December 20th 2019. To appear in the Cambridge Handbook of Experimental Syntax, Grant Goodall, ed.

Theories All The Way Down:
Remarks on “theoretical” and “experimental” linguistics

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

1 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.

Some of our previous work has, perhaps correctly, been seen as painting a 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.

The situation is reminiscent of 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 questions that we were unaware of previously. These are theoretically interesting. They are just not the theoretical questions that we started with.

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.

2 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 more opaque linking hypothesis. We all know that acceptability judgments are not a transparent reflection of grammatical well-formedness -- the unacceptability of double center-embeddings is a parade case. 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. However, acceptability judgments are necessarily filtered through the language comprehension system, and most syntacticians are adamant that the structure-building mechanisms that they describe are 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 were the same, aside from the uncertainty that is specific to the comprehension task, i.e., the fact that comprehenders have to figure out what the speaker is trying to convey.

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).

2.1 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. 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 (Wagers et al. 2009). 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 (Chow et al., 2016; Parker & Phillips 2016; Vasishth et al. 2008; Wagers et al. 2009; Wellwood et al. 2019, among others) -- 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 mechanisms are used to generate and access that representation. These are important linking assumptions for all of syntax, but also interesting questions in their own right.

Turning our attention to the online generation and access of syntactic structures calls for measurements that tap into online processes. In the case of agreement attraction and other illusions, self-paced reading and speeded acceptability are often useful tools, since they allow us to infer representations and processes that are entertained before a final, careful acceptability judgment. Note, however, that both of these measurements require a series of linking hypotheses of their own. In self-paced reading, the progression from one word to the next is mediated by not only the linguistic representations of interest but also by visual processing of the stimulus, decision-making, and motor planning to execute a keypress.

While we do not intend to advocate for any particular account of agreement attraction phenomena, we find this area to be a useful example how probing the "linking hypotheses" of theoretical syntax can open up interesting new questions. The various explanations for illusions essentially differ in where in the process of detecting ungrammaticality the blame lies. One class of hypotheses posits that the representation of linguistic information *prior to* the verb 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 entire 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. Hammerly et al. (2019) argue that response bias towards acceptance might be responsible for the observed grammatical asymmetry in speeded acceptability tasks. When they created scenarios where participants expected to see a high proportion of ungrammatical sentences, illusions of ungrammaticality were observed.

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 negative polarity items 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.

2.2 Resumptive Pronouns

Research on resumptive pronouns (RPs) is another area where the consistency of acceptability judgments has come under close scrutiny. As in the case of linguistic illusions, the focus is on sentences that are judged as surprisingly acceptable. But whereas the illusions literature treats those surprising acceptances as errors caused by lower-level mechanisms, in the RPs literature a key question is how to classify cases of surprising acceptability. Experimental studies on RPs have deepened the puzzle by showing that judgments that linguists have taken for granted for decades are more elusive than expected.

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) *The detective interrogated a man who the prosecutor knows why the officer arrested __.
(gap in embedded wh-island)
(6) The detective interrogated a man who$_1$ the prosecutor knows why the officer arrested him$_1$.
(resumptive pronoun) (from Han et al. 2012, ex. 1)

Recent research has challenged the standard view of RPs. In a number of languages, behavioral measures have not consistently reproduced the contrast in pairs like (5) and (6). This is surprising, since informal linguistic judgments reported by linguists typically converge with ratings given by naive participants in large scale acceptability judgment studies (Phillips 2010; Sprouse & Almeida 2012; Sprouse et al. 2013).

Efforts to reconcile this difference between received wisdom and the experimental record have led to a debate about the syntactic status of RPs in island configurations. A popular interpretation of the discrepant acceptability ratings holds that RPs are actually ill-formed. However, because RPs explicitly indicate where the dependency ends and what morphosyntactic features their antecedents have, they provide a production or comprehension advantage, which is responsible for the perception of acceptability in informal introspection (Alexopoulou & Keller 2007; Heestand et al. 2011; Beltrama & Xiang 2016; Chacón 2015, among others).

We consider it unlikely that this proposal is correct. It implies that linguists are bad at distinguishing percepts of well-formedness from plausibility and comprehensibility, and that typical experimental participants are better at this. Most other evidence suggests the opposite. Also, we consider it unlikely that linguists have been fooling themselves for decades over the intuition that RPs improve acceptability for island-crossing dependencies.

A review of a broader set of studies suggests that typical acceptability rating measures might not be the best way to tap into the percept of improvement that linguists report. Ackerman et al. (2018) argue that the choice of task affects judgments of sentences with RPs. Studies that do not consistently find a contrast between RPs and their gapped counterparts tend to be those that use standard acceptability ratings, while studies that do find a contrast tend to be those where participants make an explicit choice between ending an island-violating wh-dependency with a gap or an RP. In this setting, participants prefer island-violating wh-dependencies that end with an RP (e.g. Zukowski & Larsen 2004; Ferreira & Swets 2005; Ackerman et al. 2018). Acceptability ratings may 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 an 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.

The question of how to reconcile linguists' intuitions with findings from large-scale judgment studies is at most an intermediate question. And it is probably an over-simplification to regard the question as figuring out whether unbounded dependencies with RPs are genuinely well-formed or genuinely ill-formed.

Discrete notions of well-formedness applied to entire sentences are probably no more than a useful simplification that helps us to build generalizations at a high level of description. What is really at stake is more likely the question of what are the representations and processes that are involved in RP dependencies, and what is it specifically about the insertion of an RP that (sometimes) leads speakers to be happier with those dependencies. The underlying mental representations and computations are what we most care about. How those are mapped onto quantifiable behavioral responses as percepts of (un)acceptability is an important linking question, but its interest is justified mostly by how it leads us to a clearer understanding of the representations and computations.

There is also a useful methodological lesson here. 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.

A good example of a sub-literature where these issues are worked out in more detail is the literature on voice mismatches in ellipsis, where there are competing detailed hypotheses about the computations that are responsible for gradient judgments (Arregui et al. 2006, Kim et al. 2011).

3 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 debates in 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.

3.1 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, new theoretical questions began to emerge 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 data 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 was the finding 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 have documented 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 Kaufman 1988, a largely forgotten contemporary of the famous Chien & Wexler (1990) study. 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 closely with 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 relatively 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 first led us to study children's pronoun interpretations, but they are at least as interesting, and they bear on theories of human linguistic interpretation at least as much as the questions that we first started with.

3.2 Lexical Activation & Response Probability

A similar story plays out in the study of the role of linguistic context in language understanding. We started with simple high level generalizations and some apparently anomalous findings, but we were led to discover new theoretical questions that we were not even aware of when we started.

There is much recent interest in the role of context in linguistic and psycholinguistic theory. 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).

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, how does that constraint affect the process of recognizing the next word? 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…*) is the phonological input matched against all compatible words in the mental lexicon, or only those that are verbs? 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 of the questions. The TRACE model (McClelland & Elman 1986) treats auditory word recognition as a process of 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 this model. 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.

Although TRACE is not designed to directly model context effects, it is informative to translate context effects into the lexical activation currency of the model. In discussions of syntactic category constraints that we are aware of, it is generally assumed that if syntactic category constrains word recognition, this would mean that only words that match the expected category are considered. This amounts to a strong inhibitory constraint, one that in a verb context would effectively turn off the link between *br…* and *brownie*. It would mean that upon encountering a syntactic context that predicts a verb, the abstract expectation for a verb would need to be translated into an instruction that inhibits links between sounds and lexical entries for all items that are not a verb. This kind of inhibitory mechanism is potentially difficult to implement, and it could be problematic in that a word might become unrecognizable if the context is misheard and the word is then mistakenly ruled out as a candidate. A promising alternative is for links from syntactic category to lexical entries to be facilitatory, such that a verb context boosts the activation of all verbs, while leaving other categories unaffected. The theoretical contrast between inhibitory and facilitatory context effects makes a big difference to the empirical consequences of context effects. And yet the contrast had eluded us until we tried to capture context effects in terms of an explicit model. As we then discovered, the empirical record fits well with a facilitatory effect of context.

Explicit modeling also proved invaluable for understanding the link between lexical activation, i.e., what we really care about, and the various behavioral and neural measures that we use to try to infer lexical activation. 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 is tempting to think of response probabilities as direct reflections of changes in activation, just as acceptability judgments are often treated as transparent reflections of grammatical well-formedness. But an important idea that models like TRACE make clear is that lexical activation and response probability are not the same thing. The transformations that map activations to response probabilities are affected by the details of task and response candidate set.

For example, the visual world paradigm typically tracks visual fixations to a small set of pictured items while participants listen to auditory input. In this paradigm a participant can only fixate on one item at any moment in time. Fixation probabilities on a picture are dependent on the lexical activation of the picture's name. A straightforward mapping is to assume that fixation probability is computed by dividing lexical activation for a given candidate by the sum of the activation values for all response candidates. This means that shifts in fixations to one picture could reflect a shift in activation of that picture's name. Or it could reflect a shift in activation of other pictures' names. A change in activation of the item we care about will only lead to a change in fixation probability if it changes more or less than the rest of the set. This principle turns out to be crucial in interpreting behavioral results, but it is lost without a transparent linking hypothesis between activation and response probability. Insights of this nature certainly do not necessarily require computational modeling techniques. They are in the category of insights that make sense with the benefit of hindsight, but that are easily missed without the aid of modeling.

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 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 (Magnuson, Tanenhaus, & Aslin 2008; Strand, Brown, Brown, & Berg, 2018).

However, we found that when we further controlled the visual world design 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). For example, in a context like *She wanted to browse* we found increased fixations to a picture of a broom, just as in a context like *She chose the brownie*. In neither case was a broom mentioned in the utterance, but the phonological onset *br…* led to increased looks in either case.

Our visual world study showed that incoming sounds can activate any compatible words in the mental lexicon, even when this conflicts with a syntactic constraint. This could mean that the syntactic constraint is simply ineffective. Or it could reflect that the effect of the syntactic constraint is a facilitatory one. Our visual world results do not distinguish those possibilities, but results from other studies lend support to the facilitatory mechanism. For example, a meta-analysis of classic cross-modal priming studies on homophone processing (Lucas, 1999) showed that both meanings of a homophone are activated, even when one meaning conflicts with the context. Lucas found that across the literature (though this effect is very subtle in individual studies) there is more priming for the meaning that is consistent with the context. Most of the studies in this meta-analysis are concerned with semantic context, but the study on syntactic context that is included (Tanenhaus et al. 1979) also fits this pattern. For example, this would mean more priming for the word “look” after hearing “I began to watch” (same category) than after hearing “I bought the watch” (different category).

Furthermore, the facilitatory mechanism makes sense of otherwise puzzling findings from an MEG study of contextual constraints. Earlier studies had discovered that MEG activity patterns correlate with properties of the set of words that are under consideration during auditory word recognition (Ettinger, Linzen, & Marantz, 2014; Gagnepain, Henson, & Davis, 2012; Gwilliams & Marantz, 2015). Gaston & Marantz (2018) built on this finding by testing whether activity elicited in a syntactically constraining context correlates with a syntactically constrained set of words, e.g., only verbs, or with a syntactically unconstrained set of words, i.e., all words compatible with the phonological onset of the word. Gaston and colleagues found that the MEG activity correlated with both the constrained and the unconstrained word sets. This seemed oddly contradictory at first, but in retrospect it is just as predicted by a facilitatory effect of syntactic context on lexical activations.

This example resembles the other examples, in the respect that the starting point was high level generalizations that we regarded as straightforward, but that led us to new theoretical questions that we had not been aware of. What is notable in this case is the role of (relatively simple) computational modeling in helping us to recognize the new questions, and the role of diverse experimental methods in revealing the difference between surface measures and underlying mental processes.

3.3 Role Reversals and the Semantic P600

Sometimes our starting point has been a theoretical controversy in the linguistics literature. In other cases the starting point is a claim that is largely uncontroversial, but that is challenged by a psycholinguistic finding. As elsewhere, what we initially regard as important is often not what turned out to be most important.

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. This finding drew our group's attention because it challenged the largely uncontroversial assumption that syntactic and semantic combinatorics are tightly coupled, i.e., the syntactic structure of a sentence guides how the meanings of words are combined to form larger meanings. 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 proposed that this pattern arose because comprehenders perceived (10) as syntactically anomalous, despite the fact that it is syntactically well-formed. 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 active rather than 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 strong 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 effects reflect syntactic processing. Combining Kim and Osterhout's data with these assumptions leads to the conclusion that semantic interpretation can proceed independent of syntax. This challenges such a basic theoretical assumption that our group began 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 compatible with Kim and Osterhout's claims. However, the P600 effect is 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; Paczynski & 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, in Kim and Osterhout's key second experiment *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 a verb with role-reversed arguments fails to elicit an N400 effect. This 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 see 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 began as an investigation into long-standing claims about syntax-semantics relations in a standard linguistic architecture turned into new questions about the relationship between grammatical information and world knowledge. Still theoretical, certainly interesting, but not the question that we started with.

4 In Search of Linking Hypotheses

In the examples described in Sections 2 and 3 we highlighted the interest of theoretical questions at finer grains of analysis than traditional linguistic theories. But one could object that this side steps the question of what experiments can contribute to theoretical questions at the traditional higher level. Wouldn’t it be good if we could experimentally target the higher level without needing to get tied down in lower level questions? A number of recent lines of work claim to do that using ingenious experimental arguments. Our impression is that the experimental arguments are only ever as good as the linking hypotheses.

4.1 Neurosyntax

A growing body of work uses computational models to predict neural activity during naturalistic reading or listening. For example, a study might model brain activity while participants listen to passages from stories like *Alice in Wonderland* or *The Little Prince*. This is a radically different approach than traditional experiments that use tightly controlled materials that are manipulated to isolate an effect of interest. In naturalistic reading or listening studies the choice of linguistic material is relatively arbitrary, and instead all the action is in the analysis of the linguistic material that is used to model the neural activity.

To take an over-simplified toy example, neural recordings taken during listening to a story could be modeled using a super simple analysis that consists only of a sequence of words, and a slightly more complicated analysis that distinguishes lexical categories. Each model is then fit to the neural data, and the analyst can ask whether the model that includes lexical category distinctions better explains the variability in the data.

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, and a parsing algorithm (top-down, left-corner, bottom-up, etc) determines how to apply the grammar to incrementally presented 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 such dimensions as 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 incoming 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 number of interesting findings have emerged from this approach. Brennan et al. (2016) reeport that node counts from an audiobook story's proposed syntactic structure predict the time-course of participants' fMRI BOLD signal while they listen to that story. Brennan & Pylkkänen (2017) 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 conclusions that can be drawn from this approach are only ever as good as the hypotheses (models) that are used to model the naturalistic input. Typically, the models of parsing operations that are used specify limited detail, and there is a high degree of correlation between competing grammatical and parsing models. This is evident 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.

To take one example, an experiment that compares sequence-based grammars with grammars with hierarchical syntactic structure might find that the hierarchical grammar better captures the observed neural recordings. This supports claims of a hierarchical grammar, or any other linguistic system that better correlates with the hierarchical grammar than the sequential grammar. Since the existence of hierarchical structure in meanings is fairly obvious -- in *two dogs barked*, the expression *two dogs* is a unit to the exclusion of *barked* -- that could capture the advantage of a model with hierarchical structure, regardless of the form of the grammar.

We do not claim that modeling neural responses to naturalistic language input is a fruitless activity. It is a rapidly developing area and it holds much promise. We merely caution that there is no magic solution to isolating the processes or level of analysis that we want to know about. In traditional experimental designs our arguments are only as good as our experimental materials and our (all too often vague) linking hypotheses. We obsess about identifying and removing confounds from our experimental materials. We work hard to ensure that our experimental conditions are represented by diverse items that are representative of the more abstract category that we are interested in. Naturalistic studies abandon the focus on the design of materials, but this does not remove the need to obsess about materials and linking hypotheses. The linking hypotheses are made explicit in the models that are used to describe the language materials. Anything that is not included in the model is not controlled, and hence a potential confound. And the question of whether the abstract categories in the model are associated with diverse examples, e.g., are the nouns in the materials a diverse and representative sample, is often overlooked.

4.2 Structural Priming

Syntactic priming is an experimental technique that has been put forward as a measure that might offer a privileged view into abstract structural properties of sentences, independent of questions about how they are constructed and how they are encoded at more fine-grained levels of analysis. Branigan and Pickering (2017) argue that this allows structural priming to arbitrate some long-standing debates in traditional syntactic theories.

Syntactic priming is the name for the facilitation observed when a recently used structure is re-used, even when there is no lexical overlap between the initial use (prime) and the re-use (target). For example, a passive is more easy to produce when another passive sentence has been recently encountered, even if the two passive sentences have no words in common, aside from closed class morphology (Bock 1989, Bock et al. 1992). Abstract syntactic priming effects have mostly been observed in measures of production probability, e.g., how often speakers describe a given picture prompt with a passive rather than an active. But syntactic priming effects have also been observed in comprehension and production timing measures (Traxler et al. 2014, Momma et al. 2017).

Branigan and Pickering (2017) argue that syntactic priming can be used to arbitrate between transformational and non-transformational grammatical theories. Since this dispute has attracted so much attention over the past few decades, the evidence deserves attention.

Transformational theories of grammar have always maintained that there are several levels of syntactic representation; 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).

Branigan and Pickering (2017) argue for a single level ("monostratral") syntax without transformations, based on evidence for priming between sentences that are surface identical but that are argued to be structurally different in transformational theories.

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 & Rappaport Hovav 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 verb 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. More specifically, they argue that this shows that 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 arguments against transformational theories are only as good as the linking hypothesis that connects structural similarity to syntactic priming effects. These linking hypotheses are underdeveloped, despite the large amount of empirical research on structural priming. All grammatical theories agree that there is some degree of shared structure between English unaccusatives and unergatives. And all also agree that there is some difference between unaccusatives and unergatives, in order to capture their semantic differences, which also impact various structural diagnostics (Levin & Rappaport Hovav 1995), and also gives rise to differences in production planning (Momma et al. 2018). The similarities between unaccusatives and unergatives that everybody agrees upon could be sufficient to drive syntactic priming effects. Branigan and Pickering would like to use syntactic priming as evidence for a lack of differences between a pair of structures. But there is little evidence that syntactic priming is an effective tool for such arguments.

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).

4.3 Memory Access Diagnostics

A different experimental approach to diagnosing transformations is pursued by Xiang and colleagues using evidence from speed-accuracy tradeoff (SAT) and memory interference paradigms in studies on Mandarin Chinese (Xiang et al. 2014; Xiang et al. 2015). A strength of these studies is that they rely on explicit and independently motivated linking hypotheses.

Mandarin differs from English in the surface form of wh-questions. Whereas English fronts wh-phrases to a position that indicates the scope of the question as a direct or indirect question, leaving the thematic position of the wh-phrase empty, Mandarin adopts a wh-in-situ strategy, where the wh-phrase occupies the thematic position and the scope of the question must be recovered from other cues.

Transformational analyses since at least Huang (1982) have argued that English and Mandarin wh-questions are more structurally similar than they appear on the surface. Under these accounts, Mandarin wh-questions involve a structural dependency between the thematic position and the scope position, just as in English. The only difference between the languages lies in which piece of this structural dependency is signaled overtly.

Xiang and colleagues apply to Mandarin two paradigms that have been used to diagnose memory access processes in sentence comprehension. They argue that the processing of Mandarin wh-in-situ constructions is sensitive to the length of the dependency between the thematic position and the scope position of the wh-phrase, thus indirectly providing evidence for the online construction of (invisible) wh-dependencies. To probe this question experimentally, they assume that covert wh-movement involves the retrieval of previously-encountered syntactic structures, specifically, a clause-edge position that marks the scope position of the wh-phrase.

By assuming that wh-dependency formation involves memory retrieval, Xiang and colleagues (2014) justify their use of SAT, a paradigm where there is a relatively clear 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 generally accepted interpretations of these parameters. Distance manipulations that give rise to rate differences implicate a serial search and retrieval process. Distance manipulations that lead to asymptote differences implicate parallel access in content-addressable memory (CAM). See McElree (2006) for more extensive discussion. Xiang and colleagues find that the distance between scope and thematic positions in Mandarin is associated with asymptote differences, motivating the claim that invisible wh-dependencies are formed via a parallel access process. Similarly, the assumption of a memory retrieval process justifies the use of a memory interference logic in self-paced reading paradigms, leading to a similar conclusion (Xiang et al. 2015).

The conclusions from these studies on Mandarin wh-dependencies are open to question, as always. But the linking hypotheses are sufficiently explicit that it is clear what is at stake.

5 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 fact, many of these questions are amenable to experimental investigation and have led us to new insights.

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 limited theoretical insight. But once we allow that new empirical approaches reveal questions that we were not able to address or 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*, 83, 110-160.

Arregui, A., Clifton, C. Jr., Frazier, L., & Moulton, K. (2006). Processing verb phrases with flawed antecedents: The recycling hypothesis. *Journal of Memory and Language,* 55, 232-246.

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. (1989). Closed-class immanence in sentence production. *Cognition*, 31, 163-186.

Bock, K., Loebell, H., & Morey, R. (1992). From conceptual roles to structural relations: Bridging the syntactic cleft. *Psychological Review, 99,* 150-171.

Bock, K., & Miller, C. A. (1991). Broken agreement. *Cognitive Psychology*, 23, 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 edn*.* Wiley Blackwell.

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

Chacón, D. A. (2015). *Comparative psychosyntax*. Doctoral dissertation, University of Maryland.

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*. 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.

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

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.

Ettinger, A., Linzen, T., & Marantz, A. (2014). The role of morphology in phoneme prediction: Evidence from MEG. *Brain and Language*, *129*, 14-23.

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. In A. Cutler, ed., *Twenty-first century psycholinguistics: Four cornerstones*, 263-278.

Fodor, J. A. (1983). *The modularity of mind.* Cambridge, MA: MIT Press.

Gagnepain, P., Henson, R. N., & Davis, M. H. (2012). Temporal predictive codes for spoken words in auditory cortex. *Current Biology*, *22*(7), 615-621.

Gaston, P., Huang, 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.

Gaston, P., & Marantz, A. (2018). The time course of contextual cohort effects in auditory processing of category-ambiguous words: MEG evidence for a single "clash" as noun or verb. *Language, Cognition and Neuroscience*, *33*, 402-423.

Gwilliams, L., & Marantz, A. (2015). Non-linear processing of a linear speech stream: The influence of morphological structure on the recognition of spoken Arabic words. *Brain and Language*, 147, 1-13.

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, 69-101.

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

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.

Han, C., Elouazizi, N., Galeano, C., Görgülü, E., Hedberg, N., Hinnell, J., Jeffrey, M., Kim, K., & Kirby, S. Processing Strategies and Resumptive Pronouns in English. (2012). In N. Arnett & R. Bennett, eds., *Proceedings of the 30th West Coast Conference on Formal Linguistics*, Somerville, MA: Cascadilla Proceedings Project, pp. 153-161.

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, MA.

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.

Kim, C., Kobele, G. M., Runner, J. T., & Hale, J. T. (2011). The acceptability cline in VP-ellipsis. *Syntax,* 14, 318-354.

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.

Levin, B. & Rappaport Hovav, M. (1995). *Unaccusativity: At the syntax - lexical semantics interface*. Cambridge, MA: MIT Press.

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.

Lucas, M. (1999). Context effects in lexical access: A meta-analysis. *Memory & Cognition*, *27*(3), 385-398.

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.

Mantegna, 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., Kraut, R., Slevc, L. R., & Phillips, C. (2017). Timing of syntactic and lexical priming reveals structure building mechanisms in production. Talk at the 30th annual CUNY Conference on Human Sentence Processing. Cambridge, MA.

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 on Human Sentence Processing. Gainesville, FL.

Momma, S., Slevc, L. R., & Phillips, C. (2018). Unaccusativity in sentence production. *Linguistic Inquiry, 49*, 181-194.

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., Cash, S.S., Naccache, L., Hale, J.T., Pallier, C. & Dehaene, S. (2017). Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences*, 114, 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, 47-88.

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

Paczynski M. & Kuperberg G. R. (2011). Electrophysiological evidence for use of the animacy hierarchy, but not thematic role assignment, during verb argument processing. *Language and Cognitive Processes*, 26(9):1402–1456.

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*. Somerville, MA: Cascadilla Press, pp. 576-587

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.* Stanford, CA: CSLI Publications, pp. 49-64.

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. In G. Gaskell, ed., *Oxford Handbook of Psycholinguistics*, pp. 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, 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, 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.

Traxler, M. J., Tooley, K. M., & Pickering, M. J. (2014). Syntactic priming during sentence comprehension: Evidence for the lexical boost. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*: 905-918.

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.

Wellwood, A., Pancheva, R., Hacquard, V., and Phillips, C. (2018). The anatomy of a comparative illusion. *Journal of Semantics*, 35, 543–583.

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, 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). The production of sentences that we fill their gaps. *Poster presented at the 17th annual CUNY Sentence Processing Conference, University of Maryland.*