PNAS

LETTER

Another model not for the learning of language

Jordan Kodner[a,1], Spencer Caplan[b], and Charles Yang[c]

It is laudable for Yang and Piantadosi (ref. 1, henceforth YP) to tackle the important question of how language is generated and learned. Their reported success in learning a range of grammars from small amounts of data are striking, especially since some of these grammars belong to classes which are provably unlearnable under such conditions (2). However, their project is severely undermined by their evaluation method and by their conception of language.

Following standard machine-learning practice, models are trained and tested on disjoint sets of data so that their generalization efficacy can be accurately assessed. YP, however, tested their model on training data, and moreover only on its top 25 most probable strings. This cuts against the very heart of learning: A successful grammar must extend to novel (and rare) sentences and must also reject sentences not generated by the grammar. Under YP's scheme, a grammar that massively over/undergenerates would nevertheless be deemed successful as long as its 25 most probable strings match the training target. Even granting these unconventional methodological choices, YP's model fails to learn natural and artificial languages that humans easily learn. Fig. 1 displays the results of an n-gram model, a demonstrably inadequate model for human language (3), on the same data. It can appear successful—sometimes more successful than YP—under their evaluation method: a *reductio ad absurdum*.

YP's conception of language as strings fails to recognize that language has internal structures not manifested by the surface sequence of words. For example, the sentence "They are flying airplanes" has two meanings: An English speaker can assign two distinct structures and thus interpretations to the same string (3). Some of the successful grammars (by YP's metric) accept a similar set of strings but assign very different structures to them (see supporting information in ref. 1). For example, the grammars obtained by training on a set of toy English sentences (figure 5 of ref. 1) show no consistency on sentences not used for training. These results are strikingly at odds with human language learning: Children form grammars with "a high degree of uniformity in both the categorical and variable aspects" (4) as revealed by the quantitative study of language use.

YP's model transforms candidate hypotheses into probabilistic context-free grammars that are evaluated against the training data via Bayesian inference. This method was proposed by Horning over 50 y ago (5). However, as Horning himself noted, the method is neither psychologically plausible nor computationally practical, as it needs to enumerate and evaluate an astronomically large space of grammars. What is novel to YP is the advancement in computing hardware: Intractable solutions can now be approximated even though the model still failed to learn a 35-sentence fragment of English (6) after 7 d (supporting information in ref. 1). Hence, we surmise, YP's disclaimer that "we do not claim that they [humans] necessarily use the same methods as our implementation" (ref. 1, p. 9). But if a model for the learning of language is not for understanding how language is actually learned, what is it for?

Author affiliations: [a]Stony Brook University, Stony Brook, NY 11794; [b]Swarthmore College, Swarthmore, PA 19081; and [c]University of Pennsylvania, Philadelphia, PA 19104

Author contributions: J.K., S.C., and C.Y. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no competing interest.

[1] To whom correspondence may be addressed. Email: jordan.kodner@stonybrook.edu.

Published July 5, 2022.

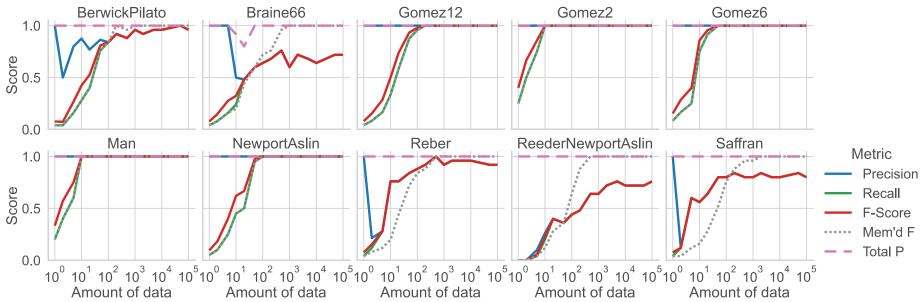

Fig. 1. The apparent success of an inadequate model: Performance of a trigram model on natural and artificial languages from the published literature; see figure 2 of ref. 1 for references. Based on YP's evaluation scheme, the trigram model performs well on several of the languages, even outperforming YP on some (e.g., ref. 6). "Total P" measures performance of the model against the entire string set rather than just the top 25, which more accurately corresponds to grammar learning. This is clearest for Saffran, which was designed to be learnable by a bigram model. The n-gram model learns it perfectly but is unfairly penalized under YP's evaluation because it generates strings that are in the language but outside the 25 most probable ones. All language data are from https://github.com/piantado/Fleet.

1. Y. Yang, S. T. Piantadosi, One model for the learning of language. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2021865119 (2022).
2. M. Kearns, L. Valiant, Cryptographic limitations on learning Boolean formulae and finite automata. *J. Assoc. Comput. Mach.* **41**, 67–95 (1994).
3. N. Chomsky, Three models for the description of language. *IRE Trans. Inf. Theory* **2**, 113–124 (1956).
4. W. Labov, (2014) "What is to be learned" in *Cognitive Sociolinguistics: Social and Cultural Variation in Cognition and Language Use*, M. Pütz, J. A. Robinson, M. Reif, Eds. (John Benjamins, Amsterdam), pp. 23–51.
5. J. J. Horning, "A study of grammatical inference," PhD thesis, Stanford University, Stanford, CA (1969).
6. R. C. Berwick, S. Pilato, Learning syntax by automata induction. *Mach. Learn.* **2**, 9–38 (1987).