MG Parsing as a Model of Effort in Online RC Processing

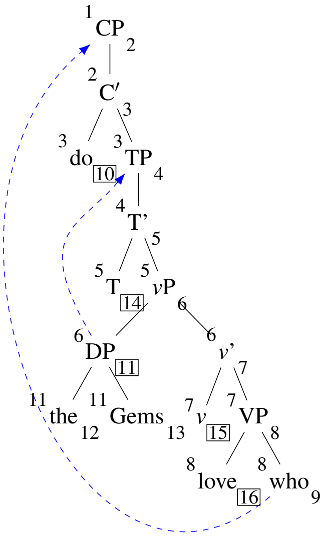
Aniello De Santo
Dept. of Linguistics, University of Utah
aniello.desanto@utah.edu

Overview. Past research has shown that a top-down parser for Minimalist grammars [MGs; 1] captures sentence processing preferences across an array of languages and phenomena, when combined with complexity metrics connecting its behavior to memory usage [2, 3, 4, a.o.]. This approach (henceforth: **MG Model**) helps probe the link between generative syntactic theory and sentence processing, by offering a fully-specified theory of how fine-grained grammatical structure affects cognitive cost. While work in this framework has focused on modeling *off-line* asymmetries, here we show how measures of effort that explicitly consider minimalist-like structure-building operations can account for word-by-word (*online*) behavioral data.

The MG Model. We adopt a model linking structural details to processing load by associating the stack states of a (deterministic) top-down parser [1] to memory burden [5]. This parser is *string-driven*: when encountering a displaced word (e.g., "*who*"), it prioritizes finding a path to its base position. Here, we measure memory usage based on how long a node is kept in memory through a derivation, tracking how the derivational operations interact with fine-grained structural details to affect linear word order (*Tenure*). The annotation schema of Fig. 1 captures how the parser's tree traversal strategy affects memory: the superscript (index) of a node n encodes the moment n was predicted and put in memory. The subscript (outdex) encodes the moment n is confirmed and frees up memory. Tenure for n is $outdex(n) - index(n)$: e.g. *Tenure*$($*do*$) = 10 - 3 = 7$. While past work has leveraged offline metrics estimating effort for a full derivation, it is straightforward to derive online measures by extracting Tenure values for every (pronounced) lexical item (Fig. 1).

Evaluating Tenure Online. Offline subject/object relative clause (SRC/ORC) asymmetries have been extensively probed with the MG Model [3, 4]. Because of this, we ask whether structure-building effort as captured by Tenure improves model fit to the self-paced reading data made available for English SRCs/ORCs in the Syntactic Ambiguity Processing Benchmark [6], beyond established expectation-based predictors. First, we fit a baseline linear mixed-effects model to the RTs, with several lexical control predictors as in [6]. We then add to the baseline model surprisal predictors, fitting two models with surprisal values derived either from an LSTM [7] or GPT-2 small [8]. Then, we add to the baseline model word-by-word Tenure values computed via the MG model for each RC item in the benchmark. The MG trees follow standard generative assumptions for the main clause of each sentence, and a wh-movement analysis for the structure of RCs [9]. Finally, we fit two models adding these MG Tenure values to the two surprisal models. We found that the Tenure-only model outperforms both surprisal-only models, and that best-fit comes from the *GPT-surprisal* + *Tenure* model (Table 1). Taking Tenure into account significantly improves model fit to RT data, as we found Tenure (of both the current word and the preceding two words) associated with significantly slower RTs independently of surprisal (Table 2).

Discussion. Our results show that predictors explicitly sensitive to structure building models of word-by-word RTs, beyond the contribution of surprisal measures — providing support to a growing body of computational work arguing for the role of structure-building operations in developing plausible cognitive models of human sentence comprehension, and to the use of the MG Model in investigating the interaction of generative syntax and human sentence processing. As the model's sensitivity to grammatical assumptions implies that analytical choices have a significant impact on the derived Tenure values, future work could exploit online behavioral data to distinguish competing syntactic proposals (and different syntactic formalisms) based on their psycholinguistic predictions, thus clarifying how/which aspects of sentence structure modulate processing difficulty.



	Who	do	the	Gems	love
Tenure	1	7	1	2	8

Figure 1: Example of an MG derivation tree for *Who do the Gems love?* with annotated parse steps, and tenure values for pronounced lexical items. Unary branches indicate movement landing sites.

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

	df	**AIC**	**BIC**
Baseline	14	977122.5	977250.8
+ LSTM Surprisal	19	976309.1	976483.1
+ GPT-2 Small Surprisal	19	976301.9	976475.9
+ **Tenure**	**19**	**974413.7**	**974587.7**
+ LSTM Surprisal + Tenure	23	974174.8	974385.5
+ **GPT Surprisal + Tenure**	**24**	**974106.3**	**974326.2**

Table 2: Model Comparison.

[1] Stabler, E. P. (n.d.). Two models of minimalist, incremental syntactic analysis. *Topics in cognitive science*.

[2] Gerth, S. (2015). *Memory limitations in sentence comprehension: A structural-based complexity metric of processing difficulty* (Vol. 6). Universitätsverlag Potsdam.

[3] Graf, T., Monette, J., & Zhang, C. (2017). Relative clauses as a benchmark for Minimalist parsing. *Journal of Language Modelling*, *5*, 57–106. https://doi.org/10.15398/jlm.v5i1.157

[4] De Santo, A. (2020). *Structure and memory: A computational model of storage, gradience, and priming*.

[5] Kobele, G. M., Gerth, S., & Hale, J. (2013). Memory resource allocation in top-down minimalist parsing.

[6] Huang, K.-J., Arehalli, S., Kugemoto, M., Muxica, C., Prasad, G., Dillon, B., & Linzen, T. (2024). Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, *137*, 104510.

[7] Gulordava, K. (2018). Colorless green recurrent networks dream hierarchically. *arXiv preprint arXiv:1803.11138*.

[8] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

[9] Chomsky, N. (1977). On wh-movement.