Implicational universals with negated consequents in Maximum Entropy Grammars

- Typologies of categorical grammars can be large and sometimes infinite, as in HG. Typologies of probabilistic grammars are always infinite, as in ME (MaxEnt), SOT (Stochastic OT), and SHG (Stochastic or Noisy HG). A natural strategy to investigate the linguistic structure of large typologies is to focus on the **implicational universals** they encode rather than on the grammars they list. Anttila & Magri (AM; 2018, 2019, 2022, 2023) say that $(\mathsf{x}, \mathsf{y}) \rightarrow (\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ is a universal of a typology of **categorical** grammars (OT, HG) provided every grammar in the typology that realizes the antecedent underlying form x as the antecedent surface form y, also realizes the consequent underlying form $\widehat{\mathsf{x}}$ as the consequent surface form $\widehat{\mathsf{y}}$. Furthermore, AM say that $(\mathsf{x}, \mathsf{y}) \rightarrow (\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ is a universal of a typology of **probabilistic** grammars (ME, SOT, SHG) provided every grammar G in the typology assigns as much probability to the consequent mapping $(\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ than to the antecedent mapping (x, y), as in fig. 1. Thus, in both categorical and probabilistic settings, this implicational universal $(\mathsf{x}, \mathsf{y}) \rightarrow (\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ captures the intuition that the mapping $(\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ is "better" because "easier to get" than the mapping (x, y).

$$\boxed{G(\mathsf{y}\,|\,\mathsf{x}) \leq G(\widehat{\mathsf{y}}\,|\,\widehat{\mathsf{x}})}$$

Figure 1

- AM found that ME validates surprisingly few universals. We illustrate, with Prince and Smolensky's Basic Syllable System (BSS). The OT/HG typology satisfies the universals represented by the solid and dotted arrows in fig. 2. Yet, all of the many dotted arrows fail in ME. Many of these failures are paradoxical. For instance, failure of the yellow arrows means that VC (with both marked onset and coda) is not worse (can have larger ME probability) than CVC and V (with either a marked coda or onset). Furthermore, fig. 2 only plots mappings that are HG possible. When the antecedent mapping (x, y) is HG impossible (no HG grammar contains it), the implication $(\mathsf{x}, \mathsf{y}) \rightarrow (\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ is an HG universal, no matter the choice of the consequent mapping $(\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ or the phonological connection between the two mappings. Yet, AM found that most of these HG universals with impossible antecedents fail in ME. To illustrate with the BSS, ME misses 66 of the HG universals with impossible antecedents. Again, many of these failures are paradoxical: (/CV/, [CVC]) $\rightarrow$ (/CVC/, [CV]) fails in ME because we can construct ME weights such that the ME probability of coda epenthesis in the impossible antecedent is larger than the ME probability of coda deletion by a staggering 0.5!

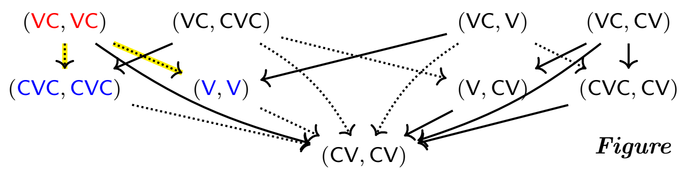

Figure 2

- AM conclude that ME typologies encode **little** linguistic structure. Yet, ME typologies must encode **some** structure after all and AM's universals are unable to extract and describe it. This talk develops the theory of implicational universals **with negated consequents** and argues they are better suited to study ME typologies. We say that $(\mathsf{x}, \mathsf{y}) \rightarrow \overline{(\widehat{\mathsf{x}}, \widehat{\mathsf{y}})}$ is a universal of a typology of **categorical** grammars (HG, OT) provided every grammar in the typology that realizes x as y, **does not** also realize $\widehat{\mathsf{x}}$ as $\widehat{\mathsf{y}}$. Furthermore, we say that $(\mathsf{x}, \mathsf{y}) \rightarrow \overline{(\widehat{\mathsf{x}}, \widehat{\mathsf{y}})}$ is a universal of a typology of **probabilistic** grammars (ME, SOT, SHG) provided the probability of realizing x as y is never larger than the probability of **not** realizing $\widehat{\mathsf{x}}$ as $\widehat{\mathsf{y}}$, as in fig. 3. We present four results on this new class of universals.

$$\boxed{G(\mathsf{y}\,|\,\mathsf{x}) \leq 1 - G(\widehat{\mathsf{y}}\,|\,\widehat{\mathsf{x}})}$$

Figure 3

- For HG, the computation of these universals is straightforward. But for ME, the problem is not trivial. Our first result is a complete characterization of ME universals with negated consequents. Indeed, suppose that the antecedent mapping (x, y) comes with m loser candidates $\mathsf{z}_1, \ldots, \mathsf{z}_m$ and the consequent mapping $(\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ with $\widehat{m}$ loser candidates $\widehat{\mathsf{z}}_1, \ldots, \widehat{\mathsf{z}}_{\widehat{m}}$. Then, $(\mathsf{x}, \mathsf{y}) \rightarrow \overline{(\widehat{\mathsf{x}}, \widehat{\mathsf{y}})}$ is a ME universal if and only if there exist m coefficients α_i and $\widehat{m}$ coefficients β_j all non-negative but not all equal to zero that satisfy the inequality in fig. 4a for every constraint C as well as the identity in fig.4b. We provide some intuition behind these conditions. Since these are all linear conditions in α_i, β_j, they can be solved with any linear programming library. Python code that uses this result to compute ME universals with negated consequents is made available at `[omitted]`. We thus improve on AM, who only provide necessary and (computationally costly) sufficient conditions for their universals, but are unable to close the gap between them.

Figure 4a

$$\sum_i \alpha_i \Big(C(\mathsf{x}, \mathsf{z}_i) - C(\mathsf{x}, \mathsf{y}) \Big) + \sum_j \beta_j \Big(C(\widehat{\mathsf{x}}, \widehat{\mathsf{z}}_j) - C(\widehat{\mathsf{x}}, \widehat{\mathsf{y}}) \Big) \leq 0$$

Figure 4b

$$\sum_i \alpha_i = \sum_j \beta_j$$

- □ Furthermore, we show that $(\mathsf{x},\mathsf{y}) \to \overline{(\widehat{\mathsf{x}},\widehat{\mathsf{y}})}$ is an HG universal if and only if there exist α_i, β_j as in fig. 4a, irrespectively of fig. 4b. Hence, our second result is that, whenever $(\mathsf{x},\mathsf{y}) \to \overline{(\widehat{\mathsf{x}},\widehat{\mathsf{y}})}$ is an ME universal (satisfies both figs. 4a and 4b), it is also an HG universal (satisfies fig. 4a). This replicates AM's analogous result for their original universals. What about the reverse: do HG universals with negated consequents survive in ME, contrary to AM's universals?
- □ In order to address this question, we have systematically computed the universals with negated consequents predicted by ME for the same test cases used by AM. In reporting our results, we distinguish between possible and impossible mappings. Like AM's original universals, our universals with negated consequents also hold trivially in HG when one of the mappings is HG **impossible**, no matter what the other mapping looks like or the phonological connection between the two mappings. Thus, we expect many HG universals with negated consequents that feature an HG impossible mapping to fail in ME, just as AM's original universals, which is indeed what we found.
- □ Yet, our third result is that the situation is rather different when we restrict ourselves to HG **possible** mappings. In this case, we indeed find that in almost all of AM's test cases (see below for discussion of the counterexamples), HG and ME share exactly the same implicational universals with negated consequents. This is the opposite of what AM observe for their universals, as recalled above. We conclude that our new universals with negated consequents do a better job than AM's original universals at extracting the linguistic structure encoded by ME typologies.
- □ We illustrate again with the BSS. Fig. 5 repeats AM's original universals for HG from fig. 1, the only difference being that we now ignore (CV,CV). The reason for ignoring this mapping is that: (a) it is HG necessary (every HG grammar contains it); (b) a necessary mapping can feature in a universal with negated consequents only when the other mapping is impossible; (c) we are focusing here on possible mappings only. Fig. 6 plots the HG universals with negated consequents. While fig. 5 has directed edges because AM's universal $(\mathsf{x},\mathsf{y}) \to (\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ is different from the reverse $(\widehat{\mathsf{x}},\widehat{\mathsf{y}}) \to (\mathsf{x},\mathsf{y})$, fig. 6 instead has undirected edges because the universal $(\mathsf{x},\mathsf{y}) \to \overline{(\widehat{\mathsf{x}},\widehat{\mathsf{y}})}$ with a negated consequent is equivalent to the reverse $(\widehat{\mathsf{x}},\widehat{\mathsf{y}}) \to \overline{(\mathsf{x},\mathsf{y})}$. All lines are solid in fig. 6 because these HG universals with negated consequents **all survive** in ME. To illustrate, the AM universal represented by the yellow arrow in fig. 5 says that, if an HG grammar faithfully realizes VC, then it faithfully realizes CVC. This arrow is dotted because this universal paradoxically fails in ME: the ME probability of faithfully realizing VC can be larger than the ME probability of faithfully realizing CVC. The universal with negated consequent represented by the yellow line in fig. 6 says that, if an HG grammar faithfully realizes VC, then it cannot delete the coda of CVC. This arrow is solid because this universal survives in ME: the ME probability of faithfully realizing VC is never larger than the ME probability of not deleting the coda of CVC. In general, we have found that the undirected graph of ME universals with negated consequents gets rid of the many holes and asymmetries in the directed graph of AM's original ME universals.

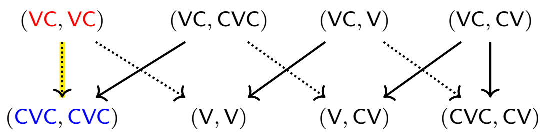

Figure 5

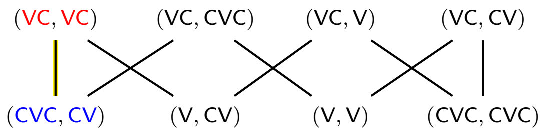

Figure 6

- □ As anticipated, we have found a few cases where HG universals with negated consequents fail in ME. Crucially, these cases share the same structure. To illustrate it, we consider $(\mathsf{x},\mathsf{y}) = (/\mathsf{da}/, [\mathsf{da}])$ and $(\widehat{\mathsf{x}},\widehat{\mathsf{y}}) = (/\mathsf{dada}/, [\mathsf{tata}])$ and the two constraints NoVoice and IdentVoice. The loser candidates $\mathsf{z}_i = [\mathsf{ta}]$ and $\widehat{\mathsf{z}}_j = [\mathsf{dada}]$ satisfy the condition in fig. 7 with $\lambda = 2$ for both constraints C. Hence, $(\mathsf{x},\mathsf{y}) \to \overline{(\widehat{\mathsf{x}},\widehat{\mathsf{y}})}$ is an HG universal because it satisfies the inequality in fig. 4a with $\alpha_i = \lambda\beta_j$ (and remaining α, β's equal to zero). It is not an ME universal because the identity in fig. 4b fails. Yet, our fourth result is that in all cases where fig. 7 holds, the sum of the ME probabilities is upper bounded by $\mathbf{S}(-\xi) + \mathbf{S}(\lambda\xi)$, where $\mathbf{S}$ is the sigmoid function and ξ is a proper argument. This upper bound $\mathbf{S}(-\xi) + \mathbf{S}(\lambda\xi)$ is plotted in fig. 8 to show that it can only barely get larger than 1 in a small region of weight space. The inequality in fig. 3 is thus **only barely flouted**, contrary to some of AM's counterexamples, where instead the inequality in fig. 1 was flouted by a large margin and for a large swath of weight space.

Figure 7

$$\boxed{C(\mathsf{x},\mathsf{z}_i) - C(\mathsf{x},\mathsf{y}) + \lambda\big(C(\widehat{\mathsf{x}},\widehat{\mathsf{z}}_j) - \mathbf{C}(\widehat{\mathsf{x}},\widehat{\mathsf{y}})\big) = 0}$$

Figure 8