Quantifying Parsing Complexity as a Function of Grammar

Sarah VanWagenen · Jonathan Brennan · Edward P. Stabler

Recent research has used measures of syntactic complexity generated by computational models to predict human processing effort in behavioral and neurophysiological experiments. These models often rely on simplifying grammatical assumptions and the dominant view in the literature has been that such simplifications are adequate approximations for processing accounts. However, the effect of chosen grammar on processing complexity has not been empirically investigated in a natural text. We compared the impact of grammar on estimated processing complexity for a 12 minute English narrative. One grammar was based on the Penn Treebank 2 schema, and the second was based in the Minimalist Grammar framework and represented as an X-bar schema. The impact of grammar was contrasted against the impact of parsing strategy using two stack-based strategies and complexity was measured in terms of the number of nodes built and the stack depth word-by-word. The impact of grammar choice matched or exceeded the impact of parsing strategy. Thus even for a single, fixed corpus and a fixed parsing strategy, the grammar matters.

This paper was presented at the 2012 meeting of the CUNY Conference on Human Sentence Processing in New York City as part of a special session on the relationship between the grammar and the parser. Sarah VanWagenen and Jonathan Brennan were listed jointly as first authors. The text and figures are minimally changed from that presentation.

Introduction

There has been growing interest in the psycholinguistic community in using computational models of language comprehension to predict human processing effort during more-or-less naturalistic processing (Bachrach 2008; Brennan et al. (2012); Boston et al. (2008); Demberg and Keller (2008); Demberg (2010); Roark et al. (2009). Such research necessarily involves developing language models that have rather broad coverage, and it has therefore been practical to rely on relatively simplified grammars for such work. For example, one common choice has been to use context-free grammars that are in accordance with the Penn Treebank 2 annotation scheme (Demberg and Keller 2008; Roark et al. 2009; Brennan et al. 2012; Levy 2005).

A question that arises is to what degree might the choice of such simplified grammars bear on processing estimates? In other words, do generalizations about processing effort...
hold across different grammatical assumptions? Differing opinions on this question can be found in the literature, though answers are usually limited to specific constructions. Gibson offers one well-known complexity metric based on parser memory states purported to be independent of any grammatical assumptions (Gibson 1998:fn 1, 8). On the other hand, Hale (2006) provides an example of work demonstrating that even fine-grained grammatical assumptions may have a substantial impact on processing within a specific domain, showing that different grammatical analyses of relative clauses make very different predictions for incremental processing effort due to syntactic uncertainty.

Developing adequate models of language performance requires making choices on a number of parameters. Not only must the modeler choose the grammar $G$ that describes the linguistic representations being used but one must also make choices concerning the parsing algorithms $A$ used to map from input to syntactic representation (varying with regards to derivation expansion and indeterminacy handling), and on the probability distributions $P$ over possible parses in the case of ambiguity. The claim we would like to explore today is that the choice of grammar matters at least as much as any other parameter when using models to make inferences about processing effort. We aim to quantify the intuition that for reasonable choices of $P$ and $A$, the choice of $G$ matters. In practical terms, we will do this using a pair of grammars, two different parsing algorithms, and two different measures of complexity, all of which will be discussed in greater detail below.

Our claim is summarized in the two following two equations, where IN stands for a complexity estimate based on incremental node count, IM for one based on memory load, $CF_T$ and $X_{MG}$ stand for specific grammatical proposals, U stands for a uniform probability model, TD for a top-down parsing algorithm, and 1 indicates the beam size.¹ We will show that complexity estimates for two language models that differ only in grammar are significantly and substantially different; estimates when using a context-free treebank grammar are not proportional to those from a mildly context-sensitive X-bar Minimalist Grammar. Finally, we will provide a small demonstration that the differences argued for here have consequences for modeling human syntactic processing data.

\[
\begin{align*}
(1) \quad & \text{IN}(CF_T, U, TD, 1) \neq \text{IN}(X_{MG}, U, TD, 1) \\
(2) \quad & \text{IM}(CF_T, U, TD, 1) \neq \text{IM}(X_{MG}, U, TD, 1)
\end{align*}
\]

1 Model Parameter Values and Text

1.1 Text

The text used for our models is a segment from Alice in Wonderland by Lewis Carroll containing approximately 160 sentences.² The text offers a rich variety of naturally occurring sentence structures requiring a broad coverage language model, allowing us to generalize the effect of grammar choice across sentence types. Defining our models over this particular text also allows us to connect with processing data in the form of neuroimaging results from a set of subjects who passively listened to this story during MRI scanning (Brennan et al. 2012).

¹ A beam size of one is easy to analyze and approximates a one-path no ambiguity situation.
² The segment does not include the jabberwocky poem.
1.2 Parameter Values

We constructed four language models by varying two factors: the grammar and the parser used. We used a context-free penn treebank grammar, and a hand-written minimalist grammar with X-bar structures, and we crossed grammar choice with a parser that was either top-down or bottom-up. We will discuss each of these choices briefly.

1.2.1 Treebank grammar, context-free models ($CF_TB$)

The Penn Treebank was not designed as a linguistically, or psychologically plausible grammatical theory (Marcus et al. 1993; Seagull and Schubert 1991; Bikel 2004; Honnibal et al. 2010). Rather, the treebank scheme was designed to offer rich syntactic annotation for a number of computational applications while requiring a minimum of grammatical assumptions. In practice, this leads to a grammar in which structures are relatively flat and there is little generalization, and thus many very similar constructions are listed separately. In other words, much like lexical items, the number of rules is an inverse exponential function of rank (i.e. Zipf-distributed). A substantial proportion of rules occur only once (40% in the treebank for the wall street journal corpus.) The $CF_TB$ was made using a parser (Bikel 2002) trained on Treekbank2 annotated text from the Wall Street Journal (Marcus et al. 1993) and manually reviewed for correctness.

1.2.2 Minimalist Grammars, X-Bar models ($\mathcal{X}_{MG}$)

To provide a contrast to the theory neutral treebank style grammar, we developed a hand-built minimalist grammar using X-bar style structures. MGs are lexicalist grammars in which well-formed structures are those that are generated by recursive application of merge rules and move rules over feature bundles (Stabler 2010). The structure of the feature bundles, and the possibility of granting syntactic features to phonologically null terms, offers means to encode underlying grammatical generalizations, which means in practice that a wide range of syntactic structures are encoded significantly more concisely, leading to more regularly structured and deeper trees.

In our grammatical fragment, structural generalizations were based on relatively uncontroversial analyses from the theoretical linguistics literature listed here, which included coverage for relative clauses, case marking, head movement, empty lexical items (e.g. PRO v, p) quantification etc. The analyses encoded in the grammar follow well-known theoretical proposals for constructions including ditransitives (Larson 1988), relative clauses (Kayne 1994), case checking (Haegeman 1991), the dative alternation (Baker 1997), passives (Baker et al. 1989), head movement, genetivies, raising, ECM, control, quantification, and wh-movement (Sportiche et al. 2014).

The trees in (3–4) illustrate the different grammatical analyses provided by our two grammars for the same sentence from the text, “There are no mice in the air, I’m afraid”. (3) shows the tree generated by the context free treebank grammar, and (4) the tree from the X-bar minimalist grammar. Note the differences in both depth and empty structure; also note the right-branchingness of both grammars.

---

3This grammar, containing $\approx 1200$ lexical items, was first presented in VanWagenen et al. (2011).
(3)

S

EX

There

VP

are

NP

do

mice

IN

the

NP

air

(4)

CP

CeP(4)

Ce

TP

there

T

BeP

T

Be

 DP(5)

T'

Be'

DP(5)

Be

Be

m

d

 NumP

 NP

 mice

 Be

 pP

 p

 P

 DP(1)

 D

 NumP

 NP

 air

P'
1.3 Parsers

In order to evaluate the magnitude of the effect of grammar on estimates of processing complexity, we crossed choice of grammar with choice of parsing algorithm. Parsing strategies form a lattice of predictiveness, with the most predictive, fully-top-down (TD) parser forming the greatest upper bound and the non-predictive bottom-up (BU) parser, forming the least lower bound. We created language models with both TD and BU parsing algorithms, these choices offer the greatest possible contrast for the effect of this parameter. As is familiar, these parsing strategies have different memory requirements for parsing different kinds of structure (Abney and Johnson 1991): TD parsers handle right-branching structures of unbounded length with finite memory and BU parsers handle left-branching structures of unbounded length with finite memory. Both have memory requirements proportional to sentence length for center-embedded structures.

1.4 Complexity Measures

The final choice we made for our analysis was in the complexity metrics. We calculated, word-by-word, two metrics that are familiar from psycholinguistic and computational applications. Incremental Node Count (IN) simply tracks the number of new nodes constructed by the parser between the current term and the previous (Frazier 1985; Hawkins 1994; Miller and Chomsky 1963). In computational terms it stands in for parser run time.

As a measure of Incremental Memory (IM) we counted the number of elements being held in memory at a given leaf node, i.e. the number of elements in the stack word-by-word (Yngve 1960; Martin and Roberts 1966; Abney and Johnson 1991; Resnik 1992; Johnson-Laird 1983). We have not calculated complexity metrics based on information theoretic measures, though there is significant interest in these. We note, however, that (Hale 2006) provides some evidence that such metrics are sensitive to grammatical assumptions at a relatively fine degree of granularity.

1.5 Modeling Summary

To summarize, we constructed four language models for a narrative text which was chosen to permit generalizations across a range of sentence structures. Our models used one of two grammars, a CFG based on the Penn Treebank that is relatively common for broad-coverage parsing (\textit{CFG}_{TB}), and a more sophisticated yet relatively uncontroversial MG-based X-bar grammar (\textit{XX}_{MG}). We also varied whether the parser was fully-predictive TD, or non-predictive BE and, finally we computed two simple and standard complexity measures: incremental node count (IN) and incremental memory load (IM).

\[^{4}\text{More precisely, IN counts the nodes in the tree after parsing word } n \text{ and subtracts from that number the count from } n - 1.\]
Figure 1: Pairwise correlations for IN. X and Y axes indicate node counts from each model and each cell plots a different pairing between models, with the best linear fit shown by the blue lines. The two grayed out cells represent comparisons of non-interest. Pearson’s $r$ values are given at top.

2 Results: Correlations

To evaluate the effect of varying grammar choice and parser choice, we computed the pairwise correlations for IN and IM counts across the narrative between four model pairings of interest:

- TD, $\hat{X}_{MG}$ (TDX)
- TD, $CF_{TB}$ (TDTB)
- BU, $\hat{X}_{MG}$ (BUX)
- BU, $CF_{TB}$ (BUTB)

We first look at the results for incremental node count.

2.1 Incremental Node Count (IN)

Figure 1 shows the global pattern of pairwise correlations for IN. Looking first an on the left, little to no correlation was seen between parser estimates for the TD language models
Figure 2: (a) Correlation for IN between BU and TD models for the $X_{MG}$ grammar for 8 randomly selected sentences. (b) Word-by-word IN estimates for the same sentences.

when grammar was varied, or for the CF$_T$B models when parser was varied. Interestingly, a high correlation was seen between node count measures from the $X_{MG}$ grammar across parser choice, shown in the bottom right. That is, estimates of node count when using an X-bar grammar were relatively similar regardless of whether a TD or BU algorithm was used. A moderately high correlation was also seen between models with the BU parser when grammar was varied, shown to the left on the bottom. To get insight into how reliable these correlations are across different structures, we examined these patterns at the sentence level.

Figure 2a shows the correlation between BU and TD models for the $X_{MG}$ grammar for 8 randomly selected sentences. A relatively strong correlation between estimates from these two models is apparent in all sentences. Viewing the word-by-word traces for the same two $X_{MG}$ models (Figure 2b) also shows a strong relationship between node count estimates throughout each sentence. These results suggest that for the X-bar grammar, changing in parsing algorithm did not have much effect on incremental node count.

There was also an apparent correlation between models with different grammars for the BU parser (Figure 3). Looking at individual sentences, while all sentences show a positive correlation, it looks as if these correlations are strongly driven by one or two outliers. The word-by-word traces indicate that there is a sharp up-swing on the final word of the sentence for both models. This is to be expected for a BU parser of predominantly right
Figure 3: (a) Correlation for IN between $CF_B$ and $\hat{X}_{MG}$ for the BU parser for 8 randomly selected sentences. (b) Word-by-word IN estimates for the same sentences.
branching structures which by design waits until all daughters have been encountered before constructing any higher nodes. While the values of the node counts differ across grammars, the relative increase at the end is similar. This effect appears to drive the moderate correlation between these models overall. Interestingly, for each of the example sentences, little to no relationship is seen internal to the sentences. These results suggest that despite the presence of a moderate global correlation, these two models do not in fact yield similar on-line, sentence internal, estimates of incremental node count.

We summarize results for the IN measure by plotting the distribution of sentence-by-sentence correlation coefficients for each pairing of interest (Figure 4). Here, per-sentence correlation coefficients are plotted on the y-axis, while each column indicates a different comparison between models. Given the strong effect of end-of-sentence outliers we observed, this plot uses correlation values for which the last word in each sentence was excluded (though the role of end-of-sentence terms should be considered in future investigations.) When the parsing algorithm is held constant, the choice of grammar radically changes node count estimates, leading to little to no consistent correlations across sentences for models both TD and BU models, shown in the middle two columns of the plot. The pattern is quite different when grammar is held constant and parsing algorithm varied. For the $\bar{X}_{MG}$ grammar, there was remarkably little effect of parser choice (the right-most column). In contrast, for the $\text{CF}_{TB}$ grammar, choice of parser strongly influences the node count estimates (the left-most column).

In sum, for the node-count measure of complexity is there is a striking interaction where the influence of parsing algorithm in fact depends on the grammar used.
2.2 **Incremental Memory (IM)**

We next turn to incremental memory, quantified in terms of the depth of the parser’s stack at a given word. Again we start with a global overview (Figure 5) showing pairwise correlations across the entire narrative. As with IN, little to no significant correlation is seen for IM when grammar is varied for TD models, or when parser is varied for CF models, shown on the left. On the other hand, moderate correlations are seen between TD models when grammar is varied, shown in the bottom left, as well as between the $X_{MG}$ models when the parser is varied, shown to the right on the bottom.

Looking sentence-by-sentence, Figure 6a compares BU models across the two grammars for a sub sample of sentences. In contrast to what was seen for node count, we see strong correlations (the magnitude of which was not clearly evident at the global level.) While there is a strong correlation in all sentences, note that the slope of the relationship differs substantially across this subset of sentences, giving some indication of why the global pattern

**Figure 5:** Pairwise correlations for IM. X and Y axes indicate node counts from each model and each cell plots a different pairing between models, with the best linear fit shown by the blue lines. The two grayed out cells represent comparisons of non-interest. Pearson’s $r$ values are given at top.
Figure 6: (a) Correlation for IM between $\bar{X}_{MG}$ and $CF_{TB}$ models for the BU parser in 8 randomly selected sentences. (b) Correlation for IM between TD and BU models for $\bar{X}_{MG}$ in the same sentences.
Figure 7: Distribution of sentence-by-sentence IM correlation values for each pairing of interest. Each data-point represents the correlation coefficient between the named pairing for a single sentence in the text. The horizontal bars indicate the median, also given at top; boxes span the inter-quartile range. Sentence-final words were removed prior to calculating these correlations.

was reduced compared to what is seen at the sentence-level. Figure 6b shows correlations between the $\bar{X}_{MG}$ models across parser type. We see that while some sentences show a relatively robust correlation across parser choice, this relationship breaks down on others. This indicates that the influence of parser choice for the X-bar grammar does not generalize across the range of sentences.

Turning finally to the distributions of sentence-by-sentence correlations for incremental memory (Figure 7), the simple summary is that the effect of grammar on processing estimates, while apparent, is less pronounced than was the case with node count. For bottom-up models, there in fact little influence of grammar choice (with a few outliers) on estimates of IM (mid-right column). Choice of grammar has a much stronger, but still somewhat attenuated effect for TD models; while a majority of structures show at least some positive correlation between IM estimates across grammar types, the distribution remains relatively wide (mid-left column). Not surprisingly, the effect of parsing algorithm has a strong effect on IM. For the $CF_{TB}$ models, algorithm choice radically changes memory estimates (left column); the effect of algorithm is less pronounced for the $\bar{X}_{MG}$ grammar (right column). Similar to the comparison of TD models, there is a moderate degree of consistency between estimates for the $\bar{X}_{MG}$ models, though this again has a wide distribution across individual sentences.

In sum, incremental memory estimates are affected by the grammar used for modeling, but this effect is reduced compared to what is seen for parsing algorithm choice.
Comparing the models with a neural measure of on-line processing

Our primary focus has been on the theoretical influence of the choice of grammar on estimates of processing complexity. We find that grammar choice matters to a striking degree for node count estimates, and to a more moderate degree for memory-based processing estimates. Other work has shown that grammar choices also influence uncertainty-based measures of complexity Hale (2006). Before concluding, we would like to briefly discuss the role these choices play when evaluated against on-line processing measures taken to reflect syntactic processing.

Evaluating which language model offers the “best” account of human performance is a familiarly tricky problem, in part due to the fact that there is not any clear gold standard processing metric that has been tied to specific parsing computations (e.g. structure-building, memory usage, ambiguity resolution). Acknowledging this limitation, we focus our attention on just a single computation, that of syntactic structure building, and one plausible processing correlate: neural activity in a brain region that has been associated with this computation.

A number of studies have suggested that the anterior temporal lobe of the left hemisphere is importantly involved in the construction of syntactic phrase structure (Mazoyer et al. 1993; Stowe et al. 2005; Friederici et al. 2000; Fertl et al. 2008; Grodzinsky and Friederici 2006; Hickok and Poeppel 2007; Brennan et al. 2012; Brennan and Pyllkkänen 2012; Dronkers et al. 2004), though we will not review the arguments if favor of this hypothesis. We evaluated the degree to which estimates of incremental node count across different models was predictive of brain activity in this region. We used data from a previously published fMRI experiment (Brennan et al. 2012). In that experiment, nine subjects passively listened to 30 minutes of Alice in Wonderland, a 12-minute subset of which corresponded to the fragment that we have analyzed with our X-bar and Treebank grammars.

Given the current question about the role of model choice, we focus our attention on two regions of interest (ROIs) in the left anterior temporal lobe which were identified using an separate data subset and independent contrast design to find language-related regions broadly. Details of data collection and analysis are given in Brennan et al. (2012).

Figure 8 shows the two anterior temporal regions of interest, one along the superior temporal gyrus of the left temporal lobe, marked 1 and a second encompassing the left...
temporal pole, marked (2). Figure 9a plots the regression coefficients and their standard errors between incremental node count from each of our four models and activity in the superior temporal gyrus region (1). Coefficients that are greater than two standard errors from zero indicate a statistically significant effect. We find that activity in this region strongly correlates with node count estimates from both of the X-bar models, and less strongly related to estimates from the Tree-bank models. That is, there is a strong influence of the grammar on estimates for neural activity in this region. The influence of parsing algorithm appears negligible.

Interestingly, a different pattern is observed in a second ROI located in the left temporal pole (Figure 9b). There we see significant relationship between node count and activity in this region across models with different grammars, and also with different parsing algorithms. It is clear, however, that these four estimates are not significantly different from each other. We tentatively suggest that the pattern in the temporal pole region (2) may reflect processing at major phrase boundaries that is preserved across all of our models.

In sum, while there was little influence of grammar for one of our neural regions, a second region showed strong associations with node counts from the X-bar grammar, but only weakly correlated with node counts from the treebank grammar. While we have focused rather narrowly on just one measure of processing effort and two candidate indexes of human syntactic processing, the take-home message is that choice of grammar can strongly influence the types of inferences one might draw about a particular processing correlate.

**Conclusions**

The guiding aim for much psycholinguistic research is determining the best (Grammar, Probability Distribution, Algorithm) triple to account for human performance. What we have addressed here is the degree to which choice of $G$ matters for reasonable choose of $(P, A)$. We find that $G$ matters substantially for estimates based on IN and moderately for IM. Furthermore, these changes have consequences for predictions about human performance.
Our analysis also reveals a rather interesting interaction effect such that \((G, P, A)\) choices are not independent.

**Acknowledgements**

We are grateful for funding to EC 7th Framework Programme, Marie Curie ITN-LAN (214570) (SVW), The NYU Abu Dhabi Research Institute Neuroscience of Language Lab, dir. Alec Marantz (JB) and to Timothy P. L. Roberts and the Children’s Hospital of Philadelphia Research Institute (JB).

**References**


**Affiliations**

Sarah VanWagenen  
Department of Linguistics  
UCLA  

Jonathan Brennan  
Department of Linguistics  
University of Michigan  
jobrenn@umich.edu  

Edward P. Stabler  
Department of Linguistics  
UCLA  
stabler@ucla.edu