Modeling markedness with a split-and-merger model of sound change

Andrea Ceolin and Ollie Sayeed

ceolin@sas.upenn.edu, sayeedo@sas.upenn.edu

1st International Workshop on Computational Approaches to Historical Language Change - 2019

Introduction

Phonological theories have described phonemes as marked or unmarked according to language-internal and language-external properties (Jakobson 1941, Haspelmath 2006, Dresher 2009). Marked segments are rare within and across languages, while unmarked segments are common (Gordon 2016).
But **what does it mean for a segment to be marked**? We explore the possibility that markedness is the expected outcome of phonetically grounded sound change.

Evolutionary Phonology

Our model is a formalization of Evolutionary Phonology (EP, Blevins 2004), where generalizations about sound patterns are reduced to generalizations about how those sound patterns come into existence. We define two new properties of 'splitwise' and 'mergerwise' markedness:

- **Splitwise marked** segments have a low probability of being created by sound change.
- **Mergerwise marked** segments have a high probability of being destroyed by sound change.

In EP, x is splitwise marked if it's hard to misperceive another sound as x; and x is mergerwise marked if it's easy to misperceive x as another sound.

A mathematical model

Our model is a variant of a class of random fragmentation and aggregation models (Banavar et al. 2004), which have power-law distributions as their fixed points. We see these power-law distributions in attested type and token frequencies of phonemes within a language (Yule-Simon, Simon 1955, Tambovtsev and Martindale 2007, Martin 2007).

Following the traditional typology in historical linguistics, we start with a frequency distribution over phonemes, and apply a **split** or a **merger** with equal probability:

- To apply a split, pick a random pair of segments x_i, x_j with $i \neq j$. Take away half of x_i's probability mass and add it to the probability mass of x_j.

$$p_i^{t+1} := \frac{p_i^t}{2} \qquad p_j^{t+1} := \frac{p_i^t}{2} + p_j^t$$

- To apply a merger, transfer *all* of x_i's probability mass to x_j.

$$p_i^{t+1} := 0 \qquad p_j^{t+1} := p_i^t + p_j^t$$

To encode biases in the actuation of sound change, we define **splitwise markedness** $P_S(x_j)$ and **mergerwise markedness** $P_M(x_i)$ as probability distributions over segments. In applying a split, we bias the choice of x_j according to $P_S(x_j)$, and in applying a merger, we bias the choice of x_i according to $P_M(x_i)$.

A computational implementation

We ran 1,000 simulations of the model for 500 generations, each with the same starting conditions:

- 20 phonemes, arbitrarily labelled $\{a,\ b,\ c \ldots\ldots t\}$;
- Uniform initial frequency in the language (0.05);
- Six phonemes $\{u,\ v,\ w,\ x,\ y,\ z\}$ with frequency 0 (i.e. don't exist yet, but can be created through a split).

Simulations of the split-and-merger model in action show long-tailed distributions emerging out of an initial flat distribution (Figure 1).

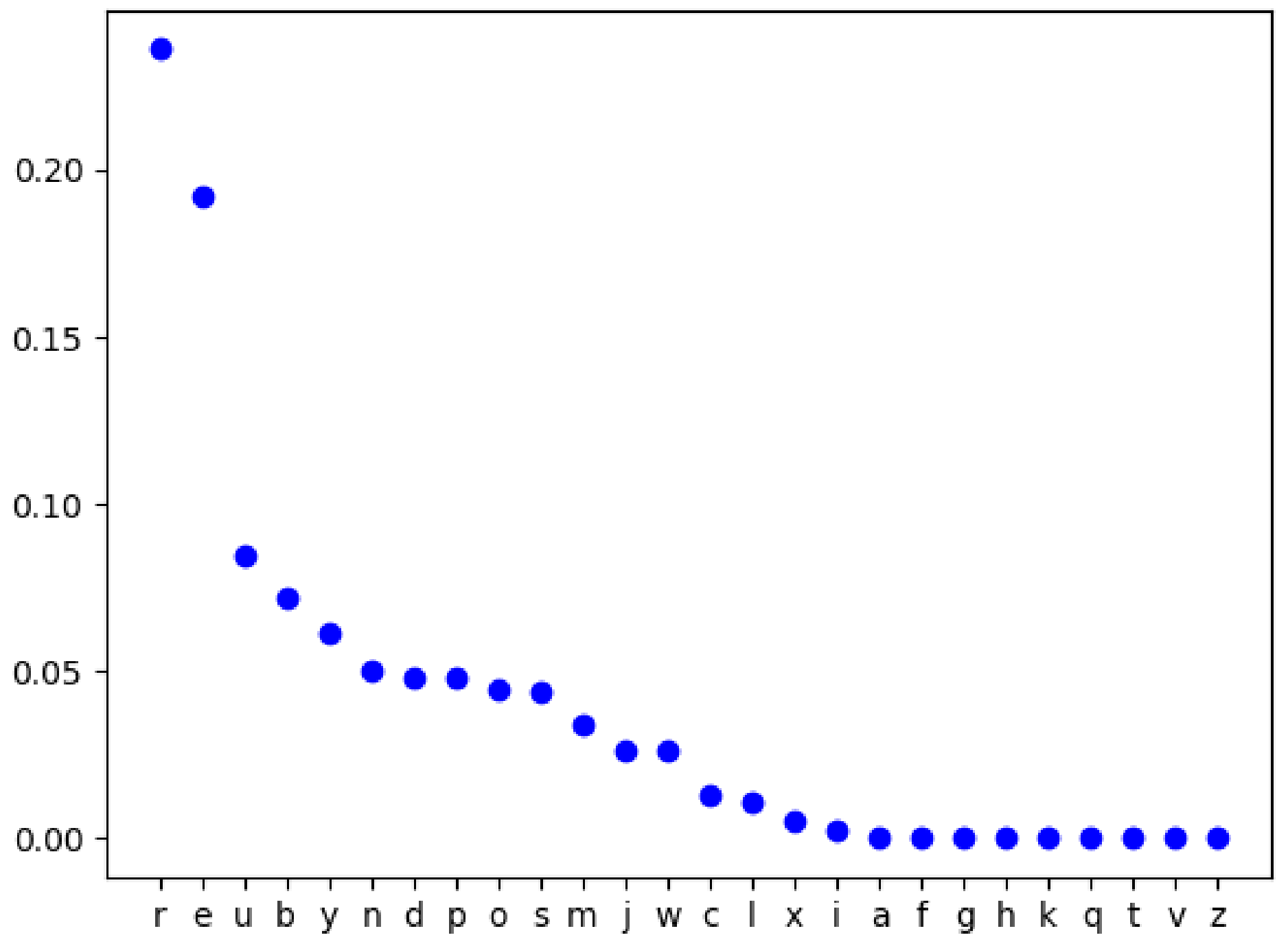

Figure 1: A typical run of our simulation after 500 iterations.

Splitwise markedness

We re-run the simulation first implementing splitwise markedness. 'a' is **unmarked** (i.e., high probability of resulting from a split), 'b' is **marked** (i.e., low probability of resulting from a split), and 'c' is neutral. Figure 2 shows the average frequencies in the languages in which 'a', 'b' and 'c' survive, and it shows that 'a' has a higher average than 'c' and 'b', while these latter segments **do not exhibit a clear difference**. Across-language frequencies are instead distinct (a=0.773, b=0.423, c=0.311).

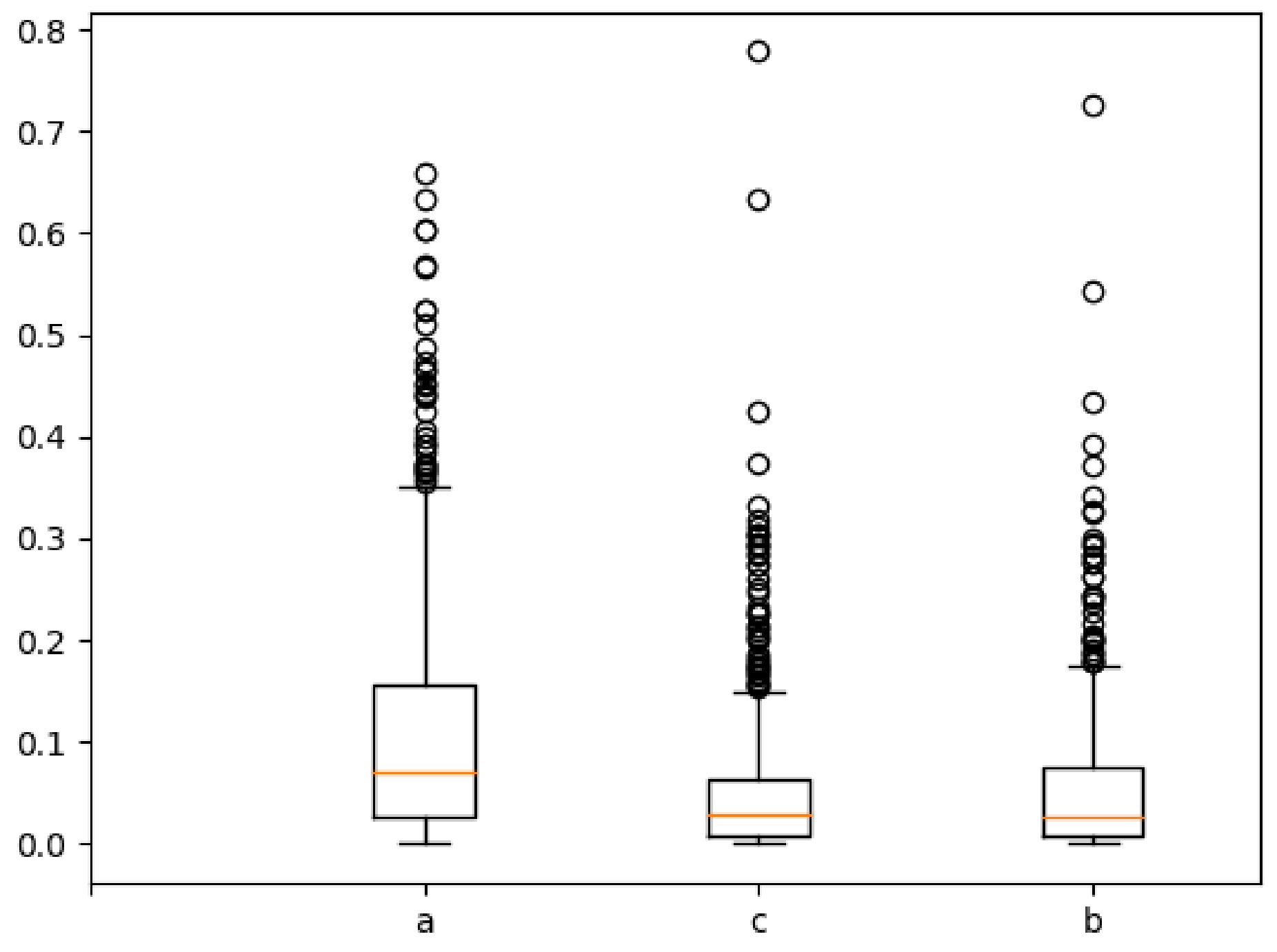

Figure 2: Summary of the final within-language frequencies of 'a', 'c' and 'b', which are modeled in terms of splitwise markedness, after 1000 parallel runs.

Mergerwise markedness

We repeat the simulation modeling mergerwise markedness, using the same symbols to represent marked and unmarked segments. Figure 3 shows the average frequencies in the languages in which 'a', 'b' and 'c' survive, and it shows that the three segments have **different distributions** (whose statistical significance depends on the magnitude of the bias). Within- and across-language frequencies line up (a=0.924, b=0.548, c=0.109), exhibiting a correlation.

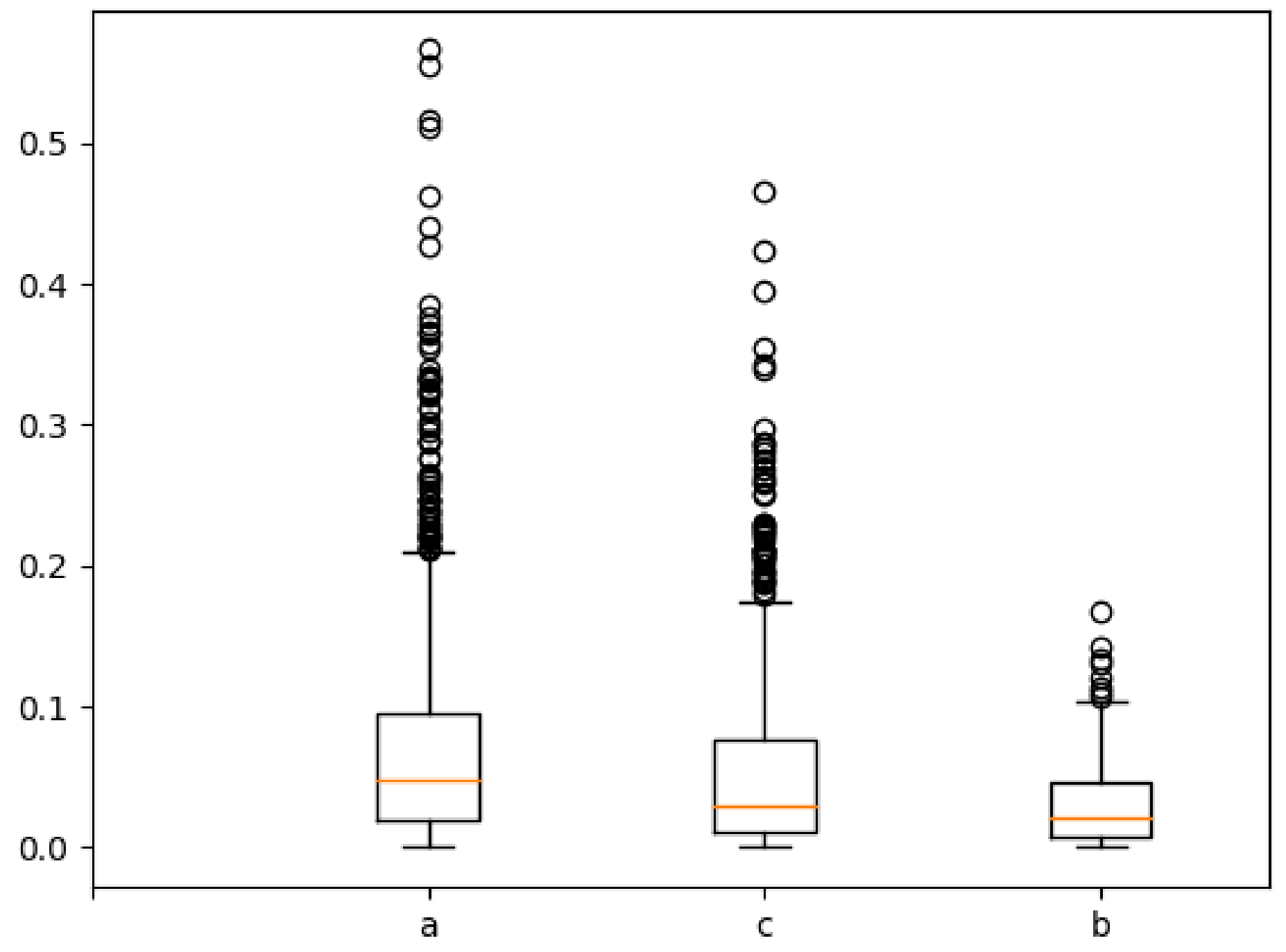

Figure 3: Summary of the final within-language frequencies for 'a', 'c' and 'b', which are modeled in terms of mergerwise markedness, after 1000 parallel runs.

Conclusions

Both the power-law frequency distribution of phonemes in a language and the cluster of properties associated with markedness can be thought of as **epiphenomena** of phonetically grounded sound change. In particular, mergerwise markedness appears to be responsible for higher within- and across-language frequencies for unmarked segments and lower frequencies for marked segments, while splitwise markedness mainly affects unmarked segments. Directions for further work:

- extend the model to strings, allowing conditioned sound changes;
- add a miniature lexicon and bias changes to affect contrasts with low functional load;
- allow multiple segments to change at once, either through feature-based natural classes or through a chain shift;
- derive the asymmetry between the behaviour of splitwise and mergerwise markedness.

Selected References

Banavar, J. R. et al. (2004). Scale-free behavior and universality in random fragmentation and aggregation. **Blevins, J.** (2004). Evolutionary phonology: The emergence of sound patterns. **Dresher, E.** (2009). The contrastive hierarchy in phonology. **Gordon, M. K.** (2016). Phonological typology. **Haspelmath, M.** (2006). Against markedness (and what to replace it with). **Martin, A. T.** (2007). The evolving lexicon. **Simon, H. A.** (1955). On a class of skew distribution functions. **Tambovtsev, Y. & Martindale, C.** (2007). Phoneme frequencies follow a Yule distribution.