Grammatical Predictors for fMRI Timecourses

Jixing Li, John Hale

Department of Linguistics, Cornell University

1 Introduction

There is widespread agreement that a network of brain regions surrounding the Sylvian fissure supports human language comprehension (Stowe et al. 2005, Dronkers et al. 2004, Pallier et al. 2011). Less clear is what the individual anatomical sites of this network actually do. Towards a more precise functional anatomy of language comprehension, we correlated time-series predictions from a variety of grammatical predictors with fMRI data from several well-known brain regions. The results categorize the types of the language-processing these areas carry out. In particular, they confirm a statistically-significant role for a predictor based on Minimalist Grammars (Chomsky 1995, Stabler 1997).

We model blood-oxygen-level dependent (BOLD) signals from fMRI that are time-locked to each spoken word in an audiobook. The case for such naturalistic stimuli in neuroscience has been made by Hasson and Honey (2012), who argued that findings within a controlled laboratory setup may not be ecologically valid in real-life contexts. We analyzed the freely-available region of interest (ROI) timecourses from Brennan et al. (2016) with two additional regressors not considered before: a memory-based metric 'structural distance' and a distributional-semantic metric indicating 'conceptual combination'. We found that even with these covariates, the

predictor based on Minimalist Grammars still significantly improved a regression model of the BOLD signal in the posterior temporal region, roughly corresponding to Wernicke's area.

Our methodology follows Brennan et al. (2016), which itself responds to Sprouse and Hornstein's (2016) exhortation to collaboratively construct a cognitive neuroscience of syntactic structure-building: First identifying the structure-building computations word-by-word, then asking whether there is evidence for those computations in neural signals. This approach allows an investigator to examine cognitive hypotheses about the role of grammar in processing (for foundational discussion of this point, see Stabler, 1983).

The remainder of the chapter is organized into four sections: Section 2 lays out some assumptions about grammar, parsing and processing complexity in our neuro-computational models; Section 3 reviews known effects of word-to-word associations and lexical-semantic coherence in human sentence processing; Section 4 details the calculation of the complexity metrics that are used as predictors of fMRI timecourses in this work. This section presents the data, statistical models and results. Section 5 discusses the implications of these results for the cognitive science of language more broadly.

2 Parameters in neuro-computational models of sentence processing

Under Brennan's (2016) formulation, a neuro-computational model involves an incremental parser as well as some kind of linking hypothesis that connects the states visited by that parser to potentially-observable neural signals. Table 1 identifies the particular combinations that we consider in this work; the fourth column is the linking hypothesis.

<insert Table 1 here>

Brennan further subdivides the parser into a grammar G, a parsing algorithm A and an oracle O for resolving the inevitable nondeterminism that attends human language processing. The first and second columns of Table 1 specify G and A respectively. All grammatical models assume a perfect oracle O, and the response function is always the default hemodynamic response function provided by the SPM software package (see e.g., Henson and Friston 2007). The following sections demonstrate in more detail how the parameters in Table 1 influence the predicted hemodynamic responses of word-to-word processing difficulty.

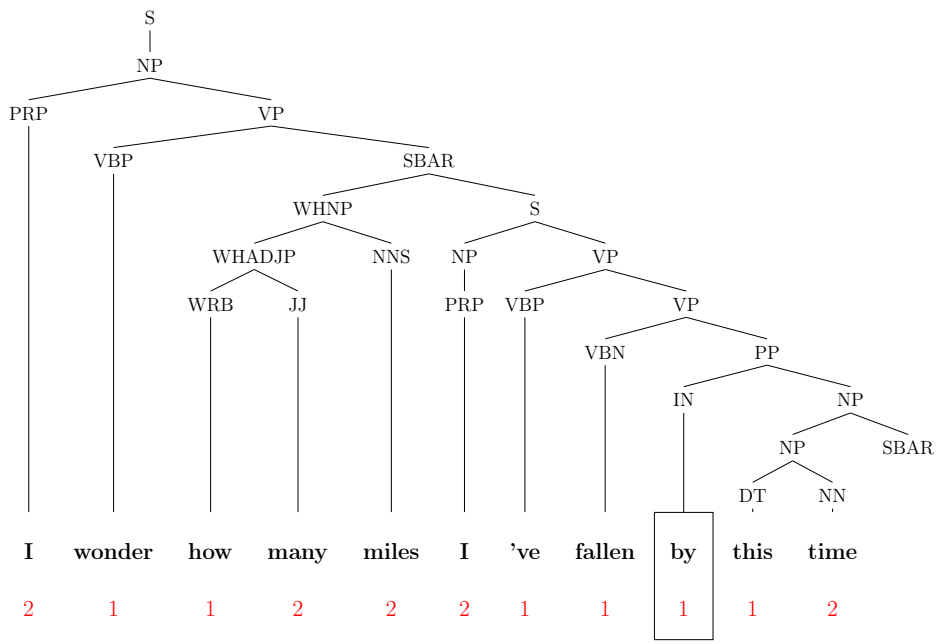
2.1 Grammar

Constituency grammars

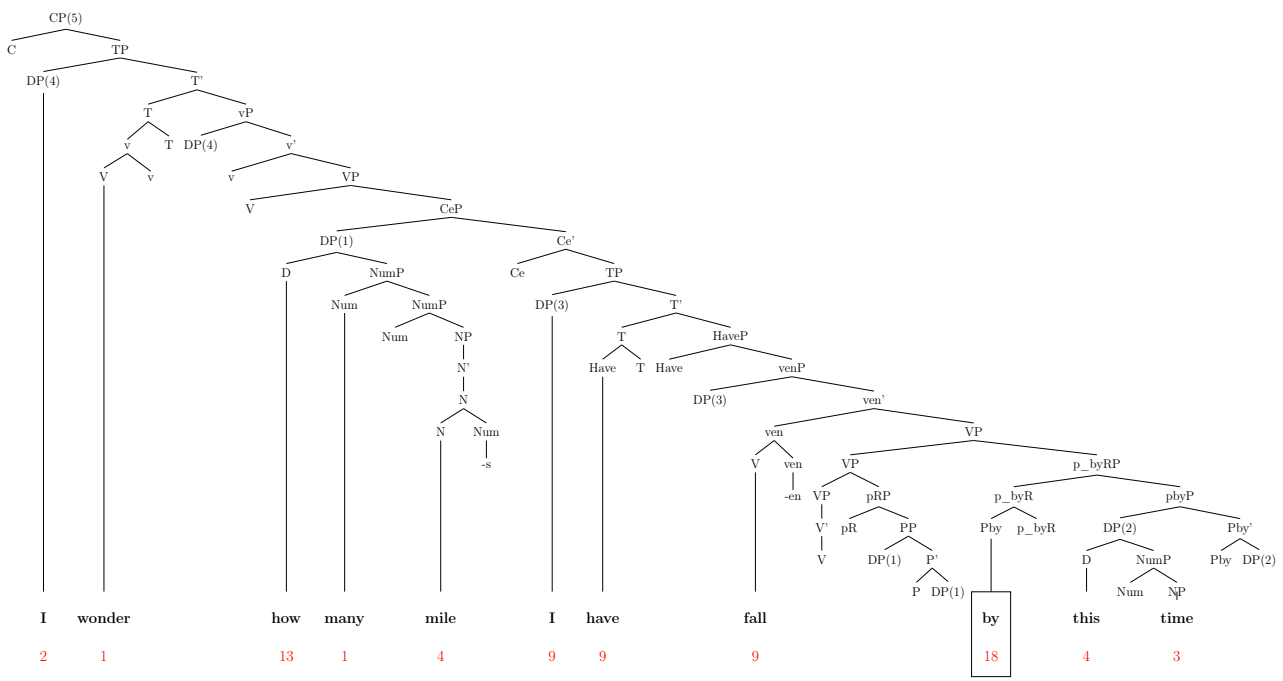
We compared models based on Minimalist Grammars (MG) to simpler Context-Free Grammars (CFG). The particular CFGs that we consider lack empty categories and have no explicit representation of headedness. As Fong and Berwick (2008) point out, CFGs list related constructions separately. This can obscure linguistically-significant generalizations, for instance about argument structure. MGs are a more expressive formalism, inspired by Chomsky's Minimalist Program (for a review see Stabler 2011). Integrating constituency, dependency and movement information, MG derivations can be viewed as X-bar structures (see e.g., Haegeman 1999). MGs make it convenient to express a variety of well-known analyses, for instance of ditransitives (Larson 1988), relative clauses (Kayne 1994), passives (Baker et al. 1989), head movement, genitives, raising, ECM, control, quantifiers, and Wh-movement (Sportiche et al. 2013).

Much recent work in computational psycholinguistics has applied CFGs in modeling human processing difficulty (see e.g. Demberg and Keller 2008, Roark et al. 2009). Yet, as shown in van Wagenen et al. (2014), the choice of grammar formalism can have a major impact on processing difficulty predictions. Tree (1) and (2) illustrate this difference. They contrast bottom-up node counts based on CFG and MG for the same sentence from Alice in Wonderland. At the word *by* the prediction would be 1 under a naive CFG analysis, and 18 under a richer MG analysis.

(1)



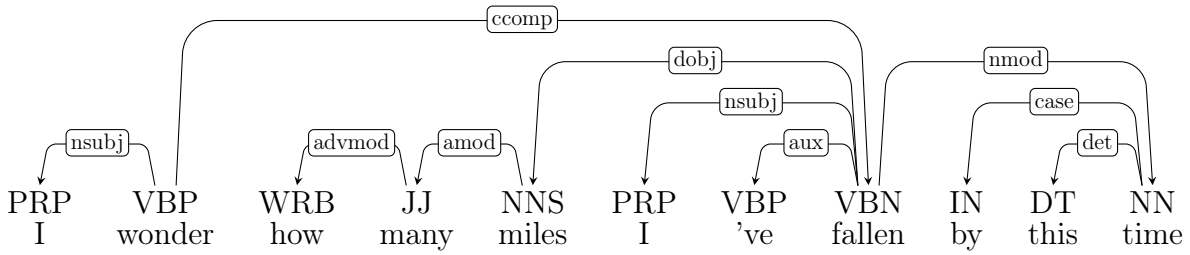
(2)



Dependency grammars

Apart from the constituency-based hypotheses, we also defined a 'structural distance' metric that reflects aspects of both constituency and dependency. This predictor is inspired by earlier work in which linguistic dependency relations correspond to memory retrieval actions that themselves carry a processing cost (Wanner and Maratsos 1978, Gibson 1998, Lewis and Vasishth 2005). Diagram (3) shows dependency relations, as recovered by the Stanford Parser (de Marneffe et al. 2006).

(3)

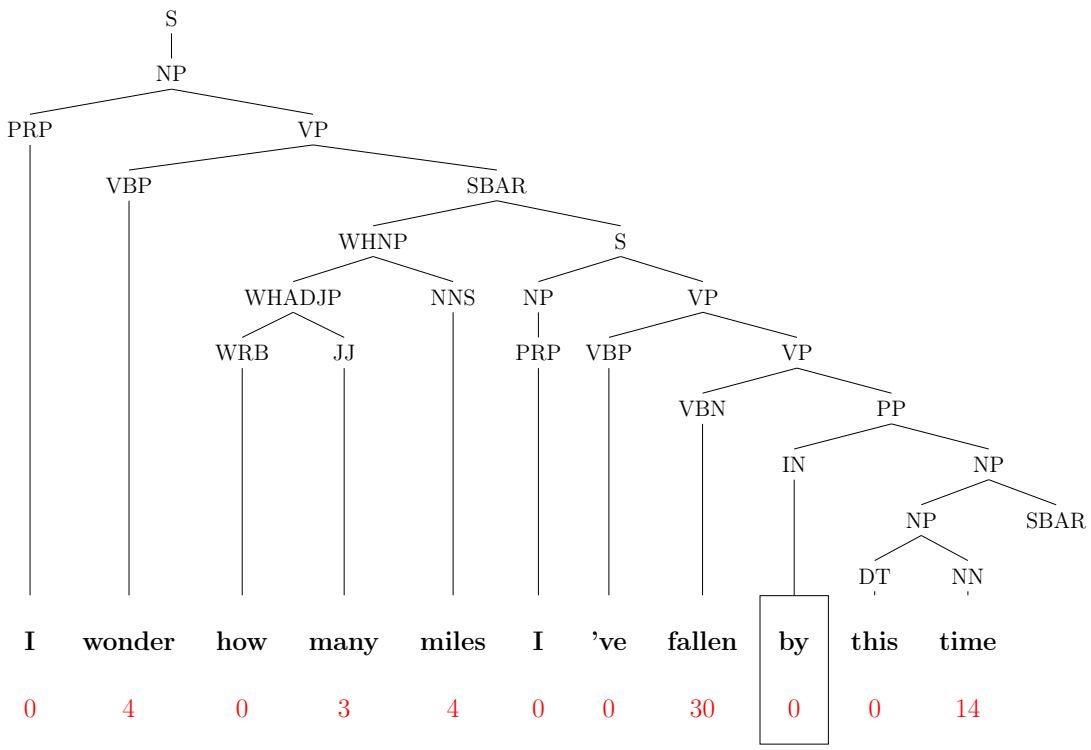


To fully specify a complexity metric, it is necessary to quantify how difficult the induced

memory retrievals are. A typical approach is to define some function of the distance between words that stand in a dependency relation. This distance could be quantified as the number of intervening discourse referents, the number of intervening words, i.e. 'linear distance', or the number of nodes crossed when traversing the syntactic tree structure from a dependent to a head, i.e., 'structural distance'. Structural distance and linear distance make contrasting predictions on relative clauses in head-final languages, and in fact only structural distance seems to derive the observed pattern (O'Grady 1997, Yun et al. 2010). Indeed, Baumann (2014) compared the three distance measures mentioned above and suggested that structural distance is the only significant predictor of reading times in an eye-tracking corpus.

Compared to node counts based on phrase structure, structural distance based on both dependency and constituency grammars predicts a distinctive set of word-by-word processing difficulty. Tree (4) illustrates the structural distance between dependent words in the same sentence in Tree (1) and (2). We considered only the rightmost word in any dependency relation. For words in multiple dependency relations, we summed the structural distances. For instance, the number of non-terminal nodes between *fallen* and *'ve* is 4 (VBP, VP, VP, VBN), between *fallen* and *I* is 6 (PRP, NP, S, VP, VP, VBN), between *fallen* and *I* is 6 (PRP, NP, S, VP, VP, VBN), between *fallen* and *miles* is 7 (NNS, WHNP, SBAR, S, VP, VP, VBN), between *fallen* and *wonder* is 7 (VBP, VP, SBAR, S, VP, VP, VBN), so the summed structural distance at *fallen* is 30. We can see that under the 'structural distance' metric, the words *fallen* and *time*, which participate in multiple dependencies, are predicted to induce the most processing effort. This is very different from the bottom-up node count metric based on Minimalist Grammar, which predicts *how* and *by* to be the most difficult words to process (see Tree (3)).

(4)



2.2 Parsing strategy

Different parsing strategies lead to different predictions about processing effort on a particular word. A top-down parser starts from a mother node and makes decisions about phrase structure before checking them against the input string. A bottom-up parser starts with the first terminal word and has to check all the evidence before applying a phrase structure rule. A left-corner parser combines both top-down and bottom-up directions, and it applies a grammatical rule after seeing the very first symbol on the right-hand side of the rule (see e.g., Hale 2014, Chap.3). Table 2 illustrates the steps of the three parsing strategies for the sentence *John loves Mary*. The numbers in red indicate the node count based on the three parsing strategies respectively.

<insert Table 2 here>

2.3 Complexity metrics

Node count

Node count is the number of parsing steps between successive words under a parsing strategy. This is related to some forms of Yngve's (1960) Depth hypothesis (see also Frazier 1985). As shown in Table 2, even for the simple sentence *John loves Mary*, the node counts based on the top-down, bottom-up and left-corner parsing can be different. We calculated the CFG- and MG-based node counts for first chapter of Alice in Wonderland using the three parsing strategies. The results show very different counts for each parsing strategy based on CFG grammars, as shown in the correlation coefficients between `cfg.bu, cfg.td` and `cfg.lc` in Figure 2.

Surprisal

The metric 'surprisal' is motivated by information theory (see Hale 2016, Armeni et al. 2017, for a review). It quantifies the transition probability from an initial substring to the next word. High surprisal simply means that the next word is improbable on some particular language model.

Surprisals from certain probabilistic grammars predict a 'subject preference' for Chinese relative clauses (Jäger et al. 2015). This prediction diverges from an account based on the Dependency Locality Theory (Hsiao and Gibson 2003). Such controversy makes it more interesting to compare the regression results of the surprisal predictor and the DLT-like predictor structural distance against the fMRI data.

2.4 Summary

We formalized syntactic processing during naturalistic comprehension using several different neuro-computational models in an effort to discern their neural bases, if any. The CFG models include both the node count and surprisal metrics. The MG models include only the node count metrics. Surprisal, though well-defined for MGs (Hale 2003), is not available for the current study. The structural distance model is based on both the dependency grammar and the CFG; its parsing strategy is bottom-up, and its complexity metric is the sum of node counts between dependent words.

3 Other factors influencing sentence processing

Apart from the syntactic factors discussed in Section 2, other factors such as linear, word-to-word association and semantic information also influence processing complexity. We formalize linear order expectancy as trigram surprisal, and semantic information as cosine similarity between words and its previous context using distributional semantic models.

Word-to-word associations

Linear, word-to-word, surface dependencies, as reflected in N-gram models, have been shown to influence online sentence comprehension at least at some level of processing (but see Everaert et al. 2015). For instance, Ferreira and Patson (2007) report that syntactic structure is largely ignored when it conflicts with other information; Christiansen and MacDonald (2009) suggest that some ungrammatical structures can still be processed with ease. Frank and Bod (2011) compare phrase structure grammar models with sequential-structural models, i.e, Markov models and connectionist

models in predicting eye-fixation measures, and find better performance for the sequential-structural models. Similarly, Frank et al. (2015) find sequential-structural models fit the EEG amplitudes better than does a phrase-structure grammar.

This N-gram predictor, based on linear word-to-word relationships, contrasts with the 'structural distance' predictor that is based on hierarchical structural dependency. A number of behavioral studies have shown that violation of hierarchically-based rules leads to increased reading times (e.g., Sturt and Lombardo 2005, Yoshida et al. 2012, Kush et al. 2015), and expectations of word category based on hierarchical grammars predict eye-fixation times (e.g., Boston et al. 2008, 2011). Event-related potential (ERP) studies have also revealed early negativity for structurally unexpected stimuli (e.g., Xiang et al. 2009). It is therefore interesting to compare the effects of both the linear and hierarchical structural models in sentence processing.

Lexical-semantic coherence

Apart from grammatical information, word meaning also influences sentence processing. As suggested by Landauer (2007), very little of this word-order information may actually be used by human readers, perhaps only 10% to 15%. On the other hand, meaning is obviously involved otherwise communication would not be possible. In the famous example *He spread the warm bread with socks*, although *socks* is a well-expected grammatical category 'NP', it hinders comprehension and elicits a large N400 effect.

Following Firth (1957)'s distributional hypothesis, semantic coherence could be quantified by distributional semantic models, which represent words as high-dimensional vectors based on co-occurrence statistics from a large text corpus (e.g., Baroni et al. 2014, Erk 2012). Similar vectors are assigned to words that usually occur in similar contexts, hence the cosine similarity

between two vectors represents the semantic distance between the two words.

Behavioral studies suggest that cosine similarity between word vector and its previous context vector accounts for a certain amount of variance in eye-fixation times (Pynte et al. 2008) and word pronunciation duration (Sayeed et al. 2015). More recently, Ettinger et al. (2016) showed that cosine distance between the critical word and its context simulates the N400 effect from a previous ERP sentence-reading experiments (Federmeier and Kutas 1999).

4 Correlating fMRI timecourses with various metrics during natural story listening

4.1 Complexity metrics

CFG node counts

We first obtained the CFG trees using the Stanford Parser (Klein and Manning 2003), then we counted the number of nodes in the CFG trees that would be visited by a bottom-up, top-down and left-corner parser respectively.

MG node counts

Analogous to the node count predictors based on CFG trees, we also counted the number of nodes in the X-bar trees that would be visited by a bottom-up, top-down and left-corner parser respectively. The X-bar structures were the derived trees generated by Minimalist Grammars in the sense of Stabler (1997). These structures reflect grammatical analysis by van Wagenen et al. (2014).

CFG surprisal

CFG surprisal is a structural notion of expectedness of the next word as described in Section 2.3. We used the EarleyX implementation of Stolcke's probabilistic Earley parser to compute surprisal values (Luong et al. 2013, Stolcke 1995). The probabilities of grammatical rules were estimated using the entire Alice in Wonderland text, with punctuation removed.

Structural distance

To examine the memory-related complexity metrics sketched earlier in Section 2.3, the dependency relations for every sentence were also obtained using the Stanford Parser (de Marneffe et al. 2006). Structural distance is then the number of nodes traversed between the head and the dependent in the phrase structural tree. We considered only the rightmost word in any dependency relation. For words in multiple dependency relations, we summed the structural distances.

N-gram surprisal

As a kind of control (see Section 3), we used the freely-available trigram counts from the Google Books project (see e.g., Michel 2011) and restricted consideration to publication years 1850-1900, i.e., the year surrounding the publication of Alice in Wonderland. We backed off to lower-order grams where necessary: coverage was 1725/2045 for trigrams and 1640/1694 for bigrams. We then used surprisal of the trigram probabilities to link the probability of a word in its left-context to BOLD signals (see Hale 2001, 2016).

Lexical-semantic coherence

We used latent semantic analysis (LSA; Landauer and Dumais 1997) to build our semantic coherence metric. The training data comprised Alice in Wonderland in its entirety. We first built the type-by-document matrix where the rows are all the words in the book and the documents are all the paragraphs. The input vector space was transformed by singular value decomposition

(SVD), and truncated to a *100*-dimensional vector space. The context vector was the average of the previous 10 word vectors. We used negative cosine between the target word vector and the context vector to represent lexical-semantic coherence: higher negative cosine value indicates less semantic coherence.

4.2 Data acquisition

We used the freely available ROI timecourses from Brennan et al. (2016). The data come from twenty-five native English speaker (17 female, 18-24 years old, right-handed) listening to a story while in the scanner. The story was the first chapter of Alice in Wonderland, lasting for about 12.4 minutes. Participants completed twelve multiple-choice questions after scanning to verify their comprehension.

Four regions of interest (ROIs) were used to evaluate the syntactic models, including the left anterior temporal lobe (LATL), the right anterior temporal lobe (RATL), the left inferior frontal gyrus (LIFG) and the left posterior temporal lobe (LPTL). Figure 1 shows the ROIs from 4 participants (from Brennan et al. 2016).

Both functional and anatomic criteria guided the precise positioning of these ROIs. The functional criterion derives from an atheoretical word rate regressor (`rate`), which has value 1 at the offset of each word in the audio stimulus, and 0 elsewhere. This localizer identified regions whose BOLD signals were sensitive to word presentation. Each ROI sphere (10 mm radius) was centered on a peak t-value of at least 2.0 within the anatomical areas.

Imaging was performed using a 3T MRI scanner with a 32-channel head coil at the Cornell MRI facility; the detailed imaging parameters and preprocessing procedures are described in

Brennan et al. (2016).

<insert Figure 1 here>

4.3 Data analysis

Estimating hemodynamic response

Following Just and Varma (2007), we convolved each complexity metric's time series with SPM12's canonical hemodynamic response function (HRF) to get the estimate hemodynamic responses for each language model. This estimated response is what should be observed if a brain region were processing the information specified in each neuro-computational model. The time series are made orthogonal to the convolved `rate` vector, since it is our localizer for defining the ROIs.

Stepwise regression

We tested the unique contribution of each model by conducting stepwise model comparisons against the ROI timecourses. We used the forward selection approach which starts with no variables in the model, and tests whether adding one variable would give statistically significant improvement of the fit. Our null model included fixed effects for head movements (`dx, dy, dz, rx, ry, rz`) and `rate`; We also included fixed effects for word frequency (`freq`), which were also convolved with the same HRF. The frequency count was estimated using the SUBTLEXus corpus (Brysbaert and New 2009), which contains 51 million words from the subtitles of American films and television series. f0 (`f0`), and root mean square (RMS) intensity (`intensity`) of the speech were also included in our null model. The raw f0 of the 12 minutes speech in the audio were extracted using the fxrapt function from the Voicebox toolbox for Matlab; each contour was further processed by removing any f0 values exceeding 4 *s.d.* from the mean, and filling the gaps by spline interpolation. The RMS intensity for every 20 ms of the audio was calculated to get the

intensity vectors. `f0` and `intensity` were also convolved and then sub-sampled to 0.5 Hz to matching the time resolution (TR) of the fMRI measurements. The random effects included a random intercept by participant and a random slope for `rate`:

$$\text{BOLD}_{\text{null}} \sim \text{dx} + \text{dy} + \text{dz} + \text{rx} + \text{ry} + \text{rz} + \text{rate} + \text{f0} + \text{intensity} + \text{frequency}\,(1 + \text{rate}\,|\,\text{subject})$$

We then added regressors in the following order: surprisal of lexical trigram probability (`trigram`), negative cosine similarity between word vector and context vector (`lsa10`), structural distance between dependent words based on CFG (`struct`), surprisal of word-category probability based on CFG (`cfg.surp`), node count based on bottom-up parsing on CFG (`cfg.bu`) and node count based on bottom-up parsing on MG (`mg.bu`), namely, `trigram` > `lsa10` > `struct` > `cfg.surp` > `cfg.bu` > `mg.bu`. This order reflects a general direction from predictors with least syntactic information (`trigram, lsa10`) to predictors with the richest syntactic information (`mg.bu`).

Model fit was assessed using chi-square tests on the log-likelihood values to compare different models. All of the predictors were converted to z-scores before statistical analysis. Statistical significance was corrected for multiple comparisons across four ROIs with the Bonferroni method (the adjusted alpha-level is 0.05/4=0.0125).

4.4 Results

Correlations between predictors

As a first step, we checked the correlations between the 14 regressors: `rate`, `f0`, `intensity`, `freq`, `trigram`, `lsa10`, `struct`, `cfg.surp`, `cfg.bu`, `mg.bu`, `cfg.td`, `mg.td`, `cfg.lc` and `mg.lc`. The correlation matrix is shown in Figure 2. As can be seen, the correlation between `rate` and

`intensity` is relatively high ($r = 0.58, p < .001$); this is expected because word rate tracks the occurrence of each word in speech, which leads to higher intensity as compared to silence. `freq` is negatively correlated with `cfg.surp` (r(`freq`, `cfg.surp`)=-0.5, $p < .001$).

Among the CFG node count regressors, `cfg.td` and `cfg.lc` are highly correlated (r(`cfg.td`, `cfg.lc`)=0.93, $p < .001$), while the correlation coefficients between `cfg.bu` and `cfg.td` (r(`cfg.bu`, `cfg.td`)=0.93, $p < .001$) and between `cfg.bu` and `cfg.lc` (r(`cfg.bu`, `cfg.lc`)=0.59, $p < .001$) are much smaller.

The MG node counts based on different parsing algorithms are all highly correlated with each other (r(`mg.bu`, `mg.td`)=0.9, $p < .001$; r(`mg.bu`, `mg.lc`)=0.78, $p < .001$; r(`mg.td`, `mg.lc`)=0.87, $p < .001$). To avoid collinearity in hierarchical regression analysis, we include only `cfg.bu` and `mg.bu` for comparison between CFG and MG models.

<insert Figure 2 here>

Model comparison

Complexity metrics based on each of the neuro-computational models are subsequently added to the four baseline regressions. In the ATLs, an improvement in the goodness of fit is obtained for lexical-semantic coherence, but structural distance is also significant for the RATL. All the parameters are significant for the LPTL, roughly corresponding to the traditional 'Wernicke's area'. Trigram and CFG Surprisal significantly improve model fit in all of the ROIs, but no other regressors are significant in the LIFG except trigram and CFG surprisal. The statistical details for the model comparisons are shown in Table 3.

<insert Table 3 here>

Figure 3 shows the estimated coefficients and 95% confidence intervals for each of the

predictors when added to the null model. The signs of the coefficients suggest that: (1) higher trigram and CFG surprisal are correlated with more activation in all the ROIs; (2) changes in distributional meaning correlate with more activation in the ATLs and the LPTL; (3) greater structural distances between heads and dependents are associated with increased activation in the RATL and LIPL; (4) higher bottom-up node counts based on CFG and MG both correlate with increased activity in the LPTL.

<insert Figure 3 here>

5 Towards a functional anatomy of sentence comprehension

We found a significant effect for MG and CFG bottom-up node count in the LPTL on top of trigram probability. Contra suggestions in Frank and Bod (2011) and Frank et al. (2015), this is evidence for sensitivity to hierarchical structure, at least in the posterior temporal lobe. Node counts from MG-derived structures are also significant over and above node counts based on CFGs — suggesting that the LPTL is also involved in the processing of long-distance dependency.

Structural distance is significant in the RATL and the LPTL. This result aligns well with Baumann's (2014) finding based on eye-tracking corpus that structural distance predicts eye-fixation times in reading, supporting a role for 'integration cost' in the memory-based models of sentence processing.

The lexical-semantic coherence metric is a significant predictor in the ATLs. This is consistent with previous findings implicating the ATLs in conceptual combination (Rogalsky and Hickok 2009, Wilson et al. 2014, Pylkkänen 2015). However, we also found a significant effect of CFG surprisal

in the LATL, and a significant effect of structural distance in the RATL. This confirms the idea that the ATLs are involved in syntactic processing as well (Humphries et al. 2006, Brennan et al. 2012, 2016).

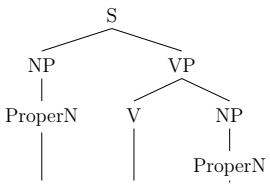
The LPTL activity is highly correlated with all the syntactic and semantic complexity metrics. As shown in Wehbe et al. (2014), multiple regions spanning the bilateral temporal cortices represent both syntax and semantics. Our results further confirms their suggestion that syntax and semantics might be non-dissociable at the level of neurobiology.

No semantic or syntactic metric is significantly correlated with the LIFG except for CFG surprisal. The failure to find effects of grammar has led to reevaluations of deficit-lesion studies that have long associated syntactic computation with the 'Broca's area' (e.g., Ben-Shachar et al. 2003, Caplan et al. 2008, Just et al. 1996, Stromswold et al. 1996). However, both this study and Brennan et al. (2016) found significant effects of n-gram models in the LIFG. It is possible that with easy-to-understand stimuli this region is not particularly strained.

To sum up, our correlational results from fMRI suggest that the temporal lobes perform a kind of computation that is both syntactic in the classical sense of phrase structure, and semantic in the sense of distributional word-embeddings. One set of questions this work leaves open is the precise relationships between these two predictors — for instance, temporal precedence. Other methods, such as MEG, may provide further insight here.

Model	Grammar	Parsing strategy	Complexity metric
`cfg.td`	Context-Free Grammar	top-down	node count
`cfg.bu`	Context-Free Grammar	bottom-up	node count
`cfg.lc`	Context-Free Grammar	left-corner	node count
`mg.td`	Minimalist Grammar	top-down	node count
`mg.bu`	Minimalist Grammar	bottom-up	node count
`mg.lc`	Minimalist Grammar	left-corner	node count
`cfg.surp`	Context-Free Grammar		surprisal
`struct`	Dependency Grammar Context-Free Grammar	bottom-up	nodes between dependencies

Table 1: Parameters in our neuro-computational models.



John	**loves**	**Mary**
3	2	2
2	1	4
3	2	2

Top-Down:		**Bottom-Up:**		**Left-Corner:**	
expand by	S → NP VP	shift	John	shift	John
expand by	NP → ProperN	reduce by	ProperN → John	project	ProperN → John
expand by	ProperN → John	reduce by	NP → ProperN	project	NP → ProperN
scan	John	shift	loves	Project + complete	S → NP VP
expand by	VP → V NP	reduce by	V → loves	shift	loves
expand by	V → loves	shift	Mary	project	V → loves
scan	loves	reduce by	ProperN → Mary	project + complete	VP → V NP
expand by	NP → ProperN	reduce by	NP → ProperN	shift	Mary
expand by	ProperN → Mary	reduce by	VP → V NP	project	ProperN → Mary
scan	Mary	reduce by	S → NP VP	project + complete	NP → ProperN

Table 2: Steps of top-down, bottom-up and left-corner parsing for the sentence *John loves Mary*. The numbers below the terminal nodes indicate the node count for that word based on the three parsing strategies respectively.

(a) LATL

	Parameter	df	LogLik	χ^2	p
∅		15	-11659		
A	`trigram`	16	-11622	72.9	**<.001**
B	`lsa10`	17	-11611	22.9	**<.001**
C	`struct`	18	-11611	0.7	0.41
D	`cfg.surp`	19	-11553	115.6	**<.001**
E	`cfg.bu`	20	-11553	0.0	0.83
F	`mg.bu`	20	-11551	3.2	0.07

(b) RATL

	Parameter	df	LogLik	χ^2	p
∅		15	-11222		
A	`trigram`	16	-11210	22.7	**<.001**
B	`lsa10`	17	-11202	16.3	**<.001**
C	`struct`	18	-11195	14.1	**<.001**
D	`cfg.surp`	19	-11176	37.8	**<.001**
E	`cfg.bu`	20	-11176	0.5	0.48
F	`mg.bu`	21	-11174	3.4	0.07

(c) LIFG

	Parameter	df	LogLik	χ^2	p
∅		15	-10653		
A	`trigram`	16	-10648	10.0	**0.002**
B	`lsa10`	17	-10646	3.0	0.086
C	`struct`	18	-10646	0.0	0.832
D	`cfg.surp`	19	-10633	25.9	**<.001**
E	`cfg.bu`	20	-10632	2.0	0.158
F	`mg.bu`	21	-11630	5.2	0.022

(d) LPTL

	Parameter	df	LogLik	χ^2	p
∅		15	-11900		
A	`trigram`	16	-11870	60	**<.001**
B	`lsa10`	17	-11853	33	**<.001**
C	`struct`	18	-11842	23	**<.001**
D	`cfg.surp`	19	-11809	65	**<.001**
E	`cfg.bu`	20	-11804	11	**0.001**
F	`mg.bu`	21	-11793	22	**<.001**

Table 3: Step-wise model comparison results for all ROIs.

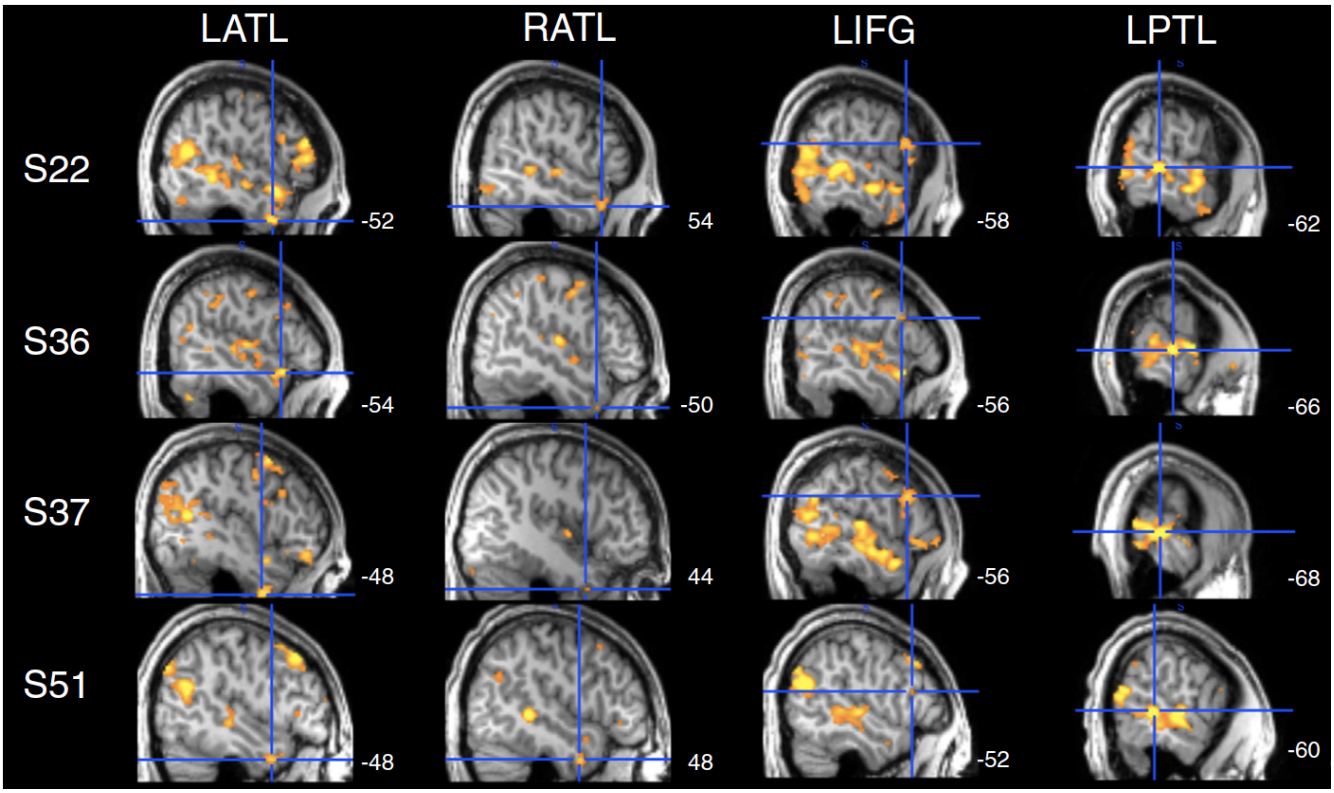


Figure 1: ROIs (column) from four participants (rows). From Brennan et al. (2016)

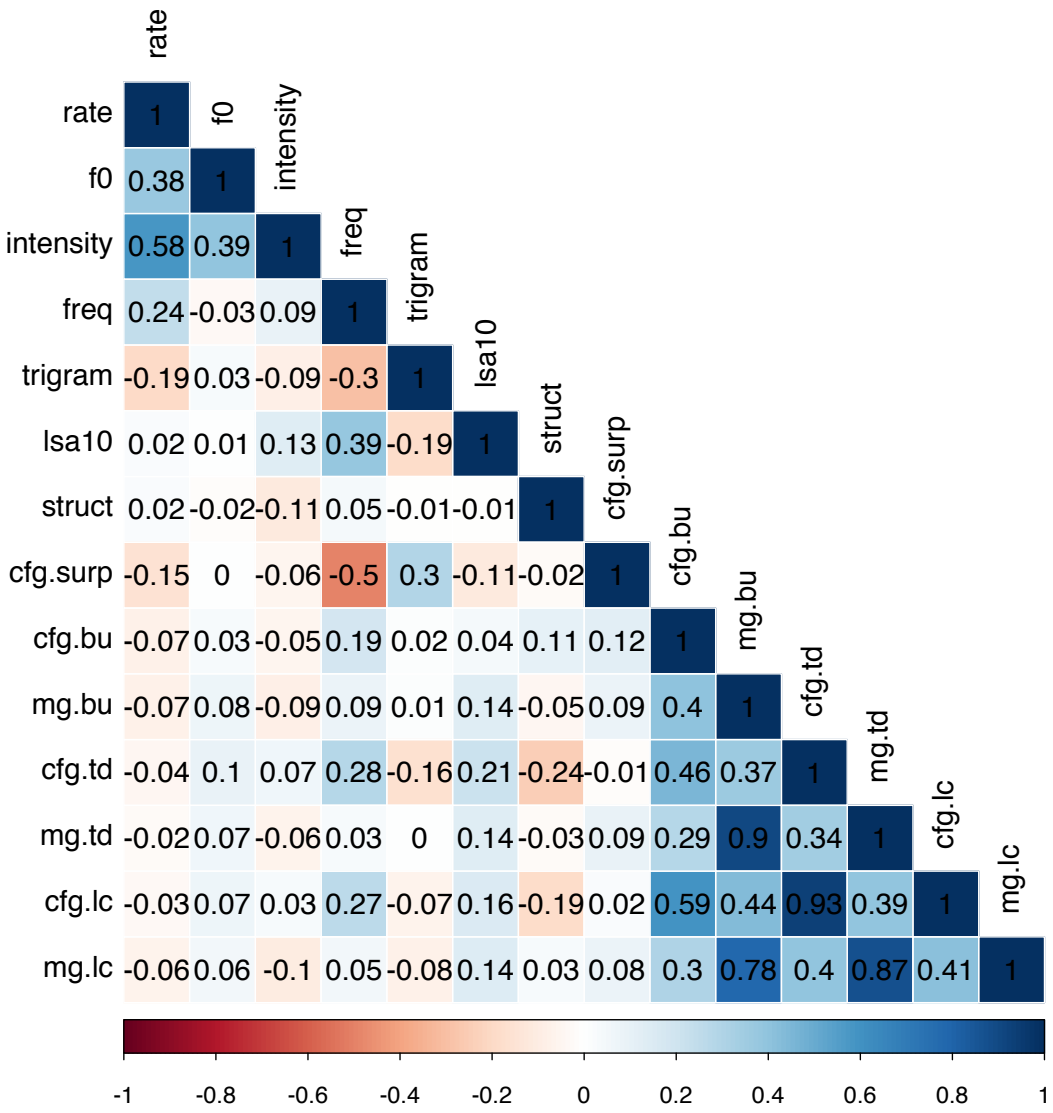


Figure 2: Pearson's correlation coefficients r between each predictor.

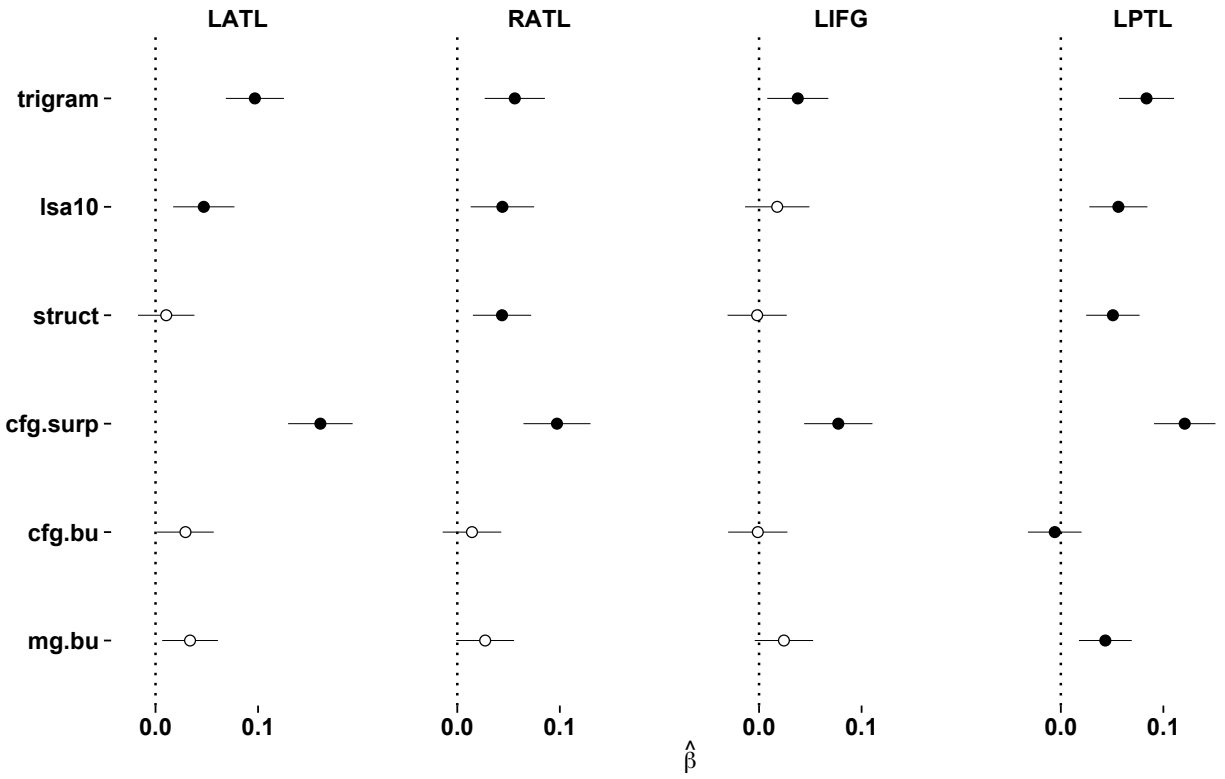


Figure 3: Fitted coefficients for all linguistic predictors across six ROIs. Coefficients show the estimated change in BOLD signal per unit change in the linguistic predictor (x-axis). Error bars show 95% confidence intervals. Filled points indicate models that made a statistically significant contribution in step-wise comparison.

References

Armeni, K., Willems, R. M. and Frank, S. L.: 2017, Probabilistic language models in cognitive neuroscience: Promises and pitfalls, *Neuroscience and Biobehavioral Reviews* **83**, 579–588.

Baker, M., Johnson, K. and Roberts, I.: 1989, Passive argument raised, *Linguistic Inquiry* **20**, 219–251.

Baroni, M., Bernardi, R. and Zamparelli, R.: 2014, Frege in space: A program of compositional distributional semantics, *Linguistic Issues in language technology.* **9**, 241–346.

Baumann, P.: 2014, Dependencies and hierarchical structure in sentence processing, *Proceedings of CogSci 2014*, pp. 152–157.

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

Boston, M., Hale, J., Kliegl, R., Patil, U. and Vasishth, S.: 2008, Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam sentence corpus, *Journal of Eye Movement Research* **2**.

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

Brennan, J.: 2016, Naturalistic sentence comprehension in the brain, *Language and Linguistics Compass* **10**, 299–313.

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

Brennan, J., Stabler, E., Van Wagenen, S., Luh, W. and Hale, J.: 2016, Abstract linguistic structure correlates with temporal activity during naturalistic comprehension, *Brain and Language* **157-158**, 81–94.

Brysbaert, M. and New, B.: 2009, Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English, *Behavior Research Methods* **41**, 977–990.

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

Chomsky, N.: 1995, *The minimalist program*, MIT Press, Cambridge, Massachusetts.

Christiansen, M. and MacDonald, M.: 2009, A usage-based approach to recursion in sentence processing, *Language Learning* **59**, 129–164.

de Marneffe, M., MacCartney, B. and Manning, C.: 2006, Generating typed dependency parses from phrase structure parses, *LREC 2006*.

Demberg, V. and Keller, F.: 2008, Data from eye-tracking corpora as evidence for theories of syntactic processing complexity, *Cognition* **101**, 193–210.

Dronkers, N. F., Wilkins, D. P., Van Valin, R. D., Redfern, B. B. and 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.

Erk, K.: 2012, Vector space models of word meaning and phrase meaning: A survey, *Language and Linguistics Compass.* **6**, 635–653.

Ettinger, A., Feldman, N., Resnik, P. and Phillips, C.: 2016, Modeling N400 amplitude using vector space models of word representation, *in* A. Papafragou, D. Grodner, D. Mirman and J. Trueswell (eds), *Proceedings of the 38th annual conference of the cognitive science society*, Austin, TX: Cognitive Science Society, pp. 1445–1450.

Everaert, A., Huybregts, M., Chomsky, N., Berwick, R. and Bolhuis, J.: 2015, Structures, Not Strings: Linguistics as Part of the Cognitive Sciences, *Trends in Cognitive Sciences* **19**, 729–743.

Federmeier, D. and Kutas, M.: 1999, A rose by any other name: Long-term memory structure and sentence processing, *Journal of Memory and Language* **41**, 469–495.

Ferreira, F. and Patson, N.: 2007, The 'good enough' approach to language comprehension, *Current Directions in Psychological Science* **1**, 71–83.

Firth, J.: 1957, A synopsis of linguistic theory, 1930-1955, *Studies in linguistic analysis*, Oxford: Philological Society, pp. 1–32.

Fong, S. and Berwick, R.: 2008, Treebank parsing and knowledge of language: A cognitive perspective, *in* B. Love, L. McRae and V. Sloutsky (eds), *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, pp. 539–544.

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

Frank, S., Otten, L., Galli, G. and 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. Dowty, L. Karttunen and A. Zwicky (eds), *Natural language parsing: Psychological, computational, and theoretical perspectives*, Cambridge, UK: Cambridge University Press, pp. 129–189.

Gibson, E.: 1998, Syntactic complexity: Locality of syntactic dependencies, *Cognition* **68**, 1–76.

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

Hale, J.: 2001, A probabilistic Earley parser as a psycholinguistic model, *Proceedings of NAACL*, Vol. 2, pp. 159–166.

Hale, J.: 2003, *Grammar, uncertainty and sentence processing*, PhD Thesis, Johns Hopkins University.

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

Hale, J.: 2016, Information-theoretical complexity metrics, *Language and Linguistics Compass.* **10**, 397–412.

Hasson, U. and Honey, C.: 2012, Future trends in neuroimaging: Neural processes as expressed within real-life contexts, *Neuroimgae* **62**, 1272–1278.

Henson, R. and Friston, K.: 2007, Convolution models for fMRI, *in* K. Friston, J. Ashburner, S. Kiebel, T. Nichols and W. Penny (eds), *Statistical Parametric Mapping*, Academic Press, London, pp. 178 – 192.

Hsiao, F. and Gibson, E.: 2003, Processing relative clauses in Chinese, *Cognition* **90**, 3–27.

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

Jäger, L., Chen, Z., Li, Q., Lin, C. and Vasishth, S.: 2015, The subject-relative advantage in Chinese: Evidence for expectation-based processing, *Journal of Memory and Language* **79-80**, 97–120.

Just, M., Carpenter, P., Keller, T., Eddy, W. and Thulborn, K.: 1996, Brain activation modulated by sentence comprehension, *Science* **274**, 114–116.

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

Kayne, R.: 1994, *The antisymmetry of syntax*, Cambridge, MA: MIT Press.

Klein, D. and Manning, C.: 2003, Accurate unlexicalized parsing, *Proceedings of the 41st Meeting of the association for computational linguistics.*, pp. 423–430.

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

Landauer, T.: 2007, LSA as a theory of meaning, *in* T. Landauer, D. McNamara, S. Dennis and W. Kintsch (eds), *Handbook of latent semantic analysis*, Mahwah, NJ: Lawrence Erlbaum Associates, pp. 1–34.

Landauer, T. and Dumais, S.: 1997, A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge, *Psychological Review* **104**, 211–240.

Larson, R.: 1988, On the double object construction, *Linguistic Inquiry* **19**, 335–391.

Lewis, R. and Vasishth, S.: 2005, An activation-based model of sentence processing as skilled memory retrieval, *Cognitive Science* **29**, 375–417.

Luong, M., Fran, M. and 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.

Michel, J. e. a.: 2011, Quantitative analysis of culture using millions of digitized books, *Science* **331**, 176–182.

O'Grady, W.: 1997, *Syntactic development*, Chicago, IL: University of Chicago Press.

Pallier, C., Devauchelle, A. and Dehaene, S.: 2011, Cortical representation of the constituent structure of sentences, *Proceedings of the National Academy of Sciences* **108**, 2522–2527.

Pylkkänen, L.: 2015, Composition of complex meaning: Interdisciplinary perspectives on the left anterior temporal lobe, *in* G. Hickok and S. Small (eds), *Neurobiology of Language*, Elsevier, pp. 622–629.

Pynte, J., New, B. and Kennedy, A.: 2008, A multiple regression analysis of syntactic and semantic influences in reading normal text, *Journal of Eye Movement Research* **2**, 1–11.

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

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

Sayeed, A., Fischer, S. and Demberg, V.: 2015, Vector-space calculation of semantic surprisal for Vector-space calculation of semantic surprisal for predicting word pronunciation duration, *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing.*, Vol. 1, pp. 763–773.

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

Sprouse, J. and Hornstein, N.: 2016, Syntax and the cognitive neuroscience of syntactic structure

nuilding, *in* G. Hickok and S. Small (eds), *Neurobiology of language*, Amsterdam : Elsevier/Academic Press, pp. 165–174.

Stabler, E.: 1983, How are grammars represented?, *Behavioral and Brain Sciences* **6**, 391–421.

Stabler, E.: 1997, Derivational minimalism, *in* Retoré (ed.), *Logical aspects of Logical aspects of computational linguistics*, Springer, pp. 68–95.

Stabler, E.: 2011, Computational perspectives on minimalism, *in* C. Boeckx (ed.), *The Oxford handbook of linguistic minimalism*, Cambridge, UK: Oxford University Press, pp. 616–641.

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

Stowe, L. A., Haverkort, M. and Zwarts, F.: 2005, Rethinking the neurological basis of language, *Lingua* **115**, 997 – 1042.

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

Sturt, P. and Lombardo, V.: 2005, Processing coordinated structures: Incrementality and connectedness, *Cognitive Science* **19**, 291–305.

van Wagenen, S., Brennan, J. and Stabler, E.: 2014, Quantifying parsing complexity as a function of grammar, *in* C. Schütze and L. Stockall (eds), *UCLA working papers in linguistics.*, Vol. 18, UCLA Linguistics Department, pp. 31–47.

Wanner, E. and Maratsos, M.: 1978, An ATN approach to comprehension, *in* M. Halle, J. Bresnan and G. Miller (eds), *Linguistics theory and psychological reality.*, The MIT Press.

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

Wilson, S., DeMarco, A., Henry, M., Gesierich, B., Babiak, M., Mandelli, M., Miller, B. and Gorno-Tempini, M.: 2014, What role does the anterior temporal lobe play in sentence-level processing? Neural correlates if syntactic processing in sematic PPA, *Journal of Cognitive Neuroscience* **26**, 970–985.

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

Yngve, V.: 1960, A model and a hypothesis for language structure, *Proceedings of the American Philosophical Society.*, Vol. 104, pp. 444–466.

Yoshida, M., Dickey, M. and 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., Whitman, J. and Hale, J.: 2010, Subject-object asymmetries in Korean sentence comprehension, *in* R. Catrambone and S. Ohlsson (eds), *Proceedings of the Annual Meeting of the Cognitive Science Societry*, pp. 2152–2157.