Evaluating Domain-General Learning of Parametric Stress Typology

Gaja Jarosz (University of Massachusetts Amherst) and Aleksei Nazarov (University of Toronto)

Introduction The existence and characteristics of an innate endowment for language are a matter of ongoing debate. Accepting the idea of innate building blocks of grammar (parameters, Chomsky 1982; constraints, Prince and Smolensky 1993), one may ask whether the learning mechanisms that allow a child to find language-specific configurations of these building blocks are a part of the innate language endowment as well. Such domain-specific learning mechanisms have been proposed for learning of parametric stress: innate cues (data patterns innately specified by the learner), default settings, and/or innately-specified ordering on parameter setting have been argued to be crucial for successful learning (Dresher and Kaye 1990, Pearl 2007). In this paper we present the first evaluations of two domain-general statistical learners, on the complete typology of stress systems generated by Dresher and Kaye's parameter system. On the basis of these quantitative results and an analysis comparing learning difficulty to empirically attested stress systems, we argue that domain-specific learning mechanisms are not necessary for successful learning.

Learning Background The standard probabilistic learner for parameter grammars is Yang's (2002) Naïve Parameter Learner (NPL), which assumes that grammars are defined by a finite set of parameters with probabilities over parameter settings. We examine Pearl's (2011) generalized variant of the NPL. This online learner processes each data point by sampling a setting for each parameter from the probabilistic grammar and using the selected settings to generate a stress pattern for that data point. A match between the predicted and observed pattern adds 1 to the match counter $m(\psi)$ of every sampled parameter setting ψ, while a mismatch decreases the selected settings' $m(\psi)$ by 1. When $|m(\psi)|$ reaches the threshold b, a reward value $R(\psi)$ of 1 or 0 is calculated for that parameter setting as in (1a), and the grammar is updated as in (1b), after which $m(\psi)$ and $m(\neg\psi)$ are reset to 0, where $\neg\psi$ is the opposite setting of the same parameter as ψ; e.g., ψ = FootHead(L), $\neg\psi$ = FootHead(R).

(1) a. $R(\psi) = 1$ iff $m(\psi) \geq b$; $R(\psi) = 0$ iff $m(\psi) \leq -b$; otherwise, $R(\psi) = P_{old}(\psi)$

b. $P_{new}(\psi) = \lambda R(\psi) + (1-\lambda)P_{old}(\psi)$, where λ in [0,1] is the learning rate

We compare the NPL with the Expectation Driven Parameter Learner (EDPL; Nazarov and Jarosz 2017), which uses the same grammar and update rule as the NPL, but determines the reward value individually for each parameter to improve fit with the data. In EDPL, $R(\psi)$ is defined as the probability of that parameter setting according to the current grammar given a match to the current data point: $P(\psi|match)$. This is estimated by sampling r times from the current stochastic parameter grammar while temporarily fixing ψ. The proportion of matches in the sample yields an estimate for $P(match|\psi)$, from which $P(\psi|match)$ is derived by Bayesian reasoning, as in (2).

(2) $R(\psi) = P(\psi|match) = \frac{P(match|\psi)P_{old}(\psi)}{P(match|\psi)P_{old}(\psi)+P(match|\neg\psi)P_{old}(\neg\psi)}$

Computationally, the main difference between the EDPL and NPL is that the EDPL takes r samples per parameter value per data point, while the NPL takes 1 sample for all parameter settings. This, and the distinct reward definition, allows the EDPL to extract more information about the utility of each parameter setting from each word (see Nazarov & Jarosz 2017 for further details).

Simulations Pearl (2011) showed that the NPL is not effective for learning stress parameters for English. Nazarov and Jarosz (2017) showed that the EDPL performed better than the NPL (with b = 0) on a handful of typologically diverse stress systems. Here, we present tests of the NPL and

Proceedings of the Society for Computation in Linguistics (SCiL) 2019, pages 383-384.
New York City, New York, January 3-6, 2019

the EDPL on all 280 unique stress systems possible with the 2048 combinations of the 11 parameters in Dresher and Kaye (1990) and Dresher (1999). Every stress system was presented on all possible 3-to-6-syllable combinations of CV, CVV, and CVC. We defined a run as successful when the stress patterns produced by the learner on each word were 99% correct (out of 100 samples). EDPL simulations were run for a maximum of 100,000 iterations, with $\lambda = 0.1$, $r = 50$. NPL simulations were run for a maximum of 10,000,000 iterations, $\lambda = 0.1$, $b = 0$, 5, 10 (cf. Pearl 2011). Both learners were run 10 times on each of the 280 languages. Results are in Table (3).

(3)	EDPL	NPL, $b = 0$	NPL, $b = 5$	NPL, $b = 10$
% successful runs (median iterations needed to reach success)	94.4% (200)	0.8% (200,000)	6.3% (70,000)	5.3% (4,100)
% languages sometimes (always) learned successfully	95.7% (91.1%)	1.1% (0.7%)	8.9% (3.6%)	8.6% (4.3%)

Discussion The EDPL is far more successful than the NPL in terms of the proportion of successful runs, the number of stress systems learned, and the number of iterations needed for learning. The NPL with $b = 0$ reliably learns only two systems: absolute initial stress, and absolute final stress. The other two variants of the NPL perform somewhat better thanks to the memory buffer created by the match counter, but still learn a small subset of the languages that the EDPL learns. The small proportion of stress systems (12 languages, 4.3% of the typology) that the EDPL reliably fails to learn are typologically anomalous:2 of these resemble Cairene Arabic (McCarthy 1979) and Manobo (Dubois 1976), respectively, in some aspects, while the other 10 are unattested variations on these typologically rare themes (Goedemans et al. 2015). This suggests that the characteristics of stress systems that pose difficulty to the EDPL may also be affected by a general learning pressure in human learners. In that case, the inability of the machine learners investigated to learn these languages models this learning pressure and could help explain the near-absence of stress systems with these characteristics among the worlds' languages.

Conclusion Our results corroborate and extend Pearl's (2011) finding that the NPL fails to learn attested stress systems; however, we also show that an alternative domain-general learning model (EDPL) performs well on a complete stress system typology. Further, while the EDPL fails to learn some of the stress systems, we show there is a strong correspondence between the languages that pose difficulty for the learner and their empirical (un)attestedness. Together, these findings suggest that domain-specific learning mechanisms may not be necessary when domain-general learners are equipped with more nuanced statistical learning mechanisms.

Select References Dresher, B.E. 1999. "Charting the Learning Path: Cues to Parameter Setting." *LI* 30.27-67. ♦ Dresher, B.E. and J.D. Kaye. 1990. "A computational learning model for metrical phonology." *Cognition* 34.137-195. ♦ Dubois, C. 1976. *Sarangani Manobo: An Introductory Guide*. Linguistic Society of the Philippines. ♦ Goedemans, R., J. Heinz, and H. van der Hulst. 2015. *StressTyp2, version 1*. http://st2.ullet.net. ♦ Nazarov, A. and G. Jarosz. 2017. "Learning Parametric Stress without Domain-Specific Mechanisms." In K. Jesney, C. O'Hara, C. Smith, and R. Walker (eds.), *Proceedings of the 2016 Meeting on Phonology*, LSA. ♦ Pearl, L. 2007. *Necessary Bias in Natural Language Learning*. PhD dissertation, University of Maryland. ♦ Pearl, L. 2011. "When Unbiased Probabilistic Learning is Not Enough: Acquiring a Parametric System of Metrical Phonology." *Language Acquisition* 18(2).87-120. ♦ Yang, C. 2002. *Knowledge and Learning in Natural Language*. Oxford University Press.