Paradoxes of MaxEnt markedness

Giorgio Magri[1] and Arto Anttila[2]
[1]*CNRS,* [1]*MIT,* [2]*Stanford University,*

1 Introduction

Over the past two decades, theoretical linguistics has taken a probabilistic turn. In phonology, categorical data collected through introspection and fieldwork are nowadays routinely complemented with probabilistic data from corpora and experiments (Hayes, 2022 and references therein). 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 & Finley, to appear). These developments raise 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).

Maximum entropy (ME; Goldwater & Johnson, 2003; Hayes & Wilson, 2008) has been endorsed as a model of probabilistic phonology because of its classical guarantees for grammatical inference (Huang et al. 2010 and Malouf 2013). Yet, we know little about the basic organizing principles of ME phonology, beyond circumstantial evidence of ME's ability to fit specific patterns of empirical frequencies (Zuraw & Hayes, 2017; Smith & Pater, 2020; Breiss & Albright, 2022). The reason is that ME typologies cannot be exhaustively listed and directly inspected because they consist of infinitely many grammars. New tools are needed to analyze ME probabilistic typologies—that is, to do probabilistic ME phonology with the same theoretical ambition that has characterized categorical phonology in the past seventy years.

Anttila & Magri (2018) introduce UNIFORM PROBABILITY INEQUALITIES as a new tool for probabilistic typological analysis. 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 has been reported to hold uniformly across dialects of English (Coetzee, 2004; Coetzee & Kawahara, 2013). These uniform probability inequalities can be interpreted as UNIVERSALS of typologies of probabilistic grammars.

Section 2 starts the investigation of the basic organizing principles of ME phonology by establishing a new generalization about ME uniform probability inequalities. The software $\mathbb{C}$o$\mathbb{G}$e$\mathbb{T}$o (*Convex Geometry Tools for phonological analysis*, available at `https://cogeto.stanford.edu/about`), implements this new generalization (plus additional generalizations developed in Anttila & Magri 2023b,a; Magri 2023c) and thus allows the users to compute the uniform probability inequalities predicted by ME on their own data. Sections 3 and 4 then argue that this new mathematical generalization is phonologically paradoxical and indeed prunes the set of ME uniform probability inequalities down to almost nothing. We conclude in section 5 that ME is not a suitable model of phonology.

2 A new generalization about ME phonology

2.1 *Harmony-based grammars* 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 that are relevant for the description of the phonological system of interest (Prince & Smolensky, 1993/2004). $Gen(\mathsf{x})$ denotes the set of CANDIDATE surface realizations y such that the mapping (x, y) belongs to Gen. A CATEGORICAL GRAMMAR G chooses a surface realization $\mathsf{y} = G(\mathsf{x})$ from the candidate set $Gen(\mathsf{x})$ of each underlying form x. A PROBABILISTIC GRAMMAR G assigns to each mapping (x, y) from Gen a number $G(\mathsf{y} \,|\, \mathsf{x})$ that is interpreted as the probability that the underlying form x is realized as the surface candidate y.

Proceedings of *AMP 2019*

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 PROBABILISTIC TYPOLOGY is a collection of probabilistic grammars for the same set Gen of mappings.

Here is perhaps the simplest strategy to define a probabilistic grammar. First, 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. Then, we define the probability $G_H(\mathsf{y}\,|\,\mathsf{x})$ as this score $H(\mathsf{x},\mathsf{y})$, simply divided by a constant $Z(\mathsf{x})$ that ensures normalization, namely $G_H(\mathsf{y}\,|\,\mathsf{x}) = H(\mathsf{x},\mathsf{y})/Z(\mathsf{x})$. A family $\{H, H' \dots\}$ of harmonies thus yields the typology $\{G_H, G_{H'} \dots\}$ of corresponding harmony-based grammars.

To ensure that the score $H(\mathsf{x},\mathsf{y})$ reflects the phonological harmony of the mapping (x,y), we presuppose a set $\boldsymbol{C}$ consisting of a finite number n of CONSTRAINTS $C_1,\dots,C_n$. 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 & Smolensky, 1993/2004). We then take the harmony score $H(\mathsf{x},\mathsf{y})$ to be a DECREASING function $H\big(\boldsymbol{C}(\mathsf{x},\mathsf{y})\big)$ of the vector $\boldsymbol{C}(\mathsf{x},\mathsf{y}) = (C_1(\mathsf{x},\mathsf{y}),\dots,C_n(\mathsf{x},\mathsf{y}))$ 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 $\boldsymbol{C}(\mathsf{x},\mathsf{y}) \geq \boldsymbol{C}(\mathsf{x},\mathsf{z})$ entails $H(\boldsymbol{C}(\mathsf{x},\mathsf{y})) \leq H(\boldsymbol{C}(\mathsf{x},\mathsf{z}))$.

2.2 *MaxEnt* The position $G_H = H/Z$ above reduces the problem of devising probabilistic grammars to the problem of choosing harmony functions. Here is a principled way to tackle the latter problem. Suppose that a phonological process (such as vowel harmony) is conditioned by two phonological factors (such as vowel quality and number of intervening consonants) that are independent in the sense that no constraint is sensitive to both factors (no constraint is sensitive to both vowel quality and number of intervening consonants). Hayes (2021, 2022) provides extensive evidence that the empirical frequencies of that process applying versus not applying to underlying forms that differ for those two factors sit on shifted sigmoids. Magri (2023a) then shows that a harmony-based grammar G_H satisfies this Shifted Sigmoids Generalization if and only if the harmony H factorizes into the product $H(\boldsymbol{C}(\mathsf{x},\mathsf{y})) = \prod_{k=1}^n f_k(C_k(\mathsf{x},\mathsf{y}))$ of n functions $f_1,\dots,f_n$.[1] In other words, the decision of which harmony score to assign to a mapping (x,y) is broken down into n decisions f_k, each of which takes into account only the number of violations $C_k(\mathsf{x},\mathsf{y})$ assigned by the corresponding constraint C_k, but ignores all other constraints.

This result reduces the problem of choosing harmony functions to the problem of choosing the factor functions f_k. Here is a way to tackle the latter problem. Let us suppose that candidate sets are countably infinite because they consist of all strings of finite but arbitrary length obtained by concatenating a finite number of symbols (such as the symbols listed in the IPA table). Consider the harmony-based grammar $G_H = H/Z$ corresponding to a factorizable harmony $H = \prod_{k=1}^n f_k$. By reasoning as in Daland (2015), Magri (2023a) shows that (under reasonable assumptions on the growth of the constraints with the length of the candidate surface strings), the normalization constant $Z(\mathsf{x})$ is finite (whereby the position $G_H = H/Z$ makes sense) if and only if, for every factor function f_k there exists a non-negative constant w_k such that $f_k(x)$ decreases at least as fast as $\exp(-w_k x)$ as x grows. When we choose the factor function f_k equal to this upper bound $f_k(x) = \exp(-w_k x)$, the factorizable harmony H becomes $H(\boldsymbol{C}(\mathsf{x},\mathsf{y})) = \prod_{k=1}^n f_k(C_k(\mathsf{x},\mathsf{y})) = \prod_{k=1}^n \exp(-w_k C_k(\mathsf{x},\mathsf{y})) = \exp\big(-\sum_{k=1}^n w_k C_k(\mathsf{x},\mathsf{y})\big)$. This is the MAXIMUM ENTROPY (ME) harmony score of the mapping (x,y), namely the exponential of the opposite of its weighted sum of constraint violations.

The derivation of ME from Hayes' Shifted Sigmoids Generalization briefly sketched here shows that, besides being mathematically sound, ME is also a phonologically principled implementation of harmony-based probabilistic phonology. We thus focus on the probabilistic typology consisting of the ME grammars corresponding to all non-negative weights $w_k \geq 0$.

2.3 *Implicational universals as uniform probability inequalities* ME typologies consist of infinitely many grammars that cannot be exhaustively listed and directly inspected. An indirect strategy is needed to understand the phonological principles encoded by ME typologies. A natural such strategy is to investigate ME typologies by extracting 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

[1] Equivalently, the logarithm $\log H = \sum_{k=1}^n \log f_k$ of the harmony H is a SEPARABLE UTILITY FUNCTION in the sense of mathematical economics (Debreu, 1960; Wakker, 1988).

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 & 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 stated in (1).

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

For example, dialects of English that delete /t/ at the end of a coda cluster before vowels, also delete it before consonants. The implication (/cost#us/, [cos us]) $\to$ (/cost#me/, [cos me]) is thus a universal in the sense of (1) 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, Anttila & Magri (2018) 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 stated in (2). Condition (1) is a special case of condition (2), when categorical grammars are construed as probabilistic grammars that assign probabilities equal to zero and one.

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

The universal implications in (1) are a special case of Evans and Levinson's (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 defined in (2). Indeed, although the latter 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 mapping, with no exceptions.

2.4 *Markedness implications* These conditions (1) and (2) 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, whose underlying and surface forms are identical (under the assumption that the distinction between these two levels of representation can be blurred because they are "made up of the same stuff"; Moreton, 2008). What is the proper interpretation of an implication $(\mathsf{y},\mathsf{y}) \to (\widehat{\mathsf{y}},\widehat{\mathsf{y}})$ 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 $(\widehat{\mathsf{y}},\widehat{\mathsf{y}})$ is better than the faithful antecedent mapping (y,y) is that the consequent form $\widehat{\mathsf{y}}$ is less marked than the antecedent form y. Universal implications $(\mathsf{y},\mathsf{y}) \to (\widehat{\mathsf{y}},\widehat{\mathsf{y}})$ between faithful mappings are thus called MARKEDNESS implications because they summarize the markedness hierarchies encoded into the typology. We will now use these universal implications between identity mappings to probe ME markedness hierarchies.

2.5 *A generalization about ME markedness* Let us say that a phonological form is only ONE STEP AWAY (in the direction of some faithfulness constraint F_0) FROM VIOLATING some markedness constraint M_0 provided it does not actually violate M_0 but it is very close to violating it in the following sense: all its non-faithful candidates that closely resemble it (because they violate only the faithfulness constraint F_0 and only once) do violate M_0. To illustrate, let us suppose that candidates are constructed by changing the underlying specifications of voicing and spread glottis in all possible ways. The phonological form ta thus comes with the three non-faithful candidates $\mathsf{t^ha}$, da, $\mathsf{d^ha}$. The only non-faithful candidate that is only one step away from ta in the direction of the faithfulness constraint $F_0 = \text{IDENT}_{[\text{voice}]}$ is da (the other two non-faithful candidates violate $\text{IDENT}_{[\text{spread glt}]}$ as well). Although ta does not violate the markedness constraint

$M_0 = *[+\text{voice}, -\text{son}]$, its one-step-away candidate [da] does violate it. We conclude that ta is only one step away (in the direction of F_0) from violating the markedness constraint M_0.

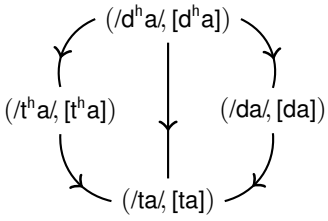
The main result of this paper is the general principle of ME phonology boxed below (for a proof, see Anttila & Magri 2023b). It uses this notion of one-step-away markedness to circumscribe ME markedness hierarchies: a form counts as less marked in ME only if it has more one-step-away markedness violations! This principle of ME markedness makes little phonological sense. In fact, the rest of the paper shows that it rules out many empirically well-grounded markedness asymmetries, making the theory empirically vacuous.

> *Suppose that a markedness implication* $(\mathsf{y}, \mathsf{y}) \to (\widehat{\mathsf{y}}, \widehat{\mathsf{y}})$ *is a universal of a ME typology. If the antecedent mapping* (y, y) *is only one step away in the direction of some faithfulness constraint* F_0 *from violating some markedness constraint* M_0*, then the consequent mapping* $(\widehat{\mathsf{y}}, \widehat{\mathsf{y}})$ *as well is only one step away in the direction of* F_0 *from violating* M_0*.*

3 Paradoxes of voicing and aspiration

3.1 *Basic system* To explore the paradoxical implications of the ME generalization boxed above, we start with the simplest system for laryngeal phonology in (3a). The representational framework Gen consists of a voiceless stop (ta), a voiced stop (da), an aspirated stop (t$^{\text{h}}$a), and a voiced aspirated stop (d$^{\text{h}}$a), each a candidate of the other. The constraint set $\boldsymbol{C}$ consists of the two markedness constraints *[+voice] and *[+spread glottis] (that penalize voiced stops and aspirated stops, respectively) and the two faithfulness constraints IDENT$_{[\text{voice}]}$ and IDENT$_{[\text{spread}]}$ (that penalize discrepancies in voicing and spread glottis).

(3) a. $Gen = \left\{ \begin{matrix} \text{/ta/} \\ \text{/da/} \\ \text{/t}^{\text{h}}\text{a/} \\ \text{/d}^{\text{h}}\text{a/} \end{matrix} \right\} \times \left\{ \begin{matrix} \text{[ta]} \\ \text{[da]} \\ \text{[t}^{\text{h}}\text{a]} \\ \text{[d}^{\text{h}}\text{a]} \end{matrix} \right\}$ $\quad \boldsymbol{C} = \left\{ \begin{matrix} *[+\text{voice}] \\ \text{IDENT}_{[\text{voice}]} \\ *[+\text{spread}] \\ \text{IDENT}_{[\text{spread}]} \end{matrix} \right\}$

b. 

The four identity mappings made available by the representational framework are ordered as in (3b) by the implicational universals predicted by the categorical OT/HG typology according to condition (1). These universals make good sense. In fact, they say that every OT/HG grammar that faithfully realizes the most marked form d$^{\text{h}}$a, faithfully realizes the less marked forms da and t$^{\text{h}}$a as well. Furthermore, every OT/HG grammar that faithfully realizes the latter two forms, faithfully realizes the unmarked form ta as well. All these markedness implications survive in ME—and are therefore plotted as solid arrows in (3b). Every ME grammar assigns less probability to the faithful realization of the most marked form d$^{\text{h}}$a than to the faithful realization of the less marked forms da and t$^{\text{h}}$a. Furthermore, every ME grammar assigns less probability to the faithful realization of the latter two forms than to the faithful realization of the unmarked form ta. In conclusion, ME seems to preserve the generalization that voiced stops and aspirated stops are marked. The next subsection shows that ME's success is ephemeral.

3.2 *When voicing and aspiration interact* Classical Greek and Vietnamese (Thompson 1965) allow voiced stops and aspirated stops but not stops that are both voiced and aspirated. Such languages motivate the additional markedness constraint $M_0 = *[+\text{voice}, +\text{spread glottis}]$ that penalizes voiced aspirated stops at the exclusion of stops that are only voiced or only aspirated,[2] yielding the slightly extended phonological system in (4a). This additional markedness constraint M_0 does not compromise the universal markedness implications (4b) predicted by categorical OT/HG, that are therefore repeated in (4b). Indeed, M_0 does not conflict with the markedness of voicing and aspiration. If anything, it reinforces it.

[2] This constraint $M_0 = *[+\text{voice}, +\text{spread}]$ 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 constraints *[+voice] and *[+spread glottis]. The HG typology without M_0 does not contain a grammar like Vietnamese.

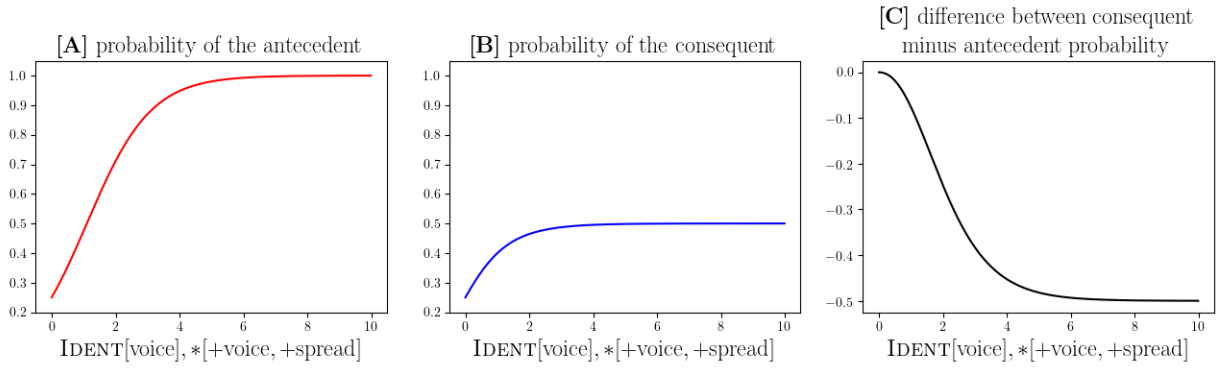
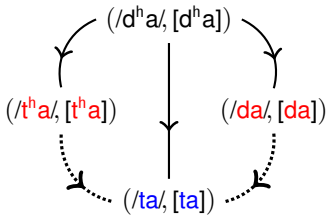
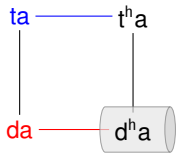


Figure 1: Failure of the markedness implication (/da/, [da]) → (/ta/, [ta]) in ME

(4) a. $Gen = \left\{ \begin{matrix} \text{/ta/} \\ \text{/da/} \\ \text{/t}^\text{h}\text{a/} \\ \text{/d}^\text{h}\text{a/} \end{matrix} \right\} \times \left\{ \begin{matrix} \text{[ta]} \\ \text{[da]} \\ \text{[t}^\text{h}\text{a]} \\ \text{[d}^\text{h}\text{a]} \end{matrix} \right\}$ $\boldsymbol{C} = \left\{ \begin{matrix} \text{*[+voice]} \\ \text{IDENT}_{\text{[voice]}} \\ \text{*[+spread]} \\ \text{IDENT}_{\text{[spread]}} \\ \text{*[+voice, +spread]} \end{matrix} \right\}$

b. 

The situation is different in ME. To illustrate, we focus on the implication (/da/, [da]) → (/ta/, [ta]). Neither the antecedent form da nor the consequent form ta violate the additional markedness constraint M_0 = *[+voice, +spread], plotted as a gray cylinder in (5). If we move right, just one step away from the antecedent form da (in the direction of F_0 = IDENT$_{\text{[spread]}}$), we get to the candidate d$^\text{h}$a which does violate M_0 (it belongs to the cylinder). But if we move right, one step away from the consequent form ta, we get to the candidate t$^\text{h}$a that does *not* violate M_0 (it does not belong to the cylinder). In conclusion, the antecedent form da is only one step away (in the horizontal direction of F_0 = IDENT$_{\text{[spread]}}$) from violating M_0, but the consequent form ta is not. Since the implication (/da/, [da]) → (/ta/, [ta]) flouts the generalization boxed above, it is not a universal of ME markedness—whereby it is plotted as a dotted arrow in (4b).

(5) 

Indeed, figure 1 shows that the marked antecedent mapping (/da/, [da]) can paradoxically have a larger ME probability mass (plotted on the vertical axis) than the unmarked consequent mapping (/ta/, [ta]) when the two constraints M_0 = *[+voice, +spread glottis] and IDENT$_{\text{[voice]}}$ share the same weight (plotted on the horizontal axis) while the other constraints have zero weight.[3] ME thus fails to capture the generalization that voicing is marked. The markedness implication (/t$^\text{h}$a/, [t$^\text{h}$a]) → (/ta/, [ta]) also flouts the boxed generalization (just replace IDENT$_{\text{[spread]}}$ with IDENT$_{\text{[voice]}}$) and thus fails in ME—whereby it is plotted as a dotted arrow in (4b). ME thus also fails to capture the generalization that aspiration is marked: the marked antecedent mapping (/t$^\text{h}$a/, [t$^\text{h}$a]) can have a larger ME probability than the unmarked consequent mapping (/ta/, [ta]).

[3] Setting these many weights to zero is not necessary to construct counterexample weights. In fact, the ME probability of the antecedent mapping (/t$^\text{h}$a/, [t$^\text{h}$a]) is larger than the ME probability of the consequent mapping (/ta/, [ta]) in particular when the constraint weights satisfy the two inequalities $w(\text{Ident}_{\text{[voice]}}) > \log 3 + w(\text{IDENT}_{\text{[spread]}}) + w(*[+\text{voice}]) + w(*[+\text{spread}])$ and $w(*[+\text{voice}, +\text{spread}]) > \log 3$. These inequalities are satisfied in particular when IDENT$_{\text{[voice]}}$ and *[+voice, +spread] have a weight larger than log 3 each and all other constraints have zero weight, as plotted in figure 1.

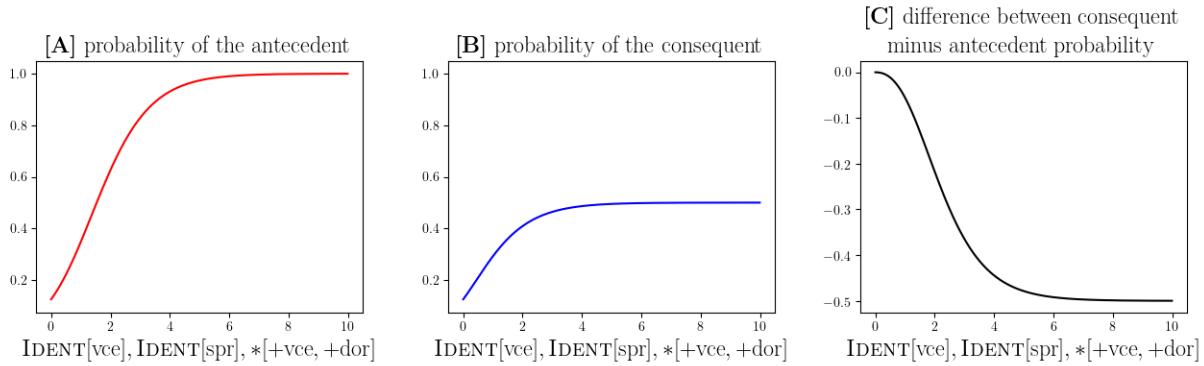
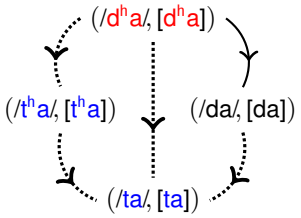


Figure 2: Failure of the markedness implication (/d^{h}a/, [d^{h}a]) → (/ta/, [ta]).

Not everything is lost in ME, though. In fact, the three markedness implications with the shared antecedent (/d^{h}a/, [d^{h}a]) and the consequents (/da/, [da]), (/t^{h}a/, [t^{h}a]), and (/ta/, [ta]) turn out to be ME universals—whereby they are plotted as solid arrows in (4b). Every ME grammar assigns a smaller probability to the faithful realization of the most marked form d^{h}a than to the faithful realization of the other three, less marked forms. In other words, ME does capture the generalization that voicing and aspiration gang up to yield the worst of the worst. The sparse ME markedness universals in (4b) could thus be made sense of as follows. Voicing and aspiration in isolation are marked, but not enough for ME to record it as a universal. Yet, when they gang up, markedness is boosted to a degree that even ME cannot fail to see the following universal: a double markedness violation is universally worse than a single violation or no violation at all. The next subsection shows that even this more modest success of ME is ephemeral.

3.3 *When voicing and place interact* 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). These considerations motivate the additional markedness constraint M_0 = *[+voice, +dorsal] that penalizes velar voiced stops to the exclusion of labial and coronal voiced stops. This markedness constraint is counterbalanced in (6a) by a faithfulness constraint IDENT$_{[place]}$ for place of articulation. In order for these constraints to have bearing, the shared candidate set is enriched with the velar series ka, ga, k^{h}a, and g^{h}a. Once again, this additional markedness constraint M_0 does not compromise the universal markedness implications (4b) predicted by categorical OT/HG, that are therefore repeated in (6b). Indeed, M_0 does not conflict with the markedness of voicing and aspiration. If anything, it reinforces it.

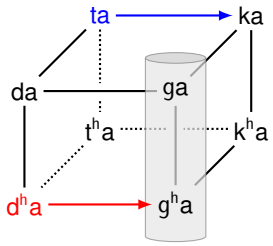
(6) a. $Gen = \left\{ \begin{matrix} \text{/ta/} \\ \text{/da/} \\ \text{/t}^{h}\text{a/} \\ \text{/d}^{h}\text{a/} \end{matrix} \right\} \times \left\{ \begin{matrix} \text{[ta]} & \text{[ka]} \\ \text{[da]} & \text{[ga]} \\ \text{[t}^{h}\text{a]} & \text{[k}^{h}\text{a]} \\ \text{[d}^{h}\text{a]} & \text{[g}^{h}\text{a]} \end{matrix} \right\}$ $\boldsymbol{C} = \left\{ \begin{matrix} \text{*[+voice]} \\ \text{IDENT}_{[\text{voice}]} \\ \text{*[+spread]} \\ \text{IDENT}_{[\text{spread}]} \\ \text{*[+voice, +sprd]} \\ \text{*[+voice, +dor]} \\ \text{IDENT}_{[\text{place}]} \end{matrix} \right\}$

b. 

The situation is different in ME. To illustrate, we focus on the implication (/d^{h}a/, [d^{h}a]) → (/ta/, [ta]). Neither the antecedent form d^{h}a nor the consequent form ta violate the markedness constraint M_0 = *[+voice, +dorsal], plotted as a gray cylinder in (7). If we move right, just one step away from the antecedent form d^{h}a (in the direction of F_0 = IDENT$_{[place]}$), we get to the candidate g^{h}a which does violate M_0 (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_0 (it does not belong to the cylinder). In conclusion, the antecedent form d^{h}a is

only one-step-away (in the horizontal direction of $F_0 = \textsc{Ident}_{[\text{place}]}$) from violating M_0, but the consequent form ta is not. The markedness implication (/dʰa/, [dʰa]) → (/ta/, [ta]) flouts the generalization boxed above and therefore fails in ME—whereby it is plotted as a dotted arrow in (6b).

(7)



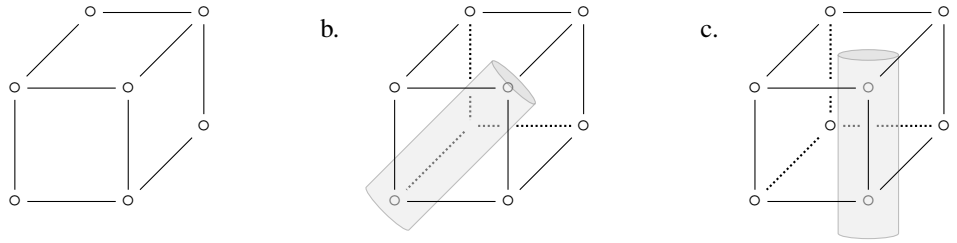
Indeed, figure 2 (virtually identical to figure 1) shows that the ME probability (plotted on the vertical axis) of the marked antecedent mapping (/dʰa/, [dʰa]) can paradoxically be larger than the ME probability of the unmarked consequent mapping (/ta/, [ta]) when the three constraints $\textsc{Ident}_{[\text{voice}]}$, $\textsc{Ident}_{[\text{spread}]}$, and $*[+\text{voice}, +\text{dorsal}]$ share the same weight (plotted on the horizontal axis) while the other constraints have zero weight.[4] ME thus fails to capture even the generalization that voicing and aspiration gang up to yield the worst of the worst.

The markedness implication (/dʰa/, [dʰa]) → (/tʰa/, [tʰa]) fails analogously—whereby it is plotted as a dotted arrow in (6b). In the end, the markedness implication (/dʰa/, [dʰa]) → (/da/, [da]) is the only one that survives in ME—whereby it is plotted as a solid arrow in (6b). But this lonely survivor 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.

4 Paradoxes everywhere

4.1 *Markedness implications must be parallel to feature co-occurrence constraints* To better understand the proposed generalization, let us make explicit the abstract logic of the examples considered so far. We start from a set of phonological features relevant for the phonological system of interest (voicing, aspiration, place, ...). We consider the phonological forms described by all feature value combinations. We assume that they are candidates of each other. We plot these forms as points in a lattice whose dimensions correspond to the features, as in (8a). A markedness constraint M_0 can be represented as a gray cylinder that singles out the forms that violate it. When M_0 is a feature co-occurrence constraint, this cylinder is never diagonal to the directions of the lattice, as in (8b), but always aligned with one direction, as in (8c).

(8) a. b. c.

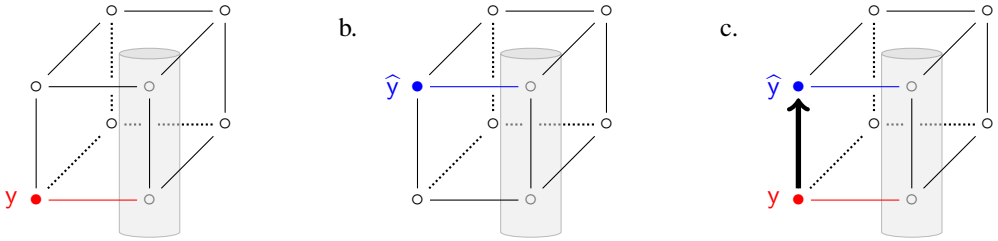


Let us now suppose that the antecedent form y of a ME markedness implicational universal $(y, y) \rightarrow (\widehat{y}, \widehat{y})$ is one step away from violating this feature co-occurrence constraint M_0. This means that y does not belong to the cylinder but is closest to it, namely it is connected to the cylinder through a single step along a singe direction of the lattice. For concreteness, let us suppose that the antecedent form y is the point highlighted as a filled circle in (9a), that is connected to the cylinder through a single step in the horizontal direction. The generalization boxed above requires the consequent form $\widehat{y}$ to be one step away from violating

[4] Once again, setting these many weights to zero is not necessary to construct counterexample weights.

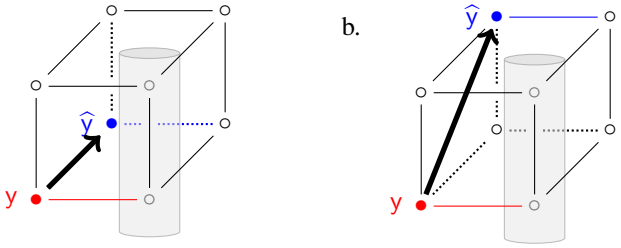
M_0 as well, in the same direction. This means that $\widehat{\text{y}}$ does not belong to the cylinder but is connected to it through a single step in the horizontal direction. The only consequent form $\widehat{\text{y}}$ that complies with the generalization is therefore the point highlighted as a filled circle in (9b). As a result, the the markedness implication $(\text{y}, \text{y}) \to (\widehat{\text{y}}, \widehat{\text{y}})$, plotted as the thick arrow in (9c), runs parallel to the cylinder.

(9) a. b. c.



The consequent form $\widehat{\text{y}}$ cannot be, say, as in (10a) or (10b), that would yield a markedness implication $(\text{y}, \text{y}) \to (\widehat{\text{y}}, \widehat{\text{y}})$ that is not parallel to the cylinder. In conclusion, the generalization boxed above effectively says that markedness implications that are universals of ME must run parallel to any feature co-occurrence constraints that the antecedent form is one step away from violating.

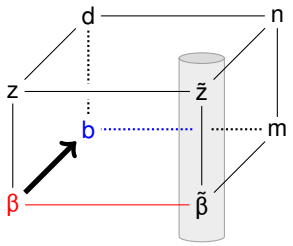
(10) a. b.



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 $(\text{y}, \text{y}) \to (\widehat{\text{y}}, \widehat{\text{y}})$ in (10a) and (10b) paradoxically fails because of a markedness feature co-occurrence constraint that does not distinguish between the antecedent and consequent forms y and $\widehat{\text{y}}$ (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.

4.2 *Obstruent paradoxes* The combinations of values of the features [nasal], [continuant], and [place] are organized into a lattice in (11). The markedness implication $(/\beta/, [\beta]) \to (/\text{b}/, [\text{b}])$ 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 & Maddieson, 1996:§4.4; Shosted, 2006), motivating the feature co-occurrence constraint $M_0 = *[+\text{nasal}, +\text{continuant}, -\text{sonorant}]$, plotted as the cylinder in (11). Intuitively, M_0 is irrelevant to the comparison between β and b (neither is nasal). Yet, the markedness implication $(/\beta/, [\beta]) \to (/\text{b}/, [\text{b}])$ fails in ME because it is not parallel to M_0 (while its antecedent β is one step away from violating it).

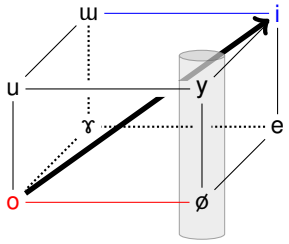
(11)



4.3 *Vowel paradoxes* The combinations of values of the vowel features [back], [high], and [round] are organized into a lattice in (12). The markedness implication $(/\text{o}/, [\text{o}]) \to (/\text{i}/, [\text{i}])$ captures four generalizations.

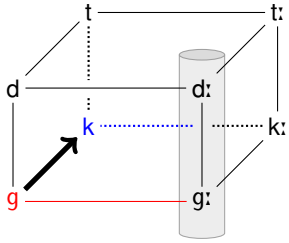
First, rounding is marked (epenthetic vowels are never rounded: Lombardi 2003; de Lacy 2006:§7.2.5). Second, rounding is particularly marked for non-high vowels (Kaun, 2004). Third, back vowels are marked (they are rarely epenthetic: de Lacy, 2006:§7.2.5). Finally, non-high vowels are marked (at least 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 $M_0 = *[+\text{round}, -\text{back}]$ (known as *RoFro; Kaun, 2004), plotted as the cylinder in (12). Intuitively, M_0 is irrelevant to the comparison between o and i (neither violates it). Yet, the implication $(/\text{o}/, [\text{o}]) \to (/\text{i}/, [\text{i}])$ fails in ME because it is not parallel to M_0 (while its antecedent o is one step away from violating it).

(12)



4.4 *More voicing paradoxes* The combinations of values of the features [voice], [place], and [length] are organized into a lattice in (13) (featural encoding of phonological length is not crucial to the argument). The markedness implication $(/\text{g}/, [\text{g}]) \to (/\text{k}/, [\text{k}])$ 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 $M_0 = *[+\text{voice}, +\text{long}]$, plotted as the cylinder in (13) (for an alternative, see Kawahara 2006). Intuitively, M_0 is irrelevant to the comparison between g and k (neither is geminated). Yet, the markedness implication $(/\text{g}/, [\text{g}]) \to (/\text{k}/, [\text{k}])$ fails in ME because it is not parallel to M_0 (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.

(13)



5 Conclusions

In this paper, we have focused on implications of the form $(\text{y}, \text{y}) \to (\widehat{\text{y}}, \widehat{\text{y}})$ that compare the faithful realizations of two phonological forms y and $\widehat{\text{y}}$. Building on Anttila & Magri (2018), we have said that this implication $(\text{y}, \text{y}) \to (\widehat{\text{y}}, \widehat{\text{y}})$ is a universal of a typology of probabilistic phonological grammars provided the probability of the faithful realization of the antecedent form y is never larger than the probability of the faithful realization of the consequent form $\widehat{\text{y}}$ and this probability inequality holds uniformly for every single grammar in the typology. The implicational universal $(\text{y}, \text{y}) \to (\widehat{\text{y}}, \widehat{\text{y}})$ thus defined intuitively captures the generalization that the antecedent form y is more marked than the consequent form $\widehat{\text{y}}$.

We have presented a new generalization about the markedness universals $(\text{y}, \text{y}) \to (\widehat{\text{y}}, \widehat{\text{y}})$ predicted by ME: whenever the antecedent form y is only one step away (in the direction of some faithfulness constraint F_0) from violating some markedness constraint M_0, also the consequent form $\widehat{\text{y}}$ must be one step away (in the same direction F_0) from violating M_0. When M_0 is a feature co-occurrence constraint that the antecedent form y is only one step away from violating, this generalization effectively requires the ME markedness universal $(\text{y}, \text{y}) \to (\widehat{\text{y}}, \widehat{\text{y}})$ to run "parallel" to the markedness constraint M_0 in the lattice of feature-value combinations.

This formal parallelism condition is paradoxical because it has intuitively nothing to do with the phonological substance of markedness. Indeed, we have shown that this paradoxical parallelism condition

condemns ME to failure even in the case of the most basic and empirically well-supported markedness universals that seem intuitively directly encoded in the constraints.

We close by observing that Stochastic (or Noisy) Harmonic Grammar (SHG; Boersma & Pater, 2016) is immune to the paradoxes documented here for ME (Magri 2023b). Indeed, an implication $(x, y) \rightarrow (\widehat{x}, \widehat{y})$ between two arbitrary mappings is a universal of the typology of categorical HG grammars in the usual sense of condition (1) if and only if it is a universal of the typology of probabilistic SHG grammars in the sense of the uniform probability inequality of condition (1). Thus in particular, SHG misses none of HG's markedness universals. We conclude that ME and SHG are very different probabilistic extensions of categorical HG, when compared in terms of their implicational universals.

References

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

Anttila, Arto & Curtis Andrus (2006). T-orders, URL `www.stanford.edu/~anttila/research/torders/t-order-manual.pdf`. Stanford University.

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

Anttila, Arto & Giorgio Magri (2023a). More paradoxes of ME phonology.

Anttila, Arto & Giorgio Magri (2023b). Paradoxes of ME phonology.

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

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

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. & 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.

Debreu, G. (1960). Topological methods in cardinal utility theory. Arrow, K. J., S. Karlin & P. Suppes (eds.), *Mathematical methods in the social sciences*, Stanford University Press, 16–26.

Dryer, M. (1998). Why statistical universals are better than absolute universals. Singer, Kora, Randall Eggert & Gregory Anderson (eds.), *Papers from the 33rd Regional Meeting of the Chicago Linguistic Society*, Chicago Linguistics Society, 123–145.

Evans, Nicholas & 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 & Mark Johnson (2003). Learning OT constraint rankings using a Maximum Entropy model. Spenader, Jennifer, Anders Eriksson & Östen Dahl (eds.), *Proceedings of the Stockholm Workshop on Variation Within Optimality Theory*, Stockholm University, 111–120.

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. Jesney, Karen, Charlie O'Hara, Caitlin Smith & Rachel Walker (eds.), *Proceedings of the 2016 Annual Meeting in Phonology*, Linguistic Society of America, Washington, DC.

Hayes, Bruce (2021). Deriving the wug-shaped curve: A criterion for assessing formal theories of linguistic variation (longer version), URL `https://linguistics.ucla.edu/people/hayes/papers/HayesWugShapedCurve2021LongVersion.pdf`.

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 & 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 & Chih-Jen Lin (2010). Iterative scaling and coordinate descent methods for maximum entropy models. *Journal of Machine Learning Research* 11, 815–848.

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

Kaun, Abigail (2004). The typology of rounding harmony. Hayes, Bruce, Robert Kirchner & Donca Steriade (eds.), *Phonetically based phonology*, Cambridge University Press, 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, URL `http://www.stanford.edu/~kiparsky/Papers/nwave94`. Stanford University.

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

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

Locke, J.L. (1983). *Phonological Acquisition and Change*. Academic Press, New York.

Lombardi, Linda (2003). Markedness and the typology of epenthetic vowels. *Linguistics and Phonetics 2002 proceedings: Prosody and phonetics*. Rutgers Optimality Archive 578.

Maddieson, Ian (1984). *Patterns of Sounds*. Cambridge University Press.

Magri, Giorgio (2023a). A characterization of the probabilistic grammars that satisfy Hayes' (2021) Shifted Sigmoids Generalization.

Magri, Giorgio (2023b). Stochastic Harmonic Grammar: an appraisal.

Magri, Giorgio (2023c). Sufficient conditions for ME uniform probability inequalities.

Malouf, Robert (2013). Maximum entropy models. Clark, Alexander, Chris Fox & Shalom Lappin (eds.), *The Handbook of Computational Linguistics and Natural Language Processing*, Wiley-Blackwell, 133–153.

Moreton, Elliott (2008). Non-computable functions in Optimality Theory. McCarthy, John J. (ed.), *Optimality Theory in Phonology: A reader*, Wiley-Blackwell, Malden: MA, 141–163.

Ohala, John J. (1983). The origin of sound patterns in vocal tract constraints. MacNeilage, Peter F. (ed.), *The production of speech*, Springer-Verlag, New York, 189–216.

Prince, Alan & Paul Smolensky (1993/2004). *Optimality Theory: constraint interaction in generative grammar*. Blackwell, Oxford.

Sherman, D. (1975). Stop and fricative systems: A discussion of paradigmatic gaps and the question of language sampling. *Stanford working papers in language universals*, vol. 17, 1–31.

Shosted, Ryan Keith (2006). *The aeroacoustics of nasalized fricatives*. Ph.D. thesis, University of California, Berkeley, URL `https://escholarship.org/uc/item/00h9g9gg`.

Smith, Brian W. & 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.

Thompson, L. C. (1965). *A Vietnamese Grammar*. University of Washington Press, Seattle.

Wakker, Peter P. (1988). The algebraic versus the topological approach to additive representations. *Journal of Mathematical Psychology* 32, 421–435.

Zuraw, Kie & Bruce Hayes (2017). Intersecting constraint families: an argument for Harmonic Grammar. *Language* 93.3, 497–546.