Probabilistic syllable structure

Emiyare Ikwut-Ukwa[1], Kushal Thaman[1], Annalisa Welinder[1], Arto Anttila[1], and Giorgio Magri[2]

Stanford University[1]
SFL, CNRS, University of Paris VIII[2]

WCCFL 42

UC Berkeley | 13 April 2024

Overview

- Two questions that may seem independent:
  - What types of syllables are possible in a language?
  - What types of syllables are favored?
- The message:
  - These questions are deeply connected.
  - The same phonological principles predict both . . .
  - . . . given the right theory of phonology.
- Given the same set of syllable structure constraints
  - **Optimality Theory (OT)** predicts universals that are empirically supported.
  - **Maximum Entropy Grammars (MaxEnt)** are so unrestrictive that no syllable is predicted to be universally worse than any other syllable.

Data

- Finnish data:
  - Corpus: 16 million words of newspaper text (*Aamulehti 1999*)
  - Sample: Words with a frequency of ≥ 100, about 15,000 words
  - Machine-syllabified by FINNSYLL (Shapiro et al. 2017)
  - Syllabification manually verified for correctness
  - Approximately 48,000 syllables total

- Dagaare data:
  - A dictionary with 7,075 lemmas (Ali et al. 2021) syllabified by us
  - Some rules: CC $\rightarrow$ C.C, VCV $\rightarrow$ V.CV, VV $\rightarrow$ V.V if the vowels were not identical or known diphthongs (Kennedy 1966, Bodomo 1997)
  - Digraphs were interpreted as single segments (e.g., /ŋm/)
  - Approximately 18,000 syllables total

- Finnish example words:

spelling	syllabification	translation	template
alkaen	al.ka.en	beginning from	VC.CV.VC
torstai	tors.tai	Thursday	CVCC.CVV
poliisilaitos	po.lii.si.lai.tos	police station	CV.CVV.CV.CVV.CVC

- Dagaare example words:

spelling	syllabification	translation	template
yiri	jí.rì	house	CV.CV
yoɔraa	jʊ́ɔ.ráà	tourist	CVV.CVV
kpageloo	kpág.lʊ́ʊ́	firm	CVC.CVV

- Top 10 rows of Finnish syllabary (by type frequency)

Syllable	Stress	Template	Weight	Vowel	Frequency
ta	U	CV	L	A	2208
si	U	CV	L	I	1190
ti	U	CV	L	I	1179
sa	U	CV	L	A	1138
a	U	V	L	A	864
tä	U	CV	L	Ä	815
li	U	CV	L	I	791
la	U	CV	L	A	764
nen	U	CVC	H	E	684
le	U	CV	L	E	612

- Top 10 rows of Dagaare syllabary (by type frequency)

Syllable	Initial	Template	Weight	+ATR	Frequency
rì	false	CV	L	false	517
rí	false	CV	L	false	430
rì	false	CV	L	true	296
rí	false	CV	L	true	264
lì	false	CV	L	false	240
gì	false	CV	L	false	213
lí	false	CV	L	false	199
gí	false	CV	L	false	183
ní	false	CV	L	false	170
ráá	false	CVV	H	false	167

- Exclusions: A few marginal syllable types were omitted when evaluating the theories.
- In Finnish, we excluded consonant clusters longer than two (CCC, CCCC) as non-native.
- In Dagaare, we excluded the following:
  - Apparent VVV-syllables left intact by our syllabification rules. They seem to be either VV.V or V.VV but we currently don't know which.
  - 80 instances of a C-syllable, mostly /m/ and /l/, that may be syllabic sonorants comparable to V-nuclei.

Modeling

- A working phonologist might expect to find some basic syllable structure asymmetries. We fitted linear regression models to the data to verify that those asymmetries are indeed there in both Finnish and Dagaare.
  - **Response variable**: the syllable's log type frequency
  - **Predictor variables**:

Predictor	Expected effect
onset	presence increases frequency
coda	presence decreases frequency
complex onset	presence decreases frequency
complex coda	presence decreases frequency
number of segments	more segments decreases frequency

- Onset vs. coda in Finnish:
  - An onset increased the log frequency of a syllable by 0.64 ($b = 0.64$, $SE = 0.11$, $t = 5.834$, $p < 0.001$)
  - A coda decreased it by 0.35 ($b = -0.35$, $SE = 0.07$, $t = -4.781$, $p < 0.001$).
- Onset vs. coda in Dagaare:
  - An onset increased the log frequency of a syllable by 1.05 ($b = 1.05$, $SE = 0.09$, $t = 11.580$, $p < 0.001$).
  - A coda decreased it by 0.29 ($b = -0.29$, $SE = 0.04$, $t = -7.605$, $p < 0.001$).

- The expected effects are seen for the other predictors as well.
- Complex onsets and complex codas in Finnish (Dagaare has no complex margins) and the number of segments in both languages significantly decrease the syllable's frequency.
- We also fitted more complex models, including mixed models with syllable as a random intercept, with similar results.

- The regression modeling shows that there is something to study. But are these facts **predicted** by any theory of phonology?
- In particular, given a set of syllable structure constraints and an arbitrary ranking (OT) or weighting (HG, MaxEnt), do the empirical asymmetries follow?
- Constraints: [Prince and Smolensky 1993/2004]
  - Onset, *Coda, *CxOnset, *CxCoda, *Seg
  - MaxV, MaxC, DepV, DepC
- Candidates:
  - 17 syllable types generated by the template $(C)(C)V(V)(C)(C)$
  - $CCVVCC$ omitted because not found in either language
  - all syllable types are candidates of each other

- We focus on IMPLICATIONAL UNIVERSALS that compare faithful mappings with identical underlying and surface forms:

$$\begin{aligned} (y, y) &\rightarrow (\hat{y}, \hat{y}) \\ (/CCVCC/, [CCVCC]) &\rightarrow (/CV/, [CV]) \end{aligned}$$

- This seems appropriate because our data are phonotactic (comparative well-formedness of syllable types) and do not directly involve alternations.

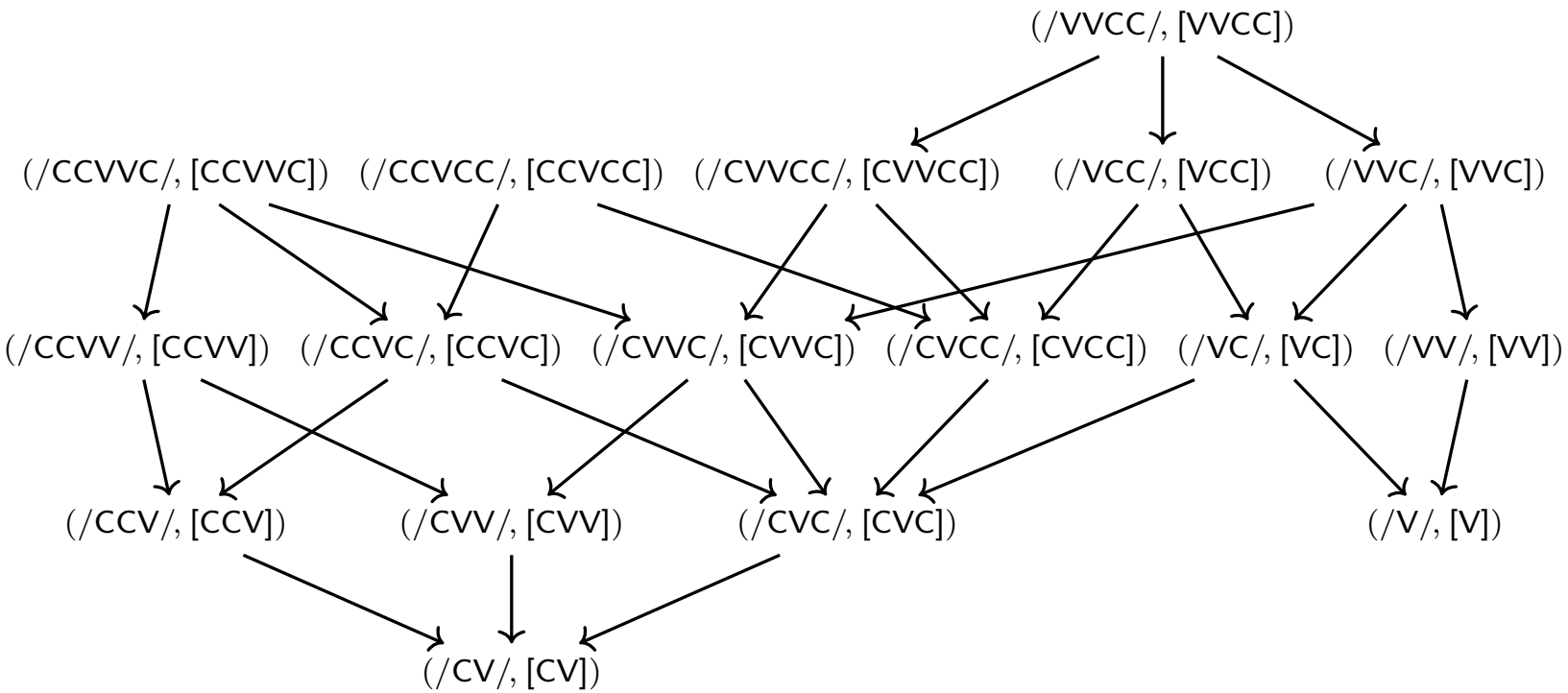
- We interpret these universals as follows:
  - every CATEGORICAL grammar that realizes /*CCVCC*/ faithfully also realizes /*CV*/ faithfully
  - every PROBABILISTIC grammar realizes /*CCVCC*/ faithfully with probability no larger than the probability with which it realizes /*CV*/ faithfully

- We used CoGeTo (`https://cogeto.stanford.edu/`) to compute the predicted universals. [Magri and Anttila 2019]

Theoretical results

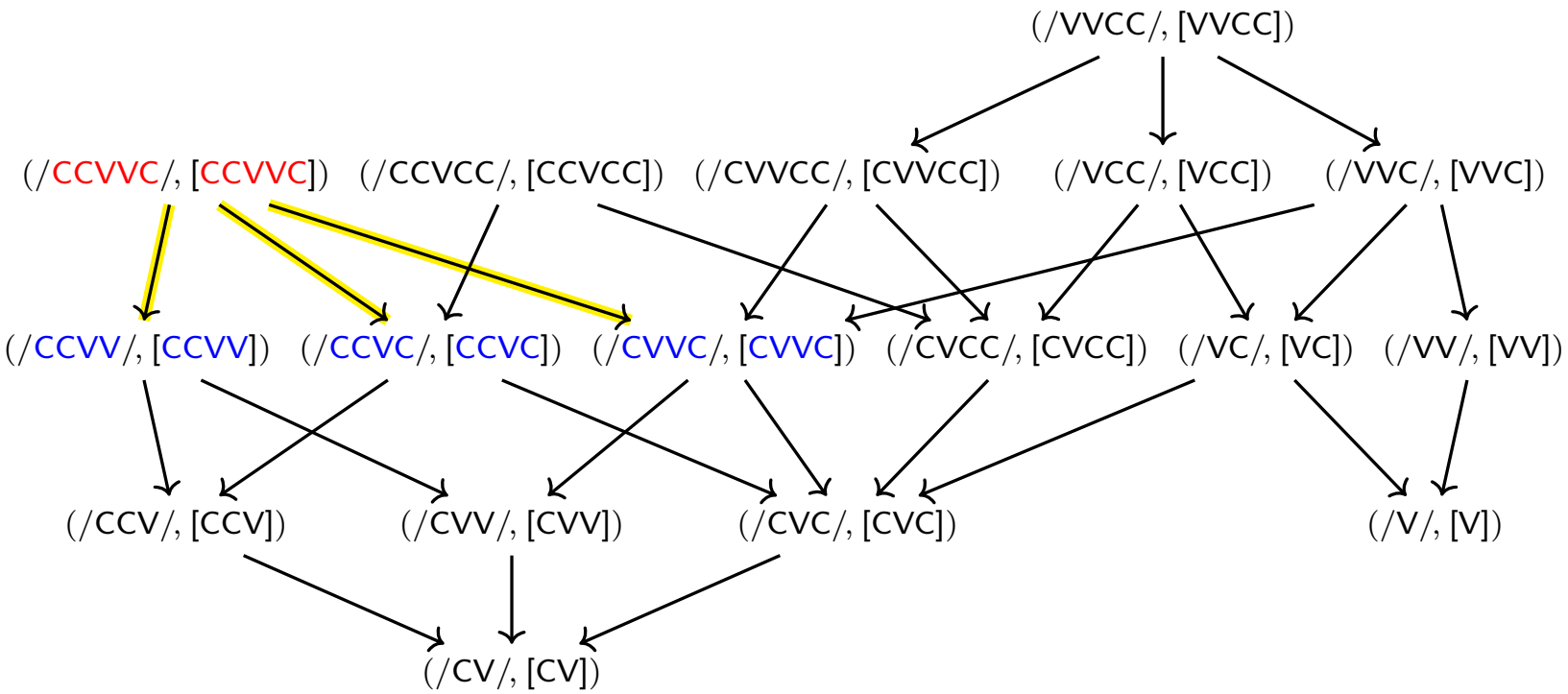
First result:

- Categorical OT predicts 58 implicational universals
- Each arrow corresponds to a one-step improvement
- OT universals thus recapitulate the markedness hierarchy



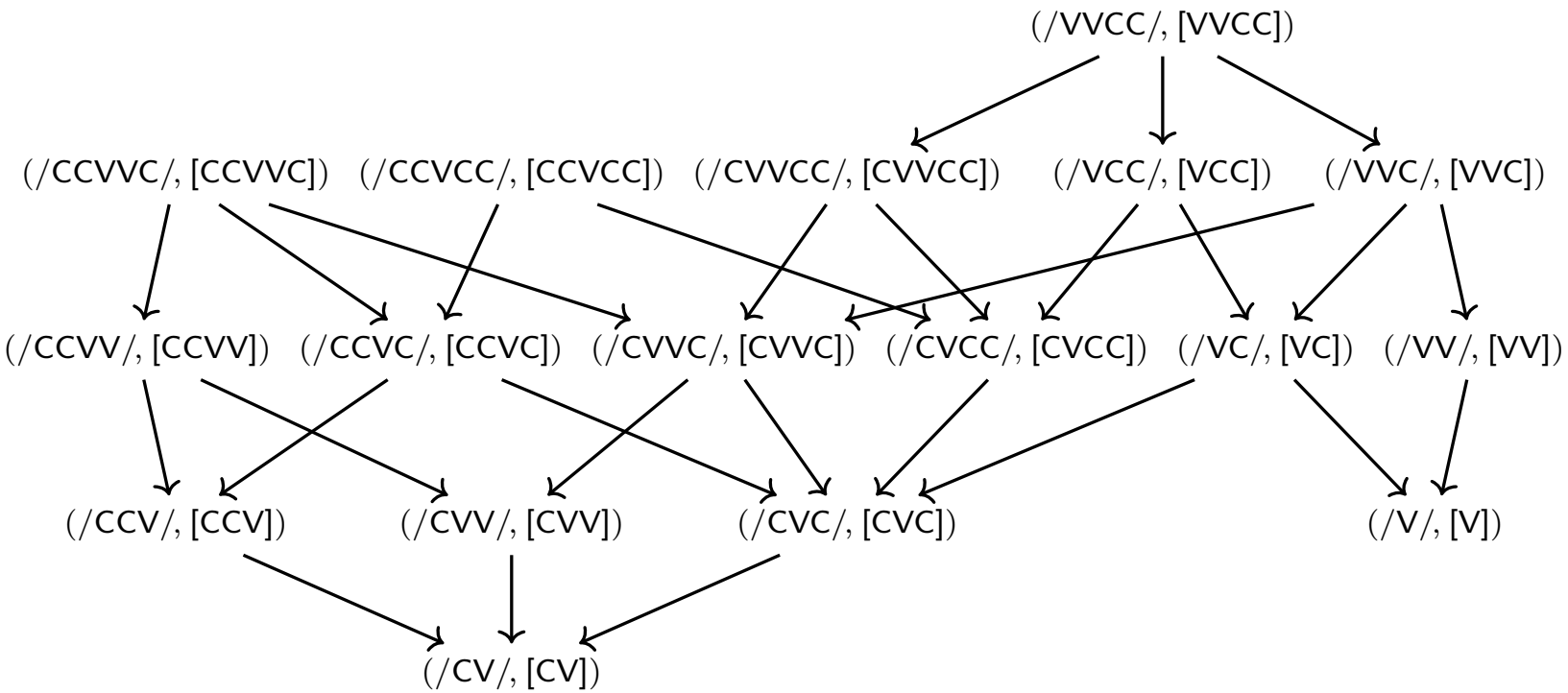
First result:

- Categorical OT predicts 58 implicational universals
- Each arrow corresponds to a one-step improvement
- OT universals thus recapitulate the markedness hierarchy



First result:

- Categorical OT predicts 58 implicational universals
- Each arrow corresponds to a one-step improvement
- OT universals thus recapitulate the markedness hierarchy

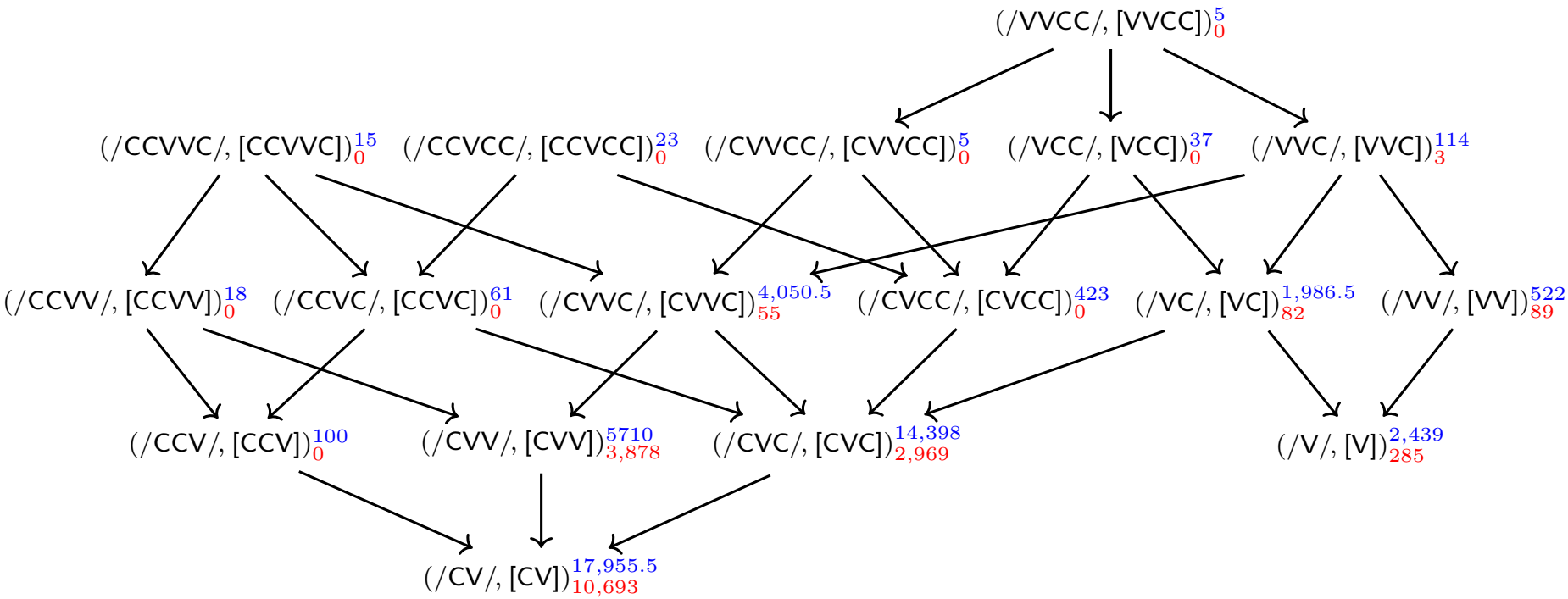


Second result:

- Each arrow admits a probabilistic interpretation:

  probability of antecedent $\leq$ probability of consequent

- To test this interpretation, we annotate each faithful mapping with the number of occurrences of the corresponding syllable type in Finnish and Dagaare
- Assuming that these counts reflect probabilities, all the predicted implications turn out empirically true in both data sets



Third result:

- ME (with these candidates and constraints) does not predict a single one of the implicational universals plotted
- Since a ME universal is also an OT universal, we conclude that ME predicts no implicational universals among faithful mappings

[Anttila and Magri 2018a; Magri and Anttila 2024]

- Each faithful mapping can have a larger ME probability than any other faithful mapping
  - E.g.: some ME weights predict $VVCC$ (5 tokens in Finnish) to have a higher probability than CV (~17,000 tokens in Finnish)
- *ME predicts no markedness asymmetries*

Why are ME's predictions empty?

First step:

- To diagnose this ME pathology, we denote by $\overline{F}(x)$ the average number of violations assigned by a faithfulness constraint F to the candidates of the underlying form x:

$$\overline{F}(x) = \frac{1}{|Gen(x)|} \sum_{y \in Gen(x)} F(x, y)$$

- If an implication $(x, x) \rightarrow (\hat{x}, \hat{x})$ between two faithful mappings is a ME universal, the average number of antecedent faithfulness violations cannot be larger than the average number of consequent faithfulness violations: [Magri and Anttila 2024]

$$\overline{F}(x) \leq \overline{F}(\hat{x})$$

Second step:

- The average number $\overline{\text{MaxV}}$ of vowel deletions grows as the number of underlying vowels grows
- Conversely, the average number $\overline{\text{DepV}}$ of vowel epentheses decreases as the number of underlying vowels grows
- Thus the inequalities $\overline{\text{MaxV}}(x) \leq \overline{\text{MaxV}}(\hat{x})$ and $\overline{\text{DepV}}(x) \leq \overline{\text{DepV}}(\hat{x})$ entail that, if $(x, x) \rightarrow (\hat{x}, \hat{x})$ is a ME universal, the two forms x and $\hat{x}$ compared must have the same number of vowels
- By reasoning analogously for MaxC and DepC, we conclude that they also must have the same number of consonants
- Out of the 58 implications in the figure above, 56 compare antecedent and consequent strings that differ in the number of either vowels or consonants
- Their failure is thus straightforwardly predicted

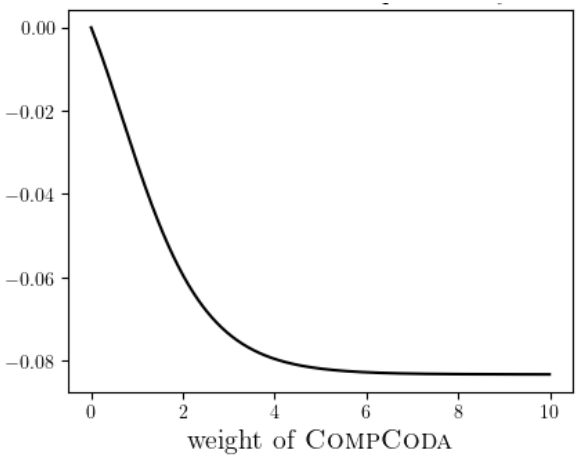
Third step:

- The remaining implications

  (/*VC*/, [*VC*]) → (/*CV*/, [*CV*]) and (/*VVC*/, [*VVC*]) → (/*CVV*/, [*CVV*])

  compare strings that have the same number of vowels and the same number of consonants

- They fail because the difference (vertical axis) between the ME probability of (/*CV*/, [*CV*]) minus that of (/*VC*/, [*VC*]) is negative when MaxC and *CxCoda share the same weight (horizontal axis) while the other weights are small



- The diagnosis of their failure in MaxEnt is more complex

[Magri and Anttila 2024]

Conclusions

- It is well known that ranked constraints (Prince and Smolensky 1993/2004) predict universals, e.g., factorial typologies of syllable types like {CV, CVC, VC, V}.
- It is less well known that probabilistic grammars also predict universals, e.g., by arranging syllables by their relative probability

  [Anttila and Magri 2018b]
- Evidence from Dagaare and Finnish shows that the same syllable structure universals (categorical, probabilistic) hold true in two unrelated languages, as predicted by Optimality Theory (OT) but not by Maximum Entropy (MaxEnt).

Thank you!

References

Anttila, Arto, and Giorgio Magri. 2018a. Does MaxEnt overgenerate? Implicational universals in Maximum Entropy grammar. In *AMP 2017: Proceedings of the 2017 Annual Meeting on Phonology*, ed. Gillian Gallagher, Maria Gouskova, and Yin Sora. Washington, DC: Linguistic Society of America.

Anttila, Arto, and Giorgio Magri. 2018b. T-orders across categorical and probabilistic constraint-based phonology. Manuscript (Stanford, CNRS).

Magri, Giorgio, and Arto Anttila. 2019. CoGeTo: Convex geometry tools for typological analysis in categorical and probabilistic constraint-based phonology (version 1.0). Available at `https://cogeto.stanford.edu`.

Magri, Giorgio, and Arto Anttila. 2024. Principles of maximum entropy phonology.

Prince, Alan, and Paul Smolensky. 1993/2004. *Optimality Theory: Constraint interaction in generative grammar*. Oxford: Blackwell. URL `http://roa.rutgers.edu`, original version, Technical Report CU-CS-696-93, Department of Computer Science, University of Colorado at Boulder, and Technical Report TR-2, Rutgers Center for Cognitive Science, Rutgers University, April 1993. Available from the Rutgers Optimality Archive as ROA 537.