Modeling typological frequency with a grammatical learner

Vighnesh Subramaniam (Millard North High School, Omaha NE) and Adam Albright (MIT)

Introduction: Why are some patterns common cross-linguistically, while others are rare or unattested? Most work assessing the typological predictions of phonological grammars have focused on the binary distinction between attested vs. unattested, and do not account for the relative frequency of attested patterns. Anttila (1997) and Bane and Riggle (2008) hypothesize that the relative frequency of a pattern is proportional to the number of constraint rankings that generate it. This is an intuitive hypothesis, but as a model of typological frequency, it assumes that the set of languages has been generated randomly, and transmitted perfectly. However, Staubs (2014), Stanton (2016), and others observe that some rankings are difficult to acquire, because the evidence that distinguishes them from other rankings is rare in the input to learners. Thus, patterns corresponding to these rankings may be mislearned, becoming rare or even unattested over time. Staubs and Stanton show that certain unattested stress patterns are indeed likely to be reanalyzed as other, attested patterns, given "ordinary" data. In this paper, we examine the broader predictions of this approach for relative typological frequency. We test whether a learning model, iterated over many generations, gradually approximates the attested frequencies of stress patterns in StressTyp2 (Goedemans, Heinz, and van der Hulst 2014).

Factorial Typology Model of Single Stress: Following Gordon (2002), Bane and Riggle (2008), Staubs (2014), and Stanton (2016), we model the typology of quantity-insensitive single-stress systems. We adopted the set of 9 non-foot-based constraints employed by Stanton (2016), and used OTSoft (Hayes, Tesar and Zuraw 2013) to calculate the factorial typology for input words of 2–8 syllables. Out of 40320 logically possible stress patterns, only 48 are predicted to occur, while the remaining 40272 cannot be derived. When we compare these patterns to the StressTyp2 database, we find (replicating Gordon, and Bane and Riggle) that none of the underivable patterns are attested. However, out of the 48 derivable patterns, only 9 are actually attested, some much more frequently than others.

Ranking-Based Frequencies: As a baseline, we first modeled the relative frequency of patterns as a function of the number of rankings that generate each (r-volume; Bane and Riggle 2008). For each of the 48 patterns in the typology, we computed the r-volume from the elementary ranking conditions, using a method proposed in Riggle (2010). The results replicate Bane and Riggle's finding that r-volume correlates positively with typological frequency (r=.89), but the fit is qualitatively imperfect: 39 patterns that are predicted to occur are unattested, while the remaining 9 are overattested.

Learning-Based Frequencies: We next tested the hypothesis that some patterns are rare or unattested because they are difficult to learn (Staubs 2014, Stanton 2016). We assume an initial distribution as predicted by r-volume, and test how mislearning shifts probability onto other, better-attested patterns. As in Stanton (2016), we employed the Gradual Learning Algorithm (GLA; Boersma 1997, Magri 2012) to simulate learning. We trained the model on each of the 48 possible stress patterns, for words of 2–8 syllables. Shorter words were presented more frequently than longer words, according to the mean frequencies reported in Stanton (2016). We assume that all constraints start out ranked equally, and constraints are demoted in response to data. In order to simulate mislearning due to limited data, we trained each learner for 2000 trials with a small plasticity. Due to the paucity of long words in the training data, this is not always sufficient to arrive at the target ranking. Each stress pattern was trained 1000 times, yielding

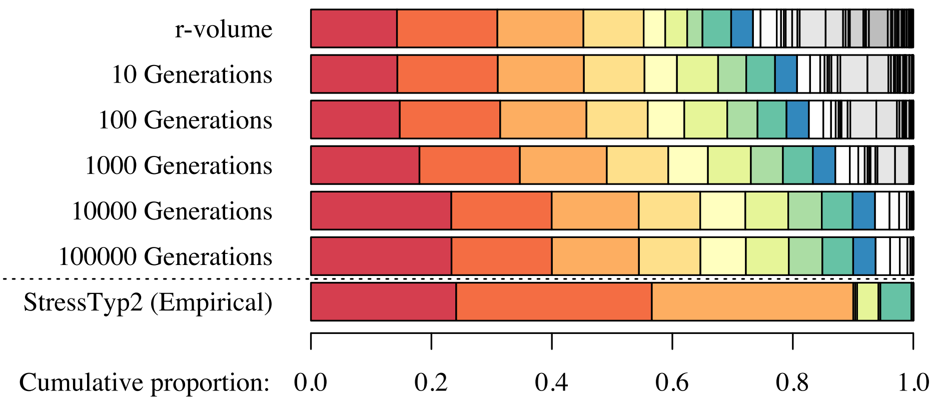

Figure 1: Iterated learning gradually eliminates unattested stress patterns (in gray)

a probability distribution over learning outcomes. This distribution was used to redistribute typological probability: starting with the r-volume as a baseline, the probability of each stress pattern was adjusted up or down, depending on how often the model acquired it in favor of other targets, or failed to acquire it. We iterated learning-based redistributions of probability for 100,000 "generations", to model the aggregated effect of mislearning over time.

Results: Figure 1 shows the effect of successive iterations of learning on the distribution predicted by r-volume. Over time, the patterns that are attested in StressTyp2 (colored bars) increase in frequency, while the patterns that are unattested (gray bars) decrease. This is consistent with the claim that certain unattested stress patterns may be eliminated by considerations of learnability (Staubs 2014, Stanton 2016). To test whether a learning-based model does better than r-volume alone, we conducted a zero-inflated Poisson regression, predicting StressTyp2 counts based on r-volume, and whether the pattern is predicted to increase or decrease over time (Learnability). The model shows that Learnability is a significant predictor of whether a pattern is attested ($p<.001$), while r-volume is a significant predictor of typological frequency of attested patterns. A model comparison reveals that adding learnability does not quite yield significant improvement over r-volume alone (Vuong $z=-1.383$, $p=.08$), but Figure 1 shows that qualitatively, the learning model succeeds in reducing or eliminating many unattested stress patterns. It is interesting to note that the effect is not complete, however; in fact, some unattested stress patterns are predicted to remain, to a limited extent ($<10\%$).

We then tested whether the predicted typological frequencies after many generations do a better job than r-volume at predicting StressTyp2 counts, using Poisson regression. The results show that although r-volume and learned frequencies are fairly strongly correlated, each makes significant independent contributions in predicting typological frequency. Notably, redistributing probability through learning does not provide a straightforward improvement over r-volume. This may reflect the fact that an idealized 'single creation event + generations of learning' model is overly simplistic; in fact, languages without stress may acquire it over time, which may help maintain a more "random" (r-volume based) typological distribution. In addition, real languages often provide less data than our training assumes concerning long words, because long words tend to be morphologically complex, and are thus governed by additional principles. Thus, some patterns that are preserved in our simulations may actually be less learnable than the model predicts.