Probabilistic phonology is intrinsically categorical

Giorgio Magri*
CNRS

Arto Anttila**
Stanford University

Phonological theory has recently extended its empirical coverage from categorical to quantitative, probabilistic data. What is the proper model of probabilistic phonology? We distinguish two classes of models. Intrinsically probabilistic models *such as maximum entropy (ME) postulate phonological grammars that directly assign probabilities to linguistic forms and are therefore very different from traditional categorical grammars.* Extrinsically probabilistic models *such as stochastic (or noisy) harmonic grammar (SHG) instead rely on traditional categorical grammars and derive the probabilistic behavior indirectly from the assumption that speakers have not converged on one specific categorical grammar but maintain whole probability distributions over these categorical grammars. We develop a three-pronged argument in favor of the latter approach. (I) We show that SHG comes with inference properties comparable to ME: a log-concave likelihood function and efficient techniques to obtain good approximations of its derivatives. (II) We extend the notion of implicational universals to the probabilistic setting in terms of probability inequalities that hold uniformly across a typology of probabilistic grammars. We derive several generalizations about ME uniform probability inequalities, providing the first attempt at understanding the general principles of ME phonology. We then show that these generalizations are phonologically paradoxical because they prune ME universals down to nothing, even for small sub-typologies. (III) Finally, we show that these paradoxes extend beyond ME to any intrinsically probabilistic model. We conclude that probabilistic natural language phonology is intrinsically categorical after all.*

1. Introduction

Over the past two decades, theoretical linguistics has taken a probabilistic turn. In phonology, categorical data collected through introspection and field work are now routinely complemented with probabilistic data from corpora and experiments. Probabilistic phonological data come in at least three different flavors (Hayes 2022 and references therein). First, there is VARIATION where the same underlying form admits multiple surface realizations with predictable frequencies. An example is the variable deletion of /t,d/ in English coda clusters, whereby the underlying form /cost#me/ is variably realized as [cost me] or [cos me] (Coetzee and Kawahara 2013). Second, speakers show awareness of STATISTICAL LEXICAL PATTERNS. For instance, in Hungarian, short stems with a neutral vowel /i/ mostly take front vowel suffixes, e.g., *címnek* [tsi:m-nɛk] 'address-DAT', while just a few dozen take back vowel suffixes, e.g., *hídnak* [hi:d-nɔk] 'bridge-DAT' (Siptár and Törkenczy 2000, 68). Crucially, this proportional difference emerges when

* SFL, 59 rue Pouchet, 75017 Paris, France
** Department of Linguistics, Stanford University, Stanford CA 94305-2150, U.S.A.

native speakers are tested with phonologically similar nonce words (Hayes and Londe 2006, 71-73). Third, speakers judge the well-formedness of phonological forms on a GRADIENT SCALE, beyond the dichotomy acceptable versus ill-formed. To illustrate, in Finnish, nominals like /etelä/ 'south' with two neutral vowels [e] and a non-back vowel [ä] sound acceptable, whereas nonce words like /etela/ with two neutral vowels and a back vowel [a] sound un-Finnish (Karlsson 1982, 103) although not ill-formed.

This empirical extension has required a corresponding theoretical extension. Categorical phonological models based on discrete building blocks (such as SPE rules or OT rankings) are being replaced with quantitative, probabilistic models with continuous parameters (Alderete and Finley to appear). This raises a new theoretical question: what is the proper probabilistic model of natural language phonology? "The choice [among] probabilistic frameworks is really part of linguistic theory" (Hayes 2017). This is the question addressed in our paper.

To formulate this question explicitly, section 2 distinguishes two classes of models. INTRINSICALLY PROBABILISTIC MODELS postulate phonological grammars that *directly* assign probabilities, and are therefore very different from traditional categorical grammars. EXTRINSICALLY PROBABILISTIC MODELS instead rely on traditional categorical grammars and derive the probabilistic behavior *indirectly* through the assumption that a speaker has not converged on a specific grammar but samples one whenever phonological computation needs to be performed. MAXIMUM ENTROPY (ME; Goldwater and Johnson 2003; Hayes and Wilson 2008) and STOCHASTIC (or NOISY) HARMONIC GRAMMAR (SHG; Boersma and Pater 2016) are the poster children of these two approaches.

ME has been endorsed in the phonological literature because of its classical guarantees for grammatical inference: the ME likelihood is a log-concave function of the weights that can be climbed with gradient ascent methods (Huang et al. 2010; Malouf 2013). Section 3.1 starts the comparison between these two approaches by establishing comparable guarantees for SHG. Also the SHG likelihood is log-concave (a straightforward consequence of Prékopa's theorem). And it can be climbed through gradient ascent methods because computing its derivatives means computing polyhedral volumes, something that can be done efficiently with good approximation. Guarantees for inference therefore provide no ground to choose between ME and SHG.

We thus turn to a comparison of their typological predictions. Unfortunately, ME and SHG typologies cannot be exhaustively inspected and directly compared, because they consist of infinitely many grammars. New tools are needed to analyze the linguistic structure encoded by these probabilistic typologies—that is, to do probabilistic phonology with the same theoretical ambition that has characterized categorical generative phonology in the past seventy years.

The new tool for probabilistic typological analysis developed in section 3.2 is UNIFORM PROBABILITY INEQUALITIES. The idea is to study a typology of probabilistic grammars by characterizing cases where one phonological mapping has a probability smaller than another mapping and this probability inequality holds uniformly for every grammar in the typology. To illustrate, the probability of t-deletion in coda clusters is smaller before a vowel (/cost#us/, [cos us]) than before a consonant (/cost#me/, [cos me]) and this probability inequality holds uniformly across dialects of English (Coetzee 2004; Coetzee and Kawahara 2013). These uniform probability inequalities can be interpreted as UNIVERSALS of typologies of probabilistic grammars.

These universals reveal a surprising difference between ME and SHG. Section 3.4 shows that SHG typologies predict a rich set of uniform probability inequalities. ME typologies are instead so unrestrictive that they yield almost no uniform probability

inequalities: no mapping is universally worse than any other mapping, because any mapping can have a larger ME probability than any other mapping.

This conclusion about ME's unrestrictiveness is established in sections 4-8 in two steps. The *first step* contributes a number of mathematical generalizations about ME uniform probability inequalities. These generalizations represent the first attempt we are aware of to understand the organizing principles of ME phonology, beyond circumstantial evidence of ME's ability to fit specific patterns of empirical frequencies (Zuraw and Hayes 2017; Smith and Pater 2020; Breiss and Albright 2022). The software ℂoGeTo (*Convex Geometry Tools for phonological analysis*, available at `https://cogeto.stanford.edu/about`), implements these results and thus allows the user to compare the uniform probability inequalities predicted by ME and SHG on their own data. The *second step* of our argument then shows that these mathematical generalizations are phonologically paradoxical and indeed prune the set of ME uniform probability inequalities down to nothing. We conclude that ME is not a suitable model of phonology.

How specific to ME is this conclusion? To address this question, section 9 takes a more principled look at intrinsically probabilistic phonology, beyond ME. We formalize the intuition that a sensible probabilistic grammar should not make spurious phonological distinctions and we formally characterize the harmonies that yield such grammars. Using this characterization, we show that the paradoxical generalizations uncovered in sections 4-8 for ME typologies extend to typologies of intrinsically probabilistic grammars that make no spurious distinctions. In other words, the paradoxes are not an idiosyncrasy of ME but an unavoidable feature of intrinsically probabilistic models. Section 10 concludes that the architecture of probabilistic natural language phonology is intrinsically categorical, after all.

2. Two approaches to probabilistic phonology

A PHONOLOGICAL MAPPING is a pair (x, y) consisting of an UNDERLYING FORM x and a corresponding SURFACE REALIZATION y. Gen denotes the set of mappings relevant for the description of the phonological system of interest (Prince and Smolensky 1993/2004). $Gen(\mathsf{x})$ denotes the set of CANDIDATE surface realizations y such that the mapping (x, y) belongs to Gen. We allow Gen to list countably infinitely many underlying forms. But we require the candidate set $Gen(\mathsf{x})$ of a underlying form x to be finite.

A CATEGORICAL GRAMMAR G assigns to each underlying form x a unique surface realization $\mathsf{y} = G(\mathsf{x})$. Categorical grammars are thus TOTAL (they specify a surface realization for each underlying form in Gen) and STRICT (they specify a unique surface realization per underlying form). A PROBABILISTIC GRAMMAR G assigns to each mapping (x, y) in Gen a number $G(\mathsf{y} \,|\, \mathsf{x})$ that is interpreted as the probability that x is realized as y. This probabilistic interpretation requires these numbers $G(\mathsf{y} \,|\, \mathsf{x})$ to be non-negative and NORMALIZED across the candidate set $Gen(\mathsf{x})$ of each underlying form x, namely $\sum_{\mathsf{y} \in Gen(\mathsf{x})} G(\mathsf{y} \,|\, \mathsf{x}) = 1$. A (categorical or probabilistic) TYPOLOGY is a collection of (categorical or probabilistic) grammars. We now explore two classes of models of the typology of probabilistic natural language phonology.

2.1 Intrinsically probabilistic models

We assign to each mapping (x, y) a positive numerical score $H(\mathsf{x}, \mathsf{y})$ that reflects its phonological HARMONY: better mappings have larger harmony scores. We then define the probability $G_H(\mathsf{y} \,|\, \mathsf{x})$ as this score $H(\mathsf{x}, \mathsf{y})$ divided by a constant $Z(\mathsf{x})$ that

ensures normalization, as in (1). A family $\{H, H' \ldots\}$ of harmonies yields the typology $\{G_H, G_{H'} \ldots\}$ of corresponding HARMONY-BASED grammars.

$$G_H(\mathsf{y} \,|\, \mathsf{x}) = \frac{H(\mathsf{x}, \mathsf{y})}{Z(\mathsf{x})} \tag{1}$$

This approach is intrinsically probabilistic because it assumes that a speaker has internalized a grammar G_H that directly assigns probabilities to mappings.

Implementation. To ensure that the score $H(\mathsf{x}, \mathsf{y})$ reflects the phonological harmony of the mapping (x, y), we presuppose a set $\mathbf{C}$ consisting of a finite number n of CONSTRAINTS $C_1, \ldots, C_n$. Each constraint C_k assigns to each mapping (x, y) in Gen a non-negative integer $C_k(\mathsf{x}, \mathsf{y})$ that counts how many times that mapping violates the phonological desideratum encoded by that constraint (Prince and Smolensky 1993/2004). We then take the harmony score $H(\mathsf{x}, \mathsf{y})$ to be a DECREASING function $H\big(\mathbf{C}(\mathsf{x}, \mathsf{y})\big)$ of the vector $\mathbf{C}(\mathsf{x}, \mathsf{y}) = (C_1(\mathsf{x}, \mathsf{y}), \ldots, C_n(\mathsf{x}, \mathsf{y}))$ of constraint violations: if a mapping violates each constraint at least as much as another mapping, the former cannot have a larger harmony than the latter, whereby $\mathbf{C}(\mathsf{x}, \mathsf{y}) \geq \mathbf{C}(\mathsf{x}, \mathsf{z})$ entails $H(\mathbf{C}(\mathsf{x}, \mathsf{y})) \leq H(\mathbf{C}(\mathsf{x}, \mathsf{z}))$.

How should we choose this decreasing harmony function H? Section 9 will address this question systematically. For now, we focus on the most popular choice of ME (Goldwater and Johnson 2003; Hayes and Wilson 2008). The ME grammar corresponding to a weight vector $\mathbf{w} = (w_1, \ldots, w_n)$ is the grammar (1) whose harmony score $H(\mathsf{x}, \mathsf{y}) = \exp - \sum_{k=1}^{n} w_k C_k(\mathsf{x}, \mathsf{y})$ is the exponential of the opposite of the weighted sum of constraint violations. H is decreasing provided the weights w_k are non-negative. In conclusion, our first model of probabilistic natural language phonology is the typology of ME grammars corresponding to all vectors of non-negative weights $w_k \geq 0$.

2.2 Extrinsically probabilistic models

We now turn to a fundamentally different model. We start from a categorical typology $\mathcal{T}$ of (strict and total) categorical grammars familiar from traditional, pre-probabilistic generative phonology. We assign to each grammar G in $\mathcal{T}$ a probability mass $P(G)$. Finally, we define $G_P(\mathsf{y} \,|\, \mathsf{x})$ as the probability (2) of sampling from this categorical typology $\mathcal{T}$ according to P a categorical grammar G that realizes the underlying form x as the surface form y. Normalization over candidate sets is easily verified in appendix A.1. A family $\{P, P', \ldots\}$ of probability mass functions on the categorical typology $\mathcal{T}$ yields the typology $\{G_P, G_{P'}, \ldots\}$ of corresponding SAMPLING-BASED grammars.

$$G_P(\mathsf{y} \,|\, \mathsf{x}) = P\big(\{G \in \mathcal{T} \,|\, G(\mathsf{x}) = \mathsf{y}\}\big) \tag{2}$$

According to this approach, the grammars internalized by a speaker are the traditional categorical grammars collected into the typology $\mathcal{T}$. The probabilistic behavior G_P arises only as an epiphenomenon of the fact that the speaker has not converged on a specific grammar from $\mathcal{T}$ but entertains a whole distribution P used to sample at random a grammar whenever phonological computation needs to be applied.

Implementation. Intuitively, a categorical grammar G ought to choose the best candidate surface realization $\mathsf{y} = G(\mathsf{x})$ of each underlying form x. And the best surface realization y ought to be the one with the largest harmony. For ease of comparison, we use again the ME harmony. Thus, we say that a strict and total categorical grammar G is a HARMONIC

GRAMMAR (HG; Legendre, Miyata, and Smolensky 1990b,a; Smolensky and Legendre 2006) provided G realizes any underlying form x as that surface candidate $\mathsf{y} = G(\mathsf{x})$ that maximizes the ME harmony corresponding to some non-negative weights $v_k \geq 0$. Equivalently, that violates the constraints the least because it has the smallest weighted average of constraint violations, as stated in (3). From now on, the categorical typology $\mathcal{T}$ in (2) consists of all strict and total HG grammars.

$$\mathsf{y} = \operatorname*{argmax}_{\mathsf{u} \in Gen(\mathsf{x})} H(\mathsf{x}, \mathsf{u}) = \operatorname*{argmin}_{\mathsf{u} \in Gen(\mathsf{x})} \sum_{k=1}^{n} v_k C_k(\mathsf{x}, \mathsf{u}) \tag{3}$$

We denote by $\mathcal{W}(G)$ the set of the non-negative weight vectors $\mathbf{v} = (v_1, \dots, v_n)$ that CORRESPOND to a strict and total grammar G in the sense that condition (3) holds for any mapping (x, y) in G. It is then natural to define the probability mass $P(G)$ of the HG grammar G as the volume (4) of this set $\mathcal{W}(G)$ of corresponding weights relative to some probability density function $\mathbf{p}$ on the set $\mathbb{R}^n_+$ of non-negative vectors.

$$P(G) = \int_{\mathcal{W}(G)} \mathbf{p}(\mathbf{v})\, \mathrm{d}\mathbf{v} \tag{4}$$

We assume that the density $\mathbf{p}$ STARTS at some weight vector $\mathbf{w}$ because it is equal to zero unless its argument is larger than $\mathbf{w}$, namely $\mathbf{p}(\mathbf{v}) = 0$ unless $v_1 \geq w_1, \dots, v_n \geq w_k$. When the starting vector $\mathbf{w}$ is non-negative, (4) yields a probability mass function P on the categorical typology $\mathcal{T}$, under mild constraint assumptions discussed in appendix A.3. The probabilities $G_P(\mathsf{y} \,|\, \mathsf{x})$ in (2) are therefore normalized over candidate sets. Examples of densities that start at a weight vector $\mathbf{w} = (w_1, \dots, w_n)$ are the uniform density $\mathbf{p}^{\text{unif}}_{\mathbf{w}}(\mathbf{v}) = \prod_{k=1}^{n} \mathbb{I}_{[w_k, w_k+1]}(v_k)$, the exponential density $\mathbf{p}^{\text{exp}}_{\mathbf{w}} = \prod_{k=1}^{n} \exp(w_k - v_k)\mathbb{I}_{[w_k, +\infty)}(v_k)$, and the half-Gaussian density $\mathbf{p}^{\text{gauss}}_{\mathbf{w}} = \prod_{k=1}^{n} \frac{2}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(v_k - w_k)^2\right\} \mathbb{I}_{[w_k, +\infty)}(v_k)$ (where $\mathbb{I}_S$ is the INDICATOR FUNCTION of a set S, that takes the value 1 on the elements of S and the value zero elsewhere).

The SHG grammar corresponding to a weight vector $\mathbf{w}$ is the grammar (2) obtained by sampling from the categorical typology $\mathcal{T}$ of (strict and total) categorical HG grammars according to the mass function P induced through (4) by a (uniform, exponential, or half-Gaussian) density that starts at $\mathbf{w}$. Our second model of probabilistic phonology is the typology of the SHG grammars corresponding to all non-negative weight vectors $\mathbf{w}$ (Boersma and Pater 2016, building on Boersma 1998; Boersma and Hayes 2001).

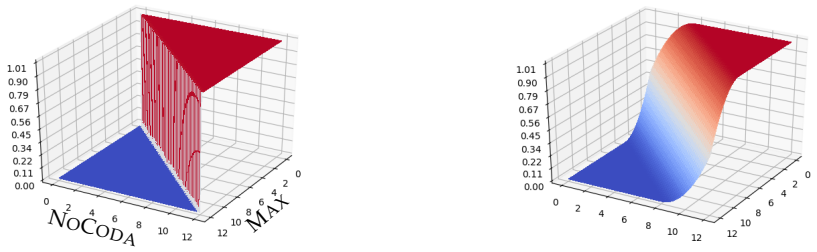
3. Comparing the two models

We now want to compare the ME and SHG approaches to probabilistic phonology in terms of their guarantees for grammatical inference and typological predictions, the two halves of any generative linguistic theory (Chomsky 1965).

3.1 Inference

Inference in probabilistic phonology is usually formulated as maximum likelihood estimation. ME has been endorsed in the recent phonological literature (Hayes and Wilson 2008) because the ME likelihood admits a closed-form expression that reveals its log-concavity as a function of the constraint weights. In other words, the ME likelihood is a smooth surface without local maxima. Its global maximum can be reached

Figure 1: Translation invariant densities



(a) (b)

Figure 2: The SHG probability of a mapping (b) as convolutional smoothing of the indicator function (a) of the set of weights HG consistent with that mapping

through standard gradient methods (Huang et al. 2010; Malouf 2013). No literature has investigated maximum likelihood estimation within the sampling-based model. Yet, we now show that SHG enjoys inference guarantees analogous to those of ME.

Our starting point is the observation that the uniform, exponential, and half-Gaussian densities are TRANSLATION INVARIANT: the value $\mathbf{p_w}(\mathbf{v})$ assigned by the density $\mathbf{p_w}$ that starts at $\mathbf{w}$ to a vector $\mathbf{v}$ is equal to the value $\mathbf{p_0}(\mathbf{w}-\mathbf{v})$ assigned by the density $\mathbf{p_0}$ that starts at the origin $\mathbf{0}$ to the vector $\mathbf{v}-\mathbf{w}$, as illustrated in figure 1 for the half-Gaussian. As shown in appendix A.4, it follows that the SHG probability mass of a mapping (x, y) (as a function of the weight vector $\mathbf{w}$) is the CONVOLUTION (more precisely, the CORRELATION) PRODUCT $\mathbf{p_0} * \mathbb{I}_{\mathcal{W}(x,y)}$ between the density $\mathbf{p_0}$ and the indicator function $\mathbb{I}_{\mathcal{W}(x,y)}$ of the set $\mathcal{W}(x, y)$ of those weight vectors $\mathbf{v} = (v_1, \ldots, v_n)$ that CORRESPOND to the mapping (x, y) in the sense that they satisfy condition (3).

To illustrate, we consider the underlying form /pat/ with the two candidates [pat] and [pa]. The constraint set consists of only NoCoda and Max, that penalize syllable codas and segment deletion. A categorical HG grammar realizes /pat/ faithfully as [pat] provided NoCoda has a smaller weight than Max, yielding the indicator function $\mathbb{I}_{\mathcal{W}(/\text{pat}/,[\text{pat}])}$ in figure 2a. The SHG probability of the mapping (/pat/, [pat]) in figure 2b is obtained through convolutional smoothing of this indicator function.

The uniform, exponential, or half-Gaussian density $\mathbf{p_0}$ is log-concave. The indicator function $\mathbb{I}_{\mathcal{W}(x,y)}$ is log-concave as well, because the set $\mathcal{W}(x, y)$ is convex. Since the convolution of log-concave functions is log-concave by Prekopa's theorem (Prékopa 1971, 1973), appendix A.4 concludes from the characterization of SHG probabilities as convolutions that also the SHG likelihood is log-concave:

Theorem 1
The likelihood is a log-concave function of the constraint weights in SHG just as in ME.

Appendix A.5 then shows that computing the derivatives of the SHG likelihood function boils down to computing volumes of polyhedra (relative to a density). Unfor-

tunately, exact volume computation cannot be performed efficiently (Dyer and Frieze 1988), plausibly not even for the special polyhedra that arise here. Yet, a number of efficient randomized techniques are available for approximate volume computation (Dyer, Frieze, and Kannan 1991; Bollobás 1997) that make robust gradient ascent techniques viable in SHG just as in ME. We conclude that inference does not provide an argument for choosing between these two models of probabilistic phonology. We thus turn from inference to the models' typological predictions.

3.2 Implicational universals

Both ME and SHG typologies consist of infinitely many grammars and therefore cannot be directly inspected. An indirect strategy is needed to compare them. A natural such strategy is to compare ME and SHG typologies through the universals they predict. We focus here on IMPLICATIONAL UNIVERSALS (Greenberg 1963), namely implications $P \to \widehat{P}$ that hold of a given typology whenever *every* grammar in the typology that satisfies the antecedent property P also satisfies the consequent property $\widehat{P}$. Which antecedent and consequent properties P and $\widehat{P}$ should we focus on?

To answer this question, we step back to categorical phonology. The simplest property that can be predicated of a categorical grammar is the property of realizing a certain specific underlying form as a certain specific surface form. We thus focus on implications $(\mathsf{x}, \mathsf{y}) \to (\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ between two specific mappings (Anttila and Andrus 2006). This implication is a universal of a categorical typology provided every grammar 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 $\widehat{\mathsf{y}}$, as in (5).

$$(\mathsf{x}, \mathsf{y}) \to (\widehat{\mathsf{x}}, \widehat{\mathsf{y}}) \text{ means that, if } G(\mathsf{x}) = \mathsf{y}, \text{ then } G(\widehat{\mathsf{x}}) = \widehat{\mathsf{y}} \text{ for every grammar } G \tag{5}$$

For example, dialects of English that delete /t/ at the end of a coda cluster before vowels, also delete it before consonants. The implication $(/\mathsf{cost\#us}/, [\mathsf{cos\ us}]) \to (/\mathsf{cost\#me}/, [\mathsf{cos\ me}])$ is thus a universal in the sense of (5) of the typology of English dialects with categorical t-deletion (Guy 1991; Kiparsky 1993). Implicational universals can also be statistical. For instance, dialects of English where /t/-deletion applies variably always delete more frequently before consonants than before vowels (Coetzee 2004).

To capture such statistical generalizations, we say that the implication $(\mathsf{x}, \mathsf{y}) \to (\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ is a universal of a probabilistic typology provided the probability of the consequent mapping $(\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ is at least as large as the probability of the antecedent mapping (x, y) and this probability inequality holds UNIFORMLY for any grammar in the typology, as in (6).

$$(\mathsf{x}, \mathsf{y}) \to (\widehat{\mathsf{x}}, \widehat{\mathsf{y}}) \text{ means that } G(\mathsf{y} \,|\, \mathsf{x}) \leq G(\widehat{\mathsf{y}} \,|\, \widehat{\mathsf{x}}) \text{ for every grammar } G \tag{6}$$

Condition (5) is a special case of condition (6), when categorical grammars are construed as probabilistic grammars that only assign probabilities equal to zero and one.

3.3 Remarks

The universal implications in (5) are a special case of Evans and Levinson's (Evans and Levinson 2009) "type 3" or "absolute conditional" universals (see their table 1). They contrast this type of universals with "type 4" or "statistical conditional" universals, that they define (after Dryer 1998) through the scheme "if a language has property

X, it will tend to have property Y". Crucially, their type 4 universals have nothing to do with the universal implications we have defined in (6). Indeed, although our implications are statistical universals (because they are about probabilistic grammars), they are exceptionless universals: the probability of the consequent mapping is never smaller than that of the antecedent, with no exceptions.

Definitions (5) and (6) capture the intuition that the consequent of an implicational universal is a "better" mapping than the antecedent. There are various phonological reasons why a mapping counts as better than another. Consequently, different types of universal implications admit different phonological interpretations. To illustrate, let us focus on fully FAITHFUL phonological mappings with identical underlying and surface forms. What is the proper interpretation of an implication $(y, y) \rightarrow (\widehat{y}, \widehat{y})$ between two faithful mappings? Obviously, considerations of faithfulness cannot distinguish between faithful antecedent and consequent mappings. If faithfulness and markedness are the only two perspectives relevant for phonology, the only sense in which the faithful consequent mapping $(\widehat{y}, \widehat{y})$ is better than the faithful antecedent mapping (y, y) is that the consequent form $\widehat{y}$ is less marked than the antecedent form y. Universal implications $(y, y) \rightarrow (\widehat{y}, \widehat{y})$ between faithful mappings are thus called MARKEDNESS implications because they summarize the markedness hierarchies encoded into the typology.

3.4 The argument in a nutshell

We now compare ME and SHG typologies by comparing their universal implications. Appendix A.6 shows that the universal implications of SHG defined in terms of uniform probability inequalities as in (6) always coincide with the universal implications of categorical HG defined as in (5), which can in turn be easily computed (Magri 2018). Furthermore, appendix A.8 shows that, if an implication is a universal of ME, then it is also a universal of categorical HG. We thus conclude that

Theorem 2
No matter what the set of mappings Gen and the constraint set $\mathbf{C}$ *look like, the set of ME universal implications is always a subset of the set of SHG universal implications.*

How much smaller is the ME subset of universal implications relative to the SHG superset? In this article, we mostly restrict our comparison to markedness implications, as this restriction allows us to make our point with a reduced apparatus (see Magri in progress for the extension to arbitrary universal implications). We develop our argument in sections 4-8 through two steps. The FIRST STEP consists of a number of theorems on ME (markedness) implications, that represent the first attempt (to the best of our knowledge) to understand the general principles of ME phonology, beyond reports of specific successes on specific patterns of data given specific choices of candidates and constraints (Zuraw and Hayes 2017; Smith and Pater 2020; Breiss and Albright 2022).

The SECOND STEP of our argument deduces from these theorems a number of phonological paradoxes. As a result of these paradoxes, ME is shown to validate no universal (markedness) implications on a battery of test cases drawn from the literature. In other words, ME predicts no uniform probability inequalities and therefore no (markedness) asymmetries: any form can have a larger ME probability than any other form. In all these cases, HG/SHG instead capture a rich system of universal (markedness) asymmetries. We conclude that SHG provides a better model of probabilistic natural language phonology than ME. Section 9 then extends our criticism of ME into a more fundamental argument against intrinsically probabilistic phonological models.

3.5 What if there are no universals?

Before starting on our argument, we need to address a fundamental objection. Evans and Levinson (2009) maintain that "languages differ so fundamentally from one another [...] that it is very hard to find any single structural property they share." This skepticism towards universals has leaked into generative phonology. Indeed, Hyman (Hyman 2008) warns us that "the study of universals is fraught with difficulties". Maddieson (Maddieson 1984) states that "there appear to be so few absolute universals". And van Oostendorp (van Oostendorp 2013) flatly concludes that "the quest for universals has failed" in phonology. Obviously, our argument that ME fails to predict universals of markedness would be void if it turned out that there are no phonological universals!

Let us take a closer look at this objection. Section 8 will show that the syllable CV can have a smaller probability than the syllable VC under ME. This looks paradoxical because only VC has two marked properties: a coda and no onset (Blevins 1995). Yet, some languages have been described as having VC but not CV (Breen and Pensalfini 1999). This would suggest that some constraints prefer VC to CV. The fact that CV can have a smaller ME probability than VC might thus be no paradox after all. More generally, a form that counts as more marked than another relative to some markedness perspective, often counts as less marked relative to a different markedness perspective. In conclusion—so goes the objection—ME's failure to predict markedness asymmetries, far from being paradoxical, might be empirically supported after all.

This objection misses our point. For the sake of the argument, let us grant that there exist some exotic markedness constraints that prefer VC over CV, next to more familiar constraints such as ONSET and NOCODA that penalize absence of onsets and presence of codas, and therefore prefer CV over VC. Now, suppose that we restrict ourselves from the entire typology to the SUB-TYPOLOGY of grammars that attach no importance whatsoever to the exotic markedness constraints that prefer VC over CV (say, by assigning them zero weight). It stands to reason that this sub-typology should satisfy the universal that VC is more marked (has uniformly smaller probability) than CV. Our point is that ME fails to predict this universal even for this sub-typology: it allows VC to have a larger probability than CV even when the constraint set contains the familiar markedness constraints that penalize VC but no exotic markedness constraints that penalize CV.

In conclusion, comparing probabilistic phonological models in terms of their ability to predict universals is warranted, no matter one's position in the debate on language universals. In fact, that debate only concerns the question of whether there exist universals that hold across the entire typology of natural languages. In contrast, the existence of universals for sub-typologies is a logical consequence of the fact that sub-typologies are subsets of the entire typology of natural languages. Our argument against ME based on universals developed in the rest of the paper thus has a logical nature and it stands independently of the empirical debate on universals.

4. Harmonic bounding generalization

We start our argument with the following generalization about ME universal implications (not just markedness implications), established in appendix A.10. It says that the consequent mapping never violates a constraint more than the antecedent mapping. With standard terminology (Prince and Smolensky 1993/2004), we say that the consequent mapping HARMONICALLY BOUNDS the antecedent mapping.

Theorem 3
If an implication $(\mathsf{x},\mathsf{y}) \rightarrow (\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ *is a ME universal, the inequality* $C(\mathsf{x},\mathsf{y}) \geq C(\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ *holds for every constraint* C *in the constraint set used to define the ME typology such that* C *is not violated by at least one candidate of the consequent underlying form* $\widehat{\mathsf{x}}$ *.*

At first sight, this ME harmonic bounding generalization seems to formalize the reasonable intuition that the consequent mapping of an implicational universal ought to be better (to violate the constraints less) than the antecedent mapping. To argue that this generalization is nonetheless paradoxical, we look at implications with impossible antecedent mappings.

Paradox of impossible antecedents

The Basic Syllable System (Prince and Smolensky 1993/2004) focuses on the four syllables CV, CVC, V, and VC, each a candidate of the other. They are evaluated by the two markedness constraints ONSET and NOCODA already mentioned above, that penalize absence of onsets and presence of codas, together with the two faithfulness constraints DEP and MAX, that penalize segment epenthesis and deletion. The implication $(/\mathsf{CV}/,[\mathsf{CVC}]) \rightarrow (/\mathsf{CVC}/,[\mathsf{CV}])$ is a universal of categorical HG in the sense of condition (5) simply because the antecedent mapping $(/\mathsf{CV}/,[\mathsf{CVC}])$ is impossible in HG given these constraints: no HG grammar epenthesizes codas because no alternations increase markedness (Moreton 1999). This implication is therefore also a universal of SHG in the sense of condition (6), because HG and SHG share the same implicational universals. Yet, this implication is not a universal of ME: the constraint $C = \text{MAX}$ flouts the harmonic bounding generalization because only the consequent mapping violates it. Harmonic bounding thus condemns ME to the paradoxical prediction that coda epenthesis in the antecedent $(/\mathsf{CV}/,[\mathsf{CVC}])$ can have a larger ME probability than coda deletion in the consequent $(/\mathsf{CVC}/,[\mathsf{CV}])$. The paradox can be replicated straightforwardly with just any markedness or faithfulness constraint C.

5. Cardinality generalization

When ME weights are all equal (or, by continuity, close) to zero, all differences among the candidates of an underlying form x are wiped away. All candidates thus receive the same share of probability, which is equal to the inverse $1/|Gen(\mathsf{x})|$ of the cardinality of the candidate set $Gen(\mathsf{x})$. Since a uniform probability inequality must hold in particular for weights equal (or close) to zero, an ME implication $(\mathsf{x},\mathsf{y}) \rightarrow (\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ requires the antecedent underlying form x to have at least as many candidates as the consequent underlying form $\widehat{\mathsf{x}}$, yielding the cardinality inequality $|Gen(\mathsf{x})| \geq |Gen(\widehat{\mathsf{x}})|$.

This sensitivity of ME to the sheer number of candidates looks *prima facie* innocuous, because this cardinality inequality is satisfied when antecedent and consequent candidate sets $Gen(\mathsf{x}) = Gen(\widehat{\mathsf{x}})$ coincide (Blaho, Bye, and Krämer 2007). Yet, let us choose a subset S of the constraint set and let us denote by $Gen_S(\mathsf{x})$ the result of pruning the original candidate set $Gen(\mathsf{x})$ of those candidates that violate some constraint in this set S. Appendix A.16 refines the original cardinality inequality $|Gen(\mathsf{x})| \geq |Gen(\widehat{\mathsf{x}})|$ as follows (the original inequality is a special case of the refined inequality with S empty).

Theorem 4
Suppose that an implication $(\mathsf{x},\mathsf{y}) \rightarrow (\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ *is a ME universal. The (refined) cardinality inequality* $\left|Gen_S(\mathsf{x})\right| \geq \left|Gen_S(\widehat{\mathsf{x}})\right|$ *holds for any subset* S *of constraints such that the antecedent surface form* y *belongs to* $Gen_S(\mathsf{x})$*.*

We now use this refined inequality to argue that ME's sensitivity to the number of candidates is indeed phonologically paradoxical (see section 8.1 for further discussion).

Paradox of few-versus-many triggers

In languages with backness harmony, suffix vowels must agree in backness with root vowels, as in the Finnish alternations (/kuuma-nA/, [kuu.ma.na]) ('hot-ESS') and (/kylmä-nA/, [kyl.mä.nä]) ('cold-ESS'). To explore the universals of backness harmony, we consider two roots /B/ with one back vowel and /BB/ with two back vowels, combined with a suffix /**F**/ with an underlying front vowel. Intuitively, one would expect /BB/ to be a stronger harmony trigger than /B/: the pressure to harmonize the suffix should not decrease when the number of root back vowel triggers increases. In other words, the probability of (/BB+**F**/, [BB**B**]) should never be smaller than the probability of (/B+**F**/, [B**B**]), whereby the implication (/B+**F**/, [B**B**]) → (/BB+**F**/, [BB**B**]) ought to be a universal of probabilistic phonology. This implication indeed holds in HG/SHG for many analyses of harmony such as the one in Hayes and Londe (2006).

To study this implication in ME, we need to spell out the antecedent and consequent candidate sets. A reasonable initial assumption is that candidates are obtained by changing underlying vowel backness in all possible ways. Yet, this assumption entails that the antecedent underlying form /B+**F**/ has only four candidates while the consequent underlying form /BB+**F**/ has eight candidates. The implication (/B+**F**/, [B**B**]) → (/BB+**F**/, [BB**B**]) then fails in ME because it flouts the original cardinality inequality $|Gen(\mathsf{x})| \geq |Gen(\widehat{\mathsf{x}})|$.

We thus revise our initial assumption to ensure that the two underlying forms /B+**F**/ and /BB+**F**/ share the same candidate set. This shared candidate set must consist of the eight strings [XXX] with X equal to B or F. The original cardinality inequality $|Gen(\mathsf{x})| \geq |Gen(\widehat{\mathsf{x}})|$ is now secured as an identity. Yet, some surface vowels are epenthetic relative to the antecedent underlying form /B+**F**/. This fact is recorded by the faithfulness constraint DEP that penalizes epenthesis. Let S be the singleton constraint subset consisting of DEP only. The corresponding pruned candidate sets Gen_S(/B+**F**/) and Gen_S(/BB+**F**/) are our initial sets consisting of the four and eight strings obtained by changing underlying backness. And the refined cardinality inequality $\left|Gen_S(\mathsf{x})\right| \geq \left|Gen_S(\widehat{\mathsf{x}})\right|$ thus fails. In the end, no amount of tweaking with candidate sets rescues this implication (/B+**F**/, [B**B**]) → (/BB+**F**/, [BB**B**]). ME thus paradoxically predicts that the probability of a phonological process can decrease when the number of triggers increases.

6. One-step-away generalization

We now restrict ourselves from arbitrary universal implications to markedness implications. As explained in section 3.3.3, these implications compare fully faithful mappings and thus capture markedness asymmetries. Any non-faithful surface candidate must be penalized by some faithfulness constraint. We say that a non-faithful candidate is only ONE STEP AWAY (in the direction of some faithfulness constraint F_0) provided it violates the faithfulness constraints as little as possible because it violates F_0 only once and it violates no other faithfulness constraints (see below for examples). Appendix A.11 shows that one-step-away candidates of ME universal markedness implications must comply with the following generalization.

Theorem 5
Suppose that a markedness implication $(\mathsf{y}, \mathsf{y}) \to (\widehat{\mathsf{y}}, \widehat{\mathsf{y}})$ *is a ME universal. If the consequent*

underlying form $\widehat{\text{y}}$ *has one-step-away non-faithful surface candidates in the direction of some faithfulness constraint* F_0*, the antecedent underlying form* y *does as well.*

6.1 Paradox of unmarked forms

Within the Basic Syllable System described in section 4, the markedness implications $(/\text{CVC}/, [\text{CVC}]) \rightarrow (/\text{CV}/, [\text{CV}])$ and $(/\text{V}/, [\text{V}]) \rightarrow (/\text{CV}/, [\text{CV}])$ are universals of HG/SHG. These implications make good sense, given that the syllable CV violates none of the markedness constraints in the system. Yet, both markedness implications are lost in ME. Indeed, consider the faithfulness constraint $F_0 = \text{DEP}$ that penalizes segment epenthesis. The consequent form /CV/ has a one-step-away non-faithful surface candidate [CVC] in the direction of F_0: the mapping $(/\text{CV}/, [\text{CVC}])$ violates DEP only once and does not violate any other faithfulness constraints. The antecedent form /CVC/ instead has no surface candidates one step away in the direction of $F_0 = \text{DEP}$: since CVC is the longest form in the Basic Syllable System, it admits no epenthesis. The markedness implication $(/\text{CVC}/, [\text{CVC}]) \rightarrow (/\text{CV}/, [\text{CV}])$ thus fails in ME because it flouts the one-step-away generalization. Analogous considerations hold for $(/\text{V}/, [\text{V}]) \rightarrow (/\text{CV}/, [\text{CV}])$ and $F_0 = \text{MAX}$. In conclusion, syllables with codas and missing onsets do not count as marked in ME: they can paradoxically have larger probabilities than the unmarked syllable CV.

6.2 A closer look at the paradox

One can fairly object that assuming CVC to be the longest syllable is an idiosyncratic property of the Basic Syllable System. With this in mind, let us enrich our inventory of syllables with complex onsets CCV— and complex codas —VCC. The antecedent form /CVC/ now has a surface candidate such as [CVCC] that is indeed one step away in the direction of $F_0 = \text{DEP}$ and the markedness implication $(/\text{CVC}/, [\text{CVC}]) \rightarrow (/\text{CV}/, [\text{CV}])$ thus complies with the one-step-away generalization (although it still fails in ME for independent reasons discussed in section 8.1).

Yet, since the syllable CCVCC (with a complex onset and a complex coda) is now longest, it admits no one-step-away candidates in the direction of $F_0 = \text{DEP}$. This time, it is the implication $(/\text{CCVCC}/, [\text{CCVCC}]) \rightarrow (/\text{CV}/, [\text{CV}])$ that fails in ME because of the one-step-away generalization. Enlarging the inventory of syllables does not rescue the generalization that CV is least marked: the paradox has been shifted, not solved.

The logic of the paradox is now clear. ME probabilities are defined over finite candidate sets (Hayes and Wilson 2008, footnote 5; Daland 2015). This is logically unpleasant because it requires an arbitrary cut-off to the unlimited combinatorics of phonological representations. Yet one might still hope that this unpleasant assumption of finite candidate sets is phonologically inconsequential in practice. That is not the case. Finite candidate sets contain forms that count as longest and therefore admit no epenthesis. As soon as the constraint set contains the faithfulness constraint $F_0 = \text{DEP}$, the one-step-away generalization predicts that unmarked forms such as CV can have a smaller ME probability than these longest forms. This prediction is paradoxical because longest forms have multiple opportunities to display marked structures, such as complex consonant clusters.

7. Reverse harmonic bounding generalization

The one-step-away generalization is coarse: it is oblivious to the phonological quality of the one-step-away candidates, but only regulates their existence. Appendix A.12 derives the following refinement: if the consequent has *good* one-step-away candidates (good because they do not violate some markedness constraint), the antecedent does as well.

Theorem 6
Suppose that a markedness implication $(\mathrm{y}, \mathrm{y}) \to (\widehat{\mathrm{y}}, \widehat{\mathrm{y}})$ *is a ME universal. Consider a markedness constraint* M *that assigns the same number of violations to the antecedent and consequent forms, namely* $M(\mathrm{y}) = M(\widehat{\mathrm{y}})$*. If the consequent underlying form* $\widehat{\mathrm{y}}$ *has some candidate that is only one step away (in the direction of some faithfulness constraint* F_0*) and does not violate* M*, then the antecedent underlying form* y *as well has some candidate that is only one step away (in the same direction* F_0*) and does not violate* M*.*

Let us now unpack this rather unintuitive generalization. Consider a form that does not violate a markedness constraint M. Yet, it is close to violating it. In the sense that all the candidates that resemble it because they are only one step away (in the direction of some faithfulness constraint F_0) do violate M. In this case, we say that the form considered is ONE STEP AWAY (in the direction of F_0) FROM VIOLATING the markedness constraint M. The theorem then says that, if the antecedent form y is one step away from violating a markedness constraint M, the consequent form $\widehat{\mathrm{y}}$ is as well.

For comparison, the harmonic bounding generalization in section 4 says instead that, if the consequent form $\widehat{\mathrm{y}}$ violates a markedness constraint M, the antecedent form y does as well. The roles of the antecedent and consequent forms are reversed in the two generalizations. In other words, a form counts as less marked in ME only if it has fewer actual markedness violations but more one-step-away markedness violations! We now deduce a score of ME paradoxes from this reverse harmonic bounding generalization.

7.1 Laryngeal paradoxes

Let us consider stops that differ in voicing, aspiration, and place and are candidates of each other. We posit the faithfulness constraints $\textsc{Ident}_{[\text{voice}]}$, $\textsc{Ident}_{[\text{spread}]}$, and $\textsc{Ident}_{[\text{place}]}$, that penalize discrepancies in voicing, spread glottis, and place. As for the markedness constraints, we start with only $*[+\text{voice}]$ and $*[+\text{spread glottis}]$, that penalize voiced stops and aspirated stops (Lombardi 1999). For concreteness, we focus on coronal stops. In HG/SHG, their four identity mappings satisfy the five markedness implications at the bottom of figure 3a. These markedness implications all survive in ME. Thus, ME captures the generalization that voicing and aspiration are marked: dʰa can never have larger ME probability than da or tʰa which in turn can never have larger ME probability than ta. Yet, we now show that ME's success is short-lived.

Classical Greek and Vietnamese (Thompson 1965) allow voiced stops and aspirated stops but not stops that are both voiced and aspirated. These languages motivate the markedness constraint $M = *[+\text{voice}, +\text{spread}]$ that penalizes voiced aspirated stops (singled out by the cylinder at the top of figure 3b) to the exclusion of stops that are only voiced or only aspirated.[1] This additional markedness constraint M does not affect

1 This constraint $M = *[+\text{voice}, +\text{spread}]$ is well motivated even within an architecture such as HG that allows for some additive effects, because the ban against voiced aspirated stops to the exclusion of simply voiced and simply aspirated stops does not follow as an additive interaction of the simple markedness

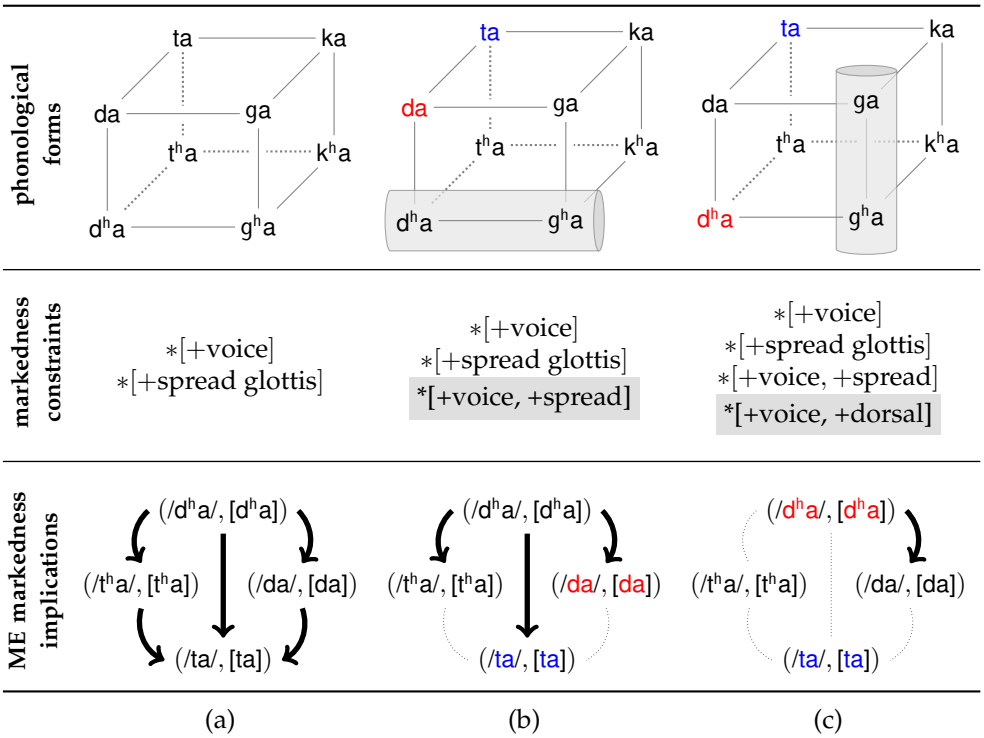


Figure 3: Pruning ME markedness implications for laryngeal phonology by adding non-conflicting markedness constraints

the HG/SHG universal markedness implications because M does not conflict with the markedness of voicing and aspiration. If anything, it reinforces it.

The situation is different in ME. To illustrate, we focus on the implication $(/\text{da}/, [\text{da}]) \rightarrow (/\text{ta}/, [\text{ta}])$. Neither the antecedent form da nor the consequent form ta violates the markedness constraint M (neither form belongs to the cylinder). If we move down, just one step away from the antecedent form da (in the direction of $F_0 = \text{IDENT}_{[\text{spread}]}$), we get to the candidate dʰa which does violate M (it belongs to the cylinder). But if we move down, one step away from the consequent form ta, we get to the candidate tʰa that does *not* violate M (it does not belong to the cylinder).

In conclusion, the antecedent form da is only one step away from violating M, but the consequent form ta is not. The markedness implication $(/\text{da}/, [\text{da}]) \rightarrow (/\text{ta}/, [\text{ta}])$ fails in ME because it flouts the reverse harmonic bounding generalization. The marked antecedent mapping $(/\text{da}/, [\text{da}])$ can paradoxically have a larger ME probability than the consequent mapping $(/\text{ta}/, [\text{ta}])$. The markedness implication $(/\text{tʰa}/, [\text{tʰa}]) \rightarrow (/\text{ta}/, [\text{ta}])$ fails analogously. In conclusion, adding the markedness constraint $M = *[+\text{voice}, +\text{spread}]$ leads to the paradoxical result that voicing and aspiration are no longer marked in ME (while M sensibly has no effect on their markedness in HG/SHG).

constraints $*[+\text{voice}]$ and $*[+\text{spread glottis}]$. The HG typology without M does not contain a grammar like Vietnamese.

7.2 More laryngeal paradoxes

Not everything is lost in ME though: the three markedness implications at the bottom of figure 3b that share the antecedent faithful mapping (/dʰa/, [dʰa]) are universals of ME relative to the constraints considered so far. ME thus does capture the generalization that voicing and aspiration gang up to yield the worst of the worst. Yet, once again, ME's success is short-lived.

Thai has voicing contrast at labial and coronal place but lacks a voiced velar stop (Sherman 1975; Locke 1983), presumably because voicing is harder to sustain for stops at the velar place (Ohala 1983). We thus add the markedness constraint $M = *[+\text{voice}, +\text{dorsal}]$ that penalizes voiced velar stops (singled out by the cylinder at the top of figure 3c) to the exclusion of voiced labial and coronal stops. Again, this additional markedness constraint M does not affect the HG/SHG markedness implications because M does not conflict with the markedness of voicing and aspiration. If anything, it reinforces it.

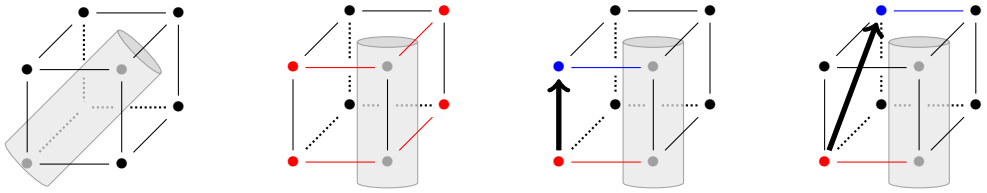
The situation is different in ME. To illustrate, we focus on the implication (/dʰa/, [dʰa]) → (/ta/, [ta]). Neither the antecedent form dʰa nor the consequent form ta violate the markedness constraint M (neither form belongs to the cylinder). If we move right, just one step away from the antecedent form dʰa (in the direction of $F_0 = \text{IDENT}_{[\text{place}]}$), we get to the candidate gʰa which does violate M (it belongs to the cylinder). But if we move right, one step away from the consequent form ta, we get to the candidate ka that does *not* violate M (it does not belong to the cylinder).

In conclusion, the antecedent form dʰa is only one step away from violating M, but the consequent form ta is not. The markedness implication (/dʰa/, [dʰa]) → (/ta/, [ta]) fails in ME because it flouts the reverse harmonic bounding generalization. ME fails to capture the generalization that the ganging-up of voicing and aspiration yields the worst of the worst: it can assign a larger probability to (/dʰa/, [dʰa]) than to (/ta/, [ta]).

The markedness implication (/dʰa/, [dʰa]) → (/tʰa/, [tʰa]) fails analogously. In the end, only the markedness implication (/dʰa/, [dʰa]) → (/da/, [da]) survives in ME. But this lonely survivor at the bottom of figure 3c makes little phonological sense, for two reasons. First, the fact that dʰa always has smaller ME probability than da but can have larger ME probability than ta makes little sense: it gets the markedness asymmetry between da and ta all wrong. Second, the fact that dʰa always has smaller ME probability than da but can have larger ME probability than tʰa makes little sense: it predicts an asymmetry between voicing (da) and aspiration (tʰa) that is phonologically spurious as it is in no way encoded into the constraint set. We conclude that the markedness universals of laryngeal phonology predicted by ME are paradoxical.

7.3 Parallelism

Let us make explicit the abstract logic of these paradoxes. We start from a set of relevant phonological features (voicing, aspiration, place, ...). We consider the phonological forms corresponding to all feature value combinations. We assume they are candidates of each other. We plot these forms as dots in a lattice whose dimensions correspond to the features, as in figure 4. A markedness constraint M can be represented as a cylinder that singles out the forms that violate it. When M is a feature co-occurrence constraint, this cylinder is never diagonal to the directions of the lattice (as in figure 4a) but always aligned with one of them (as in figure 4b-d).



(a) (b) (c) (d)

Figure 4: Interpreting the reverse harmonic bounding generalization as a parallelism condition between ME markedness implications and feature co-occurrence constraints

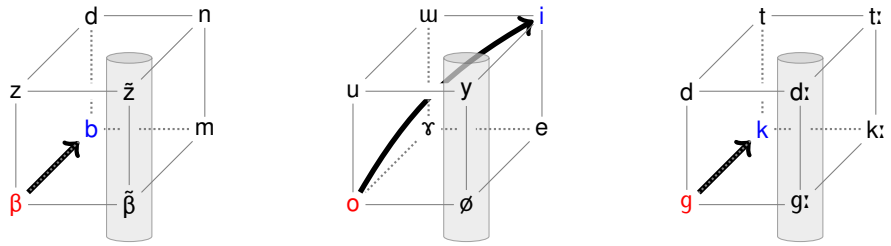
Let us now suppose that the antecedent form y of a ME markedness implication $(y, y) \rightarrow (\widehat{y}, \widehat{y})$ is one step away from violating this feature co-occurrence constraint M. This means that y does not belong to the cylinder but is closest to it, namely is one of the four red dots in figure 4b that are only one step away from the cylinder. The reverse harmonic bounding generalization requires the consequent form $\widehat{y}$ to be one step away from violating M as well, in the same direction. This means that $\widehat{y}$ does not belong to the cylinder but is connected to it through a single parallel step. Equivalently, $\widehat{y}$ belongs to the direction of the lattice that goes through y and runs parallel to the cylinder.

To illustrate, suppose that the antecedent form y is the red dot in figure 4c-d. The only consequent form $\widehat{y}$ that complies with the reverse harmonic bounding generalization is the blue dot in figure 4c, yielding the markedness implication $(y, y) \rightarrow (\widehat{y}, \widehat{y})$ plotted as the thick arrow parallel to the cylinder. The consequent form $\widehat{y}$ cannot be, say, the blue dot in figure 4d, that would yield a markedness implication $(y, y) \rightarrow (\widehat{y}, \widehat{y})$ not parallel to the cylinder. In conclusion, the reverse harmonic bounding generalization says that ME markedness implications must run parallel to any feature co-occurrence constraints that the antecedent form is one step away from violating.

But this parallelism condition is a phonological paradox: parallelism in the lattice of feature value combinations has nothing to do with the substance of phonological markedness. Indeed, the non-parallel implication $(y, y) \rightarrow (\widehat{y}, \widehat{y})$ in figure 4d paradoxically fails because of a markedness feature co-occurrence constraint that does not distinguish between the antecedent and consequent forms y and $\widehat{y}$ (neither of them violates it) and should therefore be irrelevant to their comparison. Indeed, we now show that this parallelism condition yields paradoxes in every corner of segmental phonology.

7.4 Paradoxes everywhere

Figure 5a organizes the combinations of values of the features [nasal], [continuant], and [place] into a lattice. The markedness implication (/β/, [β]) → (/b/, [b]) plotted as the thick arrow captures the generalization that (non strident) voiced fricatives are more marked than the corresponding stops (Jakobson 1941): they are typologically rarer (Maddieson 1984), more difficult to produce (Ohala 1983), and acquired later (Smith 1973). Yet, this sensible markedness implication cannot be a ME universal. Here is why. Nasal fricatives are particularly marked (they are rare, almost never contrastive, usually resulting from nasal spreading: Ladefoged and Maddieson 1996, §4.4; Shosted 2006), motivating the feature co-occurrence constraint $M = *[+\text{nasal}, +\text{continuant}, -\text{sonorant}]$, plotted as the cylinder in figure 5a. Intuitively, M is irrelevant to the comparison between β and b



(a) (b) (c)

Figure 5: Sensible markedness implications that fail in ME because of lack of parallelism

(neither is nasal). Yet, the markedness implication (/β/, [β]) → (/b/, [b]) fails in ME because it is not parallel to M (while β is one step away from violating it).

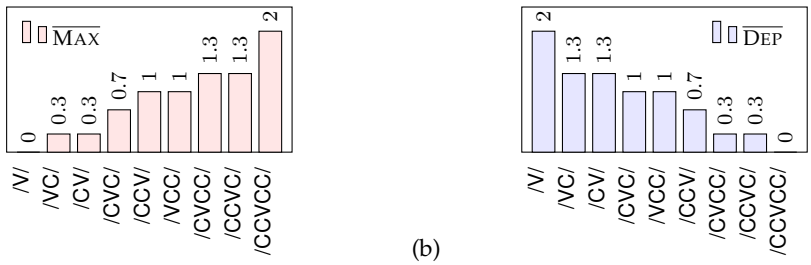
Figure 5b organizes the combinations of values of the vowel features [back], [high], and [round] into a lattice. The markedness implication (/o/, [o]) → (/i/, [i]) captures four generalizations. First, rounding is marked (epenthetic vowels are never rounded: Lombardi 2003; de Lacy 2006, §7.2.5). Second, rounding is particularly marked for non-high vowels (Kaun 2004). Third, back vowels are marked (they are rarely epenthetic: de Lacy 2006, §7.2.5). Finally, non-high vowels are marked (at least outside of prosodic heads: de Lacy 2006, p. 68). Yet, this sensible markedness implication cannot be a ME universal. Here is why. Rounding is particularly marked for front vowels, motivating the feature co-occurrence constraint $M = *[+\text{round}, -\text{back}]$ (known as *ROFRO; Kaun 2004), plotted as the cylinder in figure 5b. Intuitively, M is irrelevant to the comparison between o and i (neither violates it). Yet, the implication (/o/, [o]) → (/i/, [i]) fails in ME because it is not parallel to M (while its antecedent o is one step away from violating it).

Finally, figure 5c organizes the combinations of values of the features [voice], [place], and [length] into a lattice (featural encoding of phonological length is not crucial to the argument). The markedness implication (/g/, [g]) → (/k/, [k]) captures the markedness of voicing at the velar place, already mentioned above. Once again, this sensible markedness implication cannot be a ME universal. Here is why. Voicing is particularly hard to sustain for geminates (Ohala 1983), motivating the feature co-occurrence constraint $M = *[+\text{voice}, +\text{long}]$, plotted as the cylinder in figure 5c (for an alternative, see Kawahara 2006). Intuitively, M is irrelevant to the comparison between g and k (neither is geminated). Yet, the markedness implication (/g/, [g]) → (/k/, [k]) fails in ME because it is not parallel to M (while its antecedent g is one step away from violating it). By applying this logic systematically to a variety of feature co-occurrence constraints, analogous paradoxes can be uncovered in every corner of segmental phonology.

8. Average faithfulness generalization

Let $\overline{C}(x)$ denote the AVERAGE number of violations $\frac{1}{|Gen(x)|}\sum_{y\in Gen(x)} C(x,y)$ assigned by a constraint C to the candidates of an underlying form x. Appendix A.14 shows that ME markedness implications obey the following condition on faithfulness averages.

Theorem 7
Suppose that a markedness implication $(x,x) \to (\hat{x},\hat{x})$ *is a ME universal and that the antecedent and consequent underlying forms share the same candidate set. The average number* $\overline{F}(x)$ *of*



(a) (b)

Figure 6: Average number of deletions and epentheses for the nine underlying strings of the Extended Syllable System, ordered according to their length

antecedent faithfulness violations is at most as large as the average number $\overline{F}(\hat{x})$ *of consequent faithfulness violations, namely* $\overline{F}(x) \leq \overline{F}(\hat{x})$ *for every faithfulness constraint* F.

8.1 Paradox of sheer length

We focus on the faithfulness constraint MAX that penalizes segment deletions. When all underlying forms share the same candidate set, the average number $\overline{\text{MAX}}$ of deletions increases with the length of the underlying strings: a longer underlying string means more segments to delete. To illustrate, we consider the Extended Syllable System (Prince and Smolensky 1993/2004). It supplements the syllables CV, CVC, V, and VC of the Basic Syllable System described in section 4 with the complex edges of CCV, CCVC, CCVCC, VCC, and CVCC (all candidates of each other). Furthermore, it supplements the constraints ONSET, CODA, DEP, and MAX with COMPONSET and COMPCODA, that penalize complex onsets and codas. The average number $\overline{\text{MAX}}$ of deletions plotted in figure 6a grows with the length of the underlying strings.[2] The average faithfulness inequality $\overline{F}(x) \leq \overline{F}(\hat{x})$ for $F = \text{MAX}$ thus entails that the antecedent string x of a ME markedness implication $(x, x) \rightarrow (\hat{x}, \hat{x})$ *cannot be longer* than the consequent string $\hat{x}$.

Opposite considerations hold for the faithfulness constraint DEP that penalizes segment epentheses. The average number $\overline{\text{DEP}}$ of epentheses decreases with the length of the underlying strings, as shown in figure 6b. The average faithfulness inequality $\overline{F}(x) \leq \overline{F}(\hat{x})$ for $F = \text{DEP}$ thus entails that the antecedent string x *cannot be shorter* than the consequent string $\hat{x}$. In conclusion, if the constraint set contains both MAX and DEP, any markedness implication $(x, x) \rightarrow (\hat{x}, \hat{x})$ fails in ME when the antecedent and consequent strings x and $\hat{x}$ *differ in sheer length* (but share the same candidate set).

To illustrate, the nine identity mappings of the Extended Syllable System are ordered into sixteen sensible universal markedness implications by HG/SHG. Yet, fifteen of these implications compare forms (such as CCVCC and CV) with different sheer length. All fail in ME because ME markedness implications can only compare forms with the same sheer length. We are left with the lonely HG/SHG markedness implication (/VC/, [VC]) → (/CV/, [CV]) that compares forms VC and CV with the same sheer

2 We assume that onset consonants are never in correspondence with coda consonants. Thus, (/VC/, [CV]) violates both MAX and DEP, because the surface onset must be epenthetic and the underlying coda must be deleted. Apart from that, we assume correspondence relations that minimize deletions and epentheses. These choices are made for concreteness to draw figure 6, but they are not crucial to the argument developed in this subsection.

length. Yet, even this implication fails in ME because it flouts the cardinality inequality $\left|Gen_S(\mathsf{x})\right| \geq \left|Gen_S(\widehat{\mathsf{x}})\right|$ of section 5 when the constraint subset S singles out MAX and COMPCODA. In fact, the antecedent candidate set $Gen_S(/\mathsf{VC}/)$ pruned of the candidates that violate one of the two constraints in S consists of only *three* candidates [VC], [CVC] and [CCVC] (because [V], [CCV], and [CV] violate MAX while [VCC], [CVCC], and [CCVCC] violate COMPCODA). But the pruned consequent candidate set $Gen_S(/\mathsf{CV}/)$ consists of *four* candidates [CV], [CVC], [CCV], and [CCVC].

In conclusion, ME predicts no universals of syllable markedness. No syllable counts as more marked than any other syllable in ME because any syllable can have larger ME probability than any other syllable. This conclusion is independent of the specifics of the Extended Syllable System considered here: it holds robustly also when we take into account consonantal quality (e.g., coronals versus velars) and vowel complexity (e.g., long versus short vowels) or we allow for vowel epenthesis besides consonant deletion and epenthesis. We conclude that ME's insistence that markedness comparisons be restricted to strings of the same sheer length is a phonological paradox.

8.2 Paradox of sheer markedness counts

One of the two values (say, the value $[+]$ for concreteness) of a binary feature φ often counts as marked in the sense that a markedness constraint $M = *[+\varphi]$ specifically penalizes every occurrence of the marked feature value $[+\varphi]$. Because of the harmonic bounding inequality in section 4, the consequent form $\widehat{\mathsf{x}}$ of a ME markedness implication $(\mathsf{x}, \mathsf{x}) \to (\widehat{\mathsf{x}}, \widehat{\mathsf{x}})$ cannot violate this constraint M more than the antecedent form x. In other words, the consequent form $\widehat{\mathsf{x}}$ *cannot contain more* occurrences of the marked feature value $[+\varphi]$ than the antecedent form x.

Natural language phonology has been argued to take special care of this vulnerable marked feature value $[+\varphi]$ though a dedicated faithfulness constraint $\text{MAX}_{[+\varphi]}$ that penalizes the loss of an underlying occurrence of $[+\varphi]$ but not the loss of an underlying occurrence of $[-\varphi]$ (Kiparsky 1994; de Lacy 2006). For instance, interactions between voicing and nasality in Austronesian languages have been argued to motivate a faithfulness constraint $\text{MAX}_{[+\text{nasal}]}$ that specifically protects the marked value $[+]$ of the feature $\varphi = [\text{nasal}]$, penalizing de-nasalization but not nasalization (Pater 1999).

When all underlying forms share the same candidate set, the average number of violations $\overline{\text{MAX}}_{[+\varphi]}$ increases with the number of underlying occurrences of the value $[+\varphi]$. The average faithfulness inequality $\overline{F}(\mathsf{x}) \leq \overline{F}(\widehat{\mathsf{x}})$ for $F = \text{MAX}_{[+\varphi]}$ thus says that the consequent string $\widehat{\mathsf{x}}$ of a ME markedness implication $(\mathsf{x}, \mathsf{x}) \to (\widehat{\mathsf{x}}, \widehat{\mathsf{x}})$ *cannot contain fewer* occurrences of the feature value $[+\varphi]$ than the antecedent string x. In conclusion, if the constraint set contains both $M = *[+\varphi]$ and $F = \text{MAX}_{[+\varphi]}$, a markedness implication $(\mathsf{x}, \mathsf{x}) \to (\widehat{\mathsf{x}}, \widehat{\mathsf{x}})$ fails in ME when the antecedent and consequent strings x and $\widehat{\mathsf{x}}$ have different numbers of occurrences of the marked value $[+\varphi]$.

To illustrate, we consider the forms atta, atna, anta, and anna, each a candidate of the other. We supplement the constraints $*[+\text{nasal}]$ and $\text{MAX}_{[+\text{nasal}]}$ motivated above with two more: $*\text{N}\mathring{\text{C}}$, that penalizes a nasal stop followed by an oral voiceless stop (Pater 1999); and SYLLABLECONTACT, that penalizes a low sonority coda (such as a voiceless stop) followed by a higher sonority onset (such as a nasal stop). HG and SHG predict the sensible universal markedness implications in figure 7a. But they all fail in ME because the antecedent and consequent strings have different numbers of nasals.

As another example, we consider the forms an, ãn, ad, and ãd, each a candidate of the other. We supplement the constraints $*[+\text{nasal}]$ and $\text{MAX}_{[+\text{nasal}]}$ motivated above

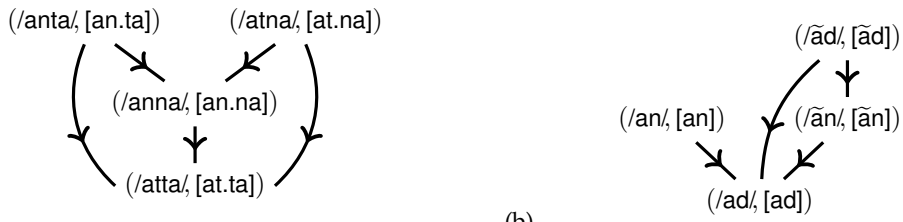


Figure 7: Markedness implications predicted by HG/SHG for two simple nasal systems

with two more: $*[-\text{nasal}, +\text{syllabic}][+\text{nasal}]$, that penalizes an oral vowel immediately followed by a nasal consonant (Kager 1999); and $\text{MAX}^{\text{consonant}}_{[+\text{nasal}]}$, that protects underlying nasal consonants at the exclusion of nasal vowels. HG and SHG predict the sensible universal markedness implications in figure 7b. Again, they all fail in ME because the antecedent and consequent forms have different numbers of nasals. We conclude that ME's insistence that markedness comparisons be restricted to strings with the same sheer number of occurrences of each marked feature value is a phonological paradox.

9. Beyond ME

Section 2.1 has introduced ME as a specific, popular example of intrinsically probabilistic models. According to these models, grammars directly assign probabilities to mappings as normalized harmony scores. Do the paradoxes documented in sections 4-8 extend beyond ME to the entire approach of intrinsically probabilistic phonology? Or can they be solved simply through a more judicious choice of the harmony?

9.1 No spurious probabilistic distinctions

To address this question, we tackle the problem of the choice of a harmony H from first principles. We start from the following question: when should the harmony-based grammar G_H in (1) assign the same probability to two different mappings (x, y) and $(\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$? To reason concretely, we focus on the two mappings $(\mathsf{x}, \mathsf{y}) = (/\mathsf{pat}/, [\mathsf{pa}])$ and $(\widehat{\mathsf{x}}, \widehat{\mathsf{y}}) = (/\mathsf{put}/, [\mathsf{pu}])$. To reason in the simplest case, we suppose that they come with only one additional candidate each, say $\mathsf{z} = [\mathsf{pat}]$ and $\widehat{\mathsf{z}} = [\mathsf{put}]$, respectively. A constraint C such as NOCODA or MAX (that penalize codas and deletions) makes no distinctions ACROSS CANDIDATE SETS: it does not distinguish either between $\mathsf{y} = [\mathsf{pa}]$ and $\widehat{\mathsf{y}} = [\mathsf{pu}]$ or between $\mathsf{z} = [\mathsf{pat}]$ and $\widehat{\mathsf{z}} = [\mathsf{put}]$, as stated in (7a).

$$\text{(a)}\quad \begin{aligned} C(\mathsf{x},\mathsf{y}) &= C(\widehat{\mathsf{x}},\widehat{\mathsf{y}}) \\ C(\mathsf{x},\mathsf{z}) &= C(\widehat{\mathsf{x}},\widehat{\mathsf{z}}) \end{aligned} \qquad \text{(b)}\quad \begin{array}{cc} C(\mathsf{x},\mathsf{y}) & C(\widehat{\mathsf{x}},\widehat{\mathsf{y}}) \\ \| & \| \\ C(\mathsf{x},\mathsf{z}) & C(\widehat{\mathsf{x}},\widehat{\mathsf{z}}) \end{array} \tag{7}$$

Suppose that the constraint set consists of only these two constraints NOCODA and MAX. Since harmonies are functions of constraint violations, the constraint identities (7a) translate into corresponding harmony identities: the candidates $\mathsf{y} = [\mathsf{pa}]$ and $\widehat{\mathsf{y}} = [\mathsf{pu}]$ have the same harmony scores as do the candidates $\mathsf{z} = [\mathsf{pat}]$ and $\widehat{\mathsf{z}} = [\mathsf{put}]$. The assumption (1) that probabilities are defined in terms of harmonies thus predicts that

[pa] and [pat] split the probability mass between the two of them exactly as [pu] and [put] split it. In conclusion, the mappings (/pat/, [pa]) and (/put/, [pu]) are predicted to share the same probability $G_H(\text{[pa]} \,|\, \text{/pat/}) = G_H(\text{[pu]} \,|\, \text{/put/})$. This result makes good sense in the simple scenario considered here.

Next, let us suppose that the constraint set contains the additional constraint $C = \text{NoRound}$ (that penalizes round vowels). This constraint makes no distinctions within candidate sets: it does not distinguish neither between $\mathsf{y} = \text{[pa]}$ and $\mathsf{z} = \text{[pat]}$ (neither violates it) nor between $\widehat{\mathsf{y}} = \text{[pu]}$ and $\widehat{\mathsf{z}} = \text{[put]}$ (both violate it), as stated in (7b). In other words, this additional constraint $C = \text{NoRound}$ is irrelevant to how the probability mass is split between [pa] and [pat] or between [pu] and [put]. Hence, the addition of this irrelevant constraint should not compromise the initial result that the mappings (/pat/, [pa]) and (/put/, [pu]) share the same probability $G_H(\text{[pa]} \,|\, \text{/pat/}) = G_H(\text{[pu]} \,|\, \text{/put/})$.

These considerations motivate the following axiom. Suppose for simplicity that two mappings (x, y) and $(\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ come with only one additional candidate z and $\widehat{\mathsf{z}}$ each. Suppose furthermore that every constraint C satisfies either the two "horizontal" identities (7a) (C makes no distinctions across candidate sets) or the two "vertical" identities (7b) (C makes no distinctions within candidate sets). Under these assumptions, the grammar G_H in (1) should not make a spurious distinction between these two mappings: it should assign them the same probability $G_H(\mathsf{y} \,|\, \mathsf{x}) = G_H(\widehat{\mathsf{y}} \,|\, \widehat{\mathsf{x}})$.

9.2 A general weighted model

Appendix A.17 provides the following characterization of harmony-based grammars that make no spurious distinctions:

Theorem 8
A harmony function H yields a harmony-based grammar G_H in (1) that makes no spurious distinctions if and only if there exist n factor functions $f_1, \dots, f_n$ such that the harmony of a mapping (x, y) is the product $H(\mathbf{C}(\mathsf{x}, \mathsf{y})) = \prod_{k=1}^{n} f_k(C_k(\mathsf{x}, \mathsf{y}))$ of the values assigned by the factor functions to the individual constraint violations.

Each factor function f_k corresponds to a constraint C_k. Differences among factor functions thus encode differences among the contributions of the constraints to the harmony. In the interest of a restrictive model of constraint interaction, we assume that the factor functions differ only minimally: they are all powers $f_k = f^{w_k}$ of the same base function f with different exponents w_k, as in (8).

$$H\big(\mathbf{C}(\mathsf{x}, \mathsf{y})\big) = \prod_{k=1}^{n} \Big(f\big(C_k(\mathsf{x}, \mathsf{y})\big) \Big)^{w_k} \tag{8}$$

In order for this harmony (8) to yield a normalization constant $Z(\mathsf{x})$ in (1) that is always different from zero, the base function f must be strictly positive. Furthermore, this harmony (8) is decreasing (as motivated in section 2.1) provided the weights w_k are non-negative and f is decreasing. Thus, f assumes its largest value at zero. Finally, we assume that f is "normalized": its largest value $f(0)$ is equal to 1. Normalization of f can be assumed without loss of generality, because rescaling f does not affect the corresponding harmony-based grammar G_H in (1). The weighted model of intrinsically probabilistic phonology is the typology consisting of the grammars G_H in (1) corresponding to the harmony (8) for a positive, decreasing, normalized base function f and all non-negative weights $w_k \geq 0$.

Figure 8: Some examples of base functions for the weighted harmony (8)

9.3 The paradoxes extend from ME to probabilistic phonology by harmony

Examples of positive, decreasing, normalized base functions include the exponential $f(x) = \exp(-x)$ and the inverse $f(x) = \frac{1}{1+x}$ functions plotted in figure 8. The weighted harmony (8) with the exponential base function $f(x) = \exp(-x)$ is the ME harmony, because $\prod_{k=1}^{n} f(C_k(\mathsf{x},\mathsf{y}))^{w_k} = \exp - \sum_{k=1}^{n} w_k C_k(\mathsf{x},\mathsf{y})$. The exponential and other base functions have similar shapes, as illustrated in figure 8. They thus yield similar grammars, as shown in appendix A.18. Appendices A.19-A.20 thus conclude that:

Theorem 9
The paradoxes derived in sections 4-8 from the harmonic bounding, the cardinality, the one-step-away, the reverse harmonic bounding, and the average faithfulness generalizations extend from ME phonology to the weighted model of intrinsically probabilistic phonology, no matter the choice of the (positive, decreasing, normalized) base function f.

We conclude that not just ME but actually the entire approach of intrinsically probabilistic phonology is not viable, because (despite the assumption of a restrictive model of constraint interaction through factor functions that differ only minimally) it yields typologies so unrestrictive that they effectively capture no linguistic universals.

10. Conclusions

We have compared two approaches to probabilistic phonology. One approach is INTRINSICALLY PROBABILISTIC. It assumes that the grammars internalized by speakers *directly* assign probabilities to mappings, in the form of normalized numerical harmony scores, as in (1). These grammars differ radically from the traditional grammars of categorical phonology. Variation is the default phonological behavior. Categorical patterns only arise as the limit of probabilities that approach the boundary values zero and one.

Another approach to probabilistic phonology is only EXTRINSICALLY PROBABILISTIC. It assumes that the grammars internalized by speakers are traditional categorical grammars. Yet, speakers are not certain where exactly they sit in the categorical typological space. This uncertainty is modeled by the probability mass function P in (2). The default phonological behavior is categorical. Variation with respect to an underlying form only arises *indirectly* when P encodes uncertainty among two or more grammars that happen to differ in the surface realization of that underlying form.

We have argued that extrinsically probabilistic phonological models such as SHG combine solid guarantees for grammatical inference with tight typological predictions. The situation is different for intrinsically probabilistic models. They must rest on harmonies that factorize into the product of n factor functions, to avoid making spurious distinctions. In the interest of restrictiveness, we have allowed these factor functions to differ only minimally (they are all powers of the same base function). In spite of this restriction, the resulting weighted model of intrinsically probabilistic phonology paradoxically fails to predict any universals of markedness, even for sub-typologies

induced by the small sets of mappings and constraints considered here. We conclude that probabilistic natural language phonology is intrinsically categorical, after all.

1. Appendix

A.1 Normalization of probabilistic grammars defined by sampling

Let $\mathcal{T}(\mathsf{x},\mathsf{y})$ denote the subset (9) of the categorical typology $\mathcal{T}$ consisting of those grammars G that realize the underlying form x as the surface candidate y. Because of the assumption that the grammars in $\mathcal{T}$ are all total and strict, the sets $\mathcal{T}(\mathsf{x},\mathsf{y})$ partition the categorical typology $\mathcal{T}$ into disjoint sets as y spans the candidate set $Gen(\mathsf{x})$.

$$\mathcal{T}(\mathsf{x},\mathsf{y}) = \big\{G \in \mathcal{T} \,\big|\, G(\mathsf{x}) = \mathsf{y}\big\} \tag{9}$$

The reasoning in (10) shows that the numbers $G_P(\mathsf{y}\,|\,\mathsf{x})$ defined in (2) are normalized over the candidate set $Gen(\mathsf{x})$ of an underlying form x. Step (10a) holds because of the definition (2) of $G_P(\mathsf{y}\,|\,\mathsf{x})$, restated with the notation in (9). Step (10b) holds because the sets $\mathcal{T}(\mathsf{x},\mathsf{y})$ in (9) partition the categorical typology $\mathcal{T}$. Finally, step (10c) holds because of the assumption that P is a probability mass function on $\mathcal{T}$.

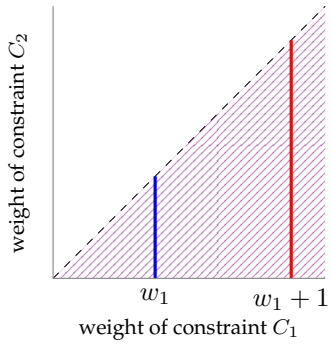
$$\sum_{\mathsf{y}\in Gen(\mathsf{x})} G_P(\mathsf{y}\,|\,\mathsf{x}) \overset{(a)}{=} \sum_{\mathsf{y}\in Gen(\mathsf{x})} \sum_{G\in\mathcal{T}(\mathsf{x},\mathsf{y})} P(G) \overset{(b)}{=} \sum_{G\in\mathcal{T}} P(G) \overset{(c)}{=} 1 \tag{10}$$

A.2 Weight vectors that correspond to no grammars

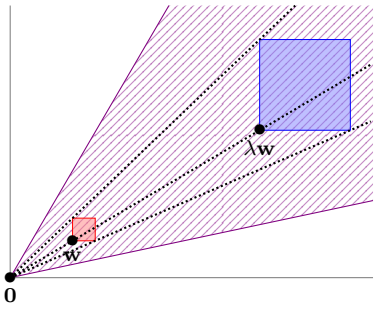
A non-negative weight vector $\mathbf{v} = (v_1, \ldots, v_n)$ CORRESPONDS to a strict and total grammar G provided a mapping (x,y) belongs to G if and only if (x,y) satisfies condition (3). We have collected these weight vectors into the set $\mathcal{W}(G)$. A weight vector $\mathbf{v}$ can also correspond to no strict and total grammar. That happens whenever the candidate set $Gen(\mathsf{x})$ of some underlying form x contains two surface realizations y_1 and y_2 that are different but have the same weighted sum of constraint violations which is in turn smaller than or equal to the weighted sum of constraint violations of any other candidate surface realization z of x. We denote by $\mathcal{N}$ the collection of these non-negative weight vectors that correspond to no strict and total grammar, as in (11).

$$\mathcal{N} = \bigcup_{\substack{(\mathsf{x},\mathsf{y}_1),(\mathsf{x},\mathsf{y}_2)\in Gen\\ \mathsf{y}_1\neq\mathsf{y}_2}} \left\{ \mathbf{v}\in\mathbb{R}^n_+ \;\middle|\; \sum_k v_k C_k(\mathsf{x},\mathsf{y}_1) = \sum_k v_k C_k(\mathsf{x},\mathsf{y}_2) = \min_{\mathsf{z}\in Gen(\mathsf{x})} \sum_k v_k C_k(\mathsf{x},\mathsf{z}) \right\} \tag{11}$$

This set $\mathcal{N}$ of weight vectors that correspond to no strict and total grammars is small, in the following sense. Let us assume that the constraint set $\mathbf{C}$ is DISTINCTIVE: whenever two mappings $(\mathsf{x},\mathsf{y}_1)$ and $(\mathsf{x},\mathsf{y}_2)$ from Gen pair the same underlying form x with two different candidate surface realizations y_1 and y_2, at least one constraint from $\mathbf{C}$ distinguishes these two mappings, whereby $\mathbf{C}(\mathsf{x},\mathsf{y}_1) \neq \mathbf{C}(\mathsf{x},\mathsf{y}_2)$. Thus, the vector $\mathbf{C}(\mathsf{x},\mathsf{y}_1) - \mathbf{C}(\mathsf{x},\mathsf{y}_2)$ is different from the zero vector. In other words, there exists a hyperplane through the origin orthogonal to this vector $\mathbf{C}(\mathsf{x},\mathsf{y}_1) - \mathbf{C}(\mathsf{x},\mathsf{y}_2)$. Each set in the union in (11) is a subset of one such hyperplane. Since Gen contains at most countably



(a)



(b)

Figure 9: Plotting weight vectors when there are only $n = 2$ constraints

infinitely many mappings, the set $\mathcal{N}$ is small in the sense that it has Lebesgue measure equal to zero, because it is the union of countably many subsets of measure zero.

To illustrate, we suppose that Gen consists of only the underlying form x = /CVC/ with the two candidate surface realizations y = [CV] and z = [CVC]. Furthermore, the constraint set **C** consists of only $n = 2$ constraints $C_1 = \text{NoCoda}$ and $C_2 = \text{Max}$, that penalize codas and segment deletions, respectively. The set $\mathcal{W}(G)$ of weight vectors corresponding to the deletion grammar G that realizes /CVC/ as [CV] is the violet region in figure 9a. And the set $\mathcal{N}$ consists of the vectors along the dashed diagonal. Since the constraint set is distinctive, this set $\mathcal{N}$ is small, namely has zero area.

A.3 Normalization of SHG grammars

Let us suppose that the set $\mathcal{N}$ in (11) has zero Lebesgue measure, as discussed in appendix A.2. The reasoning in (12) then shows that the position (4) yields a probability mass function P on the typology $\mathcal{T}$ of strict and total HG grammars. Step (12a) holds because of the definition (4) of P. Step (12b) holds because the sets $\mathcal{W}(G_1)$ and $\mathcal{W}(G_2)$ are disjoint whenever the two strict and total grammars G_1 and G_2 are different. Step (12c) holds because any non-negative weight vector from $\mathbb{R}^n_+$ either corresponds to some strict and total grammar G and thus belongs to the corresponding set $\mathcal{W}(G)$ or else it corresponds to no grammar and thus belongs to the set $\mathcal{N}$ in (11). Step (12d) holds under the assumption that this set $\mathcal{N}$ has zero Lebesgue measure. Finally, step (12e) holds because a density $\mathbf{p}$ that starts at a non-negative weight vector $\mathbf{w}$ concentrates all the probability mass on the set $\mathbb{R}^n_+$ of non-negative weight vectors.

$$\begin{aligned}\sum_{G\in\mathcal{T}} P(G) &\overset{(a)}{=} \sum_{G\in\mathcal{T}} \int_{\mathcal{W}(G)} \mathbf{p}(\mathbf{v})\,\mathrm{d}\mathbf{v} \overset{(b)}{=} \int_{\bigcup_{G\in\mathcal{T}}\mathcal{W}(G)} \mathbf{p}(\mathbf{v})\,\mathrm{d}\mathbf{v} = \\ &\overset{(c)}{=} \int_{\mathbb{R}^n_+\setminus\mathcal{N}} \mathbf{p}(\mathbf{v})\,\mathrm{d}\mathbf{v} \overset{(d)}{=} \int_{\mathbb{R}^n_+} \mathbf{p}(\mathbf{v})\,\mathrm{d}\mathbf{v} \overset{(e)}{=} 1\end{aligned} \tag{12}$$

A.4 Proof of theorem 1

Let us denote by $G^{\text{SHG}}_{\mathbf{w}}(\mathsf{y}\,|\,\mathsf{y})$ the SHG probability of a mapping (x,y) relative to a non-negative weight vector $\mathbf{w}$. Since the product of log-concave functions is log-concave, to show that the SHG likelihood is a log-concave function of the weight vector $\mathbf{w}$, it suffices to show that the SHG probability $G^{\text{SHG}}_{\mathbf{w}}(\mathsf{y}\,|\,\mathsf{x})$ is a log-concave function of $\mathbf{w}$.

Step 1. Let $\mathcal{W}(\mathsf{x},\mathsf{y})$ be the set of non-negative weight vectors that CORRESPOND to the mapping (x,y) because they satisfy condition (3). The reasoning in (13) shows that the SHG probability $G^{\text{SHG}}_{\mathbf{w}}(\mathsf{y}\,|\,\mathsf{x})$ of the mapping (x,y) is the volume of this set $\mathcal{W}(\mathsf{x},\mathsf{y})$. Step (13a) holds because of the definition (2) of probabilistic grammars by sampling, restated using the notation (9). Step (13b) holds because of the definition (4) of the SHG probability mass function P. Step (13c) holds because the sets $\mathcal{W}(G_1)$ and $\mathcal{W}(G_2)$ are disjoint whenever the two strict and total grammars G_1 and G_2 are different. Steps (13d) and (13f) hold under the assumption that the set $\mathcal{N}$ in (11) of weight vectors that correspond to no strict and total grammar has measure zero. Finally, step (13e) holds because of the identity $\mathcal{W}(\mathsf{x},\mathsf{y}) \cup \mathcal{N} = \bigcup_{G\in\mathcal{T}(\mathsf{x},\mathsf{y})} \mathcal{W}(G) \cup \mathcal{N}$. This identity says that a weight vector outside of $\mathcal{N}$ happens to correspond to a mapping (x,y) if and only if it corresponds to a grammar G that realizes the underlying form x as the surface form y.

$$\begin{aligned} G^{\text{SHG}}_{\mathbf{w}}(\mathsf{y}\,|\,\mathsf{x}) &\overset{(a)}{=} \sum_{G\in\mathcal{T}(\mathsf{x},\mathsf{y})} P(G) \overset{(b)}{=} \sum_{G\in\mathcal{T}(\mathsf{x},\mathsf{y})} \int_{\mathcal{W}(G)} \mathbf{p}_{\mathbf{w}}(\mathbf{v})\,\mathrm{d}\mathbf{v} \overset{(c)}{=} \int_{\bigcup_{G\in\mathcal{T}(\mathsf{x},\mathsf{y})}\mathcal{W}(G)} \mathbf{p}_{\mathbf{w}}(\mathbf{v})\,\mathrm{d}\mathbf{v} \\ &\overset{(d)}{=} \int_{\bigcup_{G\in\mathcal{T}(\mathsf{x},\mathsf{y})}\mathcal{W}(G)\cup\mathcal{N}} \mathbf{p}_{\mathbf{w}}(\mathbf{v})\,\mathrm{d}\mathbf{v} \overset{(e)}{=} \int_{\mathcal{W}(\mathsf{x},\mathsf{y})\cup\mathcal{N}} \mathbf{p}_{\mathbf{w}}(\mathbf{v})\,\mathrm{d}\mathbf{v} \overset{(f)}{=} \int_{\mathcal{W}(\mathsf{x},\mathsf{y})} \mathbf{p}_{\mathbf{w}}(\mathbf{v})\,\mathrm{d}\mathbf{v} \end{aligned} \tag{13}$$

Step 2. The reasoning in (14) shows that the SHG probability $G^{\text{SHG}}_{\mathbf{w}}(\mathsf{y}\,|\,\mathsf{x})$ of the mapping (x,y) as a function of the weight vector $\mathbf{w}$ is the CONVOLUTION (more precisely, the CORRELATION) product $\mathbf{p_0} * \mathbb{I}_{\mathcal{W}(\mathsf{x},\mathsf{y})}$ between the probability density function $\mathbf{p_0}$ that starts at the origin $\mathbf{0}$ and the indicator function $\mathbb{I}_{\mathcal{W}(\mathsf{x},\mathsf{y})}$ of the set $\mathcal{W}(\mathsf{x},\mathsf{y})$ of weight vectors corresponding to the mapping (x,y). Step (14a) holds because of the reasoning (13). Step (14b) holds under the assumption that the density $\mathbf{p}_{\mathbf{w}}$ is translation invariant as illustrated in figure 1 because it satisfies the identity $\mathbf{p}_{\mathbf{w}}(\mathbf{v}) = \mathbf{p_0}(\mathbf{v}-\mathbf{w})$. Step (14c) holds because the indicator function $\mathbb{I}_{\mathcal{W}(\mathsf{x},\mathsf{y})}$ is equal to one on the weight vectors $\mathbf{v}$ that belongs to the set $\mathcal{W}(\mathsf{x},\mathsf{y})$ and equal to zero elsewhere.

$$\begin{aligned} G^{\text{SHG}}_{\mathbf{w}}(\mathsf{y}\,|\,\mathsf{x}) &\overset{(a)}{=} \int_{\mathcal{W}(\mathsf{x},\mathsf{y})} \mathbf{p}_{\mathbf{w}}(\mathbf{v})\mathrm{d}\mathbf{v} \overset{(b)}{=} \int_{\mathcal{W}(\mathsf{x},\mathsf{y})} \mathbf{p_0}(\mathbf{v}-\mathbf{w})\mathrm{d}\mathbf{v} \\ &\overset{(c)}{=} \int \mathbf{p_0}(\mathbf{v}-\mathbf{w})\mathbb{I}_{\mathcal{W}(\mathsf{x},\mathsf{y})}(\mathbf{v})\mathrm{d}\mathbf{v} = \mathbf{p_0} * \mathbb{I}_{\mathcal{W}(\mathsf{x},\mathsf{y})}(\mathbf{w}) \end{aligned} \tag{14}$$

Step 3. The SHG probability $G^{\text{SHG}}_{\mathbf{w}}(\mathsf{y}\,|\,\mathsf{x})$ can now be shown to be a log-concave function of the weight vector $\mathbf{w}$ through the following reasoning from Boyd and Vandenberghe (2004, pages 106-107). The set $\mathcal{W}(\mathsf{x},\mathsf{y})$ of weight vectors that correspond to the mapping (x,y) is convex because defined through the linear inequalities in (3). The function $(\mathbf{v},\mathbf{w}) \mapsto \mathbb{I}_{\mathcal{W}(\mathsf{x},\mathsf{y})}(\mathbf{v})$ is thus log-concave, because the indicator function of a convex set is log-concave. The uniform, exponential, and half-Gaussian densities are log-concave (Boyd and Vandenberghe 2004, example 3.40, pages 104-105). The function

$(\mathbf{v}, \mathbf{w}) \mapsto \mathbf{p_0}(\mathbf{v} - \mathbf{w})$ is thus log-concave because the affine function $(\mathbf{v}, \mathbf{w}) \mapsto \mathbf{v} - \mathbf{w}$ is log-concave and the composition of log-concave functions is log-concave. In conclusion, the function $(\mathbf{v}, \mathbf{w}) \mapsto \mathbf{p_0}(\mathbf{v} - \mathbf{w})\mathbb{I}_{\mathcal{W}(\mathsf{x},\mathsf{y})}(\mathbf{v})$ is log-concave because it is the product of two log-concave functions. We conclude that the function $\mathbf{w} \mapsto G^{\mathrm{SHG}}_{\mathbf{w}}(\mathsf{y}\,|\,\mathsf{x})$ is log-concave because Prékopa's theorem (Prékopa 1971, 1973) ensures that the function $\mathbf{w} \in \mathbb{R}^n \mapsto \int f(\mathbf{v}, \mathbf{w})\mathrm{d}\mathbf{v}$ is log-concave whenever $f : (\mathbf{v}, \mathbf{w}) \in \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}$ is log-concave.

A.5 Derivatives of the SHG likelihood function

The derivatives of the SHG log-likelihood function are sums of derivatives of the logarithm of the SHG probability $G^{\mathrm{SHG}}_{\mathbf{w}}(\mathsf{y}\,|\,\mathsf{x})$ of a mapping (x, y). To understand the geometric interpretation of these derivatives, we denote by $\mathsf{Vol}^n_{w_1,w_2,\ldots,w_n}(\mathcal{W})$ the volume (15) of the set $\mathcal{W}(\mathsf{x}, \mathsf{y})$ of corresponding weight vectors relative to the product $p_{w_1} \cdot p_{w_2} \cdot \ldots \cdot p_{w_n}$ of n probability density functions on $\mathbb{R}_+$ starting at $w_1, w_2, \ldots, w_n$

$$\mathsf{Vol}^n_{w_1,w_2,\ldots,w_n}(\mathcal{W}) = \int p_{w_1}(v_1) \int p_{w_2}(v_2) \cdots \int p_{w_n}(v_n)\mathbb{I}_{\mathcal{W}}(v_1, v_2, \ldots, v_n)\mathrm{d}v_1\mathrm{d}v_2 \cdots \mathrm{d}v_n \tag{15}$$

Let H_{w_1} be the hyperplane consisting of those vectors whose first coordinate is equal to w_1. The intersection between the set $\mathcal{W}(\mathsf{x}, \mathsf{y})$ and this hyperplane H_{w_1} can be construed as a subset of (namely, its affine hull is) $\mathbb{R}^{n-1}$. We thus denote by $\mathsf{Vol}^{n-1}_{w_2,\ldots,w_n}\left(\mathcal{W} \cap H_{w_1}\right)$ the volume (16) of this intersection relative to the product density $p_{w_2} \cdot \ldots \cdot p_{w_n}$ on $\mathbb{R}^{n-1}_+$. To illustrate, if $\mathcal{W}(\mathsf{x}, \mathsf{y}) = \mathcal{W}(/\mathsf{CVC}/, [\mathsf{CV}])$ is the violet region in figure 9a, the intersection $\mathcal{W}(\mathsf{x}, \mathsf{y}) \cap H_{w_1}$ is the thick blue segment.

$$\mathsf{Vol}^{n-1}_{w_2,\ldots,w_n}(\mathcal{W} \cap H_{w_1}) = \int p_{w_2}(v_2) \cdots \int p_{w_n}(v_n)\mathbb{I}_{\mathcal{W}}(w_1, v_2, \ldots, v_n) \quad \mathrm{d}v_2 \cdots \mathrm{d}v_n \tag{16}$$

The following theorem computes the first derivative of the logarithm of the SHG probability $G^{\mathrm{SHG}}_{\mathbf{w}}(\mathsf{y}\,|\,\mathsf{x})$. The numerator in the right-hand side of (17) is the difference between the $(n-1)$-dimensional volume of the intersection between $\mathcal{W}$ and the hyperplane H_{w_1+1} (consisting of vectors with first component equal to $w_1 + 1$) minus the $(n-1)$-dimensional volume of the intersection between $\mathcal{W}$ and the hyperplane H_{w_1} (consisting of vectors with first component equal to w_1). Continuing with the example in figure 9a, this is the difference between the length (relative to the uniform density) of the red segment minus the length of the blue segment. The numerator of (18) is the difference between the n-dimensional volume of $\mathcal{W}$ minus the $(n-1)$-dimensional volume of the intersection between $\mathcal{W}$ and the hyperplane H_{w_1}. Continuing with figure 9a, this is the difference between the area (relative to the exponential density) of the violet region minus the length of the red segment.

The set $\mathcal{W}$ is a polyhedron under the assumption made here that the candidate set $Gen(\mathsf{x})$ is finite. Furthermore, the intersection between $\mathcal{W}$ and a hyperplane is a polyhedron as well (because the intersection of polyhedra is a polyhedron). The theorem thus says that that the problem of computing the derivatives of the logarithm of the SHG likelihood function effectively boils down to the problem of computing polyhedral volumes in n or $n-1$ dimensions.

Theorem 10
When SHG is defined using uniform densities $\boldsymbol{p}^{unif}_{\mathbf{w}}$ or exponential densitites $\boldsymbol{p}^{exp}_{\mathbf{w}}$, the first derivative with respect to w_1 of the logarithm of the SHG probability $G^{SHG}_{(w_1,w_2,\ldots,w_n)}(\mathsf{y}\,|\,\mathsf{x})$ can

be made explicit as in (17) and (18), respectively.

$$\frac{\partial}{\partial w_1} \log G^{SHG}_{(w_1,w_2,\dots,w_n)}(\mathsf{y}\,|\,\mathsf{x}) = \frac{\mathsf{Vol}^{n-1}_{w_2,\dots,w_n}(\mathcal{W} \cap H_{w_1+1}) - \mathsf{Vol}^{n-1}_{w_2,\dots,w_n}(\mathcal{W} \cap H_{w_1})}{\mathsf{Vol}^{n}_{w_1,w_2,\dots,w_n}(\mathcal{W})} \tag{17}$$

$$\frac{\partial}{\partial w_1} \log G^{SHG}_{(w_1,w_2,\dots,w_n)}(\mathsf{y}\,|\,\mathsf{x}) = \frac{\mathsf{Vol}^{n}_{w_1,w_2,\dots,w_n}(\mathcal{W}) - \mathsf{Vol}^{n-1}_{w_2,\dots,w_n}(\mathcal{W} \cap H_{w_1})}{\mathsf{Vol}^{n}_{w_1,w_2,\dots,w_n}(\mathcal{W})} \tag{18}$$

Proof. We start by computing the derivative of the logarithm of the SHG probability as in (19). Step (19a) holds because of basic properties of logarithms and derivatives. Step (19b) holds because (13) ensures that the SHG probability of a mapping (x,y) is the volume $\mathsf{Vol}^{n}_{w_1,w_2,\dots,w_n}(\mathcal{W})$. Step (19c) holds because of the expression (15) of this volume. Finally, step (19d) holds because of the expression (16) of $\mathsf{Vol}^{n-1}_{w_2,\dots,w_n}(\mathcal{W} \cap H_{w_1})$.

$$\begin{aligned}
&\frac{\partial}{\partial w_1} \log G^{\mathrm{SHG}}_{(w_1,w_2,\dots,w_n)}(\mathsf{y}\,|\,\mathsf{x}) = \\
&\overset{(a)}{=} \frac{1}{G^{\mathrm{SHG}}_{w_1,w_2,\dots,w_n}(\mathsf{y}\,|\,\mathsf{x})} \frac{\partial}{\partial w_1} G^{\mathrm{SHG}}_{w_1,w_2,\dots,w_n}(\mathsf{y}\,|\,\mathsf{x}) \\
&\overset{(b)}{=} \frac{1}{\mathsf{Vol}^{n}_{w_1,w_2,\dots,w_n}(\mathcal{W})} \frac{\partial}{\partial w_1} \mathsf{Vol}^{n}_{w_1,w_2,\dots,w_n}(\mathcal{W}) \\
&\overset{(c)}{=} \frac{1}{\mathsf{Vol}^{n}_{w_1,w_2,\dots,w_n}(\mathcal{W})} \frac{\partial}{\partial w_1} \int p_{w_1}(v_1) \int p_{w_2}(v_2) \cdots \int p_{wn}(v_n) \mathbb{I}_{\mathcal{W}}(\mathbf{v})\, \mathrm{d}v_1\, \mathrm{d}v_2 \cdots \mathrm{d}v_n \\
&\overset{(d)}{=} \frac{1}{\mathsf{Vol}^{n}_{w_1,w_2,\dots,w_n}(\mathcal{W})} \underbrace{\frac{\partial}{\partial w_1} \int p_{w_1}(v_1) \mathsf{Vol}^{n-1}_{w_2,\dots,w_n}(\mathcal{W} \cap H_{v_1})\, \mathrm{d}v_1}_{(*)}
\end{aligned} \tag{19}$$

To compute the derivative (*), we recall that the fundamental theorem of calculus (Rudin 1953, theorem 6.18, page 98) ensures that $\frac{\mathrm{d}}{\mathrm{d}x}\int_a^x f(t)\mathrm{d}t = f(x)$ whenever f is a real-valued function defined on a closed interval $[a,b]$ which is continuous at $x \in [a,b]$. The expression (20) then follows for every $x \geq 0$.

$$\frac{\mathrm{d}}{\mathrm{d}x} \int_x^{x+\lambda} f(t)\, \mathrm{d}t = \frac{\mathrm{d}}{\mathrm{d}x} \left\{ \int_0^{x+\lambda} f(t)\, \mathrm{d}t - \int_0^{x} f(t)\, \mathrm{d}t \right\} = f(x+\lambda) - f(x) \tag{20}$$

When p_{w_1} is the uniform density $p_{w_1}(v) = \mathbb{I}_{[w_1,w_1+1]}(v)$, the derivative (*) becomes (21). Step (21a) holds because of the definition of uniform densities. Step (21b) holds because of (20), that applies because the function $v_1 \mapsto \mathsf{Vol}^{n-1}_{w_2,\dots,w_n}\left(\mathcal{W} \cap H_{v_1}\right)$ is continuous, as it is obvious from its geometric interpretation.

$$\begin{aligned}
(*) &\overset{(a)}{=} \frac{\partial}{\partial w_1} \int_{w_1}^{w_1+1} \mathsf{Vol}^{n-1}_{w_2,\dots,w_n}(\mathcal{W} \cap H_{v_1})\, \mathrm{d}v_1 \\
&\overset{(b)}{=} \left\{ \mathsf{Vol}^{n-1}_{w_2,\dots,w_n}(\mathcal{W} \cap H_{w_1+1}) - \mathsf{Vol}^{n-1}_{w_2,\dots,w_n}(\mathcal{W} \cap H_{w_1}) \right\}
\end{aligned} \tag{21}$$

When p_{w_1} is the exponential density $p_{w_1}(v) = \exp(w_1 - v)\mathbb{I}_{[w_1,\infty)}(v)$, the derivative (*) becomes (22). Steps (22a) and (22c) hold because of the definition of exponential

densities. Step (22b) holds because of (20) and the rule for the derivative of products.

$$
\begin{aligned}
(*) &\overset{(a)}{=} \frac{\partial}{\partial w_1} e^{w_1} \int_{w_1}^{\infty} e^{-v_1} \mathsf{Vol}^{n-1}_{w_2,\dots,w_n}(\mathcal{W} \cap H_{v_1})\, \mathrm{d}v_1 \\
&\overset{(b)}{=} e^{w_1} \int_{w_1}^{\infty} e^{-v_1} \mathsf{Vol}^{n-1}_{w_2,\dots,w_n}(\mathcal{W} \cap H_{v_1})\, \mathrm{d}v_1 - e^{w_1} e^{-w_1} \mathsf{Vol}^{n-1}_{w_2,\dots,w_n}(\mathcal{W} \cap H_{w_1}) \\
&\overset{(c)}{=} \int p_{w_1}(v_1) \mathsf{Vol}^{n-1}_{w_2,\dots,w_n}(\mathcal{W} \cap H_{v_1})\, \mathrm{d}v_1 - \mathsf{Vol}^{n-1}_{w_2,\dots,w_n}(\mathcal{W} \cap H_{w_1}) \\
&= \left\{ \mathsf{Vol}^{n}_{w_1,w_2,\dots,w_n}(\mathcal{W}) - \mathsf{Vol}^{n-1}_{w_2,\dots,w_n}(\mathcal{W} \cap H_{w_1}) \right\}
\end{aligned}
\tag{22}
$$

Expressions (17) and (18) finally follow by plugging into (19) the expressions for the derivative (*) obtained in (21) and (22). □

A.6 First half of the proof of theorem 2: comparing SHG and HG universals

Let us suppose that an implication $(\mathsf{x}, \mathsf{y}) \to (\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ is a universal of the categorical HG typology $\mathcal{T}$ in the sense of condition (5). Thus every grammar in $\mathcal{T}$ 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 $\widehat{\mathsf{y}}$. With the notation in (9), this means that the categorical HG typology $\mathcal{T}$ satisfies the inclusion $\mathcal{T}(\mathsf{x}, \mathsf{y}) \subseteq \mathcal{T}(\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$. This inclusion entails in turn the inequality $P(\mathcal{T}(\mathsf{x}, \mathsf{y})) \leq P(\mathcal{T}(\widehat{\mathsf{x}}, \widehat{\mathsf{y}}))$, because a probability mass function P is monotone relative to set inclusion. Finally, this inequality can be rewritten as $G_P(\mathsf{y} \,|\, \mathsf{x}) \leq G_P(\widehat{\mathsf{y}} \,|\, \widehat{\mathsf{x}})$, because of the definition (2) of the probabilistic grammar G_P obtained by sampling according to P. Since the latter inequality holds uniformly for any probability mass function P, the implication $(\mathsf{x}, \mathsf{y}) \to (\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ is also a universal of the probabilistic SHG typology in the sense of condition (6).

To conclude that HG and SHG share the same implicational universals, let us assume by contradiction that the implication $(\mathsf{x}, \mathsf{y}) \to (\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ is a universal of the SHG typology but not of the HG typology. Thus, some HG grammar G realizes the antecedent underlying form x as the antecedent surface form y but realizes the consequent underlying form $\widehat{\mathsf{x}}$ as some surface form $\widehat{\mathsf{z}}$ different from the consequent form $\widehat{\mathsf{y}}$, as in (23).

$$
G(\mathsf{x}) = \mathsf{y}, \qquad G(\widehat{\mathsf{x}}) = \widehat{\mathsf{z}} \tag{23}
$$

Since G is a HG grammar, it corresponds to some non-negative weight vector $\mathbf{w}$. The assumption $G(\mathsf{x}) = \mathsf{y}$ in (23) then entails that $\mathbf{w}$ belongs to the set of non-negative weight vectors $\mathcal{W}(\mathsf{x}, \mathsf{y})$ that correspond to this mapping (x, y) in the sense that they satisfy condition (3). Since this condition is stated in terms of STRICT inequalities, the set $\mathcal{W}(\mathsf{x}, \mathsf{y})$ is OPEN relative to the topology induced by $\mathbb{R}^n$ on $\mathbb{R}^n_+$. As a result, since $\mathcal{W}(\mathsf{x}, \mathsf{y})$ contains $\mathbf{w}$, there exists $\delta > 0$ small enough that $\mathcal{W}(\mathsf{x}, \mathsf{y})$ contains the entire cube $\mathbf{w} + [0, \delta)^n$ that starts at $\mathbf{w}$ and has side δ, plotted as a red square in figure 9b.

Since condition (3) is stated in terms of HOMOGENEOUS inequalities, the set $\mathcal{W}(\mathsf{x}, \mathsf{y})$ is also CONIC: if it contains a vector $\mathbf{v} = (v_1, \dots, v_n)$, it also contains the rescaled vector $\lambda\mathbf{v} = (\lambda v_1, \dots, \lambda v_n)$ for any coefficient $\lambda > 0$. As a result, since $\mathcal{W}(\mathsf{x}, \mathsf{y})$ contains the cube $\mathbf{w} + [0, \delta)^n$ that starts at $\mathbf{w}$ and has side δ, it also contains the cube $\lambda\mathbf{w} + [0, \lambda\delta)^n$ that starts at $\lambda\mathbf{w}$ and has side $\lambda\delta$, plotted as a blue square in figure 9b. In other words, the side of the cube can be arbitrarily increased when the starting vertex is shifted accordingly.

The reasoning in (24) then shows that the SHG probability $G^{\text{SHG}}_{\lambda\mathbf{w}}(\mathsf{y}\,|\,\mathsf{x})$ corresponding to the shifted vector $\lambda\mathbf{w}$ is larger than $1/2$ as long as λ is large enough. Step (24a) holds because (13) has shown that the SHG probability of the mapping (x,y) corresponding to a certain weight vector is equal to the volume of the set $\mathcal{W}(\mathsf{x},\mathsf{y})$ relative to the density that starts at that weight vector. Step (24b) holds because $\mathcal{W}(\mathsf{x},\mathsf{y})$ contains the cube $\lambda\mathbf{w}+[0,\lambda\delta)^n$ for any $\lambda>0$. Finally, step (24c) holds because the density $\mathbf{p}_{\lambda\mathbf{w}}$ starts at $\lambda\mathbf{w}$ and therefore concentrates more than half of the probability mass on the cube $\lambda\mathbf{w}+[0,\lambda\delta)^n$ that starts at $\lambda\mathbf{w}$ and has side $\lambda\delta$, as long as λ is large enough.

$$G^{\text{SHG}}_{\lambda\mathbf{w}}(\mathsf{y}\,|\,\mathsf{x}) \overset{(a)}{=} \int_{\mathcal{W}(\mathsf{x},\mathsf{y})} \mathbf{p}_{\lambda\mathbf{w}}(\mathbf{v})\,\mathrm{d}\mathbf{v} \overset{(b)}{\geq} \int_{\lambda\mathbf{w}+[0,\lambda\delta)^n} \mathbf{p}_{\lambda\mathbf{w}}(\mathbf{v})\,\mathrm{d}\mathbf{v} \overset{(c)}{\geq} \frac{1}{2} \tag{24}$$

An analogous reasoning deduces from the assumption $G(\widehat{\mathsf{x}})=\widehat{\mathsf{z}}$ in (23) that also $G^{\text{SHG}}_{\lambda\mathbf{w}}(\widehat{\mathsf{z}}\,|\,\widehat{\mathsf{x}})>1/2$ when λ is large enough. Since $G^{\text{SHG}}_{\lambda\mathbf{w}}(\widehat{\mathsf{z}}\,|\,\widehat{\mathsf{x}})>1/2$, then $G^{\text{SHG}}_{\lambda\mathbf{w}}(\widehat{\mathsf{y}}\,|\,\widehat{\mathsf{x}})<1/2$, because of normalization. In conclusion, this SHG grammar $G^{\text{SHG}}_{\lambda\mathbf{w}}$ corresponding to the shifted vector $\lambda\mathbf{w}$ assigns a probability larger than $1/2$ to the antecedent mapping (x,y) but a probability smaller than $1/2$ to the consequent mapping $(\widehat{\mathsf{x}},\widehat{\mathsf{y}})$, contradicting the assumption that the implication $(\mathsf{x},\mathsf{y})\to(\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ is a universal of the SHG typology.

A.7 ME basic lemma

The following lemma characterizes the implications $(\mathsf{x},\mathsf{y})\to(\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ that are ME universals. Let $\mathsf{z}_1,\dots,\mathsf{z}_m$ be the additional candidates of the antecedent underlying form x besides the antecedent candidate y. We define the vector $\mathbf{C}^{\mathsf{y}}(\mathsf{x},\mathsf{z}_i)$ of DISCOUNTED constraint violations of one of these additional candidates z_i as the difference between its vector $\mathbf{C}(\mathsf{x},\mathsf{z}_i)$ of constraint violations minus the vector $\mathbf{C}(\mathsf{x},\mathsf{y})$ of constraint violations of y. Analogously, let $\widehat{\mathsf{z}}_1,\dots,\widehat{\mathsf{z}}_{\widehat{m}}$ be the additional candidates of the consequent underlying form $\widehat{\mathsf{x}}$ besides the consequent candidate $\widehat{\mathsf{y}}$. We define the vector of consequent discounted constraint violations $\mathbf{C}^{\widehat{\mathsf{y}}}(\widehat{\mathsf{x}},\widehat{\mathsf{z}}_j)$ analogously as in (25).

$$\mathbf{C}^{\mathsf{y}}(\mathsf{x},\mathsf{z}_i)=\mathbf{C}(\mathsf{x},\mathsf{z}_i)-\mathbf{C}(\mathsf{x},\mathsf{y}),\qquad \mathbf{C}^{\widehat{\mathsf{y}}}(\widehat{\mathsf{x}},\widehat{\mathsf{z}}_j)=\mathbf{C}(\widehat{\mathsf{x}},\widehat{\mathsf{z}}_j)-\mathbf{C}(\widehat{\mathsf{x}},\widehat{\mathsf{y}}) \tag{25}$$

Finally, we denote by $h_{\mathbf{w}}(\mathbf{x})$ the opposite of the average of the components of a vector $\mathbf{x}=(x_1,\dots,x_n)$ weighed by the weights $\mathbf{w}=(w_1,\dots,w_n)$, as stated in (26).

$$h_{\mathbf{w}}(\mathbf{x})=-\sum_{k=1}^{n} w_k x_k \tag{26}$$

The theory of ME universals developed here consists of a careful analysis of the following inequality (27), that will thus be referred to as the ME BASIC inequality.

Lemma 1
An implication $(\mathsf{x},\mathsf{y})\to(\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ is a ME universal in the sense of condition (6) if and only if every non-negative weight vector $\mathbf{w}$ satisfies the inequality (27) between discounted violations.

$$\sum_{i=1}^{m}\exp h_{\mathbf{w}}(\mathbf{C}^{\mathsf{y}}(\mathsf{x},\mathsf{z}_i))\geq\sum_{j=1}^{\widehat{m}}\exp h_{\mathbf{w}}(\mathbf{C}^{\widehat{\mathsf{y}}}(\widehat{\mathsf{x}},\widehat{\mathsf{z}}_j)) \tag{27}$$

Proof. The reasoning in (28) unpacks the ME probability $G_{\mathbf{w}}^{\mathrm{ME}}(\mathsf{y}\,|\,\mathsf{x})$ of the antecedent mapping (x,y) relative to a weight vector $\mathbf{w}$. Step (28a) holds because of the definition (1) of harmony-based probabilistic grammars. Step (28b) holds because the normalization constant $Z(\mathsf{x})$ is the sum of the ME harmonies of all the candidates of the underlying form x, namely y plus $\mathsf{z}_1,\ldots,\mathsf{z}_m$. Step (28c) holds because the ME harmony corresponding to a weight vector $\mathbf{w}$ is the exponential of the function $h_{\mathbf{w}}$ in (26) applied to the constraint violation vectors. Step (28d) holds by dividing both numerator and denominator by the quantity $\exp h_{\mathbf{w}}(\mathbf{C}(\mathsf{x},\mathsf{y}))$, which is different from zero. Finally, step (28e) holds because of the definition (25) of the discounted constraint violations.

$$\begin{aligned} G_{\mathbf{w}}^{\mathrm{ME}}(\mathsf{y}\,|\,\mathsf{x}) &\overset{(a)}{=} \frac{H(\mathsf{x},\mathsf{y})}{Z(\mathsf{x})} \overset{(b)}{=} \frac{H(\mathsf{x},\mathsf{y})}{H(\mathsf{x},\mathsf{y}) + \sum_{i=1}^{m} H(\mathsf{x},\mathsf{z}_i)} \\ &\overset{(c)}{=} \frac{\exp h_{\mathbf{w}}(\mathbf{C}(\mathsf{x},\mathsf{y}))}{\exp h_{\mathbf{w}}(\mathbf{C}(\mathsf{x},\mathsf{y})) + \sum_{i=1}^{m} \exp h_{\mathbf{w}}(\mathbf{C}(\mathsf{x},\mathsf{z}_i))} \\ &\overset{(d)}{=} \frac{1}{1 + \sum_{i=1}^{m} \frac{\exp h_{\mathbf{w}}(\mathbf{C}(\mathsf{x},\mathsf{z}_i))}{\exp h_{\mathbf{w}}(\mathbf{C}(\mathsf{x},\mathsf{y}))}} \overset{(e)}{=} \frac{1}{1 + \sum_{i=1}^{m} \exp h_{\mathbf{w}}(\mathbf{C}^{\mathsf{y}}(\mathsf{x},\mathsf{z}_i))} \end{aligned} \tag{28}$$

The ME probability $G_{\mathbf{w}}^{\mathrm{ME}}(\widehat{\mathsf{y}}\,|\,\widehat{\mathsf{x}})$ of the consequent mapping $(\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ admits the analogous expression (29) in terms of consequent discounted constraint violations.

$$G_{\mathbf{w}}^{\mathrm{ME}}(\widehat{\mathsf{y}}\,|\,\widehat{\mathsf{x}}) = \frac{1}{1 + \sum_{j=1}^{\widehat{m}} \exp h_{\mathbf{w}}(\mathbf{C}^{\widehat{\mathsf{y}}}(\widehat{\mathsf{x}},\widehat{\mathsf{z}}_j))} \tag{29}$$

In conclusion, the implication $(\mathsf{x},\mathsf{y}) \to (\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ is a ME universal in the sense of condition (6) if and only if the probability inequality $G_{\mathbf{w}}^{\mathrm{ME}}(\mathsf{y}\,|\,\mathsf{x}) \leq G_{\mathbf{w}}^{\mathrm{ME}}(\widehat{\mathsf{y}}\,|\,\widehat{\mathsf{x}})$ holds for any non-negative weight vector $\mathbf{w}$. Condition (27) then follows by plugging in the expressions for the antecedent and consequent ME probabilities obtained in (28) and (29). □

A.8 Second half of the proof of theorem 2: comparing ME and HG universals

We suppose that an implication $(\mathsf{x},\mathsf{y}) \to (\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ is a ME universal in the sense of condition (6) and we want to show that it is also a HG universal in the sense of condition (5). To this end, we consider an arbitrary HG grammar that realizes the antecedent underlying form x as the antecedent surface form y and we show that it also realizes the consequent underlying form $\widehat{\mathsf{x}}$ as the consequent surface form $\widehat{\mathsf{y}}$. Equivalently, we consider an arbitrary weight vector $\mathbf{w} = (w_1,\ldots,w_n)$ that satisfies condition (3) for the antecedent mapping (x,y) and we show that $\mathbf{w}$ also satisfies (3) for the consequent mapping $(\widehat{\mathsf{x}},\widehat{\mathsf{y}})$.

The assumption that $\mathbf{w}$ satisfies (3) for the antecedent mapping (x,y) can be restated in terms of the discounted violations in (25) as the inequality $\max_i h_{\mathbf{w}}\big(\mathbf{C}^{\mathsf{y}}(\mathsf{x},\mathsf{z}_i)\big) < 0$. Since the latter inequality is strict, it entails (30) for some large $\lambda > 0$.

$$\lambda \max_{i=1,\ldots,m} h_{\mathbf{w}}\big(\mathbf{C}^{\mathsf{y}}(\mathsf{x},\mathsf{z}_i)\big) < -\log m \tag{30}$$

The reasoning in (31) then holds for any additional consequent candidate $\widehat{\mathsf{z}}_j$. Step (31a) holds because a sum of non-negative terms is always larger than each of its terms. Steps (31b) and (31d) hold because $\lambda h_{\mathbf{w}} = h_{\lambda \mathbf{w}}$. Step (31c) holds because the ME implication $(\mathsf{x}, \mathsf{y}) \to (\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ satisfies the ME basic inequality (27) for any non-negative weight vector and thus in particular for the rescaled weight vector $\lambda \mathbf{w}$. Finally, step (31e) holds because of (30).

$$\begin{aligned}
\exp \lambda h_{\mathbf{w}}\big(\mathbf{C}^{\widehat{\mathsf{y}}}(\mathsf{x}, \widehat{\mathsf{z}}_j)\big) &\overset{(a)}{\leq} \sum_{j=1}^{\widehat{m}} \exp \lambda h_{\mathbf{w}}\big(\mathbf{C}^{\widehat{\mathsf{y}}}(\mathsf{x}, \widehat{\mathsf{z}}_j)\big) \overset{(b)}{=} \sum_{j=1}^{\widehat{m}} \exp h_{\lambda \mathbf{w}}\big(\mathbf{C}^{\widehat{\mathsf{y}}}(\mathsf{x}, \widehat{\mathsf{z}}_j)\big) \\
&\overset{(c)}{\leq} \sum_{i=1}^{m} \exp h_{\lambda \mathbf{w}}\big(\mathbf{C}^{\mathsf{y}}(\mathsf{x}, \mathsf{z}_i)\big) \overset{(d)}{=} \sum_{i=1}^{m} \exp \lambda h_{\mathbf{w}}\big(\mathbf{C}^{\mathsf{y}}(\mathsf{x}, \mathsf{z}_i)\big) \\
&\leq \sum_{i=1}^{m} \exp \lambda \max_{i=1,\dots,m} h_{\mathbf{w}}\big(\mathbf{C}^{\mathsf{y}}(\mathsf{x}, \mathsf{z}_i)\big) \overset{(e)}{<} \sum_{i=1}^{m} \exp(-\log m) \;=\; 1
\end{aligned} \tag{31}$$

Since $\lambda > 0$, the inequality $\exp \lambda h_{\mathbf{w}}\big(\mathbf{C}^{\widehat{\mathsf{y}}}(\mathsf{x}, \widehat{\mathsf{z}}_j)\big) < 1$ thus obtained holds if and only if $h_{\mathbf{w}}(\mathbf{C}^{\widehat{\mathsf{y}}}(\widehat{\mathsf{x}}, \widehat{\mathsf{z}}_j)) < 0$. Equivalently $\sum_{k=1}^{n} w_k C_k(\widehat{\mathsf{x}}, \widehat{\mathsf{y}}) < \sum_{k=1}^{n} w_k C_k(\widehat{\mathsf{x}}, \widehat{\mathsf{z}}_j)$. Since this conclusion holds for every additional consequent candidate $\widehat{\mathsf{z}}_j$, the weight vector $\mathbf{w}$ satisfies condition (3) also for the consequent mapping $(\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$.

A.9 ME one-versus-convex-sum lemma

The inequality (32) compares a vector of consequent discounted constraint violations with a convex sum of the vectors of antecedent discounted constraint violations. The harmonic bounding, one-step-away, and reverse harmonic bounding generalizations will be derived from this one-versus-convex-sum inequality.

Lemma 2
Suppose that an implication $(\mathsf{x}, \mathsf{y}) \to (\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ is a ME universal. Then, the vector $\mathbf{C}^{\widehat{\mathsf{y}}}(\widehat{\mathsf{x}}, \widehat{\mathsf{z}}_j)$ of discounted constraint violations of each additional consequent candidate $\widehat{\mathsf{z}}_j$ is at least as large as some convex sum of the vectors $\mathbf{C}^{\mathsf{y}}(\mathsf{x}, \mathsf{z}_1), \dots, \mathbf{C}^{\mathsf{y}}(\mathsf{x}, \mathsf{z}_m)$ of discounted constraint violations of the m additional antecedent candidates $\mathsf{z}_1, \dots, \mathsf{z}_m$. In other words, the inequality (32) holds for some non-negative coefficients $\lambda_1, \dots, \lambda_m \geq 0$ (that depend on the additional consequent candidate $\widehat{\mathsf{z}}_j$ considered) that add up to 1, namely $\sum_{i=1}^{m} \lambda_i = 1$.

$$\mathbf{C}^{\widehat{\mathsf{y}}}(\widehat{\mathsf{x}}, \widehat{\mathsf{z}}_j) \geq \sum_{i=1}^{m} \lambda_i \mathbf{C}^{\mathsf{y}}(\mathsf{x}, \mathsf{z}_i) \tag{32}$$

Proof. Let $\mathbf{c}_i$ denote the vector $\mathbf{C}^{\mathsf{y}}(\mathsf{x}, \mathsf{z}_i)$ of antecedent discounted constraint violations and $\widehat{\mathbf{c}}_j$ denote the vector $\mathbf{C}^{\widehat{\mathsf{y}}}(\widehat{\mathsf{x}}, \widehat{\mathsf{z}}_j)$ of consequent discounted constraint violations. The proof relies on the following straightforward formula (33), whereby the log-sum-exp function $\mathrm{LSE}(x_1, \dots, x_M) = \log \sum_{h=1}^{M} \exp x_h$ provides a smooth approximation of the maximum function $\max\{x_1, \dots, x_M\}$ (Boyd and Vandenberghe 2004, p. 72-73).

$$\max\{x_1, \dots, x_M\} \overset{(a)}{\leq} \underbrace{\log \sum_{h=1}^{M} \exp x_h}_{\mathrm{LSE}(x_1, \dots, x_M)} \overset{(b)}{\leq} \max\{x_1, \dots, x_M\} + \log M \tag{33}$$

Step 1. If the ME basic inequality (27) holds for some weight vector $\mathbf{w}$, then the following inequality (34) holds for that weight vector $\mathbf{w}$ as well.

$$\max_{j=1,\dots,\widehat{m}} h_{\mathbf{w}}(\widehat{\mathbf{c}}_j) \leq \max_{i=1,\dots,m} h_{\mathbf{w}}(\mathbf{c}_i) + \log m \tag{34}$$

In fact, steps (35a) and (35c) hold because of the LSE approximation formulas (33a) and (33b), respectively. Furthermore, step (35b) holds because of the ME basic inequality (27), rewritten with the simplified notation $\mathbf{c}_i, \widehat{\mathbf{c}}_j$ adopted here (and with the addition of logarithms at both sides).

$$\max_{j=1,\dots,\widehat{m}} h_{\mathbf{w}}(\widehat{\mathbf{c}}_j) \overset{(a)}{\leq} \log \sum_{j=1}^{\widehat{m}} \exp h_{\mathbf{w}}(\widehat{\mathbf{c}}_j) \overset{(b)}{\leq} \log \sum_{i=1}^{m} \exp h_{\mathbf{w}}(\mathbf{c}_i) \overset{(c)}{\leq} \max_{i=1,\dots,m} h_{\mathbf{w}}(\mathbf{c}_i) + \log m \tag{35}$$

Step 2. If the inequality (34) holds for *every* non-negative weight vector $\mathbf{w}$, the term $\log m$ on its right-hand side can be dropped: the following inequality (36) holds for every non-negative weight vector $\mathbf{w}$ as well.

$$\max_{j=1,\dots,m} h_{\mathbf{w}}(\widehat{\mathbf{c}}_j) \leq \max_{i=1,\dots,m} h_{\mathbf{w}}(\mathbf{c}_i) \tag{36}$$

In fact, suppose by contradiction that this inequality (36) fails for some non-negative weight vector $\mathbf{w}$, namely $h_{\mathbf{w}}(\widehat{\mathbf{c}}_j) - \max_i h_{\mathbf{w}}(\mathbf{c}_i) > 0$ for some $j = 1, \dots, \widehat{m}$. This means in turn that $\lambda\{h_{\mathbf{w}}(\widehat{\mathbf{c}}_j) - \max_i h_{\mathbf{w}}(\mathbf{c}_i)\} > \log m$ for some $\lambda > 0$ large enough. Equivalently, $h_{\lambda\mathbf{w}}(\widehat{\mathbf{c}}_j) > \max_i h_{\lambda\mathbf{w}}(\mathbf{c}_i) + \log m$, because $\lambda h_{\mathbf{w}} = h_{\lambda\mathbf{w}}$. The original inequality (34) thus fails for the weight vector $\lambda\mathbf{w}$, providing the desired contradiction.

Step 3. The inequality (36) holds for every non-negative weight vector $\mathbf{w}$ if and only if for every $j = 1, \dots, \widehat{m}$, there exist m non-negative coefficients $\lambda_1, \dots, \lambda_m \geq 0$ that satisfy condition (37) and add up to 1, namely $\sum_{i=1}^{m} \lambda_i = 1$.

$$\widehat{\mathbf{c}}_j \geq \sum_{i=1}^{m} \lambda_i \mathbf{c}_i \tag{37}$$

In fact, the inequality (36) holds for every non-negative weight vector $\mathbf{w}$ if and only if for every $j = 1, \dots, \widehat{m}$, there exists no non-negative weight vector $\mathbf{w}$ such that $h_{\mathbf{w}}(\widehat{\mathbf{c}}_j)$ is strictly larger than $h_{\mathbf{w}}(\mathbf{c}_i)$ for every $i = 1, \dots, m$. Equivalently, the inequalities (38) are inconsistent: no non-negative weight vector $\mathbf{w}$ solves them all simultaneously.

$$h_{\mathbf{w}}(\widehat{\mathbf{c}}_j) > h_{\mathbf{w}}(\mathbf{c}_1), \quad \dots \quad h_{\mathbf{w}}(\widehat{\mathbf{c}}_j) > h_{\mathbf{w}}(\mathbf{c}_m) \tag{38}$$

We now recall that MOTZKIN'S TRANSPOSITION THEOREM (Bertsekas 2009, proposition 5.6.2) ensures that conditions (C1) and (C2) below are mutually exclusive (one and only one of them holds) for any two matrices $\mathbf{A} \in \mathbb{R}^{p \times n}$ and $\mathbf{B} \in \mathbb{R}^{q \times n}$.

(C1) There exists a vector $\mathbf{w} \in \mathbb{R}^n$ such that $\mathbf{A}\mathbf{w} < \mathbf{0}$ and $\mathbf{B}\mathbf{w} \leq \mathbf{0}$.

(C2) There exist two vectors $\boldsymbol{\xi} \in \mathbb{R}^q_+$ and $\boldsymbol{\mu} \in \mathbb{R}^p_+$ with $\boldsymbol{\mu} \neq \mathbf{0}$ such that $\mathbf{A}^\mathsf{T}\boldsymbol{\mu} + \mathbf{B}^\mathsf{T}\boldsymbol{\xi} = \mathbf{0}$.

Let $\mathbf{A}$ be the matrix whose $p = m$ rows are the vectors $(\widehat{\mathbf{c}}_j - \mathbf{c}_1)^\mathsf{T}, \dots, (\widehat{\mathbf{c}}_j - \mathbf{c}_m)^\mathsf{T}$. Let $\mathbf{B}$ be the matrix whose $q = n + 1$ rows are the vectors $-\mathbf{e}_1^\mathsf{T}, \dots, -\mathbf{e}_n^\mathsf{T}, \mathbf{0}$, where $\mathbf{0}$ has

components all equal to zero and $\mathbf{e}_i \in \mathbb{R}^n$ has components equal to 0 but for the ith component which is equal to 1. Conditions (C1) and (C2) thus become (C1′) and (C2′).

(C1′) There exists a non-negative vector $\mathbf{w} \in \mathbb{R}^n_+$ that solves all the inequalities in (38).

(C2′) There exist coefficients $\mu_1, \dots, \mu_m \geq 0$ that are all non-negative but not all equal to zero such that $\sum_{i=1}^m \mu_i(\mathbf{c}_i - \widehat{\mathbf{c}}_j)^\mathsf{T} \leq \mathbf{0}$.

The sum $\mu = \sum_{i=1}^m \mu_i$ is different from zero, because the coefficients $\mu_1, \dots, \mu_m$ are all non-negative but not all equal to zero. I can therefore divide both sides of the inequality $\sum_{i=1}^m \mu_i(\mathbf{c}_i - \widehat{\mathbf{c}}_j)^\mathsf{T} \leq \mathbf{0}$ by μ to obtain (37) with $\lambda_i = \mu_i/\mu$.

Step 4. Let us suppose that the implication $(\mathsf{x}, \mathsf{y}) \to (\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ is a universal of ME. Thus, the ME basic inequality (27) holds for every non-negative weight vector $\mathbf{w}$. By chaining steps 1-3, we conclude that every vector $\widehat{\mathbf{c}}_j$ satisfies the inequality (37), which is indeed the desired inequality (32) with the simpler notation $\mathbf{c}_i, \widehat{\mathbf{c}}_j$. □

A.10 Proof of theorem 3

Step (39a) holds because the vector $\mathbf{C}^{\widehat{\mathsf{y}}}(\widehat{\mathsf{x}}, \widehat{\mathsf{z}}_j)$ of discounted constraint violations of the additional consequent candidate $\widehat{\mathsf{z}}_j$ is the difference (25) between its vector $\mathbf{C}(\widehat{\mathsf{x}}, \widehat{\mathsf{z}}_j)$ of constraint violations minus the vector $\mathbf{C}(\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ of constraint violations of the consequent mapping $(\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$. Step (39b) holds because of the one-versus-convex-sum inequality (32). Step (39c) holds because the vector $\mathbf{C}^{\mathsf{y}}(\mathsf{x}, \mathsf{z}_i)$ of discounted constraint violations of the additional antecedent candidate z_i is the difference (25) between its vector $\mathbf{C}(\mathsf{x}, \mathsf{z}_i)$ of constraint violations minus the vector $\mathbf{C}(\mathsf{x}, \mathsf{y})$ of constraint violations of the antecedent mapping (x, y). Finally, step (39d) holds because the coefficients λ_i add up to 1.

$$\begin{aligned} \mathbf{C}(\widehat{\mathsf{x}}, \widehat{\mathsf{z}}_j) - \mathbf{C}(\widehat{\mathsf{x}}, \widehat{\mathsf{y}}) &\overset{(a)}{=} \mathbf{C}^{\widehat{\mathsf{y}}}(\widehat{\mathsf{x}}, \widehat{\mathsf{z}}_j) \\ &\overset{(b)}{\geq} \sum_{i=1}^m \lambda_i \mathbf{C}^{\mathsf{y}}(\mathsf{x}, \mathsf{z}_i) \qquad \text{with } \lambda_i \geq 0 \text{ and } \sum_{i=1}^m \lambda_i = 1 \\ &\overset{(c)}{=} \sum_{i=1}^m \lambda_i \mathbf{C}(\mathsf{x}, \mathsf{z}_i) - \sum_{i=1}^m \lambda_i \mathbf{C}(\mathsf{x}, \mathsf{y}) \overset{(d)}{=} \sum_{i=1}^m \lambda_i \mathbf{C}(\mathsf{x}, \mathsf{z}_i) - \mathbf{C}(\mathsf{x}, \mathsf{y}) \end{aligned} \tag{39}$$

By reordering (39), we obtain the restatement (40) of the one-versus-convex-sum inequality (32). It says that the ME implication $(\mathsf{x}, \mathsf{y}) \to (\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ fails in particular when the antecedent candidate y is too good relative to its competitors z_i, because $C(\mathsf{x}, \mathsf{y})$ is smaller than the convex sum $\sum_{i=1}^m \lambda_i C(\mathsf{x}, \mathsf{z}_i)$; and the consequent candidate $\widehat{\mathsf{y}}$ is too bad relative to an arbitrary competitor $\widehat{\mathsf{z}}_j$, because $C(\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ is larger than $C(\widehat{\mathsf{x}}, \widehat{\mathsf{z}}_j)$.

$$C(\widehat{\mathsf{x}}, \widehat{\mathsf{y}}) - C(\mathsf{x}, \mathsf{y}) \leq C(\widehat{\mathsf{x}}, \widehat{\mathsf{z}}_j) - \sum_{i=1}^m \lambda_i C(\mathsf{x}, \mathsf{z}_i) \qquad \text{with } \lambda_i \geq 0 \text{ and } \sum_{i=1}^m \lambda_i = 1 \tag{40}$$

Since the number $C(\mathsf{x}, \mathsf{z}_i)$ of constraint violations and the coefficients λ_i are all non-negative, this inequality (40) is not compromised by dropping the sum $\sum_{i=1}^m \lambda_i C(\mathsf{x}, \mathsf{z}_i)$ on its right-hand side. Furthermore, the candidate set of the consequent underlying form $\widehat{\mathsf{x}}$ contains at least one candidate that does *not* violate the constraint C. If this unoffending candidate is some additional consequent candidate $\widehat{\mathsf{z}}_j$, the term $C(\widehat{\mathsf{x}}, \widehat{\mathsf{z}}_j)$ is equal to zero and can be ignored in the right-hand side of the inequality (40),

which therefore becomes the desired harmonic bounding inequality $C(\widehat{\mathsf{x}},\widehat{\mathsf{y}}) \leq C(\mathsf{x},\mathsf{y})$. If instead the candidate that does not violate the constraint C is the consequent candidate $\widehat{\mathsf{y}}$, this harmonic bounding inequality holds trivially, because its left-hand side is equal to zero and its right-hand side is non-negative (constraint violations are non-negative).

A.11 Proof of theorem 5

Since we are looking at a markedness implication $(\mathsf{y},\mathsf{y}) \rightarrow (\widehat{\mathsf{y}},\widehat{\mathsf{y}})$ between two faithful mappings, each additional antecedent candidate z_i and each additional consequent candidate $\widehat{\mathsf{z}}_j$ is non-faithful. We thus assume that each of these non-faithful antecedent and consequent additional candidates is penalized by at least some faithfulness constraint in the constraint set used to define the ME typology.

We consider an additional consequent candidate $\widehat{\mathsf{z}}_j$ that is one-step-away in the direction of some faithfulness constraint F_0, in the sense that $F_0(\widehat{\mathsf{y}},\widehat{\mathsf{z}}_j) = 1$ and $F(\widehat{\mathsf{y}},\widehat{\mathsf{z}}_j) = 0$ for every faithfulness constraint F different from F_0. Appendix A.9 ensures that there exist non-negative coefficients λ_i that satisfy the one-versus-convex-sum inequality (40) for every constraint C and are normalized, namely $\sum_{i=1}^{m} \lambda_i = 1$.

Step 1. When C is a faithfulness constraint F different from F_0, this one-versus-convex-sum inequality (40) reduces to (41). In fact, $F(\widehat{\mathsf{x}},\widehat{\mathsf{z}}_j) = 0$, because the additional consequent candidate $\widehat{\mathsf{z}}_j$ is one-step-away in the direction of F_0 and thus violates no other faithfulness constraint F. Furthermore, $F(\mathsf{y},\mathsf{y}) = F(\widehat{\mathsf{y}},\widehat{\mathsf{y}}) = 0$, because the antecedent and consequent mappings are faithful.

$$0 \geq \sum_{i=1}^{m} \lambda_i F(\mathsf{y},\mathsf{z}_i) \tag{41}$$

This inequality (41) ensures that the coefficient λ_i is equal to zero whenever the corresponding mapping $(\mathsf{x},\mathsf{z}_i)$ violates a faithfulness constraint F other than F_0.

Step 2. When C is instead the faithfulness constraint F_0, the one-versus-convex-sum inequality (40) reduces to (42). In fact, $F_0(\widehat{\mathsf{x}},\widehat{\mathsf{z}}_j) = 1$, because the additional consequent candidate $\widehat{\mathsf{z}}_j$ is one-step-away in the direction of F_0. Furthermore, $F_0(\mathsf{y},\mathsf{y}) = F_0(\widehat{\mathsf{y}},\widehat{\mathsf{y}}) = 0$, because the antecedent and consequent mappings are faithful.

$$1 \geq \sum_{i=1}^{m} \lambda_i F_0(\mathsf{y},\mathsf{z}_i) \tag{42}$$

This inequality (42) ensures that the coefficient λ_i is equal to zero whenever the corresponding mapping $(\mathsf{x},\mathsf{z}_i)$ violates the faithfulness constraint F_0 more than once. In fact, let us assume by contradiction that some additional antecedent candidate $\mathsf{z}_{\widehat{\imath}}$ violates F_0 more than once and that its coefficient $\lambda_{\widehat{\imath}}$ is larger than zero, namely $F_0(\mathsf{y},\mathsf{z}_{\widehat{\imath}}) > 1$ and $\lambda_{\widehat{\imath}} > 0$. The reasoning in (43) then shows that (42) fails. Step (43a) holds because $\lambda_i \neq 0$ entails that z_i does not violate any faithfulness constraint different from F_0, as established in step 1. Thus, z_i must violate the faithfulness constraint F_0 at least once, under the assumption made here that every antecedent additional candidate z_i is non-faithful and therefore violates at least one faithfulness constraint. Step (43b) holds because of the hypotheses $F_0(\mathsf{y},\mathsf{z}_{\widehat{\imath}}) > 1$ and $\lambda_{\widehat{\imath}} > 0$. Finally, step (43c) holds because of

the hypothesis that the coefficients λ_i add up to 1.

$$\begin{aligned}\sum_{i=1}^{m} \lambda_i F_0(\mathsf{y}, \mathsf{z}_i) \;&=\; \lambda_{\widehat{\imath}} F_0(\mathsf{y}, \mathsf{z}_{\widehat{\imath}}) + \sum_{\substack{i=1 \\ i\neq\widehat{\imath}}}^{m} \lambda_i F_0(\mathsf{y}, \mathsf{z}_i) \;=\; \lambda_{\widehat{\imath}} F_0(\mathsf{y}, \mathsf{z}_{\widehat{\imath}}) + \sum_{\substack{i=1 \\ i\neq\widehat{\imath}, \lambda_i\neq 0}}^{m} \lambda_i F_0(\mathsf{y}, \mathsf{z}_i) \\ &\overset{(a)}{\geq} \lambda_{\widehat{\imath}} F_0(\mathsf{y}, \mathsf{z}_{\widehat{\imath}}) + \sum_{\substack{i=1 \\ i\neq\widehat{\imath}, \lambda_i\neq 0}}^{m} \lambda_i \;\overset{(b)}{>}\; \lambda_{\widehat{\imath}} + \sum_{\substack{i=1 \\ i\neq\widehat{\imath}, \lambda_i\neq 0}}^{m} \lambda_i \;=\; \lambda_{\widehat{\imath}} + \sum_{\substack{i=1 \\ i\neq\widehat{\imath}}}^{m} \lambda_i \;\overset{(c)}{=}\; 1\end{aligned} \tag{43}$$

Step 3. In conclusion, when the additional consequent candidate $\widehat{\mathsf{z}}_j$ is one-step-away in the direction of some faithfulness constraint F_0, the one-versus-convex-sum inequality (40) requires the coefficient λ_i to be equal to zero whenever the corresponding additional antecedent candidate z_i violates a faithfulness constraint F different from F_0 or it violates the faithfulness constraint F_0 more than once. Since these coefficients λ_i cannot be all equal to zero (because they must add up to one), there exists some additional antecedent candidate $\mathsf{z}_{\widehat{\imath}}$ that does not violate any faithfulness constraint F different from F_0 and does not violate F_0 more than once. Since $\mathsf{z}_{\widehat{\imath}}$ must violate some faithfulness constraint, we conclude that $\mathsf{z}_{\widehat{\imath}}$ violates F_0 exactly once. In other words, $\mathsf{z}_{\widehat{\imath}}$ is one-step-away in the direction of F_0.

A.12 Proof of theorem 6

Since $M(\mathsf{y}) = M(\widehat{\mathsf{y}})$, the one-versus-convex-sum inequality (40) for this markedness constraint M boils down to (44).

$$M(\widehat{\mathsf{z}}_j) \geq \sum_{i=1}^{m} \lambda_i M(\mathsf{z}_i) \tag{44}$$

If the consequent additional candidate $\widehat{\mathsf{z}}_j$ does not violate this markedness constraint M, the left-hand side of (44) is equal to zero. The coefficient λ_i on the right-hand side must therefore be equal to zero whenever the corresponding antecedent additional candidate z_i violates M. Furthermore, appendix A.11 has shown that, when the consequent additional candidate $\widehat{\mathsf{z}}_j$ is one-step-away in the direction of some faithfulness constraint F_0, the coefficient λ_i can be different from zero only if the corresponding additional antecedent candidate z_i is one-step-away in the same direction F_0. Since the coefficients λ_i cannot be all equal to zero (because they must add up to one), we conclude that there exists some additional antecedent candidate $\mathsf{z}_{\widehat{\imath}}$ that does not violate the markedness constraint M and is only one-step-away in the direction of F_0.

A.13 ME sum-versus-sum lemma

The inequality (45) compares the sums of vectors of antecedent and consequent discounted constraint violations. The ME average faithfulness generalization will be derived from this sum-versus-sum inequality.

Lemma 3
Suppose that an implication $(\mathsf{x}, \mathsf{y}) \rightarrow (\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ *is a ME universal and that the antecedent mapping*

(x, y) and the consequent mapping $(\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ have the same number m of (possibly different) additional candidates. Then, the sum of the vectors of discounted constraint violations of the additional consequent candidates $\widehat{\mathsf{z}}_1, \ldots, \widehat{\mathsf{z}}_m$ is at least as large as the sum of the vectors of discounted constraint violations of the additional antecedent candidates $\mathsf{z}_1, \ldots, \mathsf{z}_m$, as in (45).

$$\sum_{j=1}^{m} \mathbf{C}^{\widehat{\mathsf{y}}}(\widehat{\mathsf{x}}, \widehat{\mathsf{z}}_j) \geq \sum_{i=1}^{m} \mathbf{C}^{\mathsf{y}}(\mathsf{x}, \mathsf{z}_i) \tag{45}$$

Proof. We consider a weight vector $\mathbf{w}$ that assigns some non-negative weight $w \geq 0$ to some constraint C and zero weights to the other constraints. Thus $h_{\mathbf{w}}(\mathbf{C}^{\mathsf{y}}(\mathsf{x}, \mathsf{z}_i)) = -wC^{\mathsf{y}}(\mathsf{x}, \mathsf{z}_i)$ and $h_{\mathbf{w}}(\mathbf{C}^{\widehat{\mathsf{y}}}(\widehat{\mathsf{x}}, \widehat{\mathsf{z}}_j)) = -wC^{\widehat{\mathsf{y}}}(\widehat{\mathsf{x}}, \widehat{\mathsf{z}}_j)$. The ME basic inequality (27) for this weight vector $\mathbf{w}$ thus reduces to (46).

$$\sum_{j=1}^{m} \exp\big\{ -wC(\widehat{\mathsf{x}}, \widehat{\mathsf{z}}_j)\big\} \leq \sum_{i=1}^{m} \exp\big\{ -wC(\mathsf{x}, \mathsf{z}_i)\big\} \tag{46}$$

Equivalently, the function $f(w) = \sum_{i=1}^{m} \exp\{-wC^{\mathsf{y}}(\mathsf{x}, \mathsf{z}_i)\} - \sum_{j=1}^{m} \exp\{-wC^{\widehat{\mathsf{y}}}(\widehat{\mathsf{x}}, \widehat{\mathsf{z}}_j)\}$ must satisfy the inequality $f(w) \geq 0$ for every $w \geq 0$. Because of the assumption that the antecedent and the consequent mappings have the same number m of additional candidates, $f(w = 0) = 0$. The function f therefore cannot decrease when we move away from $w = 0$. Equivalently, its derivative at $w = 0$ must be non-negative, namely $f'(w = 0) \geq 0$. This derivative is $f'(w) = -\sum_{i=1}^{m} C^{\mathsf{y}}(\mathsf{x}, \mathsf{z}_i) \exp\{-wC^{\mathsf{y}}(\mathsf{x}, \mathsf{z}_i)\} + \sum_{j=1}^{m} C^{\widehat{\mathsf{y}}}(\widehat{\mathsf{x}}, \widehat{\mathsf{z}}_j) \exp\{-wC^{\widehat{\mathsf{y}}}(\widehat{\mathsf{x}}, \widehat{\mathsf{z}}_j)\}$. The condition $f'(w = 0) \geq 0$ thus becomes $\sum_{i=1}^{m} C^{\widehat{\mathsf{y}}}(\widehat{\mathsf{x}}, \widehat{\mathsf{z}}_j) \geq \sum_{i=1}^{m} C^{\mathsf{y}}(\mathsf{x}, \mathsf{z}_i)$, which is indeed the desired sum-versus-sum inequality (45) for the constraint C considered. □

A.14 Proof of theorem 7

Let us consider an arbitrary ME universal implication $(\mathsf{x}, \mathsf{y}) \rightarrow (\mathsf{x}, \mathsf{y})$ whose antecedent and consequent underlying forms x and $\widehat{\mathsf{x}}$ come with the same number m of additional candidates $\mathsf{z}_1, \ldots, \mathsf{z}_m$ and $\widehat{\mathsf{z}}_1, \ldots, \widehat{\mathsf{z}}_m$. We recall from (25) that $C^{\mathsf{y}}(\mathsf{x}, \mathsf{z}_i)$ is the difference between the number of violations $C(\mathsf{x}, \mathsf{z}_i)$ of z_i minus the number of violations $C(\mathsf{x}, \mathsf{y})$ of the antecedent mapping. And $C^{\widehat{\mathsf{y}}}(\widehat{\mathsf{x}}, \widehat{\mathsf{z}}_j)$ is defined analogously. The sum-versus-sum inequality (45) can therefore be rewritten as in (47).

$$mC(\widehat{\mathsf{x}}, \widehat{\mathsf{y}}) - mC(\mathsf{x}, \mathsf{y}) \leq \sum_{j=1}^{m} C(\widehat{\mathsf{x}}, \widehat{\mathsf{z}}_j) - \sum_{i=1}^{m} C(\mathsf{x}, \mathsf{z}_i) \tag{47}$$

By adding the quantity $C(\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ to both sides of this inequality (47), by subtracting the quantity $C(\mathsf{x}, \mathsf{y})$ from both sides, and by dividing both sides by $m + 1$, we finally obtain the inequality $C(\widehat{\mathsf{x}}, \widehat{\mathsf{y}}) - C(\mathsf{x}, \mathsf{y}) \leq \overline{C}(\widehat{\mathsf{x}}) - \overline{C}(\mathsf{x})$, as shown in (48).

$$C(\widehat{\mathsf{x}}, \widehat{\mathsf{y}}) - C(\mathsf{x}, \mathsf{y}) \leq \underbrace{\frac{\sum_{j=1}^{m} C(\widehat{\mathsf{x}}, \widehat{\mathsf{z}}_j) + C(\widehat{\mathsf{x}}, \widehat{\mathsf{y}})}{m+1}}_{\overline{C}(\widehat{\mathsf{x}})} - \underbrace{\frac{\sum_{i=1}^{m} C(\mathsf{x}, \mathsf{z}_i) + C(\mathsf{x}, \mathsf{y})}{m+1}}_{\overline{C}(\mathsf{x})} \tag{48}$$

In a markedness implication $(\mathsf{x},\mathsf{x}) \to (\widehat{\mathsf{x}},\widehat{\mathsf{x}})$, the underlying and surface forms in the antecedent and consequent mappings coincide, namely $\mathsf{x}=\mathsf{y}$ and $\widehat{\mathsf{x}}=\widehat{\mathsf{y}}$. When C is a faithfulness constraint F, the inequality $C(\widehat{\mathsf{x}},\widehat{\mathsf{y}}) - C(\mathsf{x},\mathsf{y}) \leq \overline{C}(\widehat{\mathsf{x}}) - \overline{C}(\mathsf{x})$ just obtained thus reduces to the average faithfulness inequality $\overline{F}(\widehat{\mathsf{x}}) \geq \overline{F}(\mathsf{x})$, because the antecedent and consequent mappings do not violate $C=F$, namely $F(\widehat{\mathsf{x}},\widehat{\mathsf{y}}) = F(\mathsf{x},\mathsf{y}) = 0$.

A.15 ME pruning lemma

The following lemma explains how to prune the sets Gen and $\mathbf{C}$ of mappings and constraints down to subsets $Gen_2 \subseteq Gen$ and $\mathbf{C}_2 \subseteq \mathbf{C}$ without compromising ME universals. The cardinality generalization will be derived in appendix 5 from this lemma.

Lemma 4
We split the constraint set $\mathbf{C}$ *into two halves* $\mathbf{C}_1$ *and* $\mathbf{C}_2$*. We consider an assignment* $\mathbf{w}_1$ *of non-negative weights to the constraints in* $\mathbf{C}_1$*. We define* Gen_2 *as the subset (49) of those mappings* (x,y) *from* Gen *such that* y *achieves the largest ME harmony within the candidate set* $Gen(\mathsf{x})$ *relatively to the constraints in* $\mathbf{C}_1$ *and the weights assigned to them by* $\mathbf{w}_1$*.*

$$Gen_2 = \left\{ (\mathsf{x},\mathsf{y}) \in Gen \;\middle|\; h_{\mathbf{w}_1 \mathbf{C}_1}(\mathsf{x},\mathsf{y}) = \max_{\mathsf{u} \in Gen(\mathsf{x})} h_{\mathbf{w}_1 \mathbf{C}_1}(\mathsf{x},\mathsf{u}) \right\} \tag{49}$$

If the implication $(\mathsf{x},\mathsf{y}) \to (\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ *is a universal of the ME typology corresponding to the original sets* Gen *and* $\mathbf{C}$ *of mappings and constraints and if the antecedent mapping* (x,y) *belongs to the subset of mappings* Gen_2*, then the consequent mapping* $(\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ *belongs to* Gen_2 *as well and the implication* $(\mathsf{x},\mathsf{y}) \to (\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ *is also a universal of the ME typology corresponding to the subsets* Gen_2 *and* $\mathbf{C}_2$ *of mappings and constraints.*

Proof. For every assignment $\mathbf{w}_2$ of non-negative weights to the constraints in $\mathbf{C}_2$, we denote by $(\lambda\mathbf{w}_1, \mathbf{w}_2)$ the weight vector for the original constraint set $\mathbf{C} = \mathbf{C}_1 \cup \mathbf{C}_2$ that assigns to any constraint in $\mathbf{C}_1$ the weight prescribed by $\mathbf{w}_1$ rescaled by λ and assigns to any constraint in $\mathbf{C}_2$ the weight prescribed by $\mathbf{w}_2$.

Step 1. The identity (50) holds for any mapping (x,y) in Gen_2. It says that its ME probability relative to the original sets Gen and $\mathbf{C}$ of mappings and constraints and the weight vector $(\lambda\mathbf{w}_1, \mathbf{w}_2)$ converges as λ grows to the ME probability relative to the subsets Gen_2 and $\mathbf{C}_2$ of mappings and constraints and the weight vector $\mathbf{w}_2$.

$$\lim_{\lambda \to \infty} G^{\mathrm{ME},Gen,\mathbf{C}}_{\lambda\mathbf{w}_1,\mathbf{w}_2}(\mathsf{y}\,|\,\mathsf{x}) = G^{\mathrm{ME},Gen_2,\mathbf{C}_2}_{\mathbf{w}_2}(\mathsf{y}\,|\,\mathsf{x}) \tag{50}$$

In fact, steps (51a) and (51f) hold because of the expression of ME probabilities in terms of discounted constraint violations obtained in (28). Step (51b) holds because $h_{\lambda\mathbf{w}_1,\mathbf{w}_2}\mathbf{C}^{\mathsf{y}}(\mathsf{x},\mathsf{z}_i)$ is equal to $\lambda h_{\mathbf{w}_1}\mathbf{C}_1^{\mathsf{y}}(\mathsf{x},\mathsf{z}_i) + h_{\mathbf{w}_2}\mathbf{C}_2^{\mathsf{y}}(\mathsf{x},\mathsf{z}_i)$. Step (51c) holds by splitting the additional antecedent candidates $\mathsf{z}_1,\ldots,\mathsf{z}_m$ into those that survive into $Gen_2(\mathsf{x})$ and those that do not. Step (51d) holds because (*) is equal to one because $h_{\mathbf{w}_1}\mathbf{C}_1^{\mathsf{y}}(\mathsf{x},\mathsf{z}_i)$ is equal to zero for every z_i from $Gen_2(\mathsf{x})$, given that both z_i and y achieve the largest ME harmony. Finally, step (51e) holds because (**) converges to zero, because $h_{\mathbf{w}_1}\mathbf{C}_1^{\mathsf{y}}(\mathsf{x},\mathsf{z}_i)$

is negative for every z_i not in $Gen_2(\mathsf{x})$, given that z_i has a smaller ME harmony than y.

$$\begin{aligned}
&\lim_{\lambda\to\infty} G^{\mathrm{ME},Gen,\mathbf{C}}_{\lambda\mathbf{w}_1,\mathbf{w}_2}(\mathsf{y}\,|\,\mathsf{x}) = \\
&\overset{(a)}{=} \lim_{\lambda\to\infty} \frac{1}{1+\sum_{i=1}^{m} \exp h_{\lambda\mathbf{w}_1,\mathbf{w}_2}\mathbf{C}^{\mathsf{y}}(\mathsf{x},\mathsf{z}_i)} \overset{(b)}{=} \lim_{\lambda\to\infty} \frac{1}{1+\sum_{i=1}^{m} \exp \lambda h_{\mathbf{w}_1}\mathbf{C}_1^{\mathsf{y}}(\mathsf{x},\mathsf{z}_i) \exp h_{\mathbf{w}_2}\mathbf{C}_2^{\mathsf{y}}(\mathsf{x},\mathsf{z}_i)} \\
&\overset{(c)}{=} \lim_{\lambda\to\infty} \frac{1}{1+\sum_{\mathsf{z}_i\notin Gen_2(\mathsf{x})} \exp \lambda h_{\mathbf{w}_1}\mathbf{C}_1^{\mathsf{y}}(\mathsf{x},\mathsf{z}_i) \exp h_{\mathbf{w}_2}\mathbf{C}_2^{\mathsf{y}}(\mathsf{x},\mathsf{z}_i) + \sum_{\mathsf{z}_i\in Gen_2(\mathsf{x})} \underbrace{\exp \lambda h_{\mathbf{w}_1}\mathbf{C}_1^{\mathsf{y}}(\mathsf{x},\mathsf{z}_i)}_{(*)} \exp h_{\mathbf{w}_2}\mathbf{C}_2^{\mathsf{y}}(\mathsf{x},\mathsf{z}_i)} \\
&\overset{(d)}{=} \lim_{\lambda\to\infty} \frac{1}{1+\sum_{\mathsf{z}_i\notin Gen_2(\mathsf{x})} \underbrace{\exp \lambda h_{\mathbf{w}_1}\mathbf{C}_1^{\mathsf{y}}(\mathsf{x},\mathsf{z}_i)}_{(**)} \exp h_{\mathbf{w}_2}\mathbf{C}_2^{\mathsf{y}}(\mathsf{x},\mathsf{z}_i) + \sum_{\mathsf{z}_i\in Gen_2(\mathsf{x})} \exp h_{\mathbf{w}_2}\mathbf{C}_2^{\mathsf{y}}(\mathsf{x},\mathsf{z}_i)} \\
&\overset{(e)}{=} \frac{1}{1+\sum_{\mathsf{z}_i\in Gen_2(\mathsf{x})} \exp h_{\mathbf{w}_2}\mathbf{C}_2^{\mathsf{y}}(\mathsf{x},\mathsf{z}_i)} \overset{(f)}{=} G^{\mathrm{ME},Gen_2,\mathbf{C}_2}_{\mathbf{w}_2}(\mathsf{y}\,|\,\mathsf{x})
\end{aligned} \tag{51}$$

Step 2. The identity (52) holds for any mapping (x,y) that does not belong to Gen_2. It says that its ME probability relative to the original sets Gen and $\mathbf{C}$ of mappings and constraints and the weight vector $(\lambda\mathbf{w}_1,\mathbf{w}_2)$ converges to zero as λ grows to infinity.

$$\lim_{\lambda\to\infty} G^{\mathrm{ME},Gen,\mathbf{C}}_{\lambda\mathbf{w}_1,\mathbf{w}_2}(\mathsf{y}\,|\,\mathsf{x}) = 0 \tag{52}$$

In fact, step (53a) holds because the ME probabilities $G^{Gen,\mathbf{C}}_{\lambda\mathbf{w}_1,\mathbf{w}_2}(\cdot\,|\,\mathsf{x})$ are normalized: the ME probability of y is equal to 1 minus the sum of the ME probabilities of the additional candidates $\mathsf{z}_1,\ldots,\mathsf{z}_m$ of x. Step (53b) holds because $Gen_2(\mathsf{x})$ does not contain y by hypothesis. Step (53c) holds because we have just established the limit identity (50) for the mappings in Gen_2. Finally, step (53d) holds because the ME probabilities $G^{\mathrm{ME},Gen_2,\mathbf{C}_2}_{\mathbf{w}_2}(\cdot\,|\,\mathsf{x})$ are normalized, whereby their sum over $Gen_2(\mathsf{x})$ is equal to 1.

$$\begin{aligned}
\lim_{\lambda\to\infty} G^{\mathrm{ME},Gen,\mathbf{C}}_{\lambda\mathbf{w}_1,\mathbf{w}_2}(\mathsf{y}\,|\,\mathsf{x}) &\overset{(a)}{=} 1 - \lim_{\lambda\to\infty} \sum_{i=1}^{m} G^{\mathrm{ME},Gen,\mathbf{C}}_{\lambda\mathbf{w}_1,\mathbf{w}_2}(\mathsf{z}_i\,|\,\mathsf{x}) \\
&\overset{(b)}{\leq} 1 - \lim_{\lambda\to\infty} \sum_{\mathsf{u}\in Gen_2(\mathsf{x})} G^{\mathrm{ME},Gen,\mathbf{C}}_{\lambda\mathbf{w}_1,\mathbf{w}_2}(\mathsf{u}\,|\,\mathsf{x}) \\
&\overset{(c)}{=} 1 - \sum_{\mathsf{u}\in Gen_2(\mathsf{x})} G^{\mathrm{ME},Gen_2,\mathbf{C}_2}_{\mathbf{w}_2}(\mathsf{u}\,|\,\mathsf{x}) \overset{(d)}{=} 0
\end{aligned} \tag{53}$$

Step 3. Let us suppose that the implication $(\mathsf{x},\mathsf{y}) \to (\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ is a universal of the ME typology corresponding to the original sets Gen and $\mathbf{C}$ of mappings and constraints. By condition (6), this means that the ME probability $G^{\mathrm{ME},Gen,\mathbf{C}}_{\mathbf{w}}(\mathsf{y}\,|\,\mathsf{x})$ of the antecedent mapping is smaller than or equal to the ME probability $G^{\mathrm{ME},Gen,\mathbf{C}}_{\mathbf{w}}(\widehat{\mathsf{y}}\,|\,\widehat{\mathsf{x}})$ of the conse-

quent mapping as stated in (54), for any vector $\mathbf{w}$ of weights for the constraints in $\mathbf{C}$.

$$G_{\mathbf{w}}^{\mathrm{ME},Gen,\mathbf{C}}(\mathsf{y}\,|\,\mathsf{x}) \leq G_{\mathbf{w}}^{\mathrm{ME},Gen,\mathbf{C}}(\widehat{\mathsf{y}}\,|\,\widehat{\mathsf{x}}) \tag{54}$$

Since this inequality (54) holds for every weight vector, then it holds in particular for the weight vector $(\lambda\mathbf{w}_1, \mathbf{w}_2)$, no matter the choice of the coefficient λ. Hence, it also holds for the limit as λ goes to infinity, yielding (55).

$$\lim_{\lambda\to\infty} G_{\lambda\mathbf{w}_1,\mathbf{w}_2}^{\mathrm{ME},Gen,\mathbf{C}}(\mathsf{y}\,|\,\mathsf{x}) \leq \lim_{\lambda\to\infty} G_{\lambda\mathbf{w}_1,\mathbf{w}_2}^{\mathrm{ME},Gen,\mathbf{C}}(\widehat{\mathsf{y}}\,|\,\widehat{\mathsf{x}}) \tag{55}$$

Since the antecedent mapping (x,y) belongs to the subset of mappings Gen_2 by hypothesis, the limit identity (50) ensures that the limit on the left-hand side of the inequality (55) is equal to the ME probability $G_{\mathbf{w}_2}^{\mathrm{ME},Gen_2,\mathbf{C}_2}(\mathsf{y}\,|\,\mathsf{x})$ of the antecedent mapping (x,y) relative to the subsets Gen_2 and $\mathbf{C}_2$ of mappings and constraints. The consequent mapping $(\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ must then belong to Gen_2 as well. In fact, let us assume by contradiction that it does not. The limit identity (52) would then ensure that the limit on the right-hand side of the inequality (55) is equal to zero. In other words, this inequality (55) would become $G_{\mathbf{w}_2}^{\mathrm{ME},Gen_2,\mathbf{C}_2}(\mathsf{y}\,|\,\mathsf{x}) \leq 0$, contradicting the strict positivity of ME probabilities.

Since also the consequent mapping $(\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ belongs to Gen_2, the limit identity (50) ensures that the limit on the right-hand side of the inequality (55) is equal to the ME probability $G_{\mathbf{w}_2}^{\mathrm{ME},Gen_2,\mathbf{C}_2}(\widehat{\mathsf{y}}\,|\,\widehat{\mathsf{x}})$ of the consequent mapping $(\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ relative to the subsets Gen_2 and $\mathbf{C}_2$ of mappings and constraints, whereby (55) becomes (56).

$$G_{\mathbf{w}_2}^{\mathrm{ME},Gen_2,\mathbf{C}_2}(\mathsf{y}\,|\,\mathsf{x}) \leq G_{\mathbf{w}_2}^{\mathrm{ME},Gen_2,\mathbf{C}_2}(\widehat{\mathsf{y}}\,|\,\widehat{\mathsf{x}}) \tag{56}$$

Since this inequality (56) holds for any assignment $\mathbf{w}_2$ of weights to the constraints in $\mathbf{C}_2$, we conclude that the implication $(\mathsf{x},\mathsf{y}) \to (\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ is a universal of the ME typology corresponding to the subsets Gen_2 and $\mathbf{C}_2$ of mappings and constraints. □

A.16 Proof of theorem 4

With the notation of appendix A.15, let $S = \mathbf{C}_1$ and consider a weight vector $\mathbf{w}_1$ with strictly positive components. The inclusion $Gen_S \subseteq Gen_2$ then holds. In fact, a mapping belongs to Gen_S provided it does not violate any constraint in $S = \mathbf{C}_1$. Its ME harmony relative to the constraints in $\mathbf{C}_1$ and the weights in $\mathbf{w}_1$ is thus equal to one, which is the largest value the ME harmony can achieve (because both the weights and the constraints are non-negative). The mapping considered therefore belongs to the subset Gen_2 in (49).

Because of this inclusion $Gen_S \subseteq Gen_2$, the assumption that (x,y) belongs to Gen_S entails that (x,y) belongs to Gen_2. The pruning lemma of appendix A.15 then ensures that, since the implication $(\mathsf{x},\mathsf{y}) \to (\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ is a universal of the ME typology corresponding to Gen and $\mathbf{C}$, it is also a universal of the ME typology corresponding to Gen_2 and $\mathbf{C}_2$. The inequality (57) thus holds uniformly for any non-negative weight vector $\mathbf{w}_2$ for the constraint subset $\mathbf{C}_2$, as it is the ME basic inequality of appendix A.7, stated for Gen_2, $\mathbf{C}_2$ rather than Gen, $\mathbf{C}$.

$$\sum_{\substack{\mathsf{z}\in Gen_2(\mathsf{x})\\ \mathsf{z}\neq\mathsf{y}}} \exp h_{\mathbf{w}_2}(\mathbf{C}_2^{\mathsf{y}}(\mathsf{x},\mathsf{z})) \geq \sum_{\substack{\widehat{\mathsf{z}}\in Gen_2(\widehat{\mathsf{x}})\\ \widehat{\mathsf{z}}\neq\widehat{\mathsf{y}}}} \exp h_{\mathbf{w}_2}(\mathbf{C}_2^{\widehat{\mathsf{y}}}(\widehat{\mathsf{x}},\widehat{\mathsf{z}})) \tag{57}$$

Since this inequality (57) holds for any non-negative weight vector $\mathbf{w}_2$, then it holds in particular when $\mathbf{w}_2$ has components all equal (or close) to zero. In this case, the inequality (57) reduces to the inequality (58) between the cardinalities of the candidate sets of the antecedent and consequent underlying forms x and $\widehat{\mathsf{x}}$ relative to Gen_2.

$$\left|Gen_2(\mathsf{x})\right| \geq \left|Gen_2(\widehat{\mathsf{x}})\right| \tag{58}$$

The reasoning in (59) finally establishes the extended cardinality inequality. Step (59a) holds because of the inclusion $Gen_S(\mathsf{x}) \supseteq Gen_2(\mathsf{x})$. In fact, let us assume by contradiction that there exists a candidate y_0 that belongs to $Gen_2(\mathsf{x})$ but not to $Gen_S(\mathsf{x})$. Since (x, y) has ME harmony equal to 1 (as noted above), then also $(\mathsf{x}, \mathsf{y}_0)$ has ME harmony equal to 1 (because of the assumption that y_0 belongs to $Gen_2(\mathsf{x})$ and thus has the largest ME harmony). On the other hand, since $\mathbf{w}_1$ has strictly positive components and since y_0 violates at least one constraint in $\mathbf{C}_1$ (because of the assumption that y_0 does not belong to $Gen_S(\mathsf{x})$ with $S = \mathbf{C}_1$), then $(\mathsf{x}, \mathsf{y}_0)$ cannot have ME harmony equal to 1, yielding the desired contradiction. Step (59b) holds because of the inequality (58) established above. Finally, step (59c) holds because of the inclusion $Gen_S \subseteq Gen_2$ established at the beginning of this appendix.

$$\left|Gen_S(\mathsf{x})\right| \overset{(a)}{\geq} \left|Gen_2(\mathsf{x})\right| \overset{(b)}{\geq} \left|Gen_2(\widehat{\mathsf{x}})\right| \overset{(c)}{\geq} \left|Gen_S(\widehat{\mathsf{x}})\right| \tag{59}$$

A.17 Proof of theorem 8

The proof relies on the intermediate technical condition (60). The numerators on the left- and right-hand sides feature the harmony of the same vector $\mathbf{a}$ with components $a_1, \ldots, a_n$ but with the kth component replaced with a value x on the left-hand side and with a value y on the right-hand side. Analogously, the denominators on the left- and right-hand sides feature the harmony of the same vector $\mathbf{b}$ with components $b_1, \ldots, b_n$ but with the kth component replaced again with the value x on the left-hand side and with the value y on the right-hand side.

$$\frac{H(a_1, \ldots, a_{k-1}, \boldsymbol{x}, a_{k+1}, \ldots, a_n)}{H(b_1, \ldots, b_{k-1}, \boldsymbol{x}, b_{k+1}, \ldots, b_n)} = \frac{H(a_1, \ldots, a_{k-1}, \boldsymbol{y}, a_{k+1}, \ldots, a_n)}{H(b_1, \ldots, b_{k-1}, \boldsymbol{y}, b_{k+1}, \ldots, b_n)} \tag{60}$$

Step 1. We show that, if the probabilistic grammar G_H makes no spurious distinctions, the harmony H satisfies the technical condition (60) for any index $k = 1, \ldots, n$ and any values a_i, b_j, x, y. To this end, we consider two mappings (x, y) and $(\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$. We assume that they come with only one additional candidate each, denoted z and $\widehat{\mathsf{z}}$ respectively. And we define the constraint violation vectors of these four candidates as the four tuples of values (61) at the numerators and denominators of the two sides of condition (60).

$$\begin{array}{ll} \mathbf{C}(\mathsf{x}, \mathsf{z}) = (a_1, \ldots, a_{k-1}, x, a_{k+1}, \ldots, a_n) & \mathbf{C}(\widehat{\mathsf{x}}, \widehat{\mathsf{z}}) = (a_1, \ldots, a_{k-1}, y, a_{k+1}, \ldots, a_n) \\ \mathbf{C}(\mathsf{x}, \mathsf{y}) = (b_1, \ldots, b_{k-1}, x, b_{k+1}, \ldots, b_n) & \mathbf{C}(\widehat{\mathsf{x}}, \widehat{\mathsf{y}}) = (b_1, \ldots, b_{k-1}, y, b_{k+1}, \ldots, b_n) \end{array} \tag{61}$$

Constraint C_k does not make distinctions within the candidate sets, namely satisfies the two vertical identities in (7). In fact, C_k does not distinguish between the candidates y and z, because it assigns them the same number x of violations. Analogously, C_k does not distinguish between the candidates $\widehat{\mathsf{y}}$ and $\widehat{\mathsf{z}}$, because it assigns them the same number y of violations. Every other constraint C_h with $h \neq k$ does not make distinctions

across candidate sets, namely satisfies the horizontal identities in (7). In fact, C_h does not distinguish between the candidates z and $\widehat{\mathsf{z}}$, because it assigns them the same number a_h of violations. Furthermore, C_h does not distinguish between the candidates y and $\widehat{\mathsf{y}}$, because it assigns them the same number b_h of violations.

The assumption that G_H makes no spurious distinctions thus ensures that the mappings (x, y) and $(\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ share the same probability $G_H(\mathsf{y} \,|\, \mathsf{x}) = G_H(\widehat{\mathsf{y}} \,|\, \widehat{\mathsf{y}})$. This probability identity entails in turn the desired identity (60), as shown in (62). Step (62a) holds because of the definition (1) of the grammar G_H induced by the harmony H. Step (62b) holds because the normalization constant Z is the sum of the harmonies of the two candidates of each underlying form. Step (62c) holds by dividing the numerator and denominator of the left-hand side by $H(\mathbf{C}(\mathsf{x}, \mathsf{y}))$ and furthermore the numerator and denominator of the right-hand side by $H(\mathbf{C}(\widehat{\mathsf{x}}, \widehat{\mathsf{y}}))$. This step is licit because of the assumption that harmony score are strictly positive. Step (62d) holds because of the definition (61) of the constraint violation vectors.

$$
\begin{aligned}
G_H(\mathsf{y} \,|\, \mathsf{x}) = G_H(\widehat{\mathsf{y}} \,|\, \widehat{\mathsf{x}}) \quad &\overset{(a)}{\iff} \quad \frac{H(\mathbf{C}(\mathsf{x}, \mathsf{y}))}{Z(\mathsf{x})} = \frac{H(\mathbf{C}(\widehat{\mathsf{x}}, \widehat{\mathsf{y}}))}{Z(\widehat{\mathsf{x}})} \\
&\overset{(b)}{\iff} \quad \frac{H(\mathbf{C}(\mathsf{x}, \mathsf{y}))}{H(\mathbf{C}(\mathsf{x}, \mathsf{y})) + H(\mathbf{C}(\mathsf{x}, \mathsf{z}))} = \frac{H(\mathbf{C}(\widehat{\mathsf{x}}, \widehat{\mathsf{y}}))}{H(\mathbf{C}(\widehat{\mathsf{x}}, \widehat{\mathsf{y}})) + H(\mathbf{C}(\widehat{\mathsf{x}}, \widehat{\mathsf{z}}))} \\
&\overset{(c)}{\iff} \quad \frac{1}{1 + \dfrac{H(\mathbf{C}(\mathsf{x}, \mathsf{z}))}{H(\mathbf{C}(\mathsf{x}, \mathsf{y}))}} = \frac{1}{1 + \dfrac{H(\mathbf{C}(\widehat{\mathsf{x}}, \widehat{\mathsf{z}}))}{H(\mathbf{C}(\widehat{\mathsf{x}}, \widehat{\mathsf{y}}))}} \quad \iff \quad \frac{H(\mathbf{C}(\mathsf{x}, \mathsf{z}))}{H(\mathbf{C}(\mathsf{x}, \mathsf{y}))} = \frac{H(\mathbf{C}(\widehat{\mathsf{x}}, \widehat{\mathsf{z}}))}{H(\mathbf{C}(\widehat{\mathsf{x}}, \widehat{\mathsf{y}}))} \\
&\overset{(d)}{\iff} \quad \frac{H(a_1, \ldots, a_{k-1}, x, a_{k+1}, \ldots, a_n)}{H(b_1, \ldots, b_{k-1}, x, b_{k+1}, \ldots, b_n)} = \frac{H(a_1, \ldots, a_{k-1}, y, a_{k+1}, \ldots, a_n)}{H(b_1, \ldots, b_{k-1}, y, b_{k+1}, \ldots, b_n)}
\end{aligned}
\tag{62}
$$

Step 2. We now show that, vice versa, if the harmony H satisfies the technical condition (60) for any index $k = 1, \ldots, n$ and any values a_i, b_j, x, y, the corresponding grammar G_H makes no spurious distinctions. To this end, we consider two mappings (x, y) and $(\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$. We assume that they come with only one additional candidate each, denoted z and $\widehat{\mathsf{z}}$ respectively. And we suppose next that every constraint C in the constraint set $\mathbf{C}$ satisfies either the two horizontal or the two vertical identities in (7). Without loss of generality, we assume that the constraints $C_1, \ldots, C_m$ satisfy the horizontal identities while the remaining constraints $C_{m+1}, \ldots, C_n$ satisfy the vertical identities, as in (63).

$$
\begin{array}{cccccc}
\overbrace{\underbrace{C_1(\mathsf{x}, \mathsf{y}) = C_1(\widehat{\mathsf{x}}, \widehat{\mathsf{y}})}_{a_1}}^{b_1} & & & \overbrace{\underbrace{C_m(\mathsf{x}, \mathsf{y}) = C_m(\widehat{\mathsf{x}}, \widehat{\mathsf{y}})}_{a_m}}^{b_m} & & \\
\underbrace{C_1(\mathsf{x}, \mathsf{z}) = C_1(\widehat{\mathsf{x}}, \widehat{\mathsf{z}})}_{a_1} & \cdots & & \underbrace{C_m(\mathsf{x}, \mathsf{z}) = C_m(\widehat{\mathsf{x}}, \widehat{\mathsf{z}})}_{a_m} & & \\
C_{m+1}(\mathsf{x}, \mathsf{y}) & C_{m+1}(\widehat{\mathsf{x}}, \widehat{\mathsf{y}}) & & C_n(\mathsf{x}, \mathsf{y}) & C_n(\widehat{\mathsf{x}}, \widehat{\mathsf{y}}) \\
\| & \| & \cdots & \| & \| \\
\underbrace{C_{m+1}(\mathsf{x}, \mathsf{z})}_{a_{m+1}} & \underbrace{C_{m+1}(\widehat{\mathsf{x}}, \widehat{\mathsf{z}})}_{b_{m+1}} & & \underbrace{C_n(\mathsf{x}, \mathsf{z})}_{a_n} & \underbrace{C_n(\widehat{\mathsf{x}}, \widehat{\mathsf{z}})}_{b_n}
\end{array}
\tag{63}
$$

To establish the desired probability identity $G_H(\mathsf{y}\,|\,\mathsf{x}) = G_H(\widehat{\mathsf{y}}\,|\,\widehat{\mathsf{y}})$, it suffices to establish the identity $\frac{H(\mathbf{C}(\mathsf{x},\mathsf{z}))}{H(\mathbf{C}(\mathsf{x},\mathsf{y}))} = \frac{H(\mathbf{C}(\widehat{\mathsf{x}},\widehat{\mathsf{z}}))}{H(\mathbf{C}(\widehat{\mathsf{x}},\widehat{\mathsf{y}}))}$ between harmony ratios, as shown by the first four steps in (62). The chain of identities in (64) establishes this desired harmony ratio identity. Steps (64a) and (64d) hold because of the positions (63). Step (64b) holds because of the technical condition (60), with $k = n$, $x = a_n$, and $y = b_n$. Step (64c) holds analogously, by applying the technical condition (60) iteratively: first, for $k = n-1$, $x = a_{n-1}$, and $y = b_{n-1}$; then for $k = n-2$, $x = a_{n-2}$, and $y = b_{n-2}$; and so on.

$$\begin{aligned}
\frac{H(\mathbf{C}(\mathsf{x},\mathsf{z}))}{H(\mathbf{C}(\mathsf{x},\mathsf{y}))} &\overset{(a)}{=} \frac{H(a_1,\ldots,a_m,a_{m+1},\ldots,a_{n-1},\,a_n\,)}{H(b_1,\ldots,b_m,a_{m+1},\ldots,a_{n-1}\,a_n\,)} \\
&\overset{(b)}{=} \frac{H(a_1,\ldots,a_m,\,a_{m+1},\ldots,a_{n-1}\,,b_n)}{H(b_1,\ldots,b_m,\,a_{m+1},\ldots,a_{n-1}\,,b_n)} \\
&\overset{(c)}{=} \frac{H(a_1,\ldots,a_m,b_{m+1},\ldots,b_{n-1},b_n)}{H(b_1,\ldots,b_m,b_{m+1},\ldots,b_{n-1},b_n)} \overset{(d)}{=} \frac{H(\mathbf{C}(\widehat{\mathsf{x}},\widehat{\mathsf{z}}))}{H(\mathbf{C}(\widehat{\mathsf{x}},\widehat{\mathsf{y}}))}
\end{aligned} \tag{64}$$

Step 3. The reasoning in (65) shows that any harmony H that factorizes into the product $H = \prod_{k=1}^{n} f_k$ of factor functions $f_1,\ldots,f_n$ trivially satisfies the technical condition (60).

$$\begin{aligned}
&\frac{H(a_1,\ldots,a_{k-1},\boldsymbol{x},a_{k+1},\ldots,a_n)}{H(b_1,\ldots,b_{k-1},\boldsymbol{x},b_{k+1},\ldots,b_n)} = \\
&= \frac{f_1(a_1)\cdot\ldots\cdot f_{k-1}(a_{k-1})\cdot f_k(x)\cdot f_{k+1}(a_{k+1})\cdot\ldots\cdot f_n(a_n)}{f_1(b_1)\cdot\ldots\cdot f_{k-1}(b_{k-1})\cdot f_k(x)\cdot f_{k+1}(b_{k+1})\cdot\ldots\cdot f_n(b_n)} \\
&= \frac{f_1(a_1)\cdot\ldots\cdot f_{k-1}(a_{k-1})\cdot f_k(y)\cdot f_{k+1}(a_{k+1})\cdot\ldots\cdot f_n(a_n)}{f_1(b_1)\cdot\ldots\cdot f_{k-1}(b_{k-1})\cdot f_k(y)\cdot f_{k+1}(b_{k+1})\cdot\ldots\cdot f_n(b_n)} \\
&= \frac{H(a_1,\ldots,a_{k-1},\boldsymbol{y},a_{k+1},\ldots,a_n)}{H(b_1,\ldots,b_{k-1},\boldsymbol{y},b_{k+1},\ldots,b_n)}
\end{aligned} \tag{65}$$

Step 4. To conclude the proof, we need to show that, vice versa, if H satisfies the technical condition (60), then H factorizes into the product $H = \prod_{k=1}^{n} f_k$ of some factor functions $f_1,\ldots,f_n$. To construct these factor functions, we assume that there exists a vector $\mathbf{b} = (b_1,\ldots,b_n)$ with harmony equal to one, namely $H(\mathbf{b}) = 1$. This assumption can be made without loss of generality, because the harmony H can be rescaled without affecting the corresponding probabilistic grammar G_H. We then define the factor function f_k as in (66): the value $f_k(a)$ assigned by f_k to an argument a is the harmony score assigned by H to this designated vector $\mathbf{b}$, only with the kth component replaced with a.

$$f_k(a) = H(b_1,\ldots,b_{k-1},\,a\,,b_{k+1},\ldots,b_n) \tag{66}$$

The reasoning in (67) then holds for any n-dimensional vector $\mathbf{a} = (a_1,\ldots,a_n)$, establishing that H factorizes as desired. Step (67a) holds because of the technical condition (60) with $k = 1$, $x = a_1$, and $y = b_1$. Step (67b) holds because of the assumption $H(b_1,\ldots,b_n) = 1$ and the definition (66) of the factor function f_1. Step (67c) holds by repeating steps (67a-b) working this time on the 2nd component. Step (67d) holds by repeating again steps (67a-b), iteratively from the 3rd component to the nth component.

Finally, step (67e) holds because of the assumption $H(b_1, \dots, b_n) = 1$.

$$\begin{aligned} H(a_1, a_2, \dots, a_n) &\overset{(a)}{=} \frac{H(a_1, b_2, \dots, b_n)}{H(b_1, b_2, \dots, b_n)} H(b_1, a_2, \dots, a_n) \\ &\overset{(b)}{=} f_1(a_1) \cdot H(b_1, a_2, \dots, a_n) \\ &\overset{(c)}{=} f_1(a_1) \cdot f_2(a_2) \cdot H\big(b_1, b_2, a_3, \dots, a_n\big) \\ &\overset{(d)}{=} f_1(a_1) \cdot f_2(a_2) \cdot \ldots \cdot f_{h-1}(a_{n-1}) \cdot f_n(a_n) \cdot H(b_1, b_2, \dots, b_{n-1}, b_n) \\ &\overset{(e)}{=} f_1(a_1) \cdot f_2(a_2) \cdot \ldots \cdot f_{n-1}(a_{n-1}) \cdot f_n(a_n) \end{aligned} \tag{67}$$

A.18 Any weighted grammar is a ME grammar

Let us denote by $G^f_{\mathbf{w}}$ the harmony-based grammar corresponding through (1) to the weighted harmony $H^f_{\mathbf{w}}$ in (8). No matter the choice of the base function f, $G^f_{\mathbf{w}}$ can be construed as a ME grammar through a suitable constraint transformation, as follows. We start with some ORIGINAL constraints $C_1, \dots, C_n$ and obtain the TRANSFORMED constraints $C^{\mathsf{T}}_1, \dots, C^{\mathsf{T}}_n$ through (68). For instance, when the base function f is the inverse function $f(x) = \frac{1}{1+x}$, we obtain $C^{\mathsf{T}}_k = \log\big(1 + C_k\big)$. And when the base function f is ME's exponential function $f(x) = \exp(-x)$, we obtain $C^{\mathsf{T}}_k = C_k$

$$C^{\mathsf{T}}_k = -\log f(C_k) \tag{68}$$

Each transformed constraint C^{T}_k takes non-negative values (because the base function f is positive and normalized, namely takes values between 0 and 1). Yet, a transformed constraint C^{T}_k can take non-integral values, contrary to the original constraint C_k. But constraint integrality has played no role in the proofs presented so far. We therefore ignore this issue of non-integrality and thus say that a mapping (x, y) comes with both the vector $\mathbf{C}(\mathsf{x}, \mathsf{y})$ of original constraint violations and the vector $\mathbf{C}^{\mathsf{T}}(\mathsf{x}, \mathsf{y})$ of transformed constraint violations. The harmony score of the original constraint violation vector $\mathbf{C}(\mathsf{x}, \mathsf{y})$ according to the weighted harmony $H^f_{\mathbf{w}}$ in (8) corresponding to an arbitrary base function f coincides with the ME harmony score of the transformed constraint violation vector $\mathbf{C}^{\mathsf{T}}(\mathsf{x}, \mathsf{y})$, namely $H^f_{\mathbf{w}}(\mathbf{C}(\mathsf{x}, \mathsf{y})) = H^{\text{ME}}_{\mathbf{w}}(\mathbf{C}^{\mathsf{T}}(\mathsf{x}, \mathsf{y}))$. Because of this identity, the weighted grammar $G^{f,\mathbf{C}}_{\mathbf{w}}$ corresponding to an arbitrary base function f and the original constraint set $\mathbf{C}$ coincides with the ME grammar $G^{\text{ME},\mathbf{C}^{\mathsf{T}}}_{\mathbf{w}}$ corresponding to the set $\mathbf{C}^{\mathsf{T}}$ of transformed constraints, as stated in (69).

$$G^{f,\mathbf{C}}_{\mathbf{w}} = G^{\text{ME},\mathbf{C}^{\mathsf{T}}}_{\mathbf{w}} \tag{69}$$

A.19 First half of proof of theorem 9

Let us focus on the harmonic bounding generalization of section 4. We assume that an implication $(\mathsf{x}, \mathsf{y}) \to (\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ holds relative to the weighted model corresponding to some (positive, decreasing, normalized) base function f and some constraint set $\mathbf{C}$. We consider some constraint C and assume that some candidate of the consequent underlying form $\widehat{\mathsf{x}}$ does not violate it. If that candidate is $\widehat{\mathsf{y}}$, the harmonic bounding

inequality $C(\mathsf{x},\mathsf{y}) \geq C(\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ holds because its right-hand side is equal to zero. Thus, we assume that it is some additional consequent candidate $\widehat{\mathsf{z}}_j$ that does not violate C.

Since the implication $(\mathsf{x},\mathsf{y}) \to (\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ holds relative to the weighted model corresponding to f and $\mathbf{C}$, the identity (69) ensures that this implication holds also relative to the ME typology corresponding to the transformed constraint set $\mathbf{C}^{\mathsf{T}}$. By reasoning as in appendix A.10, we obtain the inequality (70), which is indeed the original inequality (40) for the transformed constraint C^{T}.

$$C^{\mathsf{T}}(\widehat{\mathsf{x}},\widehat{\mathsf{y}}) - C^{\mathsf{T}}(\mathsf{x},\mathsf{y}) \leq C^{\mathsf{T}}(\widehat{\mathsf{x}},\widehat{\mathsf{z}}_j) - \sum_{i=1}^{m} \lambda_i C^{\mathsf{T}}(\mathsf{x},\mathsf{z}_i) \tag{70}$$

The assumption that the additional consequent candidate $\widehat{\mathsf{z}}_j$ does not violate the original constraint C ensures that $\widehat{\mathsf{z}}_j$ does not violate the transformed constraint C^{T} either, as shown in (71). Step (71a) holds because of the definition (68) of the transformed constraint C^{T}. Step (71b) holds because of the assumption $C(\widehat{\mathsf{x}},\widehat{\mathsf{z}}_j) = 0$ that $\widehat{\mathsf{z}}_j$ does not violate the original constraint C. Step (71c) holds because of the assumption $f(0) = 1$ that the base function f is normalized.

$$C^{\mathsf{T}}(\widehat{\mathsf{x}},\widehat{\mathsf{z}}_j) \overset{(a)}{=} -\log f\big(C(\widehat{\mathsf{x}},\widehat{\mathsf{z}}_j)\big) \overset{(b)}{=} -\log f(0) \overset{(c)}{=} -\log 1 = 0 \tag{71}$$

The term $C^{\mathsf{T}}(\widehat{\mathsf{x}},\widehat{\mathsf{z}}_j)$, being equal to zero, can be ignored in the right-hand side of (70). And the term $\sum_{i=1}^{m} \lambda_i C^{\mathsf{T}}(\mathsf{x},\mathsf{z}_i)$ can be dropped because the coefficients λ_i are non-negative and the transformed constraint C^{T} is non-negative as well. This inequality (70) therefore reduces to the harmonic bounding inequality $C^{\mathsf{T}}(\widehat{\mathsf{x}},\widehat{\mathsf{y}}) \leq C^{\mathsf{T}}(\mathsf{x},\mathsf{y})$ for the transformed constraint C^{T}. The latter is equivalent to the harmonic bounding inequality $C(\widehat{\mathsf{x}},\widehat{\mathsf{y}}) \leq C(\mathsf{x},\mathsf{y})$ for the original constraint C, as shown in (72). Step (72a) holds because of the definition (68) of the transformed constraint C^{T}. Step (72b) holds because the base function f is decreasing.

$$\begin{aligned} C^{\mathsf{T}}(\widehat{\mathsf{x}},\widehat{\mathsf{y}}) \leq C^{\mathsf{T}}(\mathsf{x},\mathsf{y}) &\overset{(a)}{\iff} -\log f\big(C(\widehat{\mathsf{x}},\widehat{\mathsf{y}})\big) \leq -\log f\big(C(\mathsf{x},\mathsf{y})\big) \\ &\iff f\big(C(\widehat{\mathsf{x}},\widehat{\mathsf{y}})\big) \geq f\big(C(\mathsf{x},\mathsf{y})\big) \overset{(b)}{\iff} C(\widehat{\mathsf{x}},\widehat{\mathsf{y}}) \leq C(\mathsf{x},\mathsf{y}) \end{aligned} \tag{72}$$

This reasoning extends to the one-step-away, the reverse harmonic bounding, and the cardinality generalizations. In conclusion, the paradoxes derived from these generalizations in sections 4-7 extend from ME to the weighted model corresponding to an arbitrary base function f. The case of the average faithfulness generalization of section 8 is more subtle, as discussed in the following appendix.

A.20 Second half of the proof of theorem 9

Let us suppose that a markedness implication $(\mathsf{x},\mathsf{x}) \to (\widehat{\mathsf{x}},\widehat{\mathsf{x}})$ holds relative to the weighted model corresponding to some (positive, decreasing, normalized) base function f and some constraint set $\mathbf{C}$. The identity (69) then ensures that this markedness implication also holds relative to the ME typology corresponding to the transformed constraint set $\mathbf{C}^{\mathsf{T}}$. If the antecedent and consequent underlying forms x and $\widehat{\mathsf{x}}$ share the same candidate set, by reasoning as in appendix A.14, we obtain the average faithfulness inequality $\overline{F^{\mathsf{T}}}(\widehat{\mathsf{x}}) \geq \overline{F^{\mathsf{T}}}(\mathsf{x})$ for any transformed faithfulness constraint F^{T}.

Unfortunately, this inequality $\overline{F^{\mathsf{T}}}(\widehat{\mathsf{x}}) \geq \overline{F^{\mathsf{T}}}(\mathsf{x})$ does not entail the average faithfulness inequality $\overline{F}(\widehat{\mathsf{x}}) \geq \overline{F}(\mathsf{x})$ for the original constraint F. The problem is that averages are sums and sums of logarithms do not behave well. Yet, the paradoxes derived from the average faithfulness inequality do carry over from ME to the weighted model.

Let us start with the paradox of sheer markedness counts from subsection 8.2. Since the markedness implication $(\mathsf{x},\mathsf{x}) \to (\widehat{\mathsf{x}},\widehat{\mathsf{x}})$ holds relative to the weighted model and since the harmonic bounding generalization does extend to the weighted model as shown in appendix A.19, the consequent form $\widehat{\mathsf{x}}$ cannot contain more nasals than the antecedent form x whenever the original constraint set $\mathbf{C}$ contains the markedness constraint $*[+\text{nasal}]$. To obtain the paradox that x and $\widehat{\mathsf{x}}$ must have exactly the same number of nasals, we only need to consider the faithfulness constraint $F = \textsc{Max}_{[+\text{nasal}]}$ and show that the average faithfulness inequality inequality $\overline{F^{\mathsf{T}}}(\widehat{\mathsf{x}}) \geq \overline{F^{\mathsf{T}}}(\mathsf{x})$ for the transformed constraint secured above entails that $\widehat{\mathsf{x}}$ cannot contain less nasals than x.

To this end, we suppose that the underlying strings x and $\widehat{\mathsf{x}}$ are both concatenations of ℓ oral and nasal segments. Let x and $\widehat{x}$ denote the numbers of their nasals segments. These two underlying strings x and $\widehat{\mathsf{x}}$ share the same candidate set because their candidates are obtained by changing the underlying specifications of the feature [nasal] in all logically possible ways, yielding 2^ℓ shared candidates. The ith (oral or nasal) underlying segment is in correspondence with the ith (oral or nasal) surface segment. These assumptions are all satisfied by the two examples in figure 7.

The average number of violations $\overline{F^{\mathsf{T}}}(\mathsf{x})$ assigned by the transformed faithfulness constraint F^{T} to the candidates of the underlying form x can be expressed in terms of its number x of nasals as in (73). Step (73a) holds because of the definition of the average $\overline{C}$ of a constraint C. Step (73b) holds with the following positions: $\xi(k) = -\log(f(k))$ and $N_F(\mathsf{x},k)$ is the number of candidates of the underlying string x that violate $F = \textsc{Max}_{[+\text{nasal}]}$ exactly k times. (73c) holds because this number $N_F(\mathsf{x},k)$ is equal to $\binom{x}{k}2^{\ell-x}$. In fact, the candidates of x that yield exactly k violations of F can be constructed as follows: choose k of the x nasal segments and de-nasalize them, whereby the factor $\binom{x}{k}$; furthermore, choose a subset of the remaining $\ell - x$ oral segments and nasalize them, whereby the factor $2^{\ell-x}$.

$$\begin{aligned}\overline{F^{\mathsf{T}}}(\mathsf{x}) &\overset{(a)}{=} \frac{1}{2^\ell}\sum_{\mathsf{y}\in\mathcal{R}(\mathsf{x})} F^{\mathsf{T}}(\mathsf{x},\mathsf{y}) \overset{(b)}{=} \frac{1}{2^\ell}\sum_{k=1}^{x}\xi(k)N_F(\mathsf{x},k) \\ &\overset{(c)}{=} \frac{1}{2^\ell}\sum_{k=1}^{x}\xi(k)\binom{x}{k}2^{\ell-x} = \frac{1}{2^x}\sum_{k=1}^{x}\xi(k)\binom{x}{k}\end{aligned} \tag{73}$$

The base function f is decreasing by assumption. The function $\xi = -\log(f)$ is therefore increasing. The reasoning in (73) says that the faithfulness average $\overline{F^{\mathsf{T}}}(x)$ can be interpreted as the expected value of the random variable defined as follows: sample at random a subset of a set of cardinality x; let k be the cardinality of the subset sampled; return $\xi(k)$. Because of this interpretation together with the fact that ξ is increasing, we expect $\overline{F^{\mathsf{T}}}(x)$ to be increasing, whereby $\widehat{x} < x$, then $\overline{F^{\mathsf{T}}}(\widehat{x}) < \overline{F^{\mathsf{T}}}(x)$. The transformed average faithfulness inequality $\overline{F^{\mathsf{T}}}(\widehat{\mathsf{x}}) \geq \overline{F^{\mathsf{T}}}(\mathsf{x})$ established above thus ensures that $\widehat{\mathsf{x}}$ cannot contain less nasals than x, as desired.

For completeness, monotonicity of the function $x \mapsto \overline{F^{\mathsf{T}}}(x)$ expressed in (73) is verified analytically in (74). Steps (74a), (74c), and (74f) hold because of the expression (73) of the transformed averages. Step (74b) holds because of Pascal's identity

$\binom{x+1}{k} = \binom{x}{k} + \binom{x}{k-1}$. Step (74d) holds because the function ξ is increasing. Finally, step (74e) holds because $f(1) = 0$ and thus $\xi(1) = 0$.

$$\begin{aligned}
\overline{F^{\mathsf{T}}}(x+1) &\overset{(a)}{=} \sum_{k=1}^{x+1} \xi(k) \binom{x+1}{k} \frac{1}{2^{x+1}} = \sum_{k=1}^{x} \xi(k) \binom{x+1}{k} \frac{1}{2^{x+1}} + \xi(x+1) \frac{1}{2^{x+1}} \\
&\overset{(b)}{=} \sum_{k=1}^{x} \xi(k) \left\{ \binom{x}{k} + \binom{x}{k-1} \right\} \frac{1}{2^{x+1}} + \xi(x+1) \frac{1}{2^{x+1}} \\
&\overset{(c)}{=} \frac{1}{2} \overline{F^{\mathsf{T}}}(\mathsf{x}) + \sum_{k=1}^{x} \xi(k) \binom{x}{k-1} \frac{1}{2^{x+1}} + \xi(x+1) \frac{1}{2^{x+1}} \\
&= \frac{1}{2} \overline{F^{\mathsf{T}}}(\mathsf{x}) + \sum_{h=0}^{x} \xi(h+1) \binom{x}{h} \frac{1}{2^{x+1}} \overset{(d)}{>} \frac{1}{2} \overline{F^{\mathsf{T}}}(\mathsf{x}) + \sum_{h=0}^{x} \xi(h) \binom{x}{h} \frac{1}{2^{x+1}} \\
&\overset{(e)}{=} \frac{1}{2} \overline{F^{\mathsf{T}}}(\mathsf{x}) + \sum_{h=1}^{x} \xi(h) \binom{x}{h} \frac{1}{2^{x+1}} \overset{(f)}{=} \frac{1}{2} \overline{F^{\mathsf{T}}}(\mathsf{x}) + \frac{1}{2} \overline{F^{\mathsf{T}}}(\mathsf{x}) = \overline{F^{\mathsf{T}}}(\mathsf{x})
\end{aligned} \tag{74}$$

Analogous considerations hold for the paradox of sheer length from subsection 8.1. Indeed, let us introduce a new segment that we interpret as NULL. We suppose that the underlying strings x and $\widehat{\mathsf{x}}$ are concatenations of the same number of null or non-null segments. Let x and $\widehat{x}$ denote the numbers of their null segments. These two underlying strings x and $\widehat{\mathsf{x}}$ share the same candidate set because their candidates are obtained by changing null segments into non-null segments and vice versa in all logically possible ways. The ith (null or non-null) underlying segment is in correspondence with the ith (null or non-null) surface segment. MAX counts the number of non-null underlying segments with a null surface correspondent; and DEP counts the number of null underlying segments with a non-null surface correspondent. By reasoning as above, we conclude that the function $x \mapsto \overline{\textsc{Max}^{\mathsf{T}}}(x)$ is increasing while the function $x \mapsto \overline{\textsc{Dep}^{\mathsf{T}}}(x)$ is decreasing. The average faithfulness inequalities $\overline{\textsc{Max}^{\mathsf{T}}}(\widehat{\mathsf{x}}) \geq \overline{\textsc{Max}^{\mathsf{T}}}(\mathsf{x})$ and $\overline{\textsc{Dep}^{\mathsf{T}}}(\widehat{\mathsf{x}}) \geq \overline{\textsc{Dep}^{\mathsf{T}}}(\mathsf{x})$ thus entail that the two strings x and $\widehat{\mathsf{x}}$ have the same length (namely the same number of non-null segments).

References

Alderete, John and Sara Finley. to appear. Probabilistic phonology: a review of theoretical perspectives, applications, and problems. *Language and Linguistics*.

Anttila, Arto and Curtis Andrus. 2006. T-orders. Stanford University.

Bertsekas, Dimitri P. 2009. *Convex Optimization Theory*. Athena Scientific, Belmont, MA, USA.

Blaho, Sylvia, Patrik Bye, and Martin Krämer. 2007. *Freedom of Analysis?* Mouton de Gruyter, Berlin and New York.

Blevins, Juliette. 1995. The syllable in phonological theory. In J. Goldsmith, editor, *The Handbook of Phonological Theory*. Blackwell, Oxford.

Boersma, Paul. 1998. *Functional Phonology*. Ph.D. thesis, University of Amsterdam, The Netherlands. The Hague: Holland Academic Graphics.

Boersma, Paul and Bruce Hayes. 2001. Empirical tests for the Gradual Learning Algorithm. *Linguistic Inquiry*, 32(1):45–86.

Boersma, Paul and Joe Pater. 2016. Convergence properties of a gradual learning algorithm for Harmonic Grammar. In John McCarthy and Joe Pater, editors, *Harmonic Grammar and Harmonic Serialism*. Equinox Press, London.

Bollobás, Béla. 1997. Volume estimates and rapid mixing. In Silvio Levy, editor, *Flavors of Geometry*. MSRI Publications, pages 151–194.

Boyd, Stephen and Lieven Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press.

Breen, Gavan and Rob Pensalfini. 1999. Arrernte: a language with no syllable onsets. *Linguistic Inquiry*, 30(1):1–25.

Breiss, Canaan and Adam Albright. 2022. Cumulative markedness effects and (non-)linearity in phonotactics. *Glossa: a journal of general linguistics*, 7:1–32.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. The MIT Press.

Coetzee, Andries W. 2004. *What it Means to be a Loser: Non-Optimal Candidates in Optimality Theory*. Ph.D. thesis, University of Massachusetts, Amherst.

Coetzee, Andries W. and Shigeto Kawahara. 2013. Frequency biases in phonological variation. *Natural Language and Linguistic Theory*, 31(1):47–89.

Daland, Robert. 2015. Long words in maximum entropy phonotactic grammars. *Phonology*, 32.3:353–383.

Dryer, M. 1998. Why statistical universals are better than absolute universals. In *Proceedings of the*, pages 123–145, Chicago Linguistics Society.

Dyer, M. and A. Frieze. 1988. On the complexity of computing the volume of a polyhedron. *SIAM Journal on Computing*, 17:967–974.

Dyer, M., A. Frieze, and R. Kannan. 1991. A random polynomial-time algorithm for approximating the volume of convex bodies. *Journal of the Association for Computing Machinery*, 38:1–17.

Evans, Nicholas and Stephen C. Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32.5:429–448.

Goldwater, Sharon and Mark Johnson. 2003. Learning OT constraint rankings using a Maximum Entropy model. In *Proceedings of the Stockholm Workshop on Variation Within Optimality Theory*, pages 111–120, Stockholm University.

Greenberg, Joseph H. 1963. *Universals of Language*. MIT Press, Cambridge, MA.

Guy, G. 1991. Explanation in variable phonology. *Language Variation and Change*, 3:1–22.

Hayes, Bruce. 2017. Varieties of Noisy Harmonic Grammar. In *Proceedings of the 2016 Annual Meeting in Phonology*, Linguistic Society of America, Washington, DC.

Hayes, Bruce. 2022. Deriving the wug-shaped curve: A criterion for assessing formal theories of linguistic variation. *Annual Review of Linguistics*, 8:473–494.

Hayes, Bruce and Zsuzsa Cziráky Londe. 2006. Stochastic phonological knowledge: The case of Hungarian vowel harmony. *Phonology*, 23.1:59–104.

Hayes, Bruce and Colin Wilson. 2008. A Maximum Entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, 39:379–440.

Huang, Fang-Lan, Cho-Jui Hsieh, Kai-Wei Chang, and Chih-Jen Lin. 2010. Iterative scaling and coordinate descent methods for maximum entropy models. *Journal of Machine Learning Research*, 11:815–848.

Hyman, Larry. 2008. Universals in phonology. *The Linguistic Review*, 25:83–137.

Jakobson, Roman. 1941. *Kindersprache, Aphasie und allgemeine Lautgesetze*. Hiltop University Press, Cambridge, Mass.

Kager, René. 1999. *Optimality Theory*. Cambridge University Press, Cambridge.

Karlsson, F. 1982. *Suomen kielen äänne-ja muotorakenne [The phonological and morphological structure of Finnish]*. Werner Söderström Osakeyhtiö, Helsinki.

Kaun, Abigail. 2004. The typology of rounding harmony. In Bruce Hayes, Robert Kirchner, and Donca Steriade, editors, *Phonetically based phonology*. Cambridge University Press, pages 87–116.

Kawahara, Shigeto. 2006. A faithfulness ranking projected from a perceptibility scale: The case of [+voice] in Japanese. *Language*, 82:536–574.

Kiparsky, Paul. 1993. An OT perspective on phonological variation. Stanford University.

Kiparsky, Paul. 1994. Remarks on markedness. Handout from the Second Trilateral Phonology Weekend (TREND 2), UC Santa Cruz.

de Lacy, Paul. 2006. *Markedness: Reduction and Preservation in Phonology*. Cambridge University Press, Cambridge.

Ladefoged, Peter and Ian Maddieson. 1996. *The Sounds of the World's Languages*. Wiley-Blackwell.

Legendre, Géraldine, Yoshiro Miyata, and Paul Smolensky. 1990a. Harmonic Grammar – a formal multi-level connectionist theory of linguistic well-formedness: an application. In *Proceedings of the 12th annual conference of the Cognitive Science Society*, pages 884–891, Lawrence Erlbaum Associates, Hillsdale, NJ.

Legendre, Géraldine, Yoshiro Miyata, and Paul Smolensky. 1990b. Harmonic

Grammar – a formal multi-level connectionist theory of linguistic well-formedness: theoretical foundations. In *Proceedings of the 12th annual conference of the Cognitive Science Society*, pages 388–395, Lawrence Erlbaum, Hillsdale, NJ.
Locke, J.L. 1983. *Phonological Acquisition and Change*. Academic Press, New York.
Lombardi, Linda. 1999. Positional faithfulness and voicing assimilation in Optimality Theory. *Natural Language and Linguistic Theory*, 17:267–302.
Lombardi, Linda. 2003. Markedness and the typology of epenthetic vowels. In *Linguistics and Phonetics 2002 proceedings: Prosody and phonetics*. Rutgers Optimality Archive 578.
Maddieson, Ian. 1984. *Patterns of Sounds*. Cambridge University Press.
Magri, Giorgio. 2018. Efficient computation of implicational universals in constraint-based phonology through the hyperplane separation theorem. In *Proceedings of the 15th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*.
Magri, Giorgio. in progress. A course in categorical and probabilistic constraint-based phonology.
Malouf, Robert. 2013. Maximum entropy models. In Alexander Clark, Chris Fox, and Shalom Lappin, editors, *The Handbook of Computational Linguistics and Natural Language Processing*. Wiley-Blackwell, pages 133–153.
Moreton, Elliott. 1999. Non-computable functions in Optimality Theory. *Linguistics Department Faculty Publication Series*, 101.
Ohala, John J. 1983. The origin of sound patterns in vocal tract constraints. In Peter F. MacNeilage, editor, *The production of speech*. Springer-Verlag, New York, pages 189–216.
van Oostendorp, Marc. 2013. Phonology between theory and data. In S.R. Anderson, J. Moeschler, and F. Reboul, editors, *L'interface langage-cognition / The Language-Cognition Interface. Actes du 19e Congrès International des Linguistes (Langue et Cultures, 45)*. Libraire Droz, Genève and Paris, pages 289–306.
Pater, Joe. 1999. Austronesian nasal substitution and other NC effects. In René Kager, Harry van der Hulst, and Wim Zonneveld, editors, *The Prosody-Morphology Interface*. Cambridge University Press, pages 310–343.
Prékopa, A. 1971. Logarithmic concave measures with application to stochastic programming. *Acta Scientiarum Mathematicarum*, 32:301–315.
Prékopa, A. 1973. On logarithmic concave measures and functions. *Acta Scientiarum Mathematicarum*, 34:335–343.
Prince, Alan and Paul Smolensky. 1993/2004. *Optimality Theory: constraint interaction in generative grammar*. Blackwell, Oxford.
Rudin, Walter. 1953. *Principles of Mathematical Analysis*. McGraw-Hill Book Company. Third edition.
Sherman, D. 1975. Stop and fricative systems: A discussion of paradigmatic gaps and the question of language sampling. In *Stanford working papers in language universals*, volume 17. pages 1–31.
Shosted, Ryan Keith. 2006. *The aeroacoustics of nasalized fricatives*. Ph.D. thesis, University of California, Berkeley.
Siptár, Péter and Miklós Törkenczy. 2000. *The phonology of Hungarian*. Oxford University, Oxford.
Smith, Brian W. and Joe Pater. 2020. French schwa and gradient cumulativity. *Glossa: a journal of general linguistics*, 5:1–33.
Smith, N. V. 1973. *The acquisition of phonology: a case study*. CUP, Cambridge, England.
Smolensky, Paul and Géraldine Legendre. 2006. *The Harmonic Mind*. MIT Press, Cambridge, MA.
Thompson, L. C. 1965. *A Vietnamese Grammar*. University of Washington Press, Seattle.
Zuraw, Kie and Bruce Hayes. 2017. Intersecting constraint families: an argument for Harmonic Grammar. *Language*, 93.3:497–546.