Toward a broad-coverage implementation of self-organization for sentence comprehension

Garrett Smith (University of Potsdam)
gasmith@uni-potsdam.de

Self-organization—the idea that a parse of a whole sentence can arise solely through interactions between pairs of words—has been floating around in the sentence processing literature for at least 30 years [2]. However, previous implementations of this idea have had a number of shortcomings that have prevented broad-coverage testing of the theory. Typically, self-organization-based models face one or more of the following critiques: 1) many (usually hand-tuned) free parameters [2, 6, 7, 9], 2) unreported or incorrect predictions for incremental processing time effects [2, 7], 3) inability to handle more than one or two sentence types for a single set of parameters [6, 7, 9]. The model presented here is designed to implement word-by-word self-organization of syntactic structure in competent native speakers while addressing these concerns and paving the way for future, large-scale comparisons with other approaches.

High-level description The input to the model is a string of words to process, and the outputs are processing times for each word along with a history of configurations (dependency parses) visited during processing. The model uses a lexicalized grammar consisting of word-word dependency grammar relations and depends on three free parameters that are required by the mathematical formalism. For each word in the input, the model explores the space of configurations that are possible given the input up to that word. The amount of time it takes to sufficiently explore the configurations depends on the input string, which leads to different processing time predictions for different constructions.

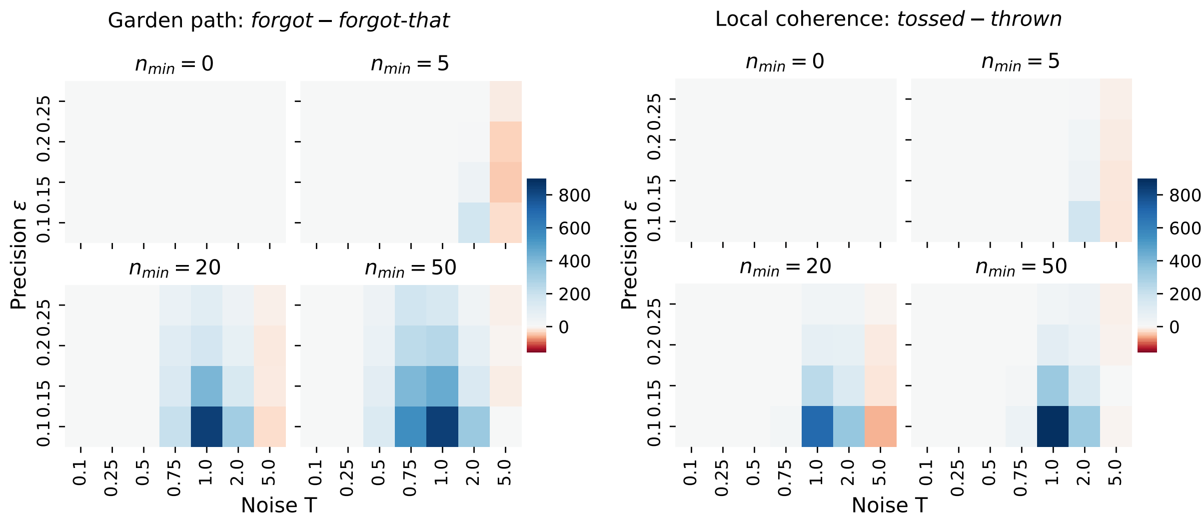
The model was tested on garden paths [e.g., 12], local coherence effects [10], and the ambiguity advantage [11] in English; see Table. 1 for materials. As shown in Fig. 1, it correctly predicts the direction of incremental processing time effects for all three constructions, i.e., slowdowns for garden paths and local coherence, speedup for ambiguity advantage. Importantly, a sensitivity analysis showed that there is a range of parameter settings where all three effects are predicted simultaneously. Thus, this implementation of self-organization improves on previous ones by having few free parameters whose values were not cherrypicked, making correct processing time predictions, and handling multiple effects in the same parameter range.

Implementation details The model uses the Metropolis algorithm [Eq. 2; 4] to stochastically explore configurations at each word. At each time step, a new configuration is proposed by either adding a new attachment link or removing an existing one. The probability of transitioning from the current configuration to the proposed one is determined by the difference in harmonies and the noise T (a free parameter controlling the model's preference for well-formed over ill-formed configurations), with transitions to better configurations more probable than to worse ones. The harmony of a configuration is the sum of the association strengths (a measure of well-formedness) of all word-word dependencies minus a penalty against partial parses, i.e., having fewer than $w-1$ dependencies in a string of w words (Eq. 1). Here, each word-word dependency in the grammar has an association strength of 1.0, but future work will learn these values from parsed corpora. Transitions between configurations continue for at least n_{min} time steps (a free parameter controlling how eager the model is to move to the next word) and until the average harmony of the possible configurations can be estimated with a precision of $\pm\varepsilon$ (the final free parameter, controlling how extensively the model explores configurations) [1]. Once this criterion is met, the next word is read in, and the system resumes exploring the now expanded space of configurations.

Outlook Future work will address whether the model can correctly predict the magnitude of effects in addition to the direction, which will allow for quantitative comparisons with other theories of sentence comprehension. Finally, this implementation is simple enough to make broad-coverage tests possible. The longer-term goal of this project is an online interface where one can enter an arbitrary sentence and get processing time predictions for each word.

Table 1: Materials and dependency grammar. Simulated words are underlined. Arrows point from governors to dependents. The implemented grammar consists of the union of the binary dependency relations in the third column. The final column gives the direction of the processing time effect at the final simulated word from the cited study.

Construction	Sentence	Dependencies	Reading times
Garden path	The student forgot [that] the solution was in the book.	forgot → solution [that-] solution ← was forgot → was	*forgot* > *forgot that* [12]
Local coherence	The coach smiled at the player tossed/thrown the frisbee.	smiled-at → player player → tossed/thrown player ← tossed	*tossed* > *thrown* [10]
Ambiguity advantage	The driver/car of the car/driver with a mustache... The son of the driver with a mustache...	driver→ of-car car → of-driver [of-] driver → mustache son → of-driver son → mustache	*son* < other conditions [11]



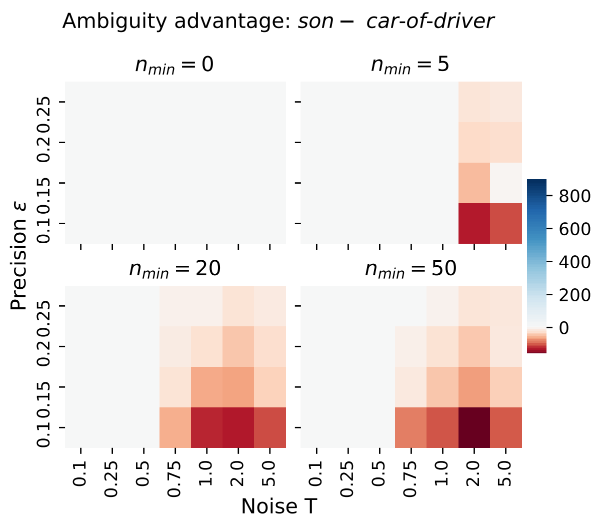


Figure 1. Sensitivity analysis results. Plotted effect sizes (color coding) are the median of the differences in processing time (arbitrary units) between test and control conditions at the final word for each construction. Deeper blues indicate a bigger slowdown, and deeper reds indicate a bigger speedup. 500 model runs per parameter combination.

Harmony function The harmony H of a configuration—a continuous measure of wellformedness [8]—is given by

$$H = \overbrace{-|n_{links} - w + 1|}^{\text{Penalty for not having } w-1 \text{ links}} + \underbrace{\sum_{l \in links} assoc(l)}_{\text{Sum of association strengths across all links}} \tag{1}$$

where n_{links} is the number of links in w-word configuration and $assoc(\cdot)$ is the strength of association between for link l. In future work the association strength will be estimated from parsed corpora; here, every link in the grammar has an association strength of 1.0.

Metropolis transition probability The probability of transitioning from the current state to the new, proposed state is given by the Metropolis probability [4]:

$$P(\text{transition}) = \min\left(1, \exp\left(-T^{-1}(H_{\text{current}} - H_{\text{proposed}})\right)\right) \tag{2}$$

where the H_* are the harmonies of the relevant configurations and T is the noise parameter. The higher the noise, the more likely a jump to the proposed configuration.

Sensitivity analysis All three free parameters were varied to determine the full range model predictions [5]. As shown in Fig. 1, the model either predicts effects in the same direction as human participants or no difference. Predictions in the wrong direction only appear when the noise is set so high that the model effectively disregards all grammatical constraints.

Measuring time To reduce correlations between consecutive Monte Carlo samples, time is measured in Monte Carlo sweeps [MCS; 3]: Here, one MCS is equal to w^{w-1} transitions. This is the number of complete, unlabelled dependency parses for sentence of w words. The stopping criterion is checked at the end of every MCS, allowing enough time (in principle) to visit each possible complete parse one time.

Stopping criterion The model uses the fixed-width Monte Carlo stopping criterion of [1] to determine when to stop processing a word. Processing continues until the half-width of the 95% confidence interval around the average harmony of visited configurations drops below a preset level ε, a free parameter. However, at least n_{min} MCS must pass before the half-width ε is checked. The smaller ε becomes, the more thoroughly the system explores configurations after at least n_{min} MCS have passed.

[1] Jones, Haran, Caffo, and Neath, 2006. *J. Am. Stat. Assoc.*, 101(476):1537–1547.

[2] Kempen and Vosse, 1989. *Connection Science*, 1(3):273–290.

[3] Landau and Binder, 2009. *A guide to Monte-Carlo simulations in statistical physics*. Cambridge University Press, 3 edition.

[4] Metropolis et al., 1953. *J. Chem. Phys.*, 21(6):1087–1092.

[5] Roberts and Pashler, 2000. *Psychological Review*, 107(2):358–367.

[6] Smith and Tabor, 2018. In Juvina, Houpt, and Myers, editors, *Proc. of the 16th ICCM*.

[7] Smith, Franck, and Tabor, 2018. *Cognitive Science*, 42(S4):1043–1074.

[8] Smolensky, 1986. In Rumelhart, McClelland, and the PDP Research Group, editors, *PDP*, volume I: Foundations. MIT Press.

[9] Tabor and Hutchins, 2004. *JEP: LMC*, 30(2):431–450.

[10] Tabor, Galantucci, and Richardson, 2004. *JML*, 50(4):355–370.

[11] Traxler, Pickering, and Clifton, 1998. *JML*, 39:558–592.

[12] Trueswell, Tanenhaus, and Kello, 1993. *JEP: LMC*, 19(3):528–553.