Capturing Online SRC/ORC Effort with Memory Measures from a Minimalist Parser

Aniello De Santo
Dept. of Linguistics, University of Utah
aniello.desanto@utah.edu

Abstract

A parser for Minimalist grammars (Stabler, 2013) has been shown to successfully model sentence processing preferences across an array of languages and phenomena when combined with complexity metrics that relate parsing behavior to memory usage (Gerth, 2015; Graf et al., 2017; De Santo, 2020b, a.o.). This model provides a quantifiable theory of the effects of fine-grained grammatical structure on cognitive cost, and can help strengthen the link between generative syntactic theory and sentence processing. However, work on it has focused on *offline* asymmetries. Here, we extend this approach by showing how memory-based measures of effort that explicitly consider minimalist-like structure-building operations improve our ability to account for word-by-word (*online*) behavioral data.

1 Introduction

Formally specifying hypotheses about how grammatical structure drives processing cost makes it possible to connect long-standing ideas about cognitive load in human language processing with representational assumptions in theoretical syntax — thus adding to the interpretability of theories of sentence comprehension, and to the plausibility of particular syntactic analyses/theories of syntactic representations (Bresnan, 1978; Berwick and Weinberg, 1982; Kaplan and Bresnan, 1982; Hale, 2001, 2011).

In this sense, recent studies have argued that the behavior of a parser for Minimalist grammars (Stabler, 1996) can link structural complexity to memory usage. In particular, this takes the form of a specific implementation of Stabler (2013)'s top-down parser, coupled with complexity metrics measuring how a tree traversal algorithm recruits memory resources (Kobele et al., 2013). This model makes fully specified commitments to (a) the nature of the structures built during the parsing process, (b) the time-course of the structure building operations connecting linear input to hierarchical representations, and (c) a psychologically plausible theory of how cognitive resources are linked to parsing operations to derive measures of sentence difficulty. Thanks do these commitments, this approach offers an insightful, empirically grounded reframing of past theories trying to bridge the study of competence and the study of performance (e.g., the Derivational Theory of Complexity; Miller and Chomsky, 1963; Fodor and Garrett, 1967; Berwick and Weinberg, 1983; De Santo, 2020b).

From an empirical perspective, computational modeling work in this framework has proved successful in accounting for a number of processing preferences across a variety of phenomena cross-linguistically (Gerth, 2015; Graf et al., 2017, a.o.). Most of this work has focused on deriving estimates of *offline* (over a whole sentence) effort, which then has been used to qualitatively evaluate categorical contrasts between minimally different sentence pairs. However, if we aim to probe the cognitive plausibility of a Minimalist Grammar model, it is important to understand its ability to capture fine-grained sentence comprehension processes, beyond broad, sentence-level complexity profiles (Demberg and Keller, 2008; Li and Hale, 2019).

In this paper, we extend this approach by extracting a metric of word-by-word effort from memory-usage measures defined in previous work on offline effects. We then evaluate this complexity metric based on its ability to capture difficulty profiles in self-paced reading from a large scale dataset. As this model implements theories of effort grounded in memory load, we also compare its predictions to those of a metric (surprisal) estimating word predictability (Hale, 2001).

2 MG Parsing and Cognitive Effort

We adopt a model combining a parser for Minimalist Grammars with metrics measuring memory usage. In the rest of this section we outline the core intu-

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 24–35
May 3, 2025 ©2025 Association for Computational Linguistics

itions behind this approach to sentence difficulty. While it is possible to implement alternative cognitive models incorporating Minimalist Grammar parsers, we refer to the specific set of choices made here as the **MG model** for ease of discussion.

2.1 A Brief introduction to MGs

Minimalist Grammars (MGs; Stabler, 1996) are a mildly-context sensitive, transformational formalism incorporating ideas from the Minimalist Program (Chomsky, 1995). An MG grammar consists of a sets of lexical items associated with a non-empty string of syntactic features and two core transformational operations — Merge and Move. Merge is a binary operation encoding subcategorization, while Move is a unary operation allowing for a movement approach to long-distance, filler-gap dependencies. Importantly for us, the central data structure of MGs is a *derivation tree*, explicitly encoding the sequence of Merge and Move operations required by a given sentence (Michaelis, 1998; Harkema, 2001; Kobele et al., 2007). Derivation trees differ from more commonly known phrase structure trees in that moving phrases remain in their base position, and thus the final, linear word order of a sentence is not directly reflected in the order of the leaf nodes in the tree (see Figure 1a).

Since MGs are able to exemplify the structurally rich analyses of modern generative syntax, they can contribute to the development of models of sentence processing that provide insights into the connection between fine-grained syntactic structure and offline processing behavior. This is the intuition behind a line of computational modeling work which, starting with Kobele et al. (2013), has shown that a top-down parser for MGs (Stabler, 2013) is successful in predicting offline processing difficulty contrasts.

2.2 MG Parsing

Stabler (2013)'s parser is adapted from a standard recursive-descent parser for CFG, accounting for the mismatch between the order of lexical items in a derivation tree and the linear surface order. Broadly, the parser scans the nodes from top to bottom and from left to right. Given the way Move is implemented however, simple left-to-right scanning of the leaf nodes yields an incorrect word order. In order to keep track of the derivational operations affecting linear order, the MG variant follows the standard approach of predicting nodes downward (toward words) and left-to-right only until a Move node is predicted. At that point, the pure top-down strategy is discarded, and the parser instead follows the shortest path towards predicting the moved item's base position (a *string-driven* strategy). After a position for the mover has been found, the parser continues from the point where the the top-down strategy had been paused (Figure 1b).

The memory stack associated to the parser plays a fundamental role in this: if a parse item is hypothesized at step i, but cannot be worked on until step j, it must be stored for $j - i$ steps in a priority queue. For instance, consider the derivation tree in Figure 1a for the sentence *Who do the Gems love __?*. Here, the node for *do* is predicted at step 3 but it is only flushed out of the parser's stack at step 10. This is because a movement dependency for *who* has been postulated at Spec,CP. Upon encountering *who* in the input string and predicting a movement operation, the parser cannot integrate the mover into the structure until a base position for it has been predicted and confirmed (at step 8 and 9). While doing so, the parser will predict intermediate structure (e.g., a position for an auxiliary in C, which could be occupied by *do*), but it will not match that prediction against the linear input until the search for *who* has been resolved.[1]

Stabler's algorithm seems to capture some core properties of human language processing strategies: it works incrementally, and it is *predictive* — it makes hypotheses about how to build the upcoming syntactic structure that need to be confirmed based on the input (Marslen-Wilson and Tyler, 1980; Tanenhaus et al., 1995; Phillips, 2003; Demberg and Keller, 2009, a.o.). As in other aspects of cognition, prediction also plays a crucial role in language processing. In the MG model, this is reflected by the fact that the predictive abilities of the string-driven top-down approach guide how the parser recruits memory resources. However, the psycholinguistic literature traditionally refers to prediction in the context of *ambiguity resolution* — the task of choosing between multiple, alternative structural hypotheses available to the parser during processing (Traxler and Pickering, 1996; Wagers and Phillips, 2009; Chambers et al., 2004; Hale, 2006). This predictive aspect has been shown to have a significant role in determining processing cost (Traxler and Pickering, 1996; Wagers and Phillips, 2009; Chambers et al., 2004), and to be modulated by past experience (Ellis, 2002; Hale, 2006; Levy, 2013).

[1]The reader in referred to (De Santo, 2020b, Chp. 2) for a deeper discussion of the differences in stack-usage between a string-driven traversal and a classic top-down traversal.

In this respect, Stabler's parser can be equipped with a search beam discarding the most unlikely predictions. Here though, we follow Kobele et al. (2013) in ignoring the beam and assuming that the parser is equipped with a perfect oracle, which always makes the right choices when constructing a tree. Essentially, the MG model adopts deterministic parsing strategy. This idealization is clearly implausible from a psycholinguistic point of view, but has a precise purpose: to ignore the cost of choosing among several possible predictions and focus on the specific contribution of structure-building strategies to processing difficulty. However, the MG model has enough flexibility to allow for the implementation and evaluation of theories of ambiguity resolution and reanalysis (Chen and Hale, 2021; De Santo and Lee, 2022; Ozaki et al., 2024). We come back to this possibility in Section 5.

2.3 Parsing Effort and Tenure

Kobele et al. (2013) introduces a tree annotation schema to make Stabler (2013)'s tree traversal strategy easy to follow (Figure 1a). Each node in a tree is annotated with the step at which it was first conjectured by the parser and placed in memory (superscript, *Index*), and the step at which it is considered completed and flushed from memory (subscript, *Outdex*). Index and Outdex thus fully encode the relation between a node and stack-states. We can then use them to link the parser's traversal strategy, syntactic structure, and memory usage. In turn, this allows us to derive predictions about sentence difficulty, based on how the structure of a derivation tree affects memory (Rambow and Joshi, 1994; Gibson, 2000; Kobele et al., 2013; Gerth, 2015).

The MG model distinguishes several cognitive notions of memory usage (Graf et al., 2017). Of interest to us is a measure of how long a node is kept in memory through a derivation (TENURE). Tenure for each node is computed considering the moment a node was first postulated into the structure (i.e., placed in the memory stack of the parser) and the moment such prediction was confirmed (i.e., the node could be taken out of memory). In practice, a node's Tenure can be computed as the difference between its index and its outdex. Considering again the annotated MG tree in Figure 1a, Tenure for *do* is $Outdex(do) - Index(do) = 10 - 3 = 7$.

As mentioned, past work has then formalized this notion in metrics of *offline* processing difficulty — for instance measuring maximum Tenure (MAXT), which ties processing difficulty to differences in grammatical structure over a whole derivation. Specifically, MAXT has been used to derive categorical processing contrasts, by comparing maximum Tenure values for derivation trees corresponding to pairs of sentences with stark asymmetries in reported offline processing preferences. For instance, Graf and Marcinek (2014) show that MAXT makes the right difficulty predictions for phenomena such as right embedding vs. center embedding, nested dependencies vs. crossing dependencies, as well as a set of cross-linguistic contrasts involving relative clauses. Following work has then strengthen the empirical support for Tenure based metrics, further demonstrating their ability to qualitatively capture offline contrasts across languages and constructions (Gerth, 2015; Graf et al., 2017; Liu, 2018; De Santo, 2019, 2020a). Evaluating this model on online patterns of effort seems then the natural next step in the enterprise. In what follows, we leverage word-by-word Tenure values as already computed by the MG model to derive online predictions.

3 Evaluating Tenure Online

Building on previous successes of the MG model in capturing offline contrasts, we ask whether structure-building effort as captured by Tenure improves estimates of word-by-word reading time patterns. We show that Tenure as computed by the model can be directly leveraged to derive predictors of processing difficulty. We then evaluate Tenure against surprisal measures extracted from two different neural architectures, as an implementation of expectation-based complexity metrics.

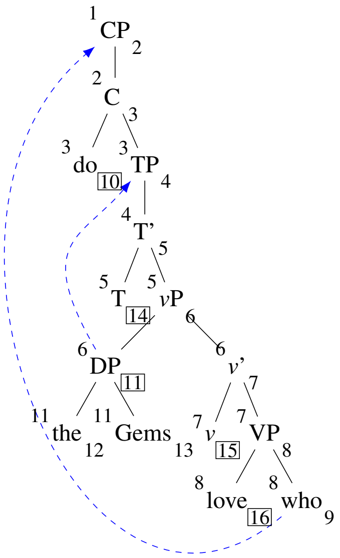
3.1 Reading Time Data

The relative comprehension difficulty of object-extracted (ORC; 2) over subject extracted (SRC; 1) relative clauses is well-attested both in English and cross-linguistically (Lau and Tanaka, 2021).

1. The Pearl who welcomed the Diamond.

2. The Pearl who the Diamond welcomed.

Additionally, while this difficulty has been partially linked to the lower frequency/predictability of ORCs (Chen and Hale, 2021; Vani et al., 2021), expectation-based approaches have been argued to fall short in accounting for the overall pattern of relative complexity. Instead (or additionally), a subject preference in RCs can be associated to the impact of memory-related processes/demands (Gibson and Wu, 2013; Levy, 2013; McCurdy and Hahn, 2024).



	Who	do	the	Gems	love
Tenure	1	7	1	2	8

(a)

Step	Parse Action
1	*CP* is conjectured
2	*CP* expands to *C'*
3	*C'* expands to *do* and *TP*
4	*TP* expands to *T'*
5	*T'* expands to *T* and *vP*
6	*vP* expands to *DP* and *v'*
7	*v'* expands to *v* and *VP*
8	*VP* expands to *love* and *who*
9	***who*** is found
10	***do*** is found
11	*DP* expands to *the* and *Gems*
12	***the*** is found
13	***Gems*** is found
14	***T*** is found
15	***v*** is found
16	***love*** is found

(b)

Figure 1: In (a): Example of an MG derivation tree for *Who do the Gems love?* with annotated parse steps as index/outdex at each node. Below it, Tenure values for pronounced lexical items computed for a node i as *Outdex*(i) – *Index*(i). Boxed nodes are those with Tenure >2. Unary branches indicate movement landing sites. In (b): Actions of a string-driven recursive descent parser for *Who do the Gems love?* as exemplified by the derivation tree in (a).

In this sense, *offline* SRC/ORC asymmetries have been extensively probed with the MG Model, with MAXT deriving the empirically reported subject advantage across languages and syntactic analyses (Graf et al., 2017; De Santo, 2021a,b; Del Valle and De Santo, 2023; Fiorini et al., 2023). Subject/Object asymmetries in RCs are then a natural venue to investigate whether structure-based complexity metrics like Tenure offer quantitative insights into online patterns of effort during sentence processing.

Thus, we use as target behavioral data the reading times (RT) for the SRC/ORC items in the Syntactic Ambiguity Processing Benchmark (SAP; Huang et al., 2024).[2] The SAP benchmark is a recent dataset of self-paced RTs from 2000 participants, covering a wide-range of complex syntactic phenomena in English. This large scale dataset has been explicitly designed in order to provide a quantitative benchmark for the evaluation of theories of sentence processing over a variety of well-studied phenomena. We focus here on the RC items in the dataset. The benchmark offers word-by-word RTs for 24 RC sets, comprising of lexically matched SRCs and ORCs taken from a classic study in the literature (Staub, 2010). Relevantly, the SAP data have already been used to probe the limited ability of expectation-based metrics (e.g., surprisal) to account for the relative difficulty of ORCs over SRCs in English.

3.2 Word-by-Word Tenure

We compute word-by-word Tenure values from derivations built for each one of the RC sentences in the benchmark. For each item, gold-standard MG derivations are built following standard generative assumptions for the main clause of each sentence, and a wh-movement analysis for the structure of RCs (Chomsky, 1977, see Figure 2). Then, derivations are annotated via Graf et al. (2017)'s implementation of Stabler (2013)'s MG parser.[3] As discussed above, Tenure is computed as $Outdex(i) - Index(i)$ for each pronounced node i in a tree (Figure 1a).

4 Model Fitting and Results

As a reminder, we want to probe whether word-by-word Tenure improves model fit to the self-paced RT data made available for English SRCs/ORCs in the SAP (Huang et al., 2024) benchmark, beyond established expectation-based predictors. Following Huang et al. (2024), in this paper we present analy-

[2] https://osf.io/b6rqh/

[3] https://github.com/CompLab-StonyBrook/mgproc

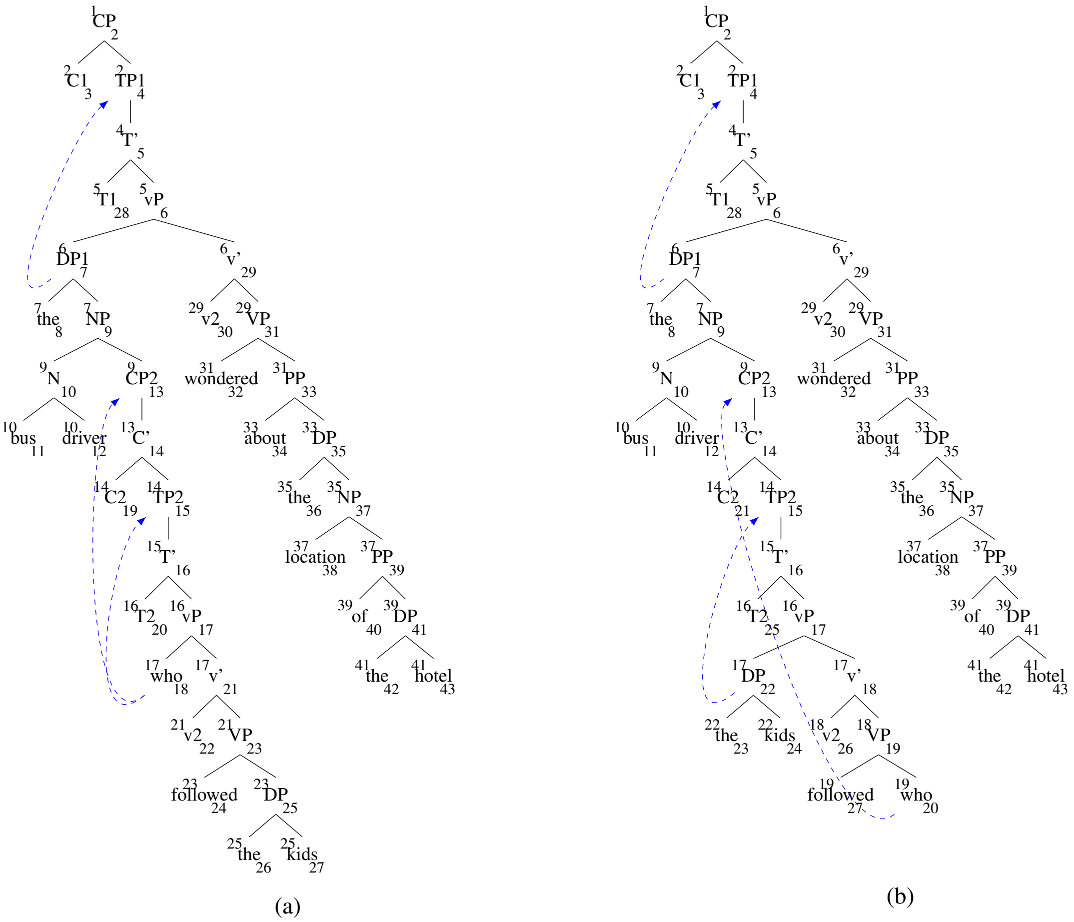


Figure 2: Annotated derivation trees for one of the subject (a) and an object (b) RCs in the dataset, modeled according to a wh-movement analysis of RCs.

ses using raw RTs, avoiding the logarithmic transformation common in the self-paced reading literature.[4] As Huang et al. (2024) argue, while this transformation reduces the right skew of RTs collected through self-paced reading, it does so by violating some theoretical assumptions about the relationship between RTs and prediction-based complexity metrics (e.g., surprisal, but also possibly Tenure).[5]

First, we fit a baseline frequentist linear mixed-effects model to the RTs, with several (scaled) lexical control predictors as computed by Huang et al. (2024):

$$\begin{aligned} RT \sim & WordPosition(i) \\ & + logfreq(i) * length(i) \\ & + logfreq(i-1) * length(i-1) \\ & + logfreq(i-1) * length(i-2) \\ & + (1|participant) + (1|item) \end{aligned}$$

[4]R scripts and data available at https://osf.io/8amqp/

[5]Analyses using log-transformed RTs are nonetheless available in our analyses scripts.

These include the position of a word in a sentence, its length and unigram frequency, and the interaction between the two. Predictors for the two preceding words are also included to account for spill-over effects common in self-paced reading (Mitchell, 2018; Vasishth, 2006).

We use surprisal as our expectation-based metric (Hale, 2006; Levy, 2008; Wilcox et al., 2023). We fit two models adding to the baseline model specified above surprisal values computed with an LSTM (Gulordava, 2018) and with GPT-2 small (Radford et al., 2019). Again, surprisal predictors are included both at the current word and at the two preceding words. We also include a random slope for surprisal by participant. Finally, we fit two models adding word-by-word Tenure (for the current word and the two preceding words) to the two surprisal models, including additional random slopes for Tenure by participant.

We select the best fitting models using AIC and BIC criteria (Akaike, 1973; Schwarz, 1978; Chakrabarti and Ghosh, 2011). Consistently with previous results, surprisal models improve fit over

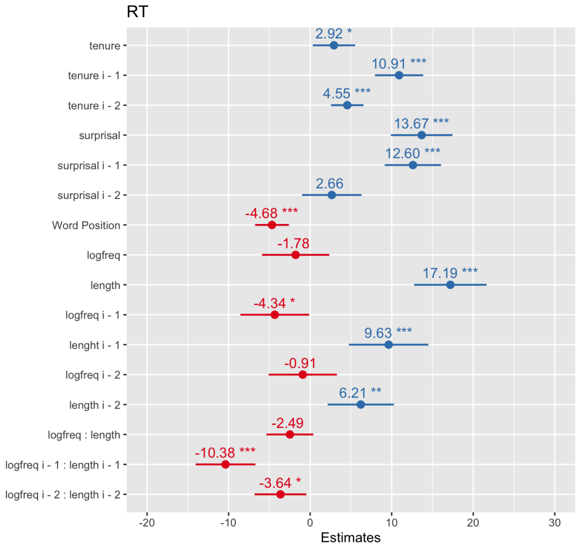


Figure 3: Estimates of coefficients for the best fitting model (GTP Surprisal + Tenure).

the baseline model (Table 1), with the GPT-2 surprisal model performs better than the LSTM model. Adding Tenure to the surprisal-only models further improves fit for both the LSTM and GPT models, showing the modeling advantage of taking memory into account explicilty. The overall best performing model was the *GPT-surprisal + Tenure* model (Table 1), consistently with GPT-2 surprisal providing a better fit than LSTM surprisal and with the structural advantage provided by Tenure. In particular, we found that Tenure of both the current word and the preceding two words is associated with significantly slower RTs independently of surprisal (Table 2 and Figure 3).

	df	**AIC**	**BIC**
Baseline	14	977122.5	977250.8
LSTM Surprisal	19	976309.1	976483.1
GPT Surprisal	19	976301.9	976475.9
LSTM Surprisal + Tenure	23	974174.8	974385.5
GPT Surprisal + Tenure	**24**	**974106.3**	**974326.2**

Table 1: Model Comparison.

5 Discussion

By combining a Minimalist grammar parser with a cognitively grounded complexity metric, the model adopted in this paper implements algorithmically theories of sentence comprehension that explicitly link comprehension difficulty to how building complex hierarchical structure affects memory usage. As discussed earlier in the paper, this approach has been successful in capturing qualitative contrasts in offline comprehension for an encouraging array of sentence processing phenomena cross-linguistically. Here, by leveraging the existing definition of Tenure, we were able to extend the evaluation of this modeling approach to quantitative word-by-word measures, providing an explicit link to the processes involved in online sentence comprehension. Importantly, Tenure does not simply measure the "raw" number of parse actions to estimate difficulty (cf. Brennan et al., 2016; Stanojević et al., 2023). It related effort to a notion of memory usage directly related to how the mismatch between the structure of the tree and the surface form of the string is navigated by the parser. By taking derivational steps seriously, Tenure ties effort to parse objects that have to be maintained "active" during the parse (e.g., partially hypothesized phrases/projections).

Our results show that predictors linking structure-building operations to memory usage improve our ability to model word-by-word RTs, beyond the contribution of expectation-based surprisal measures — adding support to the cognitive relevance of transparent structure-building measures. In particular, we found a significant positive correlation between Tenure at the current word and RTs, as well as strong effects of Tenure at the previous two words. Lingering effects of Tenure at the preceding words are compatible with known delays in RTs measured via self-paced reading. Future work could probe the plausibility of this hypothesis, and a more subtle understanding of the link between Tenure and online effort, by evaluating Tenure for similar constructions over different kinds of behavioral data (Schotter and Dillon, 2025; Boyce et al., 2020).

The recent development of broad coverage MG parsers (Torr et al., 2019) might also allow for a more fine-grained approach to the evaluation of this model's ability to capture the magnitude of the effects under study. In particular, the two-steps Bayesian approach to magnitude estimation suggested by Van Schijndel and Linzen (2021) and Huang et al. (2024) could help us leverage the modeling advantages provided by a broad coverage parser, while also retaining MGs' granular view into specific syntactic choices/details.

Similarly, building on previous offline MG results, here we only focused on the SRC/ORC asymmetry. A better understanding of the relevance of this model to theories of sentence comprehension will naturally come from evaluations over different constructions and different languages. In fact, cross-

RT						
Predictors	*Estimate*	*Std. Error*	*df*	*t value*	*Pr(>|t|)*	
(Intercept)	404.178	5.359	45.273	75.423	<2e-16	***
Tenure	2.920	1.327	3758.499	2.200	0.027899	*
Tenure $i-1$	10.907	1.507	3223.985	7.236	5.75e-13	***
Tenure $i-2$	4.553	1.018	62441.736	4.475	7.65e-06	***
Surprisal	13.675	1.924	9708.665	7.108	1.26e-12	***
Surprisal $i-1$	12.603	1.762	10126.632	7.154	9.03e-13	***
Surprisal $i-2$	2.656	1.861	59141.060	1.427	0.153489	
Word Position	-4.682	1.058	60334.657	-4.426	9.60e-06	***
logfreq	-1.782	2.102	37139.995	-0.848	0.396547	
length	17.195	2.266	22649.688	7.588	3.38e-14	***
logfreq $i-1$	-4.337	2.149	24284.605	-2.018	0.043568	*
length $i-1$	9.626	2.487	14971.417	3.871	0.000109	***
logfreq $i-2$	-0.909	2.136	46859.397	-0.425	0.670483	
length $i-2$	6.207	2.073	32905.438	2.994	0.002757	**
logfreq:length	-2.488	1.470	52063.647	-1.693	0.090503	.
logfreq $i-1$:length $i-1$	-10.378	1.871	41785.471	-5.545	2.95e-08	***
logfreq $i-2$:length $i-2$	-3.642	1.620	46877.483	-2.249	0.024533	*

$^{***}p<0.001$; $^{**}p<0.01$; $^{*}p<0.05$

Table 2: Lmer summary for the best fitting model (GTP Surprisal + Tenure).

linguistic comparison is central to the evaluation of both structure-based and expectation-based complexity metrics in cognitive modeling (Wilcox et al., 2023; Kajikawa et al., 2024). As mentioned, previous MG parsing work has proved successful in capturing the subject advantage in RCs for languages varying across several interesting structural dimensions (e.g., head-directionality, pre-nominal vs. post-nominal RCs, etc; Graf et al., 2017; De Santo, 2020b; Fiorini et al., 2023, a.o.). An investigation of this preference on cross-linguistic RT dataset would thus be a promising next step for the application of the MG model to online data.

For English specifically, the SAP benchmark offers self-paced reading data for a variety of phenomena beyond SRC/ORC contrasts (e.g., RC attachment ambiguities). Most of these phenomena involve ambiguity resolution strategies which have been used to argue in favor of single-stage, prediction based approaches — of which surprisal is one instantiation (Hale, 2001; Levy, 2013; Hale, 2016). As for the SRC advantage discussed in this paper however, surprisal has been shown unable to fully capture the magnitude of these effects within and across constructions (Van Schijndel and Linzen, 2021; Huang et al., 2024). Interesting, while this paper's model assumes a deterministic oracle and thus does not factor in ambiguity resolution explicitly, it has been shown to predict RC attachment preferences purely based on structural complexity (Lee, 2018; Lee and De Santo, 2022). More crucially, without discarding the importance of expectation/prediction in sentence comprehension, the explicit structure-building mechanisms of the MG model give us a way to implement alternative theories of ambiguity resolution — for instance two-stage approaches that consider the effort involved in structural reanalysis (Frazier and Fodor, 1978; Gorrell et al., 1995; Sturt, 1997; Pritchett, 1988; Ozaki et al., 2024).

Relatedly, the linking theory implemented by Tenure is distinct from proposals that argue for expectation-based metrics modulated/informed by syntactic structure (Demberg and Keller, 2008; Roark et al., 2009; Oh et al., 2022; Arehalli et al., 2022). As discussed, the framework described in this paper does not just argue for the relevance of syntactic structure, but for a notion of effort grounded in the direct interaction of structure building operations and memory. With this in mind, the grammar formalism adopted here is compatible with multiple ways to condition probability distributions over structural representations (Hunter and Dyer, 2013; Torr et al., 2019). Because of this, the MG approach is also flexible enough to allow for the exploration of potentially complex interactions of memory, structure, and expectation beyond the simple computation of structure-informed metrics like surprisal (Futrell et al., 2020; Brennan et al., 2020; Chen and Hale, 2021).

More generally, deeper insights into the contribution of structure-building metrics to models of sentence comprehension will come from a broader comparison between Tenure and other memory-based metrics (Kaplan, 1975; Pulman, 1986; Kaplan, 2020; Gibson, 1998; Lewis et al., 2006; Boston, 2012). For instance, an informative next step in this enterprise would be to conduct an empirical evaluation of the different predictions made by Tenure and a complexity metric like Node Count, which counts the number of syntactic operations in a tree (Brennan et al., 2016; Nelson et al., 2017; Brennan et al., 2020; Li and Hale, 2019; Stanojević et al., 2023, 2021; Kajikawa et al., 2024). It would also be fruitful to compare our results to measures of memory load relying less on rich structural information (e.g., Dependency Locality Theory; Gibson, 1998).

Similarly, through the use of MGs this work has committed explicitly to syntactic representations as hypothesized by modern generative syntax. While we made the case that the computation of particular Tenure values is deeply tied to commitments about the shape of a syntactic derivation *and* the timing of how such a derivation is built, its definition is conceptually independent of specific representational/algorithmic choices. Therefore, Tenure could be ideal for a comparison of the behavioral predictions made by different (often expressively equivalent) syntactic formalisms such as, for instance, TAG and CCG (Demberg et al., 2013; Stanojević et al., 2023, a.o.).

Relatedly, among this approach's degrees of freedom is the tree-traversal strategy adopted by the parser. This paper has followed the majority of offline MG work in extracting Tenure by evaluating the stack-usage of a top-down parser. Whether similar, or better, modeling results could be derived via different parsing strategies is thus an open question (cf. Brennan et al., 2016; Stanojević et al., 2023). In this sense, left-corner parsing algorithms have been recently proposed for MGs, and have been shown to correctly capture some interesting offline processing contrasts (Hunter, 2019; Hunter et al., 2019; Liu, 2024). Left-corner parsing's combination of top-down prediction and bottom-up "greedy" integration has also independently been argued to be more plausible as a description of human comprehension processes (Resnik, 1992). Crucially, the complex status of a parse item in Liu (2024)'s implementation of Hunter et al. (2019)'s left-corner MG parser makes adapting a word-by-word definition of Tenure non-trivial. Working out what the exact computation of online Tenure over the stack items stored by Hunter et al. (2019)'s parser would thus be the essential next step to perform this type of comparisons.

Finally, the model's sensitivity to fine-grained grammatical assumptions implies that analytical choices have a significant impact on the derived Tenure values. Conscious of this feature of the model, in this paper we have committed to one syntactic analysis for the main construction of interest. However, previous offline work has shown that alternative analyses of RCs might result in different behavioral predictions, especially when evaluated cross-linguistically (Graf et al., 2017; De Santo and Shafiei, 2019; Lee and De Santo, 2022). In this sense, the granularity of online data and the clear linking hypothesis implemented by the MG model could contribute to psycholinguistic data (and theories) bringing insights into the evaluation of analyses in theoretical syntax (Rambow and Joshi, 1994; Kobele et al., 2013; De Santo and Lee, 2022; Prasad and Linzen, 2024). Future work could then exploit online behavioral data to distinguish competing syntactic proposals based on their psycholinguistic predictions, thus clarifying how/which aspects of sentence structure modulate processing difficulty.

6 Conclusion

Extending previous work on offline contrasts, this paper provides a first evaluation of a parser for Minimalist grammars and a memory-based complexity metric over word-by-word behavioral data. While previous work in this domain evaluated offline behavior qualitatively, we provide quantitative evidence for the success of the approach by showing that the MG-based metric Tenure is a strong predictor of SRC/ORC RTs from a large scale behavioral dataset, independently of expectation-based surprisal. While many questions remain open, these results strengthen previous offline work arguing for relevance of the combination of MGs and Tenure in investigating the interaction of generative syntax and psycholinguistic results. Furthermore, they provide additional support to a growing body of computational modeling work arguing for the role of structure-building operations in developing plausible cognitive models of human sentence comprehension.

References

H Akaike. 1973. nformation theory and an extension of the maximum likelihood principle. In *Proceedings of the Second International Symposium on Information Theory. BN Petrov and F. Cs' aki, editors. Akademiai Ki'ado, Budapest*.

Suhas Arehalli, Brian Dillon, and Tal Linzen. 2022. Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities. In *26th Conference on Computational Natural Language Learning, CoNLL 2022 collocated and co-organized with EMNLP 2022*, pages 301–313. Association for Computational Linguistics (ACL).

Robert C. Berwick and Amy S. Weinberg. 1982. Parsing efficiency, computational complexity, and the evaluation of grammatical theories. *Linguistic Inquiry*, 13:165–291.

Robert C. Berwick and Amy S. Weinberg. 1983. The role of grammar in models of language use. *Cognition*, 13:1–61.

Marisa Ferrara Boston. 2012. *A COMPUTATIONAL MODEL OF COGNITIVE CONSTRAINTS IN SYNTACTIC LOCALITY*. Ph.D. thesis, Cornell University.

Veronica Boyce, Richard Futrell, and Roger P Levy. 2020. Maze made easy: Better and easier measurement of incremental processing difficulty. *Journal of Memory and Language*, 111:104082.

Jonathan R Brennan, Chris Dyer, Adhiguna Kuncoro, and John T Hale. 2020. Localizing syntactic predictions using recurrent neural network grammars. *Neuropsychologia*, 146:107479.

Jonathan R Brennan, Edward P Stabler, Sarah E Van Wagenen, Wen-Ming Luh, and John T Hale. 2016. Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language*, 157:81–94.

Joan Bresnan. 1978. A realistic transformational grammar. *Linguistic theory and psychological reality*, pages 1–59.

Arijit Chakrabarti and Jayanta K. Ghosh. 2011. Aic, bic and recent advances in model selection. In Prasanta S. Bandyopadhyay and Malcolm R. Forster, editors, *Philosophy of Statistics*, volume 7 of *Handbook of the Philosophy of Science*, pages 583–605. North-Holland, Amsterdam.

Craig G Chambers, Michael K Tanenhaus, and James S Magnuson. 2004. Actions and affordances in syntactic ambiguity resolution. *Journal of experimental psychology: Learning, memory, and cognition*, 30(3):687.

Zhong Chen and John T Hale. 2021. Quantifying structural and non-structural expectations in relative clause processing. *Cognitive Science*, 45(1):e12927.

Noam Chomsky. 1977. On wh-movement.

Noam Chomsky. 1995. *The Minimalist Program*. Cambridge, MA: MIT Press.

Aniello De Santo. 2019. Testing a Minimalist grammar parser on Italian relative clause asymmetries. In *Proceedings of the ACL Workshop on Cognitive Modeling and Computational Linguistics (CMCL) 2019*, June 6 2019, Minneapolis, Minnesota.

Aniello De Santo. 2020a. Mg parsing as a model of gradient acceptability in syntactic islands. *Proceedings of the Society for Computation in Linguistics*, 3(1):53–63.

Aniello De Santo. 2020b. *Structure and memory: A computational model of storage, gradience, and priming*. Ph.D. thesis, State University of New York at Stony Brook.

Aniello De Santo. 2021a. Italian postverbal subjects from a minimalist parsing perspective. *Lingue e linguaggio*, 20(2):199–227.

Aniello De Santo. 2021b. A minimalist approach to facilitatory effects in stacked relative clauses. *Proceedings of the Society for Computation in Linguistics*, 4(1):1–17.

Aniello De Santo and So Young Lee. 2022. Evaluating structural economy claims in relative clause attachment. In *Proceedings of the Society for Computation in Linguistics 2022*, pages 65–75.

Aniello De Santo and Nazila Shafiei. 2019. On the structure of relative clauses in Persian: Evidence from computational modeling and processing effects. In *Talk at the 2nd North American Conference in Iranian Linguistics (NACIL2)*, April 19-21 2019, University of Arizona.

Daniel Del Valle and Aniello De Santo. 2023. Processing french rcs with postverbal subjects in a minimalist parser. *Society for Computation in Linguistics*, 6(1).

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Vera Demberg and Frank Keller. 2009. A computational model of prediction in human parsing: Unifying locality and surprisal effects. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 31.

Vera Demberg, Frank Keller, and Alexander Koller. 2013. Incremental, predictive parsing with psycholinguistically motivated tree-adjoining grammar. *Computational Linguistics*, 39(4):1025–1066.

Nick C Ellis. 2002. Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in second language acquisition*, 24(2):143–188.

Matteo Fiorini, Jillian Chang, and Aniello De Santo. 2023. An mg parsing view into the processing of subject and object relative clauses in basque. In *Proceedings of the Society for Computation in Linguistics 2023*, pages 145–154.

Jerry A. Fodor and Merrill Garrett. 1967. Some syntactic determinants of sentential complexity. *Perception and Psychophysics*, 2:289–296.

Lyn Frazier and Janet Dean Fodor. 1978. The sausage machine: A new two-stage parsing model. *Cognition*, 6(4):291–325.

Richard Futrell, Edward Gibson, and Roger P Levy. 2020. Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive science*, 44(3):e12814.

Sabrina Gerth. 2015. *Memory Limitations in Sentence Comprehension: A Structural-based Complexity Metric of Processing Difficulty*, volume 6. Universitätsverlag Potsdam.

E Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. *Image, Language, Brain*, pages 95–126.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Edward Gibson and H-H Iris Wu. 2013. Processing chinese relative clauses in context. *Language and Cognitive Processes*, 28(1-2):125–155.

Paul Gorrell et al. 1995. *Syntax and parsing*, volume 76. Cambridge University Press Cambridge.

Thomas Graf and Bradley Marcinek. 2014. Evaluating evaluation metrics for minimalist parsing. In *Proceedings of the 2014 ACL Workshop on Cognitive Modeling and Computational Linguistics*, pages 28–36.

Thomas Graf, James Monette, and Chong Zhang. 2017. Relative clauses as a benchmark for Minimalist parsing. volume 5, pages 57–106.

K Gulordava. 2018. Colorless green recurrent networks dream hierarchically. *arXiv preprint arXiv:1803.11138*.

John Hale. 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*.

John Hale. 2006. Uncertainty about the rest of the sentence. *Cognitive Science*, 30(4):643–672.

John Hale. 2011. What a rational parser would do. *Cognitive Science*, 35(3):399–443.

John Hale. 2016. Information-theoretical complexity metrics. *Language and Linguistics Compass*, 10(9):397–412.

Henk Harkema. 2001. A characterization of minimalist languages. In *International Conference on Logical Aspects of Computational Linguistics*, pages 193–211. Springer.

Kuan-Jung Huang, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen. 2024. Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, 137:104510.

Tim Hunter. 2019. Left-corner parsing of minimalist grammars. *Minimalist parsing*, pages 125–158.

Tim Hunter and Chris Dyer. 2013. Distributions on minimalist grammar derivations. In *Proceedings of the 13th Meeting on the Mathematics of Language (MoL 13)*, pages 1–11.

Tim Hunter, Milos Stanojevic, and Edward Stabler. 2019. The active-filler strategy in a move-eager left-corner minimalist grammar parser. In *Cognitive Modeling and Computational Linguistics*, pages 1–10. Association for Computational Linguistics.

Kohei Kajikawa, Ryo Yoshida, and Yohei Oseki. 2024. Dissociating syntactic operations via composition count. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

R Kaplan and Joan Bresnan. 1982. Grammars as mental representations of language. *The mental representation of grammatical relations, ed. Bresnan J.. MIT Press.[rEPS]*.

Ronald M Kaplan. 1975. *Transient processing load in relative clauses*. Ph.D. thesis, Harvard University.

Ronald M Kaplan. 2020. Computational psycholinguistics. *Computational linguistics*, 45(4):607–626.

Gregory M Kobele, Sabrina Gerth, and John Hale. 2013. Memory resource allocation in top-down minimalist parsing. In *Formal Grammar*, pages 32–51. Springer.

Gregory M. Kobele, Christian Retoré, and Sylvain Salvati. 2007. An automata-theoretic approach to minimalism. In *Model Theoretic Syntax at 10*, pages 73–82. J. Rogers and S. Kepser.

Elaine Lau and Nozomi Tanaka. 2021. The subject advantage in relative clauses: A review. *Glossa: a journal of general linguistics*, 6(1).

So Young Lee. 2018. A minimalist parsing account of attachment ambiguity in English and Korean. *Journal of Cognitive Science*, 19(3):291–329.

So Young Lee and Aniello De Santo. 2022. A computational view into the structure of attachment ambiguities in chinese and korean. In *Proceedings of NELS*, volume 52.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Roger Levy. 2013. Memory and surprisal in human sentence comprehension. In *Sentence processing*, pages 90–126. Psychology Press.

Richard L Lewis, Shravan Vasishth, and Julie A Van Dyke. 2006. Computational principles of working memory in sentence comprehension. *Trends in cognitive sciences*, 10(10):447–454.

Jixing Li and John Hale. 2019. Grammatical predictors for fmri time-courses. *Minimalist parsing*, pages 159–173.

Lei Liu. 2018. Minimalist Parsing of Heavy NP Shift. In *Proceedings of PACLIC 32 The 32nd Pacific Asia Conference on Language, Information and Computation*, The Hong Kong Polytechnic University, Hong Kong SAR.

Lei Liu. 2024. Psycholinguistic adequacy of left-corner parsing for minimalist grammars. *Proceedings of the Society for Computation in Linguistics (SCiL)*, pages 275–280.

William Marslen-Wilson and Lorraine Komisarjevsky Tyler. 1980. The temporal structure of spoken language understanding. *Cognition*, 8(1):1–71.

Kate McCurdy and Michael Hahn. 2024. Lossy context surprisal predicts task-dependent patterns in relative clause processing. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 36–45.

Jens Michaelis. 1998. Derivational minimalism is mildly context–sensitive. In *International Conference on Logical Aspects of Computational Linguistics*, pages 179–198. Springer.

George A. Miller and Noam Chomsky. 1963. Finitary models of language users. In R. Luce, R. Bush, and E. Galanter, editors, *Handbook of Mathematical Psychology*, volume 2. John Wiley, New York.

Don C Mitchell. 2018. An evaluation of subject-paced reading tasks and other methods for investigating immediate processes in reading 1. In *New methods in reading comprehension research*, pages 69–90. Routledge.

Matthew J Nelson, Imen El Karoui, Kristof Giber, Xiaofang Yang, Laurent Cohen, Hilda Koopman, Sydney S Cash, Lionel Naccache, John T Hale, Christophe Pallier, et al. 2017. Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences*, 114(18):E3669–E3678.

Byung-Doh Oh, Christian Clark, and William Schuler. 2022. Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, 5:777963.

Satoru Ozaki, Aniello De Santo, Tal Linzen, and Brian Dillon. 2024. Ccg parsing effort and surprisal jointly predict rt but underpredict garden-path effects. *Proceedings of the Society for Computation in Linguistics (SCiL)*, pages 362–364.

Colin Phillips. 2003. Parsing: Psycholinguistic aspects. In *International Encyclopedia of Linguistics*, 2 edition. Oxford University Press.

Grusha Prasad and Tal Linzen. 2024. Spawning structural priming predictions from a cognitively motivated parser. *arXiv preprint arXiv:2403.07202*.

Bradley L Pritchett. 1988. Garden path phenomena and the grammatical basis of language processing. *Language*, pages 539–576.

Steven G Pulman. 1986. Grammars, parsers, and memory limitations. *Language and Cognitive processes*, 1(3):197–225.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Owen Rambow and Aravind K Joshi. 1994. A processing model for free word order languages. *Perspectives on Sentence Processing*.

Philip Resnik. 1992. Left-corner parsing and psychological plausibility. In *COLING 1992 Volume 1: The 14th International Conference on Computational Linguistics*.

Brian Roark, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 324–333.

Elizabeth R Schotter and Brian Dillon. 2025. A beginner's guide to eye tracking for psycholinguistic studies of reading. *Behavior Research Methods*, 57(2):68.

Gideon Schwarz. 1978. Estimating the dimension of a model. *The annals of statistics*, pages 461–464.

Edward P Stabler. 1996. Derivational minimalism. In *International Conference on Logical Aspects of Computational Linguistics*, pages 68–95. Springer.

Edward P Stabler. 2013. Two models of minimalist, incremental syntactic analysis. *Topics in cognitive science*, 5(3):611–633.

Miloš Stanojević, Shohini Bhattasali, Donald Dunagan, Luca Campanelli, Mark Steedman, Jonathan Brennan, and John Hale. 2021. Modeling incremental language comprehension in the brain with combinatory categorial grammar. In *Proceedings of the workshop on cognitive modeling and computational linguistics*, pages 23–38.

Miloš Stanojević, Jonathan R Brennan, Donald Dunagan, Mark Steedman, and John T Hale. 2023. Modeling structure-building in the brain with ccg parsing and large language models. *Cognitive science*, 47(7):e13312.

Adrian Staub. 2010. Eye movements and processing difficulty in object relative clauses. *Cognition*, 116(1):71–86.

Patrick Sturt. 1997. Syntactic reanalysis in human language processing.

Michael K Tanenhaus, Michael J Spivey-Knowlton, Kathleen M Eberhard, and Julie C Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632–1634.

John Torr, Milos Stanojevic, Mark Steedman, and Shay Cohen. 2019. Wide-coverage neural a* parsing for minimalist grammars. In *57th Annual Meeting of the Association for Computational Linguistics*, pages 2486–2505. ACL Anthology.

Matthew J Traxler and Martin J Pickering. 1996. Plausibility and the processing of unbounded dependencies: An eye-tracking study. *Journal of Memory and Language*, 35(3):454–475.

Marten Van Schijndel and Tal Linzen. 2021. Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive science*, 45(6):e12988.

Pranali Vani, Ethan Gotlieb Wilcox, and Roger Levy. 2021. Using the interpolated maze task to assess incremental processing in english relative clauses. In *Proceedings of the annual meeting of the cognitive science society*, volume 43.

Shravan Vasishth. 2006. On the proper treatment of spillover in real-time reading studies: Consequences for psycholinguistic theories. In *Proceedings of the international conference on linguistic evidence*, pages 96–100.

Matthew W Wagers and Colin Phillips. 2009. Multiple dependencies and the role of the grammar in real-time comprehension. *Journal of Linguistics*, 45(2):395–433.

Ethan G Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.