A Poisson model of phonological cooccurrence restrictions

Adam Albright[1], Canaan Breiss[2]

[1]*Massachusetts Institute of Technology,* [2]*University of Southern California*

A basic goal of phonological analysis is accounting for cooccurrence restrictions. When restrictions are categorical or drive repairs, it is easy to notice them and argue that they are part of phonological grammar. A major contribution of the Laboratory Phonology program has been to document numerous gradient static restrictions, in which elements cooccur less frequently than expected, but exceptions are attested and are not repaired. The known categorical restrictions may be the tip of the phonotactic iceberg, and careful study of lexica could conceivably reveal dozens or even hundreds of gradient restrictions in every language—but how do we know which are "real", even as descriptions of the lexicon? We address this question by applying Bayesian Poisson regression to the distribution of onset types in Lakhota (Siouan) roots, to find and quantify cooccurrence restrictions. Our results reveal numerous novel phonotactic restrictions, not previously documented in Lakhota or elsewhere. We further argue that these restrictions are *structured*, in that they follow from the relationship between simpler restrictions, and do not always constitute separate phonological constraints. We address some challenges involved in scaling up the approach, in hopes of encouraging more widespread use of these models.

Following [1], the existence and strength of gradient restrictions is often measured using Observed/Expected ratios. O/E attempts to control for rarity of individual subparts, but [2] show that "expected" values do not always provide an accurate baseline estimate. In addition, O/E provides no estimate of the significance of underattestation. [2] argue instead for the use of loglinear models in the Generalized Linear Model (GLM) family (Poisson, Binomial/Multinomial (a. k. a. MaxEnt models), etc.) to estimate independent occurrence and cooccurrence rates. The GLM approach predicts counts of words in the lexicon based on the structures they contain (individually, and in combination). We model the cooccurrence of onset classes (null, plain stop, aspirated stop, ejective, fricative, glide, nasal, lateral, cluster) within 2,265 one- and two-syllable roots from [3]. When reduced to these onset classes, the 2,265 roots exemplify 241 distinct root shapes. Since we are interested in the difference between root shapes that occur often vs. rarely or never, we also included 351 logically possible but unattested root shapes. We fit a Poisson model, predicting the number of attested roots for each root shape, based on 111 phonological predictors. We included 45 factors for syllable count and onset type in initial and medial position. This fine-grained approach provides the most accurate possible baseline for estimating deviations due to cooccurrence. We also included 66 factors marking (order-independent) cooccurrence of onset classes. These are similar to interaction terms, in that they measure the effect on root counts specifically when two structures cooccur. A negative coefficient for these terms indicates that roots combining the two structures are underattested.

The model results show that (1) this level of granularity suffices to provide a decent baseline model of the counts of different root shapes (Fig. 1), and (2) 25 of the 66 cooccurring combinations are underattested, to varying extents (Table 1). Some of the strongest restrictions reflect well-known OCP restrictions, such as bans on two ejectives or two laterals. Many other restrictions have not been previously observed as categorical phonotactic effects in any language, such as a restriction on two clusters or two fricatives within a root, and a restriction against fricatives cooccurring with clusters. Table 1 reveals that when an onset type is avoided in cooccurrence with itself, it is also avoided with other dispreferred onset types (negative coefficients to the right of the table, off the diagonal). Thus, although we observe many cooccurrence restrictions, they are structured, in the sense that they are not independent of one another. We interpret these results using a phonological model in which some cooccurrence restrictions follow from cumulative interactions of simpler markedness constraints [4], [5], [6].

Figure 1: Model predictions for root shapes, based on onset classes

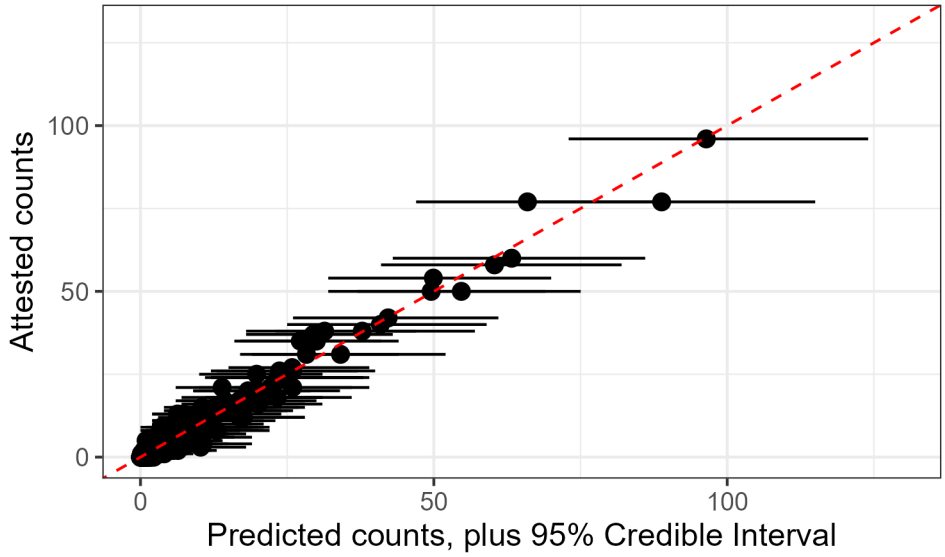

Table 1: Coefficients of cooccurrence factors. Shading indicates strength of significant restrictions. Negative values indicate that the combination occurs together less often than expected; positive values indicate that the cooccurrence is over-represented. Underlining indicates commonly observed OCP effects: cooccurrence of identical feature values that are frequently banned typologically (lateral and laryngeal features).

	Nasal	Null	Stop	Glide	Asp	CC	h	Lat	Fric	Ej Fric	Ej Stop
Nasal	5.24	4.00	4.16	3.87	3.40	2.84	3.24	2.65	2.57	1.89	1.63
Null		3.36	2.73	2.55	2.09	1.81	2.05	1.34	0.30	-0.50	0.81
Stop			2.57	2.22	1.71	1.77	2.29	2.01	1.32	0.17	-0.37
Glide				1.41	1.55	1.63	1.98	0.65	0.95	0.27	-0.23
Asp					0.35	0.07	1.12	0.09	-0.13	-19.17	-19.44
CC						-0.75	1.13	0.62	-0.38	-2.28	-2.60
h							-17.70	-0.64	-0.27	-18.34	-1.33
Lat								-16.70	-0.46	-17.83	-18.11
Fric									-2.93	-3.59	-1.03
Ej Fric										-18.97	-1.97
Ej Stop											-19.55

References

[1] Pierrehumbert, J. (1993). Dissimilarity in the Arabic verbal roots. NELS 23:367–381.

[2] Wilson, C. and Obdeyn, M. (2009). Simplifying subsidiary theory: statistical evidence from Arabic, Muna, Shona, and Wargamay. JHU ms.

[3] Buechel, E. and Manhart, P. (2002). Lakota dictionary: Lakota-English/English-Lakota. University of Nebraska Press, Lincoln, NE.

[4] Green, C. and Davis, S. (2014). Superadditivity and Limitations on Syllable Complexity in Bambara Words. In Ashley W. Farris-Trimble and Jessica Barlow (eds.) *Perspectives on Phonology, Theory, and Acquisition: Papers in honor of Daniel A. Dinnsen.* Philadelphia/Amsterdam: John Benjamins Co.

[5] Shih, S. (2016). Super additive similarity in Dioula tone harmony. In Kim, Umbal, Block, Chan, Cheng, Finney, Katz, Nickel-Thompson, Shorten (eds). *WCCFL 31.* Cascadilla Proceedings Project.

[6] Breiss, C. and Albright, A. (2022). Cumulative markedness effects and (non-)linearity in phonotactics. *Glossa* 7:1–32.