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

Towards a processing model for argument-verb computations in online sentence comprehension

Chia-Hsuan Liao[1], Ellen Lau[2], Wing Yee Chow[3]

[1]Institute of Linguistics, National Tsing Hua University, Hsinchu, Taiwan
[2] Department of Linguistics, University of Maryland College Park, College Park, MD, USA
[3] Research Department of Linguistics, University College London, London, UK

Correspondence: Chia-Hsuan Liao, National Tsing Hua University, 101, Kuang-Fu Road, Sec.2, Hsinchu 30013, Taiwan. Email: chiahsuanliao@mx.nthu.edu.tw

Data statement: The experiment stimuli, pre-processing script, and ERP data are available online at http://dx.doi.org/10.17632/g8gkmk8cwg.1

Towards a processing model for argument-verb computations in online sentence comprehension

Abstract

The current study investigated the processing stages by which the parser incorporates different pieces of information, from clausehood to argument roles, to update predictions about the main verb. Using Mandarin to match word position across relevant conditions, we extend classic ERP findings on the impact of argument role reversals ([The millionaire$_{SUBJECT}$ the servant$_{OBJECT}$ fired] vs. #[The servant$_{SUBJECT}$ the millionaire$_{OBJECT}$ fired]), by investigating cases where one of the nouns is not an argument of the verb ([The millionaire$_{SUBJECT}$ the servant$_{OBJECT}$ fired] vs. #[The millionaire thought [the servant$_{SUBJECT}$ fired...]]). The pattern of N400 responses suggest a three-stage model of argument-verb computation: An initial stage demonstrates sensitivity at the verb to semantic association only. Soon after, responses show partial structure-sensitivity, differentiating whether the noun phrases are arguments of the upcoming verb or not. Only at the last stage do the arguments' roles (e.g. agent/patient) become available to impact computations at the verb.

Keywords: Sentence processing, Argument information, Thematic relations, N400

1. Introduction

Understanding how verbs are related to noun phrases like the subject or object (i.e. arguments) is critical to building a theory of online sentence comprehension. How many such arguments we find, and what grammatical form these arguments take, depends importantly on the properties of the verb. For example, the sentence "the

farmer fled from the wolves" is acceptable, while "the wolves chased from the farmer" is not, due in part to grammatical differences between "flee" and "chase." In addition to describing the event in each clause, verbs can inform our understanding of the semantic relations associated with the subject or object. The subject of an active clause with "flee" names the agent of a fleeing event, while the subject of an active clause with "chase" names the agent of a chasing event. In these ways verbs are highly informative about both the syntax and the semantics of the dependent phrases in their grammatical context.

Since verbs are highly informative about both the syntax and semantics of the dependent phrases, what are the processes by which comprehenders compute verb-argument relations incrementally? Existing psycholinguistic work has shown that when a verb is encountered, its argument structure information can be accessed to constrain the role of an upcoming argument immediately (Bornkessel & Schlesewsky, 2006; MacDonald, Pearlmutter, & Seidenberg, 1994, Wang et al., 2020). By contrast, when an argument *precedes* a verb, its argument role will not be confirmed until the verb is presented, since argument roles are partially determined by the verb. What the predictive parser can do upon encountering the argument is to consider its structural position, case marking, and what kinds of things it denotes, and make the best estimate of what argument role will be assigned to the argument (Kamide, Altmann, & Haywood, 2003). After the verb is subsequently encountered, the predicted argument role can then be checked against the actual list of semantic relations permitted by the verb (Friederici & Frisch, 2000).

The purpose of this paper is to investigate the timecourse by which the parser incorporates different pieces of information from the arguments to predict the upcoming verb. One recent hypothesis suggests that the processing profiles can be

broken into two stages: an earlier stage in which the subset of nouns that denote the verb's arguments are identified to inform verb prediction, and a later stage in which argument role information becomes available to constrain predictions (Chow, Smith, Lau & Phillips, 2016). However, evidence for this idea is still limited. In the current study, a set of novel event-related potentials (ERP) experiments is designed to test this hypothesis more systematically, with the ultimate goal of mapping the time course of argument-verb relation computations. We will use the N400 response to index successful verb prediction, and successful verb prediction in turn as an indicator that relevant linguistic information about argument structure in the context must have been computed by that point in time. To foreshadow the results, we will propose a three-stage model of argument-verb relation computation: (1) word association without structure; (2) sensitivity to argumenthood; (3) sensitivity to argument roles.

Since we will be framing our discussion of the current investigation in terms of processing stages, it is important at the outset to acknowledge differences between the two major classes of incremental sentence processing theories that have dominated the field in recent decades: those that adopt a staged framework (Rayner, Carlson, & Frazier, 1983; Friederici, & Weissenborn, 2007; Bornkessel, & Schlesewsky, 2006) and those that adopt a strength-of-evidence framework (McRae, Spivey-Knowlton, & Tanenhaus, 1998; Kim & Osterhout, 2005, Kuperberg, 2007; Kuperberg, 2016). A staged framework holds that comprehenders pass through discrete stages of computation in the course of comprehending a sentence online. Therefore, this type of framework puts more emphasis on mapping out the time course by which different sources of information are incorporated. By contrast, a strength-of-evidence framework argues that all types of information would be evaluated in parallel. As they are differentially reliable about the underlying event and event structure being

communicated by the producer, different cues would be weighted differently. Therefore, there is no fixed order of the computational processes in sentence comprehension. Here we will largely assume a staged framework in describing our investigation, results, and conclusions. In the General Discussion section, we will compare our model with other staged frameworks. In addition, we will return to the question of how these data might be interpreted under a strength-of-evidence framework.

1.1 Fast vs. slow computations in online sentence comprehension

In working towards a model of real-time comprehension, one important principle that we begin with is the observation that online sentence comprehension is predictive (Federmeier & Kutas, 1999; Federmeier, 2007; Thornhill & Van Petten, 2012). Much evidence has shown that comprehenders actively integrate information from the context to predict what is coming next.

In these experiments, predictability of a word is often quantified by an offline cloze measure, where participants are asked to provide a continuation to a sentence frame, and the percentage of a word used to complete the sentence frame is defined as the cloze probability of the word (Taylor, 1953). For example, given the sentence frame "He bought her a necklace for her ____," a majority of participants provided "birthday" and only a small proportion provided "collection" as the best continuation to the sentence, "birthday," the high-cloze completion, is defined as a predicted word and "collection," the low-cloze one, as an unpredicted word (Federmeier, Wlotko, Ochoa-Dewald, & Kutas 2007). ERP measures then show that relative to an expected word, an unexpected word often elicits a larger ERP response known as the N400. More generally, the N400 amplitude, which peaks between 300-600 ms after the onset of the

stimulus presentation, is negatively correlated with the predictability of a target word (Kutas & Hillyard, 1984; Kutas & Federmeier, 2000; Lau, Phillips, & Poeppel, 2008). Therefore, the N400 response has been used to index the extent to which a word is pre-activated, although there are discussions about whether the N400 reflects pre-activation of conceptual features (Federmeier & Kutas, 1999) or pre-activation of a lexical form (Laszlo & Federmeier, 2009).

In the current study, we consider prediction an umbrella term, and did not differentiate the differences between feature priming and contextual facilitation, although for some researchers (e.g., Pickering and Gambi, 2018), "priming" concerns simple semantic associations whereas "prediction" is about contextual effects during comprehension. We used the N400 as a neural index of prediction, which reflected a combination of pre-activating conceptual features and pre-activating specific lexical items. More importantly, we assume that successful prediction depends on finishing the linguistic analyses of previous sentence context. Therefore, prediction can be seen as a chronometer for linguistic analysis. In other words, we can take the timing of prediction to study how long it takes to compute particular linguistic analyses (Chow, Lau, Wang, & Phillips, 2018; Liao & Lau, 2020).

Since the aim of the current study is to investigate the computation of verb-argument relations, let's turn to what we know about predictions involving verbs and their arguments. A considerable number of studies have shown that when a verb is available in the context, predictions could be updated very quickly. For example, Altmann and Kamide (1999) showed that when presented with a scene of a cake, a car, and two other distractors, participants were faster to look at the cake when they heard the sentence "the boy will eat ____" relative to "the boy will move ____." This example and many others (Altmann & Kamide, 1999; Kuperberg, Wlotko, Riley, Zeitlin, &

Cunha-Lima, 2016) reveal that even with a somewhat limited context, which contains a subject and a verb, and in a visual world paradigm accompanied by a visual scene, comprehenders could immediately access information encoded in the verb, and use it to constrain the prediction of an upcoming argument.

Then, what are the processes involved when only pre-verbal noun phrases are available in the context? What kinds of cues could be helpful to constrain the prediction of an upcoming verb? A considerable number of studies have investigated if the thematic relations of arguments can be established quickly to impact predictions of a verb. This line of research reverses the thematic roles assigned to the pre-verbal arguments and tests if the N400 is sensitive to the thematic anomaly at the verb. Although a few inconsistent results exist—which will be discussed in detail in the Discussion section—a majority of studies show that the N400 is not sensitive to thematic role reversals. In fact, the absence of N400 effect has been replicated among different languages, with various structures. For example, the N400 insensitivity is found in Chow, Smith, Lau and Phillips (2016) with objective relative clause (OSV) in English (e.g. "the customer that the waitress served" vs. "the waitress that the customer served"). It is also observed with simple SOV structure in languages that allow it, such as Mandarin and Dutch (Chow & Phillips, 2013; Chow, Lau, Wang, & Phillips, 2018; Hoeks, Stowe, & Doedens, 2004; Kolk, Chwilla, Van Herten, & Oor, 2003). In addition, the pattern still holds even when there is only one pre-verbal argument (Kuperberg, Sitnikova, Caplan, & Holcomb, 2003; Kuperberg, Kreher, Sitnikova, Caplan, & Holcomb, 2007; Kim & Osterhout, 2005; Momma, Sakai, & Phillips, 2015). The insensitivity of N400 to role reversal situations appears to be incompatible with the classic N400 observations that a low-cloze unexpected target word, or a semantically implausible word, would generate a larger N400 response relative to an expected word.

However, studies like Chow, Smith, Lau and Phillips (2016) have confirmed that there is something special about argument role assignment—even when cloze probability is collected and shown to differ, there is still no N400 difference to role reversal anomaly.

Various accounts have been proposed to explain the absence of N400 effect to role reversal situations (Kim & Osterhout, 2005; Kuperberg, Kreher, Sitnikova, Caplan, & Holcomb, 2007; Hoeks, Stowe, & Doedens, 2004; Kolk, Chwilla, Van Herten, & Oor, 2003, Brouwer, Fitz, & Hoeks, 2012; Kos, Vosse, Van Den Brink, & Hagoort, 2010). Different from most of the existing accounts, which questioned the functional interpretations of the N400 and P600 components, Chow (2013) and Chow, Momma, Smith, Lau and Phillips (2016) proposed the slow prediction hypothesis, which suggested that argument roles may impact predictive computations more slowly than other kinds of information. Further pursuing this idea, Momma, Sakai and Phillips (2015) manipulated presentation rates with two-word Japanese sentences (*bee-nominative sting* vs. *bee-accusative sting*). Their results showed that the N400 was not sensitive to role reversals when the materials were presented at 800 ms presentation rate. However, when the presentation rate was increased to 1200 ms, participants had more time to consider the thematic relations between the argument and the verb, the N400 effect emerged. In a similar spirit, Chow and her colleagues (2018) manipulated the linear distance between arguments and the verb in Mandarin. They found that when the two arguments were adjacent to the verb, the N400 was insensitive to thematic role reversal situations (*Cop ba thief arrest*, meaning "the cop arrested the thief," vs. *Thief ba cop arrest*, meaning "the thief arrested the cop"). By contrast, when a temporal adverbial was inserted between the second argument and the verb, which created a little buffer to update predictions on the verb, the N400 effect became present (*Yesterday cop ba thief arrest*, meaning "the cop arrested the thief yesterday," vs. *Thief ba cop*

yesterday arrest, meaning "the thief arrested the cop yesterday"). The above findings revealed that argument role information could constrain predictions on the verb within at least one to two seconds, although this was notably longer than many other contextual information sources.

1.2 The Bag of Words vs. the Bag of Arguments hypotheses in argument-verb computation

Prior work has shown that argument role information impacts predictions relatively slowly, but what is happening during this long time window before argument role impacts prediction? How are the necessary computations ordered within this time? Prior to argument roles becoming available, do comprehenders just compute basic lexical associations, or can some level of sentence structure be playing a role earlier? Chow, Smith, Lau and Phillips (2016) hypothesized that even before argument role impacts prediction, structure is already impacting prediction in the sense that a subset of noun phrases are identified as arguments of the upcoming verb, and this information can constrain the prediction of the verb. They called this the "Bag of Arguments" hypothesis, extending the classic metaphor by which context effects are represented as the summed associations of an unstructured "bag of words" (the "Bag of Words" hypothesis). As such, the Bag of Words hypothesis predicts quick-and-dirty feature association effects, whereas the Bag of Argument hypothesis suggests that those early associative effects on the verb might be *constrained* by structure. Comprehenders are able to use structural cues to identify if those noun phrases are arguments of the verb.

To test these hypotheses, Chow, Smith, Lau and Phillips (2016) created sentences with three noun phrases in a row (The exterminator inquired which neighbor the landlord had…). The last two noun phrases were placed in an embedded sentence

and the critical verb came at the end of the embedded sentence. N400 responses were evaluated at the embedded verb. By reversing the order of the first two noun phrases, they introduced different arguments in the embedded sentence ("The exterminator inquired which neighbor the landlord had evicted" vs. "The neighbor inquired which exterminator the landlord had evicted"). The Bag of Words hypothesis would predict no N400 differences at the verb between the two sentences. In both cases, the three noun phrases are lumped in the unstructured bag. Prediction would be facilitated as long as the upcoming verb is semantically associated with the noun phrases in the bag. By contrast, the Bag of Arguments hypothesis would predict that facilitative effects on the verb from semantic associates should be greatest when these associates are in argument positions of the verb, as in Figure 1a, compared to a case where one of the associates appears in a non-argument position, as in Figure 1b. In particular, with *neighbor* and *landlord* in the embedded clause, the predicted verb is *evict*. However, evicting would be a less likely event when the arguments in the embedded clause are *exterminator* and *landlord*. Their ERP results revealed a larger N400 response at the verb in sentences like those in Figure 1b than Figure 1a, as predicted by the Bag of Arguments hypothesis.

[Figure 1 around here]

Note that the Bag of Arguments hypothesis holds that argument *roles* do not initially impact the prediction of an upcoming verb. Metaphorically speaking, these arguments are lumped in the bag, so information about their argument roles is not distinguishable for prediction, and this is what explains the many demonstrations of N400 insensitivity to role reversals in the prior literature. Chow, Smith, Lau and Phillips

(2016) included a second experiment where the order of the last two arguments was reversed in the embedded sentence, creating role reversal scenarios ("The restaurant owner forgot which *customer* the *waitress* had served" vs. "The restaurant owner forgot which *waitress* the *customer* had served"). They successfully replicated prior studies by showing a null N400 effect between conditions.

Taken together, Chow, Smith, Lau and Phillips (2016) took their results to support the Bag of Arguments hypothesis, showing that initial verb prediction is constrained by noun phrases that are in the same clause as the target verb. What is implied by this conclusion is that the parser is able to identify which noun phrases could be arguments of the upcoming verb, potentially based on the structure cue provided by the clause boundary. Then, if additional several hundred milliseconds are provided, argument role could constrain predictions of a verb as well (Chow, Lau, Wang, & Phillips, 2018; Momma, Sakai, & Phillips, 2015). These findings imply that there are two stages of argument-verb computations. First, there exists a time window for the parser to identify if the noun phrases could be arguments of the verb, and to use that information to update predictions. Then, a later stage at which the parser is able to update predictions on the basis of argument roles, and construct detailed representations of a sentence.

However, in Chow, Smith, et al. (2016), the noun phrase outside of the embedded clause was in fact linearly further away from the embedded verb (see Figure 1b). In other words, with English sentences, whether that noun phrase could be an argument of a verb is confounded with its linear distance from the verb. The effects they observed could therefore result from a recency effect or priming, without appealing to constraints from grammatical structure like the Bag of Arguments hypothesis.

1.3 The current study

In the current study, our goal is to devise a stronger test of the Bag of Arguments hypothesis, with better control of the linear distance between the noun phrases and a verb. More broadly, our aim is to temporally dissociate different stages of argument-verb computations. We hope that by getting a better understanding of when different pieces of information contribute to the prediction of the verb, we can develop a processing model which identifies and maps out the stages comprehenders go through to compute argument-verb relations.

In the three ERP experiments reported here, the basic logic is the following. We manipulated different kinds of argument information in the context and used the N400 response to the verb, an index of the extent to which the verb is predicted, to ask whether the information has contributed to comprehenders' predictions of the verb by the time it appears. We investigated the amount of time needed for a particular type of argument information to impact verb predictions by manipulating the stimulus presentation rate. All the experiment materials were in Mandarin, which has properties that allow us to keep the linear distance between noun phrases and verbs identical regardless of whether the noun phrase could be an argument of the verb (more explanations below). In the first two experiments we tested for effects of argumenthood and argument role, establishing an initial time frame for the Bag of Arguments processing stage. In Experiment 3 we used a faster stimulus presentation rate to investigate whether a lower bound on this stage can be identified. If there is a time window at which the parser cannot tell if the noun phrases are arguments of a verb, such that only simple associative effects are present (i.e. the Bag of Words hypothesis), then we should revise the two-stage model implied by the Bag of Arguments hypothesis into a three-stage model. Note

that previous research has investigated the effect of presentation rate on language processing (Wlotko & Federmeier, 2015; Camblin, Ledoux, Boudewyn, Gordon & Swaab, 2007). They have generally shown that with a rapid presentation rate, bottom-up semantic association initially dominates processing.

1.4 Data availability

We report all data exclusions and manipulations in the study. The experiment materials, ERP pre-processing script, and ERP data of the three experiments are available on the Open Science Framework platform at http://dx.doi.org/10.17632/g8gkmk8cwg.1. This repository also contains N400 averaged data necessary to reproduce the analyses reported in this paper.

2.1 Experiment 1

In Experiment 1, we tested whether identifying noun phrases as arguments of the verb can be a useful cue to constrain predictions of the verb, when linear distance between the noun phrases and the verb is better controlled. Specifically, the Mandarin *ba* construction places two arguments before the verb (e.g. *Millionaire ba servant <u>fired</u>* meaning "Millionaire fired the servant"). While this sentence is monoclausal, a biclausal sentence could be introduced with the same noun order simply by replacing *ba* with a clausal verb, such as *think* (*Millionaire thought servant <u>fired</u>*...), so that the verb is in the embedded clause and no longer predicted by the context (see Figure 2). The Bag of Arguments hypothesis suggests that comprehenders identify noun phrases that could be the arguments of the verb relatively quickly. In this example, if both *servant* and *millionaire* are identified as arguments of a verb, it is more likely that the verb is *fire* than if *servant* is the only argument in the "bag." If this is the process used

by comprehenders, then we expect to observe a smaller N400 response at the verb in the one-clause *ba* condition compared to the two-clause *think* condition. This is the prediction evaluated in Experiment 1. The Bag of Arguments hypothesis also suggests that arguments are identified and contribute to predictions earlier than thematic roles do; metaphorically speaking, all the relevant arguments are initially lumped in the bag, with argument roles undefined. We will test whether thematic roles impact the N400 response under these same conditions in Experiment 2.

[Figure 2 around here]

We relied on previous role reversal studies to determine the presentation rate of Experiment 1. As far as we could tell, a stimulus onset asynchrony (SOA) of 800 ms was the slowest presentation rate in which argument role reversals did not modulate the N400 response (Momma, Sakai, & Phillips, 2015), and thus this rate seemed like a good place to start in narrowing in on the hypothesized time window in which argument(s) of a verb could impact prediction but not the role bounded by the argument.

2.1.1 Participants

The participants were 40 naive young adults (12 male and 28 female, 18-40 years old, mean: 24) from National Taiwan Normal University. All of them were right-handed native Mandarin speakers, with no a history of neurological or psychiatric disorders. Of the 40 participants, 7 were excluded after pre-processing because of excessive eye blinks, muscle potentials, sweat artifact and alpha waves. The reported results were obtained from the remaining 33 participants (15 male and 18 female, 19-40 years old, mean: 24). Informed consent was obtained from all participants. The

experiment protocol was approved by the Institutional Review Board Office at the University of Maryland College Park.

2.1.2 Materials

Materials were sentences adapted from Experiment 1 in Chow, Lau, Wang, and Phillips (2018). We began by selecting 60 sentences, all of which used the SOV *ba* construction in Mandarin. In particular, the construction requires a transitive verb, and the morpheme *ba* always follows an agent argument and is immediately followed by a patient argument. That is, in this construction, unambiguous and reliable cues about the arguments' syntactic roles are available before the presence of the verb. In our experiment setup, the two preverbal arguments were always animate. None of the target verbs were repeated, and the predictability of the target verb, as measured from the cloze norming in Chow, Lau, Wang and Phillips (2018), was 38%. From these 60 baseline sentences, we replaced the morpheme *ba* with the verb *think* to create another 60 sentences as the critical complement sentences. In other words, the two conditions for the experiment were (1) Baseline condition, with the two noun phrases presented in a canonical SOV word order (*Millionaire ba servant fired*, meaning the millionaire fired the servant) and (2) Complement condition, with the verb *think* separating the two noun phrases into different clauses (*Millionaire thought servant fired …,* meaning the millionaire thought the servant fired …) (see Table 1). Since replacing *ba* with *think* would introduce a clause boundary between two noun phrases, the critical verb, which was then embedded in a subordinate clause, became much less predictable based on the second noun phrase alone. A post-hoc cloze norming experiment showed that the predictability of the target verb in the Complement condition was 0%. Note that the two conditions had different post-target verb continuations, as they had very different

structure requirements. For the Baseline condition, the two pre-verbal arguments had satisfied the argument structure requirements of a transitive verb. By contrast, when the transitive verb was embedded in a subordinate clause, such as in the Complement condition, another argument was still needed in the subordinate clause to make the sentence grammatical. Depending on the length of the continuations, the length of our sentences ranged between six to nine words long. Even though the length of the sentence varied, the number of words was always identical up to reaching the target verb between conditions. Lastly, we adapted the materials to accommodate small lexical differences in language use between Mandarin speakers in China and Taiwan. The 120 sentences were divided into two lists in a Latin square design.

Condition	Sentence context	Post target verb continuation
Baseline	富翁 把 僕人 **解雇了**	之後 立即 請來了 新的 管家
	Millionaire ba servant **fired**	then immediately hired new housekeeper
	"The millionaire had fired the servant and then immediately hired a new housekeeper."	
Complement	富翁 認為 僕人 **解雇了**	童工 很 不 應該
	Millionaire thought servant **fired**	kid very not should
	"The millionaire thought that it was inappropriate for the servant to fire the kid."	
High cloze	駭客 忘掉了 **密碼**	，無法 執行 任務
	The hacker forgot **passwords**	, failed execute plan
	"The hacker forgot the passwords, so he failed to execute the plan."	

Low cloze	駭客 忘掉了 **登出**	，不小心 露出馬腳
	The hacker forgot **logout**	, accidentally gave the game away
	"The hacker forgot to log out, and that gave the game away."	

Table 1: Example stimulus in each condition in Experiment 1

To check that comprehenders did engage predictive mechanisms during the experiment that modulated N400 amplitude, we also included 30 pairs of sentences instantiating a cloze contrast (High cloze: 38% vs. Low cloze: 9%) as our control items. The cloze contrast in the control sentences was slightly smaller compared with the critical experimental items (Baseline: 38% vs. Complement: 0%). All of the control sentences were grammatical and semantically plausible. Different from the experimental conditions, the control sentences were of simple SVO structure, with predictability being examined at the object noun position (e.g. *The hacker forgot the passwords / logging out*) (See Table 1). Here, prediction was updated based on the information provided by a subject and a verb. The 30 pairs of sentences were counterbalanced between two lists.

Two presentation lists were constructed such that no sentence context or target word was presented twice within either list. Each list consisted of 240 sentences, including 30 sentences in the Baseline condition, 30 sentences in the Complement condition, 30 sentences of high-cloze target in the High cloze condition, 30 sentences of low-cloze target in the Low cloze condition, and an additional 120 grammatical and plausible filler sentences that were reported in Liao & Lau (2020). Participants were randomly assigned to one of the two lists. The presentation order of the sentences was randomized.

2.1.3 Procedure

Participants sat in front of a computer screen with their hands on a keyboard. Sentences were segmented into words, which were presented in a rapid serial visual presentation (RSVP) paradigm in a white font (traditional Chinese characters) on a black background at the center of the screen (see Figure 3). Each sentence was preceded by a fixation cross that appeared for 600 ms. Each word appeared on the screen for 600 ms, with a 200 ms inter-stimulus interval, for a stimulus-onset asynchrony (SOA) of 800 ms. At the end of 20% of the trials, a sentence would appear on the screen. Participants were asked to judge if it was a good paraphrase of the sentence they just read by pushing one of two buttons to proceed to the next trial.

Prior to the experimental session, participants were presented with six practice trials with feedback to familiarize themselves with the task. The experimental session was divided into 4 blocks of 60 sentences each, with short pauses in between. Including set-up time, an experimental session lasted around 90 minutes.

[Figure 3 around here]

2.1.4 Data acquisition and analysis

E-prime 2.0 (Psychology Software Tools Incorporated) was used to present the experimental stimuli, record participants' behavioral data, and send the event codes to the digitization computer. EEG was recorded from 30 electrodes placed according to the 10/20 system (FP1, FP2, F7, F3, FZ, F4, F8, FT7, FC3, FCZ, FC4, FT8, T3, C3, CZ, C4, T4, TP7, CP3, CPZ, CP4, TP8, T5, P3, PZ, P4, T6, O1, OZ, O2). Each channel was referenced to an average of the left and right mastoids for both online and off-line

analyses. Four additional electrodes were placed (two on the outer canthus of each eye and two on the upper and lower ridge of the left eye) to monitor blinks and horizontal eye movements. The impedance of all the electrodes was kept below 5 kΩ. EEG signals were continuously digitized at 1000 Hz, filtered between DC to 100 Hz (NuAmps, NeuroScan Incorporated).

ERP analyses were time-locked to the onset of the verb for the critical conditions and to the onset of the noun for the control items. The EEG data were processed with EEGLAB (Delorme & Makeig, 2004) and ERPLAB (Lopez-Calderon & Luck, 2014) in Matlab (MathWorks, Inc.). A linear derivation file was first imported to convert the four monopolar eye-movement monitoring channels to two bipolar channels (VEOG and HEOG). We applied a notch filter at 60 Hz and an Infinite Impulse Response (IIR) filter with the band-pass value set between 0.1 Hz to 30 Hz, 12 dB/oct. Then we extracted epochs of length from 100 ms before to 800 ms after stimulus onset. Baseline correction was applied with the pre-stimulus -100 to 0 ms interval. After baseline correction, artifact rejection was carried out by reviewing the epochs both automatically and manually: At each channel, a 200-ms window was moved across the data (100 ms before and 800 ms after the stimulus) in 100-ms increments and any epoch where the peak-to-peak voltage exceeded 70 μV was rejected. We then reviewed the data, and adjusted the voltage threshold for individual subjects, to ensure that epochs contaminated by excessive blinking, body movements, skin potentials, and amplifier saturation were rejected. The mean rejection rate across participants was 19.2 ± 11.9% (mean ± SD); participants with more than 40% of the trials rejected were excluded from further analysis. The following were the rejection rates for each condition: Baseline: 20.0 ± 12.4%; Complement: 17.9 ± 12.7%; High cloze: 21.1 ± 12.1% and Low cloze: 18.0 ± 10.5%.

Our hypotheses centered around the N400 response at the verb for the critical comparisons and at the noun for the control items, so we selected nine electrodes over the central-parietal area (C3, CZ, C4, CP3, CPZ, CP4, P3, PZ, P4), known to show the most prominent N400 effect. We carried out a paired t-test on the mean amplitudes in the measurement time window of 300-500 ms, evaluating effects of Predictability (Baseline, Complement). The control items were designed to replicate standard N400 effects of cloze probability, and we carried out a paired t-test over the same set of electrodes evaluating the effect of Cloze probability (High cloze, Low cloze). When a null effect was observed, we supplemented our analysis of the target words with a Bayes Factor analysis to quantify the likelihood of the null hypothesis relative to the alternative one (BF_{01}, H0: H1). Since the goal was to quantify evidence of our null effect, the mean of the prior distribution of a null effect should be zero. We used data from Experiment 3 in Chow, Lau, Wang, and Phillips (2018) to define the width of the prior distribution, as the materials in the current study were adapted from that experiment. The standard error of the N400 effect in that experiment was 0.72. Thus, we use an informative prior with a mean of zero and standard deviation of 0.72 in the current study. Below we will report the BF_{01} based on such an informative prior. A figure of sensitivity analysis that shows how the BF_{01} values change depending on different standard deviation values are available in the supplementary materials. If the result of a study was $BF_{01}= 5$, that means the null hypothesis (H0) was five times more probable than the alternative hypothesis (H1). We follow the guidelines from Dienes (2018), which suggests that when the Bayes Factor is greater than 3, it represents substantial evidence. All the statistical computations were conducted using JASP software version 0.9.2 (JASP team, 2021).

2.1.5 Results

2.1.5.1 Behavioral data

The overall accuracy rate for the paraphrase task was 92% (79%-100%, Baseline: 94%, Complement: 86%, High cloze: 95%, and Low cloze: 93%). Although the accuracy rate of the Complement condition was slightly lower (Baseline vs. Complement: $t(32)= -2.87$, $p <.05$; High- vs. Low-cloze: $t(32) = -0.83$, $p = 0.41$), the overall high accuracy rates suggested that participants were paying attention during the experiment.

2.1.5.2 ERP data

Figure 4 below presents the grand average ERPs to N400 effect of Predictability in the critical sentences (Baseline, Complement). Visual inspection suggests that the Complement condition elicited larger N400 amplitude than the Baseline condition. The results of the pairwise comparison show a significant effect ($t(32) = 2.09$, $p < 0.05$).

Figure 5 shows the grand average ERPs for the Cloze probability effect in the control items (High cloze vs. Low cloze). Visual inspection finds that the N400 response to the High cloze condition is reduced relative to the Low cloze condition. The results of the paired t-test show a significant effect ($t(32) = 2.21$, $p < 0.05$).

[Figures 4 and 5 around here]

2.1.6 Discussion

The Bag of Arguments hypothesis predicts that there is an early stage at which structural information about which noun phrases are arguments of the verb can constrain prediction of that verb, even when the thematic roles of those arguments do

not. In this experiment, we observed a larger N400 to the verb *fired* in the Complement condition than in the Baseline condition, even though both contexts contained the same noun phrases in the same linear position. These results suggest that comprehenders were able to use the structural information to determine that *fired* was a more predictable event when *millionaire* and *servant* were both participants than when *servant* was the only participant provided by the prior context, and to use this information in time to update predictions of the verb prior to the N400. The next question is whether argument role is available to impact predictions of verbs during the same time window. We will address this question in Experiment 2.

2.2 Experiment 2

In Experiment 2, we tested the second prediction of the Bag of Arguments hypothesis. To recap, the metaphor that the arguments are "lumped in a bag" is meant to express the hypothesis that there is a stage at which identifying the arguments of a verb could constrain prediction but the argument role information bound by the argument does not. In Experiment 1 we showed that at a presentation rate of 800 ms SOA, argumenthood did constrain the prediction of the verb in time to impact the N400 response. Therefore, in Experiment 2 we asked whether argument role information can also impact prediction of the verb with the same presentation rate. We chose to use this kind of between-subject design because it would have been difficult to generate a full set of 120 role-reversal sentences without repeating the target verbs and reducing the strength of the predictability manipulation. Critically, across Experiments 1 and 2, we tested the impact of argument identification and argument role with exactly the same timing and tightly matched experimental items. In Experiment 2 we kept the same items for the Baseline condition as Experiment 1. We kept the morpheme *ba* in the Baseline

condition, and reversed the order of the two arguments in the Reversal condition (*Millionaire ba servant fired* vs. *Servant ba millionaire fired*).

2.2.1 Participants

The participants were 37 naive young adults (13 male and 24 female, 18-31 years old, mean: 23) from National Taiwan Normal University. All of them were right-handed native Mandarin speakers, with no a history of neurological or psychiatric disorders. Of the 37 participants, 10 were excluded after pre-processing because of excessive eye blinks, muscle potentials, sweat artifact and alpha waves. The reported results were obtained from the remaining 27 participants (9 male and 18 female, 18-31 years old, mean: 23). Informed consent was obtained from all participants. The experiment protocol was approved by the Institutional Review Board Office at the University of Maryland College Park.

2.2.2 Materials

The experimental materials were 60 pairs of sentences comprising the two conditions: Baseline and (role) Reversal. We began with the same 60 Baseline sentences from Experiment 1. To create the role reversal sentences, we reversed the order of the two arguments, for example: Baseline condition (*Millionaire ba servant fired*, meaning the millionaire fired the servant) and Reversal condition (*Servant ba millionaire fired*, meaning the servant fired the millionaire) (See Table 2). Note that these 60 sentences were normed in Chow, Lau, Wang and Phillips (2018), and the cloze contrast was Baseline: 38% vs. Reversal: 0%. The 60 pairs of items were divided into two lists with latin square method. To check that participants did engage predictive mechanism during the experiment, we included the same 30 pairs of cloze items in

Experiment 1 as our control items in Experiment 2. The same 120 filler sentences used in Experiment 1 were included here as well.

Two lists were constructed such that no sentence context or target word was repeated in either list. Each list consisted of a total of 240 sentences, including 30 sentences in Baseline condition, 30 sentences in Reversal condition, 30 sentences of high-cloze target in High cloze condition and 30 sentences of low-cloze target in Low cloze condition, and 120 filler sentences. Participants were randomly assigned to one of the two lists.

Condition	Sentence	Post-target continuation
Baseline	富翁 把 僕人 **解雇了**	之後 立即 請來了 新的 管家
	Millionaire ba servant **fired**	then immediately hired new housekeeper
	"The millionaire had fired the servant and then immediately hired a new housekeeper."	
Reversal	僕人 把 富翁 **解雇了**	之後 立即 請來了 新的 管家
	Servant ba millionaire **fired**	then immediately hired new housekeeper
	"The servant had fired the millionaire and then immediately hired a new housekeeper."	
High cloze	駭客 忘掉了 **密碼**	，無法 執行 任務
	The hacker forgot **passwords**	, failed execute plan
	"The hacker forgot the passwords, so he failed to execute the plan."	
Low cloze	駭客 忘掉了 **登出**	，不小心 露出馬腳
	The hacker forgot **logout**	, accidentally gave the game away

	"The hacker forgot to log out, and that gave the game away."

Table 2: Example stimulus in each condition in Experiment 2.

2.2.3 Procedure

The procedure was identical to Experiment 1. As in Experiment 1, 20% of the sentences would be followed by a comprehension question.

2.2.4 Data acquisition and analysis

Data acquisition and analysis were identical to Experiment 1. The overall mean rejection rate across participants was 19.8 ± 10.3% (mean ± SD). Like Experiment 1, participants with rejection rate greater than 40% were excluded from further analysis. Rejection rates for each condition were summarized below: Baseline: 19.0 ± 11.8%; Reversal: 16.9 ± 8.6%; High cloze: 22.6 ± 14.5% and Low cloze: 20.9 ± 12.1%.

2.2.5 Results

2.2.5.1 Behavioral data

The overall accuracy rate to the paraphrase task was 90 % (75%-100%; Baseline: 92%, Reversal: 83%, High cloze: 94%, and Low cloze: 93%). Although the accuracy rate of the Reversal condition was slightly lower (Baseline vs. Reversal: $t(26)= -2.25$, $p < .05$; High- vs. Low-cloze: $t(26) = -0.33$, $p = 0.75$), the overall high accuracy rates suggested showing that participants were paying attention during the experiment.

2.2.5.2 ERP data

Figure 6 shows the grand average ERPs for the Predictability effect to Baseline and Reversal conditions. Visual inspection suggested that there was no N400 difference between the two conditions. The results of the paired t-test similarly showed no significant difference ($t(26) = 0.47$, $p = 0.64$). Bayes factor analysis yields a value of $BF_{01} = 3.47$, suggesting substantial evidence for the null hypothesis. For a sensitivity analysis of BF_{01} values as a function of different standard deviations of the normally distributed prior, please see the supplementary materials.

By contrast, Figure 7 shows the grand average ERPs to the High cloze and Low cloze conditions in the control items. Visual inspection showed that the N400 was reduced to the High cloze relative to the Low cloze condition. The results of the paired t-test showed a significant difference between conditions ($t(26) = 2.32$, $p < 0.05$).

[Figures 6, and 7 around here]

2.2.6 Discussion

The Bag of Arguments hypothesis predicts there should be a period of time in which identifying the arguments of a verb could exert an effect on prediction but not argument role information bound by the arguments. In Experiment 1 we had observed that with an 800 ms SOA presentation rate, comprehenders could tell if the noun phrases could be arguments of a verb. In Experiment 2, we tested if argument role information could impact prediction within the same time frame. In particular, given *millionaire-as-an-agent* and *servant-as-a-patient*, the predicted verb would be *fired*, but the role reversal scenario (i.e. *servant-as-an-agent* and *millionaire-as-a-patient*) would not predict the verb *fired*. Interestingly and in line with previous findings, the N400 was not sensitive to role reversal situations, as if the verb *fired* were a good fit of

event for a servant to act on a millionaire. As discussed in Section 1.1.2, the insensitivity of N400 to role reversal situations has been replicated in many languages with various verb final sentence structures (Kim & Osterhout, 2005; Kuperberg, Kreher, Sitnikova, Caplan, & Holcomb, 2007; Momma, Sakai, & Phillips, 2015; Chow & Phillips, 2013; Chow, Smith, Lau and Phillips, 2016). The null effect could not be attributed to lack of engaging predictive mechanism during the experiment, as we did observe an N400 effect to the cloze manipulation in our control items. A more likely explanation to the null effect of the role reversal situations, as suggested by Chow, Momma, Smith, Lau and Phillips (2016), is that it takes longer for prediction to be updated on the basis of argument role. For example, Momma, Sakai, and Phillips (2015) have found that the N400 effect emerged when the presentation rate was as slow as 1200 ms.

In sum, in Experiments 1 and 2, we tested the Bag of Arguments hypothesis, which suggested that there existed a time window where identifying the arguments of a verb could constrain prediction, but not argument roles bound by the argument. With Mandarin, we were able to manipulate whether noun phrases were arguments of a verb while keeping the linear distance between the noun phrases and the verb identical. Results from Experiments 1 and 2 allowed us to narrow down the time window to compute different levels of argument-verb relations. Specifically, given a slower presentation rate at 800 ms, the parser was able to identify noun phrases that were arguments of a verb, and to use that information to update predictions, but not argument roles.

2.3 Experiment 3

The goal of Experiment 3 was to identify if there is a lower time limit for arguments of a verb to be identified to constrain predictions. If there is a time window

at which the parser cannot tell if the noun phrases are arguments of a verb, such that only word associative effects are present (i.e. the Bag of Words hypothesis), then we should revise the two-stage model implied by the Bag of Arguments hypothesis into a three-stage model. We tested the same materials as in Experiment 1 (*Millionaire ba servant fired* vs. *Millionaire thought servant fired…*) with a faster presentation rate of 600 ms. Except for the presentation rate, other settings remained identical as Experiment 1.

2.3.1 Participants

The participants were 48 naive young adults (26 male and 22 female, 18-33 years old, mean: 23) from National Taiwan Normal University. All of them were right-handed native Mandarin speakers, with no a history of neurological or psychiatric disorders. Of the 48 participants, 10 were excluded after pre-processing because of excessive eye blinks, muscle potentials, sweat artifact and alpha waves. The reported results were obtained from the remaining 38 participants (18 male and 20 female, 18-33 years old, mean: 23). Informed consent was obtained from all participants. The experiment protocol was approved by the Institutional Review Board Office at the University of Maryland College Park.

2.3.2 Materials

The materials were identical to those in Experiment 1.

2.3.3 Procedure

The procedure was identical to Experiment 1, except for the presentation rate. The presentation rate was increased to 600 ms, with 500 ms stimulus duration and a 100 ms blank interval. See Figure 8 for details.

[Figure 8 around here]

2.3.4 Data acquisition and analysis

Data acquisition and analysis were identical to Experiment 1, except that we extracted epochs of length from 100 ms before to 600 ms after stimulus onset, which was identical to the time of the word presentation. The mean rejection rate across participants was 23.1 ± 12.7% (mean ± SD); participants with rejection rate greater than 40% were excluded from further analysis. The following were the rejection rates for each condition: Baseline: 22.0 ± 13.5%; Complement: 21.9 ± 12.7%; High cloze: 22.7 ± 13.3% and Low cloze: 22.2 ± 13.2%.

2.3.5 Results

2.3.5.1 Behavioral data

The overall accuracy rate for the paraphrase task was 95% (83%-100%; Baseline: 97%, Complement: 89%, High cloze: 96%, and Low cloze: 96%), Although the accuracy rate of the Complement condition was slightly lower (Baseline vs. Complement: $t(39)= -3.64$, $p <.05$; High- vs. Low-cloze: $t(39) = 0.00$, $p = 1$), the overall high accuracy rates suggested that participants were paying attention during the experiment.

2.3.5.2 ERP data

Figure 9 below is the grand average ERPs illustrating the N400 response in Baseline and Complement sentences. Visual inspection suggested that there was little N400 amplitude difference between the Complement condition and the Baseline condition. The results of the pairwise comparison showed no significant differences between conditions ($t(37) = 0.32$, $p = 0.75$). Bayes factor analysis yields a value of $BF_{01} = 4.33$, suggesting substantial evidence for the null hypothesis. For a sensitivity analysis of BF_{01} values as a function of different standard deviations of the normally distributed prior, please see the supplementary materials.

Figure 10 shows the grand average ERPs to High cloze and Low cloze for the control items. Visual inspection suggested that the N400 amplitude was reduced for the High cloze relative to the Low cloze ones. Paired t-test also confirmed the visual inspection ($t(37) = 2.52$, $p < 0.05$).

[Figures 9, and 10 around here]

2.3.6 Discussion

In Experiment 3, we aimed at investigating whether we could observe a lower time limit on the argumenthood effect we observed in Experiment 1, by using a slightly faster presentation rate (600 ms SOA). Prior studies have already reported the absence of argument role effects on the N400 at a 600 ms presentation rate (Chow & Phillips, 2013; Kuperberg, Kreher, Sitnikova, Caplan, & Holcomb, 2007). Here we also found no significant argumenthood effects at the 600 ms presentation rate. Whereas the same materials elicited an N400 difference between Complement and Baseline conditions with a slower presentation rate (800 ms) in Experiment 1, we found that this effect was

not significant with the faster presentation rate (600 ms). In other words, under time pressure, prediction of the verb was no longer constrained by arguments of a verb.

The absence of N400 effects of argumenthood in Experiment 3 suggests that a certain amount of time is required to identify whether a noun phrase is argument of a verb or not; if the time lapse is not long enough, then the parser cannot tell. When the presentation rate was increased to 600 ms, the N400 did not differ between the Complement and the Baseline conditions, suggesting that the two noun phrases in the Complement condition were parsed as if they were arguments of the verb just like in the Baseline condition. The patterns observed here are compatible with predictions from the Bag of Words hypothesis, which suggests that structure played a limited role in initial verb prediction; word associations were sufficient to account for the effects.

One alternative explanation for different results between Experiments 1 and 3 is that the 600 ms rate was simply too fast for processing the sentences in general. However, a 600 ms SOA is common in Mandarin ERP studies (e.g. Chow & Phillips, 2013, Li, Zhao, Zheng, & Yang (2015). More importantly, we still obtained an N400 effect of cloze contrast in our control items. This finding is crucial, because it shows that participants did engage predictive mechanisms during the experiment, even with the faster presentation rate.

3. General Discussion

In the current study, three ERP experiments were conducted to map the time course of argument-verb relation computations. We placed two noun phrases before a verb, and systematically evaluated the timing for different pieces of argument information to impact the prediction of a verb. Results from Experiments 1 and 2 showed that with the slower presentation rate at 800 ms, comprehenders were able to update predictions based on the argumenthood of the noun phrases, but prediction based

on argument roles was not yet effective. By contrast, when the presentation rate was increased to 600 ms per word in Experiment 3, comprehenders could no longer detect if the noun phrases could be arguments of an upcoming verb or not. Under time pressure, verb prediction was mainly based on nearby words.

Our work provides important support for the Bag of Arguments hypothesis (Chow, Smith, Lau & Phillips, 2016), which suggests that there exists a time window at which argument role information does not inform the prediction of an upcoming verb, but information about which noun phrases are in the same clause as the verb can. One important limitation of their initial experiments was that the argumenthood of a noun phrase with respect to the upcoming verb or not was confounded with linear distance between the noun phrases and the verb. By controlling linear distance between the noun phrases and the verb, we rule out this alternative explanation and provide further evidence in support for distinguishing argument identification and argument role computation at different temporal stages.

Our work also goes beyond Chow, Smith, Lau and Phillips (2016), as we were able to temporally dissociate the computation of different levels of argument-verb information. In particular, we suggest that there was a time window for the argument identification computation, during which the parser was able to identify if noun phrases could be arguments of an upcoming verb, and update predictions based on that, but not on the basis of argument role. In addition, on the lower end, we saw no evidence that the parser had identified if the noun phrase was an argument of the verb. Since we did see evidence of some kinds of predictions at this stage in the high/low cloze control conditions, we suggest that this time-window represents an early "Bag of Words" stage of verb prediction not constrained by structure at all; the mechanism at work here is simply word associations.

Chow, Smith, Lau and Phillips (2016) stimulated a lively public discussion about predictive mechanisms in argument structure computation. Kim, Oine and Sikos (2016) proposed that predictions could be modulated by event knowledge, on top of semantic associations. Kuperberg (2016) suggested an alternative explanation, where different cues could be weighted differently depending on the context, as these cues provide different sources of evidence about the meaning of specific event being conveyed. Below we will outline a processing model inspired by the Chow et al. (2016) approach and the current data, compare our model with other staged frameworks, and will then return to the question of how these alternative approaches might interpret these effects.

3.1 Toward a processing model of argument-verb relation computations

Based on the results of the three experiments and the findings from prior research (Momma, Sakai, & Phillips, 2015; Chow, Lau, Wang, & Phillips, 2018), we would like to propose an expanded processing model of computing argument-verb relations (see Figure 11). As depicted in Figure 11, our model suggests that there are three stages for different levels of argument information to be computed in argument-verb relation computations. At an early stage, initial verb prediction is based on word associations. The parser does not differentiate whether these noun phrases are arguments of an upcoming verb; the comprehension system simply probes memory for events that are associated with all the noun phrases (bag of words). For example, as *fire* is a plausible and likely event among all those events involving both *a millionaire* and *a servant*, then when under time pressure the system does not consider other cues in the context beyond the semantic relatedness between the noun phrases and the event described by the verb. Then, at an intermediate stage, the parser becomes more sensitive

to structural cues. The parser is able to identify whether noun phrases are arguments of the verb and use that information to update predictions (the Bag of Arguments hypothesis). It is only at a later stage that the parser starts to compute argument role information (e.g. servant-as-an-agent and millionaire-as-a-patient) and construct the full structure of the sentence.

[Figure 11 around here]

Note that although our model suggests that readers do not commit to an argument role initially, it is fully compatible with the possibility that argument role information can be computed before the presence of a verb. Such a perspective is in line with Kim, Oines and Sikos (2016) and Chow, Momma, Smith, Lau and Phillips (2016). To be clear, we suggest that some information about the arguments can be computed more quickly than others. Before argument role relations are established, the parser has identified whether the noun phrases are arguments of the verb.

Along with Chow, Momma, Smith, Lau and Phillips (2016), we believe that while clause boundaries could be a useful cue to constrain predictions of an upcoming verb, our own data currently do not speak to whether and how verb predictions are affected by arguments outside the clause boundary. We propose that at this stage, comprehenders are sensitive to clause boundaries and they can identify which noun phrases are the arguments. It is likely that they could also use other information to inform their prediction (including noun phrases that are in another clause or the larger discourse context).

3.1.1 Comparing the current model with other staged frameworks

Since we assume a staged framework in describing our results, we would like to discuss how our proposal is different from other staged frameworks. To begin with, our model suggests that initial verb prediction is mainly driven by semantic associations (i.e., "bag of words"), and structural information exerts its influence at a later stage. At first glance, this "semantic-first" proposal seems not to be in line with many staged frameworks which advocate the "syntax-first" proposal. The syntax-first proposal was motivated by findings from reading structural ambiguous sentences (Frazier & Clifton, 1997; Frazier & Rayner, 1982; Rayner, Carlson, & Frazier, 1983). This line of work suggests that the parser initially builds a simplistic structure based on syntactic category information, which is autonomous and independent from lexical semantic information. It is at a later stage that semantic features, thematic relations and other contextual information are taken into consideration. Although the exact details differ, the neurocognitive model of sentence comprehension (Friederici, 2002) and the extended argument dependency model (the eADM, Bornkessel & Schlesewsky, 2006) generally endorse this view. That is, local phrase structure building, which relies heavily on the syntactic category of words, precedes the processing of other types of information during online sentence comprehension.

While it seems that our model is at odds with proposals from other stage-based frameworks, it should be noted that using argument information to predict the verb and using verb information to predict its arguments involve very different processes (Friederici & Frisch, 2000). In fact, Bornkessel and Schlesewsky (2006) have incorporated such differences in their eADM model. It is also essential to highlight that syntax-first proposal was mainly motivated by studies that investigated structurally ambiguous sentences, whose disambiguating regions usually came after critical verbs. As discussed in the Introduction section, when a verb is encountered, information about

its argument structure can be accessed to constrain predictions of upcoming words immediately. It would then not be too surprising to observe a dominating role of syntax early on. By contrast, in the current study, we focus on how comprehenders use pre-verbal argument information to predict an upcoming verb. Although we reported a strong effect of word associations for initial prediction of a verb, we shared a similar assumption with other staged frameworks. That is, we assumed that comprehenders had to access the syntactic category of the words first, in order to further evaluate the relations of the arguments (semantic features and syntactic structures alike). Since the scope of our model and earlier frameworks are different, our interpretations and arguments are not necessarily in conflict.

3.1.2 Staged framework or strength-of-evidence framework for argument-verb computations?

From the beginning of this paper we have chosen to frame our logic and discussion in terms of a staged framework of processing—e.g. there is an initial "stage" at which semantic association cues are primary in predicting the verb. In line with Chow, Smith, Lau and Phillips (2016), we suggest that identifying whether noun phrases are arguments of a verb is a prerequisite of argument role assignment, and one way of stating this claim is that the earliest, "semantic association" stage is followed by a subsequent stage at which argumenthood information contributes to prediction, and only at a later stage does role information exerts its effect. Within such a framework, one can still straightforwardly accommodate strong effects of broader discourse context on processing: the context can just be taken to impact which possibilities are weighted more strongly at each stage, and/or the speed at which a comprehender transitions from one stage to the next. In other words, the weighting of candidates or the temporal scale

could be modulated by different variables (such as discourse contexts, experiment tasks, and presentation methods); a staged framework only assumes that the *order* of the stages continues to hold.

However, these data will be understood differently within a strength-of-evidence framework, as suggested by Kuperberg (2016), which is not committed to this last assumption. Under a strength-of-evidence framework, different types of information (e.g., semantic and syntactic cues) provide evidence that is differentially reliable about the underlying event and event structure. In other words, multiple sources of evidence from the contexts are evaluated in parallel and in combination for comprehenders to infer the event being conveyed. Therefore, within this strength-of-evidence framework, the reason why a bag-of-words mechanism has such a rapid effect is that certain combinations of arguments provide very reliable evidence about the specific event being conveyed, and this evidence overrides other cues. To be more specific, the reason why the combination of arguments, "servant-millionaire-fired", has such a rapid influence on comprehension is that these words provide strong evidence that the communicator is describing a canonical event, stored within long-term memory, where millionaires are more likely to fire servants than vice versa, and thus reliability of this evidence is stronger than the syntactic evidence. If there are other cues in the context that provide stronger evidence for an alternative event, it would be possible to override this highly reliable "bag of words" cues. This explains why in such a framework, there is not a fixed order for sentence comprehension in real time.

The findings in the current study could not distinguish the two types of frameworks. Future work can evaluate predictions of the two frameworks by manipulating different types of cues (e.g., discourse contexts and focus). For example, if we set up a context like after a revolution, servants gain all the power to fire

millionaires, then the reversal condition (*Servant ba millionaire fired*) would be more plausible than the baseline (*Millionaire ba servant fired*) in Experiment 3. Similarly, the complement condition (*Millionaire thought the servant fired*…) would be more plausible than the baseline (*Millionaire ba servant fired*) in Experiments 1 and 2. Under the staged framework, one would expect comprehenders to go through the same stages as reported in the current study—semantic association followed by argument identification and finally argument role assignment, although the temporal scale could be shifted and the N400 patterns between conditions would be reversed as the global discourse context has led to the opposite prediction. By contrast, under the strength-of-evidence framework, given the revolutionary context, the inferred event that servants could fire millionaires could be much more likely. Therefore, when reading a sentence concerning what servants could do to millionaires after a revolution, comprehenders may have already generated strong predictions for the semantic features of *fire* before encountering the verb. Such a facilitation was not there when the roles of servants and millionaires were flipped. If so, the N400 would show sensitivity to the manipulations at an early stage. The finding then would indicate that the cue from discourse context is so reliable that it overrides influences from other cues.

3.1.3 Slow parsing or slow prediction?

As our model suggests that some information about arguments can be computed more quickly than others, the next question to be addressed would be whether it is parsing itself that is slow, or just the updating of the predictions. Before further discussing this question, we would like to reiterate what we mean by "prediction" in the current study. Here we take "prediction" as an umbrella term, and do not necessarily differentiate it from "priming." This view is very different from researchers that use

"priming" and "prediction" as labels for distinctive processes. For example, for Pickering and Gambi (2018), "priming" concerns simple semantic associations whereas "prediction" is about contextual effects during comprehension. In Brouwer and Hoeks (2012), who also studied thematic role reversal sentences, the absence of N400 effects is attributed to "priming" effects on lexical access. For the current discussion, however, we will subsume all these effects under the umbrella term of prediction.

We consider this model to illustrate the processing profile by which different levels of argument-verb information is computed to feed prediction, and here we've chosen to pursue the implication that parsing is slow; the parser is only able to compute sophisticated structural information when more time is granted. An alternative to the slow parsing view of these phenomena is a slow prediction view, which holds that computing the relations of an argument and its argument role is not taxing; what slows down prediction is the memory search process to retrieve the best fit of the context (Chow, Momma, Smith, Lau & Phillips 2016). Under the slow prediction view, it would not be too challenging to compute *millionaire-as-a-patient* and *servant-as-an-agent*; what slows down prediction is to search for an event that involves them. Momma, Sakai and Phillips (2015) provide one argument in favor of the slow prediction account, examining ERP responses to pre-verbal arguments, coupled with different case markers, such as *bee-accusative* vs. *bee-nominative*. They found that the N400 amplitude is larger in arguments with an accusative case relative to a nominative case, and interpret the patterns as showing that the relation between an argument and its argument role could be established very early. However, this N400 effect could also reflect other kinds of lexical processing differences between different case markers (*-accusative* vs. *-nominative*). Therefore, we think the existing evidence is neutral on whether the observed delays reflect slow prediction or slow parsing, and thus for now

we prefer to couch the current model in terms of slow parsing. However, we do not have direct evidence from the current study to argue for or against either of these views. This will be an interesting avenue of future work.

3.2 Reconciling these results with prior work

3.2.1 N400

In the current study, we propose a staged model of how different levels of argument information are integrated to feed the prediction of a verb. However, we suggest that its temporal course could be flexible. That is, while comprehenders might go through the same stages of computations, under different parameters, different levels of argument-verb computation could be facilitated, and the timing to capture an N400 effect could vary. Below we review some role reversal studies that have reported an N400 effect, and discuss possible parameters that have facilitated argument-verb computation.

To begin with, as the current model was based on data in Mandarin, we would like to draw attention to Bornkessel-Schlesewsky et al. (2011), where the authors report an N400 effect in Mandarin role reversal manipulations. In addition to modality differences, as their experiment was conducted aurally, it should be noted that they only found an N400 effect in passive *bei* constructions in Mandarin, not in *ba* constructions. Both constructions introduce two preverbal arguments (*ba*: SOV structure; *bei*: OSV structure), but according to Bornkessel-Schlesewsky et al. (2011), only *ba* construction involves structural ambiguity at the verb. The parser might not consider the verb anomalous as it permits a continuation as a relative clause (see Example 1), so the N400 effect is absent at the verb in *ba* constructions. However, we do not find such an interpretation very convincing, as in fact both *ba* and *bei* constructions could take a

relative clause continuation after the verb (see Examples 1 and 2). The absence of N400 effect could not be attributed to the potential structure ambiguity in *ba* constructions. In fact, we believe that the N400 effect in *bei* constructions is more likely to have resulted from a language-specific pragmatic principle in Mandarin. Specifically, Mandarin passive *bei* involves a negative connotation. The patient of a passive *bei* sentence always bore a negative consequence of an event, which is reflected as a bigger N400 as early as the presence of the second argument (Philipp, Bornkessel-Schlesewsky, Bisang, & Schlesewsky, 2008). What this means is that the pragmatic cue encoded in the passive marker *bei* could facilitate the computation of verb-argument relation, such that the parser was able to detect the role reversal situation more quickly. In the future, we could investigate if the "negative" implication of *bei* is a different kind of information than thematic role information.

(1) 偵探 把 [[子彈 擊中 的] 罐頭] 拿走了。

Detective ba [[bullet hit de] tin] take-away

"The detective took away the tin which the bullet hit."

(2) 偵探 被 [[子彈 保存 的] 方法] 嚇到了。

Detective bei [[bullet kept de] way] shock

"The detective was shocked by the way which kept the bullet."

In addition to the Mandarin experiment discussed above, Bornkessel-Schlesewsky et al. (2011) also found an N400 effect in role reversal materials in Turkish. Both experiments were conducted aurally, in contrast with most other role-reversal studies in the literature. Although all the studies time lock their ERPs to the

onset of their target word, auditory presentation provides phonological cues, such as coarticulation (and tone sandhi in Mandarin), which are not available in visual presentation. In addition, spoken words unfold in time whereas with visual presentation, the whole word appears at the same time. In our opinion, the impact from lower-level phonetic cues on argument-verb computations might not be significant, but with different durations of the arguments and the verb, cross-modality comparison does not seem very feasible. It is possible that in natural listening/reading, argument-verb relations could be computed faster than in an RSVP paradigm. We will return to the comparison between natural presentation and RSVP at the end of this section.

Bourguignon, Drury, Valois and Steinhauer (2012) show that verb types could modulate the N400 effect in role reversal situations, at least in English. The authors on one hand replicate Kuperberg, Kreher, Sitnikova, Caplan and Holcomb (2007), showing an absent N400 effect of role reversal with action verbs ("The boys have eaten" vs. "The fries have eaten"); on the other hand, they examine role reversal with psych-verbs, and did obtain an N400 effect at the verb ("The judges have despised" vs. "The movies have despised"). It is possible that the contrast between the sentient and the nonsentient entities is psychologically salient, such that given a subject that is nonsentient, the verb is less likely to be a psych verb. By contrast, for the action verbs, the finer distinction (e.g. edible vs. not edible) is not immediately available to the comprehenders; it is not a major division in how comprehenders immediately see the world. Either way, this intriguing data point suggests a future direction to examine the broader question of how verb types interact with argument features identified in the model, such as argument identification and argument roles.

3.2.2 P600

Previous studies generally report a P600 effect, instead of an N400 effect, in role reversal sentences (Chow & Phillips, 2013; Chow, Lau, Wang, & Phillips, 2018; Hoeks, Stowe, & Doedens, 2004; Kolk, Chwilla, Van Herten, & Oor, 2003; Kuperberg, Sitnikova, Caplan, & Holcomb, 2003; Kuperberg, Kreher, Sitnikova, Caplan, & Holcomb, 2007; Kim & Osterhout, 2005). We provide the grand averaged waveforms of all the electrodes from Experiments 1 to 3 in the supplementary materials. Although our materials were adapted from Chow, Lau, Wang, and Phillips (2018), who also found a P600 effect, we do not observe a tendency towards the presence of a P600 effect among our three experiments. We suspect that these differences result from differences in the tasks used (Brouwer, Fitz, & Hoeks, 2012). In particular, participants performed a plausibility judgment task at every sentence in Chow, Lau, Wang and Phillips (2018) whereas in the current study, participants had to do a paraphrase judgment on just 20% of the sentences. Although the accuracy rate of the anomalous conditions across the three conditions was slightly lower than the baseline conditions (Anomalous vs. Baseline, Experiment 1: 86% vs. 94%; Experiment 2: 83% vs. 92%; Experiment 3: 89% vs. 98%), they were always above 80%, showing that the participants did process the experiment materials fully. Therefore, despite the fact that neither an N400 nor a P600 effect was observed in Experiments 2 and 3, it seems rather unlikely that our participants did not detect the anomalies.

A related question is whether P600 component overlap with the N400 could have held differentially across different SOAs, perhaps contributing to the reduced N400 in Experiment 3. Unfortunately, it was not possible to evaluate this possibility in the current dataset because the post-600 ms time-window in Experiment 3 covered the presentation of post-target stimuli that differed significantly on both visual and linguistic dimensions: the baseline condition was often continued with the presentation

of a comma alone (unremarkable in RSVP for Mandarin characters) because the clause was finished, while the complement condition always contained a full post-target word to continue the clause. However, as we never saw any positive evidence in the ERPs for a P600 effect, we have no reason to think that component overlap drove the N400 modulation. As discussed in the paragraph above, the fact that we did not use an acceptability judgment task, which is known to increase the likelihood and amplitude of P600 effects, also makes this possibility less likely.

3.2.3 Making generalizations about sentence comprehension in natural contexts

A final critical question is whether we can make a generalization about the dynamics of sentence processing computation in natural contexts based on the use of the seemingly artificial RSVP paradigm. We used two presentation rates in the current study: 800 ms/word (75 words/minute) in Experiments 1 and 2 and 600 ms/word (100 words/minute) in Experiment 3. According to Brysbaert's (2019) meta analysis of reading rates across different languages, fluent Mandarin readers are estimated to read 260 words per minute in silent reading. Note that the value was computed based on a 1.5 characters for 1 word ratio. In the current study, each word had an average of 2.2 characters (range 1-4), the equivalent natural reading speed in Brysbaert (2019) would be 177 words/minute. Therefore, the stimulus presentation rates in the current study were slower than natural reading, and this raises the question of whether comprehenders would ever use anything other than "bag-of-words" prediction in natural reading.

We think this is an interesting and important question. One possibility is that, indeed, comprehenders just rarely benefit in real life from the kind of processing facilitation indexed by the structure-informed N400 effect at slower SOAs; predictions based on pre-verbal argument information could be infrequent in natural reading. On

that interpretation, our manipulation may be informative about the underlying structure of the language processing system, but it is less informative about real-life prediction in language comprehension. However, we note that the reading experiences of participants in eye-tracking and ERP studies are extremely different by nature. In normal reading, the reading rate is controlled by readers and it can be varied. Readers gain parafoveal preview information, skip words, or regress on prior texts, and none of these are possible under the RSVP paradigm in ERP (Wlotko & Federmeier, 2016). Therefore, it is likely that the temporal scale of our model could be shifted in natural reading. Although it remains an empirical question, it seems possible to us that in natural reading, the time scale of these stages might well be shorter than in RSVP.

Despite being unnatural, the RSVP paradigm is still commonly used in EEG studies, because it allows researchers to fully control the timing of stimulus presentation and to time-lock comprehenders' brain response to specific pieces of information. There are some attempts to co-register EEG with methodologies that present stimuli more naturally. For example, Ditman, Holcomb and Kuperberg (2007) used simultaneous self-paced reading and EEG to study the processing profiles of sentences containing pragmatic and morphsyntactic violations. While participants read the stimuli at their own pace, Ditman and her colleagues (2007) were able to replicate the findings reported in Kuperberg, Caplan, Sitnikova, Eddy, and Holcomb, (2006), where the stimuli were presented in RSVP. Other studies that co-registered eye-tracking and EEG generally showed a robust N400 predictability effect from fixation-related brain potentials on target words (Dimigen, Sommer, Hohlfeld, Jacobs, & Kliegl, 2011; Kretzschmar, Schlesewsky, & Staub, 2015). Taken together, results from existing coregistration studies are generally compatible with findings reported from RSVP

paradigms. For these reasons, we are hopeful that the patterns observed in the current study can be generalized to sentence comprehension in natural contexts.

4. Conclusion

Based on the results of prior studies and our three experiments, we have proposed a model of the processing profile of argument-verb relation computation. At an initial stage, the system does not differentiate the noun phrases by structural position, and only simple word association effects are observed at the verb. At a second stage, contextual facilitation is now sensitive to whether the noun phrases are arguments of the upcoming verb, but not to their thematic role (the Bag of Arguments hypothesis). It is only at a later stage that the parser starts to consider argument roles in computing argument-verb relations. Our model thus delineates the stages for the context-based mechanisms that support online sentence comprehension.

Figure Captions

Figure 1: Visual illustrations for stimuli in Chow, Smith et al (2016). Dotted line indicates semantic associations and color shading shows argument positions of the embedded verb.

Figure 2: Visual illustrations for stimuli in the current study. Dotted line indicates semantic associations and color shading shows argument positions of the embedded verb.

Figure 3: Presentation of stimuli in Experiment 1.

Figure 4: Grand average ERPs to predictability effect of Baseline and Complement at Cz and the topographic distribution of ERP effects in the 300-500 ms interval in Experiment 1 (Complement minus Baseline).

Figure 5: Left: Grand average ERPs to cloze control items at Cz in Experiment 1. Right: The topographic distribution of ERP effects in the 300-500 ms interval in Experiment 1 (Low minus High cloze).

Figure 6: Grand average ERPs to predictability effect of Baseline and Reversal at Cz and the topographic distribution of ERP effects in the 300-500 ms interval in Experiment 2 (Reversal minus Baseline).

Figure 7: Left: Grand average ERPs to cloze control items at Cz in Experiment 2. Right: The topographic distribution of ERP effects in the 300-500 ms interval in Experiment 2 (Low minus High cloze).

Figure 8: Presentation of stimuli in Experiment 3.

Figure 9: Grand average ERPs to predictability effect of Baseline and Complement at Cz and the topographic distribution of ERP effects in the 300-500 ms interval in Experiment 3 (Complement minus Baseline).

Figure 10: Left: Grand average ERPs to cloze control items at Cz in Experiment 3. Right: The topographic distribution of ERP effects in the 300-500 ms interval in

Experiment 3 (Low minus High cloze).

Figure 11: The three-stage processing model of argument-verb computations.

Acknowledgment

We would like to thank Alexander Williams and Colin Phillips for helpful discussions and Shiao-Hui Chan, Shih-Chiang Hu and Aymeric Collart for the support for EEG data collection in Taiwan.

Funding

This research was supported by the William Orr Dingwall Dissertation Fellowship and by the Ministry of Science and Technology of Taiwan grant MOST 110-2410-H-007-095-MY2 to Chia-Hsuan Liao and by the National Science Foundation grant BCS-1749407 to Ellen Lau.

Reference

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

Bornkessel-Schlesewsky, I., Kretzschmar, F., Tune, S., Wang, L., Genç, S., Philipp, M., Roehm, D., & Schlesewsky, M. (2011). Think globally: Cross-linguistic variation in electrophysiological activity during sentence comprehension. *Brain and language*, *117*(3), 133-152.

Bornkessel, I., & Schlesewsky, M. (2006). The extended argument dependency model: a neurocognitive approach to sentence comprehension across languages. *Psychological review*, *113*(4), 787.

Bourguignon, N., Drury, J. E., Valois, D., & Steinhauer, K. (2012). Decomposing animacy reversals between agents and experiencers: an ERP study. *Brain and language, 122*(3), 179-189.

Brouwer, H., Fitz, H., & Hoeks, J. (2012). Getting real about semantic illusions: rethinking the functional role of the P600 in language comprehension. *Brain research, 1446*, 127-143.

Brysbaert, M. (2019). How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of memory and language, 109*, 104047.

Camblin, C. C., Ledoux, K., Boudewyn, M., Gordon, P. C., & Swaab, T. Y. (2007). Processing new and repeated names: Effects of coreference on repetition priming with speech and fast RSVP. *Brain Research, 1146*, 172-184.

Wing Yee Chow (2013). The Temporal Dimension of Linguistic Prediction. PhD Dissertation. University of Maryland.

Chow, W. Y., & Phillips, C. (2013). No semantic illusions in the "Semantic P600" phenomenon: ERP evidence from Mandarin Chinese. *Brain research, 1506*, 76-93.

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), 1-26.

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.

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

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.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in psychology, 5*, 781.

Dimigen, O., Sommer, W., Hohlfeld, A., Jacobs, A. M., & Kliegl, R. (2011). Coregistration of eye movements and EEG in natural reading: analyses and review. *Journal of experimental psychology: General, 140*(4), 552.

Ditman, T., Holcomb, P. J., & Kuperberg, G. R. (2007). An investigation of concurrent ERP and self-paced reading methodologies. *Psychophysiology, 44*(6), 927-935.

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, *44*(4), 491-505.

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.

Frazier, L., & Clifton, C. (1997). Construal: Overview, motivation, and some new evidence. *Journal of Psycholinguistic Research, 26*(3), 277-295.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive psychology*, *14*(2), 178-210.

Friederici, A. D., & Frisch, S. (2000). Verb argument structure processing: The role of verb-specific and argument-specific information. *Journal of Memory and language*, *43*(3), 476-507.

Friederici, A. D., & Weissenborn, J. (2007). Mapping sentence form onto meaning: The syntax–semantic interface. *Brain research*, *1146*, 50-58.

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.

JASP Team (2021). JASP (Version 0.9.2) [Computer software].

Kamide, Y., Altmann, G. T., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, *49*(1), 133-156.

Kim, A. E., Oines, L. D., & 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.

Kolk, H. H., Chwilla, D. J., Van Herten, M., & Oor, P. J. (2003). Structure and limited capacity in verbal working memory: A study with event-related potentials. *Brain and language, 85*(1), 1-36

Kos, M., Vosse, T. G., Van Den Brink, D., & Hagoort, P. (2010). About edible restaurants: conflicts between syntax and semantics as revealed by ERPs. *Frontiers in psychology*, *1*, 222.

Kretzschmar, F., Schlesewsky, M., & Staub, A. (2015). Dissociating word frequency and predictability effects in reading: Evidence from coregistration of eye movements and EEG. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(6), 1648.

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

Kuperberg, G. R. (2016). Separate streams or probabilistic inference? What the N400 can tell us about the comprehension of events. *Language, cognition and neuroscience, 31*(5), 602-616.

Kuperberg, G. R., Caplan, D., Sitnikova, T., Eddy, M., & Holcomb, P. J. (2006). Neural correlates of processing syntactic, semantic, and thematic relationships in sentences. *Language and cognitive processes, 21*(5), 489-530.

Kuperberg, G. R., Kreher, D. A., Sitnikova, T., Caplan, D. N., & Holcomb, P. J. (2007). The role of animacy and thematic relationships in processing active English sentences: Evidence from event-related potentials. *Brain and Language, 100*(3), 223-237.

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.

Kuperberg G. R., Wlotko E., Riley S., Zeitlin M., & Cunha-Lima ML. (2016) "The brain dissociates between different levels of prediction during language comprehension." *Psychonomic Society's 57th annual meeting*.

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

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

Laszlo, S., & Federmeier, K. D. (2009). A beautiful day in the neighborhood: An event-related potential study of lexical relationships and prediction in context. *Journal of Memory and Language, 61*(3), 326-338.

Lau, E., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics:(de) constructing the N400. *Nature Reviews Neuroscience, 9*(12), 920.

Li, X. Q., Zhao, H. Y., Zheng, Y. Y., & Yang, Y. F. (2015). Two-stage interaction between word order and noun animacy during online thematic processing of sentences in Mandarin Chinese. *Language, Cognition and Neuroscience, 30*(5), 555-573.

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.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). Lexical nature of syntactic ambiguity resolution. *Psychological review, 101*(4), 676.

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language, 38*(3), 283-312.

Momma, S., Sakai, H., & Phillips, C. (2015, March). *Give me several hundred more milliseconds: the temporal dynamics of verb prediction.* Paper presented at the 28th annual CUNY Conference on Human Sentence Processing, Los Angeles, CA.

Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin, 144*(10), 1002.

Philipp, M., Bornkessel-Schlesewsky, I., Bisang, W., & Schlesewsky, M. (2008). The role of animacy in the real time comprehension of Mandarin Chinese: Evidence from auditory event-related brain potentials. *Brain and Language, 105*(2), 112-133.

Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Bulletin, 30*(4), 415-433.

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.

Wang, L., Wlotko, E., Alexander, E., Schoot, L., Kim, M., Warnke, L., & Kuperberg, G. R. (2020). Neural Evidence for the Prediction of Animacy Features during Language Comprehension: Evidence from MEG and EEG Representational Similarity Analysis. *Journal of Neuroscience, 40*(16), 3278–3291.

Wlotko, E. W., & Federmeier, K. D. (2015). Time for prediction? The effect of presentation rate on predictive sentence comprehension during word-by-word reading. *Cortex, 68*, 20-32.