Comparing Symbolic Models of Language via Bayesian Inference (Student Abstract)

Annika Heuser,[1] **Polina Tsvilodub**[2]

[1] Departments of Brain & Cognitive Sciences and Electrical Engineering & Computer Science
Massachusetts Institute of Technology
77 Massachusetts Avenue, Cambridge MA, 02139, USA
[2] Institute of Cognitive Science, Osnabrück University
aheuser@mit.edu, ptsvilodub@uos.de

Abstract

Given recurring interest in structured representations in computational cognitive models, we extend a Bayesian scoring procedure for comparing symbolic models of language grammar. We conduct a case-study of modeling syntactic principles in German, providing preliminary results consistent with linguistic theory. We also note that dataset and part-of-speech (POS) tagger quality should not be taken for granted.

Recent advances in AI have brought back the controversy between symbolic and connectionist approaches to knowledge representation and learning (e.g., Lake et al. 2017). While great progress in AI has been brought about by neural models, they in general require a lot of training data to achieve human-like generalization, while children make correct generalizations and learn their native language based on substantially fewer examples (Tenenbaum et al. 2011; Xu and Tenenbaum 2007; Chomsky 1965). One position holds that this sample size problem is remedied by resorting to an initial bias. Specifically, many linguists argue that symbolic hierarchical representations and compositionality principles are indispensable for human-like language processing and production (e.g., Crain and Nakayama 1987; Chomsky 1965; Pelletier 1994; Lake et al. 2017). These principles could potentially be incorporated as structural constraints in models of natural language. One approach to building and evaluating human-like symbolic language models comes from the increasingly influential domain of *Bayesian learning* (e.g., Xu and Tenenbaum 2007; Perfors, Tenenbaum, and Regier 2011). In particular, Perfors, Tenenbaum, and Regier (2011) indicate that a learner equipped with domain-general Bayesian inference capacities favors hierarchical phrase structure over linear phrase structure given linguistic data representing input available to human learners (i.e., from a child-directed speech corpus). Perfors, Tenenbaum, and Regier (2011) represent several explicit hypotheses about syntactic structure as probabilistic grammars which are compared via their *Bayesian posterior scores* given training data, assuming a meta-grammar which generated those particular hypotheses. We note that the models were all able to parse the entirety of the data they were compared on, making them directly comparable via posterior scores. However, the inability to evaluate competing models that are supported by different data subsets severely limits the power of modeling results due to the inevitability of noise within the training data. Therefore, in this work we aim to depart from strict Bayesian inference of the system proposed by Perfors, Tenenbaum, and Regier (2011) to enable comparison of models corresponding to different data subsets. To this end, we employ a case study of word-order modeling in German, to additionally assess the quality of a widely used child-language dataset in a language less studied than English.

Data Processing

Following Perfors, Tenenbaum, and Regier (2011), we designed four probabilistic context-free grammars (PCFGs) implementing rules consistent with different hypotheses about the word-order of German: Subject-Verb-Object (SVO), Subject-Object-Verb (SOV), SVO+Verb-Second (SVO+V2) and SOV+V2 word orders, drawing from a specific linguistic theory, among multiple, to specify different conceivable word orders. The SOV+V2 grammar was hand-designed to approximate the basic syntactic structure of standard German. The other three CFGs were derived from the SOV+V2 grammar in order to preserve the encoded syntactic knowledge unrelated to verb positioning phenomena; about 80% of the CFG was identical across models. We note that context-free grammars encode hierarchical sentence structure, meaning that model length does not necessarily correlate with the sentence complexity that they can generate.

The grammars were trained and compared on adult speech from the Leo corpus of the CHILDES dataset (MacWhinney 2000). We preprocessed the dataset such that sentences consisted of part-of-speech (POS) tags from the STTS tag set rather than words. Upon closer inspection of the dataset, we found many tagging and grammatical errors. Therefore, we further preprocessed the data by excluding, among others, typically incorrectly tagged filler words, nonwords and short ungrammatical sentences which are unlikely to be relevant for learning syntactic generalizations, leaving 160,635 utterances in the dataset. The remaining sentences still included more tagging errors than we would have expected.

Though the SOV+V2 CFG was designed to parse German, it was still only able to parse 48.98% of all POS-tagged

sentences in the preprocessed dataset and only 21.29% of 74,758 unique POS strings. We found that it fit the data sufficiently tightly in that it tended not to overgenerate, producing 1.45 parses per POS string. The other grammars were able to parse even smaller proportions of the corpus. To determine the extent to which residual errors left in the corpus were to blame for this, we randomly sampled 100 POS strings and their corresponding sentences that the SOV+V2 CFG could parse and 100 that it could not, and manually determined their grammaticality as native German speakers.

We calculated the error proportion upper bounds for the POS strings that the SOV+V2 CFG could not parse to be 0.19 ($p < 0.05$), and 0.20 for the POS strings that the CFG could parse. Using these error proportions, we determined the sensitivity of the SOV+V2 CFG to be 0.808 and its specificity to be 0.802, concluding that the low proportion of parsed sentences was due to errors in the data. Through this statistical analysis of the SOV+V2 CFG's ability to capture grammatical German POS strings and reject ungrammatical ones, we approximated the percentage of grammatical errors in the processed corpus itself to be 68% of the total number of unique POS strings. With the assumption that each unique ungrammatical POS string only occurs once in the preprocessed Leo corpus, we conservatively estimated that 31.65% (50,854/160,635 total POS strings) of the entire preprocessed corpus were ungrammatical POS strings. Similarly, the proportion of ungrammatical or fragment string tokens in the corpus used by Perfors, Tenenbaum, and Regier (2011) was 30.3%. Therefore, we argue that upon recording, transcribing and processing, this data might not accurately represent a child's linguistic input because we deem adults extremely unlikely to produce ungrammatical speech over 30% of the time in their native language, questioning the suitedness of such a dataset for accurate language learning modeling.

Bayesian Model Evaluation

Finally, we compared how likely a Bayesian learner is to arrive at a particular hypothesis about the word order of German from observing primary linguistic data only. Although we found that the data on which we evaluated our grammars might not have sufficient quality to draw strong conclusions about the models' performance or about language learning in general, we nevertheless propose methods for applying Bayesian scoring to symbolic linguistic models trained on different datasets.

The different grammars were compared using their posterior probability given the preprocessed data they could parse, computed via Bayes' rule from their prior probability under a generative meta-grammar and the likelihood of the data under each grammar (as proposed by Perfors, Tenenbaum, and Regier 2011). This prior computation generally preferred simpler grammars and, importantly, assumed a uniform prior over the different word-order hypotheses. Crucially, the likelihood components of the posterior grammar scores were not directly comparable under strict Bayesian inference because they were each calculated over the different subsets parsed by each of the grammars, penalizing models generalizing to more data (i.e., the SOV+V2 grammar). Further, we note that the likelihood component strongly reflected the number of rules necessary to generate a particular sentence, thereby penalizing the SOV+V2 grammar for parsing more complicated sentences. To combat this, we propose to compute a conservative estimator for the tightness-of-fit of a grammar to the data by normalizing the average sentence log-likelihood by the length of the sentence. To create an estimator for the model's generalization potential to different grammatical phenomena, we additionally weighed this mean of the normalized sentence log-likelihood by the proportion of unique sentence types that the PCFG can parse. The results of this scoring procedure are reported in Table 1. With the additional calculations called for by our system to counteract the compounding effects of parsing different subsets of the data, our system preferred the expected SOV+V2 grammar.

G	$P(G,T\|D)$	$P(D\|G,T)$	$P(G\|T)$
SOV	-851,853.98	-849,792.21	-2,061.77
SVO	-418,402.09	-416,430.19	-1,971.90
SVO+V2	-322,738.54	-320,747.86	-1,990.68
SOV+V2	-223,738.53	-221,574.62	-2,163.91

Table 1: Log prior of grammar $G \mid$ type T, log likelihood of data D with applied corrections and PCFGs' log posteriors

Conclusion

We propose a computational language acquisition modeling approach that combines structured representations and statistical inductive inference in a more flexible way. It provides a starting point for developing scalable comparison methods for models involving symbolic components. In addition, we provide evidence that corpus and POS tagger quality should not be taken for granted.

References

Chomsky, N. 1965. Aspects of the theory of syntax. *Cambridge, MA: MIT Press* (1977): 71–132.

Crain, S.; and Nakayama, M. 1987. Structure dependence in grammar formation. *Language* 522–543.

Lake, B. M.; Ullman, T. D.; Tenenbaum, J. B.; and Gershman, S. J. 2017. Building machines that learn and think like people. *Behavioral and brain sciences* 40.

MacWhinney, B. 2000. *The CHILDES Project: Tools for analyzing talk. Transcription format and programs*, volume 1. Psychology Press.

Pelletier, F. J. 1994. The principle of semantic compositionality. *Topoi* 13(1): 11–24.

Perfors, A.; Tenenbaum, J. B.; and Regier, T. 2011. The learnability of abstract syntactic principles. *Cognition* 118(3): 306–338.

Tenenbaum, J. B.; Kemp, C.; Griffiths, T. L.; and Goodman, N. D. 2011. How to grow a mind: Statistics, structure, and abstraction. *Science* 331(6022): 1279–1285.

Xu, F.; and Tenenbaum, J. B. 2007. Word learning as Bayesian inference. *Psychological review* 114(2): 245.