Enough time to get results? An ERP investigation of prediction with complex events

Chia-Hsuan Liao, Ellen Lau

Department of Linguistics, University of Maryland College Park, Maryland, USA

Contact: Chia-Hsuan Liao, cliao@umd.edu

Enough time to get results? An ERP investigation of prediction with complex events

How quickly can verb-argument relations be computed to impact predictions of a subsequent argument? We take advantage of the substantial differences in verb-argument structure provided by Mandarin, whose compound verbs encode complex event relations, such as resultatives (*Kid bit-broke lip*: a kid bit his lip such that it broke) and coordinates (*Store owner hit-scolded employee*: a store owner hit and scolded an employee). We tested sentences in which the object noun could be predicted on the basis of the preceding compound verb, and used N400 responses to the noun to index successful prediction. By varying the delay between verb and noun, we show that prediction is delayed in the resultative context (*broken-BY-biting*) relative to the coordinate one (*hitting-AND-scolding*). These results present a first step towards temporally dissociating the fine-grained subcomputations required to parse and interpret verb-argument relations.

Keywords: Sentence processing, Prediction, Argument structure, N400

Introduction

Language comprehension can be conceived as a bottom-up or reactive process in which the processing of each input word is tied to its position in the input string, or it can be conceived as a top-down or predictive process in which expectations about the linguistic form or message are generated early and updated as new input arrives. In recent decades, much work has argued that comprehension involves some degree of prediction. Listeners can respond to another interlocutor immediately in conversations, and sometimes can even fill in a particular word that the other person fails to produce (Schegloff, 2000). Behavioral and ERP work has shown that predictive sentence contexts have a robust facilitatory influence on the processing of the subsequent word (see Kutas, Delong & Smith, 2011; Van Petten & Luka, 2012, for review), and eye tracking work has been able to demonstrate such effects prior to the critical word; for example, given a scene of a man, a girl, a motorcycle, and a carousel, and presented with the sentence frame "the man likes to ride ____," participants tend to look at the picture

of a motorcycle, whereas given the context “the girl likes to ride ____”, participants tend to look more at the picture of the carousel (Kamide, Altmann and Haywood, 2003). These examples and others have been taken to suggest that comprehenders quickly integrate information from the context to predict what is coming next. These examples and others have been taken to suggest that comprehenders quickly integrate information from the context to predict what is coming next.

Recent studies have noted that predictions are not always fast, and that in fact, the speed with which predictions are updated can be used as a chronometer for determining the time course of the linguistic and interpretive computations required to do so (Federmeier, Kutas, & Schul, 2010; Chow et al., 2018; Chow et al., 2016b; Momma et al., 2015). Recent work by Chow and colleagues has focused on argument structure computation, using an ERP approach to estimate how long it takes to compute noun argument roles and update predictions about upcoming verbs accordingly (at least ~1800ms or so, based on these results discussed further below). These kinds of data provide an initial framework for developing a detailed timecourse model of top-down interpretation processes, and give important insights for longstanding debates about how interpretive errors arise in 'role-reversed’ contexts (e.g. *the waitress that the customer served*).

The current study adopts this approach to investigate a different but equally important component of argument structure computation that has received much less attention in the psycholinguistic literature—that is, the mechanisms by which complex verb relations are computed online. For example, Mandarin Chinese has a highly productive system of compound verbs—such as coordinate verbs (X hit-scolded Y, meaning X hit and scolded Y) and resultative verbs (X bit-broke Y, meaning X bit Y and in doing so caused Y to break)—which require mechanisms to combine the verbs into a single complex predicate denoting a

complex event, and to derive the corresponding set of argument roles. Another body of work has asked about how comprehenders deal with simple verb-object structures that require more complex semantics such as coercion (Kuperberg et al., 2010), or light-verb constructions (Wittenberg et al., 2014). However, to our knowledge, relatively little is known thus far about the processing algorithms by which complex verb structures are interpreted, even though they are pervasive in many languages. Here we take a preliminary step towards disentangling the fine-grained linguistic and conceptual subcomputations that are likely to be required, by comparing the speed of prediction update associated with coordinate compounds and resultative compounds in Mandarin. If each individual verb of the compound predicts a very different object than the compound as a whole, how long does that update take, and does it depend on the nature of the meaning relation? While many studies will be needed to develop a full model of these highly complex processes, we hope to show that this new method provides a way to successfully dissociate some of them experimentally.

Predictive mechanisms in sentence processing

The amplitude of the N400 response in ERP has frequently been used to track the prediction of lexical and conceptual material (Kutas & Federmeier, 2000; Lau et al., 2008). The N400 component peaks between 300-600 ms after the onset of the stimulus presentation, and is negatively correlated with the predictability of a target word (Kutas & Hillyard, 1984). Predictability is usually operationalized in these experiments by the construct of cloze probability, which is the percentage of responses that a given word occurred in a separate offline completion task (Taylor, 1953). For example, given a sentence context like "He was afraid that doing drugs would damage his ____," a majority of participants in the offline norming might complete the sentence with "brain" and a minority with "reputation," and the predicted high-cloze continuation "brain" would then elicit a significantly smaller N400 than

the less predicted low-cloze "reputation" (Thornhill & Van Petten, 2012). These kinds of effects have been frequently replicated (Federmeier & Kutas, 1999; Federmeier et al., 2007; Thornhill & Van Petten, 2012). Many authors have taken these results to indicate that linguistic input is predicted in context (although it is worth noting that a non-predictive explanation is possible in which these effects are due to variations in integration difficulty after the bottom-up input is encountered).

Whether N400 reductions reflect conceptual pre-activation, lexical pre-activation, or both, is still an open question. Consistent with a conceptual component, N400 responses are observed for meaningful pictures and environmental sounds as well as spoken and written words, and N400 modulations are observed when sentences are completed with pictures (Kutas & Federmeier, 2011). Federmeier and Kutas (1999) observed N400 reduction for unexpected completions that were semantically related to the expected completion ('They wanted to make the hotel look more like a tropical resort, so along the driveway they planted rows of palms/pines/tulips'), consistent with the idea that the conceptual features themselves were pre-activated by the context, although other accounts are also possible (see Thornhill & Van Petten, 2012, for similar results). Consistent with a lexical pre-activation component, work by Laszlo and Federmeier (2009) showed N400 sensitivity to unexpected words that are orthographically related to the expected ending, and Brothers, Swaab, and Traxler (2015) showed that on a trial-by-trial basis, N400s were reduced earlier and to a greater extent for words that participants had specifically predicted than words that were simply contextually supported. Together, we take these different lines of work to suggest that N400 effects reflect a combination of pre-activating conceptual features and pre-activating specific lexical items.

'Slow prediction' and argument structure

Although prior work has shown that comprehenders use contextual information to predict specific lexical forms, recent studies have argued that predictions are not always fast and

accurate (Chow et al., 2018; Chow et al., 2016b; Momma et al., 2015). These studies were investigating a longstanding puzzle in the literature about why reversing the thematic roles of noun phrases in a sentence usually does not modulate N400 amplitude. For example, Chow et al. (2018) tested Mandarin sentences such as *Cop ba thief arrest* "the cop arrested the thief" and the 'role-reversed' *Thief ba cop arrest* "the thief arrested the cop". Although the cloze probability in the canonical sentences was much higher than that of the role-reversal sentences, there was no N400 difference between the two conditions. This insensitivity of the N400 to differences in predictability caused by argument role reversals has been observed many times across many languages, and numerous hypotheses have been proposed to account for it (Hoeks et al., 2004; Kim & Osterhout, 2005; Kuperberg et al., 2007; Brouwer, Fitz, & Hoeks, 2012).

The work by Chow and colleagues argues that the explanation for the role-reversal results at least partly depends on how *quickly* predictions can be generated on the basis of the context. Chow et al. (2018) manipulated the linear distance between an argument and a verb, via varying the position of an adverbial temporal phrase in a sentence. In the short-distance conditions, the adverbial temporal phrase was placed at the beginning of the sentence, and the argument was immediately followed by the verb (*Last week, cop ba thief arrest*); in the long-distance condition, the adverbial phrase was placed between the argument and the verb (*Cop ba thief last week arrest)*, which created a buffer (around 1800 ms) to formulate the prediction of an upcoming verb. The results showed that there was no N400 effect in short-distance conditions, but critically the N400 response was recovered in long-distance conditions. Momma et al. (2015) report similar findings in Japanese. Together, these data argue that argument roles can be used to predict an upcoming verb if sufficient time is provided; the corollary implication is that not all information in the context impacts prediction immediately. Chow et al. (2016a) discuss several reasons that argument roles

might impact predictive computations slower than other kinds of information: (1) in the absence of the verb, argument roles like agent and patient are not directly observable from the syntactic structure but must be inferred, (2) the semantic memory database of event schemas that support correct verb predictions may not be organized in such a way that it can be rapidly probed with cues like *cop-as-agent* or *thief-as-patient*. In a separate paper, they are able to use similar logic to demonstrate that it is the argument *roles* in particular that are slow to impact prediction, as comprehenders appear quick to identify which noun phrases in the sentence are arguments of the verb at all and to preferentially weight these arguments in computing predictions for the verb (Chow et al., 2016b).

For the current research, the key takeaway from the prior work by Chow et al. (2018) is that we can estimate the temporal dynamics of argument structure computations by using N400 designs that vary the timing between the word-to-be-predicted and the critical elements of the context that could contribute to that prediction, and thus gain insight into the processes that relate the linguistic input with conceptual representations in long-term memory. In the current study, our goal is to use the same kind of approach to investigate the online computation of more complex argument structures and their corresponding event structures, by taking advantage of some convenient properties of compound verbs in Mandarin.

The current study

Compounding is a very productive word formation process in Mandarin. In fact, according to Huang (1998), stems of all lexical categories, except for prepositions, could be combined to form a compound. In the current study, we investigate the argument and event structure computations required to process compound verbs composed of two verbal morphemes (V1-V2); in particular, compound verbs whose two verbal morphemes are involved in a causal/resultative relation (i.e., V1 resulting in V2). In the most common type of resultative compound verb (Shen & Mochizuki, 2010), V1 is a transitive verb and V2 predicates the

object of V1, indicating how the object of V1 was affected by the event described by V1. For example, in sentence (1), the complex predicate *washed-ruined* introduced a subject agent that performed a washing event, and an object patient that was ruined as a result of washing.

(1) 媽媽 洗壞了 衣服. (Transitive)

Mom washed-ruined le the clothes

"Mom washed the clothes so that the clothes were ruined.

While the literature on argument structure processing often characterizes the problem as a relatively straightforward one of mapping arguments to a predicate and participants to an event, resultatives are one of many cases that remind us that languages regularly make use of structures that go beyond this simple characterization. In the interpretation of resultatives and other compound verbs, participants are related to events, but events are also related to other events. In resultatives, this relation has a specifically causal dimension: the result described by V2 is in some way caused by the event described by V1. The goal of our study was to begin to map the time course of the syntactic and semantic processes that are engaged by these more complex relations, in order to bring new insights to our understanding of the components of argument structure computation in general. As a starting point, we hypothesized that the extra complexity of the argument and event structure in resultatives would require extra processing time, delaying updates to predictions about upcoming arguments.

In the three experiments reported here, the basic logic was the following. We created subject-verb-object item sets where the amplitude of the N400 response to the object noun was the dependent measure of interest. All versions of a given item had the same object noun, which was carefully selected to have a relatively low cloze probability in control conditions

(~10%), but a relatively high cloze probability in the resultative condition (e.g., Table 1). The key question of interest across the three experiments is how much processing time is required for comprehenders to be able to take advantage of the predictability of the resultative context to reduce N400 responses on the object.

[Table 1 near here]

Given the prior literature discussed above, we assume that in a simple context like "*A kid had bitten _____,*" upon recognizing the word *bite* as a simple verb and retrieving its meaning from the lexicon, comprehenders can rapidly generate a prediction for the object based on the verb alone, searching for frequent *biting* events in semantic and episodic memory and identifying the patient of the event as the likely upcoming object noun. If Chow et al. (2016a) are correct, comprehenders can also rapidly identify the pre-verbal noun as an argument and use it to search memory specifically for *biting* events that have *kid* as a participant. By contrast, if the verb is a resultative compound verb, then comprehenders would have to additionally analyze the correct structure of the compound, evaluate the event relation of the two verbal morphemes, disambiguate the thematic structure, and generate a representation of the complex event, such that the parser could probe memory for schemas or episodes involving the proper agents and patients for the complex event. For example, in *The kid bit-broke _____,* comprehenders would have to recognize the verb-verb sequence as a resultative compound verb, evaluate the relations of *biting* and *breaking* events, disambiguate thematic structure involved with *biting* and *breaking*, and generate a representation of a broken-by-biting event, where the subject should be an animate agent to perform the biting event, and the object should be a patient that could be broken by biting. Then they need to be able to successfully probe long-term memory for broken-by-biting events, potentially constrained to those involving *a kid*. Our goal was thus to begin to home in on how much

time it might take to use this extra information coded by a resultative compound verb to generate predictions about the object.

Before proceeding to the experiments, some basic background on the resultative construction in Mandarin is in order. Although our study focuses on the *washed-ruined* or 'transitive' type of resultative, which is the most common one and assigns an agent role to the subject and a patient-theme role to the object, it is worth noting the existence of other resultative types with different argument relations. In 'unergative' resultatives, V1 is still a transitive verb, but V2 predicates the subject of V1, which thus bears an agent-experiencer role, as in *Mom washed-tired the clothes (*媽媽 洗累了 衣服). Argument roles assigned by V1 are also not restricted to agent and patient roles; in *Boy upset-cried Mom (*男孩 氣哭了 媽媽), V1 assigns an experiencer role to the object. V1 is also not restricted to being a transitive verb; in *Mom coughed-hoarse voice* (媽媽 咳啞了 嗓子), both V1 *coughed* and V2 *hoarse* are intransitive verbs, but combining them together forms a transitive complex predicate. In our materials, the intended parse of the 'transitive' resultative was encouraged through the higher frequency of this type of resultative, verb subcategorization preferences, and plausibility.

A considerable number of existing studies have investigated the role of factors such as lexical frequency, semantic transparency, morphological headedness in Mandarin compound verb word recognition (Kuo, 2006; Zhang & Peng, 1992; see Myers, 2006 for a review), but few have examined the processing of these verbs in a sentence context. To our knowledge, Lin and Jaeger's conference paper (2014) is the only study that has examined the factors of structural probability and thematic role order of resultatives in sentence context. Their eye-tracking results showed that transitive resultatives had the shortest first-pass and total fixation

time at the post-verb critical region, indicating that the transitive is the easiest one to process compared with other types of resultative verbs.

Experiment 1

In Experiment 1, we asked whether comprehenders could use predictions afforded by a resultative compound verb to facilitate processing on the object noun when reading with a stimulus-onset asynchrony of 800ms from verb to object (e.g. *A kid had bit-broke his lip*). In the Resultative condition, the compound verbs were always composed of a transitive V1 and an intransitive V2, in which V2 predicated the object of V1 and indicated the result of V1. Objects were selected to be strongly predictable by the resultative context, as determined by offline completion norming. In this first experiment we included two baseline conditions in which the context did not strongly predict the object. The Simple condition contained a simple verb (V1-asp, e.g. *A kid had bit his lip*). The Causative condition was included to rule out the possibility that any facilitation in the Resultative condition was due to unintended associative priming from V2 alone. Since V2 itself was intransitive, we added a transitive light verb *to make*, to form a transitive complex predicate (e.g. *A kid had made-broken his lip*). To match the number of characters of the verbs in the Resultative and the Causative conditions, we added an experiential aspect marker *guo* after V1 in the Simple condition. All of the verbs thus had the same word length.

Experiment 1 used an 800ms stimulus-onset asynchrony (SOA) between each word (600ms on, 200ms off), where the compound verb was presented on a single screen, as is natural in Mandarin. In other words, from the onset of the verb, comprehenders had 800ms to process the verb and to predict an upcoming object noun. We note that although in English studies the typical SOA used for RSVP is shorter than 800ms, such a slow presentation rate is relatively common in Mandarin (e.g., Zhou et al., 2010; Su et al., 2018). With no clear prior evidence about what time range might be required for complex argument/event structure

processing, we chose to begin with an 800ms SOA as it is a slow enough presentation rate to not be consciously taxing, but has been successfully used to identify certain slower aspects of argument role computation/prediction (Momma et al., 2015). If 800ms SOA is enough time for participants to compute the resultative structure and use it to generate predictions about the object, then ERP responses to the critical object noun should track the offline cloze probability, with reduced N400 amplitude in the Resultative condition relative to the Causative and Simple conditions. However, if prediction on the basis of the Resultative takes longer than is afforded by an 800ms SOA, then we would see no N400 differences among the three types of verbs. As this second case predicts a null effect, we also included a sanity check comparison in a separate set of items to show that N400 effects are indeed elicited for predictable and unpredictable object nouns following simple verbs.

Participants

Forty-nine naïve young adults (28 females, 20-35 years old, mean: 24) participated in the study at National Taiwan Normal University. All of them were right-handed native Mandarin speakers, without a history of neurological or psychiatric disorders. Of the 49 participants, 20 were excluded after pre-processing because of excessive eye-blinking, muscle potentials, sweat artifact and alpha waves[1]. The reported results were obtained from the remaining 29 participants (15 females, 20-34 years old, mean age: 24). All of them consented to participate in the experiment. The experiment protocol was approved by the Institutional Review Board Office at the University of Maryland, College Park.

Materials

Our stimuli were sentences of SVO structure, with the verbs varying among the following three conditions: Resultative (*bit-broken*), Causative (*made-broken*) and (*bit-asp*), and the rest of the sentence being the same. Note that even though the subject and objects were kept

identical, we intended to make the object in the Resultative context more predictable than that in the Causative and Simple contexts (see Table 1).

We started by finding resultative verbs from *A dictionary of Chinese verb-resultative complement phrases* (Wang, Jiao & Pang, 1987). We selected an initial list of high frequency resultative compound verbs (n= 186) as the critical verbs for our Resultative condition. Based on the verbs (V1-V2) in the Resultative condition, we created our Causative and Simple conditions. The verbs in the Causative condition were resultative complex predicates whose V1 was a causative light verb *make* and V2 was taken from the Resultative condition. As for the Simple condition, its verbs were literally simple predicates. We took V1 from the Resultative condition and added an experiential aspect marker *guo* after V1 to match the number of characters in Resultative and Causative conditions. Note that resultative compound verbs in Mandarin are usually accomplishment or achievement verbs which denote telic events (Tai, 1984), and they frequently occur with the perfective aspect marker *le*. We thus added the perfective aspect marker *le* at the end of each verb in all experiment conditions to make them sound more natural in a sentence context.

In total we created 186 triplets of verbs, with one Resultative verb, one Causative verb and one Simple verb in each triplet. We added a subject noun phrase in each set, such that the subject noun phrase was the same among different conditions. The 186 triplets of subject-verb frames, 558 sentence frames in total, were divided into nine lists. Each list had 62 frames that were critical to the current study, and none of the frames were repeated among the lists. We had another 360 filler sentence frames, which were stimuli for another experiment, were divided into nine lists (so 40 filler frames per list) to pair up with the current study. Therefore, each list contained 102 sentences. 225 participants were recruited for the cloze norming (25 participants per list); none of the participants took part in the ERP experiment. Cloze norming data were collected online via Ibex Farm

(http://spellout.net/ibexfarm/). We presented the context of a sentence frame all at once and the sentence frame would remain on the screen. Participants were instructed to provide the best continuations for the sentence frames. When computing the cloze probability of the target objects, we counted near synonyms (e.g., 馬路 *road* and 道路 *roadway*), nouns that were further specified by a modifier, (e.g., 秀髮 *beautiful hair* and 頭髮 *hair*), and words that contained a functional morpheme (e.g., 刀 and 刀子 *knife*) as the same lexical item.

Through cloze norming, our goal was to select sentence frames in which a given object noun phrase was more predicted by the Resultative condition than by the Causative or Simple conditions. Sentence frames that did not meet this criterion were excluded. The finalized stimuli were comprised of 90 triplets, with the average cloze probability for the target noun being 39% (range: 16%-80%) in the Resultative condition, 9% (range: 0%-36%) in the Causative condition, and 11% (range: 0%-36%) in the Simple condition (See Table 1). After finalizing the target words, we added more contexts following the target object nouns to make the sentences slightly longer and sound more natural. Each sentence consisted of six to nine words, with each word being one to four characters long. As a sanity check, we also included a set of 60 sentences from Liao and Chan (2016) with a similar cloze probability contrast (high cloze: 40%; low cloze: 0%), but where the predictability was driven by multiple features of the context and not just the verb. However, note that the cloze target of these sanity check sentences was in the sentence final position.

Due to the fact that we had three lists for the experiment manipulations and two lists for the sanity check sentences, six experimental lists were constructed such that no sentence context or target was repeated within the same list. The presentation order of the sentence stimuli was randomized within each list. Participants were randomly assigned to one of the six lists.

Procedure

Participants sat in front of a computer screen with their hands on a keyboard. Sentences were segmented into words; the complex verb and aspect marker were always presented as a single word on the same screen (see example in (2)), which were presented one word at a time in a white font (traditional Chinese characters) on a black background at the center of the screen. Each sentence was preceded by a fixation cross that appeared for 600ms. Each word appeared on the screen for 600ms, with a 200ms inter-stimulus interval, for a stimulus-onset asynchrony (SOA) of 800ms (See Figure 1 for details). At the end of 20% of the trials, a comprehension question would show up on the screen, and the participant had to answer via button pressing in order 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 three blocks of 50 sentences each, with short pauses in between. Including set-up time, an experimental session lasted around 90 minutes.

(2) Sentence segmentation for stimulus presentation:

小孩/咬破了/嘴唇/

The kid/ bit-broke le/ (his) lip/

(Figure 1 near here)

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 (two on the outer canthus of each eye and two on the upper and lower ridge of the left eye) were placed 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 from the onset of the verb. 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 the continuous EEG file was epoched (1) from -100 to 1600 ms, from the onset of the verb until the end of the object noun phrase, for all the experimental conditions and (2) from -100 to 800 ms for the sanity check items. 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 1600 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 if needed, adjusted the voltage threshold for individual subjects. Epochs contaminated by excessive blinking, body movements, skin potentials, and amplifier saturation were rejected. The overall rejection rates (including sanity check items) across participants was 20.3 ± 11.3% (mean ± SD); participants with greater than 40% trials rejected were excluded from further analysis. The rejection rates of each critical condition were: Resultative: 22.6 ± 11.9% (mean ± SD), Causative: 22.1 ± 11.0%, Simple: 22 ± 10.3%.

Our hypotheses centered around the N400 response at the object noun phrase. 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, and averaged them as our single clustered region of interest (ROI). We carried out a repeated-measure Type III ANOVA on the mean amplitudes in the measurement time windows of 1100-1300 ms, which was 300-500 ms after the onset of the noun, evaluating effects of Verb type (Resultative, Causative, Simple). When Mauchly's test of Sphericity was violated, Greenhouse-Geisser correction (Greenhouse & Geisser, 1959) was applied to adjust the p-values.

In the sanity check items that were designed to replicate standard N400 effects of cloze probability, we carried out a paired t-test over the same set of electrodes evaluating the effect of predictability (High-cloze, Low-cloze).

Results

Behavioral data

The overall accuracy rate to the comprehension questions was 93 % (80%-100%), showing that participants were paying attention during the experiment.

ERP data

In order to ensure a clean baseline, we time-locked ERPs to the onset of the verb, where the three conditions first differed from one another, even though our interest of analysis would focus on the N400 responses at the noun. We ran statistic analyses on a pre-defined clustered ROI. However, when we visually inspected the data, we observed somewhat inconsistent patterns across electrodes: although the N400 responses to Causative condition were numerically more negative than Resultative among electrodes in our ROI, the N400 responses to Simple were more negative than Resultative over some electrodes (e.g., Cz) but

not others (e.g., Pz). Figure 2 shows the grand average ERPs to Resultative, Causative and Simple conditions across several electrodes (Cz, Pz) that usually show robust N400 effects. We included the whole-head grand averaged ERPs in the supplementary materials.

(Figure 2 near here)

Statistical analyses during the N400 time-window showed a main effect of Verb type ($F(2, 56) = 3.70$, $p < 0.05$). Follow-up paired-t-tests reveal that the N400 response to the object in the Resultative condition was significantly smaller than the Causative ($t(28)=-2.55$, $p < 0.05$), but when compared the N400 response to the object in the Resultative condition to the Simple condition, there was no significant difference ($t(28) = -1.09$, $p = 0.28$).

Plotted in Figure 3 are the grand average ERPs to the Predictability effect in High- and Low-cloze sanity check sentences. During the N400 time window, there was a significant main effect of cloze ($t(28) = 26.10$, $p < 0.001$), which confirms the clear impression from visual inspection that the high-cloze continuations elicited reduced N400 amplitude than the low-cloze continuations.

(Figure 3 near here)

Discussion

Experiment 1 was designed to investigate how quickly the computation of a resultative compound verb can impact predictions of an upcoming noun. We used an 800ms stimulus-onset asynchrony rate and asked whether the cues encoded in the resultative compound verbs could be used to update predictions in time to facilitate processing of the subsequent noun. We used materials in which offline cloze probability was high for the Resultative condition and low for the Causative and Simple conditions, so that rapid use of resultative cues for prediction should result in a reduced N400 for the noun in the Resultative condition relative to the other two. In contrast, if an 800ms SOA is not enough time to use resultative cues to

update predictions, we expected that all three conditions should elicit relatively similar N400 amplitudes.

However, it is difficult to draw strong conclusions about either possibility from these results, as they fit neither of these predicted patterns. As shown in Figure 2, centro-parietal electrodes did show a reduced N400 response to the object in the Resultative condition than in the Causative condition, and this difference was significant in a follow-up pairwise comparison. However, the N400 contrast between the Resultative and the Simple conditions were not significant, even though the cloze probability to the object of the Causative and the Simple conditions were quite similar (Causative: 9%; Simple: 11%). In fact, if we took a closer look, we found that some anterior electrodes seemed to fit the 'fast' prediction pattern, with smaller N400 in Resultative relative to Simple, whereas more posterior electrodes seemed to fit the slow prediction hypothesis, with no N400 difference between Resultative and Simple conditions. It remained unclear to us why the Simple patterned differently than the Causative condition, since both of their object nouns were relatively unpredictable based on the offline cloze norming. Such a finding was not consistent with any hypothesis we were aware of.

Although this pattern of data is equally unexpected on both hypotheses, both hypotheses are also consistent with reasonable post-hoc explanations that can inform improvements in the design. If resultative cues can be used rapidly to update predictions, it is possible that we failed to detect a true N400 difference between Simple and Resultative conditions because our cloze probability contrast was not robust enough across items, or that the 1 x 3 design limited power for detecting our effect of interest. In particular, it could be that the N400 to the object in the Causative condition was not reduced for a different reason, perhaps due to properties specific to the Causative construction. In Experiment 2, we worked to mitigate these possibilities by selecting a more tightly controlled subset of Resultative and

Simple verb items from Experiment 1, in a different design that compared two different types of compound verbs and their corresponding Simple controls.

Although our sanity check sentences demonstrated a classic N400 effect, showing that participants did engage prediction during the experiment, we note that these items were qualitatively different than experiment materials: the cloze contrast was not closely matched to the experimental materials, target words were placed at sentence final position, and the predictability of target words was driven by multiple sources of contexts, not just subject and a verb. Therefore, in Experiment 2, we also modified the items in the simple predictability contrast to be more comparable to experimental materials.

Experiment 2

In Experiment 2, we aimed to improve on the design and materials of Experiment 1. We selected a more tightly controlled subset of Resultative and Simple verb items from Experiment 1, and created a 2 x 2 design in which the predictive effect of the resultative was compared against the predictive effect of a different type of compound verb, coordinate verbs. This allowed us to test different sources of online prediction difficulty in complex predicates.

Similar to resultative verbs, coordinate verbs are compound verbs that are composed of two contentful verbal morphemes, V1 and V2. Whereas the verbal morphemes in a Resultative are involved in a causal relation (V2 by V1), the two morphemes of Coordinate are in a coordinate relation (V1 and V2). For example, in the sentence *The store owner hit-scolded the employee*, the interpretation is that the store owner hit and scolded the employee. Although coordinates and resultatives bear a surface similarity in both being composed of two predicates, comprehenders can distinguish them online through cues provided by the meaning of the two verbs and by the subcategorization of V2. For example, given the

compound verb *hit-scold*, the V1 *hit* is a transitive verb, which requires an agent and a patient, and so is the V2 *scold.* Since V1 *hit* and V2 *scold* have the same subcategorization, they naturally form a coordinate relation, and both of them are the head of the compound verb. It should be noted that because V2 *scold* is not a stative verb, it cannot denote the state of change after the V1 hitting event. The compound verb *hit-scold* cannot be a resultative verb.

The goal of Experiment 2 was to identify potential sources of online prediction difficulty in complex predicates. As in Experiment 1, the target nouns were more predictable in the complex predicate contexts compared with the simple predicate contexts. If computing a complex predicate is generally hard in a way that causes delays in prediction, then we would expect to observe no N400 effect to the objects in both Resultative and Coordinate contexts. However, if it is resultative predicates specifically that are costly, because of the causal relationship between V1 and V2, then we would expect to obtain an interaction between Set type and Predictability effect, with a significant N400 contrast in Coordinate context, but not Resultative one.

Participants

The participants were 40 naive young adults (28 females, 18-40 years old, mean: 24) from National Taiwan Normal University. All of them were right-handed native Mandarin speakers, without a history of neurological or psychiatric disorders. Of the 40 participants, 7 were excluded after pre-processing because of excessive eye-blinking, muscle potentials, sweat artifact and alpha waves. The reported results were obtained from the remaining 33 participants (18 females, 19-40 years old, mean: 24). All of them consented to participate in the experiment. The experiment protocol was approved by the Institutional Review Board Office at the University of Maryland College Park.

Materials

Similar to Experiment 1, the materials were sentences of SVO structure, with the verbs varying among different conditions. Two sets of compound verbs were created: Resultative set and Coordinate set. Within each set, in addition to a compound verb condition, we included a simple verb condition as a baseline condition. The verbs in the simple verb conditions were the V1 from the compound verb conditions, followed by an experiential aspect marker *guo* to match the number of characters in the compound verb conditions. In other words, the Resultative set contained Resultative (V1-V2), and R-Simple (V1-asp) conditions whereas the Coordinate set contained Coordinate (V1-V2) and C-Simple (V1-asp) conditions. Note that resultative compound verbs in Mandarin are usually accomplishment or achievement verbs which denote telic events (Tai, 1984), and they frequently occur with the perfective aspect marker *le*. We thus added a perfective aspect marker *le* at the end of each verb in all experiment conditions.

Although the verbs varied, the subject and object were identical between conditions in the same set. We intended to make the object in Resultative context and Coordinate context more predictable than that their Simple cotrols. Materials for the Resultative set were 60 Resultative verbs and corresponding Simple verbs selected from Experiment 1. For the Coordinate set, the procedure to finalize the materials was similar to the procedure to Resultative verbs in Experiment 1. Coordinate verbs were chosen from *An Online Revised Mandarin Dictionary by the Ministry of Education, R.O.C.* (http://dict.revised.moe.edu.tw/cbdic/index.html). We did not include Coordinate compound verbs whose V1 and V2 are synonyms. In addition, we excluded Coordinate verbs whose V1 was identical to the V1 of the Resultative verbs, because in this case the baseline condition to the Coordinate condition (C-Simple) and the baseline condition to the Resultative condition (R-Simple) would be identical. We selected 119 coordinate compound verbs and created 119

pair of verbs, with each pair containing one Coordinate verb, and one Simple verb. We added a subject noun phrase in each set, such that the subject noun phrase was the same between conditions. Then, the 119 sets of subject-verb frames (each set contained 2 subject-verb frames, so 238 frames in total), were divided into two lists such. Each list contained 119 sentences. Fifty participants were recruited for the cloze norming (25 participants per list); none of the participants took part in the ERP experiment. Cloze norming data were collected online via Ibex Farm (http://spellout.net/ibexfarm/). We presented the context of a sentence frame all at once and the sentence frame would remain on the screen. Participants were instructed to provide the best continuations for the sentence frames. The presentation order of the sentence stimuli was randomized.

To demonstrate that participants were able to generate predictions based on a minimal sentence context, we also created predictability sentence frames that only contained a subject and a simple verb, which were a better match to the experimental conditions. One hundred subject-verb frames were subject to online cloze norming. Another 25 participants were recruited to perform sentence completion task. None of the participants took part in the ERP experiment. The presentation order of the sentence stimuli was randomized.

Through cloze norming, our goal was to select sentence frames in which a given object noun phrase was highly predicted by Resultative condition and Coordinate condition, but not by their baseline R-Simple and C-Simple conditions. Sentence frames that did not meet this criterion were excluded. The finalized stimuli were 60 items in the Resultative set and 60 items in the Coordinate set. The averaged cloze probability to the target nouns in the Resultative set was 39% for Resultative (range: 16%-80%) and 9% for R-Simple (range: 0%-36%) and in the Coordinate set was 38% for Coordinate (range: 16%-72%) and 10% for C-Simple (range: 0%-44%). The cloze sanity check items were of similar contrast to the experimental materials (High-cloze: 38% vs. Low-cloze: 9%) (See Table 2). After finalizing

the target nouns, we added more contexts following the target nouns to make the sentences slightly longer and sound more natural. Each sentence consisted of six to nine words, with each word being one to four characters long.

(Table 2 near here)

Two experimental lists were constructed such that no sentence context or target was repeated within the same list. Each list consisted of 240 sentences, including 60 items of Resultative set, 60 items of Coordinate set, 60 items of cloze sanity check items, and additional 60 filler items that were of similar length for an unrelated experiment that will not be described here. The presentation order of the sentence stimuli was randomized within each list. Participants were randomly assigned to one of the two lists.

Procedure

The procedure was identical to Experiment 1.

Data acquisition and analysis

Data acquisition and analysis, including the regions of interest, were identical to Experiment 1. The overall mean rejection rate (including sanity check items) across participants was 24.1 ± 12.4% (mean ± SD); participants with greater than 40% trials rejected were excluded from further analysis. Rejection rates of experimental conditions were summarized as follows: Resultative: 28.6 ± 12.9%, R-Simple: 30.8 ± 16.3%, Coordinate: 27.8 ± 9.4%, and C-Simple: 28.4 ± 12.8%. We carried out a repeated-measure ANOVA on the mean amplitudes in the measurement time windows of 1100-1300ms, which was 300-500ms since the onset of the noun, and evaluated effects of Set type (Resultative, Coordinate) and Predictability (High-cloze, Low-cloze). Follow-up paired t-tests were performed when an interaction was observed.

In the sanity check items that were designed to replicate standard N400 effects of cloze probability, we carried out a paired t-test over the same set of electrodes evaluating the effect of predictability (High-cloze, Low-cloze).

Results

Behavioral data

The overall accuracy rate to the comprehension questions was 91 % (83%-96%), showing that participants were paying attention in the experiment.

ERP data

Plotted in Figure 4 shows the grand average ERPs to the verb and object noun in the Resultative and R-Simple conditions and in the Coordinate and C-Simple conditions. Visual inspection suggested that there was no N400 cloze difference to the objects in the Resultative set, but that there was a difference in the Coordinate set. A repeated-measure Type III ANOVA analyses demonstrated a significant Set type by Predictability interaction ($F(1,32) = 4.346$, $p < 0.05$). Follow-up pairwise analyses revealed that that there was a significant difference between Coordinate and its C-Simple baseline ($t(32) = 2.96$, $p < 0.01$), but not between Resultative and its R-Simple baseline ($t(32) = 0.56$, $p = 0.58$).

(Figure 4 near here)

It is worth noting that visual inspection suggested that the coordinate comparison also showed an earlier increased negativity for the C-Simple condition relative to the Coordinate condition that onset approximately 500ms into the verb region (more negative for simple verbs than coordinated verbs). Although we did not have any specific hypotheses about what ERP differences might emerge at the verb, this difference might raise the question of whether the N400 difference observed at the object noun in the coordinate conditions might be partly

due to ongoing negativity for the C-Simple condition from the verb region. We think this is unlikely as the waveforms appear to come back together prior to the N400 window on the noun, but we will return to this point in examining the results of Experiment 3.

Figure 5 shows the grand average ERPs for the Predictability effect in the sanity check items (High-cloze vs. Low-cloze). Visual inspection suggested that the high-cloze continuations had reduced N400 amplitude than the low-cloze continuations. The results of the pairwise comparison also showed a significant effect of cloze ($t(33) = 4.89$, $p < 0.05$). (Figure 5 near here)

Discussion

The results of Experiment 2 suggest that prediction for the object noun is not immediately updated by information from Resultative verbs. The interaction between Set type and Predictability indicated that predictability was modulated differentially by the two types of compound verbs we tested: Coordinate verbs and Resultative verbs. Specifically, we found an N400 effect in the Coordinate set, indicating that information encoded in coordinate verbs can impact prediction in time. However, there was no N400 effect in the Resultative set. Based on the significant interaction, we could infer that the computation of Resultative was too slow to impact prediction in time.

Our finding that Coordinate verbs immediately contributed to object predictions, above and beyond V1 alone, is important for ruling out several possible explanations of the failure for Resultative verbs to do so. In Experiment 1, we observed a larger N400 difference between Resultative (V1-V2) and Causative (V2) than between Resultative and Simple (V1). One possible explanation of this pattern could have been simply that predictions were rapidly updated on the basis of V1 only, with V2 contributing little to constrain predictions. However, in Experiment 2, we showed that although both Resultative and Coordinate are

compound verbs, comprehenders were only able to quickly incorporate V1 and V2 into their prediction when they form a coordinate relation. Therefore, we would like to argue what slowed down prediction in Resultative is a process that was specific to Resultative verbs. We suggest that it could be the process of computing the causal relationship between V1 and V2 that slowed down prediction, but other alternatives are also possible. We will discuss these alternatives in the General Discussion section.

In Experiment 2, we made the sanity check sentences more comparable to the experimental sentences: The sentence context of sanity check items consisted of a subject and a simple verb, and the cloze contrast was matched to that of experiment conditions. With these manipulations, the N400 effect was still significant, showing that participants were engaged to update their predictions given the minimal contexts. However, it is essential to note that the N400 effect of the sanity check sentences was much smaller than that in Experiment 1. These results suggest that a cloze difference of this magnitude based on subject and a verb corresponds to a relatively small effect size on N400 amplitude.

Based on the results of Experiment 2, we could infer that the computation of Resultative was too slow to impact prediction in time when words are presented with an 800ms SOA. This hypothesis would predict that with enough time the N400 contrast should emerge. Experiment 3 was designed to test this hypothesis.

Experiment 3

Experiment 2 showed that participants could quickly update predictions based on Coordinate verbs but not Resultative verbs. In Experiment 3 we asked, could predictions be updated if comprehenders were given several hundred more milliseconds? We used the same materials as in Experiment 2 except that we added a buffer to allow additional processing time by inserting a prenominal modifier with minimal conceptual content, such as a possessive or a

quantifier, between the compound verb and its object noun, which resulted in an extra 400ms of processing time compared to Experiment 2 (see details below). Our hypothesis predicts that the N400 cloze effect should re-emerge in the Resultative set when sufficient processing time is provided.

Participants

The participants were 48 naive young adults (22 females, 18-33 years old, mean: 23) from National Taiwan Normal University. All of them were right-handed native Mandarin speakers, without a history of neurological or psychiatric disorders. Of the 48 participants, 10 were excluded after pre-processing because of excessive eye-blinking, muscle potentials, sweat artifact and alpha waves. The reported results were obtained from the remaining 38 participants (20 females, 18-33 years old, mean: 23). All of them consented to participate in the experiment. The experiment protocol was approved by the Institutional Review Board Office at the University of Maryland College Park.

Materials

The materials were identical to Experiment 2, except that we inserted a modifier (either a possessive or a quantifier) between a verb and a noun, such that participants might have a little buffer to update their predictions.

Although we assumed the predictability of the target noun would remain the same despite the insertion of a modifier, we conducted post-hoc cloze norming to confirm this. Our norming focused on the 60 Resultative sets of subject-verb-modifier frames (each set contained Resultative and R-Simple conditions, so 120 frames were normed in total). They were divided into two lists. Fifty participants were recruited for the cloze norming (25 participants per list); none of the participants took part in the ERP experiment. Cloze norming data were collected online via Ibex Farm (http://spellout.net/ibexfarm/). We presented the

context of a sentence frame all at once and the sentence frame would remain on the screen; participants were instructed to provide the best continuations for the sentence frames. The presentation order of the sentence stimuli was randomized. Surprisingly, our norming revealed that the cloze contrast between the Resultative and R-Simple conditions actually became smaller (in Experiment 3, Resultative: 40% vs. R-Simple: 18% whereas in Experiment 2, Resultative: 39% vs. R-Simple: 9%). Fortunately, this difference goes against our hypothesis (a smaller cloze difference in Experiment 3 than 2, although we expect the N400 effect to re-emerge in Experiment 3) and therefore only acts to provide a more conservative test of that hypothesis.

Procedure

The procedure was identical to Experiment 2, except for the presentation rate. With presentation rate of 800ms in Experiment 2, the EEG recording time was about 40 minutes. As we added a modifier between the verb and the noun in all conditions, the EEG recording time could be even longer. To keep participants from being too tired during the experiment, which could introduce artifacts such as alpha waves, we increased the presentation rate from 800ms to 600ms in Experiment 3. Each word appeared on the screen for 500ms, with a 100ms inter-stimulus interval (See Figure 6 for details). Given the new SOA, participants had up to 1200ms (i.e., the duration from a verb to a modifier) to update predictions whereas in Experiment 2, only 800ms (i.e., the duration of a noun) was available to make predictions.

(Figure 6 near here)

Data acquisition and analysis

Data acquisition and analysis were identical to Experiment 2. We time-locked ERPs to the onset of the verb, where experimental conditions started to differ, with the epoch ranging from -100ms to 1800ms, to cover the brainwave responses from the onset of the verb to the

end of the object noun (600ms each for the verb, modifier, and object). As for the sanity check items, the epoch was from -100 to 600ms. The overall mean rejection rate (including sanity check items) across participants was 23.1 ± 12.7%; participants with greater than 40% trials rejected were excluded from further analysis. Rejection rates of the experimental conditions were summarized as follows: 24.1 ± 11.4%, R-Simple: 24.4 ± 13.2%; Coordinate: 23.8 ± 12.6%, C-Simple: 23.7 ± 11.6%.

Results

Behavioral data

The overall accuracy rate to the comprehension questions was 92 % (83%-98%), showing that participants were paying attention in the experiment.

ERP data

Plotted in Figure 7 shows the grand average ERPs to the Resultative and R-Simple conditions and to the Coordinate and C-Simple conditions. Visual inspection suggested that there was an N400 effect to the objects in the Resultative set as well as in the Coordinate set. Statistically, we found a Predictability main effect ($F(1,37) = 10.73$, $p < 0.005$) and no evidence of a Set type by Predictability interaction ($F(1,37) = 0.05$, $p = 0.82$).

(Figure 7 near here)

Figure 8 shows the grand average ERPs to High-cloze and Low-cloze sanity check sentences. Visual inspection suggested that the high-cloze continuations elicited a reduced N400 response than the low-cloze continuations. Statistics also showed a significant effect ($t(37) = 6.35$, $p < 0.05$).

(Figure 8 near here)

Discussion

In Experiment 3, we investigated if predictions on the basis of Resultatives can be updated when participants were given several hundred more milliseconds. A modifier was inserted between the verb and the object noun to create a little buffer for participants to update predictions. Under these conditions, we did not obtain an N400 reduction at the object in Resultative set in Experiment 2, but it emerged in Experiment 3. By contrast, the N400 reduction was observed at the object in Coordinate set in both experiments. These effects held even though the addition of the modifier unintentionally made the cloze contrasts slightly smaller than Experiment 2 and Experiment 1 (mean differences of ~20% in Experiment 3 and ~30% in Experiment 2 and Experiment 1). Overall, these results showed that the causal relations between V1 and V2 could constrain predictions, if participants were provided with sufficient time—here, a buffer of 1200 ms between complex verb and object noun.

During the verb region in Experiment 3, we showed the same numerical pattern from Experiment 2 of more negativity for the C-Simple condition than the Coordinate condition at around 500ms post-verb onset. One concern from Experiment 2 was whether the apparent N400 effect at the noun could rather be due to differences carried over from the verb region. The prenominal modifier in Experiment 3 allowed us to see that the verb-elicited differences appeared to subside by about 800ms post-verb onset, as illustrated in Figure 7. In order to confirm that there were no reliable differences between Coordinate and C-Simple conditions immediately prior to noun onset, we ran an additional paired t-test on 100ms before the onset of the noun. Results showed that the Coordinate condition did not differ from the C-Simple condition ($t(37) = 1.29$, $p = 0.20$). Therefore, it unlikely that early differences on the verb are responsible for the significant N400 effect observed on the subsequent object noun for the coordinate comparison in Experiment 2.

General Discussion

Three ERP experiments were conducted to investigate the predictive mechanism of online sentence comprehension through properties of Mandarin compound verbs. We focused on resultative compound verbs whose V2 predicates the object of V1, featuring that the object is affected by V1. We asked if the causal relationship of a resultative compound verb could rapidly constrain predictions of a subsequent object. The predictive effect of resultative compound verbs was compared against that of coordinate compound verbs, which allowed us to test different sources of online prediction difficulty in complex predicates.

The N400 was used as a neural indicator of what is predicted in the current study. Although results from Experiment 1 were inconclusive, the better-controlled design in Experiment 2 suggested that predictions on the basis of the resultative were not updated in time to impact processing of the object when verb and object onset were separated by 800ms. This 'timing' hypothesis was supported by Experiment 3, where the N400 predictability effect was recovered when participants were provided with up to 1200ms between verb and object onset to update predictions.

Two classes of explanation for why prediction update is delayed are (1) the computation of a resultative predicate is slow and/or (2) using the resultative predicate to generate predictions—to retrieve entities/nouns from memory that are likely to complete the message/clause—is slow. We do not have strong evidence to favor one over the other. In fact, as the two classes of explanation target different stages of processing, it is likely that they are not mutually exclusive. In the following, we consider the two accounts in more detail.

Slow prediction due to the computation of a complex resultative predicate

One possibility is that predictions based on the resultative predicate take significant time because computing the predicate itself takes time. As discussed in earlier sections, complex

predicates are different from simple verbs in many aspects. For example, with the combination of two verbal morphemes, comprehenders could be struggling with lexical processing, such as accessing the meaning of V1 and V2, constructing the mental representation of the complex predicate, and decomposing the internal structure of the complex predicate. Any of the above computations could take longer and slow down predictions. To identify sources of online prediction difficulty in complex predicates, we introduced coordinate compound verbs in Experiments 2 and 3, and compare the predictability effect of resultative compound verbs against coordinate ones. Our results revealed that not all complex predicates yield equally slow predictions: with an 800ms SOA, predictability effects were observed after coordinate verbs but not resultative verbs. These data suggest that what slowed down the predictive mechanism were processes that were specific to resultative compound verbs.

Different theoretical frameworks differ in exactly what kinds of complex predicate representations are computed over compound verbs. Li (1990) proposed that when the two verbal morphemes were merged into a complex predicate, theta roles from V1 and V2 should merge into a composite theta role. For example, in the complex predicate *bit-broke* "broken by biting," whose V1 *bite* required an agent and a patient and V2 *break* required a theme, the theme role from V2 should be merged with the patient role from V1, and then the composite theta role, patient-theme, would be assigned to the object noun phrase. Different from Li, Williams (2007) treats resultatives as an event that involved a causer and causee, or an agent and a patient that underwent an event of change. In other words, V1-V2 together is not the same event as V1 or V2 alone, nor does it equal a conjunction of V1 and V2. The meaning of a resultative rather refers to events that had V2 by means of V1. Since our experiments were not set up to test any of the above frameworks, we do not have a stand to argue for one analysis than the other. However, both frameworks feature unique properties in the

resultative structure. If some of these properties are particularly costly to compute, we could explain why updating predictions subsequent to a resultative verb took longer than other types of verbs.

Slow prediction due to memory search for an optimal candidate

We also entertained the slow prediction hypothesis proposed by Chow et al. (2018), which would hold that predictions were slow because it takes longer to use the cues from resultatives to retrieve the best fitting word or concept for the context.

Chow et al. (2016a), specifically propose that lexical prediction can be seen as a two-step memory retrieval process, which involves (1) a fast parallel process that activates all the event schemas associated with the individual context words, and (2) a slow serial search through this initial set for the schemas that match the argument role assignment of the nouns in the context. For example, in *Cop ba thief arrest* (the cop arrest the thief'), it was fast for comprehenders to recognize that *cop* was an agent and *thief* was a patient. Nevertheless, because the information of *taking-cop-as-an-agent* and *taking-thief-as-a-patient* were compound cues and not simplex cues, comprehenders would have to serially search through the semantic space for an item that satisfied all the features, delaying successful prediction. On the other hand, other authors point out the challenges in formulating a principled distinction between simplex and compound cues that captures the semantic retrieval phenomena, and instead suggest that delays in contextual prediction may reflect differential weighting of cue certainty across time (Kuperberg, 2016). What these accounts have in common is that they all place the locus of the timing effects in the process of prediction update, rather than the process of parsing and interpreting the context.

To explain the results of the current study, these kinds of account would posit that the computation of complex predicates, including the configuration of argument structures, was

completed rapidly, but what slowed down prediction update was the process of retrieving the candidate that best satisfied the context. Consider predictions generated from a coordinate compound verb first. When perceiving the verb *hit-scold,* comprehenders recognized the verb-verb sequence as a coordinate verb, and consider both verbs heads of the complex predicate. Therefore, both verbs served as retrieval cues at an initial stage, with a set of hittable and a set of scold-able candidates being activated, and those that matched both cues being the most activated. In this case, *employee* was the most preferred answer to the context *The store owner hit-scolded* _____(See Figure 9 for a Venn diagram of comprehenders' prediction space). In this case, a simple summation of the activation elicited by each verb weighted equally would be likely to yield successful retrieval of the best-fitting candidate. Our ERP results indeed indicated that participants could make use of the cues provided by a coordinate verb to update their prediction promptly.

However, prediction on the basis of a resultative verb could be more complicated. When perceiving the verb *bit-broke* in *The kid bit-broke* _____*,* comprehenders should recognize the verb-verb sequence as a resultative verb, in which only V1 *bite* is the head of the complex predicate. Given identification of this structural relationship, it could be that at an initial stage, only the V1 event is used as a retrieval cue, or that this cue is initially weighted more strongly. In this case, initial activation would be focused on entities involved as participants in biting events involving a kid, such as toy, corn and lip. In a two-stage model, it might be only in a second, later stage that these initial candidates are serially searched for one that could be broken by biting, in this case, *lip* (see Figure 9). In a more continuous model, the delayed memory access of the predictable candidates in the resultative case might reflect the lower frequency, and thus lower resting activation level, of the complex real-world events that support the prediction (that is, biting events in general will be more frequent than biting events that result in breaking). It could also be that generating a

prediction about the likely result of an event does not solely depend on retrieving memories of existing events, but also requires an extra processing step of inference or simulation. All of these explanations predict the dissociation in timing from the coordination contexts, in which identifying predictable candidates can be done with reference to simple events in memory.

(Figure 9 near here)

ERP responses to verbs

Our results suggest that the computations required to generate predictions following resultative verbs take longer than following coordinate verbs. While our design focused on neural activity during the target noun, these results raise the question of whether traces of those costly computations could be observed during the ERP to the verb itself. To our knowledge, we are the first group that used EEG responses to study the processing of Mandarin resultatives. Since we did not have any a priori hypothesis about the processing of resultative verbs, we plotted the topographic distribution of ERP effects in P200 (150-300ms), N400 (300-500ms) and P600 (500-800ms) intervals at the verb (see Figure 10). If we could observe any pattern across the three experiments, we might get an initial clue about which stage of processing drives slower prediction update in response to resultative verbs, and whether this would be a useful avenue to pursue in future work.

(Figure 10 near here)

As depicted by Figure 10, relative to the Simple condition, the Resultative condition seemed to elicit a larger negativity over the central-parietal sites in the N400 time window. The topographic distribution and the peak latency resembled an N400 effect. We ran post-hoc paired t-tests to examine the effect, with the same region of interests as what we defined for the N400 effect at the noun. Our analyses showed that the N400 effect was not significant in Experiment 1 ($t(28) = 2.20$, $p = 0.15$), but was significant both in Experiment 2 ($t(32) = 5.47$, $p < 0.05$) and Experiment 3 ($t(37) = 10.63$, $p < 0.001$). We also plotted the topographic maps

of the Coordinate set (see Figure 11). However, unlike the Resultative set, we found no effects at the N400 time window when considering Coordinate relative to C-Simple conditions (Experiment 2: $t(32) = 0.26$, $p = 0.62$; Experiment 3: $t(37) = 0.23$, $p = 0.64$).

(Figure 11 near here)

Although resultative verbs and coordinate verbs are both compound verbs, the post-hoc analyses reveal that only resultative verbs elicited a larger N400 response. As the N400 is primarily associated with lexical or conceptual processing, these data then tentatively suggest that resultative verbs require additional lexical or conceptual computations that could be tied to the delayed prediction effects at the subsequent noun. However, it could also be the case that these effects on the verb are unrelated to the effects on the object and simply differences between the lexical properties of the resultative and coordinate verbs, such as different degrees of semantic association between V1 and V2, lexical frequency, number of brush strokes, neighborhood density, etc. Since the current study was not aimed at evaluating responses at the verb, we did not attempt to control these properties, and so we leave a more systematic investigation of these differential verb responses for future work.

Implications for L2 acquisition and processing

Finally, as the current study took advantage of language-specific properties by which argument structure is encoded in Mandarin, we would like to briefly discuss potential implications of the current study for L2 acquisition and processing. It is interesting to note that Mandarin resultative compound verbs are notably challenging for L2 learners. In Yuan and Zhao (2010), English L2 speakers of different proficiency levels, as well as native Mandarin speakers, were asked to rate the acceptability of sentences that were composed of different resultative compound verbs. Ratings from Mandarin native speakers were collected as a baseline. The results revealed that regardless of proficiency levels in Mandarin, L2 learners rejected almost all of the sentences that contained a resultative compound verb. An

exception was sentences of transitive resultative compound verbs, where advanced L2 learners showed the same acceptance rate as native speakers. The authors attributed the benefit of Mandarin transitive resultative compound verbs to similar thematic configurations in English resultative constructions (i.e., Mom washed the clothes ruined): in both languages, the object noun phrase of a resultative complex predicate received a patient role from V1 and a theme role from V2. Such a transfer effect from learners' first language could also explain why resultative compound verbs of other thematic relations were rejected by the learners, although it still remained a puzzle why thematic role reconfiguration was challenging for L2 learners. To further explore the mental representation of Mandarin resultatives in L2 learners, we suggest that a better understanding of the computation involved in L1 resultative comprehension should be developed. We believe that the current study constitutes one such step.

Conclusion

The current study investigated how quickly two types of complex predicates associated with verb-verb compounds—coordinates and resultatives—could be computed and used to update predictions for the subsequent input, using the N400 response as a measure of online prediction. If processing speed were mainly a function of syntactic complexity, then we would expect both conditions to demonstrate the same temporal dynamics, but if the computations required by certain semantic relations are particularly costly, the two verb types should dissociate. Results from our three experiments indicate that predictions afforded by a resultative verb do not impact processing of the subsequent noun at an effective verb-noun SOA of 800ms, but that predictive effects emerge with a verb-noun SOA of 1200ms. This contrasts with the case of coordinate verbs, which impacted predictions at the verb at both SOAs. We discussed two broad families of accounts for the dissociation: (1) the computation of a resultative compound verb is more taxing and/or (2) retrieving a candidate that fits the resultative context requires longer time. Our study shows that evaluating the speed of prediction update with the N400 is an effective approach for dissociating some of the fine-grained

subcomputations required for the interpretation of complex verb constructions. Future work using this method, in combination with other tools, can help to lay the groundwork for a detailed timecourse model of argument structure computation.

Note

[1.] The rejection rate was unusually high because (1) the epoch was fairly long (-100ms to 1600ms), and (2) the air conditioner in the lab was broken during data collection section. 10 out of the 20 excluded participants were removed because of sweat artifact.

Acknowledgements

This research was supported by NSF grant (BCS-1749407) to Ellen Lau and the William Orr Dingwall Dissertation Fellowship to Chia-Hsuan Liao. We thank Dr. Shiao-Hui Chan and the research assistants in the lab for the support for EEG data collection in Taiwan. We would like to thank the two anomalous reviewers for their constructive comments for an earlier version of this paper.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

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

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.

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*, 1-26.

Chow, W. Y., Momma, S., Smith, C., Lau, E., & Phillips, C. (2016a). Prediction as memory retrieval: timing and mechanisms. *Language, Cognition and Neuroscience, 31*(5), 617-627.

Chow, W. Y., Smith, C., Lau, E., & Phillips, C. (2016b). 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.

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., Kutas, M., & Schul, R. (2010). Age-related and individual differences in the use of prediction during language comprehension. *Brain and language, 115*(3), 149-161.

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24*(2), 95-112.

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.

Huang, C. T. J. (1998). *Logical relations in Chinese and the theory of grammar*. Taylor & Francis.

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., & Osterhout, L. (2005). The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of Memory and Language, 52*(2), 205-225.

Kuo, Grace. (2006). Processing Chinese resultative compounds: a study on its morphological head- edness. e 4th Conference of the European Association of Chinese Linguistics (EACL-4), Budapest, Hungary.

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., Choi, A., Cohn, N., Paczynski, M., & Jackendoff, R. (2010). Electrophysiological correlates of complement coercion. *Journal of cognitive neuroscience, 22*(12), 2685-2701.

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.

Kutas, M., DeLong, K. A., & Smith, N. J. (2011). A look around at what lies ahead: Prediction and predictability in language processing. *Predictions in the brain: Using our past to generate a future, 190207.*

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

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

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

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

Li, Y. (1990). On VV compounds in Chinese. *Natural language & linguistic theory, 8*(2), 177-207.

Liao, C-H., & Chan, S-H. (2016). Direction matters: Event-related brain potentials reflect extra processing costs in switching from the dominant to the less dominant language. *Journal of Neurolinguistics, 40*, 79-97.

Lin C-J., and Jäger L. Reading resultative verb compounds in Chinese sentences: An eye-tracking study. In 2nd East Asian Psycholinguistics Colloquium (EAPC2), Chicago, IL, 2014. University of Chicago.

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.

Momma, S., Sakai, H., & Phillips, C. (2015). Give me several hundred more milliseconds: the temporal dynamics of verb prediction. Talk presented at the 28th annual CUNY Conference on Human Sentence Processing, Los Angeles, CA. March 19-21.

Myers, J. (2006). Processing Chinese Compounds: A Survey of the Literature.

Shen, Y., & Mochizuki, K. (2010). Inheritance of argument structure and compounding constraints of resultative compound verbs in Chinese and Japanese. In *North American Conference on Chinese Lingui* (p. 341).

Su, J. J., Molinaro, N., Gillon-Dowens, M., Tsai, P. S., Wu, D. H., & Carreiras, M. (2016). When “he” can also be “she”: An ERP study of reflexive pronoun resolution in written mandarin Chinese. *Frontiers in Psychology, 7*, 151.

Schegloff, E. (2000). Overlapping talk and the organization of turn-taking for conversation. *Language in Society*, *29*(1), 1-63.

Tai, J. H. (1984). Verbs and times in Chinese: Vendler’s four categories. *Parasession on Lexical Semantics, 20*, 289-296.

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.

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

Wang, Jiao & Pang, (1987). A dictionary of Chinese verb-resultative complement phrases.

Williams, A. (2007). Objects in resultatives. Ms., University of Maryland, College Park.

Wittenberg, E., Paczynski, M., Wiese, H., Jackendoff, R., & Kuperberg, G. (2014). The difference between “giving a rose” and “giving a kiss”: Sustained neural activity to the light verb construction. *Journal of Memory and Language, 73*, 31-42.

Yuan, B., & Zhao, Y. (2011). Asymmetric syntactic and thematic reconfigurations in English speakers’ L2 Chinese resultative compound constructions. *International Journal of Bilingualism, 15*(1), 38-55.

Zhang, B., & Peng, D. (1992). Decomposed storage in the Chinese lexicon. In *Advances in Psychology* (Vol. 90, pp. 131-149). North-Holland.

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

Table titles

Table 1: Example stimulus in each condition in Experiment 1.
(averaged cloze probability in parenthesis)

Table 2: Example stimulus in each condition in Experiment 2.
(averaged cloze probability in parenthesis)

Figure captions

Figure 1: Presentation of stimuli in Experiment 1 and Experiment 2.

Figure 2: Top: Grand average ERPs from the verb to the noun at Cz and Pz in Experiment 1. Bottom: Topographic distribution of ERP effects in the 300-500 ms intervals at the noun in Experiment 1 (Left: Causative minus Resultative; Right: Simple minus Resultative).

Figure 3: Left: Grand average ERPs to cloze sanity check sentences at Cz in Experiment 1. Right: The topographic distribution of ERP effects in the 300-500 ms interval in Experiment 1 (Low cloze minus High cloze).

Figure 4: Top: Grand average ERPs from the verb to the noun of the Resultative set (Left) and Coordinate set (Right) at the Cz electrode in Experiment 2. Bottom: Topographic distribution of ERP effects in the 300-500 ms intervals at the noun in Experiment 2 (Left: R-Simple minus Resultative; Right: C-Simple minus Coordinate).

Figure 5: Left: Grand average ERPs to cloze sanity check sentences at Cz in Experiment 2. Right: The topographic distribution of ERP effects in the 300-500 ms interval in Experiment 2 (Low cloze minus High cloze).

Figure 6: Presentation of stimuli in Experiment 3.

Figure 7: Top: Grand average ERPs from the verb to the noun of the Resultative set (Left) and Coordinate set (Right) at the Cz electrode in Experiment 3. Bottom: Topographic distribution of ERP effects in the 300-500 ms intervals at the noun in Experiment 3 (Left: R-Simple minus Resultative; Right: C-Simple minus Coordinate).

Figure 8: Left: Grand average ERPs to cloze sanity check at Cz in Experiment 3. Right: The topographic distribution of ERP effects in the 300-500 ms interval in Experiment 3 (Low cloze minus High cloze).

Figure 9: Venn diagrams of prediction space of Coordinate verb (Left) and Resultative verb (Right).

Figure 10: Left: Topographic distribution of Resultative effect across different time windows in Experiments 1-3. Right: Grand average ERPs from the verb to the noun of Resultative and R-Simple conditions at the Pz electrode in Experiments 1-3.

Figure 11: Left: Topographic distribution of Coordinate effect across different time windows in Experiments 1-3. Right: Grand average ERPs from the verb to the noun of Coordinate and C-Simple conditions at the Pz electrode in Experiments 1-3.