Oxford Handbook of English Grammar

Chapter 3

Grammar and the use of data

Jon Sprouse
University of Connecticut
jon.sprouse@uconn.edu

Carson T. Schütze
University of California, Los Angeles
cschutze@humnet.ucla.edu

Abstract:

Over the past twenty years, linguists have taken a renewed interest in the data that underlies grammatical theories. In this chapter we review five types of data: corpus data, acceptability judgments, reading times, electrophysiological data (EEG/MEG), and hemodynamic data (specifically fMRI). The approach we take for each is slightly different, as each data type occupies a different role in grammatical theory construction. For corpus data, we defer to chapter 4 (this volume) for a detailed review, and instead focus on the reasons why some linguists prefer experimental data over (observational) corpus data. For acceptability judgments, we review their role in theory construction because they currently form the vast majority of data used for the construction of grammatical theories. For reading time data, we review the logic that has been used to search for consequences of grammatical theories in real time sentence processing. For electrophysiological data, we observe that there is relatively little connection between the electrophysiological literature and the grammatical literature, and therefore review the basic results from the ERP literature as a first step toward encouraging closer ties between the two fields. For hemodynamic responses, we review the research into two brain areas that have been argued to be implicated in syntactic processing (left inferior frontal gyrus and left anterior temporal lobe), as this seems like the best starting place for exploring the relationship between grammatical theories and neurobiology. In the end, it is our hope that this chapter will serve as a useful starting point for thinking about the use of data in grammatical theories for both linguists and non-linguists.

Related articles: Argumentation in grammar; Grammar and corpus methodology; Generative approaches.

1. Introduction

Over approximately the past twenty years, linguists have taken a renewed interest in the data that underlies grammatical theories. This has taken the form of large scale reviews of data in syntax (e.g., Schütze 1996), textbooks (e.g., Cowart 1997), proposals for new experimental techniques (e.g., Bard et al. 1996), enticements to widespread adoption of formal experimental methods (e.g., Featherston 2007), proposals for new grammatical models that can accommodate different data types (e.g., Keller 2000, Featherston 2005), criticisms of informally collected data (e.g., Gibson and Fedorenko 2013), and finally chapters like this, which attempt to provide a useful summary of our current state of knowledge about data in syntax. Here we will discuss five types of data: corpus data, acceptability judgments, reading times (self-paced reading and eye-tracking), electrophysiological methods (EEG and MEG), and hemodynamic methods (specifically fMRI). It is important to note that in principle, there is no privileged data type in linguistics. Any data type that can bear on the nature of the grammar is a potential source of data for theorists. The wide spectrum of possible data types arises because many linguists assume that there is a single grammatical system that plays a substantive role in both language comprehension and language production (Marantz (2005) calls this the *Single Competence Hypothesis*). Given this assumption, all language behaviors, and all neurobiological responses related to language behaviors, are potential sources of information about the grammar. We will focus on these five data types because they play the largest role in modern studies of grammar. In the sections that follow, we will discuss each data type in turn. The approach that we take for each data type will be slightly different, because each data type has historically played a different role in grammatical theory construction.

2. Corpus data

Chapter 4 by Sean Wallis in this volume provides an excellent discussion of the use of corpus data in linguistics. We won't double that effort here (and couldn't do nearly as good a job). Instead, we will briefly discuss some potential disadvantages of corpus data that lead some linguists to explore other data types. For example, two fundamental questions in studies of grammar are (i) Is a given sentence grammatical or ungrammatical (i.e., possible or impossible in the language)?, and (ii) For ungrammatical sentences, what is the property that causes the ungrammaticality? Because corpus data is fundamentally observational, it cannot be used to definitively answer these two questions. For the first question, one potential strategy would be to use the presence or absence of a sentence in the corpus as a proxy for grammaticality or ungrammaticality, respectively. The problem with this approach is that presence and absence are influenced by factors other than the grammar: it is possible for grammatical sentences to be absent from a corpus simply due to sampling (the relevant sentence was accidentally not uttered during corpus generation), and it is possible for ungrammatical sentences to be present in a corpus due to speech errors or the inclusion of non-native speakers in the sample. This leads

linguists interested in the first question to seek out data types that more directly test the impact of the grammar while controlling for other factors that might influence the outcome—in other words, a controlled experimental setting. For the second question (what properties cause the ungrammaticality), the concern among some linguists is that corpus data can only reveal *correlations* between grammatical properties and presence/absence in the corpus; corpus data cannot directly reveal *causal* relationships. Again, this leads linguists who are interested in the mechanisms underlying ungrammaticality to seek out controlled experimental methods that can be used to directly manipulate grammatical properties to reveal causal relationships with grammaticality. As chapter 4 demonstrates, there are plenty of interesting research questions one can investigate using corpus data, but for linguists interested in ungrammaticality and its mechanisms, corpus data tends to be less useful than experimental methods.

3. Acceptability judgments

3.1 What are acceptability judgments?

An acceptability judgment is simply the act of judging whether a sentence is "acceptable" in a given language. But this simple definition belies the complexity of what one means by "acceptable". One common assumption in linguistics is that the act of comprehending a sentence automatically (in the sense of an automatic cognitive process) gives rise to an evaluation of that sentence along multiple dimensions: the grammaticality of the sentence, the plausibility of the meaning, the processing difficulty associated with comprehending the sentence, etc. Another common assumption is that those multiple dimensions tend to be (automatically, and therefore subconsciously) combined into a single percept. It is that multi-dimensional percept that linguists call "acceptability." Acceptability judgments can then be defined as a conscious report of the automatic evaluation of the acceptability of a sentence, elicited for experimentally designed sentence types (either in an informal setting, as has been typical in linguistics, or in a formal experiment, as has become more common over the last two decades). In short, acceptability judgments are a behavioral response that can reveal information about the grammaticality of a sentence, if the experimenter controls for other factors that influence acceptability (plausibility, processing, etc.). Therefore, the goal of an acceptability judgment experiment is to isolate a potential difference in grammaticality between two (or more) conditions, while holding differences in other properties constant.

There are at least three pieces of information that acceptability judgments can potentially provide that are relevant for constructing grammatical theories. The first is the presence/absence of an effect – whether there is a difference in acceptability between two (or more) conditions. In many ways this is the minimum piece of information that may be relevant for a grammatical theory. Assuming that the two (or more) conditions were well controlled, such that the only difference between them was the grammatical property of interest, the presence of a difference tells us that the grammatical property has an effect on acceptability. The second potential piece of information is the effect size – the size of the difference between conditions. Again, assuming that the conditions were well controlled, the effect size tells us how big an impact the grammatical manipulation has. The third potential piece of information is the location of the

conditions on the scale of acceptability.[1] These three pieces of information can be used in different combinations by linguists to construct and evaluate grammatical theories.

Figure 1 highlights these three pieces of information. The raw data for Figure 1 comes from 274 sentence types (conditions) that form 137 two-condition phenomena. For example, one of the phenomena is comprised of these two sentences (from Sobin 2004):

(1)   a.   Some frogs and a fish is in the pond.
          b.   Some frogs and a fish are in the pond.

By hypothesis, the sentence in (1a) is ungrammatical because it does not respect the subject-verb agreement properties of English (whereas as (1b) does). Sprouse, Schütze, and Almeida (2013) tested this phenomenon and 149 others that were randomly sampled from the journal *Linguistic Inquiry* in formal judgment experiments using a 7-point Likert-like scale (see also section 3.2 below). In Figure 1, we use the 137 phenomena that showed statistically significant effects to illustrate the three pieces of information that judgment experiments make available to linguists.
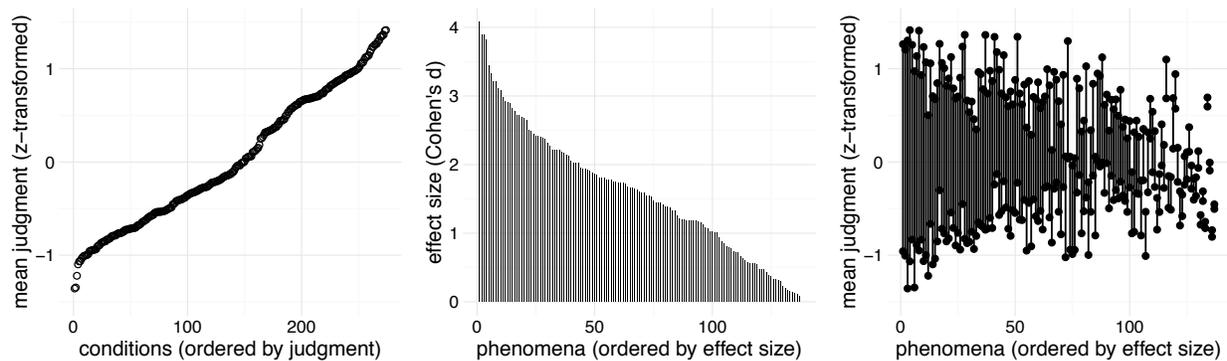

Figure 1: Three illustrations of the types of information available from acceptability judgments. Raw data is 137 pairwise phenomena randomly sampled from *Linguistic Inquiry* and tested by Sprouse, Schütze, and Almeida 2013. The leftmost panel shows location information. The middle panel shows effect size information. The rightmost panel combines the two. The middle and rightmost panels also convey the presence/absence of an effect.

The leftmost panel of Figure 1 takes all 274 sentence types, and orders them in ascending order according to their mean judgment (after a z-score transformation, which removes some types of scale bias; see Schütze and Sprouse 2013 for more discussion). This panel therefore highlights location information: some sentences are very clearly on the low end of the scale, others at the high end, and still others in the middle. The middle panel of Figure 1 highlights effect sizes. Each vertical bar represents one phenomenon, with the height of the bar representing the size of the effect, which in this case is the size of the difference between the mean ratings of the two

---

[1] Location on the scale is a complex topic in its own right, as the way that participants use a scale will be influenced by the instructions that they are given (How are the points on the scale labeled? Are example items given for the points on the scale?) and the content of the experiment itself. See Schütze and Sprouse 2013 for a discussion of the details of acceptability judgment tasks.

conditions in each phenomenon (reported here in Cohen's *d* units, which is a standardized measure of effect size that is common in the experimental literature). The rightmost panel combines these two pieces of information by plotting the ratings of the two conditions in each phenomenon in a vertical pair (thus showing location information), and connecting them with a line (thus showing effect size information). Both the middle panel and the rightmost panel also highlight the presence/absence of an effect in a way that the leftmost panel does not.

   In practice, which of the three pieces of information the linguist decides to use depends upon the specific grammatical theory being investigated, and the way it is used depends on the specific argument that the linguist wishes to make (see chapter 2 by Bas Aarts on linguistic argumentation). That said, some basic patterns of use do emerge. First, all studies using acceptability judgments report the presence/absence of an effect, as this is the minimum required to demonstrate that a grammatical manipulation is potentially relevant to a theory. Second, grammatical theories that rigidly divide sentences into two types (grammatical and ungrammatical) tend to also use location on the scale as a potential piece of information about whether a sentence should be classified as grammatical (high on the scale) or ungrammatical (low on the scale). The use of scale location information is complex in its own right because there is no necessary connection between location on the scale and grammaticality. Even if extra-grammatical properties are controlled across the conditions in the experiment as discussed above, participants can rate individual sentences high or low because of the extra-grammatical properties of that individual sentence. This is particularly salient in classic examples of mismatches between acceptability and grammaticality, such as the acceptable-but-ungrammatical comparative construction in (2) (from Montalbetti 1984), which appears acceptable but has no coherent meaning, and the unacceptable-but-(by-hypothesis)-grammatical doubly center embedded relative clause in (3) (from Miller and Chomsky 1963), which appears unacceptable, but is most likely unacceptable due to the difficulty of processing the nested relative clauses.

(2) More people have been to Russia than I have. (acceptable-but-ungrammatical)
(3) The food the dog the cat scratched ate spoiled. (unacceptable-but-grammatical)

Despite the possibility of mismatches between location information and grammaticality, it is not uncommon for location information to play a role in linguistic argumentation about grammars that divide sentences into two sets (e.g., an author might assume a transparent mapping between the two halves of the scale of acceptability and ungrammaticality/grammaticality). The third piece of information, effect size information, tends to play a role in studies that seek to compare the constraints that make up the grammar. This can either be because the grammatical theory under investigation distinguishes more than two levels of grammaticality (e.g., Keller 2000, Featherston 2005), or because the investigation is exploring properties that might distinguish two constraints from one another (e.g., Chomsky 1986).

3.2 Are the acceptability judgments published in the literature valid?

Perhaps the most important debate in the acceptability judgment literature concerns the validity of the judgments that have been published in the literature so far. Since the earliest days of acceptability-based linguistic theories there has been a concern that the relatively informal methods that linguists tend to use to collect acceptability judgments might lead to invalid data, and therefore incorrect theorizing (e.g., Hill 1961, Spencer 1973). Over the past 20 years, as

formal experimental methods for judgment collection have gained in popularity, this question has arisen more and more often, with many linguists asking whether the field should shift entirely to formal acceptability judgment collection methods (e.g., Bard et al. 1996, Schütze 1996, Cowart 1997, Ferreira 2005, Featherston 2007, Gibson and Fedorenko 2010, Gibson and Fedorenko 2013, among many others). Nearly every linguist who has written about this issue agrees that there are benefits to formal acceptability judgment experiments, making them an important tool in the syntactician's toolkit. What is less clear is whether informal judgment collection methods should co-exist with formal experimental methods in that toolkit, and whether the informally collected judgments that have been published in the literature to date should be considered valid. The central concern (e.g., from Gibson and Fedorenko 2013) is that linguists tend to solicit informal judgments from other linguists. Because professional linguists are aware of the theoretical issues at play for a given judgment, it is possible that their judgments might be (subconsciously) biased. A related concern is the fact that linguists tend to collect small numbers of judgments, leading to the possibility that the results they obtain are not representative of the population. Similarly, linguists tend to use a small number of example items, leading to the possibility that the results they obtain are not representative of all of the possible tokens of a given construction. If any of these potential problems were actual problems, the published judgment literature would not be a solid foundation for constructing grammatical theories.

Sprouse, Schütze, and Almeida (2013) and Sprouse and Almeida (2012) took a first step toward investigating these concerns by directly comparing published informal judgments with formal judgments that they collected using the best practices of formal experimental work (e.g., naïve participants, large samples, multiple items per condition, frequentist and Bayesian statistical analyses, etc.). Their argument is that, though this does not settle the question of which method is best, it does begin to address the question of how much impact these design choices might have on the data. If the two methods overlap substantially, then either both methods are valid or both are invalid. If the two methods diverge substantially, then either one method is valid and one is invalid, or both are invalid in different ways. Sprouse, Schütze, and Almeida (2013) found a 95% (±5%) overlap when they re-tested a random sample of 150 two-condition phenomena (300 sentence types) from the journal articles published in *Linguistic Inquiry* (*LI*) between 2001 and 2010. Sprouse and Almeida (2012) found a 98% overlap when they re-tested all of the data points published in Adger's (2003) *Core Syntax* textbook. Taken together, these results suggest that data from informal and formal methods overlap to a very high degree. This suggests that replacing informal judgment data with formally-collected judgment data would have little impact, at least for work on English syntax. Of course, studies such as these are just the beginning of these kinds of investigations. Future work should explore other languages, other judgment types (e.g., semantic judgments), and other facets of the data (e.g., gradience).

3.3 Using formal judgment methods to explore the nature of the grammar

Formal judgment experiments are not limited to investigating the validity of informal judgments; they can also be used to push the boundaries of the theory of grammar itself. One way formal judgment experiments can do this is by using higher-order experimental designs to isolate and quantify putative grammatical effects while controlling for many of the other factors that potentially influence acceptability judgments. As a concrete example, let us consider the *whether*

island violation in (4), which we have labeled with a star to indicate that it is generally judged to have low acceptability in US English:[2]

(4)   *What do you wonder whether Jack bought __ ?

One common analysis in the generative literature is to postulate a grammatical constraint that prevents movement of a WH-phrase like *what* out of an embedded question – the embedded question is metaphorically an "island" for this kind of movement (Chomsky 1964; Ross 1967, Chomsky 1973), so this is called a *whether island violation*. If we wanted to study the necessity of this constraint, we would first want to isolate the *effect* of moving out of the embedded *whether*-clause over and above other possible effects that are in this sentence, such as the effect of having a long-distance movement in a sentence, and the effect of having an embedded question in a sentence (both of which might lower acceptability independently of the potential *island effect*). A 2×2 factorial design can solve this problem by creating a sequence of subtractions that eliminate the extra confounds. Take the four sentences in (5):

			EMBEDDED STRUCTURE	MOVEMENT DISTANCE
(5)	a.	Who __ thinks that Jack bought a car?	declarative |	short
	b.	What do you think that Jack bought __ ?	declarative |	long
	c.	Who __ wonders whether Jack bought a car?	interrogative |	short
	d.	What do you wonder whether Jack bought __ ?	interrogative |	long

If we make the subtraction (5a – 5b), we isolate the effect of a long-distance as opposed to a short-distance (local) movement. The difference (5a – 5c) isolates the effect of the embedded question without any movement out of it. The difference (5a – 5d) combines three effects: the effect of a long-distance movement, the effect of an embedded question, and crucially, the effect of moving out of an embedded question (the *island effect*). By subtracting the first two from the third, we can isolate the island effect, as in (6):

(6)   island effect = (5a – 5d) – (5a – 5b) – (5a – 5c)

In this way, a series of three subtractions can isolate island effects even though there are two other effects present in the critical sentence. This design is called a 2×2 factorial design because there are two factors, STRUCTURE and DISTANCE, and each factor has two levels (declarative vs. interrogative and short vs. long).

    There are a number of benefits to 2×2 factorial designs. The most important, of course, is allowing us to isolate the effect of interest. They also allow us to isolate two other effects, one for each factor in the design. Further, they lend themselves to a perspicuous visual interpretation, as illustrated in the two contrasting hypothetical outcomes in the two panels of Figure 2. If there is no effect unique to the island violation in (5d) over and above the effects of the two other factors, plotting the four conditions in a pattern known as an interaction plot will yield parallel lines, as in the left panel of Figure 2. However, if there is a further effect over and above those

---

[2] In this and subsequent examples, the underscore indicates the position from which a WH-phrase has moved, according to movement-based theories of generative syntax.

two factors, such as an island effect in the design in (5), the two lines will not be parallel, as in the right panel of Figure 2.
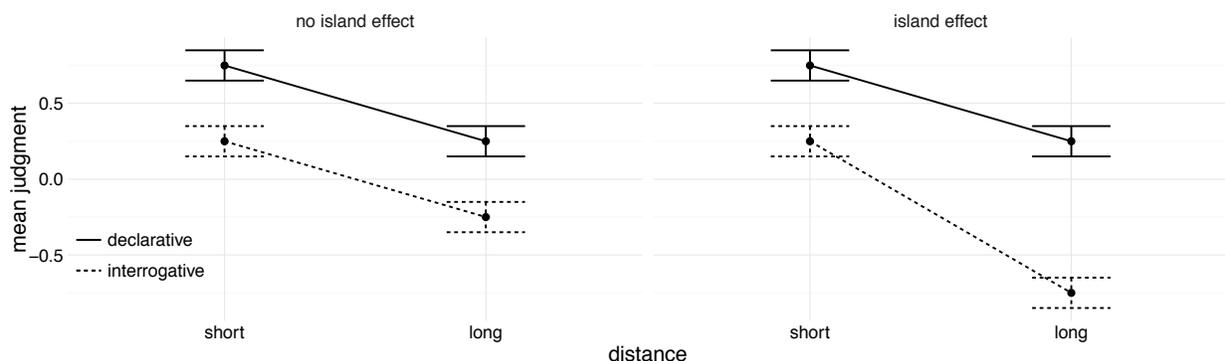

Figure 2: The visual logic of 2×2 factorial designs, illustrated using the *whether* island effect design from (5). The left panel shows main effects for the two factors (STRUCTURE and DISTANCE), but no island effect over and above those two factors. The right panel shows an island effect (interaction) over and above those two main effects.

Another benefit of 2×2 factorial designs is that they allow for the use of relatively standard statistical tests such as two-way ANOVAs and omnibus linear mixed effects models. The effect of interest (e.g., island effect) in these designs will show up as the interaction term. One final advantage of 2×2 factorial designs is that, though they only quantify three effects (one for each factor plus the interaction effect), they theoretically allow us to control for an infinite number of potential confounds as long as the confounds are distributed across conditions such that they subtract out at the end of the subtraction steps.

The theoretical value of the factorial design really becomes apparent when one considers analyses that attempt to explain acceptability judgment phenomena as being caused by factors outside the grammar. For example, Kluender and Kutas (1993b) proposed that island effects (the interaction in the right panel of Figure 2) may not be due to syntactic constraints in the grammar, but rather to the effect of the two independent processing costs of a long-distance dependency and an embedded question interacting with each other, perhaps because they both draw on the same limited pool of working memory resources, such that the parser's attempt to deploy both leads to a larger cost than one would expect from the linear sum of the two in isolation. With a 2×2 factorial design, we can isolate that effect (the superadditive interaction), and attempt to study its properties to see if it is more likely to come from the grammar or more likely to come from the working memory system. For example, Sprouse et al. (2012) looked to see if the superadditive effects for four island effects in English correlate with two working memory measures, which is one possible prediction of the Kluender and Kutas (1993b) proposal.

Factorial designs also have the potential to reveal novel data for the construction of grammatical theories – such as the existence of superadditive interaction patterns without any single sentence being in the lower half of the acceptability scale (e.g., Featherston 2005, Sprouse et al. 2016, Villata et al. 2016). What is interesting about these effects is that the superadditivity suggests the possibility of a grammatical constraint at work, but the fact that all of the sentences are in the top half of the acceptability scale suggests that this potential constraint is not causing the extreme unacceptability that is typically associated with constraint violations. The question

8

facing the field is how to capture these types of effects in grammatical theories. This is one of the current areas of debate in the acceptability judgment literature.

4. Reading times: self-paced reading and eye-tracking

The next three data types that we will discuss (reading times, electrophysiological measures, and hemodynamic measures) are far more prevalent in psycholinguistics and neurolinguistics, two domains that focus on language processing, than theoretical linguistics, which focuses on grammatical theories. This means that the value of these data types is primarily tied to the relationship between theories of language processing and theories of grammar. Nonetheless, we believe it is valuable to provide brief reviews of these three data types, not only because readers may encounter them in the literature, but also because we believe that the future of linguistics is one where there is a closer integration between theories of grammar and theories of language processing.

There are two primary reading time measures used in psycholinguistics: self-paced reading, and eye-tracking. The most prevalent version of self-paced reading uses what is called a "moving window" to more accurately mimic natural reading (Just et al. 1982). In a moving window self-paced reading task, the words of a sentence are replaced with underscores, one per character. The underscores representing the sentence are presented in their entirety on the screen. The participant can then reveal each word in succession by pressing a key on a keyboard or a button on a response box. The computer measures the amount of time between button presses, yielding a measure of the amount of time (typically in milliseconds) that it takes to read each word. In reading-based eye-tracking, the entire sentence is presented on screen (without masking by underscores). A sophisticated camera then records the movements of the participant's pupils as they read the sentence naturally. Analysis algorithms can then be used to determine how long the participant fixated on each word (or multi-word region) of the sentence, as well as several secondary measures such as if (and for how long) the participant looked back to previously read material. Though the method of collection of reading times differs between self-paced reading and eye-tracking, the logic applied to these two data types is the same: one can compare the reading times between two sentences that differ by a specific property of interest at one or more critical words; if the critical word(s) differ in reading times between the two sentences, then one can infer that the processes deployed to understand those sentences at the critical word(s) differed in either quality or quantity. Reading time measures such as self-paced reading and eye-tracking have been used in this way to build complex theories of sentence processing.

As the previous paragraph makes clear, reading times are primarily used to investigate sentence processing, not grammatical theory. But under the assumption that there is a predictable relationship between grammar and sentence processing (setting aside the exact nature of that relationship), it is possible to look for the consequences of proposals from grammatical theories in real-time sentence processing. If found, these effects can increase our confidence in such proposals, as there would now be convergent evidence from multiple sources. One concrete example of this is a series of findings showing that the human sentence processor appears to respect syntactic island constraints (briefly discussed in section 3) in real time during incremental sentence processing. These studies are predicated upon the *Active Filling Strategy* (Frazier and Flores d'Arcais 1989), a parsing strategy that appears to be operative in most, if not all, human languages. In effect, the active filling strategy says that the parser attempts to complete long-distance dependencies at the first viable location. For the wh-dependencies that

characterize questions in English, this means that the parser will attempt to associate the wh-word (called the *filler* in the sentence processing literature) with the first location that could potentially host the wh-word (this location is often called the *gap*, with the full dependencies called *filler-gap* dependencies) – typically a verb or preposition. The active filling strategy leads to several useful effects in sentence processing. For space reasons, we will use only one as an example: the filled-gap effect. In the filled-gap effect, participants experience a reading time slow-down when presented with filler-gap dependencies in which the first potential gap location is occupied (or filled) by another argument (Crain and Fodor 1985, Stowe 1986). For example, Stowe 1986 found a reading time slow-down using self-paced reading at the critical word *us* in sentence (7a), which contains a filler-gap dependency relative to sentence (7b), which contains no dependency. Under the active filling strategy, the parser associates the filler *who* with the verb *bring*, but when *us* is encountered, the parser must reanalyze the structure (and eventually associate the filler with the true gap after the word *to*), leading to a reading time slow-down, which we call the filled-gap effect.

(7) a. My brother wanted to know *who* Ruth will bring *us* home to __ at Christmas.
     b. My brother wanted to know if Ruth will bring *us* home to Mom at Christmas.

In this way, the filled-gap effect is evidence that the active filling strategy is operative, and that the location of *us* is considered a viable gap location by the parser.

Having established the filled-gap effect as consequences of the active filling strategy, psycholinguists are able to probe the processing consequences of syntactic island constraints. Syntactic island constraints prohibit gaps from occurring inside of certain structures. The active filling strategy attempts to associate fillers with the first gap location that the parser encounters. This leads to the following question: Will the parser attempt to posit a gap inside of an island structure? Stowe 1986 investigated this question for the Subject island constraint using the filled-gap paradigm. The Subject island constraint (Huang 1984) prohibits gaps inside of complex subjects in English as in (8):

(8) *What did [the joke about __] offend the audience?

Stowe 1986 looked for a filled-gap effect inside of complex subjects using (9a), which contains a filler-gap dependency and filled-gap location in the complex subject, and (9b), which contains no filler-gap dependency (thus acting as a control condition). If the active filling strategy attempts to fill gaps inside of complex subjects (contrary to the Subject island constraint), there should be a filled-gap effect (a reading time slow-down) at the noun *Greg's*. If the active filling strategy respects the Subject island constraint, there should be no filled gap effect.

(9) a. The teacher asked *what* the silly story about *Greg's* older brother was supposed to mean.
     b. The teacher asked if the silly story about *Greg's* older brother was supposed to mean anything.

Stowe 1986 found no filled-gap effect in this paradigm (but did find an effect with similar complex NPs in object position), suggesting that the parser respects the Subject island constraint

in real time by suppressing the active filling strategy when gap locations are prohibited by the grammar.

This basic finding has been replicated using both the filled-gap effect and other consequences of the active filling strategy such as the plausibility effect for a number of islands, and a number of languages (see Phillips 2006 for a comprehensive list as of that time). It has also led to a number of more sophisticated explorations of the interaction of syntactic island effects and sentence processing, including the effect of parasitic gaps (Phillips 2006), and differences between parasitic gaps and across-the-board movement (Wagers and Phillips 2009). The literature on the active filling strategy has also inspired similar investigations in the domain of pronoun coreference, where a similar active search strategy has been found with respect to the search for antecedents for pronouns (van Gompel & Liversedge, 2003), using a reading time slow-down effect similar to the filled-gap effect (called the gender mismatch effect). This literature has also explored the effect of grammatical constraints on coreference dependencies (called Binding Constraints, Chomsky 1981) on this active search for an antecedent (Sturt 2003; Kazanina et al. 2007). There is even recent work exploring a similar notion of an active search for the antecedent in ellipsis constructions (Yoshida, Dickey, and Sturt 2013). For space reasons we can't summarize them all here, but the general logic is the same in each – one can use a well-established reading time effect to probe the interaction of a parsing strategy with a constraint from the grammatical literature.

5. Electrophysiology: EEG and MEG data

Much like reading times, electrophysiological responses are primarily used in the psycholinguistics literature to investigate sentence processing. They are rarely used to investigate the grammar directly. That said, similar to reading times, the assumption of a predictable relationship between grammar and parser can allow some limited forms of inference (or at least corroboration) to flow between the two fields. To be fair, there has been much less of this in the electrophysiological literature than the reading time literature. This doesn't seem like a necessary fact, so in the spirit of making this chapter maximally useful to future researchers, we will review some of the findings in the EEG and MEG literature that show the most promise for connections with grammatical theory.

Electroencephalography (EEG) and magnetoencephalography (MEG) are two sides of the same coin: EEG measures (some of) the electrical activity generated by the brain (typically through electrodes placed on the scalp; though also potentially through electrodes placed directly in the cortex), and MEG measures the magnetic fields generated by (some of) the electrical activity of the brain (through sensors positioned around the head). The analysis of EEG and MEG results is relatively complicated, but the basics are as follows. First, the electrical activity of the brain is an alternating current, which means that both EEG and MEG data can be characterized as time-varying waves (either the oscillations of electrical voltage in EEG, or of magnetic fields in MEG). This means that all M/EEG data analysis can take advantage of the mathematics of waves. Second, nearly all research in neurolinguistics is event-related. This means that the measurements are time-locked to a specific event (such as the presentation of a word), so that neurolinguists can explore the effects of that event on cognition. Third, M/EEG activity can be divided into two types: evoked-activity and induced-activity. Both evoked and induced activity are event-related, which means that the EEG response is recorded relative to specific (experimentally controlled) event such as the presentation of a word. The difference

between the two types of activity lies entirely in the amount of phase-locking across tokens of the event. Evoked activity is activity that is phase-locked to the event: the peaks of the waves from one token of the event line up with the peaks of the waves of another token (and the same holds for the troughs of the waves). Induced activity is not phase-locked across tokens of the event. This means that it is possible for the peaks of waves from one event to line up with the troughs of waves from a second token of the event. Evoked and induced activity require different analysis techniques (because of the potential for destructive interference with induced activity). Evoked and induced activity may also represent distinct neurophysiological events, though their precise interpretation is an active area of research. The vast majority of research using M/EEG in neurolinguistics has focused on evoked-activity in the form of event-related potentials (ERPs), which are simply changes in the amplitude of electrical potentials over time in evoked EEG activity, and event-related fields (ERFs), which are simply changes in the amplitude of magnetic fields over time in evoked MEG activity (see Luck 2005 for a complete introduction to the ERP technique, and see Cohen 2014 for a complete introduction to analysis techniques for induced activity). There have been some recent explorations of induced-activity in the M/EEG literature, but that area is still relatively new inside neurolinguistics, so we won't review it here (but see Bastiaansen et al. 2011 for an excellent overview of that literature to date).

The majority of the EEG literature in neurolinguistics has focused on ERPs. Though many ERPs have been identified in the broader EEG literature, the sentence-level neurolingustics literature tends to focus on four ERPs: the early left anterior negativity (ELAN), the left anterior negativity (LAN), the N400, and the P600. The ERP literature typically describes two facets of ERPs when defining them: their eliciting conditions, and their functional interpretation. Both facets are potentially useful for linking the EEG literature to the grammatical literature. The eliciting conditions can be used to link grammatical theories and ERPs at the level of specific phenomena. For example, the grammatical theory may predict a certain kind of violation in a sentence, and thus predict a specific type of ERP at a specific location in the sentence. The functional interpretation can potentially license the kind of logic we saw in the previous section on reading times. The functional interpretation can indicate what aspect of sentence processing the ERP indexes, and then one can ask whether known sentence processing strategies predict that those aspects of sentence processing should be engaged at the relevant positions in the sentence. As mentioned above, there aren't many great examples of this being used to connect with the grammatical literature, but it is possible in theory. So here we will briefly review the eliciting conditions and potential functional interpretation of the four sentence-level ERPs that readers are most likely to encounter in the literature: ELAN, LAN, N400, and P600.

As the name suggests, the ELAN (early left anterior negativity) is a negative-going deflection that peaks in a relatively early processing window (100–250ms post-stimulus onset) and is greatest over left anterior electrode sites. The ELAN was first reported by Neville et al. 1991 to a specific phrase structure violation in which a preposition appears in an ungrammatical position (note that the critical position must contain either a noun, adjective, or adverb, but not a preposition):

(10) a. The boys heard Joe's stories *about* Africa.
      b. *The boys heard Joe's *about* stories Africa.

A similar effect was reported by Friederici et al. (1993) in German, in this case when a participle appears in a position that must contain a noun:

(11)   *Das Baby wurde im   gefürttert
          The baby was   in-the fed
        '*The baby was in the fed'

The ELAN has since been elicited to very similar phrase structure violations in Spanish (Hinojosa et al. 2003), French (Isel et al. 2007), and further replicated in English (Lau et al. 2006, Dikker et al. 2009) and German (e.g., Hahne and Friederici 1999, Hahne and Friederici 2002, Rossi et al. 2005). The ELAN is not affected by task (Hahne and Friederici 2002), by the probability of the violation in the experiment (Hahne and Friederici 1999), or by the frequency of a disambiguated structure (Ainsworth-Darnell, Shulman, and Boland 1998, Friederici et al. 1996). These results suggest that the ELAN is a very specific response to phrase structure violations, and not simply a response to difficult or unlikely structures. The functional interpretation of the ELAN is an area of much active debate. Here are four proposals that exist in the literature: (i) Friederici 2002 (among others) interprets the ELAN as a marker of syntactic-category-based violations; (ii) Lau et al. 2006 interpret the ELAN as a marker of [failed/falsified?] syntactic prediction more generally; (iii) Dikker et al. 2009 interpret the ELAN as indexing processing in the sensory cortices that occurs prior to lexical access; and (iv) Drury et al. 2012 argue that the ELAN is an artifact of specific data analysis properties of ELAN-generating experimental paradigms.

The LAN (left anterior negativity) is a negative-going deflection that is generally largest over left-anterior electrode sites (similar to the ELAN), and tends to occur in the 300–500ms time window (later than the ELAN). The LAN has been elicited by a broad array of (morpho-) syntactic violations, such as agreement violations (Coulson et al. 1998, Gunter et al. 1997, Münte et al. 1997, Kaan 2002, Osterhout and Mobley 1995), case violations (Münte and Heinze 1994), phrase structure violations (Friederici, Hahne, and Mecklinger 1996, Hagoort, Wassenaar, and Brown 2003), island constraint violations (Kluender and Kutas 1993b), and even garden-path sentences (Kaan and Swab 2003). The LAN has also been elicited during the processing of long-distance dependencies such as wh-movement, at both the displaced wh-word and the gap location (Kluender and Kutas 1993a, Phillips, Kazanina, and Abada 2005). The functional interpretation of the LAN is even less clear than that of the ELAN. One issue is that the LAN results are often relatively fragile, and do not always replicate from study to study. Another issue is that the LAN arises for very different phenomena (e.g., morphosyntactic agreement and dependency processing), suggesting either a high-level interpretation that links these two phenomena, or two distinct sources for the LAN. A final issue is that LAN effects often co-occur with P600 effects, raising the possibility that the LAN effect is really the result of a combination of an N400 and a P600, with the two canceling out in the LAN time window, except for where their distributions fail to overlap (i.e., if the N400 is left/central, and the P600 is right/posterior, the combination may yield a left/anterior negativity; see Tanner and van Hell 2014, Tanner 2015).

The N400 is a negative-going deflection that is generally largest over centro-parietal electrode sites, and tends to occur 300–500ms post-stimulus onset (with a peak amplitude occurring at 400ms, hence the name). The N400 was first observed by Kutas and Hillyard (1980) when they presented participants with sentences that ended with unexpected words. They compared baseline sentences with semantically congruent endings (9a) to sentences with

semantically incongruent endings (9b) and sentences with endings that were incongruent due to the physical properties of the stimulus such as words written in all capital letters (9c):

(12)   a.   I spread the warm bread with butter.
       b.   I spread the warm bread with socks.
       c.   I spread the warm bread with BUTTER.

Kutas and Hillyard (1980) observed a larger N400 for (12b) compared to (12a), and a larger P300 (also known as a P3b) to (12c) compared to (12a). This qualitative difference in the responses to (12b) versus (12a) suggests that the N400 is specifically related to semantic processes rather than general error detection. In the decades since its discovery, the N400 has been elicited by a broad array of linguistic and non-linguistic stimuli, with the common pattern being that they are all meaningful in some way: spoken words, written words, signed words, pseudowords, acronyms, environmental sounds, faces, and gestures (for a review see Kutas, Van Petten, and Kluender 2006). There are two leading functional interpretations of the N400: Hagoort 2008 (among others) interprets the N400 as an index of the increased difficulty of integrating incongruent words into the preceding context, while Kutas and Federmeier 2000 (among others) interpret the N400 as an index of processes related to the activation of semantic features in the lexicon (or semantic memory).

The P600 (alternatively the "syntactic positive shift") is a positive-going deflection that is generally largest over centro-parietal electrode sites and tends to occur 500–800ms post-stimulus onset (although there is a good deal of variability in its latency in the ERP literature). Like the LAN, the P600 has been reported for a broad array of syntactic violations, in many cases co-occurring with a preceding LAN. For example, P600s have been elicited to phrase structure violations (Hagoort, Brown, and Groothusen 1993, Friederici et al. 1993, Hahne and Friederici 1999, Friederici and Frisch 2000, Osterhout and Holcomb 1992), agreement violations (Hagoort, Brown, and Groothusen 1993, Kaan 2002), syntactic garden-paths (Friederici et al. 1996, Kaan and Swaab 2003, Osterhout, Holcomb, and Swinney 1994), and island violations (McKinnon and Osterhout 1996). The sheer number of violation types that elicit a P600 has led some researchers to suggest that the P600 may be a (slightly delayed) version of the P300 (or P3b), which is a general response to unexpected stimuli (Coulson et al. 1998, see Osterhout and Hagoort 1999 for a response). P600s have also been elicited by the processing of grammatical sentences with particularly complex syntactic properties, such as ambiguous structures (Frisch, Schlesewsky, Saddy, and Alpermann 2002) and wh-movement (Fiebach, Schlesewsky, and Friederici 2002, Kaan, Harris, Gibson, and Holcomb 2000, Phillips, Kazanina, and Abada 2005). Recent research has even found P600s to sentences that appear to contain one very specific type of semantic violation (Kim and Osterhout 2005, Kuperberg, Sitnikova, Caplan, and Holcomb 2003, van Herten, Kolk, and Chwilla 2005, Kuperberg 2007, Bornkessel-Schlesewsky and Schlesewsky 2008, Stroud and Phillips 2011). As for functional interpretation, Friederici 2002 (among others) interprets the P600 as indexing syntactic revision during a stage of processing requiring the integration of syntactic and semantic information, whereas Hagoort 2008 (among others) interprets the P600 as indexing the difficulty of unifying syntactic and semantic information (a subtly different interpretation from the revision approach).

MEG data can, in principle, be used the same way as EEG data, yielding ERFs to various grammatical violations. And much of the early MEG literature has that profile. However, because MEG lends itself to better spatial resolution of the source of activity than EEG, the

current trend in the MEG literature is to focus on the localization of event-related activity to specific areas of the brain. For example, using a series of experiments that investigate activity to two-word units that undergo semantic composition such as "red boat", Bemis and Pylkkänen 2011 found activation in the left anterior temporal lobe and the ventro-medial pre-frontal cortex (with much subsequent work directed at determining to what extent this activation reflects purely semantic composition versus other types of conceptual combinatorics). Similarly, using a naturalistic story-listening task, Brennan and Pylkkänen (in press) found that activation in the left anterior temporal lobe correlates with the number of parse steps in a predictive left-corner (syntactic) parser. Studies such as these suggest a potential future in which MEG is used to localize the fundamental grammatical operations that are postulated by grammatical theories, at least those that translate into distinct operations at the level of sentence processing.

6. Hemodynamic responses: functional magnetic resonance imaging

One potentially useful property of the human circulatory system is the fact that depletion of local energy stores due to neural firing in the cortex triggers a hemodynamic response wherein oxygenated blood is sent to that area to replenish those stores. Functional magnetic resonance imaging (fMRI) leverages this property of the circulatory system to localize cognitive function. Though the details are quite complex, the basic idea is as follows. Oxygenated and deoxygenated hemoglobin have different magnetic properties. The fMRI method can detect these differences (the signal is called the BOLD signal: blood oxygen-level dependent). By comparing two conditions, one with a specific cognitive process, and another without, it is possible to use the BOLD signal to localize the specific area(s) of the brain that depleted local energy stores due to that cognitive process. The fMRI method has better spatial resolution, and requires fewer controversial assumptions, than source-localization using MEG (and EEG). However, fMRI also has worse temporal resolution (because blood travels much slower than electricity and magnetic fields). As such, fMRI can have difficulty isolating the differential effects of sequences of processes; but if the researcher can isolate the relevant process in an experimental design, fMRI is unparalleled among current non-invasive technologies for spatial localization.

The localization of cognitive processes is not typically a component of grammatical theories. But as briefly discussed in the previous section, the localization of grammatical operations (by way of sentence processing theories) is a natural extension of grammatical theories as theories of cognition, and can help to integrate linguistics with the rest of cognitive neuroscience. To that end, in this section we will briefly review research that has sought to identify brain areas that underlie syntactic processing using fMRI. There are two areas that have been the focus of most neuroimaging research on syntax: Broca's and the anterior temporal lobe.

Broca's area is probably the most famous brain region to be correlated with structural properties of sentences. The term *Broca's area* usually refers to a portion of the *left inferior frontal gyrus* (LIFG) composed of the more anterior *pars triangularis* (Brodmann area 45) and the more posterior *pars opercularis* (Brodmann area 44). Paul Broca originally identified this area as central to speech processing based on the post-mortem inspection of the brains of two patients that exhibited severe aphasia: one patient could only produce the word "tan", the other only a handful of basic words. With the advent of non-invasive neuroimaging techniques such as fMRI, Broca's area has taken center stage in the investigation of the neural substrates of syntactic processing. At least two thirds of the neuroimaging studies of the brain areas involved in sentence processing (in health) over the past 15 years reveal an increased activation in (at least

part of) Broca's area for at least one of the reported contrasts, suggesting that this area indeed plays a significant role in some aspect of sentence processing.

Although there has been great debate about the key property that modulates activity in Broca's area in sentence processing, perhaps the most theory-neutral description of the central data is that Broca's area tends to respond more to sentences with non-canonical word order than sentences with canonical word order. For example, relative to controls with canonical word order, Broca's area shows increased BOLD signal for relative clauses (e.g., Just et al. 1996, Ben-Shachar et al. 2003), wh-movement (e.g., Ben-Shachar 2004, Santi and Grodzinsky 2007), topicalization (e.g., Ben-Shachar 2004), clefting (e.g., Caplan et al. 1999), and scrambling (e.g., Friederici et al. 2006, Bornkessel et al. 2005, Bornkessel-Schlesewsky et al. 2009). The question then is which cognitive processes these syntactic phenomena have in common. There is quite a bit of debate in the literature about this. For example, Grodzinsky and colleagues have argued that Broca's area seems to be more active for non-canonical word orders because Broca's area supports the syntactic mechanism of *movement* that is familiar from generative syntactic theory (Grodzinsky 1986, see Grodzinsky & Santi 2008 for a recent review). Bornkessel-Schlesewsky and colleagues have proposed that the same effects can be explained by assuming a parsing stage in which the argument relations of the sentence are computed according to several prominence hierarchies that are familiar from typological research (e.g., the animacy hierarchy, the case hierarchy, the definiteness hierarchy; Comrie 1989, Bornkessel and Schlesewsky 2006, Wolff et al. 2008). This parsing stage would require a "linearization" process that maps word order to argument structure according to these prominence hierarchies. For Bornkessel-Schlesewsky and colleagues, Broca's area supports this linearization process (Chen et al 2006, Grewe et al. 2006, Bornkessel-Schlesewsky et al. 2009). In principle, if one assumes a strong relationship between cognitive theories and neurobiology (such as the strong reductionism examined critically in Fodor 1975), it is possible that the resolution of this debate about the functional interpretation of these effects in Broca's area could be used as evidence to adjudicate between these competing theories of syntax.

Although Broca's area has been the focus of many neuroimaging studies of syntax, there is a growing literature implicating portions of the temporal lobe in syntactic processing. One of the most robust neuroimaging findings about sentence-level processing is that lateral anterior portions of the superior and middle temporal cortex show greater activation bilaterally for reading or listening to sentences than word lists (Mazoyer 1993, Stowe et al. 1998, Friederici et al. 2000, Vandenberghe et al. 2002, Humphries et al 2005, 2006, and Brennan and Pylkkänen (in press) mentioned in the previous section). Furthermore, lesion mapping has associated damage to the left lateral anterior temporal lobe with comprehension impairment for most sentences more complex than simple declaratives (Dronkers et al. 2004, although cf. Kho et al. 2008). These findings suggest that anterior portions of the temporal lobe support sentence-level computations that do not rely on lexical semantics, but this leaves open a number of possible candidate processes: syntactic processes, argument structure processes, discourse processes, and even prosodic processes. If there were a brain region dedicated to basic syntactic phrase structure computation in comprehension, one would expect it to show a profile similar to that of the anterior temporal lobe, showing more activity for processing word strings with syntactic structure than those without. However, demonstrating that this area is specifically involved in syntax as opposed to other phrase-level computations has proved challenging (but see Brennan et al. 2010, Brennan and Pylkkänen (in press), and Rogalsky and Hickok 2008 for interesting attempts to distinguish syntax and semantics).

16

As our confidence in the functional localization of the brain increases, it may become possible to use that information to test and refine grammatical theories. If a grammatical theory (as suitably integrated into a theory of sentence processing) predicts that a specific process should be deployed in a sentence, a definitive theory of functional localization in the brain would allow us to use methods such as MEG and fMRI to test whether that process is indeed deployed. Though we are still far from a definitive theory of functional localization in the brain, studies such as the ones reviewed in this section begin to demonstrate the value of integrating neuroscience and linguistics for both fields.

7. Conclusion

In principle, any data type that is related to language behavior is a potential source of information for grammatical theories. In practice, acceptability judgments form the majority of data used to construct grammatical theories, primarily due to their ability to reveal causal relationships, and the assumption that grammatical properties are one of several factors that directly impact acceptability judgments. Because of the prevalence of acceptability judgments in the literature, the past two decades have seen a number of advancements in the use of formal experimental methods (including factorial designs) for the collection of acceptability judgments. However, for many linguists, the future of grammatical theory lies in integrating grammatical theories with theories of sentence processing, both at a behavioral level and at a neurophysiological level. One potential step toward this integration is an exploration of the data types that are used to construct theories of sentence processing (e.g., reading times, M/EEG data, and fMRI data). Though the relationship between these data types and grammatical theories is less direct than that of acceptability judgments, we hope that the discussion in this chapter makes it clear that such an integration is a worthy goal for 21st century linguistics.

References

Adger, David (2003). *Core Syntax: A Minimalist Approach*. Oxford: Oxford University Press.

Ainsworth-Darnell, Kim, Harvey G. Shulman, and Julie E. Boland (1998). 'Dissociating brain responses to syntactic and semantic anomalies: Evidence from event-related potentials', *Journal of Memory and Language* 38: 112–130.

Bard, Ellen Gurman, Dan Robertson, and Antonella Sorace (1996). 'Magnitude estimation of linguistic acceptability', *Language* 72: 32–68.

Bastiaansen, Marcel, Ali Mazaheri, and Ole Jensen (2011). 'Beyond ERPs: Oscillatory Neuronal Dynamics', in Emily S. Kappenman and Steven J. Luck (eds.), *The Oxford Handbook of Event-Related Potential Components*. Oxford University Press.

Bemis, Douglas K., and Liina Pylkkänen (2011). 'Simple Composition: A magnetoencephalography investigation into the comprehension of minimal linguistic phrases', *Journal of Neuroscience* 31(8): 2801–2814.

Ben-Shachar, Michal, Talma Hendler, Itamar Kahn, Dafna Ben-Bashat, and Yosef Grodzinsky (2003). 'The neural reality of syntactic transformations: Evidence from functional magnetic resonance imaging', *Psychological Science* 14: 433–440.

Ben-Shachar, Michal, Dafna Palti, and Yosef Grodzinsky (2004). 'Neural correlates of syntactic movement: converging evidence from two fMRI experiments', *Neuroimage* 21: 1320–1336.

Bornkessel, Ina, and Matthias Schlesewsky (2006). 'The extended argument dependency model: A neurocognitive approach to sentence comprehension across languages', *Psychological Review* 113: 787–821.

Bornkessel-Schlesewsky, Ina, and Matthias Schlesewsky (2008). 'An alternative perspective on "semantic P600" effects in language comprehension', *Brain Research Reviews* 59: 55–73.

Bornkessel-Schlesewsky, Ina, Matthias Schlesewsky, and D. Yves von Cramon (2009). 'Word order and Broca's region: Evidence for a supra-syntactic perspective', *Brain and Language* 111: 125–139.

Bornkessel, Ina, Stefan Zysset, Angela D. Friederici, D. Yves Von Cramon, and Matthias Schlesewsky (2005). 'Who did what to whom? The neural basis of argument hierarchies during language comprehension', *Neuroimage* 26: 221–233.

Brennan, Jonathan, Yuval Nir, Uri Hasson, Rafael Malach, David J. Heeger, and Liina Pylkkänen (2010). 'Syntactic structure building in the anterior temporal lobe during natural story listening', *Brain and Language* 120: 163-173.

Brennan, Jonathan, and Liina Pylkkänen (in press). 'MEG evidence for incremental sentence composition in the anterior temporal lobe', *Cognitive Science*.

Caplan, David, Nathaniel Alpert, and Gloria Waters (1999). 'PET studies of syntactic processing with auditory sentence presentation', *Neuroimage* 9: 343–351.

Chen, Evan, W. Caroline West, Gloria Waters, and David Caplan (2006). 'Determinants of BOLD signal correlates of processing object-extracted relative clauses', *Cortex* 42: 591–604.

Chomsky, Noam (1957). *Syntactic structures*. The Hague: Mouton.

Chomsky, Noam (1964). *Current issues in linguistic theory*. The Hague: Mouton.

Chomsky, Noam (1973). 'Conditions on Transformations', in Stephen R. Anderson and Paul Kiparsky (eds.), *A festschrift for Morris Halle*, New York: Holt, Rinehart and Winston, 232-286.

Chomsky, Noam (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

Chomsky, Noam (1986). *Barriers*. Cambridge, MA: MIT Press.

Cohen, Mike X (2014). *Analyzing Neural Time Series Data*. Cambridge, MA: MIT Press.

Coulson, Seana, Jonathan W. King, and Marta Kutas (1998). 'Expect the unexpected: Event-related brain response to morphosyntactic violations', *Language and Cognitive Processes* 13: 21–58.

Cowart, Wayne (1997). *Experimental syntax: Applying objective methods to sentence judgments*. Thousand Oaks, CA: SAGE.

Crain, Stephen, and Janet Dean Fodor (1985). 'How can grammars help parsers?', in David R. Dowty, Lauri Karttunen, and Arnold M. Zwicky (eds.), *Natural Language Parsing: Psycholinguistic, Computational, and Theoretical Perspectives*, Cambridge: Cambridge University Press, 94–128.

Dikker, Suzanne, Hugh Rabagliati, and Liina Pylkkänen (2009). 'Sensitivity to syntax in visual cortex', *Cognition* 110: 293–321.

Dronkers, Nina F., David P. Wilkins, Robert D. Van Valin, Brenda B. Redfern, and Jeri J. Jaeger (2004). 'Lesion analysis of the brain areas involved in language comprehension', *Cognition* 92: 145–177.

Edelman, Shimon, and Morton H. Christiansen (2003). 'How seriously should we take Minimalist syntax?', *Trends in Cognitive Sciences* 7: 60–61.

Featherston, Sam (2005). 'The Decathlon Model of empirical syntax', in Marga Reis and Stephan Kepser (eds.), *Linguistic Evidence: Empirical, Theoretical, and Computational Perspectives*, Berlin: Mouton de Gruyter, 187–208.

Featherston, Sam (2007). 'Data in generative grammar: The stick and the carrot', *Theoretical Linguistics* 33(3): 269–318.

Ferreira, Fernanda (2005). 'Psycholinguistics, formal grammars, and cognitive science', *The Linguistic Review* 22: 365–380.

Fiebach, Christian J., Matthias Schlesewsky, and Angela D. Friederici (2002). 'Separating syntactic memory costs and syntactic integration costs during parsing: the processing of German wh-questions', *Journal of Memory and Language* 47: 250–272.

Fodor, Jerry. *The Language of Thought.* Cambridge, MA: Harvard University Press.

Frazier, Lyn, and Giovanni B. Flores d'Arcais (1989). 'Filler driven parsing: A study of gap filling in Dutch', *Journal of Memory and Language* 28(3): 331–344.

Friederici, Angela D. (2002). 'Towards a neural basis of auditory sentence processing', *Trends in Cognitive Sciences* 6: 78–84.

Friederici, Angela D., Jörg Bahlmann, Stefan Heim, Ricarda I. Schubotz, and Alfred Anwander (2006). 'The brain differentiates human and non-human grammars: Functional localization and structural connectivity', *Proceedings of the National Academy of Sciences of the United States of America* 103: 2458.

Friederici, Angela D. and Stefan Frisch (2000). 'Verb argument structure processing: The role of verb- specific and argument-specific information', *Journal of Memory and Language* 43: 476–507.

Friederici, Angela D., Martin Meyer, and D. Yves von Cramon (2000). 'Auditory language comprehension: An event-related fMRI study on the processing of syntactic and lexical information', *Brain and Language* 74: 289–300.

Friederici, Angela D., Erdmut Pfeifer, and Anja Hahne (1993). 'Event-related brain potentials during natural speech processing: Effects of semantic, morphological and syntactic violations', *Cognitive Brain Research* 1: 183–192.

Friederici, Angela D., Anja Hahne, and Axel Mecklinger (1996). 'Temporal structure of syntactic parsing: Early and late event-related brain potential effects', *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22: 1219–1248.

Frisch, Stefan, Matthias Schlesewsky, Douglas Saddy, and Annegret Alpermann (2002). 'The P600 as an indicator of syntactic ambiguity', *Cognition* 85: B83–B92.

Garnsey, Susan M., Michael Tanenhaus, and Robert M. Chapman (1989). 'Evoked potential and the study of sentence comprehension', *Journal of Psycholinguistic Research* 18: 51–60.

Gibson, Edward, and Evelina Fedorenko (2010). 'Weak quantitative standards in linguistics research', *Trends in Cognitive Sciences* 14: 233–234.

Gibson, Edward, and Evelina Fedorenko (2013). 'The need for quantitative methods in syntax and semantics research', *Language and Cognitive Processes* 28: 88–124.

Grodzinsky, Yosef (1986). 'Language deficits and the theory of syntax', *Brain and Language* 27: 135–159.

Grodzinsky, Yosef, and Andrea Santi (2008). 'The battle for Broca's region', *Trends in Cognitive Sciences* 12: 474–480.

Grewe, Tanja, Ina Bornkessel, Stefan Zysset, Richard Wiese, D. Yves von Cramon, and Matthias Schlesewsky (2005). 'The emergence of the unmarked: A new perspective on the language‐specific function of Broca's area', *Human Brain Mapping* 26: 178–190.

Gunter, Thomas C., Laurie A. Stowe, and Gusbertus Mulder (1997). 'When syntax meets semantics', *Psychophysiology* 34: 660–676.

Hagoort, Peter (2008). 'The fractionation of spoken language understanding by measuring electrical and magnetic brain signals', *Philosophical Transactions of the Royal Society B: Biological Sciences* 363: 1055.

Hagoort, Peter, Colin Brown, and Jolanda Groothusen (1993). 'The syntactic positive shift (SPS) as an ERP measure of syntactic processing', *Language and Cognitive Processes* 8: 439–483.

Hagoort, Peter, Marlies Wassenaar, and Colin M. Brown (2003). 'Syntax-related ERP-effects in Dutch', *Cognitive Brain Research* 16: 38–50.

Hahne, Anja, and Angela D. Friederici (1999). 'Electrophysiological evidence for two steps in syntactic analysis: Early automatic and late controlled processes', *Journal of Cognitive Neuroscience* 11: 194–205.

Hahne, Anja, and Angela D. Friederici (2002). 'Differential task effects on semantic and syntactic processes as revealed by ERPs', *Cognitive Brain Research* 13: 339–356.

Hill, Archibald A. (1961). 'Grammaticality', *Word* 17: 1–10.

Hinojosa, José, Manuel Martín-Loeches, Pilar Casado, Francisco Muñoz, and Francisco Rubia (2003). 'Similarities and differences between phrase structure and morphosyntactic violations in Spanish: An event-related potentials study', *Language and Cognitive Processes* 18: 113–142.

Huang, Cheng-Teh James (1982). *Logical Relations in Chinese and the Theory of Grammar.* PhD dissertation, MIT, Cambridge, MA.

Humphries, Colin, Jeffrey R. Binder, David A. Medler, and Einat Liebenthal (2006). 'Syntactic and semantic modulation of neural activity during auditory sentence comprehension', *Journal of Cognitive Neuroscience* 18: 665–679.

Humphries, Colin, Tracy Love, David Swinney, and Gregory Hickok (2005). 'Response of anterior temporal cortex to syntactic and prosodic manipulations during sentence processing', *Human Brain Mapping* 26: 128–138.

Isel, Frédéric, Anja Hahne, Burkhard Maess, and Angela D. Friederici (2007). 'Neurodynamics of sentence interpretation: ERP evidence from French', *Biological Psychology* 74: 337–346.

Just, Marcel A., Patricia A. Carpenter, and Jacqueline D. Woolley (1982). 'Paradigms and processes in reading comprehension', *Journal of Experimental Psychology: General* 111: 228–238.

Just, Marcel A., Patricia A. Carpenter, Timothy A. Keller, William F. Eddy, and Keith R. Thulborn (1996). 'Brain activation modulated by sentence comprehension', *Science* 274: 114.

Kaan, Edith (2002). 'Investigating the effects of distance and number interference in processing subject-verb dependencies: An ERP study', *Journal of Psycholinguistic Research* 31: 165–193.

Kaan, Edith, Anthony Harris, Edward Gibson, and Phillip Holcomb (2000). 'The P600 as an index of syntactic integration difficulty', *Language and Cognitive Processes* 15: 159–201.

Kaan, Edith, and Tamara Y. Swaab (2003). 'Electrophysiological evidence for serial sentence processing: a comparison between non-preferred and ungrammatical continuations', *Cognitive Brain Research* 17: 621–635.

Kazanina, Nina, Ellen Lau, Moti Lieberman, Masaya Yoshida, and Colin Phillips (2007). 'The Effect of Syntactic Constraints on the Processing of Backwards Anaphora', *Journal of Memory and Language* 56: 384–409.

Keller, Frank (2000). *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality*. Ph.D. dissertation, University of Edinburgh.

Kho, Kuan H., Peter Indefrey, Peter Hagoort, C. W. M. van Veelen, Peter C. van Rijen, and Nick F. Ramsey (2008). 'Unimpaired sentence comprehension after anterior temporal cortex resection', *Neuropsychologia* 46: 1170–1178.

Kim, Albert, and Lee Osterhout (2005). 'The independence of combinatory semantic processing: Evidence from event-related potentials', *Journal of Memory and Language* 52: 205–225.

Kluender, Robert, and Marta Kutas (1993a). 'Bridging the gap: Evidence from ERPs on the processing of unbounded dependencies', *Journal of Cognitive Neuroscience* 5: 196–214.

Kluender, Robert, and Marta Kutas (1993b). 'Subjacency as a processing phenomenon', *Language and Cognitive Processes* 8: 573–633.

Kuperberg, Gina R. (2007). 'Neural mechanisms of language comprehension: Challenges to syntax', *Brain Research* 1146: 23–49.

Kuperberg, Gina R., Tatiana Sitnikova, David Caplan, and Phillip J. Holcomb (2003). 'Electrophysiological distinctions in processing conceptual relationships within simple sentences', *Cognitive Brain Research* 17: 117–129.

Kutas, Marta, and Kara D. Federmeier (2000). 'Electrophysiology reveals semantic memory use in language comprehension', *Trends in Cognitive Sciences* 4: 463–470.

Kutas, Marta, and Steven A. Hillyard (1980). 'Event-related brain potentials to semantically inappropriate and surprisingly large words', *Biological Psychology* 11: 99–116.

Kutas, Marta, Cyma Van Petten, and Robert Kluender (2006). 'Psycholinguistics electrified II: 1994– 2005', in Matthew J. Traxler and Morton Ann Gernsbacher (eds.), *Handbook of psycholinguistics,* 2nd ed., Elsevier, 659–724.

Lasnik, Howard, and Mamoru Saito (1984). 'On the nature of proper government', *Linguistic Inquiry* 15: 235–289.

Lau, Ellen F., Clare Stroud, Silke Plesch, and Colin Phillips (2006). 'The role of structural prediction in rapid syntactic analysis', *Brain and Language* 98: 74–88.

Luck, Steven. (2005). *An introduction to the event-related potential technique*. MIT Press.

Marantz, Alec (2005). 'Generative linguistics within the cognitive neuroscience of language', *The Linguistic Review* 22: 429–445.

Mazoyer, B. M., N. Tzourio, V. Frak, A. Syrota, N. Murayama, O. Levrier, G. Salamon, S. Dehaene, Laurent Cohen, and J. Mehler (1993). 'The cortical representation of speech', *Journal of Cognitive Neuroscience* 5: 467–479.

McKinnon, Richard, and Lee Osterhout (1996). 'Constraints on movement phenomena in sentence processing: Evidence from event-related brain potentials', *Language and Cognitive Processes* 11: 495–524.

Miller, George, and Noam Chomsky (1963). 'Finitary models of language users', in R. Luce, R. Bush, and E. Galanter (eds.), *Handbook of Mathematical Psychology*, Vol. 2, New York: Wiley, 419–493.

Montalbetti, Mario (1984). *After binding*. PhD Dissertation, MIT, Cambridge, MA.

Münte, Thomas F., and Hans J. Heinze (1994). 'ERP negativities during syntactic processing of written words. In Hans J. Heinze, Thomas F. Münte, & G. R. Mangun (eds.), *Cognitive Electrophysiology,* Boston: Birkhauser, 211–238.

Münte, Thomas F., Mike Matzke, and Sönke Johannes (1997). 'Brain activity associated with syntactic incongruencies in words and pseudo-words', *Journal of Cognitive Neuroscience* 9: 318–329.

Neville, Helen, Janet L. Nicol, Andrew Barss, Kenneth I. Forster, and Merrill F. Garrett (1991). 'Syntactically based sentence processing classes: Evidence from event-related brain potentials', *Journal of Cognitive Neuroscience* 3: 151–165.

Osterhout, Lee, and Peter Hagoort (1999). 'A superficial resemblance does not necessarily mean you are part of the family: Counterarguments to Coulson, King and Kutas (1998) in the P600/SPS-P300 debate', *Language and Cognitive Processes* 14: 1–14.

Osterhout, Lee, and Phillip J. Holcomb (1992). 'Event-related brain potentials elicited by syntactic anomaly', *Journal of Memory and Language* 31: 785–806.

Osterhout, Lee, Phillip J. Holcomb, and David Swinney (1994). 'Brain potentials elicited by garden-path sentences: Evidence of the application of verb information during parsing', *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20: 786–803.

Osterhout, Lee, and Linda A. Mobley (1995). 'Event-related brain potentials elicited by failure to agree', *Journal of Memory and Language* 34: 739–773.

Phillips, Colin (2006). 'The real-time status of island phenomena', *Language* 82: 795–823.

Phillips, Colin, Nina Kazanina, and Shani H. Abada (2005). 'ERP effects of the processing of syntactic long-distance dependencies', *Cognitive Brain Research* 22: 407–428.

Rogalsky, Corianne, and Gregory Hickok (2008). 'Selective attention to semantic and syntactic features modulates sentence processing networks in anterior temporal cortex', *Cerebral Cortex* 19: 786–796.

Ross, John Robert (1967). *Constraints on variables in syntax*. PhD dissertation, MIT, Cambridge, MA.

Rossi, Sonja, Manfred F. Gugler, Anja Hahne, and Angela D. Friederici (2005). 'When word category information encounters morphosyntax: An ERP study', *Neuroscience Letters* 384: 228–233.

Santi, Andrea, and Yosef Grodzinsky (2007). 'Working memory and syntax interact in Broca's area', *Neuroimage* 37: 8–17.

Schütze, Carson T. (1996). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: University of Chicago Press. Reprinted by Language Science Press (Classics in Linguistics #2), Berlin, 2016.

Schütze, Carson T., and Jon Sprouse (2013). Judgment data. In Robert J. Podesva and Devyani Sharma (eds.), *Research Methods in Linguistics*, New York: Cambridge University Press, 27–50.

Spencer, Nancy J. (1973). 'Differences between linguists and nonlinguists in intuitions of grammaticality-acceptability', *Journal of Psycholinguistic Research* 2: 83–98.

Sprouse, Jon, and Diogo Almeida (2012). 'Assessing the reliability of textbook data in syntax: Adger's *Core Syntax', Journal of Linguistics* 48: 609–652.

Sprouse, Jon, Ivano Caponigro, Ciro Greco, and Carlo Cecchetto (2016). 'Experimental syntax and the variation of island effects in English and Italian', *Natural Language and Linguistic Theory* 34: 307–344.

Sprouse, Jon, Carson T. Schütze, and Diogo Almeida (2013). 'A comparison of informal and formal acceptability judgments using a random sample from *Linguistic Inquiry* 2001–2010', *Lingua* 134: 219–248.

Sprouse, Jon, Matt Wagers, and Colin Phillips (2012). 'A test of the relation between working memory and syntactic island effects', *Language* 88: 82–123.

Steinhauer, Karsten, and John Drury (2012). 'On the early left-anterior negativity (ELAN) in syntax studies', *Brain and Language* 120: 135–162.

Stowe, Laurie A. (1986). 'Parsing WH-constructions: Evidence for on-line gap location', *Language and Cognitive Processes* 1: 227–245.

Stowe, Laurie A., Cees A. J. Broere, Anne M. J. Paans, Albertus A. Wijers, Gijsbertus Mulder, Wim Vaalburg, and Frans Zwarts (1998). 'Localizing components of a complex task: sentence processing and working memory', *NeuroReport* 9: 2995–2999.

Stroud, Clare, and Colin Phillips (2011). 'Examining the evidence for an independent semantic analyzer: An ERP study in Spanish', *Brain and Language* 120: 108–126.

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

Tanner, Darren (2015). 'On the left anterior negativity (LAN) in electrophysiological studies of morphosyntactic agreement: A Commentary on "Grammatical agreement processing in reading: ERP findings and future directions" by Molinaro et al., 2014', *Cortex* 66: 149–155.

Tanner, D., and Janet G. Van Hell (2014). 'ERPs reveal individual differences in morphosyntactic processing', *Neuropsychologia* 56: 289–301.

Van Gompel, Roger P. G., and Simon P. Liversedge (2003). 'The influence of morphological information on cataphoric pronoun assignment', *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29: 128–139.

van Herten, Marieke, Herman H. J. Kolk, and Dorothee J. Chwilla (2005). 'An ERP study of P600 effects elicited by semantic anomalies', *Cognitive Brain Research* 22: 241–255.

Vandenberghe, Rik, Anna Christina Nobre, and Cathy J. Price (2002). 'The response of left temporal cortex to sentences', *Journal of Cognitive Neuroscience* 14: 550–560.

Villata, Sandra, Luigi Rizzi, and Julie Franck (2016). 'Intervention effects and Relativized Minimality: New experimental evidence from graded judgments', *Lingua* 179: 76-96.

Wagers, Matthew, and Colin Phillips (2009). 'Multiple dependencies and the role of the grammar in real-time comprehension', *Journal of Linguistics* 45: 395–433.

Wellwood, Alexis, Roumyana Pancheva, Valentine Hacquard, and Colin Phillips (2015). 'The anatomy of a comparative illusion', Ms., University of Maryland, College Park.

Wolff, Susann, Matthias Schlesewsky, Masako Hirotani, and Ina Bornkessel-Schlesewsky (2008). 'The neural mechanisms of word order processing revisited: Electrophysiological evidence from Japanese', *Brain and Language* 107: 133-157.

Yoshida, Masaya, Michael Walsh Dickey, and Patrick Sturt (2013). 'Predictive Processing of Syntactic Structure: Sluicing and Ellipsis in Real-Time Sentence Processing.' *Language and Cognitive Processes* 28: 277-302.