Brain & Language 157–158 (2016) 81–94

Contents lists available at ScienceDirect

Brain & Language

journal homepage: www.elsevier.com/locate/b&l

Abstract linguistic structure correlates with temporal activity during naturalistic comprehension

Jonathan R. Brennan [a,*], Edward P. Stabler [b], Sarah E. Van Wagenen [b], Wen-Ming Luh [c], John T. Hale [d]

[a] Department of Linguistics, University of Michigan, Ann Arbor, MI, United States
[b] Department of Linguistics, University of California, Los Angeles, CA, United States
[c] MRI Facility, Cornell University, Ithaca, NY, United States
[d] Department of Linguistics, Cornell University, Ithaca, NY, United States

ARTICLE INFO

Article history:
Received 14 July 2015
Revised 14 March 2016
Accepted 10 April 2016
Available online 19 May 2016

Keywords:
Syntax
Parsing
Prediction
fMRI
ATL
IFG
PTL
Narrative

ABSTRACT

Neurolinguistic accounts of sentence comprehension identify a network of relevant brain regions, but do not detail the information flowing through them. We investigate syntactic information. Does brain activity implicate a computation over hierarchical grammars or does it simply reflect linear order, as in a Markov chain? To address this question, we quantify the cognitive states implied by alternative parsing models. We compare processing-complexity predictions from these states against fMRI timecourses from regions that have been implicated in sentence comprehension. We find that hierarchical grammars independently predict timecourses from left anterior and posterior temporal lobe. Markov models are predictive in these regions and across a broader network that includes the inferior frontal gyrus. These results suggest that while linear effects are wide-spread across the language network, certain areas in the left temporal lobe deal with abstract, hierarchical syntactic representations.

1. Introduction

The neural bases of syntactic processing remain elusive, despite intensive study. Current models catalog the network of regions and connections involved in various sentence-related computations, including syntax, but do not specify the kind of information that flows through this network (see e.g. Friederici & Gierhan, 2013; Hagoort & Indefrey, 2014; Hickok & Poeppel, 2007; Turken & Dronkers, 2011). As Poeppel (2012) notes, it is the information encoded during incremental stages of language comprehension that is critical for mapping between the vocabulary of neurobiology and the vocabulary of linguistics. This study examines what kind of syntactic information is manipulated by brain regions involved in sentence comprehension by correlating the complexity of different syntactic structures with brain activity recorded using fMRI while participants listen to a naturalistic narrative.

The proper conception of syntactic structure is debated across the language sciences. The available models range across many different levels of detail. There are models based on word-to-word dependencies, models based on abstract, hierarchical grammars, and many alternatives in between. While mathematical linguists are in agreement regarding the level of expressive power needed for adequate natural language grammars (Joshi, Shanker, & Weir, 1990; Shieber, 1985; Stabler, 2013a) there remains a debate over the need for more abstract representations in every-day language performance (Frank & Bod, 2011; Sanford & Sturt, 2002). To address this debate, we quantify, word-by-word, the cognitive states that are implied by parsing models that assign comparatively more or less detailed syntactic analyses. We evaluate alternative theories of syntactic structure and parsing by fitting these models to brain activity from regions that have been traditionally associated with sentence comprehension. By relying on brain data collected while participants simply listen to a story, we aim to better understand the role of syntax in every-day language comprehension.

1.1. Brain regions involved in syntactic processing

The spatio-temporal characteristics of brain activity that is sensitive to sentence structure have been examined using a wide variety of experimental techniques (see Hagoort & Indefrey, 2014 for a recent review). One common approach has been to vary whether

* Corresponding author.
E-mail addresses: jobrenn@umich.edu (J.R. Brennan), stabler@ucla.edu (E.P. Stabler), wl358@cornell.edu (W.-M. Luh), jthale@cornell.edu (J.T. Hale).

http://dx.doi.org/10.1016/j.bandl.2016.04.008
0093-934X/

syntactic structure is present or not by comparing phrases or sentences with lists of words. Sentence structure reliably leads to greater activation in the anterior portion of the temporal lobes (ATL) across multiple techniques and stimulus modalities (Brennan & Pylkkänen, 2012; Friederici, Opitz, & von Cramon, 2000; Humphries, Binder, Medler, & Liebenthal, 2006; Jobard, Vigneau, Mazoyer, & Tzourio-Mazoyer, 2007; Rogalsky & Hickok, 2009; Snijders et al., 2009; Stowe et al., 1998; Vandenberghe, Nobre, & Price, 2002; Xu, Kemeny, Park, Frattali, & Braun, 2005). Many studies also show sensitivity in a broader network as well, which includes the left inferior frontal gyrus (IFG; "Broca's Area") and the posterior temporal lobe (PTL; "Wernicke's Area") in the vicinity of the temporal-parietal junction (Brennan & Pylkkänen, 2012; Friederici et al., 2000; Jobard et al., 2007; Pallier, Devauchelle, & Dehaene, 2011; Snijders et al., 2009; Vandenberghe et al., 2002; Xu et al., 2005).

These studies reveal a network of regions that are sensitive to sentence structure, with a focus on the ATL, the IFG and the PTL. Evidence suggests that these regions subserve different functions that relate to identifying or perhaps interpreting phrases, though debate is far from settled. In several of these studies, the ATL, but not the IFG or PTL, is activated even for simple sentences (Rogalsky & Hickok, 2009; Stowe et al., 1998), though others show broader activations (e.g. Pallier et al., 2011; Snijders et al., 2009). Further work using Magnetoencephalography (MEG) has shown that simple two-word phrases lead to increased ATL activation within 200–400 ms of word onset in both visual and auditory presentation (Bemis & Pylkkänen, 2011; Bemis & Pylkkänen, 2013). This effect generalizes across languages and phrase types (Westerlund, Kastner, Al Kaabi, & Pylkkänen, 2015). Shetreet, Friedmann, and Hadar (2009) report a similar sensitivity to constituent structure type in the anterior temporal lobe: more complex hierarchical structure (phrasal vs. nominal verb complements) increased activation in this region. Brennan et al. (2012) build on these observations by testing for sensitivity to incremental, word-by-word, phrase-structure complexity. In this study, the ATL is the only brain area whose activity correlates positively with phrase-structure complexity. Some models suggest that the ATL may subserve constituent structure processes, a conclusion consistent with the morphosyntactic deficits due to anterior lesions observed by Dronkers, Wilkins, Van Valin, Redfern, and Jaeger (2004) (e.g. Friederici & Gierhan, 2013). However, more recent evidence from magnetoencephalography (Westerlund & Pylkkänen, 2014; Zhang & Pylkkänen, 2015) and patient studies of Primary Progressive Aphasia (Wilson et al., 2014) point towards a more nuanced function that relates to the semantic interpretation of composed structures.

Turning to the functional role of the PTL, it has been reported to be modulated by the presence or absence of basic phrase structure in some studies (e.g. Bemis & Pylkkänen, 2013; Pallier et al., 2011), but does not uniformly show such effects across the literature. There is also evidence from neurodegenerative disorders that posterior temporal and inferior parietal atrophy is associated with syntactic deficits (Wilson et al., 2011). Some theorists have hypothesized that this region may play a role in discourse-level comprehension (e.g. Ferstl, Neumann, Bogler, & Yves von Cramon, 2008), though note also that nodes within this broad area, specifically along the posterior middle temporal gyrus, have long been implicated in lexical processing that is sensitive to sentence and discourse context (see Hickok & Poeppel, 2007, for discussion). Bornkessel-Schlesewsky, Schlesewsky, Small, and Rauschecker (2015) argue that posterior and dorsal regions, which include the PTL and extend through the inferior parietal lobule (IPL) to premotor cortex, are involved in sentence processing that is sensitive to linear order. These order-sensitive regions contrast with ventral anterior regions like the ATL, discussed above, which are associated with hierarchical processes. It remains unknown whether sentence-related activation in PTL is best attributed to a single function, such as order-related, lexical, or discourse computations, or to some combination of these or other functions.

Evidence for a functional division specifically between temporal lobe processing and the IFG comes from studies that compare processing of sentence types which differ in their constituent structure or dependency properties. Studies that compare sentences which differ in memory-load demands, such as subject and object relative clauses, yield differential activation in IFG, with variation in the precise localization (Ben-Shachar, Hendler, Kahn, Ben-Bashat, & Grodzinsky, 2003; Ben-Shachar, Palti, & Grodzinsky, 2004; Caplan, Chen, & Waters, 2008; Just, Carpenter, Keller, Eddy, & Thulborn, 1996; Santi & Grodzinsky, 2007a; Santi & Grodzinsky, 2007b; Santi & Grodzinsky, 2010; Stromswold, Caplan, Alpert, & Rauch, 1996;). This result is consistent with deficit-lesion studies suggesting that frontal lobe damage most strongly impacts the processing of syntactically complex sentences (Caramazza & Zurif, 1976; Grodzinsky, 2000; Zurif, 1995). One possibility is that the IFG is implicated in the processing of more complex syntactic operations, such as the formation of long-distance dependencies, however, the literature has yet to settle on a functional explanation that captures the broader range of observations (see Rogalsky & Hickok, 2010, for a critical review). While some models take the IFG to be implicated only in more complex syntactic operations (e.g. Grodzinsky & Friederici, 2006), in others it is positioned as a central hub for basic combinatoric processing (Hagoort, 2013). This latter view contrasts with that described above in which basic combinatorics is attributed to the ATL (e.g. Friederici & Gierhan, 2013). One avenue of current research is whether these disparate results may be reconciled in terms of fine-grained functional divisions within sub-parts of the IFG. For example, Zaccarella and Friederici (2015) report sensitivity in a sub-part of the Pars Opercularis of the IFG to very simple phrases. Similarly, different argument structure configurations have been associated with differences in IFG activation that form a spatial cline (Bornkessel-Schlesewsky & Schlesewsky, 2009).

Despite the lack of consensus about the functional division of anterior-frontal and posterior-dorsal structures in sentence comprehension, a common thread across this broad literature is that the mental representations whose processing is implicated in various regions are described at a relatively coarse-grain, for example, at the level of separating syntactic and compositional semantic representations (Westerlund & Pylkkänen, 2014) or hierarchical from non-hierarchical processing (Bornkessel-Schlesewsky et al., 2015). The level of detail of these representations remains largely underspecified.

1.2. Sensitivity to syntactic structure during incremental processing

While neural studies have become increasingly tuned to fine-grained linguistic differences between sentence and phrase types (e.g. Bornkessel, Zysset, Friederici, von Cramon, & Schlesewsky, 2005; Shetreet et al., 2009; Westerlund et al., 2015), the relationship between detailed linguistic grammars and language comprehension remains controversial. On one view, the abstract hierarchical grammars that have been developed to explain offline judgments and typological patterns should also serve to explain online comprehension (Berwick & Weinberg, 1983; Bresnan & Kaplan, 1982; Lewis & Phillips, 2015; Miller & Chomsky, 1963; Steedman, 2000). This is the *competence hypothesis*

> an explanatory model of human language performance will incorporate a theoretically-justified representation of the native speaker's linguistic knowledge

designated as such by Kaplan and Bresnan (1982, page 173), who offer the formulation quoted above as a restatement of Chomsky's original suggestion (1965, p. 9). Alternatively, interpretive shortcuts relying on surface patterns and extragrammatical heuristics might be the best characterization of on-line processing. This grammar-free alternative seems more plausible in circumstances that encourage rapid but not especially deep processing (Ferreira & Patson, 2007; Ferreira, Bailey, & Ferraro, 2002; Sanford & Sturt, 2002). On the other hand, by postulating two cognitive faculties to explain two distinct types of data, this latter view is more complex. Defenders have traditionally appealed to patterns of fallibility, such as garden path sentences, to motivate the additional heuristic system (Bever, 1970). We review evidence for both of these positions.

Evidence for the competence hypothesis comes from behavioral and event-related potential (ERP) studies involving syntactically unexpected stimuli. For example, Xiang, Dillon, and Phillips (2009) probe the processing of words whose use is licensed only in particular hierarchical configurations. Words like "any" or "ever", so-called negative polarity items, can only be used in contexts where they are embedded under phrases with restricted entailment properties, such as those that contain a negation (see (1-a)–(1-c)) (e.g. Giannakidou, 1998, but cf. Vasishth, Brüssow, Lewis, & Drenhaus, 2008).[1] Using ERPs, Xiang et al., 2009 found that such words elicited an immediate early evoked negativity in sentences like (1-b), where negation is not in the correct hierarchical position to license the negative polarity item. This contrasts with the pattern for well-formed Examples (1-a), (1-c) and indicates that the relevant hierarchical relationships are made available within a few hundred milliseconds after encountering the unlicensed target word.

(1) a. [No students [would ever say that.]]
b. ∗[A professor [with no students] [would ever say that.]]
c. [A professor [with no students] [would definitely say that.]]

The ERP response to unlicensed polarity items is but one example of the human parser's sensitivity to hierarchical structure during online comprehension, as predicted by the competence hypothesis. Related support for the competence hypothesis comes from behavioral studies that show differences in incremental reading times when a variety of structure-dependent rules are violated, such as those governing the distribution of reflexive pronouns (Sturt & Lombardo, 2005; Yoshida, Dickey, & Sturt, 2012), bound-variable pronouns (Kush, Lidz, & Phillips, 2015), and filled gaps (Phillips, 2006); see Lewis and Phillips (2015) for extensive discussion.

Further evidence for the competence hypothesis comes from eye-tracking measures collected while participants read natural texts, such as newspaper stories. In these studies, word-category expectations based on hierarchical grammars have been found to predict eye-fixation measures (Boston, Hale, Kliegl, Patil, & Vasishth, 2008; Boston, Hale, Vasishth, & Kliegl, 2011; Demberg & Keller, 2008; Fossum & Levy, 2012; Roark, Bachrach, Cardenas, & Pallier, 2009; van Schijndel & Schuler, 2015). These findings are particularly relevant to the present study, since they indicate that hierarchical structure subserves every-day comprehension.

[1] As Bornkessel-Schlesewsky and Schlewswsky (2009, p. 23–24) and Xiang et al. (2009, p. 42) point out, it remains controversial whether NPI licensing conditions should be thought of as being syntactic, semantic or pragmatic. What matters for the present point is that the conditions reflect, at least in part, hierarchical relationships that are characterized by the competence grammar.

On the other hand, there is also evidence for a two-system view that minimizes the role of hierarchical syntactic structure in comprehension. One type of evidence comes from experiments where comprehenders seem to ignore syntactic structure when syntactic cues conflict with other information (see Ferreira & Patson, 2007; Sanford & Sturt, 2002, for reviews). Another type is based on eye-tracking corpora. For instance, Frank and Bod (2011) compare predictors that are based on hierarchical grammars to those based solely on word-to-word dependencies. In this study, syntactic structure did not improve models of eye-fixation measures. This result, like the eye-tracking data described above, relies on reading data from newspaper text. Using similar models, Frank, Otten, Galli, and Vigliocco (2015) report that ERP indices of syntactic expectations are similarly insensitive to hierarchically-based predictions.

In summary, then, the literature draws conflicting conclusions regarding the role of syntactic structure in comprehension. Disagreements may reflect differences in tasks and techniques in prior work. While ERP studies and behavioral experiments have found support for the competence hypothesis, they have typically done so in a way that relies on stimuli that sharply violate syntactic expectations. By contrast, results based on naturalistic texts, such as the eye-tracking corpora mentioned above, have been mixed, with Frank and Bod (2011) coming down on the "con" side and Fossum and Levy (2012) on the "pro" side. Differences between these latter studies may reflect alternative modeling choices. In addition, behavioral measures which integrate over many stages of processing may challenge efforts to separate out effects of syntactic hierarchy from word-to-word expectations. Likewise, the ERP signatures sensitive to syntactic violations that were probed by Frank et al. (2015) have not been directly linked with sentence processing in non-violational every-day contexts. On balance, the field remains uncertain about the competence hypothesis.

1.3. Examining syntactic structure during naturalistic story listening

To better characterize the role of syntactic structure in every-day language comprehension, we examine several different types of structure. These structure types can be viewed as points on a cline of increasing syntactic detail. Each type of structure implies different partial products of the comprehension process. We focus on three levels that have received significant attention in psycholinguistics. At one end are Markov models that use linear, word-to-word, surface dependencies. These are "string-level" language models. One step further along the cline are context-free grammars (CFG). These are "tree-level" models that directly derive the sorts of immediate-constituency relationships that linguists traditionally hold up as a central part of sentence structure. The particular grammars that we use at this point in the cline are free from empty categories and lack any systematic treatment of movement. Proceeding one more step, to the deep end of the cline, are Minimalist Grammars (Stabler, 1997) (MGs). These grammars generate X-Bar structures (see e.g. Haegeman, 1999) which encode movement and make extensive use of empty categories in a drive towards greater regularity in the analysis of typologically diverse languages.

Of course, there are other formal grammars that could have been chosen to represent each distinct level of this cline. For example, certain forms of Tree-Adjoining Grammars and Categorial Grammars are weakly-equivalent to MGs (see Stabler, 2013a, and references cited therein). But such equivalences pose no particular problems for the present study. Our investigation is not intended to decide between alternative accounts of syntactic competence at any one specific level of expressivity. Rather, it probes the level of syntactic detail that is processed by the brain, using the cline as a yardstick whose notches are classes of formalisms.

We link properties of these grammars, word-by-word, with hemodynamic data collected while participants listen to a story. This linking is accomplished using two different complexity metrics. One quantifies the degree of expectation for a particular symbol, given the syntactic left-context; this is "surprisal" in the sense of Hale (2001). The second metric quantifies syntactic complexity by counting the number of tree nodes that would be visited by a surface-structure parser (Frazier, 1985, chap. 4; Miller & Chomsky, 1963).

Throughout, we focus on naturalistic sentence comprehension, since this has been an important pivot in the debate over the role of abstract structures. The passive story listening task we use does not lend itself to clearly delineated conditions and the traditional "subtraction" approach to neuroimaging analysis. Instead, we adapt the methodology of Just and Varma (2007) to construct a model of the expected hemodynamic response (see also Brennan et al., 2012; Willems, Frank, Nijhof, Hagoort, & van den Bosch, 2015). This expected response is what an experimenter should observe if a brain region were doing the work implied by each language model—as seen through the lens of the complexity metric. We evaluate these models by testing the fit between the expected hemodynamic responses and those observed in the data.

1.4. Summary

Previous work suggests that the ATL, left PTL, and left IFG form the core of a network involved in sentence comprehension, and the ATL in particular has been linked with basic constituent-level processing. The syntactic structures manipulated within this network have not been identified. We examine the nature of the syntactic structures computed by these circuits. To do so, we derive predictions of word-by-word processing complexity from a range of syntactic models, from string-level Markov models to hierarchical CFGs and MGs. These models are tested against neural timecourses recorded using fMRI while participants passively listen to a natural story. If hierarchical structures are computed during passive naturalistic listening, we predict that syntactic complexity estimates based on hierarchical language models will correlate with fMRI signal fluctuations above and beyond those derived from non-hierarchical, string-level models.

2. Methods

2.1. Participants

Twenty-nine college-age volunteers participated for pay (17 women and 12 men, 18–24 years old). All qualified as right-handed on the Edinburgh handedness inventory Oldfield, 1971. They self-identified as native English speakers and gave their informed consent. As detailed below, we excluded from our analyses data from one participant due to excessive head movement and data from two participants due to poor behavioral performance, leaving twenty-six datasets for our analyses (15 women, 11 men).

2.2. Stimuli & procedure

The audio stimulus was Kristen McQuillan's reading of the first chapter of Lewis Carroll's Alice in Wonderland from librivox.org. We chose this text because of its enjoyability, its use in prior imaging work (Brennan et al., 2012), and because of available fine-grained syntactic annotations (VanWagenen, Brennan, & Stabler, 2014). The chapter we used does not include significant word-play, such as the famous Jabberwocky poem that appears elsewhere in the story. To improve comprehensibility in the noisy scanner, the audio was normalized to 70 dB and slowed by 20% with the pitch-preserving PSOLA algorithm implemented in Praat software. This moderate amount of time-dilation did not introduce recognizable distortion and was judged by an independent rater to sound natural and to be easier to comprehend than the raw audio recording. The audio presentation lasted 12.4 min. The stimulus is available as Supplementary Material.

After giving their informed consent, participants were familiarized with the MRI facility and assumed a supine position on the scanner gurney. Auditory stimuli were delivered through MRI-safe, high-fidelity headphones (Confon HP-VS01, MR Confon, Magdeburg, Germany) inside the head coil. The headphones were secured against the plastic frame of the coil using foam blocks. Using a spoken recitation of the US Constitution, an experimenter increased the volume stepwise until participants reported that they could hear clearly. Participants then listened passively to the audio storybook. Upon emerging from the scanner, participants completed a twelve-question multiple-choice questionnaire concerning events and situations described in the story. The entire session lasted less than an hour.

2.3. Modeling syntactic effort

We constructed nine models which quantified word-by-word syntactic processing effort. These models spanned three different levels of syntactic detail and drew from two different complexity metrics.

2.3.1. Two complexity metrics

For probabilistic language models, we linked the probability of a word in its left-context to the BOLD signal using the log-reciprocal of the probability of the next word. This is "surprisal" in the sense of Hale (2001).

With non-probabilistic grammars, we linked the syntactic structure of a sentence to the BOLD signal it evokes by counting the number of tree nodes between successive words (Frazier, 1985, chap. 4; Hawkins, 1994; Miller & Chomsky, 1963). Counts included "empty" nodes such as the traces of movement. If one thinks of nodes as consuming stack cells, this becomes a kind of depth hypothesis in the sense of Yngve (1960). We considered two parsing strategies: top-down and bottom-up (see e.g. Hale, 2014, chap. 3). The top-down traversal that we used enumerates nodes in a depth-first, left to right order analogous to an LL parser. The bottom-up traversal that we used enumerates daughters before mothers in the manner of a shift-reduce LR parser. Taking the story stimulus to be largely unambiguous for native English-speaking listeners, we assume that the parser enumerates nodes of just the correct structure when faced with temporary ambiguities (i.e. a "perfect" oracle).

These complexity metrics could not both be applied at each level of syntactic detail. As described below, only the surprisal metric was defined for the least abstract Markov models, and only the node count metric was applied at the most abstract MG level of detail.

2.3.2. String-level: Markov models

Word-to-word surface dependencies were modeled using n-gram Markov models; these models involve a minimal level of syntactic abstraction. Rather, they define the probability of a word at position j, denoted w_j, in terms of the preceding $n-1$ words. A 2-gram model considers $P(w_j|w_{j-1})$ while a 3-gram model considers $P(w_j|w_{j-1}, w_{j-2})$ and so-forth. Lexicalized models define probabilities of actual words, while unlexicalized models define the probability of part-of-speech tags (POS).

We used OpenGRM to fit Markov models of various orders (Allauzen, Riley, Schalkwyk, Skut, & Mohri, 2007). Linguistic expectations have been shown to be highly sensitive to

experiment- and genre-specific idiosyncrasies (e.g. Fine, Jaeger, Farmer, & Qian, 2013). To best approximate such expectations, these models were trained on the entire text of *Alice in Wonderland* that is distributed by Project Gutenberg, etext # 11. As a preprocessing step, chapter headings were removed and all words converted to lowercase. Note that the test data for our models was fMRI signals, not corpus occurrences, and thus no circularity was introduced by including the text that corresponded to our stimulus within the training set.

Four models in total were constructed for this class. Models differed in order (2 or 3) and whether they were lexicalized or unlexicalized. These models do not allow for the representation of syntactic nodes and so only the surprisal complexity metric is defined. We identify these models as *2gram.l*, *2gram.p*, *3gram.l*, and *3gram.p*.

2.3.3. Hierarchical: Context-free phrase structure

Context-free phrase structure grammars (CFGs, defined on page 6) model sentence structure as a hierarchy of phrases. In this sense, they are more abstract than the Markov models discussed above. Although CFGs are inadequate to characterize natural languages in certain key respects—for instance, cross-serial dependencies (Joshi et al., 1990; Shieber, 1985; Stabler, 2013a)—they are very commonly used in broad-coverage parsing systems.

We constructed a family of models based on treebank CFGs.[2] In these grammars, many constructions that are syntactically similar end up being listed separately; this lack of abstraction highlights the positioning of these CFGs at a middle point on the cline of structure types. An example phrase structure, in the style of the Penn Treebank, is illustrated in Fig. 1 (top).

We used the EarleyX implementation of Stolcke's probabilistic Earley parser to compute surprisal values (Luong, Frank, & Johnson, 2013; Stolcke, 1995). Rules were read off the output of the Stanford parser (Klein & Manning, 2003b) and probabilities were trained using the entire *Alice in Wonderland* text, just as with the Markov models described above. Punctuation was removed. The node count predictors were based on the same Penn Treebank structures as in Brennan et al. (2012).

Three models were constructed for this class, differing in terms of the complexity metric used to derive effort: Surprisal, bottom-up node count, and top-down node count. We denote these models *cfg.surp*, *cfg.bu*, *cfg.td*, respectively.

2.3.4. Dependency-capable: Minimalist Grammar

At the most abstract level of syntactic detail, we used Minimalist Grammars (MG; Stabler, 1997; Stabler, 2011). This formalism derives binary-branching "X-bar" structural descriptions that integrate constituency, dependency and movement information. The particular syntactic analyses that we used extend those in Hale (2003, chap. 4) along the lines of Sportiche, Koopman, and Stabler (2013). Fig. 1 (bottom) illustrates one such tree in which the representation of a long-distance "movement" relationship leads to node counts that are different from those derived by a context-free analysis of the same sentence (Fig. 1, top). While MGs are not the only grammar formalism that adequately covers long-distance dependencies (see e.g. Müller, 2015) the fact that they include a nonconcatenative rule, one that goes beyond the mechanisms of context-free grammar, is a key part of the "hidden consensus" among the many formal approaches to grammar in modern linguistics (see e.g. Stabler, 2013a, §17.2).

The current study examines MGs only through the lens of node count linking hypotheses. While information-theoretical complexity metrics like surprisal are well-defined for MGs (Hale, 2003; Hunter & Dyer, 2013; Yun, Chen, Hunter, Whitman, & Hale, 2015), computing their values from wide-coverage grammars requires approximations analogous to those typically applied with CFGs (Charniak, Goldwater, & Johnson, 1998; Klein & Manning, 2003a). Such techniques are a current focus of MG parsing research (e.g. Stabler, 2013b), but are not available for our application. As they mature, we expect to be able to deploy them in future modeling efforts. We identify the two models in this class *mg.bu*, *mg.td* based on top-down and bottom-up enumeration, respectively.

2.4. Data collection and analysis

Imaging was performed using a 3T MRI scanner (Discovery MR750, GE Healthcare, Milwaukee, WI) with a 32-channel head coil at the Cornell MRI Facility.

Blood Oxygen Level Dependent (BOLD) signals were collected from twenty-nine participants. Thirteen participants were scanned using a T2*-weighted echo planar imaging (EPI) sequence with: a repetition time of 2000 ms, echo time of 27 ms, flip angle of 77°, image acceleration of 2X, field of view of 216 × 216 mm, and a matrix size of 72 × 72. Under these parameters we obtained 44 oblique slices with 3 mm isotropic voxels. Sixteen participants were scanned with a three-echo EPI sequence where the field of view was 240 × 240 mm resulting in 33 slices with an in-plane resolution of 3.75 mm^2 and thickness 3.8 mm. This multi-echo sequence was used for reasons that are not related to the present study. For our purposes, analyses of this second group were based exclusively on images from the second EPI echo, where the echo time was 27.5 ms. All other parameters were exactly the same. This selection of the second-echo images renders the two sets of functional images as comparable as possible.

2.4.1. fMRI preprocessing

Preprocessing was done with SPM8 (Friston, Ashburner, Kiebel, Nichols, & Penny, 2007), Data were spatially realigned based on 6-parameter rigid body transformation using the 2nd degree B-spline method. Functional (EPI) and structural (MP-RAGE) images were co-registered via mutual information and functional images were smoothed with a 3 mm isotropic gaussian filter. We used the ICBM template provided with SPM8 to put our data into MNI stereotaxic coordinates. The data were high pass filtered at 1/128 Hz and we discarded the first 10 functional volumes. Data from one participant was excluded at this stage due to head movement that exceeded an absolute threshold of 1 mm.

2.4.2. Deriving estimated BOLD signals from syntactic models

Via the models described in Section 2.3 above, we predicted the level of syntactic processing effort at each word in the stimulus text. Specifically, we defined point events at the offset of each word, whose intensity is proportional to this predicted effort. The predicted effort from each model is illustrated for three example sentences in Supplementary Fig. S1. This yields a time series of theoretical predictions for each point along the cline of syntactic structures. Following Just and Varma (2007), we convolved these time series with a canonical hemodynamic response function (HRF). Such a procedure yields an expected BOLD signal for each syntactic predictor under the assumption that the BOLD signal reflects the output of a linear system (Boynton, Engel, Glover, & Heeger, 1996). Fig. 2(A)–(E) summarizes this methodology graphically and illustrates how values derived from different models yield distinct predictors for brain activity.

Left alone, the resulting estimate for brain activity is dominated by the narrator's speech rate. Higher estimates appear for words in rapid succession and smaller estimates for segments where words are more widely spaced. Following Brennan et al. (2012), we define

[2] A treebank is a collection of hand-analyzed syntactic representations, see e.g. Marcus, Santorini, and Marcinkiewicz (1993) or Jurafsky and Martin (2009, chap. 12).

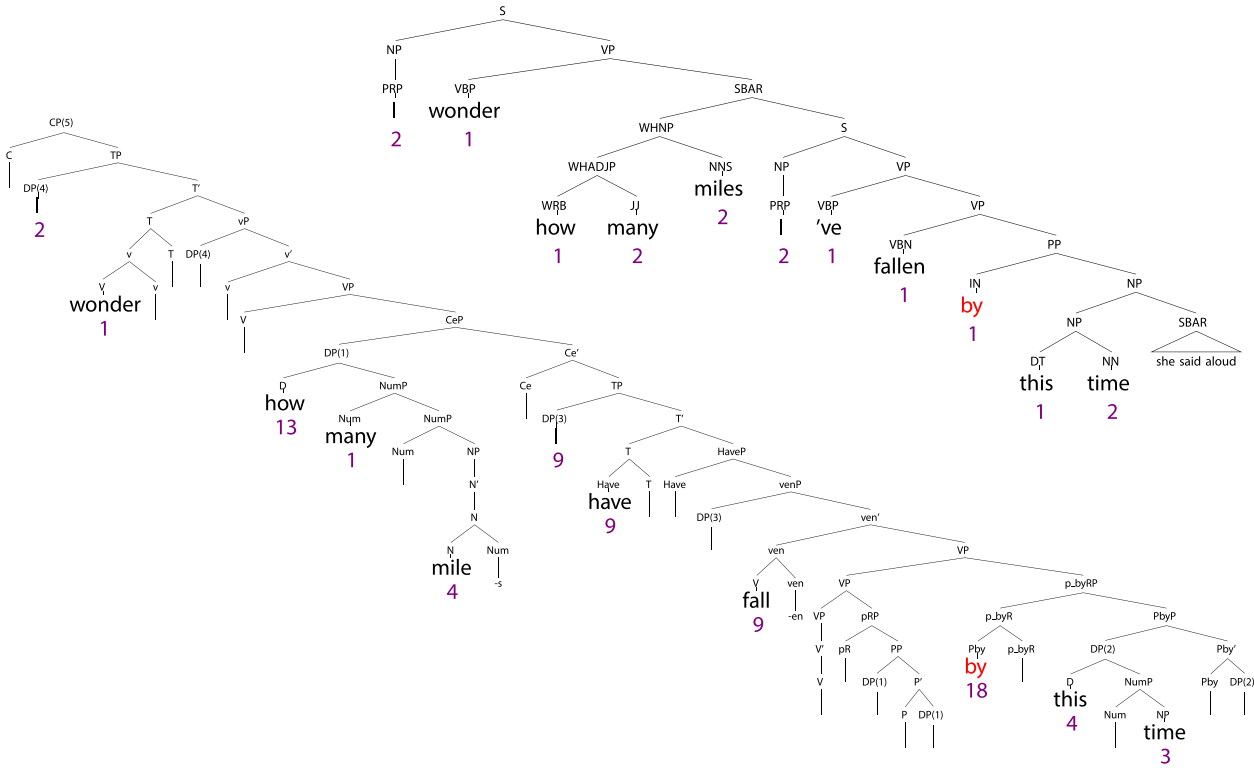

Fig. 1. Less-detailed CFG analysis (top) versus more-detailed MG analysis (bottom) of the same sentence. Numbers beneath each word (purple) are estimates of syntactic "effort" which is used to derive BOLD signal predictors from a node count based on a bottom-up enumeration. Node count reflects the presence of empty nodes in the MG but not the CFG. This aspect of the structure impacts estimates of processing effort, for example, at the word "by" which is highlighted in red.

a baseline Word Rate predictor with a value of one at the offset of each word and zero otherwise. We orthogonalize all syntactic predictors against this baseline, after it has been convolved with the HRF, in order to isolate the truly syntactic aspect of the predictor. Panel E of Fig. 2 illustrates this step.

2.4.3. Regions of interest

Models were evaluated against fMRI timecourses from six regions of interest (ROIs). We followed the theory-driven ROI analysis with an exploratory whole brain analysis that is described in the next section.

We defined ROIs on a per-participant basis using both functional and anatomical criteria (Fedorenko, Hsieh, Nieto Castanon, Whitfield-Gabrieli, & Kanwisher, 2010). The functional criteria were based on an a-theoretical language localizer using the Word Rate predictor introduced in Section 2.4.2. This predictor identifies brain regions whose BOLD signal increases each time a word is presented. Such a pattern is expected to be found in regions sensitive to any activation that is time-locked to word presentation, including those involved in incremental sentence processing as well as regions sensitive to lexical, sub-lexical, and auditory processes. Analyses at the single-participant and group levels verified that this localizer activated a broad set of temporal and frontal perisylvian regions with a left-hemisphere bias (see Figs. 3 and 5). Anatomical criteria were added to narrow our focus to brain regions associated specifically with sentence-level processing. Each ROI was a sphere with radius 10 mm centered on a peak *t*-value of at least 2.0 for the language localizer within the anatomical constraints that are described below. Data from every voxel within each sphere was averaged into a single timecourse per ROI. We discuss the (in)sensitivity of our results to size and inclusion criteria used to define the ROIs in Section S.3 of the Supplementary Materials.

Peaks that fell bilaterally within the superior, middle, or inferior temporal gyri with greater than 50% probability in the Harvard-Oxford Brain Atlas, and were anterior to Heschl's Gyrus, served to define the center of left and right anterior temporal regions (LATL; RATL) (cf. Rademacher, Galaburda, Kennedy, Filipek, & Caviness, 1992). In some participants, multiple peaks with similar statistical values fit these criteria. In these cases, anteriority (including the temporal pole) and non-contiguity with posterior temporal activity were included as additional criteria. Anterior temporal lobe has shown sensitivity to the presence vs. absence of hierarchical constituent structure (e.g. Bemis & Pylkkänen, 2011; Humphries et al., 2006; Pallier et al., 2011; Snijders et al., 2009; Stowe et al., 1998; Vandenberghe et al., 2002). Several studies report bilateral activation (Rogalsky & Hickok, 2009; Stowe et al., 1998, but cf. Humphries, Love, Swinney, & Hickok, 2005 for the role of prosody in the right hemisphere). Further, anterior temporal brain damage to this region correlates with deficits in morphosyntax (Dronkers et al., 2004) and anterior temporal atrophy has been associated with deficits in combinatorial semantics (Wilson et al., 2014). These data, and others, have led to the proposal that the anterior temporal lobe is involved in basic compositional processes (Friederici & Gierhan, 2013).

Maxima of the language localizer that fell with greater than 50% probability in the left superior temporal or middle temporal gyri, and were posterior to Heschl's Gyrus, defined a left posterior temporal region of interest (LPTL). This region shows sensitivity to the presence and complexity of phrase structure (e.g. Bemis & Pylkkänen, 2013; Pallier et al., 2011); though the functional role of this region remains poorly understood, one recent proposal links it with order-related processing (Bornkessel-Schlesewsky et al., 2015).

Peaks of the language localizer that fell within the left Inferior Frontal Gyrus with greater than 50% probability in the Harvard-

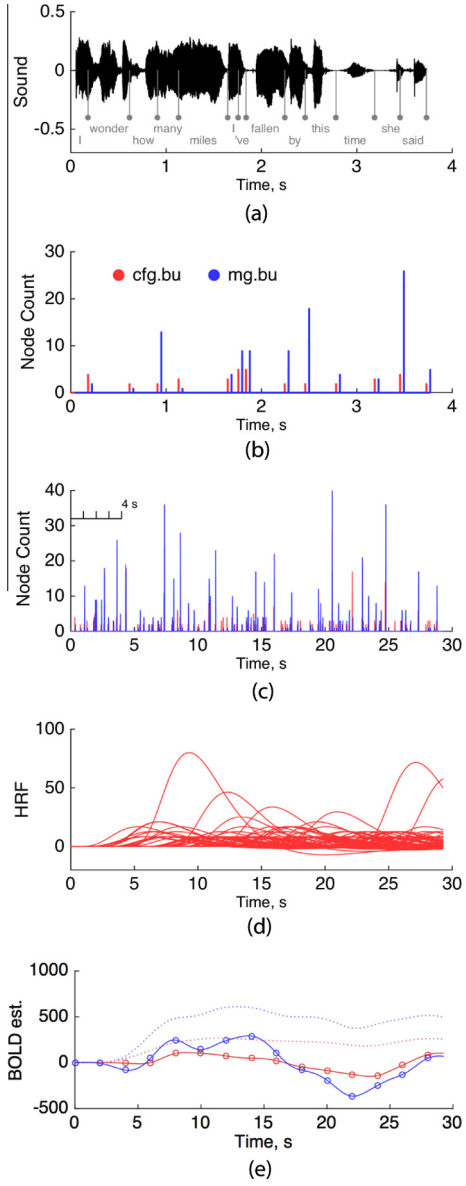

Fig. 2. Deriving an expected BOLD signal from linguistic structure: (A) The spoken narrative is segmented into words. (B) A complexity metric such as node count defines the intensity of point events at the offset of each word according to a particular grammar; examples from a context-free grammar (red) and Minimalist Grammar (blue) derived using bottom-up enumeration are shown (see Fig. 1). Panel (C) illustrates the same complexity counts over a longer interval. (D) The points are then convolved with the canonical HRF (only one grammar is illustrated in this panel). (E) Results are summed to yields estimated BOLD responses (dotted) which are then made orthogonal to the Word Rate covariate (solid) and, finally, sampled at 0.5 Hz to match the sampling rate of the collected data (open circles). The two solid lines in panel (E) illustrate how different grammatical representations yield diverging estimates for BOLD signals associated with syntactic processing.

Oxford Brain Atlas defined the center of our left inferior frontal gyrus region (LIFG). In cases where multiple peaks fit these criteria, proximity to the Pars Opercularis was included as an additional criterion. Numerous findings from lesion-induced syntactic deficits (Caramazza & Zurif, 1976; Grodzinsky, 2000) and neuroimaging of brain activations for syntactically complex sentences (inter alia Just et al., 1996; Santi & Grodzinsky, 2007b; Snijders et al., 2009; Stowe et al., 1998; Stromswold et al., 1996) have implicated this region in various aspects of grammatical processing (see Rogalsky & Hickok, 2010 for a critical review). Recent research points to a locus in the Pars Opercularis, specifically, for hierarchical composition (Zaccarella & Friederici, 2015).

Maxima that fell within either the Angular Gyrus or Supramarginal Gyrus with greater than 50% probability defined an inferior parietal region (LIPL). Finally, maxima that fell within the posterior aspect of the Middle Frontal Gyrus region of the left hemisphere with greater than 50% probability defined a premotor region (LPreM). Both of these regions have been implicated in sentence-level processing that is sensitive to linear order (see Bornkessel et al., 2005 for functional imaging evidence and Wilson et al., 2011 for evidence from neurodegenerative disorders). Bornkessel-Schlesewsky et al. (2015) propose a model which contrasts order-sensitive processing in these posterior-dorsal regions with hierarchy-sensitive processing in ventral-anterior regions such as the ATL.

Fig. 3 illustrates each of these ROIs in four representative participants. Supplementary Table S1 lists MNI coordinates and peak activation values for the center of each ROI for all participants.

2.5. ROI statistical analysis using stepwise model comparison

To provide an overview of which models correlate with measured brain activity, we first constructed a family of mixed-effects regression models using non-syntactic control predictors, described below, together with one syntax predictor drawn from each model. Fitted beta coefficients with a 95% confidence interval that did not include zero were taken to be "statistically significant" (Gelman & Hill, 2007). This first step is limited, however, as it does not take into account the relationships between the different models. Indeed, bivariate correlations between each syntactic predictor and also with non-syntactic control predictors showed non-trivial effects (Supplementary Fig. S2). Of the target syntactic predictors, *2gram.p* and *3gram.p* were highly correlated with each other ($r = 0.84$). The *cfg.surp* model was anti-correlated with unigram frequency ($r = -0.54$), and positively correlated with *2gram.p* and *3gram.p* ($r \approx 0.45$). The node count models were moderately-to-highly correlated with each other: $r(cfg.td,\ cfg.bu) = 0.43$; $r(mg.bu,\ mg.td) = 0.9$.

To evaluate the unique contribution made by each type of syntactic structure, we conducted step-wise model comparisons using likelihood ratio tests. The models were ordered both by amount of syntactic detail ($ngram \prec cfg \prec mg$),X and by a rough-and-ready characterization of the number of parsing assumptions in the complexity metric ($surp \prec bu \prec td$). Judging the number of parsing assumptions is subjective, and so we conducted an auxiliary analysis to test the sensitivity of any results to this parameter by reversing the ordering by complexity metric (see Supplementary Table S3). The list of fixed effects for each model entered into this comparison is given on the left-hand side of Table 1.

All models included fixed effects for sound power, Word Rate (defined in Section 2.4.3, above), word frequency (log-transformed values from the HAL corpus via the English Lexicon Project; Balota et al., 2007), and six parameters representing estimated head movements. A fixed effect for prosodic breaks was also included to control for correlations between acoustic variance and syntactic structure. This predictor is a perceptual judgment of break index strength made in light of ToBI annotation guidelines (Beckman, Hirschberg, & Shattuck-Hufnagel, 2005) by two independent raters. All predictors except for those representing head movements were converted to z-scores. Each model also included a random intercept by participant and a random slope for the baseline Word Rate predictor.

Statistical significance was evaluated against an alpha level of 0.05 that was corrected for multiple comparisons across six ROIs with the bonferroni method (raw *p*-values, which are reported in the Supplementary Materials, should be evaluated against an adjusted alpha-level of $0.008\bar{3}$).

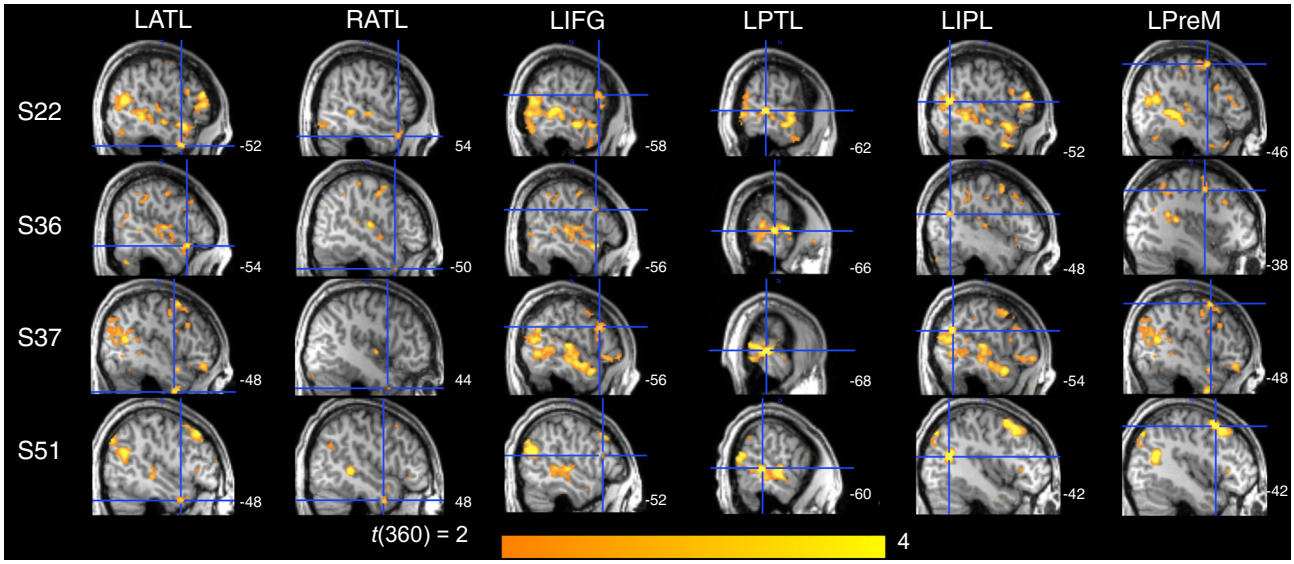

Fig. 3. Regions of interest (columns) from four representative participants (rows). Regions were defined by conjoining activation peaks based on the language localizer Word Rate predictor with anatomical definitions for each sentence-related region (see Methods). Integers indicate the MNI location of the sagittal slice shown in each frame.

2.6. Whole brain analysis

As a follow-up to the theory-driven analysis, we conducted an exploratory analysis of the whole brain using a subset of our predictors. This analysis permits us to test for activations that may fall outside of regions traditionally implicated in sentence-level processing. However doing so necessarily sacrifices power to detect possibly subtle differences between models.

A first-level General Linear Model (GLM) was fit for each voxel of each individual participant.[3] The non-syntactic predictors were sound power, word rate, word frequency, prosodic breaks, and six head movement predictors. We added to this baseline model three syntactic predictors to represent each of the three levels of syntactic detail. We selected the most robust predictor from each level based on the ROI results. As detailed in the Results section, below, the most robust predictors were *2gram.p*, *cfg.surp*, and *mg.td* (see Fig. 4 and Table 1). To align the whole brain analysis with the model comparison-based ROI analysis, the three syntactic predictors were first residualized against all lower-level predictors according to the ordering of models shown in Table 1. For example, *2gram.p* was residualized against *2gram.l*, *3gram.l*, and all non-syntactic predictors. By residualizing the predictors in this way, the whole brain analysis is sensitive to the unique contribution of each predictor independent of the contribution from lower-level predictors.

At the second, group, level, beta values from the first-level GLMs from each participant were evaluated with one-sampled *t*-tests. We report as "statistically significant" voxels with a *p*-value of at least 0.001 in clusters of at least 50 voxels that were reliable at $p < 0.05$ after correcting for the number of comparisons and the estimated smoothness of the data according to Random Field Theory (Worsley et al., 1996).

3. Results

3.1. Behavioral results

The quiz comprised twelve questions, each with four possible answers. Under the cumulative binomial distribution, $P(score \geq 7) = 0.014$. Two participants who scored lower than this threshold were discarded from further analysis. The remaining 26 participants had a median score of 10 with a range of [7 12]. This means that all participants whose data were analyzed scored higher than would be expected by chance.

3.2. fMRI Region of interest results

Fig. 4 shows the estimated coefficients and 95% confidence intervals for each of the syntax predictors when included alone in a model with only non-syntactic and physiological "nuisance" predictors. Treated independently of each other, significant correlations were observed for unlexicalized 2- and 3-gram models in LATL, RATL, LIFG and LPTL. CFG surprisal estimates were significant in all six ROIs. Node count CFG predictors were significant in the LATL and LPTL but not in any other region. Node counts derived from the MG were significant predictors in the LATL, RATL, LIFG and LPTL. Estimated parameters for all of the linguistic coefficients for each model are shown for each ROI in Supplementary Fig. S3.

These results suggest a role for both word-to-word and hierarchical dependencies in characterizing BOLD signal across the ROIs. The most abstract aspects of sentence structure were most predictive in the temporal lobe. However, this first analysis does not take into account co-dependencies between these different syntactic predictors. Accordingly, model comparison was used to evaluate the independent contribution of the more abstract models, above and beyond any effects due to more concrete string-level dependencies.

This model comparison evaluated a family of nested models, identified by the letters A through I in Table 1. The main result is that predictors based on MG and CFG each improved a mixed-effects model of the neural time course in the temporal lobe during naturalistic story comprehension.

Table 1 column 3 summarizes the model comparison in LATL. Surprisal based on word trigrams and POS bigrams improved the fit of a regression over the null model. Additional improvements were found for surprisal based on the CFG, node counts based on a top-down traversal of CFG structures, and for node counts based on a top-down traversal of X-Bar structures generated by MGs. In other words, even the most abstract grammars led to significant improvements in explaining the timecourse of the LATL response, above and beyond the variance explained by Markov models.

[3] We used a mask that includes all voxels within the envelope of the MNI 152-brain average. The default masks calculated by SPM8 for each participant exclude some voxels in the orbito-frontal cortex and from the anterior temporal lobes.

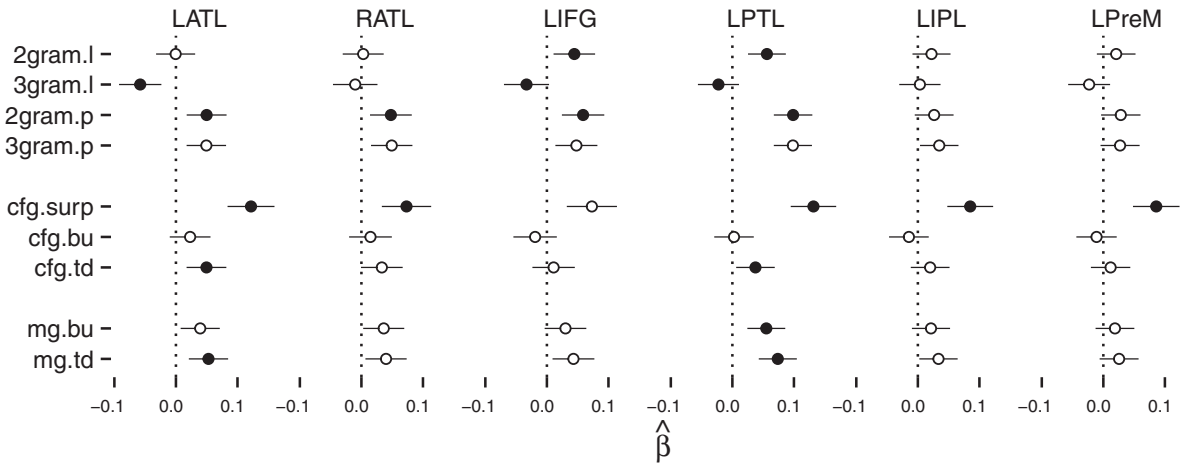

Fig. 4. Fitted coefficients for all syntax predictors across six ROIs. Coefficients show the estimated change in BOLD signal per unit change in the syntactic predictor (*x*-axis). The nine models are ordered in descending order along the *y*-axis based on syntactic detail and complexity metric. Error bars show 95% confidence intervals based on Wald's approximation. Filled points indicate models that made a statistically significant contribution in a descending step-wise comparison against simpler models (see Table 1).

Table 1
Summary of results for nested model comparison from six ROIs. Only timecourses from ROIs where the per-participant functional localizer had a peak $t \geq 2.0$ were included for analysis; these are counted in the last row. *p*-Values are corrected for multiple comparisons across ROIs. Statistical details for each ROI are given in Supplementary Table S2.

Model	Description	LATL	RATL	LIFG	LPTL	LIPL	LPreM	Syntactic detail
Ø	Sound power, word rate, word frequency, prosodic, and movement but no syntactic predictors							Less
A	add *2gram.l*			$p < 0.05$	$p < 0.05$			
B	add *3gram.l*	$p < 0.001$		$p < 0.01$	$p < 0.001$			
C	add *2gram.p*	$p < 0.001$	$p < 0.001$	$p < 0.05$	$p < 0.05$			
D	add *3gram.p*							
E	add *cfg.surp*	$p < 0.001$	$p < 0.001$		$p < 0.05$	$p < 0.05$	$p < 0.05$	
F	add *cfg.bu*							
G	add *cfg.td*	$p < 0.01$			$p < 0.01$			
H	add *mg.bu*				$p < 0.05$			
I	add *mg.td*	$p < 0.05$			$p < 0.01$			More
Number of timecourses		23	22	21	24	24	24	

Detailed model comparison statistics for this region and for the regions discussed below are given in Supplementary Table S2.

The RATL (Table 1 column 4) showed significant improvements in fit for POS bigrams and CFG surprisal but not for any other models.

In LIFG, the situation was quite different. As shown in column 5 of Table 1, we observed improved fits for word-based bigram and trigram models, and also for POS-based bigram models, but neither CFG nor MG hierarchical models led to a significant improvement in regression fits. While CFG surprisal did show a significant correlation when considered alone (see Fig. 4), this model was not significant when variance due to *n*-gram models was taken into consideration.

Results from the LPTL were similar to those of the LATL (Table 1 column 6): we observed effects for *n*-gram models, additional effects for CFG surprisals and node counts and, further, effects for MG node counts based on both a bottom-up and a top-down traversal.

Finally, fits against activation from the LIPL and LPreM regions were improved only by CFG surprisals. Neither *n*-gram models, nor MG models showed significant effects in these two ROIs. These effects are shown in columns 7 and 8 of Table 1.

3.3. Whole brain fMRI results

Following the theory-driven ROI analysis, we conducted an exploratory analysis across the whole brain with a focus on three syntactic predictors: *2gram.p*, *cfg.surp*, and *mg.td*. Whereas the ROI analysis addresses what kind of syntactic information is processed in brain regions involved in sentence-level processing, the whole brain analysis explores the complementary question of what regions are sensitive to different kinds of syntactic information. These three syntactic predictors were, respectively, the most robust in the ROI analysis for each level of syntactic detail. They were entered into a whole brain GLM along with non-syntactic control predictors, including the language localizer Word Rate predictor after being residualized against lower-level covariates. Fig. 5 illustrates a selection of the results of this analysis. Statistical maps for all predictors are shown in Supplementary Fig. S4 along with detailed results in Supplementary Table S8.

The Word Rate localizer predictor showed robust effects across the temporal and frontal lobes with more reliable effects observed in the left-hemisphere. These are shown in red on Fig. 5. This result serves as a "sanity check" showing that the localizer predictor correlates with activity across traditional language regions (e.g. Friederici & Gierhan, 2013). This result also offers a back-drop for more localized effects observed for other predictors.

The *2gram.p* predictor, representative of string-level models, showed significant effects in the posterior temporal lobe bilaterally. This pattern closely matches the results reported by Willems et al. (2015) who applied a trigram part-of-speech language model and used similarly naturalistic spoken narratives for stimuli. Unlike Willems et al., we also found significant bilateral activation in the fusiform gyrus for this predictor. These results are shown in blue on Fig. 5.

We did not observe any statistically reliable correlations for the *cfg.surp* or *mg.td* predictors. At an uncorrected $p < 0.01$ threshold,

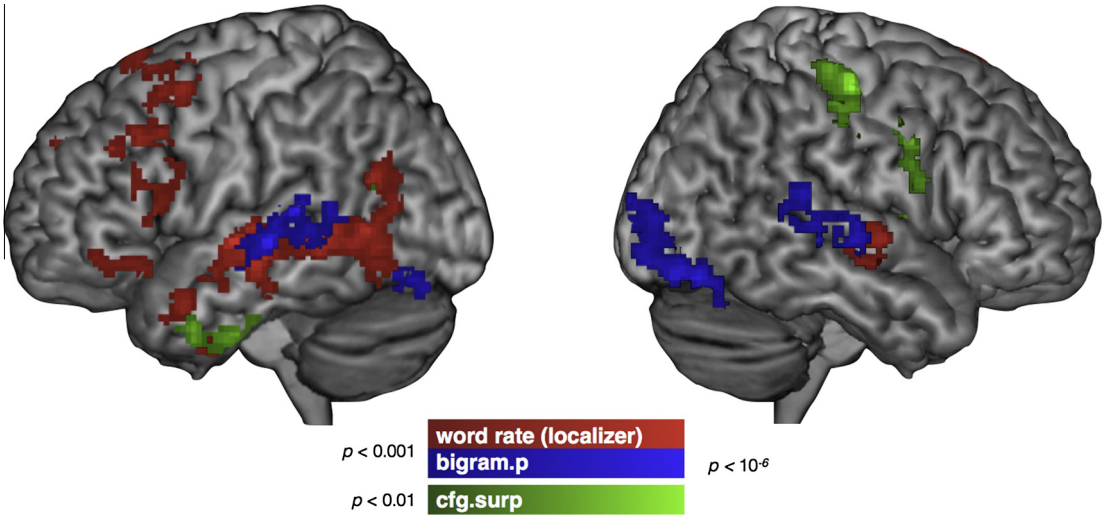

Fig. 5. Whole brain activation maps for three predictors rendered onto the surface of a template brain ($N = 26$). Maps for the language localizer Word Rate predictor (red) and *2gram.p* predictor (blue) are thresholded at $p < 0.001$ with a cluster size of at least 50 voxels (family-wise $p < 0.05$). The *cfg.surp* predictor (green) is shown at a liberal $p < 0.01$ threshold, illustrating non-significant trends in this analysis.

cfg.surp correlated with clusters of voxels in the left anterior temporal lobe ($p_{corrected} = 0.423$) and with a cluster of voxels in the right hemisphere spanning the central sulcus ($p_{corrected} = 0.277$; see Supplementary Table S8). These uncorrected observations are shown in green on Fig. 5. Even at liberal thresholds, there were no coherent clusters of activation for *mg.td* (this is shown on the bottom-right panel of Supplementary Fig. S4).

4. Discussion

Using fMRI, we evaluated alternative hypotheses regarding the syntactic information computed by neural circuits involved in naturalistic comprehension. This evaluation considered six brain regions that have been traditionally associated with sentence processing and was followed-up by an exploration across the entire brain. We correlated estimates of processing effort drawn from different syntactic models with fMRI timecourses. Comparing the fits of different models, we find support for abstract hierarchical structure in the left anterior and posterior temporal lobe, but not in the left inferior frontal gyrus or in dorsal parietal and premotor regions. These latter two regions showed sensitivity to phrase structure, but not to the most abstract structures that we considered. A whole brain analysis did not show any significant activations elsewhere. By contrast, string-level Markov models over part of speech tag sequences correlated with the fMRI-measured signal in the inferior frontal gyrus, the anterior temporal lobe bilaterally and the left posterior temporal lobe. This pattern of results constitutes support for the competence hypothesis that abstract hierarchical grammars subserve real-time natural comprehension. It also supports the characterization of the sub-parts of the temporal lobe as a kind of combinatorial hub.

4.1. Evidence for abstract hierarchy in naturalistic comprehension

Regarding the competence hypothesis, two findings in particular are significant. The first is syntactic structures generated by CFGs are helpful in predicting timecourses from all ROIs save the LIFG. This result obtained using the surprisal linking hypothesis. These CFG-based surprisals were predictive after taking into account *n*-gram and other predictors, such as unigram word frequency and prosodic break size. This suggests that abstract hierarchical structure plays a role in on-line comprehension, even in a task-free environment. This finding corroborates experimental work showing early effects of syntactic structure on on-line processing (Kush et al., 2015; Phillips, 2006; Sturt & Lombardo, 2005; Xiang et al., 2009; Yoshida et al., 2012). Our results also align well with naturalistic eye-tracking studies that demonstrate sensitivity to expectations based on hierarchical structure (Fossum & Levy, 2012; van Schijndel & Schuler, 2015).

The second finding that bears on the competence hypothesis is the pattern of fits obtained through node counts. Node counts derived from the CFG correlated with activity from LATL and LPTL but no other region, as shown in Fig. 4. This partially replicates Brennan et al. (2012). The strongest support for the competence hypothesis in this study comes from node counts in X-bar trees that are generated by MGs. These were positive predictors of BOLD signal in the LATL and LPTL when considered on top of CFG node counts, string-level expectations, and non-syntactic predictors.

The simplest interpretation is that these temporal lobe regions do a computation that is isomorphic, in some way, to the abstract structures that MGs and CFGs strive to capture. Both node count and surprisal seem to point toward the same sort of structure-dependence. This suggests that while the dynamics of processing in the temporal lobe may indeed be experience-based, as reflected by the surprisal effects, they are also correlated with the raw amount of syntactic structure.

Our findings contrast with those of Frank and Bod (2011) and Frank et al. (2015). In these studies, processing complexity predictions from hierarchical grammars turn out not to fit eye-movement measures or evoked scalp potentials any better than predictions based on string-level models. The positive results that we obtain may be attributable to the use of fMRI. By measuring a spatially-specific BOLD signal, our analysis is evidently able to detect hierarchical processing in just two regions. Indeed, the less abstract Markov models that we considered were predictive in a broader set of ROIs. It could be that word-to-word effects drown out the indicators of hierarchical processing in certain eye-tracking measures and ERP components.

4.2. Incremental syntactic parsing in the temporal lobe

Existing neurobiological models of sentence comprehension are divided as to the role of the anterior temporal lobes, posterior temporal lobe and inferior frontal gyrus in performing basic

constituent-building computations. A prominent hypothesis links the ATL with such computations based on imaging and lesion-based data (Friederici & Gierhan, 2013; Hickok & Poeppel, 2007). It suggests that ATL activation ought to be sensitive to constituency properties as they are incrementally identified. The results of the present studies confirm this suggestion, but also show that such a correlation holds in the posterior temporal lobe as well. This latter effect is consistent with patterns of syntactic deficits that have been observed in patients with atrophy to posterior temporal regions (Wilson et al., 2011). In virtue of using formal grammars, however, we gain a more specific understanding of the operations carried out by this neural circuit. What previous accounts have labeled "basic syntactic processes" or "constituent-building", our results characterize in terms of intermediate parser states.[4]

As already mentioned above, the most abstract syntactic structures that we examined were only predictive in LATL and LPTL but not in RATL, LIFG or dorsal parietal and premotor regions. In these regions, the comparatively finer-grained X-bar structures did not improve regression models the way they did in the temporal lobe. This result indicates that we did not find evidence to support a role for this most abstract level of detail outside of the left temporal lobe. The effect is consistent with earlier findings that used node count but not surprisal (Brennan et al., 2012). These predictors, based on MGs, take into account movement, empty categories and other aspects of Minimalist syntax. But they do so only through the lens of tree nodes. The node count linking hypothesis does not separately distinguish the formation, checking, or maintenance of long-distance dependencies. It could be that the LIFG plays a distinctive role during these operations (Grodzinsky & Friederici, 2006). On the other hand, our results do not straightforwardly support the idea that LIFG is the seat of syntactic and semantic integration in general (Hagoort, 2013). Rather, LIFG activity correlated with string-level *n*-gram expectations, as well as with unigram word frequency (see Supplementary Fig. S4 and Table S8). This latter result extends previous findings (e.g. Fiebach, Friederici, Muller, & Cramon, 2002) to naturalistic listening. Further work, perhaps using the unification spaces of Vosse and Kempen (2000), with careful attention to fine-grained spatial detail in the spirit of Zaccarella and Friederici (2015), may help clarify what role, if any, LIFG plays in basic sentence comprehension.

The temporal lobe results also align well with results obtained by Wehbe et al. (2014) using written (as opposed to auditory) stimuli. Wehbe and colleagues used predictors based on the labels of the dependency arcs, for instance `SBJ` for subject `OBJ` for direct object or `PMOD` for prepositional modifier. No hierarchical relationships entered into this labelling. However, as predictors they converge on some of the same brain regions as in the present work. Specifically, Wehbe et al. classified short text passages on the basis of the fMRI images they elicited in readers' brains. In a searchlight-style analysis, their classifier performed above chance level in both a left hemisphere posterior temporal area and a right hemisphere anterior temporal area. While this *right* hemisphere localization was unexpected, it coheres well with our bilateral results for CFG-based surprisal. It suggests the existence of a temporal lobe language network that normally employs both hemispheres, to some degree.

Bornkessel-Schlesewsky et al. (2015) suggest that dorsal-posterior regions may be involved in processing related to linear order, in contrast to hierarchy-sensitive processing in anterior regions. Our results are not consistent with a naïve interpretation of this proposal: we do not see robust evidence for string-level Markov models in dorsal LIPL or LPreM, and we do see robust evidence for hierarchical effects in the posterior LPTL region. However, the notion of linear processing implied by that model, such as the mapping of the ordering of noun phrases to thematic roles like "agent" and "patient", may be of a very different sort than the notion of string-based linear order encoded in an *n*-gram model.

The results from our analysis leave open several questions about the syntactic operations implemented in these brain regions. One has to do with the degree to which human parsing is "predictive" of upcoming words and phrases that have yet to be heard. Our complexity metrics distinguish non-predictive bottom-up node counts from highly predictive top-down node counts, and yet these metrics are highly correlated after being projected through the hemodynamic response function, especially for MG structures (see Supplementary Fig. S2). This could reflect temporal limitations of fMRI. Bottom-up and top-down enumeration differ not in how much structure is built, but in the dynamics of when structure-building takes place. Given the rapid unfolding of spoken language, such differences may not have been sufficiently spread out to lead to detectable effects. Future work using electrophysiological tools may be more capable of teasing out these effects.

Details about the grammar implemented in these circuits are also underspecified. We only contrasted a single MG with CFG and Markov models. A very large variety of grammatical analyses that can be described by MGs continue to be explored in theoretical linguistics (Stabler, 2011). Further, MGs are but one of a class of grammars that are suitable for describing human language (Stabler, 2013a). Our approach does not distinguish which particular analysis from this class of grammars best matches the measured brain signals.

Finally, our interpretation has been couched in terms of syntactic structure. This framing reflects the available computational models of incremental parsing. Interestingly, the anterior temporal lobes have also been linked with conceptual semantic processing. One piece of evidence for such a link is the correlation between anterior temporal atrophy and deficits in conceptual processing associated with Semantic Dementia (Patterson, Nestor, & Rogers, 2007). Related research has found that activation in these regions is sensitive to uniqueness and to the conceptual specificity with which stimuli are categorized (Gorno-Tempini & Price, 2001; Grabowski et al., 2001; Rogers et al., 2006). These findings have led to the hypothesis that ATL activity may reflect, in part, the specificity of a semantic representation being processed (Martin & Chao, 2001; Patterson et al., 2007). These conceptual and syntactic functions may be related: more complex phrases could be used to describe more specific concepts. In fact, recent work using MEG has found that LATL sensitivity to phrase-structure is modulated by conceptual specificity (Westerlund & Pylkkänen, 2014; Zhang & Pylkkänen, 2015).

While intuitive, however, such a link has not yet been formalized in a way that yields quantitative predictions. What we observe is that incremental syntactic parsing models appear to provide a good fit, quite apart from considerations of meaning. To tease apart possible connections between syntactic and conceptual composition, we await the deployment of more sophisticated quantifiable accounts of semantic composition. Those that rely on vector-semantics provide one promising avenue for research (Chang, Cherkassky, Mitchell, & Just, 2009; Mitchell & Lapata, 2008; Mitchell et al., 2008). Another avenue might draw on algorithms describing the incremental evaluation of logical semantic rules (Stabler, 1991; Steedman, 2000).

5. Conclusion

This study winnows down the type of information that flows through brain regions involved in syntactic processing. We asked

[4] Hale (2014) develops the "automaton view" of parser states as relating quite directly to syntactic structure.

whether abstract grammatical structures characterize fMRI-measured neural activity associated with sentence processing during a passive listening task. These more abstract structures indeed correlated with brain activity in the temporal lobes, but not in inferior frontal gyrus, inferior parietal lobe, or premotor areas. By contrast, predictors based on string-level Markov models correlated with brain activity frontal and temporal regions. In the most general terms, abstract linguistic structure of the sort proposed in generative grammars appears to characterize the information flowing through the temporal lobe during naturalistic comprehension.

Acknowledgements

We acknowledge with appreciation the help of David Lutz, Frederick Callaway, Elana Feldman, Jaclyn Jeffrey-Wilensky, and Adam Mahar. We thank Emily Qualls for assistance with data collection. With sadness we acknowledge the loss of our co-author, Sarah Van Wagenen, in 2014. This research was supported by NSF CAREER award #0741666 and NIH grant S10RR025145.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.bandl.2016.04.008.

References

Allauzen, C., Riley, M., Schalkwyk, J., Skut, W., & Mohri, M. (2007). OpenFst: A general and efficient weighted finite-state transducer library. In *Proceedings of the 12th international conference on implementation and application of automata* (pp. 11–23). Springer.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchinson, K. I., Kessler, B., Loftis, B., ... Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39*, 445–459.

Beckman, M. E., Hirschberg, J., & Shattuck-Hufnagel, S. (2005). The original ToBI system and the evolution of the ToBI framework. In S.-A. Jun (Ed.), *Prosodic typology: The phonology of intonation and phrasing*. Oxford University Press.

Bemis, D. K., & Pylkkänen, L. (2011). Simple composition: A magnetoencephalography investigation into the comprehension of minimal linguistic phrases. *The Journal of Neuroscience, 31*, 2801–2814.

Bemis, D. K., & Pylkkänen, L. (2013). Basic linguistic composition recruits the left anterior temporal lobe and left angular gyrus during both listening and reading. *Cerebral Cortex, 23*, 1859–1873.

Ben-Shachar, M., Hendler, T., Kahn, I., Ben-Bashat, D., & Grodzinsky, Y. (2003). The neural reality of syntactic transformations: Evidence from fMRI. *Psychological Science, 14*, 433–440.

Ben-Shachar, M., Palti, D., & Grodzinsky, Y. (2004). Neural correlates of syntactic movement: Converging evidence from two fMRI experiments. *NeuroImage, 21*, 1320–1336.

Berwick, R. C., & Weinberg, A. S. (1983). The role of grammars in models of language use. *Cognition, 13*, 1–61.

Bever, T. G. (1970). The cognitive basis for linguistic structures. In J. Hayes (Ed.), *Cognition and the development of language* (pp. 279–362). New York: Wiley.

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2009). The role of prominence information in the real-time comprehension of transitive constructions: A cross-linguistic approach. *Language and Linguistics Compass, 3*, 19–58.

Bornkessel-Schlesewsky, I., Schlesewsky, M., Small, S. L., & Rauschecker, J. P. (2015). Neurobiological roots of language in primate audition: Common computational properties. *Trends in Cognitive Sciences, 19*, 142–150.

Bornkessel-Schlesewsky, I., & Schlewswsky, M. (2009). *Processing syntax and morphology: A neurocognitive perspective*. Oxford University Press.

Bornkessel, I., Zysset, S., Friederici, A. D., von Cramon, D. Y., & Schlesewsky, M. (2005). Who did what to whom? The neural basis of argument hierarchies during language comprehension. *NeuroImage, 26*, 221–233.

Boston, M. F., Hale, J., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research, 2*, 1–12.

Boston, M. F., Hale, J., Vasishth, S., & Kliegl, R. (2011). Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes, 26*, 301–349.

Boynton, G. M., Engel, S. A., Glover, G. H., & Heeger, D. J. (1996). Linear systems analysis of functional magnetic resonance imaging in human V1. *Journal of Neuroscience, 16*, 4207–4221.

Brennan, J., Nir, Y., Hasson, U., Malach, R., Heeger, D. J., & Pylkkäen, L. (2012). Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and Language, 120*, 163–173.

Brennan, J., & Pylkkänen, L. (2012). The time-course and spatial distribution of brain activity associated with sentence processing. *NeuroImage, 60*, 1139–1148.

Bresnan, J., & Kaplan, R. M. (1982). Introduction: Grammars as mental representations of language. In J. Bresnan (Ed.), *The mental representation of grammatical relations*. Cambridge, MA: MIT Press.

Caplan, D., Chen, E., & Waters, G. (2008). Task-dependent and task-independent neurovascular responses to syntactic processing. *Cortex, 44*, 257–275.

Caramazza, A., & Zurif, E. B. (1976). Dissociation of algorithmic and heuristic processes in language comprehension: Evidence from aphasia. *Brain and Language, 3*, 572–582.

Chang, K., Cherkassky, V., Mitchell, T., & Just, M. (2009). Quantitative modeling of the neural representation of adjective-noun phrases to account for fMRI activation. In *Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP: Volume 2* (pp. 638–646). Association for Computational Linguistics.

Charniak, E., Goldwater, S., & Johnson, M. (1998). Edge-based best-first chart parsing. In *Proceedings of the Sixth Workshop on Very Large Corpora at COLING-ACL*.

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition, 109*, 193–210.

Dronkers, N. F., Wilkins, D. P., Van Valin, R. D., Redfern, B. B., & Jaeger, J. J. (2004). Lesion analysis of the brain areas involved in language comprehension: Towards a new functional anatomy of language. *Cognition, 92*, 145–177.

Fedorenko, E., Hsieh, P.-J., Nieto Castanon, A., Whitfield-Gabrieli, S., & Kanwisher, N. (2010). A new method for fMRI investigations of language: Defining ROIs functionally in individual subjects. *Journal of Neurophysiology, 104*, 1177–1194.

Ferreira, F., Bailey, K. G. D., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science, 11*, 11–15.

Ferreira, F., & Patson, N. (2007). The 'good enough' approach to language comprehension. *Language and Linguistics Compass, 1*, 71–83.

Ferstl, E., Neumann, J., Bogler, C., & Yves von Cramon, D. (2008). The extended language network: A meta-analysis of neuroimaging studies on text comprehension. *Human Brain Mapping, 29*, 581–593.

Fiebach, C. J., Friederici, A. D., Muller, K., & von Cramon, D. Y. (2002). fMRI evidence for dual routes to the mental lexicon in visual word recognition. *Journal of Cognitive Neuroscience, 14*, 11–23.

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS ONE, 8*, e77661.

Fossum, V., & Levy, R. (2012). Sequential vs. hierarchical syntactic models of human incremental sentence processing. In *Proceedings of the 3rd annual workshop on cognitive modeling and computational linguistics* (pp. 61–69). Association for Computational Linguistics.

Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science, 22*, 829–834.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language, 140*, 1–11.

Frazier, L. (1985). Syntactic complexity. In D. R. Dowty, L. Karttunen, & A. M. Zwicky (Eds.), *Natural language parsing: Psychological, computational, and theoretical perspectives* (pp. 129–189). Cambridge University Press.

Friederici, A. D., & Gierhan, S. M. (2013). The language network. *Current Opinion in Neurobiology, 23*, 250–254.

Friederici, A. D., Opitz, B., & von Cramon, D. Y. (2000). Segregating semantic and syntactic aspects of processing in the human brain: An fMRI investigation of different word types. *Cerebral Cortex, 10*, 698–705.

Friston, K. J., Ashburner, J., Kiebel, S. J., Nichols, T. E., & Penny, W. D. (Eds.). (2007). *Statistical parametric mapping: The analysis of functional brain images*. Academic Press.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Giannakidou, A. (1998). *Polarity sensitivity as (non)veridical dependency*. John Benjamins.

Gorno-Tempini, M. L., & Price, C. J. (2001). Identification of famous faces and buildings: A functional neuroimaging study of semantically unique items. *Brain, 124*, 2087–2097.

Grabowski, T. J., Damasio, H., Tranel, D., Ponto, L. L., Hichwa, R. D., & Damasio, A. R. (2001). A role for left temporal pole in the retrieval of words for unique entities. *Human Brain Mapping, 13*, 199–212.

Grodzinsky, Y. (2000). The neurology of syntax: Language use without Broca's area. *Behavioral and Brain Sciences, 23*, 1–21.

Grodzinsky, Y., & Friederici, A. D. (2006). Neuroimaging of syntax and syntactic processing. *Current Opinion in Neurobiology, 16*, 240–246.

Haegeman, L. (1999). X-bar theory. In *The MIT encyclopedia of the cognitive sciences*. Cambridge, Mass: MIT Press.

Hagoort, P. (2013). MUC (memory, unification, control) and beyond. *Frontiers in Psychology, 4*.

Hagoort, P., & Indefrey, P. (2014). The neurobiology of language beyond single words. *Annual Review of Neuroscience, 37*, 347–362.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American chapter of the association for computational linguistics*. Association for Computational Linguistics.

Hale, J. (2003). *Grammar, uncertainty and sentence processing*. Ph.D. thesis Johns Hopkins University Baltimore, Maryland.

Hale, J. (2014). *Automaton theories of human sentence comprehension*. CSLI Publications.

Hawkins, J. (1994). *A performance theory of order and constituency*. Cambridge University Press.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience, 8*, 393–402.

Humphries, C., Binder, J. R., Medler, D. A., & Liebenthal, E. (2006). Syntactic and semantic modulation of neural activity during auditory sentence comprehension. *Journal of Cognitive Neuroscience, 18*, 665–679.

Humphries, C., Love, T., Swinney, D., & Hickok, G. (2005). Response of anterior temporal cortex to syntactic and prosodic manipulations during sentence processing. *Human Brain Mapping, 26*, 128–138.

Hunter, T., & Dyer, C. (2013). Distributions on Minimalist Grammar derivations. In *Proceedings of the 13th meeting on the mathematics of language* (pp. 1–11). Association for Computational Linguistics.

Jobard, G., Vigneau, M., Mazoyer, B., & Tzourio-Mazoyer, N. (2007). Impact of modality and linguistic complexity during reading and listening tasks. *NeuroImage, 34*, 784–800.

Joshi, A. K., Shanker, K. V., & Weir, D. (1990). The convergence of mildly context-sensitive grammar formalisms. Technical Report MS-CIS-90-01 University of Pennsylvania Department of Computer and Information Science.

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing. Computational linguistics and speech recognition* (2nd ed.). Prentice-Hall.

Just, M. A., Carpenter, P. A., Keller, T. A., Eddy, W. F., & Thulborn, K. R. (1996). Brain activation modulated by sentence comprehension. *Science, 274*, 114–116.

Just, M., & Varma, S. (2007). The organization of thinking: What functional brain imaging reveals about the neuroarchitecture of complex cognition. *Cognitive, Affective, & Behavioral Neuroscience, 7*, 153–191.

Kaplan, R. M., & Bresnan, J. (1982). Lexical functional grammar: A formal system for grammatical representation. In J. Bresnan (Ed.), *The mental representation of grammatical relations* (pp. 173–281). Cambridge MA: MIT Press.

Klein, D., & Manning, C. D. (2003a). A* parsing: Fast exact viterbi parse selection. In *Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics*. Association for Computational Linguistics.

Klein, D., & Manning, C. D. (2003b). Accurate unlexicalized parsing. In *Proceedings of the 41st annual meeting of the association for computational linguistics* (pp. 423–430). Association for Computational Linguistics.

Kush, D., Lidz, J., & Phillips, C. (2015). Relation-sensitive retrieval: Evidence from bound variable pronouns. *Journal of Memory and Language, 82*, 18–40.

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

Luong, M.-T., Frank, M. C., & Johnson, M. (2013). Parsing entire discourses as very long strings: Capturing topic continuity in grounded language learning. *Transactions of the Association for Computational Linguistics, 1*, 315–323.

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics, 19*, 313–330.

Martin, A., & Chao, L. L. (2001). Semantic memory and the brain: Structure and processes. *Current Opinion in Neurobiology, 11*, 194–201.

Miller, G., & Chomsky, N. (1963). Finitairy models of language users. *Handbook of Mathematical Psychology*, 419–491.

Mitchell, J., & Lapata, M. (2008). Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT* (pp. 236–244).

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., ... Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science, 320*, 1191–1195.

Müller, S. (2015). *Grammatical theory: From transformational grammar to constraint-based approaches*. Language Science Press.

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia, 9*, 97–113.

Pallier, C., Devauchelle, A.-D., & Dehaene, S. (2011). Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences, 108*, 2522–2527.

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience, 8*, 976–987.

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

Poeppel, D. (2012). The maps problem and the mapping problem: Two challenges for a cognitive neuroscience of speech and language. *Cognitive Neuropsychology, 29*, 34–55.

Rademacher, J., Galaburda, A. M., Kennedy, D. N., Filipek, P. A., & Caviness, V. S. Jr., (1992). Human cerebral cortex: Localization, parcellation, and morphometry with magnetic resonance imaging. *Journal of Cognitive Neuroscience, 4*, 352–374.

Roark, B., Bachrach, A., Cardenas, C., & Pallier, C. (2009). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 324–333).

Rogalsky, C., & Hickok, G. (2009). Selective attention to semantic and syntactic features modulates sentence processing networks in anterior temporal cortex. *Cerebral Cortex, 19*, 786–796.

Rogalsky, C., & Hickok, G. (2010). The role of broca's area in sentence comprehension. *Journal of Cognitive Neuroscience, 23*, 1–17.

Rogers, T. T., Hocking, J., Noppeney, U., Mechelli, A., Gorno-Tempini, M. L., Patterson, K., ... Price, C. J. (2006). Anterior temporal cortex and semantic memory: Reconciling findings from neuropsychology and functional imaging. *Cognitive, Affective, & Behavioral Neuroscience, 6*, 201–213.

Sanford, A., & Sturt, P. (2002). Depth of processing in language comprehension: Not noticing the evidence. *Trends in Cognitive Sciences, 6*, 382–386.

Santi, A., & Grodzinsky, Y. (2007a). Taxing working memory with syntax: Bihemispheric modulations. *Human Brain Mapping, 28*, 1089–1097.

Santi, A., & Grodzinsky, Y. (2007b). Working memory and syntax interact in Broca's area. *NeuroImage, 37*, 8–17.

Santi, A., & Grodzinsky, Y. (2010). fMRI adaptation dissociates syntactic complexity dimensions. *NeuroImage, 51*, 1285–1293.

Shetreet, E., Friedmann, N., & Hadar, U. (2009). An fMRI study of syntactic layers: Sentential and lexical aspects of embedding. *NeuroImage, 48*, 707–716.

Shieber, S. M. (1985). Evidence against the context-freeness of natural language. *Linguistics and Philosophy, 8*, 333–343.

Snijders, T. M., Vosse, T., Kempen, G., Van Berkum, J. J., Petersson, K. M., & Hagoort, P. (2009). Retrieval and unification of syntactic structure in sentence comprehension: An fMRI study using word-category ambiguity. *Cerebral Cortex, 19*, 1493–1503.

Sportiche, D., Koopman, H., & Stabler, E. P. (2013). *An introduction to syntactic analysis and theory*. Wiley-Blackwell.

Stabler, E. P. (1991). Avoid the pedestrian's paradox. In R. C. Berwick, S. P. Abney, & C. Tenny (Eds.), *Principle-based parsing: Computation and psycholinguistics studies in linguistics and philosophy* (pp. 199–237). Dordrecht: Kluwer.

Stabler, E. P. (1997). Derivational minimalism. In C. Retoré (Ed.), *Logical aspects of computational linguistics* (pp. 68–95). Springer.

Stabler, E. P. (2011). Computational perspectives on minimalism. In C. Boeckx (Ed.), *The Oxford handbook of linguistic minimalism* (pp. 616–641). Oxford University Press.

Stabler, E. P. (2013a). The epicenter of linguistic behavior. In M. Sanz, I. Laka, & M. K. Tanenhaus (Eds.), *Language down the garden path: The cognitive and biological basis of linguistic structures* (pp. 316–323). Oxford University Press.

Stabler, E. P. (2013b). Two models of minimalist, incremental syntactic analysis. *Topics in Cognitive Science, 5*, 611–633.

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

Stolcke, A. (1995). An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Computational Linguistics, 21*, 165–201.

Stowe, L. A., Broere, C. A. J., Paans, A. M. J., Wijers, A. A., Mulder, G., Vaalburg, W., ... Zwarts, F. (1998). Localizing components of a complex task: Sentence processing and working memory. *NeuroReport, 9*, 2995–2999.

Stromswold, K., Caplan, D., Alpert, N., & Rauch, S. (1996). Localization of syntactic comprehension by positron emission tomography. *Brain and Language, 52*, 452–473.

Sturt, P., & Lombardo, V. (2005). Processing coordinated structures: Incrementality and connectedness. *Cognitive Science: A Multidisciplinary Journal, 29*, 291–305.

Turken, A. U., & Dronkers, N. F. (2011). The neural architecture of the language comprehension network: Converging evidence from lesion and connectivity analyses. *Frontiers in Systems Neuroscience, 5*, 1.

Vandenberghe, R. R., Nobre, A. C., & Price, C. J. (2002). The response of left temporal cortex to sentences. *Journal of Cognitive Neuroscience, 14*, 550–560.

van Schijndel, M., & Schuler, W. (2015). Hierarchic syntax improves reading time prediction. In *Proceedings of NAACL 2015*. Association for Computational Linguistics.

VanWagenen, S., Brennan, J., & Stabler, E. P. (2014). Quantifying parsing complexity as a function of grammar complexity. In C. T. Shütze, & L. Stockall (Eds.), *Connectedness: Papers by and for Sarah VanWagenen* (Vol. 18, pp. 31–47). UCLA Working Papers in Linguistics.

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

Vosse, T., & Kempen, G. (2000). Syntactic structure assembly in human parsing: A computational model based on competitive inhibition and a lexicalist grammar. *Cognition, 75*, 105–143.

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., & Mitchell, T. (2014). Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PLoS ONE, 9*, e112575.

Westerlund, M., Kastner, I., Al Kaabi, M., & Pylkkänen, L. (2015). The LATL as locus of composition: MEG evidence from English and Arabic. *Brain and Language, 141*, 124–134.

Westerlund, M., & Pylkkänen, L. (2014). The role of the left anterior temporal lobe in semantic composition vs. semantic memory. *Neuropsychologia, 57*, 59–70.

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & van den Bosch, A. (2015). Prediction during natural language comprehension. *Cerebral Cortex*, 1–11.

Wilson, S. M., DeMarco, A. T., Henry, M. L., Gesierich, B., Babiak, M., Mandelli, M. L., ... Gorno-Tempini, M. L. (2014). What role does the anterior temporal lobe play in sentence-level processing? Neural correlates of syntactic processing in semantic variant primary progressive aphasia. *Journal of Cognitive Neuroscience, 26*, 970–985.

Wilson, S. M., Galantucci, S., Tartaglia, M. C., Rising, K., Patterson, D. K., Henry, M. L., ... Gorno-Tempini, M. L. (2011). Syntactic processing depends on dorsal language tracts. *Neuron, 72*, 397–403.

Worsley, K. J., Marrett, S., Neelin, P., Vandal, A. C., Friston, K. J., & Evans, A. C. (1996). A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping, 4*, 58–73.

Xiang, M., Dillon, B., & Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *Brain and Language, 108*, 40–55.

Xu, J., Kemeny, S., Park, G., Frattali, C., & Braun, A. (2005). Language in context: Emergent features of word, sentence, and narrative comprehension. *NeuroImage, 25*, 1002–1015.

Yngve, V. H. (1960). A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society, 104*, 444–466.

Yoshida, M., Dickey, M. W., & Sturt, P. (2012). Predictive processing of syntactic structure: Sluicing and ellipsis in real-time sentence processing. *Language and Cognitive Processes, 28*, 272–302.

Yun, J., Chen, Z., Hunter, T., Whitman, J., & Hale, J. (2015). Uncertainty in processing relative clauses across east asian languages. *Journal of East Asian Linguistics, 24*, 113–148.

Zaccarella, E., & Friederici, A. D. (2015). Merge in the human brain: A sub-region based functional investigation in the left pars opercularis. *Frontiers in Psychology, 6*, 1818.

Zhang, L., & Pylkkänen, L. (2015). The interplay of composition and concept specificity in the left anterior temporal lobe: An MEG study. *NeuroImage, 111*, 228–240.

Zurif, E. (1995). Brain regions of relevance to syntactic processing. In L. R. Gleitman & M. Liberman (Eds.), *An invitation to cognitive science* (pp. 381–398). Cambridge, MA: MIT Press.