Theories All The Way Down

Colin Phillips, Phoebe Gaston, Nick Huang, Hanna Muller
{colin, pgaston, znhuang, hmuller}@umd.edu
University of Maryland

Abstract

It is common in linguistics to draw a contrast between “theoretical” and “experimental” research, following terminology used in other fields of science. Researchers who pursue experimental research are often asked about the theoretical consequences of their work. Such questions generally equate “theoretical” with theories at a specific high level of abstraction, guided by the questions of traditional linguistic theory. These theories focus on the structural representation of sentences in terms of discrete units, without regard to order, time, finer-grained memory encoding, or the neural circuitry that supports linguistic computation. But there is little need for the high level descriptions to have privileged status. There are interesting theoretical questions at all levels of analysis. A common experience in our group’s work is that we embark on a project guided by its apparent relevance to high-level theoretical debates. We then discover that this relevance depends on linking assumptions that are not as straightforward as we initially thought. And then we discover new theoretical questions at lower levels of analysis that we had not even been aware of previously. We illustrate this using examples from many different lines of experimental research on syntactic issues.

Introduction

We expect both too much and too little from experiments in linguistics, including the recent wave of interest in ‘experimental syntax’. We often encounter the hope that experiments will give us more precise data that will allow us to settle difficult theoretical questions, but such hopes are rarely realized. We believe that this is because we have unrealistic expectations about the ability of experiments to answer questions that syntacticians already had. Meanwhile, researchers have underappreciated the value of experiments for allowing us to address new questions that were not even on our radar previously.

Linguistic theories that are constructed based on traditional data collection methods, i.e., yes/no acceptability judgments, unsurprisingly make claims that are well suited to those methods. For example, they make claims about sentences that are well-formed and ill-formed, based on properties of their structural organization, typically with no reference to how those mental representations are constructed. This does not mean that only those claims count as ‘theories’. Nor does it mean that other data collection methods are of ‘theoretical’ interest only if they address those existing claims. In order for linguistic theories to most benefit from experimental

research, it is important to take an inclusive approach to what counts as a linguistic theory and what counts as a theoretically interesting contribution.

We have had a similar experience, repeatedly, in multiple projects, spanning many years. We have been attracted to explore a topic because of its purported impact for 'theoretical' linguistics, typically because some experimental finding bears on a generalization or claim that we are fond of. This could be a specific generalization about how constraints on anaphora are represented, or a broad generalization about syntax-semantics relations.

When we start to explore, we find that the experiment that initially drew our interest does not wear its interpretation on its sleeve. This is because understanding the conclusions of the study depends on a 'linking hypothesis'. A linking hypothesis is a theory of the experimental task that connects mental linguistic operations, i.e., the things that we really care about, to observed experimental measures such as button presses, eye-movements, scalp voltages, etc. Once we better understand the linking hypothesis, we often realize that the theoretical consequences are not as decisive as we first thought, because of additional assumptions that we had been unaware of. Armed with a clearer linking hypothesis, we often also realize that there were confounds in the experimental set-up. Once those confounds are addressed, we then often find that the theoretical conclusions are different than where we started.

Importantly, once we articulate a clearer linking hypothesis we often find that it includes interesting claims about linguistic computations, often at a more fine-grained level of analysis than we are used to thinking about. And those computations often become fruitful research themes in their own right. These are theoretically interesting. They're just not the theoretical questions that we started with.

Some of our previous work has, perhaps correctly, been seen as painting a rather negative picture of the contributions that experiments can make to questions about grammatical theory, including debates about the representation of filler-gap dependencies (Phillips & Wagers 2007) or disputes over the licensing of ellipsis (Phillips & Parker 2014). In these and other cases we have argued, for example, that timing data is of limited use for deciding among theories that make no clear timing predictions.

But the negative stance in those cases was because the focus was on how experiments bear on the traditional questions asked by theoretical syntax. We can assure the reader that we do not spend our time feeling miserable about the theoretical irrelevance of our research. On the contrary, we enjoy discovering many new theoretical questions that we weren't previously aware of, and these then become research focus topics in their own right.

This scenario has played out repeatedly in our work, and we describe a few examples here. We start with cases involving linking hypotheses that are close to standard linking hypotheses in syntax, and we then move to cases that are further afield theoretically.

Part 1: Acceptability and Well-formedness

Although we most often think about linking hypotheses in the context of sophisticated experimental methods, they are just as relevant for the simplest kinds of linguistic data, i.e., acceptability judgments. Already in this domain we have found that by looking closely at challenges to standard assumptions we have uncovered interesting new questions.

Standard practice in linguistics combines one very simple linking hypothesis with one rather more obscure linking hypothesis. We all know that acceptability judgments are not a transparent reflection of grammatical well-formedness -- the unacceptability of double center-embeddings is the parade case of this. But most of the time we assume a simple link from acceptability to well-formedness: if a sentence sounds fine, then it corresponds to a well-formed representation. At the same time, acceptability judgments are necessarily filtered through the language comprehension system, and most syntacticians are adamant that the structure-building mechanisms that they describe are quite different than the structure-building mechanisms that are invoked in language comprehension. This means that there must be a relatively complex link between real-time comprehension processes and acceptability judgments, one that is almost never spelled out in detail.

A starting point for some of our group's work was an attempt to question the standard disconnect between comprehension processes and acceptability judgments. The link would be simpler if grammatical derivations and the operations of the comprehension system are one and the same thing, aside from the uncertainty that is specific to the comprehension task.

We reasoned that a transparent link between structure building processes in comprehension and grammatical derivations should predict a straightforward alignment between the representations that are entertained during comprehension and the representations that are judged acceptable in off-line judgment tasks. This motivated a research program looking at the real-time status of various well-known grammatical constraints. Many studies did indeed reveal a close alignment between what speakers find acceptable in untimed judgment tasks and the possibilities that they entertain in real-time processes (e.g., Stowe 1986; Sturt 2003; Phillips 2006; Kazanina et al. 2007). But many other studies did not, and those mismatches turned out to be rather more interesting (Lewis & Phillips 2015).

Grammatical Illusions

Standard syntactic reasoning relies on contrasts in acceptability in minimal pairs -- we conclude that a difference in acceptability arises due to a difference in grammaticality, unless other explanations for unacceptability can be identified (such as with center embedding). With this simple linking hypothesis, we can observe that some speakers find (2) intuitively more acceptable than (1) (from Bock & Miller 1991), suggesting a contrast in the grammatical status of the two.

(1) * The key to the cabinet are rusty.
(2) * The key to the cabinets are rusty.

Accordingly, we find claims in the formal syntax literature that (some) sentences that show subject-verb mismatches are in fact grammatical for certain dialects of American English (e.g. Kimball and Aissen 1971, Baker 2008). If this contrast truly is a grammatical contrast, this would have important implications for subject-verb agreement more generally (since c-command appears to be irrelevant to the dependency).

However, the judgment that (2) is more acceptable than (1) is somewhat fragile. Measures that tap into earlier representations, such as reading times and speeded judgments, reveal a greater contrast than slower measures that require more careful judgment. In other words, the more you think about it, the worse (2) sounds. The discovery that this pattern of acceptability is in fact geographically widespread (i.e., not dialect-dependent) and that it aligns closely with well-documented production errors, led to reconceptualizing these sentences as a parser-grammar mismatch, rather than a grammatical phenomenon. That is, the grammar rules out both sentences, but a contrast arises because of an illusion of acceptability. This kind of mismatch has since been documented in many different areas -- negative polarity item (NPI) licensing, comparatives, argument roles -- and the study of the specific types of dependencies that give rise to illusions, compared to the many dependencies that the parser computes accurately, has become a fruitful line of research (Phillips et al. 2011; Dillon et al. 2013).

Once we think of these sentences as a parser-grammar mismatch, the question becomes not what the grammatical representation of the sentence is, but what processing mechanisms are used to generate and access that representation. These are interesting questions in sentence processing, but they are also linking assumptions for all of syntax.

Although agreement attraction, and illusions more broadly, can be studied with many different tools, we will use self-paced reading as a case study. In this task, participants read a sentence that appears one word at a time (all other words appear as dashes) and the participant controls the rate of progression from one word to the next with key presses. Each press advances the presentation by one word. Using this method, we can straightforwardly measure the reading time for each word. If we accept the linking assumption that detecting when the comprehender notices a problem in the sentence (such as ungrammaticality) slows reading times, then we should see slow-downs in sentences (1) and (2) at the word *are,* which is the point at which the ungrammaticality should be detectable. And, in fact, we do see slower reading times for both of these, compared to a grammatical baseline. Further, there's typically a greater slow-down for sentence (1) compared to sentence (2). Given our linking assumption, this indicates that the ungrammaticality of (2) is less detectable than the ungrammaticality of (1).

This leads us to ask what are the component processes of detecting ungrammaticality. The problem for the parser is one of matching the input (the word *are*) with the representation of the

prior input (*the key to the cabinet(s)*), and determining whether they match in the appropriate way, presumably by consulting the grammar.

The various explanations for illusions essentially differ in where in this process of detecting ungrammaticality the blame lies. One class of hypothesis posits that the representation that is generated for the prior linguistic information is defective in some way, such as in the way that features are represented on nodes of the tree. If the plural feature on *cabinets* is occasionally permitted to spread to the DP *the key to the cabinets*, and if we accept the linking assumption that detecting grammaticality is essentially a feature-matching process, then consulting this defective representation should sometimes yield the incorrect decision that the sentence is grammatical, leading to faster reading times (Pearlmutter et al. 1997; Eberhard et al. 2005; Hammerly et al. 2019). The details of this feature spreading could be spelled out in various ways.

Other work has led to the suggestion that the representations generated are in fact perfectly accurate, and the problem arises with the processes by which the representations are accessed in memory (Wagers et al. 2009). Importantly, the feature spreading account also assumes a memory retrieval process, but the consistent success of that process is an unspecified linking assumption for this theory.

The memory retrieval account points to independent research suggesting that retrieval relies on parallel cue-based activation of nodes in memory (McElree 2006; Jonides et al. 2008). This mechanism could lead to retrieval of the wrong part of the representation (i.e., *cabinets* instead of *key*) on some proportion of trials because of a partial match in retrieval cues. That is, if the retrieval cues are [subject] and [plural], then *cabinets* and *key* should each match exactly one cue. If the wrong node is accessed, this could also lead to the incorrect decision regarding the grammaticality of *are* in this position, and this incorrect decision leads to faster reading times. This explanation is motivated by the grammatical asymmetry that is often observed in agreement attraction studies: an illusion of grammaticality arises for (3), but no illusion of ungrammaticality arises for (4).

(3) * The key to the cabinets are rusty.
(4) The key to the cabinets is rusty.

The memory retrieval account naturally accounts for this contrast because a search initiated by *is* in (4) should match *key* on all features and it should match *cabinets* on none.

However, recent work has suggested that the grammatical asymmetry can be better accounted for by properties of the decision-making process itself, rather than by retrieval errors. Both the feature spreading account and the memory retrieval account typically leave unspecified this part of the linking hypothesis, and assume a trivial decision procedure: if the accessed information is consistent with the grammar, accept it and continue to the next word; if it is inconsistent with the grammar, slow down and reanalyze. Hammerly et al. (2019) show that in two-alternative

forced-choice experiments response bias towards acceptance could be responsible for the observed grammatical asymmetry. When they create scenarios where participants expect to see a high proportion of ungrammatical sentences, illusions of ungrammaticality are observed. If this reasoning carries over to the grammatical asymmetry observed in other dependent measures, such as reading times, this could neutralize the advantage of the memory retrieval account over the feature spreading account.

Similar insights have come from research on other grammatical illusions. An influential early proposal was that a number of different types of illusion could be subsumed under the same memory retrieval framework, with illusions understood as instances of mis-retrieval due to partially matching retrieval cues (Lewis et al. 2006; Phillips et al. 2011). For example, illusory licensing of NPIs could be understood as mis-retrieval of an inappropriate negative element in the same way that agreement attraction can be understood as mis-retrieval of an inappropriate number-marked noun (Vasishth et al. 2008). However, subsequent research has revealed that NPI illusions have a different temporal profile than agreement attraction (Parker & Phillips 2016) and that they have rather specific triggers (Muller et al. 2019). This has led to a new set of hypotheses and questions about the time course of semantic interpretation.

Thus, illusions have prompted investigation into several components of the linking hypothesis that underlies the use of acceptability judgments in syntax, including the nature of stored representations of linguistic input, the retrieval process by which they are accessed, and the process by which a decision regarding acceptability is reached. These processes are important as linking hypotheses, but they have also led to productive new avenues for research.

Resumptive Pronouns

The relation between acceptability and grammaticality has also come under scrutiny in the resumptive pronouns (RPs) literature. A conventional claim found in generative syntax since Ross (1967) is that wh-movement out of an island is ungrammatical, but the representation can be "repaired," with a resumptive pronoun (RP) in the place of the gap. This claim was founded on informal introspection. In English, for example, sentences like (6) are reported to be more acceptable than (5). Similar dependencies with RPs are attested in naturally-occurring contexts across many languages. In languages like Hebrew and Irish, RPs can even occur in non-island contexts.

(5) *Who did Mary meet the people that will fire __ ? (wh-movement from relative clause island)
(6) Who_1 did Mary meet the people that will fire him_1? (resumptive pronoun inside relative clause)

Recent research has questioned some of the above claims. In a number of languages, behavioral measures have not consistently reproduced the contrast reported (5) and (6). This is surprising, since informal linguistic judgments are typically highly replicable in large scale acceptability judgment studies (Phillips 2009; Sprouse & Almeida 2012; Sprouse et al. 2013).

Consequently, resumptive pronouns are challenging standard assumptions about the relationship between acceptability and well-formedness.

It appears that judgments of sentences involving RPs are impacted by the choice of task. Studies using scalar acceptability ratings do not consistently find a contrast between RPs and their gapped counterparts (Alexopoulou & Keller 2007; Heestand et al. 2011; Beltrama & Xiang 2016, Chacón 2015, among others). These findings have provided support for the claim that RPs in island configurations are actually ill-formed. In contrast, studies that have used a forced choice task more often demonstrate a difference. For example, production studies have found that when speakers have a choice of ending an island-violating wh-dependency with a gap or an RP in naturally-occurring or experimental contexts, native speakers (implicitly) choose the RP option (Zukowski & Larsen 2004; Ferreira & Swets 2005). Ackerman et al. (2018) reported that when presented with a choice between an island violation with and without an RP, native speakers prefer the sentence with an RP.

Even the choice of scalar rating may make a difference. Beltrama & Xiang (2016) used both acceptability and comprehensibility ratings for sentences with gaps and RPs. They replicated earlier findings that the presence of an RP does not improve acceptability. But they found that the RP does improve ratings of comprehensibility. They conclude that the previously reported benefits of RPs do not reflect rescuing of the grammatical well-formedness of the sentence, but simply the ability to recover a meaning for the sentence.

We take the studies on RPs to show that even simple experimental measures do not wear their interpretation on their sleeve. It is sometimes presumed that “experimental syntax” will deliver clarity to linguistics simply by gathering large quantities of scalar acceptability judgments (Ferreira 2005; Gibson & Fedorenko 2013). This should overcome the biases that surely plague decades of informal introspection by linguists. But in the case of RPs we now know that different quantitative measures point to different conclusions about the status of RPs. Simply asking lots of people does not help in this case. We also consider it unlikely that Beltrama and Xiang’s comprehensibility proposal is correct. One of the things that linguists are rather good at is distinguishing percepts of well-formedness from plausibility and comprehensibility.

We tentatively suggest that scalar acceptability ratings can be a blunt tool, especially when a rating for an entire sentence is used as a proxy for the status of one specific piece of that sentence, such as a RP. Forced choice tasks may yield greater sensitivity in this case because they direct speakers’ attention to the one part of the sentence that differs between the alternatives. However, rather more work is needed to understand why large scale rating studies, which in the vast majority of cases corroborate the results of introspective judgments, sometimes do not yield results that match linguists’ intuitions.

Part 2: Clarifying Linking Hypotheses

In this section we describe case studies involving methods that are further from standard acceptability judgments, and that require linking hypotheses in additional domains. These are scenarios in which care must be taken to rule out any confounds in the many steps between the question and the experimental data. Often we start with a question that is guided by questions from standard (high-level) linguistic theory, but once we spell out the linking hypotheses we find that we are led to different conclusions, and discover interesting new theoretical questions that we have been unaware of previously.

Children's Interpretation of Pronouns

One of the experimental findings that has most captured the imagination of researchers in (traditional) linguistic theory involves preschoolers' mastery of Principle B of the binding theory. Our group's original interest in this topic was motivated by the claim that developmental dissociations can help to decide among competing high level theories. But as we looked more closely, we discovered new theoretical questions at a finer grain of detail that we had not been aware of previously.

Classic versions of the binding theory due to Chomsky (e.g., Chomsky 1981) treated instances of coreference and bound variable anaphora equivalently, whereas Reinhart (1983) and others argued that binding constraints apply to bound variable relations but not to coreference. So, if we focus on binding Principle B, which blocks a pronoun from being bound by a co-argument, Chomsky's account treats (7-8) as equivalent. The subject NP cannot bind the object pronoun in either case, and for the same reason. Reinhart's account regards quantificational (8) as straightforwardly excluded by Principle B, because it clearly involves bound variable anaphora. But additional machinery is needed to capture (7). If (7) is treated as an instance of coreference, then it should not be subject to Principle B. Yet speakers of English clearly perceive that *Mama Bear* and *her* cannot be the same individual. So Reinhart needed to invoke an additional pragmatic constraint that forces instances of possible coreference to be treated as bound variable anaphora, all other things being equal.

(7) Mama Bear washed her.
(8) Every bear washed her.

Reinhart's theory received a significant boost from the finding that preschoolers appear to treat (7) and (8) differently. Chien & Wexler (1990) replicated earlier findings that preschoolers often entertain interpretations of (7) in which *her* refers to Mama Bear. But they found that the same group of children did not allow *her* in (8) to be bound by *every bear*. Children of this age are independently known to have difficulty in some areas of pragmatics, so the results fit remarkably well with the idea that these children are following Reinhart's Principle B. Further studies

provided further evidence for the so-called *quantificational asymmetry*. (McDaniel et al. 1990; Philip & Coopmans 1996; Thornton & Wexler 1999).

The key linking assumption behind these studies was that children consider sentence interpretations that are allowed by their grammar, and that they do not consider sentence interpretations that their grammar disallows. This seems like a reasonable starting assumption, but it presumes a tight and effective link between grammar and interpretation (Crain & Thornton 1998)

One of us swooned when, as a graduate student in the 1990s, he learned of this developmental dissociation. Experimental evidence could turn up surprising evidence that cut through difficult theoretical disputes. So the appearance of Elbourne (2005), which argued that the quantificational asymmetry reflected an experimental confound, was not welcomed. He recruited some students to help try to respond to Elbourne's concerns, fully expecting to show that the concerns were unfounded. The upshot of this is that they found that Elbourne was at least partly correct (Conroy et al. 2009).

In a series of truth-value judgment tasks Conroy et al. (2009) went to great lengths to provide a matched test of children's interpretation of pronouns with quantificational and referential antecedents. When they did this, two key findings emerged. First, as Elbourne had predicted, the quantificational asymmetry disappeared. Second, children performed rather well across-the-board, giving 85-90% adultlike judgments. Set against the many studies that had demonstrated children's non-adultlike interpretations for sentences like (7) this seemed puzzling, but further investigation revealed a more interesting picture.

A review of over 30 prior studies with children revealed that the findings of Conroy et al. were not unprecedented. In particular, they were rather similar to a largely forgotten contemporary of the famous Chien & Wexler (1990) study (Kaufman 1988). But they lay on one end of a wide range of performance by the children in different studies. Some studies showed a quantificational asymmetry, but just as many did not (Lombardi & Sarma 1989; Avrutin & Wexler 1992; Hestvik & Philip 1999/2000; Grolla 2005). Some studies showed very high rates of interpretations that violate Principle B, while other studies showed quite low rates of variation. The spread in findings was more than would be expected by chance. Furthermore, a closer look at the experimental designs revealed that the varying outcomes were somewhat predictable based on the scenarios that were used to test the children. When the grammatical interpretation of the pronoun, i.e., disjoint from the subject NP was supported by a prominent referent ("availability") and an at-issue proposition ("disputability") children were good at selecting that interpretation over an illicit bound interpretation of the pronoun.

So the empirical conclusion is that preschoolers have the linguistic knowledge needed to successfully apply Principle B. But they are very fragile. When experimental conditions are not set up just right, they can easily be pushed to entertain interpretations that violate Principle B. Moreover, there is a strikingly close alignment between children and adults. The cases where

children appear to get stuck on non-adultlike interpretations in their offline interpretations align rather closely with the cases where adults fleetingly consider illicit interpretations in their online interpretations. And cases where children's interpretations are more robustly adultlike correspond to cases where adults' online interpretations are impervious to illicit lures (Phillips & Ehrenhofer 2015).

These findings lead to new theoretical questions: why are some illicit interpretations considered fleetingly in the course of parsing while others are not? This aligns closely with questions raised by linguistic illusions. Relatedly, how are (combinatorial) interpretations generated, using a combination of grammatical and situational knowledge? These are rather different than the questions that we started with, but they are at least as interesting.

Lexical Activation & Response Probability

There is much recent interest in linguistic and psycholinguistic theory in the role of context. In psycholinguistics the focus has been on how comprehenders use contextual constraints to constrain the parsing and interpretation of upcoming input. At a high level, much of this work can be understood as an investigation of cross-talk between different parts of the language system: can information in one level of representation be used to constrain operations in another level of representation (e.g., Fodor 1983). However, we have found here, as in other areas, that our initial simplistic theoretical questions led us to new questions that we were not even aware of when we started.

Our starting point was one of the simplest contextual constraints that we can find. When a sentence context strongly predicts the syntactic category of the next word, does that constraint limit the words that are considered as the next word is recognized? For example, the word onset *br…* could turn out to be either a noun (e.g., *brownie*) or a verb (e.g., *browse*). If the context strongly predicts a verb (e.g., *She wanted to br…*) are words from the non-predicted category accessed? This is a long-standing question, and studies using different experimental measures have reached different conclusions. Some have concluded that syntactic category does not limit lexical access (Tanenhaus et al. 1979; Tyler 1984), whereas others have concluded that it does (Magnuson et al. 2008; Strand 2018). We wanted to better understand these conflicts.

In this case computational modeling provided crucial clarification for our linking hypotheses, and subsequent clarification of our theories. The TRACE model (McClelland & Elman 1986) proposes that auditory word recognition proceeds by activating feature, phoneme, and then word-level representations in a connectionist network. So, continuously varying activation levels are the key currency of lexical computation. In behavioral studies of word recognition, lexical activation levels are, of course, not directly observable. Instead, researchers have used accuracy, reaction time, eye movements, and neural activity as proxies for lexical activation levels.

In some paradigms the dependent measure can be thought of as a response probability: how likely is a participant to choose, fixate, or produce one lexical candidate rather than another? It would be easy to think of response probabilities as directly reflecting changes in activation, just as acceptability is often taken to transparently reflect grammaticality. But an important idea that TRACE makes clear is that we should not expect response probability to increase straightforwardly with lexical activation. Lexical activation -- the property we are interested in studying -- must be mapped onto response probability -- the property we have access to -- via transformations whose properties we must also take into account.

For example, the set of response candidates in a task is the set of lexical candidates made available to the participant by the dependent measure. Sometimes the participant can choose freely from the entire mental lexicon, as in production tasks, and sometimes the participant can only look among four options, as in the visual world (eye-tracking) paradigm. TRACE specifies that the makeup of the response candidate set does not affect underlying lexical activation, but it does affect response probabilities. This is because response probability is computed by dividing (transformed) lexical activation for a given candidate by the sum of the activation values for all response candidates. Lexical activation may be the property of interest, but response probability is the property being measured, and it is jointly determined by the activation of the item we care about and the activation of the rest of the set. So, in making inferences about the lexical activation of a particular item from patterns in response probability, we have to consider what is happening to the rest of the candidate set. Specifically, if we want to isolate a change in probability for the item we care about and infer a change (or lack thereof) in activation *for that item*, we need to ensure that no changes in activation are taking place in the rest of the set.

If we cannot be sure of this, we must consider that the *direction* and *latency* of changes in response probability due to changes in activation are heavily influenced by the relative changes in activation of the rest of the set. A lack of change in response probability therefore does not necessarily imply a lack of change in activation, because a change in activation of the item we care about will only lead to a change in probability if it changes more or less than the rest of the set. This principle turns out to be crucial in interpreting our behavioral results, but it is lost without a transparent linking hypothesis between activation and response probability. TRACE, as a computational model of auditory word recognition, has to be clear about this in order for simulations to be possible. Insights of this nature certainly do not necessarily require sophisticated modeling techniques, though they are often helpful.

Seriously considering the relationship between activation and response probability turns out to be extremely important for understanding the role of syntactic category in auditory word recognition. Lexical competition in the visual world paradigm is indexed by an increase in fixations to a cohort competitor of the auditory target, relative to an unrelated distractor. Such competition is well-established when the competitor is of the same syntactic category as the target. Two studies (Magnuson, Tanenhaus, & Aslin 2008; Strand, Brown, Brown, & Berg, 2018) have shown that lexical competition does not occur when the cohort competitor is of a different

syntactic category from the target, suggesting that syntactic context can prevent the activation of syntactically incompatible lexical candidates.

However, in recent work we have found that when the design is carefully controlled such that the syntactically inappropriate cohort competitor is the only candidate in the response candidate set whose activation can be expected to change in response to the auditory input, lexical competition is indeed detectable (Gaston et al. 2019). Changes in activation of the target and/or distractors may have obscured this in previous designs; there was no “competition” as measured by response probability, but this does not mean there was not increased activation for the item in question. Strategies that could lead fixations to be driven by properties of the pictures in the display other than the words associated with them may also have played a role. It was simple simulations using jTRACE (Strauss, Harris, & Magnuson, 2007) that allowed us to understand the importance of ensuring that fixations only be driven by lexical activation changes of the critical item. Evidence that syntactic category does not prevent the competition of syntactically incompatible lexical candidates then leads us to new questions about the role of syntactic category expectations in auditory word recognition. For example, it establishes that bottom-up auditory cues cannot be overridden, and that the more likely mechanism for syntactic category effects on lexical representations is facilitation of syntactically compatible items.

Role Reversals and the Semantic P600

One of the most influential ERP findings on language comprehension in the past 15 years comes from a study by Kim & Osterhout (2005) on sentences with ‘thematic role reversals’. This study appeared to show that comprehenders build interpretations that are inconsistent with the syntactic structure of the sentence. Our group became interested in these findings because they challenged a widespread assumption in linguistic and psycholinguistic theories about the tight coupling of syntax and semantics. However, as we dug deeper our original motivation faded, and we discovered new theoretical questions that we had not been aware of previously, involving the use of linguistic information to access non-linguistic information in memory.

Kim & Osterhout (2005) compared ERPs to sentences like (9-10), where (9) is a grammatically appropriate and plausible passive sentence and (10) contains the same open class words but is a grammatically well-formed but implausible active sentence. The verb *devouring* in (10) elicited the P600 effect commonly seen in response to syntactic anomalies, despite the fact that the sentence is syntactically well-formed. It did not elicit the N400 effect typically associated with semantic anomalies, despite being highly semantically anomalous.

(9) The hearty meal was devoured ...
(10) The hearty meal was devouring ...

Kim and Osterhout argued that this pattern arose because comprehenders perceived the sentence as syntactically anomalous. Under this account, comprehenders recognize that *meal* is an attractive theme of *devour* and they construct a corresponding interpretation, ignoring the

fact that the sentence is an active rather than a passive. As a result, the sentence is initially perceived as plausible and no N400 effect is elicited. Subsequently they notice that the syntactic form of the sentence mismatched the interpretation, and hence a P600 effect is elicited.

Kim and Osterhout's finding was not the first of its kind, and many others have reported similar ERP effects in response to role reversals and similar anomalies (Kuperberg et al. 2003; Kolk et al. 2003; Hoeks et al. 2004; Ye & Zhou 2008). One of the most distinctive contributions of Kim and Osterhout's study is that they present evidence that the effect specifically depends on the presence of 'semantic attraction' between the verb and the arguments. In a second experiment they compared *hearty meal* with *dusty tabletop* in sentences like (10). Both are poor agents of *devour*, but only *hearty meal* is an attractive theme. They found that *dusty tabletop* elicited a more typical pattern of an N400 effect and no P600 effect. So they concluded that their initial effect was specifically due to interpretations that are semantically attractive but syntactically unsupported. This is a well-constructed experimental argument.

The key linking assumptions for Kim and Osterhout's argument were the long-standing view that N400 effects reflect (combinatorial) semantic processing and that P600 effect reflect syntactic processing. If we adopt these assumptions then we are led to the conclusion that semantic interpretation can proceed independent of syntax. This challenges such a fundamental theoretical assumption that members of our group decided to examine the processing of role reversals in more detail. As in many other cases, this led us somewhere very different from where we had started.

Summarizing a number of years of research by our group and many others, we learned the following.

First, some of the empirical generalizations offered by Kim & Osterhout (2005) are robust, but others are not.

Role reversal sentences do consistently elicit P600 effects, despite the fact that they are syntactically well-formed. This is consistent with Kim and Osterhout's claims. However, the P600 effect is not not limited to cases of semantic attraction. Relatively few studies have manipulated the presence of attraction as Kim and Osterhout did, but those that have done so have generally found that the P600 is elicited even in the absence of semantic attraction (e.g., Van Herten et al. 2006; Kuperberg et al. 2006; Stroud 2008; Psczynski & Kuperberg 2011; Stroud & Phillips 2012; Chow & Phillips 2013).

Kim and Osterhout's N400 findings are generally robust, but they appear to reflect a broader generalization: N400 effects reflect the degree to which an incoming word is expected in context, and this in turn is influenced by lexical associations between the incoming word and prior words in the context. So, *meal … devour* elicits a smaller N400 than *tabletop … devour* because *meal* is more closely associated with *devour*. A widespread current view is that the N400 reflects lexical processes, rather than combinatorial semantic processes, and that those

processes are modulated by earlier processes that make a word more or less expected. Evidence for this view of the N400 comes from lexical priming effects, neuroanatomical evidence, predictive grammatical agreement, and phonological effects, among others (Kutas & Federmeier 2000; Lau et al. 2008; van Berkum et al. 2005; Mantegna et al. 2019).

Furthermore, our group has found that the 'blindness' of the N400 to thematic role reversals can be cured if there is more time between the predictive cues and the target verb. For example, Chow et al. (2018) replicated in Mandarin the standard finding that the verb with role-reversed arguments fails to elicit an N400 effect. This much is the same as Kim & Osterhout (2005). But they also found that when the verb is presented at a greater delay after the same arguments then the N400 effect reappears. So, timing matters. They also found that this reappearance occurs only when the arguments strongly predict the verb. So, timing matters specifically for prediction. We found similar effects in role-reversed sentences in Japanese, and in speeded cloze tasks in English (Chow et al. 2015; Momma et al. 2016; Burnsky et al. 2019).

Based on these findings, we argued that comprehenders accurately parse and interpret incoming sentences, and that they use all available information to make predictions about upcoming words. However, not all information is used equally quickly. Early predictions are based primarily on lexical associations. Further refinement of those predictions based on thematic roles does occur, but it takes more time. This is why canonical and role-reversed sentences elicit identical N400s at short latencies, because the sentence types are matched in terms of lexical associations. But when more time elapses between the arguments and the verb the N400 to canonical and role-reversed sentences differs, reflecting the emergence of more specific predictions based on thematic roles.

Why should it take time for predictions based on thematic roles to impact lexical expectations? We have suggested that this may be because lexical prediction is the result of a memory access process, that is slower when the memory access cues mismatch the format of semantic memory. We suggest that our long-term knowledge of events is not encoded in terms of abstract thematic roles like 'agent' and 'patient', and so a multi-step process is needed to map from linguistic argument role cues to event memory (Chow et al. 2016).

So, what started as an investigation into long-standing claims about syntax-semantics relations in a standard linguistic architecture turned into brand new questions about the relationship between grammatical information and world knowledge. Still theoretical, certainly interesting, but not the question that we started with.

Part 3: In Search of Linking Hypotheses

The methods described above allow us to determine the properties of some level of representation: whether a representation satisfies or violates Principle B, how constituents are encoded in memory, etc. But sometimes we are interested in more basic, architectural questions

about syntactic representations: how structures are built, what syntactic representations there are, and how they are related to each other.

Like the case studies discussed above, these architectural questions have traditionally been investigated using acceptability judgments. In more recent work, though, researchers have also employed diverse experimental techniques --- measuring neural activity, structural priming effects, and speed-acceptability tradeoffs --- to investigate these questions. On the basis of data collected via these techniques, they have begun to advance strong claims about the nature of representations and how they are built and parsed.

While these techniques are exciting in their own right, we point out that the linking hypotheses connecting these dependent measures to representational claims are often somewhat underdeveloped. Consequently, conclusions based on these methods should be taken with caution, since there are likely to be alternative interpretations of the data. That said, we are hopeful that further elaboration of the linking hypotheses will lead to more productive debates and further breakthroughs in experimental techniques.

Neurosyntax

A growing body of work uses neurocomputational models to predict neural activity during naturalistic reading or listening. This approach is based on the idea that spelling out the linking hypotheses between syntactic representations of linguistic input and neural responses to that input will allow us to narrow the hypothesis space for those syntactic representations themselves, as well as how they are parsed and where and when this occurs in the brain.

Neurocomputational models of this sort generally have multiple components: a grammar, an algorithm, an oracle, a complexity metric, and a response function. As described in Brennan's review of the approach (Brennan 2016), a grammar (sequence-based, context-free, Minimalist, etc) defines well-formed syntactic representations for the linguistic input that is provided to participants, and a parsing algorithm (top-down, left-corner, bottom-up, etc) determines how to apply the grammar to incremental input. An *oracle* is used to make decisions in the case of, for example, syntactic ambiguities, and the oracle can vary in the information it has access to. The grammar, algorithm, and oracle together make up the syntactic parser, which takes in words and returns mental states. Mental states can be, for example, syntactic trees. A complexity metric is then used to describe or quantify those mental states, in terms of the number of nodes added to the tree, the reduction in entropy over possible syntactic trees (if the oracle is not choosing a single tree at each step), the surprisal of the syntactic category of the word, or the number of open dependencies, among many other possibilities. Complexity metrics are combined with response functions that try to take into account the relationship between hypothesized neural states and the neural signals measurable in methods like fMRI, EEG, MEG, or ECoG. This final step is what allows us to compare predicted neural signal and actual neural signal, in response to specific input.

“L-studies”, as Brennan (2016) terms them, test the predictions of different versions of the model against the neural signal from a constrained set of brain areas. This approach has been used in order to ask, for example, whether sequence-based grammars or grammars allowing abstract hierarchical structure better predict neural activity in areas known to be associated with syntactic processing. “N-studies”, in contrast, take a given parametrization of the model and then ask about the location or timing of correlations between its predictions and actual neural activity in all areas of the brain.

A variety of interesting findings have emerged from this approach. Brennan et al. (2016) show that node counts from an audiobook story’s proposed hierarchical syntactic structure predict the time-course of participants’ BOLD signal while they listen to that story, and Brennan & Pylkkanen (2016) show that the number of left-corner parse steps associated with visually presented sentences predicts MEG activity in the anterior temporal lobe. Similarly, Nelson et al. (2017) describe neural evidence for a merge operation (among other claims), in ECoG, and Hale et al. (2018) argue for RNN grammars with beam search on the basis of their findings in EEG.

However, the resolution with which we can distinguish the models that yield these different predicted neural states is limited by the noisiness of neural data, by the detail of the linking hypotheses about parsing operations, and the degree of correlation between alternative models. This is evident already in the sheer number of different grammars, parsing algorithms, and complexity metrics that have found support in recent work. It can easily be the case that different hypotheses for the grammar or parser yield similar outcomes with respect to their complexity metrics, which means that predicted neural signals from many different model parameterizations can be highly correlated with each other. Complexity metrics, which are the way in which we quantify the mental states derived from the parser, can suffer from an untold number of potential confounds, where other properties of the input that are perhaps less interesting to the researcher could equally well contribute to an observed pattern. It is therefore unlikely that parsing is accurately characterized by all of the different metrics that have been found to correlate with neural activity.

But this is a rapidly developing area. Confounds are being incorporated to the extent possible, and model comparison techniques allow us to say which overall model better predicts neural activity within any given dataset. One of the primary challenges for this area of inquiry will be to account for increasingly more of the various processes that influence the response function linking the complexity metric and the dependent measure. This, in combination with increased specificity in the parsing models, will allow for more fine-grained predictions for neural activity.

Structural Priming

Closely related to the question of how a representation is built in parsing are questions about levels of syntactic representations: how many there are, what they are, and how they are related. Transformational theories of grammar, for example, have always maintained that there

are several levels; a standard view is that there are at least three within syntax: one level that has consequences for both semantics and phonology (roughly S-structure in government-binding theory, “narrow syntax” in some more recent minimalist work), one level that specifically impacts semantics ("LF"/"Logical Form"), and another level that specifically impacts morphophonology ("PF"/"Phonetic Form").

All grammatical theories assume that sentences somehow simultaneously encode multiple different syntactic, semantic, and phonological properties (e.g., thematic structure, scope relations, linear order). Many theories assume that at least some of these properties are encoded in distinct structural representations (e.g., *Lexical-functional Grammar*, Bresnan et al., 2015; *Categorial Grammar*, Steedman 2000). The distinctive claim of transformational theories has always been that the multiple representations are related by means of transformational operations that move items between different positions in a phrase marker, e.g., moving a noun phrase from a position that encodes its thematic status to a higher position that encodes its scope.

Traditionally, evidence for the different levels of representation in transformational theories, especially LF, comes from informal acceptability judgments (e.g. Huang 1982, May 1985). However, researchers have also weighed in with experimental evidence. Branigan and Pickering (2017), for example, put forward an argument for a single level (“monostratral”) syntax without transformations, offering evidence from structural priming, which is assumed to diagnose structural identity.

Branigan and Pickering cite several priming studies that show that constructions implicating transformations are primed by string-identical constructions that are, by hypothesis, not produced via transformations. For example, English and other languages distinguish two types of intransitive verbs. *Unergative* verbs have a single argument that bears an agent role, whereas *unaccusative verbs* have a single argument that bears a theme or patient role (Levin & Rapoport 1995). Unaccusatives and unergatives differ in many ways across languages, but they appear in very similar surface forms in English. Transformational theories claim that the surface subject of an unaccusative sentence is derived by moving the single argument from an underlying direct object position to the surface subject position.

Branigan and Pickering point to evidence that English unaccusatives sentences like *The snow melted* are primed to the same degree by other unaccusative sentences, such as *The water froze* and by unergative sentences like *The children sang*. They reason that the syntactic representation of unaccusatives must therefore be identical to that of unergatives. Specifically, there is no representation where the argument of an unaccusative verb is found in an object position before moving to a subject position, as suggested by transformational or Relational Grammar theories. Rather, the argument of unaccusative and unergative verbs only ever occupies the same structural position, the standard subject position. This, together with a number of similar cases, is used to argue that syntax is monostratal, in the sense that there are no distinct levels of representation connected by movement / transformation operations.

The intuition behind the syntactic priming argument is easy to grasp: pairs of sentences that share structural properties prime one another, because the same abstract structure is reused. However, there are many ways for two sentences to be identical. Branigan and Pickering's argument appears to rest on the assumption that, if two structures prime a third construction equally, then the two structures must be homomorphic, i.e. there is a one-to-one correspondence between the syntactic relations in the first structure and those in the second.

However, one needs to move beyond the basic intuition to a more detailed theory of what type of structural identity is needed to drive syntactic priming. This is an underdeveloped area, despite the great amount of empirical research on structural priming. For example, all grammatical theories agree that there is some degree of shared structure between English unaccusatives and unergatives, even if they assume additional steps in the derivation of unaccusatives. It could be that this degree of shared structure is sufficient to yield structural priming effects. Structural priming is a potentially powerful way of diagnosing the structural content of sentences, but this will only be possible once a more articulated theory of priming is available, and there is little reason to regard it as somehow more reliable or privileged as a diagnostic of syntactic structure. For further discussion see Gaston et al. (2018).

We can also contrast this with the logic that underlies another set of experimental results that have been used to test the claims of transformational theories. For example, Xiang and colleagues used the speed--accuracy tradeoff (SAT) paradigm to argue that the processing of Mandarin wh-in-situ constructions is sensitive to the length of the covert wh-dependency between the thematic position and the scope position of the wh-phrase, thus indirectly providing evidence for the online construction of covert wh-dependencies (Xiang et al. 2014; see also Xiang et al. 2015). To probe this question experimentally, they explicitly assume that covert wh-movement targets a "clause-edge position" and that "wh-in-situ constructions engage the same types of parsing procedures involved in building overt wh-dependencies," namely, the retrieval of previously-encountered syntactic structures. By assuming that covert dependencies also involve memory retrieval, they justify their use of SAT, a paradigm where there is a clearer consensus on how to analyze and interpret the data. For example, it is standard practice to convert responses into d-prime measures and to model d-primes as a function of time and three parameters: asymptote, rate, and intercept. There are also widely-accepted interpretations of these parameters, e.g. differences in rates implicate a serial search and retrieval process.

In other words, while experimental approaches open up new opportunities for probing fundamental questions about linguistic representations, the data they yield are also often more removed from the representational claims that they address. Consequently, the validity of the linking hypotheses in these approaches become more important. The structural priming and SAT examples provide an interesting contrast: SAT might not provide as much detail about representations, but its core assumptions are more worked out and the results are easier to interpret. Structural priming as an approach is less developed, but could let us directly compare between representations. We hope that future work on structural priming will lead us to a more

explicit theory that can diagnose different kinds of structural identity and interface with other methods of diagnosing structure, such as acceptability judgments.

Conclusion: Theories all the way down

Standard linguistic theory is a cognitive theory at a rather high level of analysis, one that abstracts away from many important properties of neurocognitive systems. It typically makes a series of assumptions about the discreteness of representations, and it abstracts away from issues of the real-time order and timing of cognitive processes, or how linguistic representations are encoded in memory or in neural circuitry. This high level of analysis sets aside a lot of detail, and in so doing it allows for rapid progress and broad coverage. But this is not to say that there is a shortage of interesting theories and theoretical questions at finer-grained levels of analysis,

In a famous anecdote, attributed to different philosophers and scientists over the years, the main protagonist gives a public lecture on the structure of the earth and the universe, and is then approached by an old lady who offers an alternative: the earth is flat, and it is supported on the back of a large turtle. The scientist tries to politely point to the flaw in the old lady's argument, by asking what is supporting the large turtle. The old lady replies that the first turtle is sitting on the back of a second one, and so on -- "It's turtles all the way down!".

And so it is with linguistic theories. The questions posed by standard linguistic theory are interesting and important, but they abstract away from a great amount of detail at lower levels of analysis. Once we dig deeply into those levels, we uncover many new theoretical questions that we were unaware of previously. If one is interested only in the theoretical questions that already came from traditional linguistic theory, then one could be forgiven for concluding that the advent of experimental approaches has brought little theoretical insight. But once we allow that new empirical approaches reveal questions that we were not even aware of previously, then the outlook becomes a great deal more promising.

Acknowledgments

This research was supported in part by NSF training grant DGE-1449615 to the University of Maryland (Phillips, PI), and by University of Maryland Flagship Fellowship awards to Phoebe Gaston and Hanna Muller. We are grateful to many people for helpful discussions of the issues addressed here, and especially to Ellen Lau and Iria de Dios Flores for pushing us to tackle the main challenge addressed here.

References

Ackerman, L., Frazier, M., & Yoshida, M. (2018). Resumptive pronouns can ameliorate illicit island extractions. *Linguistic Inquiry*, *49*, 847-859.

Alexopoulou, T., & Keller, F. (2007). Locality, cyclicity, and resumption: At the interface between the grammar and the human sentence processor. *Language*, 110-160.

Avrutin, S. & Wexler, K. (1992). Development of Principle B in Russian: Coindexation at LF and coreference. *Language Acquisition, 2*, 259-306.

Baker, M. C. (2008). *The syntax of agreement and concord*. Cambridge University Press.

Beltrama, A., & Xiang, M. (2016). Unacceptable but comprehensible: the facilitation effect of resumptive pronouns. *Glossa*, *1*(1), 1024

Bock, K., & Miller, C. A. (1991). Broken agreement. *Cognitive Psychology*, *23*(1), 45-93.

Boster, C. (1991). Children's failure to obey Principle B: Syntactic problem or lexical error? Ms., University of Connecticut, Storrs.

Branigan, H. P., & Pickering, M. J. (2017). An experimental approach to linguistic representation. *Behavioral and Brain Sciences*, *40,* e282.

Brennan, J. (2016). Naturalistic sentence comprehension in the brain. *Language and Linguistics Compass*, *10*, 299-313.

Brennan, J. R., & Pylkkänen, L. (2017). MEG evidence for incremental sentence composition in the anterior temporal lobe. *Cognitive Science*, *41*, 1515-1531.

Brennan, J. R., Stabler, E. P., Van Wagenen, S. E., Luh, W., & Hale, J. (2016). Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language, 157,* 81-94.

Bresnan, J., Asudeh, A., Toivonen, I., & Wechsler, S. (2015). *Lexical Functional Syntax, 2nd edition.* Wiley Blackwell.

Burnsky, J. & Staub, A. (2019). Cloze completions reveal misinterpretations of noncanonical sentences. Talk at Psycholinguistics in Iceland - Parsing and Prediction. Reykjavik, Iceland.

Chien, Y. C., & Wexler, K. (1990). Children's knowledge of locality conditions in binding as evidence for the modularity of syntax and pragmatics. *Language Acquisition*, *1*, 225-295.

Chomsky, N. (1981). *Lectures on government and binding: The Pisa lectures* (No. 9). Walter de Gruyter.

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

Chow, W.-Y., Kurenkov, I., Buffinton, J., Kraut, R., & Phillips, C. (2015). How predictions change over time: Evidence from an online cloze paradigm. Poster presented at the 28th annual CUNY Sentence Processing Conference. Los Angeles, California.

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

Chow, W.-Y., Lau, E., Wang, S., & Phillips, C. (2018). Wait a second! Delayed impact of argument roles on on-line verb prediction. *Language, Cognition and Neuroscience, 33*, 803-828.

Conroy, A., Takahashi, E., Lidz, J., & Phillips, C. (2009). Equal treatment for all antecedents: How children succeed with Principle B. *Linguistic Inquiry*, *40*, 446-486.

Dillon, B., Mishler, A., Sloggett, S., & Phillips, C. (2013). Contrasting interference profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language, 69*, 85-103.

Eberhard, K. M., Cutting, J. C., & Bock, J. K. (2005). Making syntax of sense: number agreement in sentence production. *Psychological Review, 112*, 531-559.

Elbourne, P. (2005). On the acquisition of Principle B. *Linguistic Inquiry*, *36*, 333-365.

Ferreira, F. (2005). Psycholinguistics, formal grammars, and cognitive science*. The Linguistic Review, 22*, 365-380.

Ferreira, F., & Swets, B. (2005). The production and comprehension of resumptive pronouns in relative clause “island” contexts. *Twenty-first century psycholinguistics: Four cornerstones*, 263-278.

Gaston, P., Huang, Z. N., & Phillips, C. (2017). The logic of syntactic priming and acceptability judgments. *Behavioral and Brain Sciences, 40, e289.*

Gaston, P., Lau, E., & Phillips, C. (2019). Syntactic category does not inhibit lexical competition. Proceedings of the 11th International Conference on the Mental Lexicon.

Gibson, E. & Fedorenko, E. (2013). The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes, 28*, 88-124.

Grodzinsky, Y., & Reinhart, T. (1993). The innateness of binding and coreference. *Linguistic inquiry*, *24*(1), 69-101.

Grolla, E. (2005). *Pronouns as elsewhere elements: Implications for language acquisition*. Doctoral dissertation, University of Connecticut.

Hale, J., Dyer, C., Kuncoro, A., & Brennan, J. R. (2018). Finding Syntax in Human Encephalography with Beam Search. *arXiv preprint arXiv:1806.04127*.

Hammerly, C., Staub, A., & Dillon, B. (2019). The grammaticality asymmetry in agreement attraction reflects response bias: Experimental and modeling evidence. *Cognitive Psychology, 110*, 70-104.

Heestand, D., Xiang, M., & Polinsky, M. (2011). Resumption still does not rescue islands. *Linguistic Inquiry*, *42*, 138-152.

Hestvik, A. & Philip, W. (1999/2000). Binding and coreference in Norwegian child language. *Language Acquisition, 8*, 171-235.

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, C. T. J. (1982). *Logical relations in Chinese and the theory of grammar* (Doctoral dissertation, Massachusetts Inst. of Technology Cambridge).

Jonides, J., Lewis, R. L., Nee, D. E., Lustig, C. A., Berman, M. G., & Moore, K. S. (2008). The mind and brain of short-term memory. *Annual Reviews in Psychology, 59*, 193-224.

Kaufman, D. (1988). *Grammatical and cognitive interactions in the study of children's knowledge of binding theory and reference relations*. Doctoral dissertation, Temple University, Philadelphia, PA.

Kazanina, N., Lau, E. F., Lieberman, M., Yoshida, M., & Phillips, C. (2007). The effect of syntactic constraints on the processing of backwards anaphora. *Journal of Memory and Language, 56*, 384-409.

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

Kimball, J., & Aissen, J. (1971). I think, you think, he think. *Linguistic Inquiry*, *2*, 241-246.

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.

Kuperberg, G. R., Sitnikova, T., Caplan, D., & Holcomb, P. J. (2003). Electrophysiological distinctions in processing conceptual relationships within simple sentences. *Cognitive Brain Research, 217*, 117-129.

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

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

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (de)constructing the N400. *Nature Reviews Neuroscience, 9,* 920-933.

Lewis, S. & Phillips, C. (2015). Aligning grammatical theories and language processing models. *Journal of Psycholinguistic Research, 44*, 27-46.

Lewis, R. L., Vasishth, S., & Van Dyke, J. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences, 10*, 447-454.

Lombardi, L., and Sarma, J. (1989). Against the bound variable hypothesis of the acquisition of Condition B. Paper presented at the annual meeting of the Linguistic Society of America, Washington, D.C.

May, R. (1985). *Logical form: Its structure and derivation*. MIT Press.

Magnuson, J. S., Tanenhaus, M. K., & Aslin, R. N. (2008). Immediate effects of form-class constraints on spoken word recognition. *Cognition*, *108*(3), 866-873.

Mangegna, F., Hintz, F., Ostarek, M., Alday, P. M., & Huettig, F. (2019). Distinguishing integration and prediction accounts of ERP N400 modulation in language processing through experimental design. In press, *Neuropsychologia.*

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive psychology*, *18*(1), 1-86.

McDaniel, D., Cairns, H., & Hsu, J. (1990). Binding principles in the grammars of young children. *Language Acquisition, 1*, 121-139.

McElree, B. (2006). Accessing recent events. *Psychology of Learning and Motivation*, *46*, 155-200.

Momma, S., Luo, Y., Sakai, H., Lau, E., & Phillips, C. (2016). Lexical predictions and the structure of semantic memory: EEG evidence from case changes. Talk at the 29th annual CUNY Conference in Human Sentence Processing. Gainesville, FL.

Muller, H., de Dios Flores, I., & Phillips, C. (2019). Not (just) any licensors cause negative polarity illusions. Talk at Psycholinguistics in Iceland - Parsing and Prediction. Reykjavik, Iceland.

Nelson, M. J., El Karoui, I., Giber, K., Yang, X., Cohen, L., Koopman, H., ... & Dehaene, S. (2017). Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences*, *114*(18), E3669-E3678.

Paczynski, M. & Kuperberg, G. R. (2011). Electrophysiological evidence for the use of the animacy hierarchy, but not thematic role assignment, during verb argument processing. *Language and Cognitive Processes, 26*, 1402-1456.

Parker, D. & Phillips, C. (2016). Negative polarity illusions and the format of hierarchical encodings in memory. *Cognition, 157*, 321-339.

Pearlmutter, N. K., Garnsey, S. M., & Bock, J. K. (1999). Agreement processes in sentence comprehension. *Journal of Memory and Language, 41*, 427-456.

Reinhart, T. (1983). Coreference and bound anaphora: A restatement of the anaphora questions. *Linguistics and Philosophy*, *6*(1), 47-88.

Ross, J. R. (1967). *Constraints on variables in syntax* (Doctoral dissertation, Massachusetts Institute of Technology).

Philip, W. & Coopmans, P. (1996). The double Dutch delay of Principle B effect. In A. Stringfellow, D. Cahana-Amitay, E. Hughes, & A. Zukowski (eds.), *Proceedings of the 20th annual Boston University Conference on Language Development,* pp. 576-587. Somerville, MA: Cascadilla Press.

Phillips, C. (2006). The real-time status of island phenomena. *Language, 82*, 795-803.

Phillips, C. (2010). Should we impeach armchair linguists? In S. Iwasaki, H. Hoji, P. Clancy, and S.-O. Sohn (eds.), *Japanese-Korean Linguistics 17, pp. 49-64.* Stanford, CA: CSLI Publications.

Phillips, C. & Ehrenhofer, L. (2015). The role of language processing in language acquisition. *Linguistic Approaches to Bilingualism, 5*, 409-453.

Phillips, C., & Parker, D. (2014). The psycholinguistics of ellipsis. *Lingua*, *151*, 78-95.

Phillips, C., & Wagers, M. (2007). Relating structure and time in linguistics and psycholinguistics. *Oxford handbook of psycholinguistics*, 739-756.

Phillips, C., Wagers, M. W., & Lau, E. F. (2011). *Grammatical illusions and selective fallibility in real-time language comprehension*. Experiments at the Interfaces, 37, 147-180.

Ross, J. R. (1967). *Constraints on variables in syntax*. Doctoral dissertation, MIT.

Sprouse, J. & Almeida, D. (2012). Assessing the reliability of textbook data in syntax: Adger's *Core Syntax*. *Journal of Linguistics, 48*, 609-652.

Sprouse, J., Schütze, C. T., & Almeida, D. (2013). A comparison of informal and formal acceptability judgments using a random sample from *Linguistic Inquiry* 2001-2010. *Lingua, 134,* 219-248.

Steedman, M. (2000). *The syntactic process*. Cambridge, MA: MIT Press.

Stowe, L. A. (1986). Evidence for on-line gap location. *Language and Cognitive Processes, 1*, 227-245.

Strand, J. F., Brown, V. A., Brown, H. E., & Berg, J. J. (2018). Keep listening: Grammatical context reduces but does not eliminate activation of unexpected words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*(6), 962-973.

Strauss, T. J., Harris, H. D., & Magnuson, J. S. (2007). jTRACE: A reimplementation and extension of the TRACE model of speech perception and spoken word recognition. *Behavior Research Methods*, *39*(1), 19-30.

Stroud, C. (2008). *Structural and semantic selectivity in the electrophysiology of sentence comprehension*. Doctoral dissertation, University of Maryland.

Stroud, C. & Phillips, C. (2012). Examining the evidence for an independent semantic analyzer: An ERP study in Spanish. *Brain and Language, 120*, 107-126.

Sturt, P. (2003). The time course of the application of binding constraints in reference resolution. *Journal of Memory and Language, 48*, 542-562.

Tanenhaus, M. K., Leiman, J. M., & Seidenberg, M. S. (1979). Evidence for multiple stages in the processing of ambiguous words in syntactic contexts. *Journal of Verbal Learning and Verbal Behavior, 18*, 427-440.

Thornton, R., & Wexler, K. (1999). *Principle B, VP ellipsis, and interpretation in child grammar*. MIT Press.

Tyler, L., K. (1984). The structure of the initial cohort: Evidence from gating. *Perception and Psychophysics, 36*, 417-427.

Van Berkum, J. J. A., Brown, C., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory and Cognition, 31*, 443-467.

Van Herten, M., Chwilla, D. J., & Kolk, H. H. J. (2006). When heuristics clash with parsing routines: ERP evidence for conflict monitoring in sentence perception. *Journal of Cognitive Neuroscience, 18*, 1181-1197.

Vasishth, S., Brüssow, S., Lewis, R. L, & Drenhaus, H. (2008). Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science, 32*, 685-712.

Wagers, M., Lau, E. F., & Phillips, C. (2009). Agreement attraction in comprehension: representations and processes. *Journal of Memory and Language, 61*, 206-237.

Xiang, M., Dillon, B., Wagers, M., Liu, F., & Guo, T. (2014). Processing covert dependencies: An SAT study on Mandarin wh-in-situ questions. *Journal of East Asian Linguistics*, *23*(2), 207-232.

Xiang, M., Wang, S., & Cui, Y. (2015). Constructing covert dependencies—The case of Mandarin wh-in-situ dependency. *Journal of Memory and Language*, *84*, 139-166.

Ye, Z. & Zhou, X. (2008). Involvement of cognitive control in sentence comprehension: evidence from ERPs. *Brain Research, 1203*, 103-115.

Zukowski, A., & Larsen, J. (2004, March). The production of sentences that we fill their gaps. In *Poster presented at the CUNY Sentence Processing Conference, University of Maryland*.