DISENTANGLING THE STATISTICAL PROPERTIES THAT DRIVE LANGUAGE ACQUISITION: EVIDENCE FROM MAXIMALLY DIVERSE LANGUAGES

Jekaterina Mažara[*1,2], Giuachin Kreiliger[1,2], Sabine Stoll[1,2], and Balthasar Bickel[1,2]

[*]Corresponding Author: jekaterina.mazara@uzh.ch
[1]Department of Comparative Language Science University of Zurich, Switzerland
[2]Center for the Interdisciplinary Study of Language Evolution, University of Zurich, Switzerland

A key property of the human language faculty is that any of its instantiations can be acquired by any child in roughly the same time frame. A powerful suggestion for what makes this possible is the idea that children can rely on statistical properties for segmenting and learning elements such as morphemes or words (Saffran, 2003; Saffran, Aslin, & Newport, 1996; Gerken, 2005; Pelucchi, Hay, & Saffran, 2009; Hay, Pelucchi, Estes, & Saffran, 2011; Thiessen, 2011; Höhle, 2015). However, an unresolved question is which specific aspects of statistical distributions drive or differentiate the acquisition process in real-world data: is it the sheer number of items that need to be learned? Or distributional differences across structural or semantic types, or across entire languages? Or frequency classes?

Here we address these questions by assessing potential factors that might influence the acquisition of verb forms in morphologically highly diverse languages. Our data consists of longitudinal recordings of 20 children and their surrounding speakers from five naturalistic corpora from the *ACQDIV* data base (Moran, Schikowski, Pajović, Hysi, & Stoll, 2016): Chintang (Sino-Tibetan), English (Indo-European), Japanese (Japonic), Turkish (Turkic), and Yucatec (Mayan). As potential factors we tested the following: (i) **the difference between *heads* and *dependents***. *Heads* include the main stem of a verb form that carries structurally independent information, while *dependents* comprise elements that provide modifying lexical, derivational, and grammatical information, such as affixes and secondary stems (Stoll, Mazara, & Bickel, 2017). We hypothesize that heads are easier to learn than dependents because they are cognitively more salient. (ii) **the difference between languages**, hypothesizing that some languages are generally acquired faster than others. (iii) the **deviation of the empirical input distributions from the theoretical Zipf distribution** (quantified by the contribution of a parabolic function to fitting the log-log rank-frequencies), hypothesizing that convexity in rank-frequency distributions facilitates learning because there are more high-frequency items than expected. (iv) the **log number of distinct heads and**

dependents, assuming that a higher number would indicate a more difficult system for learners. (v) the **entropy** of items (measured with the Chao-Shen estimator (Chao & Shen, 2003) to account for possible unseen elements), hypothesizing that lower entropies facilitate learning (Lavi-Rotbain & Arnon, 2022).

To evaluate the children's development of productivity in verb form use over time, we computed the (log) ratio in usage entropies between the target child and surrounding adult speakers within each recording session (Stoll et al., 2012).

We applied a hierarchical nonlinear scale/location ("distributional") model (Bürkner, 2018) to fit the development of the log entropy ratios over time using the exponential function to describe the learning curves, which allows us to evaluate the speed of acquisition. To allow for variation at an individual level, we model the children as random effects. We fitted models with one of the five factors each (plus an ancillary model of item type and language together) and compared them through model stacking, estimating the relative weight of each factor in predictive performance during leave-one-out cross-validation (Vehtari, Gelman, & Gabry, 2017).

Model stacking (Yao, Vehtari, Simpson, & Gelman, 2018) shows that the data are best predicted by the convex deviation from the theoretical Zipfian distribution, i.e. distributions with more higher-frequency types faciliate learning and increase the rate of acquisition. The (log) number of items has a slightly weaker but still appreciable impact (stacked prediction model weight .45 vs .5 for the deviation from the Zipfian distribution). All other factors predict the data much less well (receiving no weight in stacking), i.e., differences between languages, item types, and entropies have no impact on the acquisition process.

Our findings suggest that the children in our sample indeed do make use of statistical properties of the input during learning, and that it is specifically the number of high-frequency items and the total number of items that matters.

References

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*, 395–411.

Chao, A., & Shen, T.-J. (2003). Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environmental and Ecological Statistics*, *10*(4), 429–443.

Gerken, L. (2005). Decisions, decisions: infant language learning when multiple generalizations are possible. *Cognition*, *98*(3), B67–B74.

Hay, J. F., Pelucchi, B., Estes, K. G., & Saffran, J. R. (2011). Linking sounds to meanings: infant statistical learning in a natural language. *Cognitive Psychology*, *63*(2), 93–106.

Höhle, B. (2015). Crosslinguistic perspectives on segmentation and categorization in early language acquisition. In E. L. Bavin & L. R. Naigles (Eds.), *The cambridge handbook of child language* (p. 159-182). Cambridge University

Press.

Lavi-Rotbain, O., & Arnon, I. (2022). The learnability consequences of zipfian distributions in language. *Cognition*, *223*, 105038.

Moran, S., Schikowski, R., Pajović, D., Hysi, C., & Stoll, S. (2016). *The acqdiv corpus.*

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Statistical learning in a natural language by 8-month-old infants. *Child Development*, *80*(3), 674–685.

Saffran, J. R. (2003). Statistical language learning mechanisms and constraints. *Current Directions in Psychological Science*, *12*(4), 110–114.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926 - 1928.

Stoll, S., Bickel, B., Lieven, E., Banjade, G., Bhatta, T. N., Gaenszle, M., Paudyal, N. P., Pettigrew, J., Rai, I. P., Rai, M., & Rai, N. K. (2012). Nouns and verbs in chintang: children's usage and surrounding adult speech. *Journal of Child Language*, *39*, 284-321.

Stoll, S., Mazara, J., & Bickel, B. (2017). The acquisition of polysythetic verb forms in chintang. In M. Fortescue, M. Mithun, & N. Evans (Eds.), *Handbook of polysynthesis* (p. 495-517). Oxford University Press.

Thiessen, E. D. (2011). Domain-general constraints on statistical learning. *Child Development*, *82*(2), 462–470.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing*, *27*(5), 1413–1432.

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using stacking to average bayesian predictive distributions (with discussion). *Bayesian Analysis*, *13*(3), 917–1007.