PROBABILISTIC SYLLABLE STRUCTURE

Outline. What types of syllables are possible in a language? What types of syllables are favored? These two questions may seem independent, but we show that they are deeply connected. Just like the Jakobsonian categorical syllable typology {CV, CVC, VC, V} can be derived from ranked constraints (Prince and Smolensky 1993/2004), probabilistic typologies that arrange syllables by their relative well-formedness can be derived from probabilistic grammars in terms of uniform probability inequalities, also known as PROBABILISTIC IMPLICATIONAL UNIVERSALS (Anttila and Magri 2018). Evidence from Dagaare (`dga`) and Finnish (`fin`) reveals that the same probabilistic universals hold true in two unrelated languages that are sufficiently well documented for detailed comparison. Surprisingly, these universals only arise under the probabilistic interpretation of two theories: Optimality Theory (OT) and Harmonic Grammar (HG). In particular, Maximum Entropy (MaxEnt) typologies are so unrestrictive that no syllable is predicted to be universally worse than any other syllable.

Data. Our Finnish data come from a 16-million-word *Aamulehti 1999* corpus of written newspaper Finnish. Our sample contains all the words with a text frequency of at least 100, approximately 15,000 words, machine-syllabified by the FINNSYLL-syllabifier (Shapiro et al. 2017) and manually verified for correctness. This resulted in approximately 48,000 syllables. Our Dagaare data come from Ali et al. 2021, a dictionary with 7,075 lemmas, syllabified by us based on a few simple rules, in particular CC → C.C, with digraphs interpreted as single segments (e.g., /ŋm/), VCV → V.CV, and VV → V.V if the two vowels were not identical or any of the known diphthongs (Kennedy 1966, Bodomo 1997). This resulted in approximately 18,000 syllables.

To find out whether the expected phonological patterns hold in the data, we fitted linear regression models to both data sets, with each syllable's log type frequency as the response variable, with the following predictors: onset, coda, complex onset, complex coda (logical, present vs. absent) and the number of segments (integer). In Finnish, the presence of an onset increased the log frequency of a syllable by 0.64 ($b = 0.64$, $SE = 0.11$, $t = 5.834$, $p < 0.001$) and the presence of a coda decreased it by 0.35 ($b = -0.35$, $SE = 0.07$, $t = -4.781$, $p < 0.001$). In Dagaare, the presence of an onset increased the log frequency of a syllable by 1.05 ($b = 1.05$, $SE = 0.09$, $t = 11.580$, $p < 0.001$) and the presence of a coda decreased it by 0.29 ($b = -0.29$, $SE = 0.04$, $t = -7.605$, $p < 0.001$). Similar results hold for complex onsets and complex codas in Finnish (Dagaare has no complex margins) and the number of segments in both languages, all of which significantly decrease the syllable's frequency. We also fitted more complex models, including models with syllable as a random intercept, with similar results.

Theory. The typologies predicted by categorical OT and HG are often large. The typologies predicted by probabilistic Noisy HG and MaxEnt are always infinite and cannot be exhaustively inspected or enumerated. The ℂ𝕠𝔾𝕖𝕋𝕠 software (Magri and Anttila 2019, `https://cogeto.stanford.edu`) nevertheless allows one to explore such typologies in terms of implicational universals. The idea is to characterize cases where a grammar with one phonological mapping necessarily contains another mapping (the categorical case) or where one phonological mapping has a probability smaller than another mapping and this probability inequality holds uniformly for every grammar in the typology (the probabilistic case).

We adopted nine constraints familiar from the literature: ONSET, *CODA, MAX(V), MAX(C), DEP(V), DEP(C), *CXONSET, *CXCODA, and *SEG. We used ℂ𝕠𝔾𝕖𝕋𝕠 to compute the uniform probability inequalities for 17 syllable types generated by the template (C)(C)V(V)(C)(C), leaving out the most marked syllable type CCVVCC not found in either language in our current dataset. A few marginal syllable types were excluded for the purposes of evaluation. In Finnish, we excluded consonant clusters longer than two (CCC, CCCC) that we considered non-native. In Dagaare, apparent VVV-syllables left intact by our syllabification rules were excluded as misanalyses; we believe that in reality they are either VV.V or V.VV. We further excluded 80 instances of a C-syllable, mostly /m/ and /l/, that may be syllabic sonorants comparable to V-nuclei. Finnish instantiates all the 17 types, Dagaare instantiates 8.

Since our data are phonotactic and do not involve alternations, we focused on categorical and implicational universals that compare faithful mappings with identical underlying and surface forms. The categorical OT typology predicts 58 such implicational universals plotted in Figure 1. Each arrow (line between nodes, read from top to bottom) corresponds to a one-step improvement. For example, CCVVC heads three arrows because it admits three one-step improvements: the coda can be improved by deletion, yielding CCVV; the complex onset can be improved to a simplex onset, yielding CVVC; or the complex nucleus can be simplified, yielding CCVC. The implicational universals predicted by OT thus recapitulate the markedness hierarchy. Each arrow in Figure

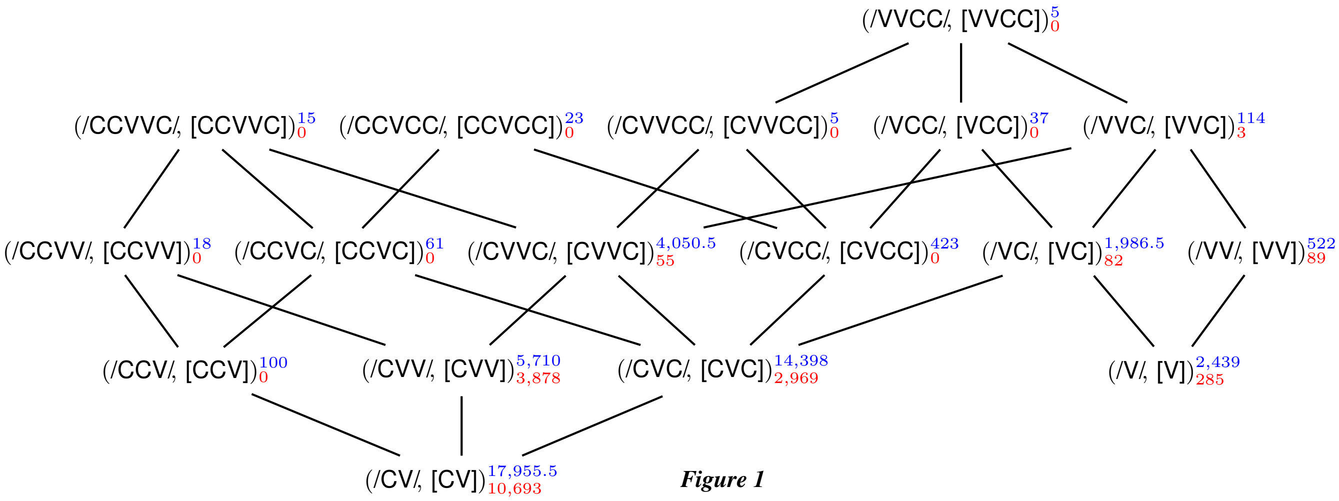
(/VVCC/, [VVCC])$^{5}_{0}$
(/CCVVC/, [CCVVC])$^{15}_{0}$
(/CCVCC/, [CCVCC])$^{23}_{0}$
(/CVVCC/, [CVVCC])$^{5}_{0}$
(/VCC/, [VCC])$^{37}_{0}$
(/VVC/, [VVC])$^{114}_{3}$
(/CCVV/, [CCVV])$^{18}_{0}$
(/CCVC/, [CCVC])$^{61}_{0}$
(/CVVC/, [CVVC])$^{4,050.5}_{55}$
(/CVCC/, [CVCC])$^{423}_{0}$
(/VC/, [VC])$^{1,986.5}_{82}$
(/VV/, [VV])$^{522}_{89}$
(/CCV/, [CCV])$^{100}_{0}$
(/CVV/, [CVV])$^{5,710}_{3,878}$
(/CVC/, [CVC])$^{14,398}_{2,969}$
(/V/, [V])$^{2,439}_{285}$
(/CV/, [CV])$^{17,955.5}_{10,693}$

Figure 1

1 admits a probabilistic interpretation: the probability of an antecedent (upper) mapping is always smaller than or equal to the probability of the consequent (lower) mapping. To test this quantitative interpretation empirically, we annotated each faithful mapping in Figure 1 with the corresponding syllable type frequency in Finnish (blue) and Dagaare (red). Assuming that these counts reflect probabilities, all the predicted implications turn out empirically true in both data sets: the frequency of the antecedent is at most as high as that of the consequent. This serves as an empirical validation of the constraints that constitute the model.

A key finding is that given these candidates and constraints MaxEnt does not predict a single one of the probabilistic implicational universals in Figure 1. Since any universal of MaxEnt is also a universal of OT, as shown in [redacted], we conclude that the MaxEnt typology predicts no implicational universals among faithful mappings. Each faithful mapping can have a larger MaxEnt probability than any other faithful mapping. In other words, MaxEnt predicts no markedness asymmetries. Concretely, this means that it is possible to find MaxEnt weights that predict, say, VVCC (5 tokens in Finnish) to have a higher probability than CV ($\approx$ 17,000 tokens in Finnish).

To diagnose this MaxEnt pathology, we denote by $\overline{F}(\mathsf{x})$ the average number of violations assigned by a faithfulness constraint F to the candidates of the underlying form x, that is, the sum of those violations divided by the number of candidates. As [redacted] shows, if an implication $(\mathsf{x}, \mathsf{x}) \rightarrow (\widehat{\mathsf{x}}, \widehat{\mathsf{x}})$ between two faithful mappings is a MaxEnt universal, the number of average antecedent faithfulness violations cannot be larger than the number of average consequent faithfulness violations, namely $\overline{F}(\mathsf{x}) \leq \overline{F}(\widehat{\mathsf{x}})$. Crucially, the average number $\overline{\textsc{Max}(\mathrm{V})}$ of vowel deletions grows as the number of underlying vowels grows. Conversely, the average number $\overline{\textsc{Dep}(\mathrm{V})}$ of vowel epentheses decreases as the number of underlying vowels grows. The average inequalities $\overline{\textsc{Max}(\mathrm{V})}(\mathsf{x}) \leq \overline{\textsc{Max}(\mathrm{V})}(\widehat{\mathsf{x}})$ and $\overline{\textsc{Dep}(\mathrm{V})}(\mathsf{x}) \leq \overline{\textsc{Dep}(\mathrm{V})}(\widehat{\mathsf{x}})$ thus entail that, if an implication $(\mathsf{x}, \mathsf{x}) \rightarrow (\widehat{\mathsf{x}}, \widehat{\mathsf{x}})$ between faithful mappings is a MaxEnt universal, the two forms x and $\widehat{\mathsf{x}}$ compared must have the same number of vowels. By reasoning analogously for $\textsc{Max}(\mathrm{C})$ and $\textsc{Dep}(\mathrm{C})$, we conclude that they also must have the same number of consonants. Out of the 58 implications in Figure 1, 56 compare antecedent and consequent strings that differ in the number of either vowels or consonants. Their failure is thus straightforwardly predicted.

Finally, the two remaining implications (/VC/, [VC]) $\rightarrow$ (/CV/, [CV]) and (/VVC/, [VVC]) $\rightarrow$ (/CVV/, [CVV]) compare antecedent and consequent strings that have the same number of consonants and vowels. They only differ in whether the consonant belongs to the onset or the coda. The diagnosis of their failure in MaxEnt is more complex. To illustrate, the vertical axis of Figure 2 plots the difference between the MaxEnt probability of the consequent (/CV/, [CV]) minus that of the antecedent (/VC/, [VC]) when Max and *CxCoda share the same weight (plotted on the horizontal axis) while the other weights are small. As this difference is negative, the implicational universal fails.

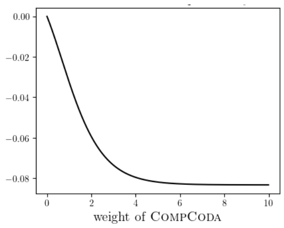

Figure 2

Conclusion. Descriptively, MaxEnt models can be very successful and often fit the data closely, but these descriptive gains often come with explanatory losses. The fact that MaxEnt predicts many unnatural and hence unattested probabilistic relations among syllable types suggests that it is not a satisfactory basis for a theory of natural language phonology.