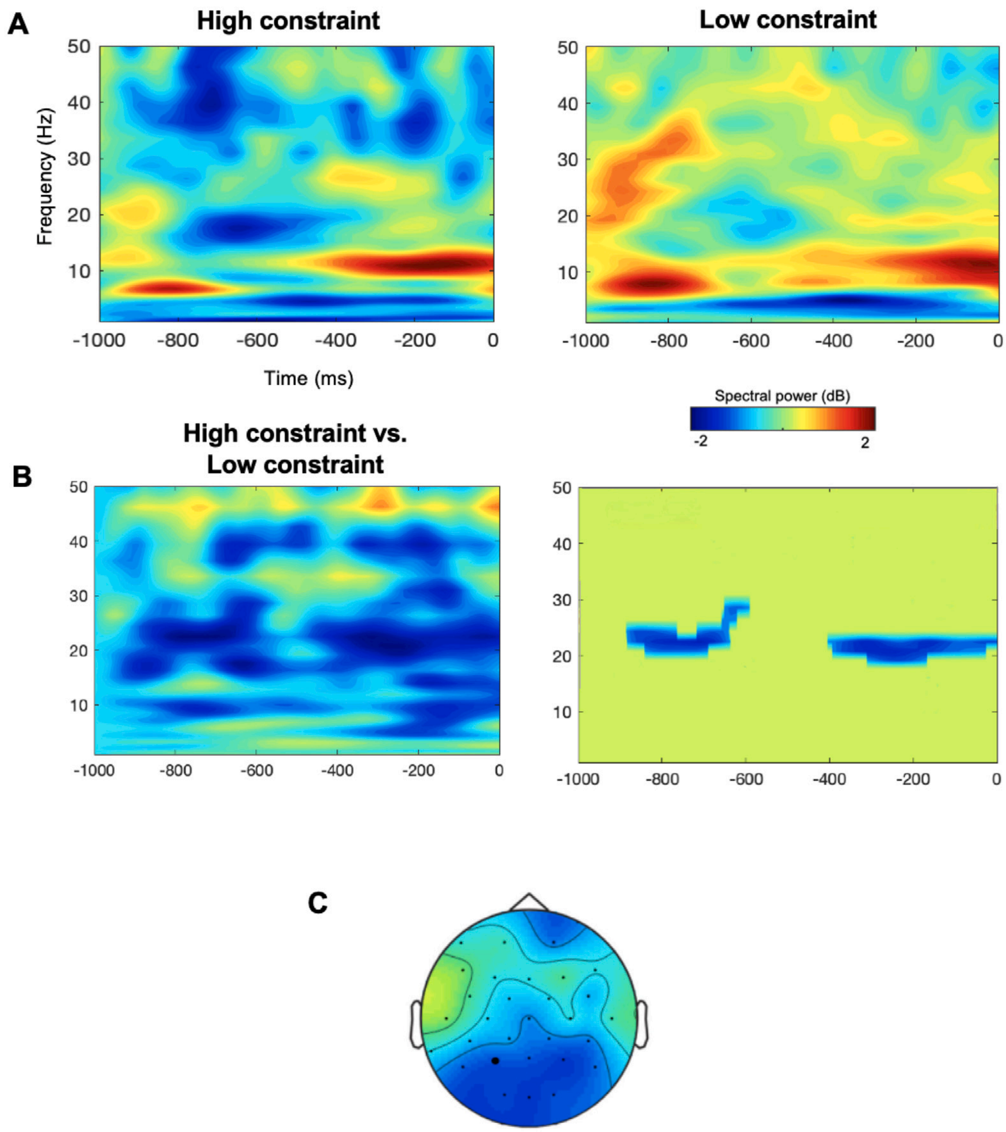

Fig. 13. **A.** Time-frequency plots of power changes in high constraint (left) and low constraint (right) in the classifier condition at a parietal channel (P3, indicated with a black dot in the scalp map). Time zero indicates the onset of the critical noun; negative values indicate time prior to its onset. **B.** Contrast between the conditions in raw difference (left) and statistically significant difference (right), showing a stronger alpha/beta decrease in response to high constraint compared with low constraint. **C.** Difference scalp topography of the alpha/beta (17 to 24 Hz) decrease from −397 ms to 0 ms (noun onset).

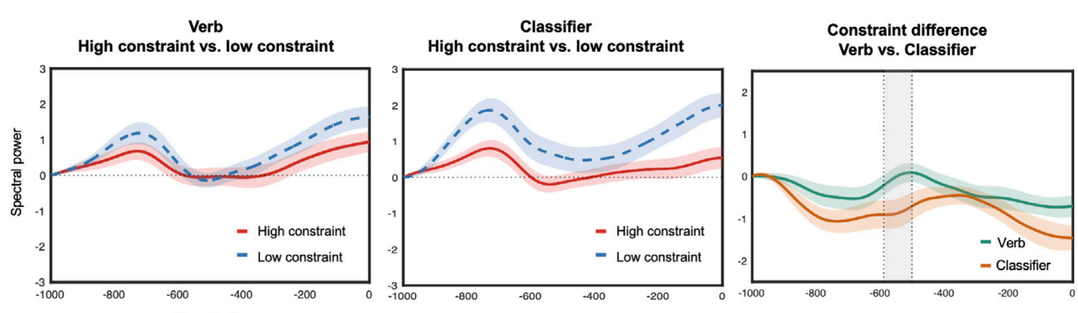

Fig. 14. **A.** Time courses of power modulations averaged from 10 to 30 Hz and from Pz, P3 and P4 channels in high and low constraint in verb (left), and classifier condition (middle), as well as the constraint difference as computed by subtracting power in low constraint from that in high constraint (right). Gray bar indicates significant between-condition difference (verb vs. classifier) in time (via permutation test, $p < .05$). Shaded regions indicate mean ±1 standard error.

the N400 on the target gradually increased in the order of expected < unexpected < anomalous noun. This pattern is expected given the large body of literature showing that N400 is modulated by word predictability and plausibility (see Kutas & Federmeier, 2011 for a review). The higher degree of semantic feature match between the target word and the contextual expectation, the smaller N400 one would observe, either due to preactivation of semantic features on the target words (Kutas & Federmeier, 2000; Van Berkum, 2009) or due to the relative ease to integrate target words that better match contextual expectations (Brown & Hagoort, 1993; Hagoort et al., 2009). In our experiment, expected target words better matched contextual expectations than unexpected but coherent words, which in turn matched context to a greater degree than incoherent anomalous target words. We therefore observed a spine of N400 increase.

Also consistent with the previous literature (Federmeier et al., 2007; Kuperberg et al., 2020; Kutas & Hillyard, 1984; Van Petten & Luka, 2012), under both the classifier-noun and the verb-noun contexts, N400 did not show sensitivity to the cost of violated expectations. Unexpected target words in strongly and weakly constraining sentences elicited N400 of comparable amplitudes. We also show that N400 amplitudes to anomalous continuations in highly constraining and weakly constraining contexts are indistinguishable. Taken together, the evidence suggests that N400 reflects the degree of match of semantic features between target word and preceding context, rather than the cost of revising the disconfirmed expectation. And since the contextual constraint and the cloze probability differences between the target words and the context were controlled to be comparable across the classifier-noun and verb-noun contexts, we observed no differences between these two types of contexts during the N400 window.

5.2. Differences between the two contexts: post-N400 window and pre-target neural oscillations

Regarding the contextual constraint effect, the neural oscillation patterns obtained prior to the target noun revealed an important difference between the classifier-noun vs. verb-noun context. The alpha/beta suppression was greater after a strongly constraining classifier than after a strongly constraining verb, relative to their respective weakly constraining baseline. Since we have controlled the contextual constraints to be similar across these two contexts, as operationalized by the maximum cloze probability of the upcoming target noun (see Table 1), the difference in alpha/beta suppression between the two contexts is informative. One possibility is that a classifier can engage participants more deeply to anticipate a specific upcoming noun. Since the relationship between a classifier and a noun is a learned collocation, it is straightforward for participants to use a classifier to anticipate a specific noun. The greater alpha/beta suppression may reflect higher working memory demands related to pre-selection and maintenance of anticipated lexical candidates (Piai et al., 2014, 2015). For the verb-noun context, on the other hand, although there was also greater alpha/beta suppression following the high-constraint vs. the low-constraint verbs, indicative of deeper engagement with the high-constraint verb context, the verb information may not be as effective as classifiers to guide a memory search of specific nouns. As mentioned earlier, using verb information to generate predictions requires memory retrieval of nuanced event structure information, which is both structurally complex and highly context dependent, reflecting people's general knowledge about the world. Retrieving and synthesizing various aspects of event structure information to make predictions, therefore, could be a less effective process than using classifier information to generate predictions. It is interesting to note that although the pre-target-noun window suggests that the ways that verb and classifier information is recruited to generate predictions may be different, the cloze probability effects on the N400 are comparable across the two contexts. This may be due to the fact that the word presentation is relatively slow in the current experiment (i.e. 1000 ms between the verb onset and the noun onset), allowing more time for the verb information to be processed to a level that can generate sufficient predictions about the upcoming noun. Various previous studies have discussed the effect of presentation speed on N400 responses, and we will come back to this issue in Section 5.4.

Another critical difference between the classifier-noun vs. verb-noun context emerged in the post-N400 time window. Unexpected target nouns following verbs elicited larger post-N400 positivities (PNP) in the strongly constraining sentences relative to the weakly constraining sentences, but no reliable difference was found in the classifier context. Not only contextual constraint showed a different effect between the two contexts, cloze probability and plausiblity both affected the post-N400 window in the verb-noun context more strongly than in the classifier-noun context (see Tables 5 and A.2). Specifically, unexpected but congruent target nouns elicited larger PNP than highly expected targets only in the verb-noun context but not in the classifier-noun context; and for both types of contexts, anomalous targets elicited a more sustained central-frontal negativities than the unexpected by congruent nouns after the N400 window, but the effect is stronger in the verb-noun context than in the classifier-noun context. These findings are informative to refine the existing interpretations of the PNP effect, which we turn to below.

5.3. Implications for the post-N400 frontal positivities

The post-N400 mid-frontal positivities observed in the verb context, and the absence of it in the classifier context, helps to shed light on the possible interpretations of previously observed post-N400 frontal positivities (PNP) effects in other studies. As mentioned in the introduction, proposals explaining the PNP effect have attributed it to the detection of the failed prediction through some sort of error signal (Van Petten & Luka, 2012), or an effort to inhibit the previously predicted words when the actual outcome mismatches the prediction (Kutas, 1993), or they may reflect a discourse update process such that the previously anticipated discourse relations and inferences can be revised when the unexpected linguistic input is ultimately integrated (Brothers et al., 2015; Kuperberg et al., 2020). These hypotheses are not totally mutually exclusive. As we discuss further below, while the present findings do not definitively eliminate any of these possibilities, exploring how the current findings can be compatible with these hypotheses contributes to the refinement of the alternative hypotheses themselves.

Let us consider the discourse-update hypothesis of the PNP effect first. Under this hypothesis, when a previously anticipated outcome turns out to be incorrect, the comprehension system needs to go through a process to revise and update its current discourse representation in order to fully integrate the less expected word into the current context. In the current study, the PNP effect was observed in the verb-noun context but not the classifier-noun context. As discussed above, the nature of the contextual constraint under these two contexts is not the same. A specific classifier regulates the space of possible following nouns through memorized idiosyncratic collocations. When a less expected noun appeared in the input, people could simply check whether the noun is on the list of the memorized nouns that can combine with the given classifier. This process does not require a revision of any high level representations. It is only a revision of a very local relation. On the other hand, a verb-noun combination describes a partial event. Even though the verb-noun combination is presented as a simple phrase, and it is not embedded under any larger linguistic context, people's perception of the event structure is still grounded in their general semantic and episodic knowledge about what action can be done to what object. A verb constrains the space of possible upcoming object noun by evoking both grammatical and world knowledge about what event is possible at which time, in what location, with what type of participants and instruments. All these information would conspire to generate an expectation for the upcoming noun. If the expectation turns out to be incorrect, and a less expected noun needs to be integrated with the verb, the event structure has to be revised, and a new set of event relations need to be established. This is a much more extensive revision process than the revision process required for the classifier-noun context. It is possible that the PNP effect is only reliably evoked as a result of the more complex event structure update in the verb-noun context.

It is worth noting that previous studies supporting the discourse-update hypothesis of the PNP effect usually make use of sentential discourse as the experimental stimuli. For instance, Brothers et al. (2020) showed that there was a greater frontal PNP to unexpected words in multi-sentence contexts with rich information to constrain the target at discourse level. In contrast, in context that contains less information to help construct a rich situation model, there was no frontal PNP effect. The current study only looked at simple two-word phrases instead of elaborated discourse contexts, but we still observed the PNP effect. One difference between the current study and Brothers et al. (2020) is that in our study each word was presented for 850 ms with a 150 ms interstimulus interval blank screen between words, whereas in Brothers et al. (2020) the presentation time was shorter, 450 ms each word plus 100 ms interval between words. It is possible that, even with simple two-word verb-noun phrases, people can still update their analysis of the event structure by retrieving the relevant semantic and episodic knowledge required for the update process; but without the help of an extensive discourse context, it is more effortful and would take longer time for the activation and accumulation of the relevant information to reach the level that can trigger the PNP effect. The longer word presentation in the current study afforded the timing that is necessary for the effect to emerge. If this is on the right track, this suggests that although PNP effects could in general index the discourse update process, the discourse context does not have to be explicitly given in order for the effect to emerge. People have the ability to accommodate a situation model based on the minimal event information presented to them in a two-word phrase, although the update process with fewer explicit contextual cues would be more effortful.

Let us now turn to the other two hypotheses, which consider PNP as an index for prediction error detection or inhibition of a previously expected outcome. Upon initial examination, the current findings may appear to cast doubts regarding these hypotheses. If PNP is simply an indication of prediction failure, it should appear in both the verb-noun and the classifier-noun contexts. Similarly, if PNP reflects the cost of inhibiting previously expected outcome, it should also appear for both the verb and the classifier contexts. The contextual constraints are controlled to be very similar across the verb and the classifier context, and the most expected words under the two contexts also share similar cloze probabilities. Inhibiting the previously expected words therefore should elicit very similar PNP effects. These conclusions, however, are not warranted. As we argued earlier, verbs and classifiers provide different types of predictive cues. In the verb-noun context, people rely on the event structure information to generate predictions about the upcoming noun; whereas in the classifier-noun context, it is the collocation regularities the shape the expectation of the upcoming noun. The oscillatory activities prior to the target noun also suggests that contextual constraint, although matched in strength between the verb and classifier contexts, was recruited in different ways to generate predictions in these two contexts. It is possible then, prediction errors will be detected in these contexts based on distinct types of information as well, leading to different strength of error signals. For example, the error signal may be greater in the verb-noun context because various aspects associated with an event may need to be examined in order to detect an error, making the error detection process a more information-intensive one than the error detection process evoked by the classifier-noun context. Similarly, if an initial expectation turns out to be incorrect and needs to be inhibited, the verb-noun context may evoke stronger inhibition than the classifier-noun context, because inhibiting complex event structure representations in the former context may be a more costly process.

5.4. Limitations of the current study and future questions

In the current study, there are fewer "expected" trials than "unexpected" or "anomalous" trials, since the "expected" trials only appeared under the high constraint condition. We did not include any filler trials to balance the proportion of expected vs. other types of trials. This may have potentially introduced some task effects into the results. Previous work has shown that the strength of a given ERP response could be sensitive to the statistical properties of the overall experimental environment (Li & Ettinger, 2023). For example, N400 has been shown to be impacted by the proportion of trials that encourage prediction (Delaney-Busch et al., 2019; Lau et al., 2013), and late positivities were sensitive to the proportion of trials that contain grammatical violations (Hahne & Friederici, 1999). Therefore it is possible that the N400 and PNP effects we observed may also have been under the influence of the unequal proportions of different types of trials. We will not be able to fully address this question with the current design. But it is worth noting that this potential issue would not affect conclusions about the differences between the verb-noun and the classifier-noun contexts.

As we mentioned earlier, the word presentation time in the current study is longer than many other studies, with each word being presented for 850 ms plus a 150 ms interstimulus interval blank screen between words. This choice was made because we had planned to examine the neural activities prior to the target word, and it was desirable to have a longer analysis window that did not overlap with the activities on the target word. The drawback of this choice is that the word presentation speed is much slower than the regular reading speed. But the benefit of this approach is that not only we were able to carry out the analysis of interest during the pre-target window, the longer time window also afforded us an opportunity to observe patterns that may not have emerged with shorter time. Some previous work has suggested that depending on the information sources underlying the predictive process, different types of information may require different amount of time in order to generate predictions of sufficient strength (Chow et al., 2018, 2016; Liao & Lau, 2020). Some of our findings are in line with this hypothesis. As discussed in Section 5.2, although the constraint effect in the verb context appeared to be weaker than in the classifier contexts, as measured by the neural activities during the pre-target-noun window, both contexts nonetheless appear to have ultimately generated predictions of comparable strength, as reflected by the cloze effect on the N400 of the target noun. This may be due to the fact that people were allowed sufficient time to fully deploy the predictive information in the verb contexts. The benefit of having more processing time may also help explain why we observed PNP effect on simple two-word verb-noun phrases without an elaborated context. As discussed in Section 5.3, without the support of an extensive discourse context, it may be necessary to have more processing time in order to reach the level of sufficient belief update that triggers the PNP effect. Since our study did not specifically manipulate the amount of presentation time, the discussion here remains exploratory. Future work could test whether under a shorter stimulus presentation time, the current experimental design will generate different results.

Finally, our discussion primarily focused on the late frontal positivities following the N400 time window. A different type of late positivity, the posterior positivity, or the P600, have been previously reported to appear on anomalous target words relative to coherent target words (Brothers et al., 2020; DeLong et al., 2014; Friederici et al., 1996; Hagoort et al., 1993; Kim & Osterhout, 2005; Kuperberg, 2007; Kuperberg et al., 2020, 2003; Osterhout & Holcomb, 1992; Osterhout et al., 1994; Van De Meerendonk et al., 2010; Van Petten et al., 1999). But in our results, the amplitude of late posterior positivity elicited by anomalous words in both classifier and verb contexts was not greater than coherent expected/unexpected words. The P600 was often thought to reflect the effort of reanalysis when comprehenders encounter (anomalous) words that are difficult to be integrated into the current context. It is not totally clear why anomalous target words in the current study did not elicit larger P600. One possibility is that the P600 is more likely to arise when the anomalous word is associated with a more elaborated discourse, whereas the current study only looked at simple two-word phrases without an extended context. In Nieuwland and Van Berkum (2005), when sentences with anomalous

Table A.1
By-ROI results of cloze probability, plausibility and constraint in N400 (300–500 ms) time window for Classifier-Noun (CL-N) and Verb-Noun (VB-N) constructions.

Construction	ROI	Cloze probability			Plausibility			Constraint		
		coef	*se*	*t*	*coef*	*se*	*t*	*coef*	*se*	*t*
CL-N	Anterior	0.77	0.36	2.13*	0.96	0.31	3.14**	0.31	0.23	1.38
	Mid-frontal	1.38	0.43	3.16**	1.45	0.35	4.16***	0.35	0.26	1.38
	Mid-posterior	2.21	0.39	5.44***	1.29	0.33	3.98***	0.18	0.27	0.67
	Parietal	1.03	0.33	3.15**	0.14	0.22	0.63	0.002	0.18	0.13
VB-N	Anterior	0.67	0.45	1.50	1.43	0.39	3.65**	0.30	0.21	1.43
	Mid-frontal	1.52	0.53	2.90**	1.84	0.37	4.97***	0.34	0.22	1.53
	Mid-posterior	2.14	0.59	3.60**	1.40	0.36	3.82***	0.31	0.23	1.35
	Parietal	1.31	0.35	3.69**	0.20	0.26	0.78	0.03	0.17	0.16

Table A.2
By-ROI results of cloze probability, plausibility and constraint in Post-N400 (600–1000 ms) time window for Classifier-Noun (CL-N) and Verb-Noun (VB-N) constructions.

Construction	ROI	Cloze probability			Plausibility			Constraint		
		coef	*se*	*t*	*coef*	*se*	*t*	*coef*	*se*	*t*
CL-N	Anterior	−0.35	0.42	−0.84	1.06	0.35	3.06**	0.27	0.25	1.08
	Mid-frontal	−0.45	0.50	−0.90	1.34	0.55	2.44*	0.22	0.29	0.76
	Mid-posterior	−0.06	0.48	−0.12	0.75	0.50	1.52	0.006	0.30	0.02
	Parietal	0.59	0.38	1.57	−0.05	0.30	−0.19	0.06	0.26	0.25
VB-N	Anterior	−1.19	0.43	−2.74**	1.94	0.43	4.43***	0.46	0.23	1.97*
	Mid-frontal	−1.13	0.44	−2.55*	2.04	0.40	5.03***	0.68	0.26	2.58*
	Mid-posterior	−0.48	0.52	−0.92	1.20	0.43	2.81**	0.61	0.30	2.05*
	Parietal	0.82	0.41	2.00^	0.46	0.36	1.27	0.16	0.21	0.76

words (*Next, the woman told the suitcase...*) were placed at the beginning of a multi-sentence discourse, there was no P600 effect. Another possibility is that comprehenders are more likely to engage in the process of reanalyzing an anomalous word when their task was to find an answer to a comprehension question. But the current experimental task was a plausibility rating task. For sentences with anomaly, participants could perform the rating task without attempting deeper reanalysis. To better assess the conditions under which the P600 effect would be elicited by anomalous words, we need more future work that more closely examine various factors.

6. Conclusion

A complete understanding of the predictive processing effect in sentence comprehension needs to understand both the facilitation effect when a word satisfies contextual expectation and the cost associated with disconfirmed expectations. By comparing the classifier-noun and verb-noun contexts in Mandarin Chinese, in which both the classifier and the verb can provide predictive cues for the upcoming target noun, the current study was able to tease apart the two types of effects. The differences between the two types of contexts we examined also shed interesting new lights on how the brain deploys different types of information when engaging in the predictive and revision processes.

Supplementary material

Data and scripts can be found on https://osf.io/cpwbt.

CRediT authorship contribution statement

Jiaxuan Li: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing, Visualization. **Jinghua Ou:** Conceptualization, Methodology, Software, Validation, Formal analysis, Writing, Visualization. **Ming Xiang:** Conceptualization, Methodology, Writing, Resources, Supervision, Project administration, Funding acquisition.

Data availability

A link to the data repository (osf) is included in the manuscript (supplementary material).

Acknowledgments

We are very grateful to those who provided invaluable support and assistance in data collection during the challenging period of the COVID-19 pandemic. Special thanks go to Professor Suiping Wang, Professor Bofeng He and the dedicated research assistants Wenjing Shi, Junhao Li and Shuling Zhang (South China Normal University and Guangzhou Medical University). We also thank the anonymous reviewers, Ellen Lau, and the Language Processing Lab at University of Chicago for their feedback. This project was partly supported by the University of Chicago Humanities Division Council.

Appendix A. Post-hoc statistical analysis across ROIs

Besides the results presented in Table 5, we carried out post-hoc analysis on our effects of interests (cloze probability, plausibility, constraint) for each ROI separately. For each of the four ROIs and under each context, we constructed a mixed effect model using the two user-defined contrasts, frequency and number of strokes as the fixed predictors and the maximal by-participant and by-item random effects.[5]. The results of this analysis were presented in Tables A.1 and A.2.

Appendix B. Time frequency analysis on the word frequency effect on pre-target segment

To examine the effect of word frequency of the verb and classifier construction on neural activity prior to the arrival of the critical nouns, trials were grouped into high and low word frequency respectively for the verb and classifier condition, using median as the cut-off for high vs. low word frequency. The median frequency counts for the verb and classifier construction were 18 and 87, respectively. In the verb condition, the mean frequencies for the low and high word frequency groups were 7 and 143, respectively. In the classifier condition, the corresponding mean frequencies were 34 and 369. Similar to the analytical procedures in TFA on the contextual constrain effect, the single-trial induced EEG were decomposed into time-frequency representations

[5] For example, the model at the anterior region is: $Amplitude \sim contrast1 + contrast2 + frequency + stroke + (1 + contrast1 + contrast2 \mid subject) + (1 \mid item)$

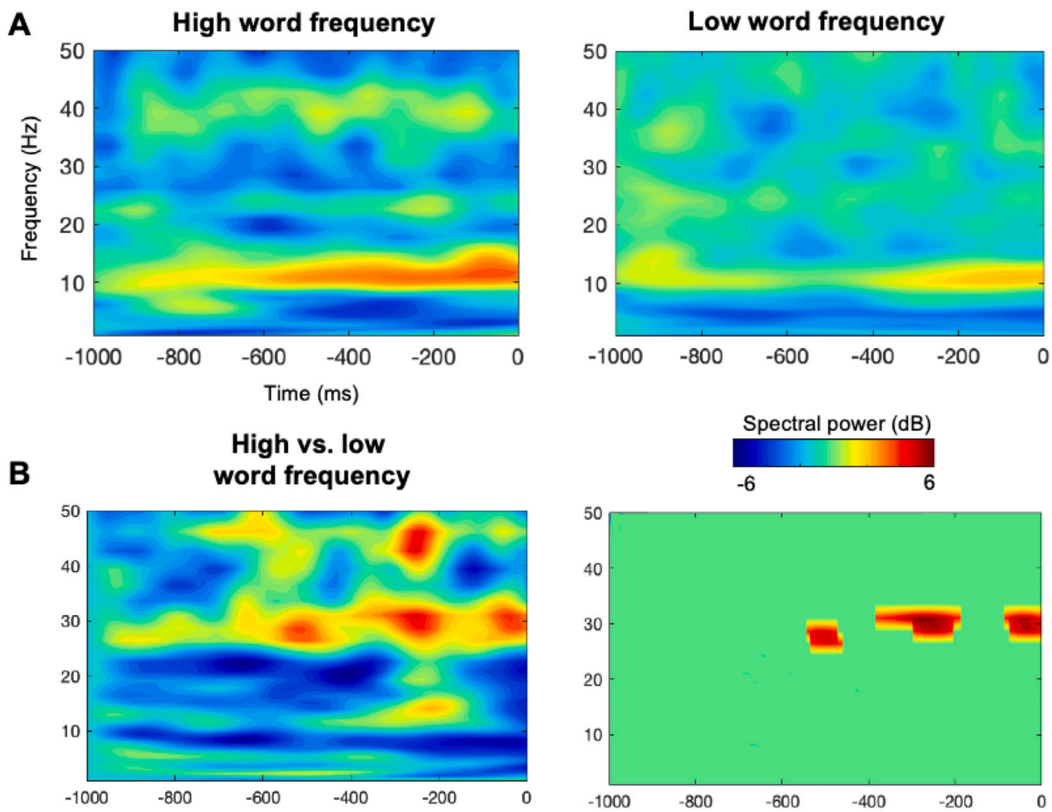

Fig. B.1. **A.** Time-frequency plots of power changes in high word frequency (left) and low word frequency (right) in the verb condition at the P3 channel. Time zero indicates the onset of the critical noun; negative values indicate time prior to the noun onset. **B.** Contrast between conditions in raw difference (left) and statistically significant difference (right), showing a stronger beta increment in high word frequency as compared to low word frequency.

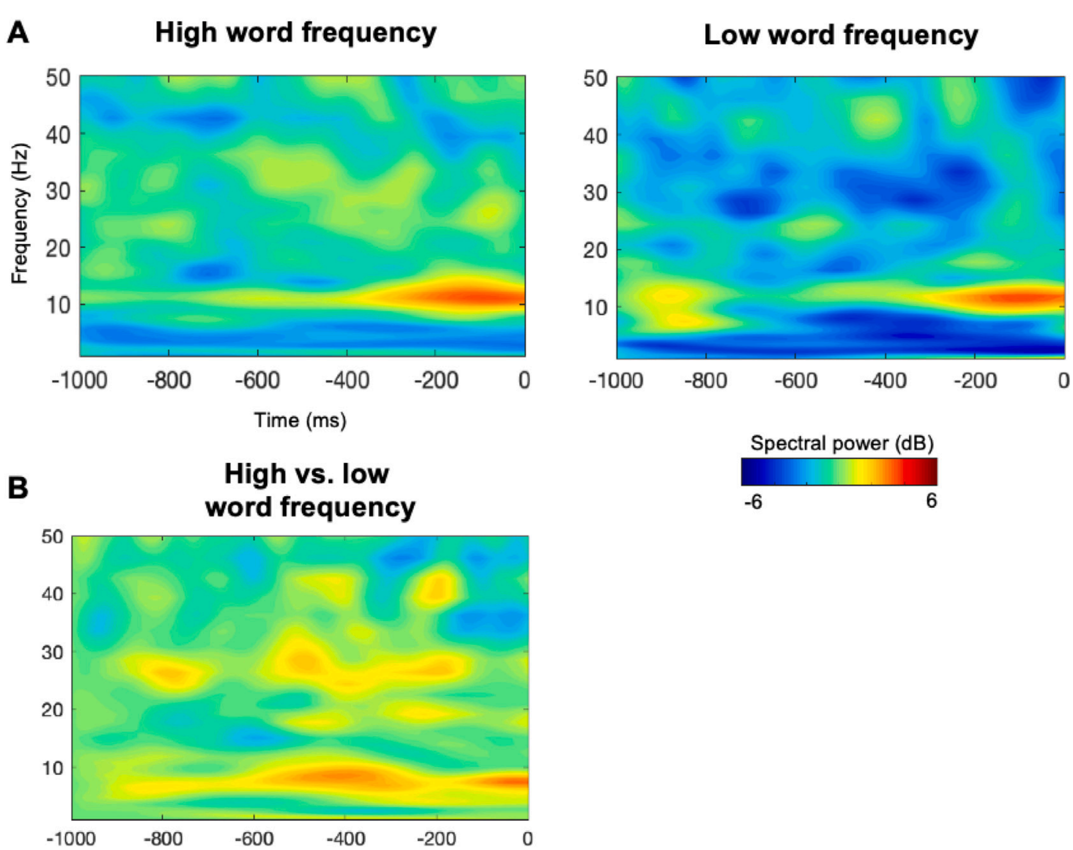

Fig. B.2. **A.** Time-frequency plots of power changes in high word frequency (left) and low word frequency (right) in the classifier condition at the P3 channel. Time zero indicates the onset of the critical noun; negative values indicate time prior to the noun onset. **B.** Contrast between conditions in raw difference, and no significant differences were revealed between the two conditions.

using a moving window Fast Fourier Transform, and the differences in time-frequency representations between conditions were assessed using permutation test (see details in Section 4.2.1).

The effects of word frequency on power modulations prior to noun onset in the verb condition are shown in Fig. B.1A, with stronger power modulation observed in the high beta range for the high word frequency condition as compared to the low word frequency condition (Fig. B.1B). Permutation test revealed significant clusters between 27 to 32 Hz around −570 ms to the noun onset ($p < .05$), suggesting high frequency verbs elicit higher beta power than low frequency verbs. The power modulations in high and low frequency prior to noun onset in the classifier condition are demonstrated in Fig. B.2A. We observed power increment in the alpha frequency range in both high and low word frequency conditions, and permutation test did not reveal any significant clusters between the two conditions of the classifier construction.

References

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

Altmann, G. T., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language*, *57*(4), 502–518.

Berkum, J. J. v., Hagoort, P., & Brown, C. M. (1999). Semantic integration in sentences and discourse: Evidence from the N400. *Journal of Cognitive Neuroscience*, *11*(6), 657–671.

Bertrand, O., Tallon-Baudry, C., & Pernier, J. (2000). Time-frequency analysis of oscillatory gamma-band activity: Wavelet approach and phase-locking estimation. In *Biomag 96* (pp. 919–922). Springer.

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

Brothers, T., Wlotko, E. W., Warnke, L., & Kuperberg, G. R. (2020). Going the extra mile: Effects of discourse context on two late positivities during language comprehension. *Neurobiology of Language*, *1*(1), 135–160.

Brown, C., & Hagoort, P. (1993). The processing nature of the N400: Evidence from masked priming. *Journal of Cognitive Neuroscience*, *5*(1), 34–44.

Bullmore, E. T., Suckling, J., Overmeyer, S., Rabe-Hesketh, S., Taylor, E., & Brammer, M. J. (1999). Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. *IEEE Transactions on Medical Imaging*, *18*(1), 32–42.

Chan, S.-h. (2019). An elephant needs a head but a horse does not: An ERP study of classifier-noun agreement in mandarin. *Journal of Neurolinguistics*, *52*, Article 100852.

Chou, C.-J., Huang, H.-W., Lee, C.-L., & Lee, C.-Y. (2014). Effects of semantic constraint and cloze probability on Chinese classifier-noun agreement. *Journal of Neurolinguistics*, *31*, 42–54.

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.

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.

Crone, N. E., Boatman, D., Gordon, B., & Hao, L. (2001). Induced electrocorticographic gamma activity during auditory perception. *Clinical Neurophysiology*, *112*(4), 565–582.

Delaney-Busch, N., Morgan, E., Lau, E., & Kuperberg, G. R. (2019). Neural evidence for Bayesian trial-by-trial adaptation on the N400 during semantic priming. *Cognition*, *187*, 10–20.

DeLong, K. A., & Kutas, M. (2020). Comprehending surprising sentences: Sensitivity of post-N400 positivities to contextual congruity and semantic relatedness. *Language, Cognition and Neuroscience*, *35*(8), 1044–1063.

DeLong, K. A., Quante, L., & Kutas, M. (2014). Predictability, plausibility, and two late ERP positivities during written sentence comprehension. *Neuropsychologia*, *61*, 150–162.

Delong, K. A., Urbach, T. P., Groppe, D. M., & Kutas, M. (2011). Overlapping dual ERP responses to low cloze probability sentence continuations. *Psychophysiology*, *48*(9), 1203–1207.

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.

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.

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.

Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research*, *1146*, 75–84.

Fields, E. (2017). Factorial mass univariate ERP toolbox [computer software].

Fischler, I. S., & Bloom, P. A. (1985). Effects of constraint and validity of sentence contexts on lexical decisions. *Memory & Cognition*, *13*(2), 128–139.

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Thousand Oaks CA: Sage, Retrieved from https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Frankowsky, M., Ke, D., Zwitserlood, P., Michel, R., & Bölte, J. (2022). The interplay between classifier choice and animacy in Mandarin-Chinese noun phrases: An ERP study. *Language, Cognition and Neuroscience*, 1–17.

Friederici, A. D., Hahne, A., & Mecklinger, A. (1996). Temporal structure of syntactic parsing: Early and late event-related brain potential effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(5), 1219.

Frisson, S., Harvey, D. R., & Staub, A. (2017). No prediction error cost in reading: Evidence from eye movements. *Journal of Memory and Language*, *95*, 200–214.

Friston, K. J., Bastos, A. M., Pinotsis, D., & Litvak, V. (2015). LFP and oscillations—What do they tell us? *Current Opinion in Neurobiology*, *31*, 1–6.

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. *Psychophysiology*, *48*(12), 1711–1725.

Hagoort, P., Baggio, G., & Willems, R. M. (2009). Semantic unification. In *The cognitive neurosciences* (4th ed.). (pp. 819–836). MIT Press.

Hagoort, P., Brown, C., & Groothusen, J. (1993). The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Language and Cognitive Processes*, *8*(4), 439–483.

Hahne, A., & Friederici, A. D. (1999). Electrophysiological evidence for two steps in syntactic analysis: Early automatic and late controlled processes. *Journal of Cognitive Neuroscience*, *11*(2), 194–205.

Hoeks, J. C., 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.

Hsu, C.-C., Tsai, S.-H., Yang, C.-L., & Chen, J.-Y. (2014). Processing classifier–noun agreement in a long distance: An ERP study on Mandarin Chinese. *Brain and Language*, *137*, 14–28.

Hubbard, R. J., Rommers, J., Jacobs, C. L., & Federmeier, K. D. (2019). Downstream behavioral and electrophysiological consequences of word prediction on recognition memory. *Frontiers in Human Neuroscience*, *13*, 291.

Husband, E. M., & Bovolenta, G. (2020). Prediction failure blocks the use of local semantic context. *Language, Cognition and Neuroscience*, *35*(3), 273–291.

Ito, A., Corley, M., Pickering, M. J., Martin, A. E., & Nieuwland, M. S. (2016). Predicting form and meaning: Evidence from brain potentials. *Journal of Memory and Language*, *86*, 157–171.

Ito, A., Gambi, C., Pickering, M. J., Fuellenbach, K., & Husband, E. M. (2020). Prediction of phonological and gender information: An event-related potential study in Italian. *Neuropsychologia*, *136*, Article 107291.

Kalcher, J., & Pfurtscheller, G. (1995). Discrimination between phase-locked and non-phase-locked event-related EEG activity. *Electroencephalography and Clinical Neurophysiology*, *94*(5), 381–384.

Kim, A., Oines, L., & Sikos, L. (2016). Prediction during sentence comprehension is more than a sum of lexical associations: The role of event knowledge. *Language, Cognition and Neuroscience*, *31*(5), 597–601.

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.

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research*, *1146*, 23–49.

Kuperberg, G. R., Brothers, T., & Wlotko, E. W. (2020). A tale of two positivities and the N400: Distinct neural signatures are evoked by confirmed and violated predictions at different levels of representation. *Journal of Cognitive Neuroscience*, *32*(1), 12–35.

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

Kuperberg, G. R., Paczynski, M., & Ditman, T. (2011). Establishing causal coherence across sentences: An ERP study. *Journal of Cognitive Neuroscience*, *23*(5), 1230–1246.

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.

Kutas, M. (1993). In the company of other words: Electrophysiological evidence for single-word and sentence context effects. *Language and Cognitive Processes*, *8*(4), 533–572.

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

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, *62*, 621–647.

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

Kwon, N., Sturt, P., & Liu, P. (2017). Predicting semantic features in Chinese: Evidence from ERPs. *Cognition*, *166*, 433–446.

Lai, M. K., Rommers, J., & Federmeier, K. D. (2021). The fate of the unexpected: Consequences of misprediction assessed using ERP repetition effects. *Brain Research*, *1757*, Article 147290.

Lau, E., Almeida, D., Hines, P. C., & Poeppel, D. (2009). A lexical basis for N400 context effects: Evidence from MEG. *Brain and Language*, *111*(3), 161–172.

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.

Lewis, A. G., & Bastiaansen, M. (2015). A predictive coding framework for rapid neural dynamics during sentence-level language comprehension. *Cortex*, *68*, 155–168.

Lewis, A. G., Wang, L., & Bastiaansen, M. (2015). Fast oscillatory dynamics during language comprehension: Unification versus maintenance and prediction? *Brain and Language*, *148*, 51–63.

Li, J., & Ettinger, A. (2023). Heuristic interpretation as rational inference: A computational model of the N400 and P600 in language processing. *Cognition*, *233*, Article 105359.

Li, F., Hong, X., & Wang, Y. (2021). The N400 and Post-N400 positivity effect in mandarin classifier-noun congruence: An ERP study. *Journal of Neurolinguistics*, *57*, Article 100958.

Li, X., Zhang, Y., Xia, J., & Swaab, T. Y. (2017). Internal mechanisms underlying anticipatory language processing: Evidence from event-related-potentials and neural oscillations. *Neuropsychologia*, *102*, 70–81.

Liao, C.-H., & Lau, E. (2020). Enough time to get results? An ERP investigation of prediction with complex events. *Language, Cognition and Neuroscience*, *35*(9), 1162–1182.

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.

Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology*, *88*, 22–60.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190.

Momma, S., Sakai, H., & Phillips, C. (2015). Give me several hundred more milliseconds: Temporal dynamics of verb prediction. In *Talk given at the 28th annual CUNY conference on human sentence processing, Los Angeles, CA*.

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.

Ness, T., & Meltzer-Asscher, A. (2021). From pre-activation to pre-updating: A threshold mechanism for commitment to strong predictions. *Psychophysiology*, *58*(5), Article e13797.

Ng, S., Payne, B. R., Steen, A. A., Stine-Morrow, E. A., & Federmeier, K. D. (2017). Use of contextual information and prediction by struggling adult readers: Evidence from reading times and event-related potentials. *Scientific Studies of Reading*, *21*(5), 359–375.

Nieuwland, M. S., & Van Berkum, J. J. (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.

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*.

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, *31*(6), 785–806.

Osterhout, L., Holcomb, P. J., & Swinney, D. A. (1994). Brain potentials elicited by garden-path sentences: Evidence of the application of verb information during parsing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 786.

Payne, B. R., & Federmeier, K. D. (2017). Pace yourself: Intraindividual variability in context use revealed by self-paced event-related brain potentials. *Journal of Cognitive Neuroscience*, *29*(5), 837–854.

Piai, V., Roelofs, A., & Maris, E. (2014). Oscillatory brain responses in spoken word production reflect lexical frequency and sentential constraint. *Neuropsychologia*, *53*, 146–156.

Piai, V., Roelofs, A., Rommers, J., & Maris, E. (2015). Beta oscillations reflect memory and motor aspects of spoken word production. *Human Brain Mapping*, *36*(7), 2767–2780.

Piai, V., Rommers, J., & Knight, R. T. (2018). Lesion evidence for a critical role of left posterior but not frontal areas in alpha–beta power decreases during context-driven word production. *European Journal of Neuroscience*, *48*(7), 2622–2629.

Qian, Z., & Garnsey, S. (2016). An ERP study of the processing of Mandarin classifiers. *Integrating Chinese Linguistic Research and Language Teaching and Learning*, 59–80.

Roland, D., Yun, H., Koenig, J.-P., & Mauner, G. (2012). Semantic similarity, predictability, and models of sentence processing. *Cognition*, *122*(3), 267–279.

Rommers, J., Dickson, D. S., Norton, J. J., Wlotko, E. W., & Federmeier, K. D. (2017). Alpha and theta band dynamics related to sentential constraint and word expectancy. *Language, Cognition and Neuroscience*, *32*(5), 576–589.

Sharoff, S. (2006). Open-source corpora: Using the net to fish for linguistic data. *International Journal of Corpus Linguistics*, *11*(4), 435–462.

Stanovich, K. E., & West, R. F. (1983). On priming by a sentence context. *Journal of Experimental Psychology: General*, *112*(1), 1.

Staub, A. (2015). The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation. *Language and Linguistics Compass*, *9*(8), 311–327.

Staub, A., Abbott, M., & Bogartz, R. S. (2012). Linguistically guided anticipatory eye movements in scene viewing. *Visual Cognition*, *20*(8), 922–946.

Tai, J. H. (1994). Chinese classifier systems and human categorization. *In Honor of William S.-Y. Wang: Interdisciplinary Studies on Language and Language Change*, 479–494.

Terporten, R., Schoffelen, J.-M., Dai, B., Hagoort, P., & Kösem, A. (2019). The relation between alpha/beta oscillations and the encoding of sentence induced contextual information. *Scientific Reports*, *9*(1), 1–12.

Thornhill, D. E., & Van Petten, C. (2012). Lexical versus conceptual anticipation during sentence processing: Frontal positivity and N400 ERP components. *International Journal of Psychophysiology*, *83*(3), 382–392.

Van Berkum, J. J. (2009). The neuropragmatics of'simple'utterance comprehension: An ERP review. In *Semantics and pragmatics: From experiment to theory* (pp. 276–316). Palgrave Macmillan.

Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 443.

Van De Meerendonk, N., Kolk, H. H., Vissers, C. T. W., & Chwilla, D. J. (2010). Monitoring in language perception: Mild and strong conflicts elicit different ERP patterns. *Journal of Cognitive Neuroscience*, *22*(1), 67–82.

Van Petten, C., Coulson, S., Rubin, S., Plante, E., & Parks, M. (1999). Time course of word identification and semantic integration in spoken language. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(2), 394.

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*(2), 176–190.

Wang, L., Hagoort, P., & Jensen, O. (2018a). Gamma oscillatory activity related to language prediction. *Journal of Cognitive Neuroscience*, *30*(8), 1075–1085.

Wang, L., Hagoort, P., & Jensen, O. (2018b). Language prediction is reflected by coupling between frontal gamma and posterior alpha oscillations. *Journal of Cognitive Neuroscience*, *30*(3), 432–447.

Wicha, N. Y., Moreno, E. M., & Kutas, M. (2004). Anticipating words and their gender: An event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *Journal of Cognitive Neuroscience*, *16*(7), 1272–1288.

Wlotko, E. W., & Federmeier, K. D. (2012). Age-related changes in the impact of contextual strength on multiple aspects of sentence comprehension. *Psychophysiology*, *49*(6), 770–785.

Wlotko, E. W., & Federmeier, K. D. (2013). Two sides of meaning: The scalp-recorded N400 reflects distinct contributions from the cerebral hemispheres. *Frontiers in Psychology*, *4*, 181.

Xiang, M., & Kuperberg, G. (2015). Reversing expectations during discourse comprehension. *Language, Cognition and Neuroscience*, *30*(6), 648–672.

Xue, N., Xia, F., Chiou, F.-D., & Palmer, M. (2005). The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, *11*(2), 207.

Zhou, X., Jiang, X., Ye, Z., Zhang, Y., Lou, K., & Zhan, W. (2010). Semantic integration processes at different levels of syntactic hierarchy during sentence comprehension: An ERP study. *Neuropsychologia*, *48*(6), 1551–1562.