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

Andrea Ceolin
University of Pennsylvania
Department of Linguistics
ceolin@sas.upenn.edu

Ollie Sayeed
University of Pennsylvania
Department of Linguistics
sayeedo@sas.upenn.edu

1 Background

The concept of 'markedness' has been influential in phonology for almost a century. Theoretical phonology used it to describe some segments as more 'marked' than others, referring to a cluster of language-internal and language-external properties (Jakobson, 1968; Haspelmath, 2006). We argue, using a simple mathematical model based on Evolutionary Phonology (EP; Blevins, 2004), that markedness is an epiphenomenon of phonetically grounded sound change.

2 Model: random splits and mergers

We propose a simple abstract model of sound change as a discrete-time stochastic process of random splitting and merging of phonemic categories. In the split-and-merger model, sound change belongs to a class of random fragmentation and aggregation processes (Banavar et al., 2004), whose fixed points are power-law frequency distributions over the elements being split and merged. It has been shown that phoneme type and token frequencies in natural languages do indeed follow a power-law distribution, specifically a Yule-Simon distribution (Simon, 1955; Tambovtsev and Martindale, 2007; Martin, 2007).

Say the phoneme inventory of a language is a set of segments $\{x_i\}$, where the ith segment x_i has frequency p_i^t at time step t. At each stage, we apply either a *split* or a *merger* to the language 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 existing probability mass of x_j.

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

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

$$p_k^{t+1} := p_k^t$$

- Mergers follow a similar algorithm, except that *all* of x_i's probability mass is transferred to x_j.

$$p_i^{t+1} := 0$$

$$p_j^{t+1} := p_i^t + p_j^t$$

$$p_k^{t+1} := p_k^t$$

- Define a function $P_S(x_j)$ such that $P_S(x_j) \geq 0$ and $\Sigma P_S(x_j) = 1$; this is a probability distribution representing the probability that the jth segment x_j will have its frequency increased when another segment splits. When the splitting algorithm calls for picking a random pair of segments x_i, x_j, pick x_j randomly according to the distribution $P_S(x_j)$.

- Define a second probability distribution $P_M(x_i)$, representing the probability that x_i is lost in a merger. When the merging algorithm calls for picking a random pair of segments x_i, x_j, pick x_i randomly according to $P_M(x_i)$.

Say that segments with *low* $P_S(x_j)$ are 'splitwise marked', and segments with *high* $P_M(x_i)$ are 'mergerwise marked'. In other words, marked segments are segments that either *do not* tend to be created after a split, or *do* tend to be lost in a merger.

3 Predictions: within-language and across-language frequency

Empirically, across-language phoneme frequencies correlate well with within-language frequencies (Gordon, 2016). We show that a split-and-merger model derives this link from stochastic sound change.

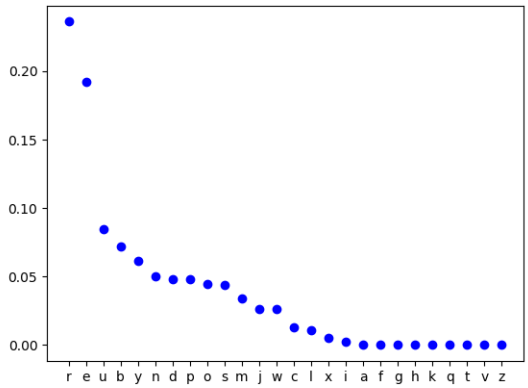

Figure 1: A typical run of our simulation after 500 iterations.

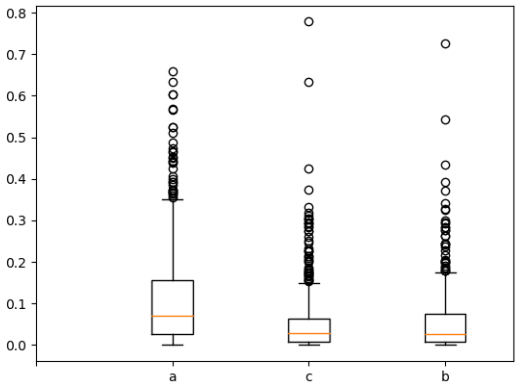

Figure 2: Summary of the final within-language frequencies of 'a', 'c' and 'b, which are modeled in terms of splitwise markedness, after 1000 parallel runs, with r=10.

We run a simulation of the split-and-merger process for 500 iterations with a set of 20 segments arbitrarily labeled $\{a, b, c,....t\}$. We assume that segment frequencies are uniform as a starting point. In addition, six segments $\{u, v,...z\}$ are assigned an initial value of zero. In the sound change simulation, either a split or a merger is applied to the phonemic inventory at each iteration with equal probability. Simulations of the split-and-merger model in action show long-tailed distributions emerging out of an initial flat distribution, qualitatively in line with the results from random fragmentation and aggregation models (Figure 1).

3.1 Splitwise markedness

We re-run the simulation first implementing splitwise markedness. In this simulation, 'a' is 'unmarked' with respect to the other segments by having a probability of increasing its frequency after a split which is higher than that of the other segments, and 'b' is 'marked' by having a probability of increasing its frequency after a split which is lower. The probabilities are determined by a parameter r, which represents the ratio between the probability of the 'unmarked' and the 'marked' segments with respect to the others. This value quantifies how 'unmarked' or 'marked' a segment is with respect to the others.

In a first experiment, we track the average frequencies of 'a' and 'b' across 1000 parallel runs, and we also track the number of runs in which they survive, interpreting each independent run as a separate language. We then compare these numbers with the frequencies exhibited by segments which are neither 'unmarked' nor 'marked', for example 'c'.

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.

Table 1 shows both within- and across-language frequencies for different values of r. Interestingly, increasing the value for r has the effect of increasing the difference between 'a' and the other symbols, but it does not have any effect on 'b'. On the other hand, across-language frequencies are clearly distinct, and 'a' and 'b' display frequencies different from the neutral segment 'c'. These differences become more salient as r increases.

This experiment shows that when we add a diachronic bias, 'unmarked' segments display higher frequencies both within- and across-languages, while the effect for 'marked' segments appears to be limited to across-language frequencies. This might follow from the fact that while splitwise marked segments tend to appear less in languages, their within-language frequencies are dependent on other factors, for instance the frequency of the segments from which they split or their likelihood of merging with other segments. In the next subsection, we investigate mergerwise markedness.

3.2 Mergerwise markedness

We repeat the simulation modeling mergerwise markedness. This time, 'a' is 'unmarked' with re-

	Markedness	Within-language	Across-language
r=2			
'a'	Unmarked	0.063 (±0.006)	0.572 (±0.003)
'c'	Neutral	0.057 (±0.007)	0.475 (±0.003)
'b'	Marked	0.056 (±0.008)	0.410 (±0.003)
r=5			
'a'	Unmarked	0.081 (±0.006)	0.702 (±0.003)
'c'	Neutral	0.058 (±0.007)	0.452 (±0.003)
'b'	Marked	0.052 (±0.008)	0.348 (±0.003)
r=10			
'a'	Unmarked	0.099 (±0.008)	0.773 (±0.002)
'c'	Neutral	0.052 (±0.007)	0.423 (±0.003)
'b'	Marked	0.058 (±0.008)	0.311 (±0.003)

Table 1: Average within- and across-language frequencies for three segments which differ in terms of splitwise markedness, with different values of r. Confidence intervals at 95% are reported for within-language frequencies. We also report confidence intervals at 95% for across-language frequencies, which we obtained by repeating the whole experiment 100 times.

spect to the other segments by having a probability of being lost after a merger which is lower than that of the other segments, and 'b' is 'marked' by having a probability of being lost after a merger which is instead higher. The probabilities are determined by the same parameter r.

As previously done, we track the average frequencies of 'a' and 'b' across 1000 parallel runs and the number of runs in which they survive, along with those of a neutral segment 'c'.

Figure 3 shows the average frequencies in the languages in which 'a', 'b' and 'c' survive, and it shows that this time the three segments have different distributions. From Table 2, we see that within- and across-language frequencies line up, exhibiting a correlation. In this case, 'marked' segments exhibit a lower within-language frequency with respect to neutral segments.

4 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. A stochastic split-and-merger model predicts the attested language-internal and typological correlations. 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.

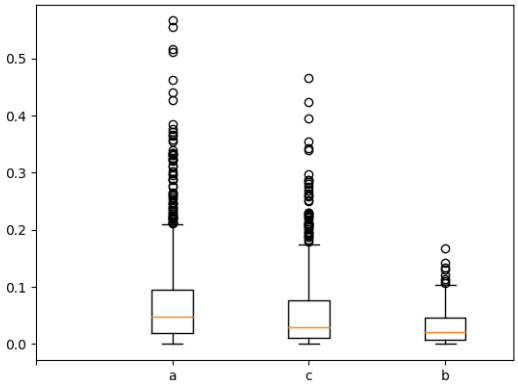

Figure 3: Summary of the final within-language frequencies for 'a', 'c' and 'b', which are modeled in terms of mergerwise markedness, after 1000 parallel runs, with r=10.

	Markedness	Within-language	Across-language
r=2			
'a'	Unmarked	0.065 (±0.006)	0.652 (±0.003)
'c'	Neutral	0.056 (±0.007)	0.485 (±0.003)
'b'	Marked	0.052 (±0.008)	0.320 (±0.003)
r=5			
'a'	Unmarked	0.071 (±0.006)	0.836 (±0.002)
'c'	Neutral	0.051 (±0.005)	0.509 (±0.003)
'b'	Marked	0.045 (±0.008)	0.173 (±0.002)
r=10			
'a'	Unmarked	0.072 (±0.005)	0.924 (±0.002)
'c'	Neutral	0.050 (±0.005)	0.548 (±0.003)
'b'	Marked	0.032 (±0.007)	0.109 (±0.002)

Table 2: Average within- and across-language frequencies for three segments which differ in terms of mergerwise markedness, with different values of r. Confidence intervals at 95% are reported for within-language frequencies. We also report confidence intervals at 95% for across-language frequencies, which we obtained by repeating the whole experiment 100 times.

References

Jayanth R Banavar et al. 2004. Scale-free behavior and universality in random fragmentation and aggregation. *Physical Review E*, 69(3):036123.

Juliette Blevins. 2004. *Evolutionary phonology: The emergence of sound patterns*. Cambridge University Press.

Matthew Kelly Gordon. 2016. *Phonological typology*, volume 1. Oxford University Press.

Martin Haspelmath. 2006. Against markedness (and what to replace it with). *Journal of linguistics*, 42(1):25–70.

Roman Jakobson. 1968. *Child language: aphasia and phonological universals*. 72. Walter de Gruyter.

Andrew Thomas Martin. 2007. *The evolving lexicon*. Ph.D. thesis, University of California, Los Angeles Los Angeles, CA.

Herbert A Simon. 1955. On a class of skew distribution functions. *Biometrika*, 42(3/4):425–440.

Yuri Tambovtsev and Colin Martindale. 2007. Phoneme frequencies follow a Yule distribution. *SKASE Journal of Theoretical Linguistics*, 4(2):1–11.