

Prediction as Memory Retrieval: Timing and Mechanisms

Wing-Yee Chow^a, Shota Momma^b, Cybelle Smith^c, Ellen Lau^b, & Colin Phillips^b

a Department of Linguistics, University College London, Chandler House, 2 Wakefield Street, London WC1N 1PF, UK

b Department of Linguistics, University of Maryland, 1401 Marie Mount Hall, College Park, MD, 20742, USA

c Department of Psychology, University of Illinois at Urbana-Champaign, 603 East Daniel St., Champaign, IL 61820, USA

* Corresponding author address:

Wing-Yee Chow
Division of Psychology and Language Sciences
University College London
Chandler House
2 Wakefield Street
London WC1N 1PF
United Kingdom

Tel: +44 (0)20 7679 4213

email: wingyee.chow@ucl.ac.uk

In our target article (Chow, Smith, Lau, & Phillips, 2015) we investigated the predictions that comprehenders initially make about an upcoming verb as they read. We studied the impacts of the lexical meaning and structural roles of preverbal arguments on comprehenders' verb predictions by examining their effects on the N400 ERP response at the verb. We provided evidence that comprehenders' initial verb predictions are sensitive to the arguments' lexical meaning but not their structural roles. This is compatible with a number of other recent findings using electrophysiology and other time-sensitive measures of prediction. The commentaries offered alternative accounts of our findings (Kim, Oines, & Sikos, 2015; Kuperberg, 2016). In this article we synthesize findings from our target article with other studies that show that verb predictions are sensitive to the arguments' structural roles if more time is available for prediction. Our first main contention is that prediction involves computations that may require differing amounts of time. Our second main contention is that prediction can be usefully framed as a memory retrieval problem, linking prediction to independently well understood memory mechanisms in language processing. We suggest that the delayed impact of argument roles on verb predictions may reflect a mismatch between the format of linguistic cues and target event memories. We clarify points of agreement and disagreement with the commentaries, and give reasons why memory access mechanisms are better able to explain the time course of prediction.

Keywords: language comprehension; prediction; thematic relations; event-related potentials; memory; N400

1. Background

In our original article (Chow, Smith, Lau, & Phillips, 2015, henceforth CSLP), we focused on the predictions that comprehenders initially make about an upcoming verb as they read. We studied the impacts of the lexical meaning and structural roles of preverbal arguments on comprehenders' verb predictions by examining their effects on the N400 ERP response at the verb. We provided experimental evidence that comprehenders' initial verb predictions are sensitive to the arguments' lexical meaning but not their structural roles. We further showed that comprehenders' initial verb predictions are driven by the verb's arguments and not associatively by all nouns in the context. Taken together with convergent evidence in Mandarin Chinese (Chow, 2013; Chow, Wang, Lau, & Phillips, submitted) and Japanese (Momma, Sakai, & Phillips, submitted) which shows that argument role information does impact comprehenders' verb predictions at a later point in time, we proposed that comprehenders quickly identify the arguments of a verb and use event knowledge to generate initial predictions about the verb based on the set of arguments (in a 'bag-of-arguments' fashion) before information about the arguments' structural roles can be integrated to refine their predictions.

We are grateful for the responses to our article, as they present interesting challenges and alternative explanations, and they have helped to shape our own perspective. Here we aim to clarify our proposal and review the broader theoretical motivations and empirical evidence that have informed our thinking, before relating them to the challenges and alternatives put forth by Kim, Oines and Sikos (2015) and Kuperberg (2016).

Our first main contention is that prediction involves computations that may require differing amounts of time. We know from arithmetic that predictions may be easier or harder to compute, even when the predicted item and its probability are identical. The expressions " $37 \times 54 =$ " and " $2000 - 2 =$ " both require the continuation 1998 with probability 1.0, but the first is harder to compute than the second. We believe that our findings demonstrate the same kind of contrast in the domain of verb predictions.

Our second main contention is that prediction can be usefully framed as a memory retrieval problem, and that we can gain a better understanding of underlying mechanisms by linking prediction to independently well understood memory mechanisms in language processing.

We believe the debate about the interpretation of N400 effects holds a key to understanding how humans predict, but the literature is large and can be impenetrable, even to experienced psycholinguists. And so we begin by clarifying the key findings that motivate our proposal.

2. The phenomenon and our proposal

In CSLP we investigated the initial stages of verb prediction during real-time comprehension. We will first clarify what we mean by 'prediction' as it has been used to refer to various distinct constructs. We take 'prediction' to be the pre-activation of stored mental representations and we are interested in the mechanisms involved in pre-activating such representations before the relevant physical stimuli arise. Crucially, we believe mental representations can be pre-activated via different mechanisms, and we make no assumptions about the number or nature of such mechanisms. For example, even though priming is often distinguished from prediction in the literature (e.g., Otten, Nieuwland, & Van Berkum, 2007), we do not exclude it as a possible predictive mechanism as long as it can result in the pre-activation of stored representations.

Although the N400 ERP response is sometimes narrowly described as a 'semantic anomaly' response, much research over the past three decades indicates that it is more accurate to view it as a general index of semantic memory access (Kutas & Federmeier, 2011). All words elicit N400 responses (Kutas, Van Petten, & Kluender, 2006), but when words are primed or predicted by the context, the amplitude of this response is sharply reduced, presumably

reflecting facilitated semantic memory access. While N400 effects of semantic plausibility alone are quite small, N400 effects of predictability are large and reliable (e.g. DeLong, Quante, & Kutas, 2014). Many studies have further demonstrated that the N400 shows graded sensitivity to *cloze probability*, or predictability as measured by offline completion norming (e.g. Kutas & Hillyard, 1984; Wlotko & Federmeier, 2012a). In CSLP we therefore used the amplitude of the N400 response to a given word as an approximate measure of the extent to which the semantic representation of that word has been pre-activated.

In recent years evidence has emerged that another ERP response, the late positivity or P600, often accompanies the detection of semantic implausibility (e.g. Sanford, Leuthold, Bohan, & Sanford, 2011; Brouwer, Fitz, & Hoeks, 2012; Van Petten & Luka, 2012). The mechanisms underlying this late positivity are still much debated and range from post-stimulus integration to general reanalysis to conflict monitoring (e.g. Kim & Osterhout, 2005; Kuperberg, 2007; van de Meerendonk, Kolk, Chwilla, & Vissers 2009; Brouwer et al., 2012; Chow & Phillips, 2013). As our proposal is primarily concerned with predictive mechanisms and the relation of the late positivity to these is unclear, the late positivity does not play a central role here. However we note that we assume that the late positivity reflects a qualitatively different process than the N400, and that when a semantically incongruous word elicits a late positivity we take it to be a good indication that the incongruity was in fact detected.

As noted above, N400 amplitude has usually been found to track cloze probability as measured by offline completion norming, which suggests that 'online' predictions often align impressively with 'offline' predictions generated without time constraints. Our investigation has focused on the perhaps more interesting cases in which online predictions as reflected by the N400 do not match the offline predictability of a word and its meaning. We believe that such prediction 'failures' may provide new insights into the processes that underlie the pre-activation of semantic representations. We next describe the two primary findings we aim to explain: (i) the roles of immediately preceding arguments do not impact online verb predictions as measured by the N400, and (ii) if a short 'buffer' is introduced between the arguments and the verb, N400 sensitivity to the roles of the arguments re-emerges. CSLP provides specific evidence for (i), but our views are guided by (i) and (ii) together.

The clearest evidence to date comes from studies that examine the effects of argument roles, as marked by word order, case, or grammatical particles. Studies across different languages have reported that reversing pre-verbal arguments did not modulate the N400 response at the verb even when it had a clear impact on the verb's offline predictability (e.g., Kolk, Chwilla, van Herten, & Oor, 2003; Oishi & Sakamoto, 2010; Chow & Phillips, 2013). For example, we found no N400 effect in sentences like (1), despite the fact that the verb has a much higher cloze probability in (1a) than (1b). Meanwhile, comprehenders readily detected the implausibility of role-reversed sentences like (1b) and showed a larger P600 response at the verb in (1b) than in (1a) (Chow & Phillips, 2013).

(1) Argument role reversal in a verb-final sentence in Mandarin Chinese:

- a. jingcha ba xiaotou zhua-le...
cop BA thief arrest...
"The cop arrested the thief ..."
- b. xiaotou ba jingcha zhua-le...
thief BA cop arrest...
"The thief arrested the cop ..."

One interpretation of these results is that, even though comprehenders reliably compute an accurate interpretation of the sentence, their expectations about the verb, as indexed by the N400 response, are not immediately sensitive to information about the arguments' structural roles. We examined this possibility in an ERP study in Mandarin Chinese by measuring the effect of argument role reversals when a short 'buffer' is introduced between the arguments

and the verb to allow more time for verb predictions (Chow et al., submitted). Using sentences like (2), we manipulated the linear distance between the arguments and the verb while holding their effect on the verb's cloze probability constant.

(2) Manipulation of linear distance between arguments and verb by varying the position of a temporal adverbial (Chow et al., submitted):

a. Short distance:

zuotian-xiawu {jingcha / xiaotou} ba {xiaotou / jingcha} zhua-le...
yesterday-afternoon {cop / thief} BA {thief / cop} arrest...

b. Long distance:

{jingcha / xiaotou} ba {xiaotou / jingcha} zai zuotian-xiawu zhua-le...
{cop / thief} BA {thief / cop} ZAI yesterday-afternoon arrest...
"The {cop / thief} arrested the {thief / cop} yesterday afternoon."

Consistent with our hypothesis, we found that argument role reversals did elicit an N400 effect at the verb when the verb was further away from its arguments. Further, this online effect was observed only when argument role reversals strongly impacted the verb's (offline) cloze probability. The N400's reemerged sensitivity suggests that information about the arguments' structural roles can impact comprehenders' verb predictions, but that this impact is not immediate.

More recently, convergent evidence has been obtained in an ERP study in Japanese (Momma, Sakai, & Phillips, submitted). Momma and colleagues manipulated the SOA between an argument and a verb in simple two-word Subject-Verb and Object-Verb sentences such as (3) and found that the N400 became sensitive to argument role reversals when the SOA was increased from 800 ms to 1200 ms. By using sentences with just one argument and a verb and directly manipulating the SOA between them, this study demonstrated that the N400's reemerged sensitivity can be attributed to timing in the absence of any other changes in the context. Further, since Japanese (unlike Mandarin Chinese) is a verb-final language and encodes argument role information using case markers, these results suggest that the delayed impact of argument role information on comprehenders' verb predictions is not limited to non-canonical sentence forms or to languages that encode argument role information via word order.

(3) Argument role reversal in simple two-word Japanese sentences (Momma, Sakai & Phillips., submitted):

a. hachi-ga sasu.

bee-NOM sting

"(the) bee stings"

b. hachi-wo sasu.

bee-ACC sting

"(the) bee gets stung"

The idea that argument role information has a delayed impact on comprehenders' predictions is also in line with observations made using other experimental techniques. For example, Kukona, Fang, Aicher, Chen and Magnuson (2011) reported that listeners' anticipatory eye-movements were not immediately sensitive to whether a thematic role had already been filled. Specifically, upon hearing a sentence like "Toby arrests the ...," listeners were equally likely to anticipatorily look at the picture of a likely agent (a police officer) as well as a likely patient (a crook) despite the fact that the agent role had already been filled. If we take anticipatory eye-movements as a measure of listeners' expectations about upcoming materials, these results are consistent with the idea that listeners cannot immediately use information about an argument's role for predictions. Meanwhile, in a recent study, we manipulated the amount of time participants had to respond in a *speeded* completion/cloze task and found that, when placed under time pressure, participants sometimes produced the 'role-reversed' completion

verb. Also, the frequency of role-reversed completions dropped significantly at a longer response latency (Chow, Kurenkov, Buffinton, Kraut, & Phillips, 2015). Taken together, research using different techniques and across multiple languages provides converging evidence that fully informed predictions are not generated instantaneously and that the impact of argument role information may be delayed.

In CSLP, we reported two key additional findings. First, in a direct comparison, we found that offline cloze probability manipulations of the same size had greatly different effects on the N400, depending on what contextual information was manipulated. Specifically, the order of the preverbal arguments (Argument Role reversals, e.g., “The restaurant owner forgot which customer/waitress the waitress/customer had served ...”) strongly modulated offline cloze probability but did not modulate the N400 response at all. In contrast, cloze probability differences of the same size elicited a clear N400 effect when one of the arguments was substituted instead (Argument Substitution, e.g., “The superintendent overheard which tenant/realtor the landlord had evicted...”). This suggests that the arguments’ meaning can impact comprehenders’ verb predictions at a point when information about their structural roles has no discernable effect. Further, we found that the substitution of one of the arguments in a clause elicited an N400 effect even when the substituted argument still appeared elsewhere in the sentence (Same-words Argument Substitution, e.g., “The exterminator inquired which neighbor the landlord had evicted ...” vs. “The neighbor inquired which exterminator the landlord had evicted ...”). This suggests that comprehenders’ verb predictions are selectively sensitive to the arguments and not just to any word in the sentence context.

Taking into consideration these findings along with the relevant evidence reviewed above, our proposal in CSLP makes four main claims: (i) online verb predictions are initially sensitive to only a subset of contextual information, (ii) the meaning of preverbal arguments has a more immediate impact on verb predictions than their structural roles, (iii) comprehenders rely on the clausal structure of a sentence to quickly identify the arguments, and use them as an unordered set (or, a “bag”) of event participants to query their event knowledge and pre-activate likely verbs (e.g., what events are likely given these participants?), and (iv) information about the arguments’ roles is incorporated into comprehenders’ verb predictions at a later time point.

3. Some claims that we are not making

In Section 4 we explore in more detail the mechanisms that might support these proposals. First, however, for the sake of clarification, we now address several claims that are often confused with our own but which we do not intend to make. For each hypothesis below that we fail to endorse, we will state why we think it is implausible or dispreferred.

I. Syntactic parsing is slow.

One interpretation of our findings has been that correct syntactic parsing of the preverbal arguments is slow, and hence that the argument role information is slow to affect comprehenders’ verb predictions. We do not endorse this interpretation because there is ample evidence that syntactic parsing proceeds rapidly, both in the general case, and in the specific case of thematic role assignment. A large body of converging evidence suggests that syntactic parsing is highly incremental (reviewed in Pickering & van Gompel, 2006). For example, reading-time studies on pronoun interpretation in Japanese, a strongly verb-final language, indicate that hierarchical relations are assigned rapidly and appropriately before the verb (Aoshima, Yoshida, & Phillips, 2009). On the specific issue of argument role assignment, eye-tracking and ERP evidence suggest that readers are rapidly sensitive to the grammatical case of an argument (e.g., Kamide, Altmann, & Haywood, 2003; Coulson, King, & Kutas, 1998; Bornkessel, Schlesewsky, & Friederici, 2002). For example, comprehenders may use the case of an argument to anticipate the likelihood of an additional argument being mentioned, immediately following the offset of a critical case marker and prior to encountering the verb

(Kamide et al., 2003). In keeping with findings like these, we believe that participants in the CSLP study had ample time to identify the likely thematic roles of the arguments prior to encountering the verb. This view is further supported by the observation of a P600 effect at the critical verb in our experiments, as we discuss further below.

Our proposal should also not be mistaken for the related idea that participants are slow to compute predictions about the likely syntactic category of upcoming words. Our studies never explicitly manipulated the predictability of the target word's syntactic category, but there is much prior evidence that comprehenders routinely and rapidly generate predictions about the upcoming syntactic category (e.g., Hale, 2001; Staub & Clifton, 2006; Lau, Stroud, Plesch, & Phillips, 2006; Levy, 2008; Boston, Hale, Patil, Kliegl, & Vasishth, 2008; Dikker, Rabagliati, & Pylkkänen, 2009; Smith & Levy, 2013) and other syntactic properties (e.g., Ilkin & Sturt, 2011; Omaki et al., 2015) of upcoming words. In our experiments the target word is always preceded by a syntactically constraining context, and so participants should be able to quickly anticipate that the next word is likely to be a verb across all conditions. Our claims are specifically focused on comprehenders' capacity to pre-activate the semantic representations of verbs.

II. The syntactic parse is fine, but semantic interpretation ignores it.

Another common interpretation of our results is that comprehenders initially interpret thematic relations based on what is most plausible given the semantic information alone, without regard to what the syntactic parse requires for interpretation. Only later do comprehenders integrate the semantic interpretation with the syntactic parse, leading them to detect a mismatch (e.g., Kolk et al., 2003; Hoeks, Stowe, & Doedens, 2004; Kim & Osterhout, 2005; Kuperberg, 2007; Bornkessel-Schlesewsky & Schlewsky, 2008). Under this type of account, the N400 is taken to directly reflect semantic/pragmatic plausibility (e.g., Hagoort, Hald, Bastiaansen, & Petersson, 2004), and its insensitivity to argument role reversals is taken to indicate that comprehenders compute a plausible interpretation in canonical and role-reversed sentences alike. Meanwhile, the presence of a P600 effect is taken to reflect increased processing difficulty when a clash is detected between the syntactic cues and the semantic interpretation that has been computed (e.g., Kolk et al., 2003; Kuperberg, 2007; Kim & Sikos, 2011).

Under this account, the clash should specifically arise when there is a tempting construal of the verb and the arguments that does not fit with the syntactic parse ("semantic attraction"). This in turn implies that the most important processes occur after the verb is encountered, because it is the semantic relation between the verb and its arguments that leads to misinterpretation. We are skeptical of this account for a couple of reasons. First, the characteristic ERP profile is not limited to cases of semantic attraction: Kim and Osterhout (2005) offered compelling evidence that the so-called "semantic P600" effect is specifically tied to semantic attraction, but subsequent studies have not confirmed this (Kolk et al., 2003; Kuperberg, Kreher, Sitnikova, Caplan & Holcomb, 2007; Stroud & Phillips, 2012; Paczynski & Kuperberg, 2012; Chow & Phillips, 2013). Second, if the main action reflects processes that start only after the verb is encountered, then it is unclear why delaying the verb should make a difference to comprehenders' sensitivity, as we and others have found in both Mandarin Chinese (Chow et al., submitted) and Japanese (Momma, Sakai & Phillips, submitted).

III. Information outside the current clause does not influence verb argument predictions.

Our findings in Experiment 2 of CSLP suggested that words outside the current clause do not impact comprehenders' initial verb predictions via lexical association. However, we did not intend to make the stronger claim that predictions are blind to information outside the current clause. There is good evidence against this stronger version. For example, van Berkum Zwitserlood, Hagoort and Brown (2003) showed that the N400 is sensitive to discourse factors outside the target sentence. Many other studies make a similar point. The key finding in Experiment 2 of CSLP is that changing the nouns within a clause affects verb predictions, even when this does not change the set of nouns in the preceding context as a whole. That

suggests that arguments of the verb have a special status, though we have not yet explained how argument information is combined with broader discourse information.

IV. All processes involving thematic roles are slow in language processing.

We do not claim that all processes involving thematic roles in language comprehension are slow. The commentators take us to task for overlooking well-known findings in psycholinguistics about rapid thematic interpretation during ambiguity resolution (e.g., Trueswell, Tanenhaus, & Garnsey, 1994), but we see no conflict. In classic ambiguity resolution studies the focus is on the information that comprehenders can rapidly access after they have been presented with a verb+noun combination. These studies show that, once comprehenders actually encounter a verb, they can rapidly access lots of information about that verb and its constraints. That could include abstract argument roles like “agent” and “patient”, though most evidence is consistent with more verb-specific role information such as “server” or “servee.” In the current debate about prediction, in contrast, the focus is on the inferences that comprehenders can make about verbs that they have not seen yet, based on argument role information that is necessarily verb-independent. These are therefore quite different processes, and so it should not be surprising if they unfold differently.

V. Initial verb prediction is driven by associative priming, with no role for event knowledge.

Kim and colleagues (2015) take issue with our “central conclusion that predictions are based only on unstructured lexical associations” and instead propose that “predictions are mediated by event knowledge.” We are in agreement on this point. We certainly agree that event knowledge plays a key role when using nouns to infer likely upcoming verbs, though it could be that associative relations between nouns and verbs contribute to accessing candidate event representations. In fact, the question of how linguistic cues are used to identify the most relevant event representations in memory is central to our view of the verb prediction process, to which we turn next.

4. Mechanisms: Prediction as Memory Access

Having described the empirical generalizations that we seek to explain, we turn to the mechanisms that could be responsible for these generalizations. At its core, we view prediction as a memory retrieval problem: the current linguistic (and non-linguistic) context is used to access representations in long-term memory, with the goal of inferring the speaker’s message, and hence what words are likely to come next.

If prediction is a memory access problem, then we need to specify the memory architecture, and in any individual situation we need to specify the memory retrieval cues and the format of the memory representations that the cues are matched against. We assume a content-addressable memory (CAM) architecture, in which the items in memory are directly accessed by matching cues to the contents of memory in parallel. This architecture is well motivated based on findings involving retrospective dependency formation in language comprehension (Lewis, Vasishth, & Van Dyke, 2006; McElree, Foraker, & Dyer, 2003; McElree, 2000; Wagers, Lau, & Phillips, 2009), though it is not normally invoked in discussions of prediction. A memory access process is implied by the widespread view that prediction involves using linguistic context to activate relevant concepts, but we encounter less discussion of how this can be implemented in terms of matching retrieval cues to target memories.

In real-world cases of linguistic prediction the contexts are so complex that it is difficult to identify what the memory retrieval cues might be. But in the very simple contexts used in our experiments this is more feasible. In these studies, the contexts consist of nouns (or noun phrases), together with their argument roles, e.g., waitress+agent, customer+patient, and the goal is to identify what verb is likely to follow, given those noun/role combinations. This task requires the comprehender to consult his knowledge of events in the world, to infer on the

basis of the input so far the event(s) that are most likely being described, and perhaps in turn to infer from this the verb or verbs that are most likely imminent.

The empirical generalization that we aim to capture is that comprehenders' expectations about upcoming verbs evolve over time. At an early time point the expectations reflect the contribution of the arguments but not their roles. Only after a delay do comprehenders' expectations reflect the combined constraints of the arguments and their roles.

In terms of retrieval cues and target memory representations, we envision three possible explanations for the delayed effects of argument roles in prediction, each of which could have a number of variants. First, the delay may originate in the initiation of memory retrieval attempts, i.e., retrieval on the basis of noun-only "bag-of-argument" cues may be engaged separately and earlier than retrieval on the basis of composite argument+role cues, giving a head start to the effects of the bag-of-argument cues. For example, the cue [+waitress] might be used for retrieval faster than the cue [+waitress-as-agent]. Second, the delay may originate in the dynamics of the process by which partially matching candidates compete for selection, such that cue weighting effectively changes across the course of the competition. Under both of these possible accounts it is assumed that the retrieval cues are compatible with the format of the target memories, i.e., event memories are encoded in terms that make it possible to access them using cues like waitress+agent and customer+patient. A third possibility is that the delay is due to a mismatch between the format of the retrieval cues and the target memories. Event knowledge surely includes specification of participants in the events. But if those participants are not encoded in terms of abstract grammatical roles such as agent/patient or subject/object, then they cannot be directly accessed using the available retrieval cues, and a more indirect access mechanism is needed.

We do not have strong evidence to favor any one of these three possibilities, but we are sympathetic to the third option (memory cue/target mismatch), as its assumptions seem plausible. We consider each account in a little more detail.

Hypothesis 1: Delayed cue initiation

Under this account, argument+role cues can be directly matched to the format of event memories, but their effects are seen at a delay because the retrieval cues themselves are issued at a delay. This delay could arise if comprehenders access the lexical representation of nouns faster than their argument roles, and if the comprehension system does not wait for the argument roles to be identified before issuing memory retrieval cues. Under this account, there is no principled difficulty in using argument+role cues to directly access relevant verbs or event types. The only problem is at the very start of the process. In some cases it is plausible that argument roles might be identified later than the lexical identity of the arguments, such as when using word order to recover argument roles in English. But in other cases the argument role information is very easy to identify, such as when it is signaled by transparent case markers (Japanese) or by pre-nominal particles (e.g., the Mandarin *ba*-construction). Therefore, even though we cannot rule out this account, we regard it as unlikely as a cause of the general delay in the effects of argument+role cues on prediction.

Hypothesis 2: Delayed cue use

A second possibility is that both bag-of-argument and argument+role cues can be directly matched to the format of event memories, and both kinds of cues are issued without a delay, but for other reasons the argument+role cues have a delayed impact on the outcome of retrieval. One version of this account might hold that the process by which partially-matching candidates compete for selection would be dynamic such that at an early stage the candidates that strongly match the bag-of-argument cues dominate the competition and only later on do the candidates that strongly match the argument+role cues win out. Although this is logically possible, we are not aware of CAM models in which the winners change over time when all

the candidates have to be retrieved from LTM and the retrieval cues are fixed, and it is not clear to us what would cause this to occur here if it is assumed that both kinds of cues can target event memories at the same time and in a qualitatively similar way.

Kuperberg's (2016) suggested explanation of our results might be understood as an interesting variant on this account. Kuperberg's proposal is framed in different terms than ours. Her Bayesian account is stated at a computational level (in the terms of Marr, 1982), whereas ours is presented in more algorithmic terms. And her account focuses on the ineffectiveness of argument+role cues that we highlighted in CSLP, with less attention to the additional timing dimension that we have emphasized here. Her account assumes that argument+role cues can be directly used to identify relevant event memories and that there is no problem, in principle, with using composite argument+role cues to predict upcoming verbs. Kuperberg also assumes that comprehenders can generate predictions using more fine-grained or more coarse-grained cues, i.e., in addition to the argument+role cues, they can also predict verbs using bag-of-argument cues alone.

A distinctive property of Kuperberg's proposal is that, in the face of uncertainty, cues might be weighted according to their reliability, based on degree of prior experience. For example, one might imagine that a comprehender has encountered many events involving waitresses and customers, and most of them involved serving (making the bag-of-arguments cue reliable) and that they have encountered many events involving waitress-as-agent, customer-as-patient, and most of them involved serving (making the argument+role cue reliable in the canonical condition). However, they may not have encountered as many events involving customer-as-agent and waitress-as-patient, and those events may have been quite heterogeneous (criticizing, paying, etc.), making the argument+role cue relatively less reliable in the role-reversed condition. According to Kuperberg's account, this would result in stronger weighting for the bag-of-argument cue than the argument+role cue in the role-reversed condition, thus correctly predicting the absence of an N400 difference between the canonical and role-reversed conditions. It is less clear how this account explains the late re-emergence of argument role effects. Kuperberg seems to briefly suggest the possibility of a mechanism more akin to the delayed cue initiation account, where certainty about the argument+role cue might increase over time and accordingly be assigned a greater weight. However, as discussed above, questions remain about why cues that are less informative at one moment should become more informative at a later point in time.

Kuperberg's multi-level prediction account appears to be computationally intensive, and it does not specify how predictions are computed at each level of detail. But it makes an interesting prediction about the findings in CSLP. Kuperberg predicts that the N400's insensitivity to argument role reversals in this study is an artifact of the specific experimental materials that we used. Roughly, if we take contextual constraint as a proxy for the reliability of argument+role cues, this account would suggest that the sentence contexts were on average less constraining in the low cloze (role-reversed) condition than in the high cloze (canonical) condition, therefore leading participants to rely more on the bag-of-argument cues in the role-reversed condition. However, this is not confirmed by our analysis of the cloze probability of the most likely verb completion ("MaxCloze") in the CSLP materials, which reveals a very similar MaxCloze profile in argument role reversal sentences (high cloze: 28.6% vs. low cloze: 23.8%) and in argument substitution sentences (high cloze: 29.3% vs. low cloze: 22.6%). These are not the cloze values of the actual verbs presented in our study; these values reflect how constraining the contexts used in each condition are. This challenges Kuperberg's account of our findings.

Hypothesis 3: Format of event memory

The third possibility is that there is a principled difficulty in directly matching composite argument+role cues with items in event memory, because information like x-as-agent or y-as-patient, where x and y represent a specific participant, are not featurally encoded in event

memory. Under this account, only the bag-of-argument cues can participate in the content-addressable retrieval procedure. In order to selectively retrieve event types or verbs that take a particular noun as an agent vs. a patient, some indirect and likely less rapid access mechanism, such as a search mechanism (Sternberg, 1975) must be invoked in addition.

In the list memory literature, the canonical example of information thought to require a search-like mechanism is relational information, such as temporal (McElree & Doshier, 1993) or positional information (Gronlund, Edwards, & Ohrt, 1997). If argument role information is encoded relationally in event memories, analogous to how positional/temporal information is encoded in list memory, this might then provide an explanation for why this information appears to have a delayed impact on predictions. For instance, an event memory representation of a *servicing* event might consist of stereotypical participants like a waitress and customer, but the role of these participants might not be coded featurally (e.g., waitress-as-agent, customer-as-patient), but rather relationally. Or according to a related variant, roles in the event memory representation might be encoded in more event-specific terms (e.g., waitress-as-server, customer-as-servee), such that they cannot be targeted for retrieval by general argument+role cues like waitress-as-agent. Either way, targeting this relational coding of argument roles might then require a search-like mechanism that follows an initial content-based retrieval of candidates based on bag-of-argument cues.

Under this account, identification of suitable verbs relies on a two-step process. Initial retrieval of candidate verbs is based on bag-of-argument cues alone, e.g., the nouns are used as cues to retrieve event memories containing both customers and waitresses, irrespective of their roles. These candidate event memories are subsequently evaluated, perhaps sequentially, to identify whether they contain appropriate participant role bindings. Kuperberg (2016) characterizes this as an account in which we attribute inherently distinct time courses to distinct processing mechanisms, but we disagree. At least, the timing differences are not arbitrary. Rather, we claim that the distinct time courses are a consequence of the (mis)match between linguistic cues and the format of target memories, and the access processes that this entails.

We do not have definitive evidence for this account, but we are sympathetic to this option, for a couple of reasons. First, this account does not presuppose that event knowledge is stored in a format that conveniently matches the abstract argument+role information that can be extracted from linguistic input. Second, the effect of argument role seems to be markedly slow, taking more than 800ms and up to 1200ms even with only one pre-verbal argument (Momma, Sakai & Phillips, submitted). It seems to us unlikely that a slight temporal privilege for lexical access over argument role identification (account #1 above) or a temporary lack of certainty about the prior context (Kuperberg's suggestion) could explain such a long delay. It is thus tempting to speculate that argument role information can only be used via a serial or near-serial (i.e., limited parallel) search-like retrieval procedure, which could be quite time-consuming given multiple candidates.

5. Conclusion and Future Directions

We appreciate the opportunity to engage with the commentaries on our paper, and we hope to have clarified here the points on which we agree and disagree with Kuperberg (2016) and with Kim and colleagues (Kim et al., 2015). Our work on this topic was inspired by discoveries made by Kuperberg, Kim, Kolk, and others around 10 years ago. All agree that comprehenders are in some way failing to make full use of argument role information, and all agree that this provides valuable clues about how we rapidly interpret sentences. But newer findings have changed the way that we all think about these problems. Kuperberg's exploration of predictive mechanisms in terms of a Bayesian model of inference is very different from the way that she approached the problem in earlier work. And our own thinking about the problem in terms of memory access mechanisms is rather different than how we first approached this topic. This progress is an encouraging sign. We are also very glad that the discussion is focused more

on underlying language processing mechanisms than on ERP componentry. This is helped by the existence of related findings from other experimental measures (eye-tracking: Kukona et al., 2011; speeded cloze: Chow et al., 2015).

The main differences between our account and the commentators are in the reasons why argument role information is not fully exploited. Kuperberg and Kim and colleagues claim that argument+role cues are, in principle, highly effective cues for predicting upcoming verbs. They do not spell out in detail how these cues can be used, but they assume that ample resources and time are available, and that information is encoded in an easily accessible format. In contrast, we argue that the argument+role cues are only indirectly useful, because they are less effective memory cues, perhaps because the abstract linguistic cues mismatch with the format of the event memories that they are used to access. Put simply, Kuperberg and Kim and colleagues assume that inappropriate verbs are predicted because they are tempting, whereas we assume that inappropriate verbs are predicted because they are a step on the path to identifying the appropriate verbs. Our account is guided by (i) additional evidence from timing, which comes from studies other than CSLP, (ii) analyses of the full set of experimental materials in CSLP, which shows that the linguistic contexts used in each condition were fairly similar in terms of how constraining they are; and (iii) speculation about memory access mechanisms and the format of event memories.

This debate depends on some assumptions about things that we currently know too little about. We made assumptions about the format of event memories, but have little direct evidence for that. Our assumptions about memory access mechanisms are guided by years of thinking how recent linguistic material is accessed in memory, but we cannot be sure that long-term event memories are accessed in a similar fashion. Meanwhile, Kuperberg and Kim and colleagues make assumptions about the availability of powerful inference mechanisms that we have only limited evidence for. And we share with the commentators the assumption that all usable predictive cues are used as soon as they are recognized, though we have little evidence about the timing of retrieval operations (as opposed to the timing of when information is retrieved). We have also assumed that predictive mechanisms are engaged in a uniform fashion, and so more needs to be said to capture emerging evidence that predictions are compromised in older adults (Wlotko & Federmeier, 2012b; Huang, Meyer & Federmeier, 2012) or in second language learners (Grüter, Lew-Williams, & Fernald, 2012; Martin et al., 2013). And while the focus of the current debate is on argument roles, we have little information on how other types of abstract grammatical constraint contribute to lexical predictions.

At a more general level, we believe that one of the main morals of this debate is that we can learn at least as much about predictive mechanisms in language when they fail as when they succeed.

Acknowledgments

This work was supported in part by an NSF NRT award to the University of Maryland (DGE-1449815, Phillips, PI) and by an NSF Graduate Research Fellowship to CS (DGE-1144245).

6. References

- Aoshima, S., Yoshida, M., & Phillips, C. (2009). Incremental processing of coreference and binding in Japanese. *Syntax*, 12(2), 93–134.
- Bock, K., Loebell, H., & Morey, R. (1992). From conceptual roles to structural relations: bridging the syntactic cleft. *Psychological Review*, 99(1), 150.
- Bornkessel, I., Schlesewsky, M., & Friederici, A. D. (2002). Grammar overrides frequency: Evidence from the online processing of flexible word order. *Cognition*, 85(2), B21-B30.
- Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2008). An alternative perspective on “semantic P600” effects in language comprehension. *Brain Research Review*, 59, 55–73.

- Boston, M. F., Hale, J. T., Patil, U., Kliegl, R., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, 2(1), 1-12.
- 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.
- Carnie, A. (2012). *Syntax: a generative introduction*, 3rd edition. Wiley.
- Chow, W. Y. (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., & Kurenkov, I., Buffinton, J., Kraut, B., & Phillips, C. (2015). How predictions change over time: evidence from an online cloze paradigm. Poster presented at the 28th annual CUNY Human Sentence Processing Conference, Los Angeles, CA.
- Chow, W., Smith, C., Lau, E., & Phillips, C. (2015). A ‘bag-of-arguments’ mechanism for initial verb predictions. *Language, Cognition & Neuroscience*, Advance online publication. doi:10.1080/23273798.2015.1066832
- Chow, W. Y., Wang, S., Lau, E. & Phillips, C. (submitted). Wait a second! Delayed impact of argument roles on on-line verb prediction.
- Coulson, S., King, J. W., & Kutas, M. (1998). Expect the unexpected: Event-related brain response to morphosyntactic violations. *Language and cognitive processes*, 13(1), 21-58.
- DeLong, K. A., Quante, L., & Kutas, M. (2014). Predictability, plausibility, and two late ERP positivities during written sentence comprehension. *Neuropsychologia*, 61, 150-162.
- Dikker, S., Rabagliati, H., Pykkänen, L. (2009). Sensitivity to syntax in visual cortex. *Cognition*, 110(3), 293-321.
- Gronlund, S. D., Edwards, M. B., & Ohrt, D. D. (1997). Comparison of the retrieval of item versus spatial position information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(5), 1261.
- Grüter, T., Lew-Williams, C., & Fernald, A. (2012). Grammatical gender in L2: A production or a real-time problem. *Second Language Research*, 28, 191-215.
- Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K.M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science*, 304, 438-441.
- Hale, J. (2001). A probabilistic early parser as a psycholinguistic model. *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics* (pp. 159–166). Pittsburgh, PA.
- Hoeks, J. C. J., Stowe, L. A., Doedens, G. (2004). Seeing words in context: the interaction of lexical and sentence level information during reading. *Cognitive Brain Research*, 19, 59-73.
- Huang, H. W., Meyer, A. M., & Federmeier, K. D. (2012). A “concrete view” of aging: Event related potentials reveal age-related changes in basic integrative processes in language. *Neuropsychologia*, 50(1), 26-35.
- Ilkin, Z., & Sturt, P. (2011). Active prediction of syntactic information during sentence processing. *Dialogue and Discourse*, 2(1), 35-58.
- 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. (2015). Prediction during sentence comprehension is more than a sum of lexical associations: the role of event knowledge. *Language, Cognition and Neuroscience*.
- 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.
- Kim, A. & Sikos, L. (2011). Conflict and surrender during sentence processing: An ERP study of syntax-semantics interaction. *Brain and Language*, 118, 15-22.

- Kolk, H.H.J., Chwilla, D.J., van Herten, M., & Oor, P. (2003). Structure and limited capacity in verbal working memory: a study with event-related potentials. *Brain and Language*, 85, 1-36.
- Kukona, A., Fang, S., Aicher, K. A., Chen, H., & Magnuson, J.S. (2011). The time course of anticipatory constraint integration. *Cognition*, 119, 23-42.
- Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research*, 1146, 23-49.
- 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. (2016). Separate streams or probabilistic inference? What the N400 can tell us about the comprehension of events. *Language, Cognition and Neuroscience*.
- 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.
- Kutas, M., Van Petten, C., & Kluender, R. (2006). Psycholinguistics electrified II (1994–2005). In M. Traxler & M. Gernsbacher (Eds.), *Handbook of psycholinguistics* (2nd edition, pp.659–724). Cambridge, MA: Academic Press.
- Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307,, 161–163.
- Lau, E. F., Stroud, C., Plesch, S., & Phillips, C. (2006). The role of structural prediction in rapid syntactic analysis. *Brain and Language*, 98, 74-88.
- Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106, 1126–1177.
- Lewis, R. L., Vasishth, S., & Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, 10(10), 447-454.
- Marr, D. (1982). *Vision*. Cambridge, MA: MIT Press.
- Martin, C. D., Thierry, G., Kuipers, J. R., Boutonnet, B., Foucart, A., Costa, A. (2013). Bilinguals reading in their second language do not predict upcoming words as native readers do. *Journal of Memory and Language*, 69, 574–588.
- McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of Psycholinguistic Research*, 29(2), 111-123.
- McElree, B., & Doshier, B. A. (1993). Serial recovery processes in the recovery of order information. *Journal of Experimental Psychology: General*, 122(3), 291–315.
- McElree, B., Foraker, S., & Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of Memory and Language*, 48(1), 67-91.
- Momma, S., Sakai, H., & Phillips, C. (submitted). Give me several hundred more milliseconds: Temporal dynamics of verb prediction.
- Momma, S., Slevc, R., & Phillips, C. (submitted). Unaccusativity in sentence production.
- Oishi, H., & Sakamoto, T., (2010). Immediate interaction between syntactic and semantic outputs: evidence from event-related potentials in Japanese sentence processing. Poster presented at the 22nd annual CUNY Human Sentence Processing Conference, Davis, CA.
- Omaki, A., Lau, E. F., White, I. D., Dakan, M. L., Apple, A., & Phillips, C. (2015). Hyperactive gap filling. *Frontiers in psychology*, 6.
- Otten, M., Nieuwland, M. S., & Van Berkum, J. J. (2007). Great expectations: Specific lexical anticipation influences the processing of spoken language. *BMC neuroscience*, 8(1), 1.
- Paczynski, M., & Kuperberg, G. R. (2012). Multiple influences of semantic memory on sentence processing: distinct effects of semantic relatedness on violations of real-world event/state knowledge and animacy selection restrictions. *Journal of Memory and Language*, 67(4), 426-448.
- Pickering, M. J., & van Gompel, R. P. G. (2006). Syntactic parsing. In M. Traxler & M. Gernsbacher (Eds.), *Handbook of psycholinguistics* (2nd edition, pp. 455-503). Academic Press, London.
- Sanford, A. J., Leuthold, H., Bohan, J., & Sanford, A. J. (2011). Anomalies at the borderline of awareness: An ERP study. *Journal of Cognitive Neuroscience*, 23(3), 514-523.

- Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128, 302–319.
- Staub, A., & Clifton, C., Jr. (2006). Syntactic prediction in language comprehension: Evidence from *either...or*. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32, 425-436.
- Sternberg, S. (1975). Memory scanning: New findings and current controversies. *The Quarterly journal of experimental psychology*, 27(1), 1-32.
- Stroud, C., & Phillips, C. (2012). Examining the evidence for an independent semantic analyzer: An ERP study in Spanish. *Brain and language*, 120(2), 108-126.
- Trueswell, J., Tanenhaus, M., & Garnsey, S. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, 33, 285-318.
- Van Berkum, J. J., Zwitserlood, P., Hagoort, P., & Brown, C. M. (2003). When and how do listeners relate a sentence to the wider discourse? Evidence from the N400 effect. *Cognitive Brain Research*, 17(3), 701-718.
- Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, 83(2), 176-190.
- Van de Meerendonk, N., Kolk, H. H., Chwilla, D. J., & Vissers, C. T. W. (2009). Monitoring in language perception. *Language and Linguistics Compass*, 3(5), 1211-1224.
- Wagers, M. W., Lau, E. F., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, 61(2), 206-237.
- Wlotko, E. W., & Federmeier, K. D. (2012a). So that's what you meant! Event-related potentials reveal multiple aspects of context use during construction of message-level meaning. *NeuroImage*, 62(1), 356-366.
- Wlotko, E. W., & Federmeier, K. D. (2012b). Age-related changes in the impact of contextual strength on multiple aspects of sentence comprehension. *Psychophysiology*, 49(6), 770-785.