Does MaxEnt Overgenerate? Implicational Universals in Maximum Entropy Grammar

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

1 Introduction

It goes without saying that a good linguistic theory should neither *undergenerate* (i.e., it should not miss any attested patterns) nor *overgenerate* (i.e., it should not predict any "unattestable" patterns). Recent literature has argued that the *Maximum Entropy* (ME; Goldwater & Johnson 2003) framework provides a probabilistic extension of categorical *Harmonic Grammar* (HG; Legendre et al. 1990; Smolensky & Legendre 2006) which is rich enough to model attested patterns of variable and gradient phonology (see for instance Hayes & Wilson 2008, Zuraw & Hayes 2017, and Smith & Pater 2017). In other words, ME is rich enough to avoid undergeneration. But does ME's richness come at the expense of overgeneration? This is the question that we would like to start investigating in this paper.

In the case of a categorical phonological theory such as HG, overgeneration can be investigated directly by exhaustively listing all the grammars predicted by the theory for certain constraint and candidate sets. That is possible because the predicted typology of grammars is usually *finite*. The situation is rather different for probabilistic theories such as ME. In this case, the typology consists of an *infinite* number of probability distributions which therefore cannot be exhaustively listed and directly inspected. A more indirect strategy is needed to glance at the boundary of the probabilistic typology and thus investigate its overgeneration.

A natural indirect strategy that gets around the problem raised by an infinite typology is to enumerate, not the individual languages/grammars/distributions in the typology, but the corresponding set of implicational universals predicted by the typology. An *implicational universal* is an implication

(1) $P \longrightarrow \widehat{P}$

that holds of a given typology whenever *every* language in the typology that satisfies the antecedent property P also satisfies the consequent property $\widehat{P}$ (Greenberg 1963). Since implicational universals involve universal quantification over the languages in the typology, the larger the typology, the harder it is for implicational universals to hold. A phonological theory overgenerates if it generates so many languages/grammars/distributions that its set of implicational universals becomes very sparse, perhaps empty, and the theory thus fails to predict many implicational universals that empirically hold of natural language.

Which antecedent and consequent properties P and $\widehat{P}$ should we focus on in the pursuit of this investigation of overgeneration? To start from the simplest case, let us consider a typology $\mathfrak{T}$ of *categorical* phonological grammars, construed traditionally as mappings from underlying representations (URs) to surface representations (SRs). Within this categorical framework, the simplest, most basic, most atomic antecedent property P is the property of mapping a certain specific UR x to a certain specific SR y. Analogously, the simplest consequent property $\widehat{P}$ is the property of mapping a certain specific UR $\widehat{\mathsf{x}}$ to a certain specific SR $\widehat{\mathsf{y}}$. We thus focus on implicational universals of the form

(2) $(\mathsf{x}, \mathsf{y}) \xrightarrow{\mathfrak{T}} (\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$

* The research reported in this paper was partially supported by the France-Stanford Center for Interdisciplinary Studies as part of the project *The Mathematics of Language Universals* (A. Anttila and G. Magri, co-coordinators) and by the Agence Nationale de la Recherche as part of the project *The mathematics of segmental phonotactics*. This material has benefited from presentations at the Annual Meeting on Phonology (AMP 2017, September 16, 2017) and at the inaugural meeting of the Society for Computation in Linguistics (SCiL 2018, January 6, 2018). We thank the audiences for helpful questions and comments. In particular, we thank Colin Wilson for discussion. We are responsible for any errors.

Proceedings of *AMP 2017*

which hold provided each grammar in the typology $\mathfrak{T}$ which succeeds at the antecedent mapping (i.e., it maps the UR x to the SR y), also succeeds at the consequent mapping (i.e., it maps the UR $\widehat{x}$ to the SR $\widehat{y}$). The relation $\overset{\mathfrak{T}}{\to}$ thus defined over mappings turns out to be a partial order (under mild additional assumptions). It is called the *T-order* induced by the typology $\mathfrak{T}$ (Anttila & Andrus 2006). For example, it has been found that any dialect of English that deletes a t/d at the end of a coda cluster before a vowel also deletes it before a consonant (Guy 1991; Kiparsky 1993; Coetzee 2004). The implication $(/\mathsf{cost.us}/, [\mathsf{cos.us}]) \to (/\mathsf{cost.me}/, [\mathsf{cos.me}])$ thus holds relative to the typology $\mathfrak{T}$ of English dialects.

Implicational universals can also be statistical. For instance, in dialects of English where t/d deletion applies variably, deletion has been found to be more frequent before consonants than before vowels. In order to model these non-categorical effects, we need to consider a typology $\mathfrak{T}$ of *probabilistic* phonological grammars, construed as functions from URs to probability distributions over SRs. How should the notion of T-order be extended from the categorical to the probabilistic setting? We propose to define T-orders for typologies of probabilistic phonological grammars as follows:

Definition *The implicational universal* $(x, y) \overset{\mathfrak{T}}{\to} (\widehat{x}, \widehat{y})$ *holds relative to a probabilistic typology* $\mathfrak{T}$ *provided each grammar in* $\mathfrak{T}$ *assigns a probability to the consequent mapping* $(\widehat{x}, \widehat{y})$ *which is at least as large as the probability it assigns to the antecedent mapping* (x, y). □

To illustrate, the implication $(/\mathsf{cost.us}/, [\mathsf{cos.us}]) \to (/\mathsf{cost.me}/, [\mathsf{cos.me}])$ considered above also holds relative to the typology $\mathfrak{T}$ of English dialects with variable deletion because the probability of the consequent $(/\mathsf{cost.me}/, [\mathsf{cos.me}])$ (i.e., the frequency of deletion before a consonant) in any dialect is at least as large as the probability of the antecedent $(/\mathsf{cost.us}/, [\mathsf{cos.us}])$ (i.e., the frequency of deletion before a vowel).

The original categorical definition of T-order is a special case of the probabilistic definition just proposed. In fact, suppose that a categorical grammar succeeds on the antecedent mapping (x, y). That grammar construed probabilistically thus assigns probability 1 to the antecedent mapping. The probabilistic definition of T-orders proposed above thus requires that grammar to also assign probability 1 to the consequent mapping $(\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$. In other words, the grammar considered categorically succeeds on the consequent mapping, as required by the original definition of T-orders in the categorical setting.

We are now well equipped to go back to our original question: does ME overgenerate? In section 2, we provide a complete characterization of T-orders in HG based on the constraint violations of the antecedent and consequent mappings together with their losers (see proposition 1). We then show that ME T-orders are a proper subset of HG T-orders, as they are subject to stricter constraint conditions (see propositions 3 and 4) as well as to a condition on the number of candidates of the antecedent and consequent URs. In section 3, we deploy these constraint and candidate conditions for T-orders on two familiar test cases related to basic syllabification (Prince & Smolensky 2004) and obstruent voicing (Lombardi 1999). We show that ME prunes the set of implicational universals in a way that appears to make little phonological sense. In other words, when the ME model of constraint interaction is deployed on these two test cases, the result is a typology of probability distributions that overgenerates by missing a number of phonologically plausible implications. Interestingly, this overgeneration is not an unavoidable drawback of infinite probabilistic typologies. In fact, *stochastic* (or *noisy*) *HG* (Boersma & Pater 2016) is another probabilistic extension of HG which also generates an infinite typology of probability distributions. Yet, the T-order generated by stochastic HG exactly coincides with the T-order generated by categorical HG (see proposition 2). Section 4 concludes the paper.

2 The formal theory of T-orders

This section provides constraint and candidate conditions for T-orders in categorical HG and two of its probabilistic extensions, namely stochastic HG and ME.

2.1 *T-orders in categorical HG* The definition of the HG T-order $(\mathsf{x}, \mathsf{y}) \xrightarrow{\text{HG}} (\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ requires every HG grammar which succeeds on the antecedent mapping to also succeed on the consequent mapping. Checking this definition directly is costly, because it requires computing the entire HG typology which can be large. Yet, one would expect the validity of the T-order to only depend on the violations of the antecedent and consequent mappings together with their losers, with other underlying forms playing no role. Thus, let's focus on the antecedent mapping (x, y) of the UR x to the winner SR y, pitting it against one of its loser mappings (x, z) of that same UR x to some loser SR z. The corresponding *antecedent difference vector* is

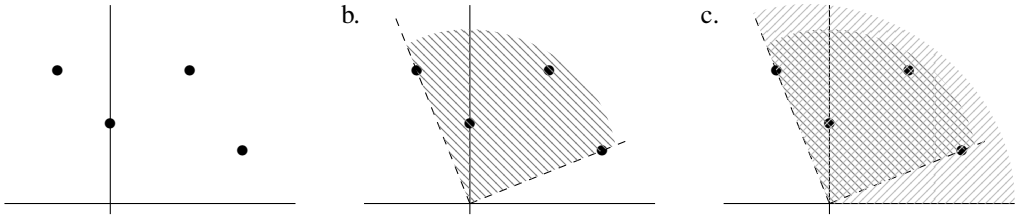
the vector $\mathbf{C}(\mathsf{x},\mathsf{y},\mathsf{z})$ which has a component for each constraint C_k and that component is defined as the number $C_k(\mathsf{x},\mathsf{z})$ of violations assigned by C_k to the loser mapping (x,z) discounted by the number $C_k(\mathsf{x},\mathsf{y})$ of violations assigned to the winner mapping, as in (3).

$$(3)\quad \mathbf{C}(\mathsf{x},\mathsf{y},\mathsf{z}) = \underset{\text{antecedent loser } (\mathsf{x},\mathsf{z})}{\text{violations of the}} - \underset{\text{antecedent winner } (\mathsf{x},\mathsf{y})}{\text{violations of the}} = \begin{bmatrix} C_1(\mathsf{x},\mathsf{z}) - C_1(\mathsf{x},\mathsf{y}) \\ \vdots \\ C_k(\mathsf{x},\mathsf{z}) - C_k(\mathsf{x},\mathsf{y}) \\ \vdots \\ C_n(\mathsf{x},\mathsf{z}) - C_n(\mathsf{x},\mathsf{y}) \end{bmatrix}$$

The *consequent difference vector* $\mathbf{C}(\widehat{\mathsf{x}},\widehat{\mathsf{y}},\widehat{\mathsf{z}})$ is defined analogously, as pitting the consequent winner mapping $(\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ against one of its losers $(\widehat{\mathsf{x}},\widehat{\mathsf{z}})$. Examples of difference vectors are given below in section 3.

Suppose that there are only $n = 2$ constraints, so that the antecedent difference vectors have only two components and can therefore be plotted as points in the plane, as in (4a). The gray region in (4b) is the *convex cone* generated by the set of antecedent difference vectors. Finally, the light gray region in (4c) consists of those points which are at least as large (component by component) than some point in the cone.

(4) a. b. c.



The following proposition says that an HG T-order holds if and only if each consequent difference vector lives in the light gray region depicted in (4c). This necessary and sufficient characterization of T-orders in HG follows from the *Hyperplane Separation Theorem* (HST; Boyd & Vandenberghe 2004) through straightforward algebra, as detailed in Anttila & Magri (2018a).

Proposition 1 *The HG entailment* $(x,y) \xrightarrow{HG} (\widehat{x},\widehat{y})$ *holds (for any number* n *of constraints) if and only if each consequent difference vector* $\mathbf{C}(\widehat{x},\widehat{y},\widehat{z})$ *is at least as large (constraint by constraint) as some vector in the cone generated by the antecedent difference vectors* $\mathbf{C}(x,y,z)$. □

This geometric characterization of HG T-orders admits the following algebraic reformulation: the HG entailment $(\mathsf{x},\mathsf{y}) \xrightarrow{\text{HG}} (\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ holds if and only if for every consequent loser $\widehat{\mathsf{z}}$, it is possible to find a non-negative coefficient $\lambda_{\mathsf{z}} \geq 0$ for each antecedent loser z in such a way that the consequent difference vector $\mathbf{C}(\widehat{\mathsf{x}},\widehat{\mathsf{y}},\widehat{\mathsf{z}})$ is at least as large (constraint by constraint) than the (conic) combination of the antecedent vectors $\mathbf{C}(\mathsf{x},\mathsf{y},\mathsf{z})$ each rescaled by its corresponding coefficient λ_{z}, as stated in (5).

$$(5)\quad \underbrace{\mathbf{C}(\widehat{\mathsf{x}},\widehat{\mathsf{y}},\widehat{\mathsf{z}})}_{\substack{\text{consequent}\\ \text{difference}\\ \text{vector}}} \geq \sum_{\mathsf{z}} \lambda_{\mathsf{z}} \underbrace{\mathbf{C}(\mathsf{x},\mathsf{y},\mathsf{z})}_{\substack{\text{antecedent}\\ \text{difference}\\ \text{vector}}}$$

The condition (5) makes intuitive sense. It says that each consequent loser $\widehat{\mathsf{z}}$ violates the constraints at least as much as (some combination of) the antecedent losers z. Thus, the consequent winner $\widehat{\mathsf{y}}$ has an easier time beating its losers than the antecedent winner y. In other words, it is easier for the consequent to win than it is for the antecedent to win, as required by the definition of T-order.

As anticipated at the beginning of this section, an HG T-order $(\mathsf{x},\mathsf{y}) \xrightarrow{\text{HG}} (\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ is *expensive* to check directly using the definition, because the definition involves a universal quantification over all non-negative weights. The algebraic characterization provided by (5) is instead *easy* to check with readily available linear programming tools. In Anttila & Magri (2018b) we extend this result from HG to OT using the HG-to-OT portability observation in Magri (2013). Thus, also in the case of OT, T-orders can be established without computing the entire OT typology by checking a simple condition on the antecedent and consequent difference vectors analogous to (5).

2.2 *T-orders in stochastic HG* The version of HG considered so far is categorical. In the rest of this section, we investigate probabilistic extensions of HG. A natural way to endow HG with probabilistic structure is to corrupt the constraint weights which parameterize the HG typology with small additive noise. Noise is sampled independently for each constraint, according to a distribution which is usually concentrated around zero (e.g., a gaussian with zero mean), thus ensuring that the noise is small. The resulting framework is called *stochastic* (or *noisy*) HG (Boersma & Pater 2016). Our second result is that the T-orders of categorical and stochastic HG exactly coincide, as stated by the following proposition. As shown in Anttila & Magri (2018b), an analogous identity result holds for the T-orders predicted by categorical OT (Prince & Smolensky 2004), stochastic OT (Boersma 1998), and partial-order OT (Anttila & Cho 1998).

Proposition 2 *The T-order predicted by categorical HG according to the original notion of T-order in the categorical setting exactly coincides (it is the same set of arrows) with the T-order predicted by stochastic HG according to the notion of T-order proposed in section 1 for the probabilistic setting.* □

It is of course reassuring that the extension of the notion of T-orders from the categorical to the probabilistic setting proposed in section 1 ensures that closely related frameworks such as categorical HG and stochastic HG yield closely related (actually identical) T-orders.

2.3 *T-orders in ME* ME can be construed as another probabilistic extension of HG, as indeed ME probabilities depend on the same notion of *harmony* as a weighted sum of constraint violations used in HG. Yet, this section shows that ME and HG induce very different typological structures when measured through T-orders. In other words, stochastic HG and ME are very different probabilistic extensions of categorical HG, as the former maintains intact HG's T-orders while the latter substantially prunes them.

Here is a simple way to start appreciating the difference between T-orders in HG and ME. According to the definition of T-order in the probabilistic setting proposed in section 1, the ME T-order $(\mathsf{x},\mathsf{y}) \xrightarrow{\text{ME}} (\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ requires the inequality (6) to hold for any non-negative weights w_k, which says that the ME probability of the antecedent mapping (x,y) is never larger than the ME probability of the consequent mapping $(\widehat{\mathsf{x}},\widehat{\mathsf{y}})$.[1]

$$(6)\qquad \underbrace{\frac{\exp\{-\sum_k w_k C_k(\mathsf{x},\mathsf{y})\}}{\sum_{\mathsf{z}} \exp\{-\sum_k w_k C_k(\mathsf{x},\mathsf{z})\}}}_{\text{probability of the antecedent } (\mathsf{x},\mathsf{y})} \leq \underbrace{\frac{\exp\{-\sum_k w_k C_k(\widehat{\mathsf{x}},\widehat{\mathsf{y}})\}}{\sum_{\widehat{\mathsf{z}}} \exp\{-\sum_k w_k C_k(\widehat{\mathsf{x}},\widehat{\mathsf{z}})\}}}_{\text{probability of the consequent } (\widehat{\mathsf{x}},\widehat{\mathsf{y}})}$$

Crucially, for weights w_k all equal to zero (or, equivalently, all very small), the inequality (6) becomes (7), as the exponential of 0 is equal to 1. Thus, a necessary condition for the ME entailment $(\mathsf{x},\mathsf{y}) \xrightarrow{\text{ME}} (\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ is that the antecedent UR x has at least as many candidates as the consequent UR $\widehat{\mathsf{x}}$. This makes intuitive sense. As the number of candidates increases, each one of them gets a smaller share of the probability mass. In order for the consequent mapping to get a sufficiently large share of the probability mass, it should not have to share with too many candidates.

$$(7)\qquad \frac{1}{\#\text{ antecedent candidates}} \leq \frac{1}{\#\text{ consequent candidates}}$$

In conclusion, the first difference between T-orders in HG and ME is that T-orders in ME are subject to a candidate condition which requires the antecedent candidate set to be at least as large as the consequent candidate set. In section 3, we will suggest that this ME candidate condition makes little phonological sense.

Assume now that the number of antecedent and consequent candidates is actually the same—a more restrictive assumption than the candidate condition just derived, which nonetheless often holds in practice. Under this assumption on the number of candidates, the antecedent and the consequent probabilities coincide when the weights are all equal to zero. In other words, the difference between the consequent probability minus the antecedent probability is equal to zero in this case. In order for that difference to never become negative, as required by the definition of T-order, it cannot decrease below zero when increasing any of the weights above zero. Through basic calculus, this condition boils down to the inequality (8) between the sums of the antecedent and of the consequent difference vectors, as detailed in Anttila & Magri (2018a).

[1] The sum in the denominator on the left hand side is over all candidates z of the antecedent UR x, including the winner candidate y. Analogously, the sum in the denominator on the right hand side is over all candidates $\widehat{\mathsf{z}}$ of the consequent UR $\widehat{\mathsf{x}}$, including the winner candidate $\widehat{\mathsf{y}}$.

Proposition 3 *If the ME T-order* $(x, y) \xrightarrow{ME} (\widehat{x}, \widehat{y})$ *holds and the antecedent UR* x *and the consequent UR* $\widehat{x}$ *have the same number of candidates, then the sum of the consequent difference vectors is at least as large (constraint by constraint) as the sum of the antecedent difference vectors, i.e.*

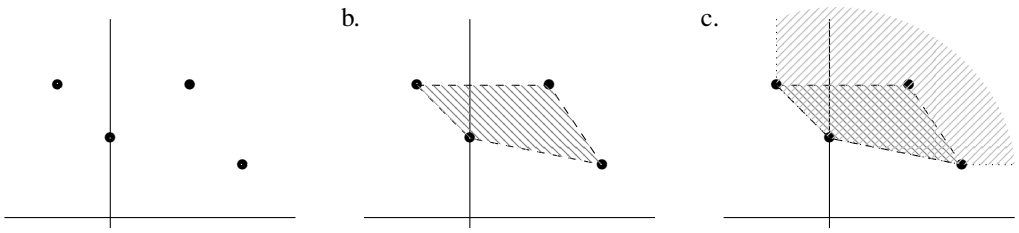
(8) $$\underbrace{\sum_{\widehat{z}} \mathbf{C}(\widehat{x}, \widehat{y}, \widehat{z})}_{\substack{\text{sum of consequent} \\ \text{difference vectors}}} \geq \underbrace{\sum_{z} \mathbf{C}(x, y, z)}_{\substack{\text{sum of antecedent} \\ \text{difference vectors}}}$$

where the sum on the left hand side is over all loser candidates $\widehat{z}$ *of the consequent UR* $\widehat{x}$ *and the sum on the right hand side is over all loser candidates* z *of the antecedent UR* x*.* □

In conclusion, the second difference between T-orders in HG and ME is that T-orders in ME are subject to the global constraint condition (8) which looks at the sum of all consequent difference vectors, while HG T-orders only depend on condition (5) which looks at a single consequent difference vector at a time. In section 3, we will suggest that this ME global constraint condition makes little phonological sense.

Let us take a closer look at constraint conditions for ME T-orders. Suppose again that there are only $n = 2$ constraints, so that the antecedent difference vectors can be plotted as points in the plane, as in (9a). The gray region in (9b) is the *convex hull* generated by the set of antecedent difference vectors, which is usually a proper subset of the convex cone depicted in (4b). Finally, the light gray region in (9c) consists of those points which are at least as large (component by component) than some point in the convex hull.

(9) a. b. c.



The following proposition says that a necessary condition for an ME T-order is that each consequent difference vector lives in the light gray region depicted in (9c). This necessary condition follows again from the HST but through a more involved derivation, detailed in Anttila & Magri (2018a).

Proposition 4 *If the ME T-order* $(x, y) \xrightarrow{ME} (\widehat{x}, \widehat{y})$ *holds (for any number* n *of constraints), then each consequent difference vector* $\mathbf{C}(\widehat{x}, \widehat{y}, \widehat{z})$ *is at least as large (constraint by constraint) as some vector in the convex hull of the antecedent difference vectors* $\mathbf{C}(x, y, z)$*.* □

As in the case of HG T-orders, this geometric characterization of ME T-orders also admits an algebraic translation: if the ME entailment $(\mathsf{x}, \mathsf{y}) \xrightarrow{\text{ME}} (\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ holds, then for every consequent loser $\widehat{\mathsf{z}}$, it is possible to find a non-negative coefficient $\lambda_{\mathsf{z}} \geq 0$ for each antecedent loser z in such a way that the inequality (5) above holds and furthermore the sum of these coefficients λ_{z} over all antecedent losers z is equal to 1, as in (10).

(10) $$\sum_{\mathsf{z}} \lambda_{\mathsf{z}} = 1$$

It follows thus in particular that ME T-orders are a subset of HG T-orders: if the entailment $(\mathsf{x}, \mathsf{y}) \xrightarrow{\text{ME}} (\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ holds in ME, then the inequality (5) holds, which is in turn sufficient to guarantee that the entailment $(\mathsf{x}, \mathsf{y}) \xrightarrow{\text{HG}} (\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ holds in HG. In conclusion, the third difference between T-orders in HG and ME is that the constraint condition (5) for T-orders in ME is subject to the additional normalization condition (10) on the coefficients λ_{z}. In section 3, we will suggest that this additional ME normalization condition makes little phonological sense.

The constraint condition provided by proposition 1 for HG T-orders is both necessary and sufficient. The constraint conditions provided by propositions 3 and 4 for ME T-orders are instead only necessary. They turn out to be also sufficient in the special case where both the antecedent UR x and the consequent UR $\widehat{\mathsf{x}}$ have at most three candidates. For the general case, in Anttila & Magri (2018a) we use the Tomic-Weyl majorization theorem (Marshall et al. 2010:157) to derive a sufficient condition which holds whenever x and $\widehat{\mathsf{x}}$ have the

same number of candidates. This sufficient condition is stronger than the necessary conditions provided by propositions 3 and 4.

3 Phonological applications

In this section, we use the preceding formal results to compute and compare the T-orders predicted by OT/HG and ME on two familiar test cases. We will see that ME prunes the set of T-orders in a way that appears phonologically counterintuitive.[2]

3.1 *Basic syllabification* This test case is described in table (11). No rankings are intended. The familiar constraints ONSET ('a syllable must have an onset'), *CODA ('a syllable must not have a coda'), MAX ('no deletion'), and DEP ('no epenthesis') are deployed on the four inputs /CV/, /CVC/, /VC/, and /V/, each paired up with the four candidates [CV], [CVC], [VC], and [V].

(11)

		ONSET	*CODA	MAX	DEP
/CV/	[CV]				
	[CVC]		1		1
	[VC]	1	1	1	1
	[V]	1		1	
/CVC/	[CV]			1	
	[CVC]		1		
	[VC]	1	1	1	
	[V]	1		2	

		ONSET	*CODA	MAX	DEP
/VC/	[CV]			1	1
	[CVC]		1		1
	[VC]	1	1		
	[V]	1		1	
/V/	[CV]				1
	[CVC]		1		2
	[VC]	1	1		1
	[V]	1			

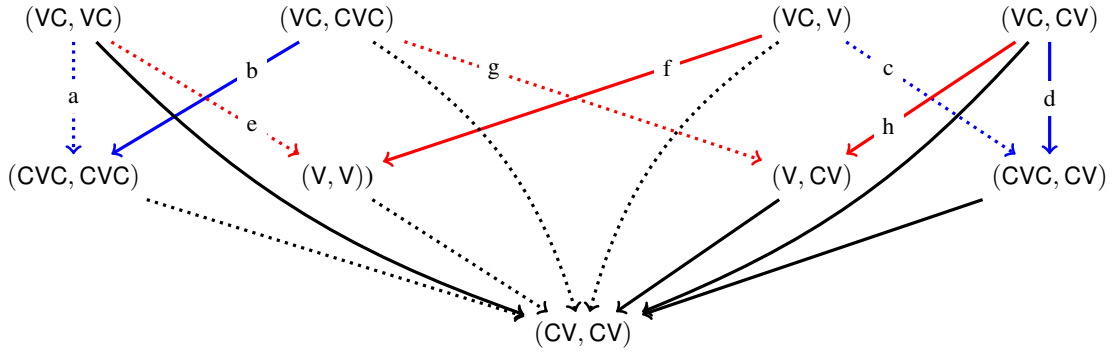
The corresponding OT and HG typologies coincide and yield the four languages listed in (12), as can be verified with OTHelp (Staubs et al. 2010). Language (12a) only allows CV-syllables, requiring onsets and banning codas; language (12b) bans codas, but does not require onsets; language (12c) requires onsets, but does not ban codas; finally, language (12d) allows any type of syllable to emerge faithfully.

(12)

	a.	b.	c.	d.
/CV/:	[CV]	[CV]	[CV]	[CV]
/CVC/:	[CV]	[CV]	[CVC]	[CVC]
/VC/:	[CV]	[V]	[CVC]	[VC]
/V/:	[CV]	[V]	[CV]	[V]

The T-order predicted by this small grammar fragment consists of a surprising 100 implicational universals. 16 of them have a *feasible antecedent* (i.e., a candidate that can potentially win), listed in (13).[3] From now on, we omit the brackets /…/ and […] (the first item in a pair is always a UR and the second item an SR).

(13)



[2] The Python software used to compute T-orders in these two test cases is available from the authors. A user-friendly version of the software will be released shortly.

[3] For reasons of space, we will not discuss the 84 implicational universals with an unfeasible antecedent, i.e., one that is harmonically bounded. In OT and HG, a harmonically bounded mapping entails any other mapping in the T-order. This makes sense: if a harmonically bounded candidate wins (it does not), then any candidate must win. Strikingly, this does not hold in ME where 56 out of the 84 implicational universals with an unfeasible antecedent fail.

These 16 implicational universals can be sorted into three groups. The **four blue arrows** (13a-d) concern codas. The arrows (13a) and (13b) say that if a coda is preserved when there is a problem in the onset position (the onset is either empty or epenthesized), then it is also preserved when there is no such problem. Analogously, the arrows (13c) and (13d) say that if a coda is deleted when there is a problem in the onset position, then it is also deleted when there is no such problem. The **four red arrows** (13e-h) concern onsets. The arrows (13e) and (13f) say that if an empty onset is allowed when there is a problem in the coda position (a coda is either allowed or deleted), then it is also allowed when there is no such problem. Analogously, the arrows (13g) and (13h) say that if an onset is epenthesized when there is a problem in the coda position, then it is also epenthesized when there is no such problem. Finally, the remaining **eight black arrows** all share the consequent $(\mathsf{CV}, \mathsf{CV})$ and thus acknowledge the fact that this mapping does not violate any constraint and is therefore entailed by any mapping. Many of the arrows make obvious phonological sense. For example, the arrows (13d) and (13h) state that if both a coda and an onset are repaired, as in the antecedent $(\mathsf{VC}, \mathsf{CV})$, then surely just the coda or just the onset must be repaired, as in the consequents $(\mathsf{CVC}, \mathsf{CV})$ and $(\mathsf{V}, \mathsf{CV})$.

The OT/HG T-order is halved in ME: the eight arrows dotted in (13) fail in ME. Are these failures problematic or can they be made sense of phonologically? The failures of (13c) and (13g) might be made sense of as additive markedness effects. Consider for instance (13c). The probability of the antecedent $(\mathsf{VC}, \mathsf{V})$ is the probability of deleting a coda when tolerating a marked empty onset. That probability could be high because without deletion the marked filled coda and the marked empty onset could gang up. The probability of the consequent $(\mathsf{CVC}, \mathsf{CV})$ is the probability of deleting a coda without a marked empty onset. One might expect the latter probability to be lower in defiance of the implication, based on the intuition that the marked filled coda could be more tolerable when it does not gang up with a marked empty onset as in the antecedent. Analogous considerations hold for (13g).

However, one might then expect by parity of reasoning that the arrows (13b) and (13f) should also fail because of comparable additive faithfulness effects. Consider for instance (13b). The probability of the antecedent $(\mathsf{VC}, \mathsf{CVC})$ can be interpreted as the probability of preserving a coda when epenthesizing an onset. That probability could be high because, if the coda were deleted instead of preserved, the two faithfulness violations (coda deletion and onset epenthesis) could gang up. The probability of the consequent $(\mathsf{CVC}, \mathsf{CVC})$ can be interpreted as the probability of preserving a coda without any onset unfaithfulness. One might expect the latter probability to be lower in defiance of the implication, based on the intuition that coda deletion could be more tolerable when it does not gang up with onset epenthesis. Analogous considerations hold for (13f).

The failures of the arrows (13a) and (13e) are particularly hard to make sense of from a phonological perspective. Consider (13a). The antecedent mapping $(\mathsf{VC}, \mathsf{VC})$ and the consequent mapping $(\mathsf{CVC}, \mathsf{CVC})$ are both faithful, hence faithfulness constraints are not at stake. The antecedent $(\mathsf{VC}, \mathsf{VC})$ preserves both a marked coda and marked empty onset. ME now makes the puzzling prediction that it can be harder for the consequent $(\mathsf{CVC}, \mathsf{CVC})$ to preserve the coda because it has an onset. It is difficult to see any phonological motivation for such a pattern. An analogous puzzle arises with the failure of (13e). Finally, consider the eight arrows with the consequent $(\mathsf{CV}, \mathsf{CV})$. These arrows capture the intuition that CV is unmarked and should therefore always surface faithfully. The fact that four of these black arrows are lost under ME is truly puzzling. We tentatively conclude that the divide between the implicational universals that survive in ME and those that fail does not admit any natural phonological interpretation. Put differently, embedding Prince and Smolensky's (2004) basic syllabification constraints into ME results in a non-Jakobsonian syllable typology.

The problem is that the geometry of T-orders uncovered in section 2 matches the phonological intuition in the case of HG but not in the case of ME. Here is a way to substantiate this claim (see Anttila & Magri 2018a for a more extensive discussion). Consider the arrow $(\mathsf{VC}, \mathsf{VC}) \rightarrow (\mathsf{CVC}, \mathsf{CVC})$ in (13a). The consequent difference vectors $\mathbf{C}(\widehat{\mathsf{x}}, \widehat{\mathsf{y}}, \widehat{\mathsf{z}})$ corresponding to the consequent winner mapping $(\widehat{\mathsf{x}}, \widehat{\mathsf{y}}) = (\mathsf{CVC}, \mathsf{CVC})$ and the three consequent losers $\widehat{\mathsf{z}} = \mathsf{CV}, \mathsf{VC}, \mathsf{V}$ are listed on the left hand side of (14a). The components of the difference vectors are ordered top-to-bottom according to the left-to-right order of the constraints in table (11): the top row corresponds to the leftmost constraint ONSET, the second row to *CODA, and so on. The first component of the consequent difference vector $\mathbf{C}(\mathsf{CVC}, \mathsf{CVC}, \mathsf{CV})$ is 0 because the first constraint ONSET assigns zero violations to both the loser mapping $(\mathsf{CVC}, \mathsf{CV})$ and the winner mapping $(\mathsf{CVC}, \mathsf{CVC})$. The second component is -1 because the second constraint *CODA assigns zero violations to the loser mapping $(\mathsf{CVC}, \mathsf{CV})$ and one violation to the winner mapping $(\mathsf{CVC}, \mathsf{CVC})$ and $0 - 1 = -1$. And so on. As shown in (14a), the three consequent difference vectors are each larger (constraint by constraint) than the antecedent difference vector $\mathbf{C}(\mathsf{x}, \mathsf{y}, \mathsf{z})$ corresponding to the antecedent winner mapping $(\mathsf{x}, \mathsf{y}) = (\mathsf{VC}, \mathsf{VC})$

and the antecedent loser $\mathsf{z} = \mathsf{V}$ which differs from the antecedent winner $\mathsf{y} = \mathsf{VC}$ only for the deletion of the coda. The implicational universal (13a) thus holds because the HG condition (5) is satisfied with the coefficient λ_{z} equal to 1 for this loser $\mathsf{z} = \mathsf{V}$ and the other coefficients equal to 0.

(14) a. $(\mathsf{VC}, \mathsf{VC}) \rightarrow (\mathsf{CVC}, \mathsf{CVC})$:

$$\begin{array}{l} \\ \textsc{Onset} \\ \textsc{*Coda} \\ \textsc{Max} \\ \textsc{Dep} \end{array} \underbrace{\overset{\mathbf{C}(\mathsf{CVC},\mathsf{CVC},\mathsf{CV})}{\begin{bmatrix} 0 \\ -1 \\ 1 \\ 0 \end{bmatrix}}, \overset{\mathbf{C}(\mathsf{CVC},\mathsf{CVC},\mathsf{VC})}{\begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}}, \overset{\mathbf{C}(\mathsf{CVC},\mathsf{CVC},\mathsf{V})}{\begin{bmatrix} 1 \\ -1 \\ 2 \\ 0 \end{bmatrix}}}_{\text{consequent difference vectors}} \geq \underbrace{\overset{\mathbf{C}(\mathsf{VC},\mathsf{VC},\mathsf{V})}{\begin{bmatrix} 0 \\ -1 \\ 1 \\ 0 \end{bmatrix}}}_{\text{antecedent difference vector}}$$

b. $(\mathsf{VC}, \mathsf{CVC}) \rightarrow (\mathsf{CVC}, \mathsf{CVC})$:

$$\underbrace{\overset{\mathbf{C}(\mathsf{CVC},\mathsf{CVC},\mathsf{CV})}{\begin{bmatrix} 0 \\ -1 \\ 1 \\ 0 \end{bmatrix}}, \overset{\mathbf{C}(\mathsf{CVC},\mathsf{CVC},\mathsf{VC})}{\begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}}, \overset{\mathbf{C}(\mathsf{CVC},\mathsf{CVC},\mathsf{V})}{\begin{bmatrix} 1 \\ -1 \\ 2 \\ 0 \end{bmatrix}}}_{\text{consequent difference vectors}} \geq \underbrace{\overset{\mathbf{C}(\mathsf{VC},\mathsf{CVC},\mathsf{CV})}{\begin{bmatrix} 0 \\ -1 \\ 1 \\ 0 \end{bmatrix}}}_{\text{antecedent difference vector}}$$

Analogous considerations hold for the implicational universal $(\mathsf{VC}, \mathsf{CVC}) \rightarrow (\mathsf{CVC}, \mathsf{CVC})$ in (13b). As shown in (14b), the three consequent difference vectors are larger (constraint by constraint) than the antecedent difference vector $\mathbf{C}(\mathsf{x}, \mathsf{y}, \mathsf{z})$ corresponding to the antecedent winner mapping $(\mathsf{x}, \mathsf{y}) = (\mathsf{VC}, \mathsf{CVC})$ and the antecedent loser $\mathsf{z} = \mathsf{CV}$ which again differs from the antecedent winner $\mathsf{y} = \mathsf{CVC}$ only for the deletion of the coda. The HG condition (5) is thus again satisfied with the coefficient λ_{z} equal to 1 for this loser $\mathsf{z} = \mathsf{CV}$ and the other coefficients equal to 0.

These two implicational universals (13a) and (13b) capture the same phonological intuition: if a coda is preserved in the presence of a markedness or faithfulness violation in the onset position, the coda is also preserved when there is no such violation. The HG formalism perfectly matches this phonological intuition: both arrows hold in HG and they hold for exactly the same formal reason, as underscored in (14).

This complete parallelism that holds between the two arrows (13a) and (13b) in HG is broken in ME. In fact, the implication $(\mathsf{VC}, \mathsf{VC}) \rightarrow (\mathsf{CVC}, \mathsf{CVC})$ in (13a) fails in ME because it does not satisfy the ME global necessary constraint condition provided by proposition 3, as shown in (15): the sum of the consequent difference vectors is not at least as large (constraint by constraint) as the sum of the antecedent difference vectors. The inequality fails for the bottom component (boldfaced) of the difference vectors, corresponding to DEP. The reason is that the consequent UR $\widehat{\mathsf{x}} = \mathsf{CVC}$ is the longest string whereby DEP is never violated by any consequent mapping. The bottom entries corresponding to DEP in the consequent difference vectors are therefore all equal to 0, thus failing at being larger than the sum of the corresponding entries in the antecedent difference vectors.

(15)

$$\underbrace{\overset{\mathbf{C}(\mathsf{CVC},\mathsf{CVC},\mathsf{CV})}{\begin{bmatrix} 0 \\ -1 \\ 1 \\ \mathbf{0} \end{bmatrix}} + \overset{\mathbf{C}(\mathsf{CVC},\mathsf{CVC},\mathsf{VC})}{\begin{bmatrix} 1 \\ 0 \\ 1 \\ \mathbf{0} \end{bmatrix}} + \overset{\mathbf{C}(\mathsf{CVC},\mathsf{CVC},\mathsf{V})}{\begin{bmatrix} 1 \\ -1 \\ 2 \\ \mathbf{0} \end{bmatrix}}}_{\text{consequent difference vectors}} = \underbrace{\begin{bmatrix} 2 \\ -2 \\ 4 \\ \mathbf{0} \end{bmatrix}}_{\text{sum of consequent vectors}} \not\geq \underbrace{\begin{bmatrix} -2 \\ -2 \\ 2 \\ \mathbf{2} \end{bmatrix}}_{\text{sum of antecedent vectors}} = \underbrace{\overset{\mathbf{C}(\mathsf{VC},\mathsf{VC},\mathsf{CVC})}{\begin{bmatrix} -1 \\ 0 \\ 0 \\ \mathbf{1} \end{bmatrix}} + \overset{\mathbf{C}(\mathsf{VC},\mathsf{VC},\mathsf{CV})}{\begin{bmatrix} -1 \\ -1 \\ 1 \\ \mathbf{1} \end{bmatrix}} + \overset{\mathbf{C}(\mathsf{VC},\mathsf{VC},\mathsf{V})}{\begin{bmatrix} 0 \\ -1 \\ 1 \\ \mathbf{0} \end{bmatrix}}}_{\text{antecedent difference vectors}}$$

In sum, the implicational universal $(\mathsf{VC}, \mathsf{VC}) \rightarrow (\mathsf{CVC}, \mathsf{CVC})$ in (13a) fails because the ME formalism is sensitive to the completely spurious property of length of the consequent UR $\widehat{\mathsf{x}} = \mathsf{CVC}$. Futhermore, the implicational universal $(\mathsf{VC}, \mathsf{CVC}) \rightarrow (\mathsf{CVC}, \mathsf{CVC})$ in (13b) does hold in ME, but for reasons which are orthogonal to the intuition captured by the implicational universal. In fact, this implication holds in ME because the antecedent winner SR, the consequent UR, and the consequent winner SR all coincide, namely $\mathsf{y} = \widehat{\mathsf{x}} = \widehat{\mathsf{y}} = \mathsf{CVC}$. In other words, this entailment has the shape $(\mathsf{x}, \mathsf{y}) \rightarrow (\mathsf{y}, \mathsf{y})$. This type of entailment turns out to always hold in ME as a consequence of the sufficient ME conditions briefly mentioned at the end of section 2.3, under mild assumptions on the faithfulness constraints. Intuitively, entailments of the shape

$(x, y) \rightarrow (y, y)$ say that ME does not yield *chain shifts*, whereby x goes to y but y is not faithfully mapped to itself. The validity of the entailment $(x, y) \rightarrow (y, y)$ is thus just another manifestation of the notorious transparency of constraint-based phonology, irrespectively of the specific mode of constraint interaction. The crucial point we want to stress here is that the two implicational universals (13a) and (13b), which capture exactly the same phonological intuition and indeed hold for exactly the same formal reason in HG, happen to live completely different lives in ME, where they are driven by the length of the consequent underlying form $\widehat{x} = \mathsf{CVC}$ on the one hand, and by the inherent transparency of the framework on the other.

3.2 *Obstruent voicing* We now turn to the typology of obstruent voicing, building on Lombardi's (1999) influential analysis. For example, in Swedish word-final obstruents maintain voicing contrast (skog, /g#/ → [g], 'forest'); if adjacent obstruents disagree in voicing, both become voiceless (vigsel, /gs/ → [ks], 'wedding', stekte, /k-d/ → [kt], 'fry-PAST'); if they agree, nothing happens (ägde, /g-d/ → [gd],'own-PAST'). Other languages show different patterns, but the variation is limited and many logically possible patterns are unattested. To account for the limited cross-linguistic variation Lombardi proposed four constraints: AGREE($[voice]$) ('consonant clusters agree in voice'); *VOICE ('no voiced consonant in the output'); ID$ent([voice])$ ('be faithful to [±voice]'); IDentON$set([voice])$ ('be faithful to [±voice] in onsets'). The tableau in (16) spells out our candidate set and shows the constraint violations. We write "T" for voiceless obstruent, "D" for voiced obstruent, and "#" for word boundary. No rankings are intended.

(16)

		AGREE	*VOICE	ID[vce]	IDON[vce]
/T#/	[T#]				
	[D#]		1		1
/D#/	[T#]				1
	[D#]		1		
/DT/	[TT]				1
	[TD]	1	1	1	2
	[DT]	1	1		
	[DD]		2	1	1

		AGREE	*VOICE	ID[vce]	IDON[vce]
/TD/	[TT]			1	1
	[TD]	1	1		
	[DT]	1	1	1	2
	[DD]		2		1
/DD/	[TT]			1	2
	[TD]	1	1		1
	[DT]	1	1	1	1
	[DD]		2		

Lombardi's analysis predicts a typology of six OT languages and nine HG languages, shown in (17). The OT languages are framed. Each OT language is annotated with a sample language. We do not know if the three additional HG patterns are attested.

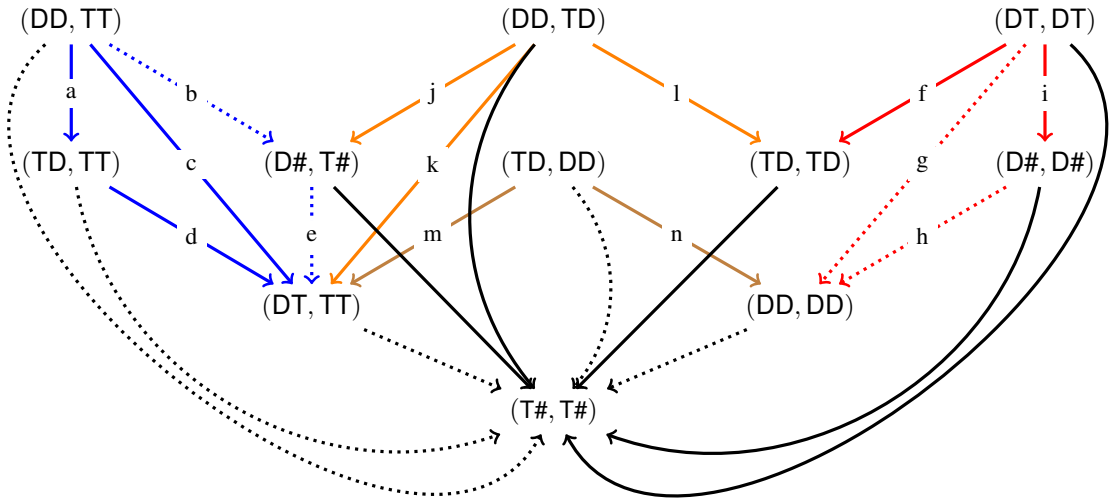
(17)

	Swedish	English	Yiddish	Polish	German	Finnish			
/T#/:	[T#]	[T#]	[T#]	[T#]	[T#]	[T#]	[T#]	[T#]	[T#]
/D#/:	[D#]	[D#]	[D#]	[T#]	[T#]	[T#]	[T#]	[D#]	[T#]
/DT/:	[TT]	[DT]	[TT]	[TT]	[TT]	[TT]	[TT]	[TT]	[TT]
/TD/:	[TT]	[TD]	[DD]	[DD]	[TD]	[TT]	[TT]	[TD]	[TD]
/DD/:	[DD]	[DD]	[DD]	[DD]	[TD]	[TT]	[DD]	[DD]	[DD]

Despite the HG typology being larger than the OT typology, the T-orders predicted by OT and HG turn out to coincide and contain 24 implications with a feasible antecedent,[4] listed in (18).

[4] In addition, there are 62 implicational universals with an unfeasible antecedent of which 29 fail in ME.

(18)



These 24 implicational universals can be sorted into five groups. The **five blue arrows** (18a-e) have to do with devoicing when agreement is not at issue. The arrows (18a), (18b), and (18c) say that if both an onset and a coda are devoiced, then just the onset or just the coda are also devoiced. The arrow (18d) states that progressive devoicing entails regressive devoicing: if an onset is devoiced, then a coda is devoiced as well. This is one of Lombardi's key predictions and can be illustrated from Swedish. Finally, (18e) says that word-final devoicing implies coda devoicing in clusters, another key implicational universal (Lombardi 1999:269) that can be illustrated from Standard German (see, e.g., Wiese 1996:200-205): word-final devoicing in Lob (/b#/ → [p]) 'praise' implies syllable-final devoicing in sagte (/gt/ → [kt]) 'said, 1st person sg.'

The **four red arrows** (18f-i) are voicing preservation universals. First, if coda voicing is preserved despite disagreement, then onset voicing is preserved no matter what (18f-g) and word-final voicing is similarly preserved (18i). Second, if word-final voicing is preserved, then voiced clusters are also preserved (18h). The **three orange arrows** (18j-l) govern devoicing in the face of disagreement. First, if a coda is devoiced when the result is disagreement, then it is devoiced when there is no disagreement (18j-k). Second, if a coda is devoiced when the result is disagreement, then faithful voiced-voiceless clusters are also acceptable (18l). The **two brown arrows** (18m-n) are regressive assimilation universals: regressive voicing implies regressive devoicing (18m), and regressive voicing implies the possibility of voiced clusters (18n). Finally, the remaining **ten black arrows** all share the consequent (T#, T#) that does not violate any constraint and state that spontaneous word-final voicing (T#, D#) is not possible.

The nine dotted arrows in (18) are lost in ME. The failure of (18g-h) (DT, DT) $\not\to$ (DD, DD) and (D#, D#) $\not\to$ (DD, DD) can be understood as additive markedness effects: two voiced obstruents are worse (i.e., yield a lower probability) than just one voiced obstruent. The failure of (18b) (DD, TT) $\not\to$ (D#, T#) can also be interpreted as an additive effect: the probability of devoicing a word-final coda may be low because the alternative (word-final voiced coda) is tolerable, but the probability of devoicing both consonants in a cluster may be higher because devoicing only one would yield a disagreement and devoicing none would yield two voiced obstruents which are less tolerable than a single voiced obstruent, another additive effect.

ME thus permits some potentially reasonable additive effects. However, this comes at a price. Several sensible predictions are lost and some of these failures appear downright strange. To start, consider the loss of five of the ten black arrows with the consequent (T#, T#). This means that ME predicts the possibility of languages where word-final spontaneous voicing has a higher probability than, say, regressive voicing, regressive devoicing, or preservation of voiced clusters. Second, it is troubling that ME fails to derive the devoicing universal by which word-final devoicing implies syllable-final devoicing, captured by the implication (D#, T#) $\not\to$ (DT, TT) in (18e) that Lombardi illustrates from Standard German.

Why does the latter prediction fail in ME? As noted in subsection 2.3, T-orders in ME are subject to a candidate condition that requires the antecedent candidate set to be at least as large as the consequent candidate set. It is precisely this candidate condition that fails here, as shown in table (16): the antecedent UR /D#/ has fewer candidates (two) than the consequent UR /DT/ (four). An easy way of satisfying the

candidate condition would be to make the candidate sets identical as in (19).[5] We also need two additional constraints: MAX and DEP.[6]

(19)

		AGREE	*VOICE	ID[vce]	IDON[vce]	MAX	DEP
/D#/	[D#]		1				
	[T#]			1			
	[TT]			1			1
	[DT]	1	1				1
	[DD]		2				1
	[TD]	1	1	1			1

		AGREE	*VOICE	ID[vce]	IDON[vce]	MAX	DEP
/DT/	[D#]		1			1	
	[T#]			1		1	
	[TT]			1			
	[DT]	1	1				
	[DD]		2	1	1		
	[TD]	1	1	2	1		

However adding candidates does not help: the universal $(\mathsf{D\#}, \mathsf{T\#}) \not\rightarrow (\mathsf{DT}, \mathsf{TT})$ remains lost, this time for the following reason. Recall that proposition 4 says that the ME implication requires each consequent difference vector to be at least as large (constraint by constraint) than some combination of the antecedent difference vectors through non-negative coefficients which add up to 1, as stated in (10). As shown in (20), this condition fails for instance for the consequent difference vector $\mathbf{C}(\mathsf{DT}, \mathsf{TT}, \mathsf{T\#})$ corresponding to the consequent loser T#. The second row corresponding to *VOICE in that consequent difference vector is equal to zero. Since the first four antecedent difference vectors on the right hand side of (20) each have a second component larger than zero, their corresponding four coefficients $\lambda_{\mathsf{D\#}}, \lambda_{\mathsf{DT}}, \lambda_{\mathsf{DD}}, \lambda_{\mathsf{TD}}$ must be each equal to zero. In order for the five coefficients to add up to 1, the fifth coefficient λ_{TT} corresponding to the antecedent loser TT must therefore be equal to 1. But the inequality then fails for the last component corresponding to DEP, as that component is equal to zero in the consequent difference vector.

(20)

$$\underbrace{\begin{array}{r}\text{AGREE}\\ \text{*VOICE}\\ \text{IDENT}\\ \text{IDON}\\ \text{MAX}\\ \text{DEP}\end{array}\overbrace{\begin{bmatrix}0\\ \mathbf{0}\\ 0\\ 0\\ 1\\ \mathbf{0}\end{bmatrix}}^{\mathbf{C}(\mathsf{DT},\mathsf{TT},\mathsf{T\#})}}_{\text{consequent difference vector}} \not\geq \underbrace{\lambda_{\mathsf{D\#}}\overbrace{\begin{bmatrix}0\\ \mathbf{1}\\ -1\\ 0\\ 0\\ 0\end{bmatrix}}^{\mathbf{C}(\mathsf{D\#},\mathsf{T\#},\mathsf{D\#})} + \lambda_{\mathsf{DT}}\overbrace{\begin{bmatrix}1\\ \mathbf{1}\\ -1\\ 0\\ 0\\ 1\end{bmatrix}}^{\mathbf{C}(\mathsf{D\#},\mathsf{T\#},\mathsf{DT})} + \lambda_{\mathsf{DD}}\overbrace{\begin{bmatrix}0\\ \mathbf{2}\\ -1\\ 0\\ 0\\ 1\end{bmatrix}}^{\mathbf{C}(\mathsf{D\#},\mathsf{T\#},\mathsf{DD})} + \lambda_{\mathsf{TD}}\overbrace{\begin{bmatrix}1\\ \mathbf{1}\\ 0\\ 0\\ 0\\ 1\end{bmatrix}}^{\mathbf{C}(\mathsf{D\#},\mathsf{T\#},\mathsf{TD})} \lambda_{\mathsf{TT}}\overbrace{\begin{bmatrix}0\\ 0\\ 0\\ 0\\ 0\\ \mathbf{1}\end{bmatrix}}^{\mathbf{C}(\mathsf{D\#},\mathsf{T\#},\mathsf{TT})}}_{\text{combination of the antecedent difference vectors}}$$

4 Conclusions

We started by noting that a good linguistic theory should neither undergenerate (i.e., it should not miss any attested patterns) nor overgenerate (i.e., it should not predict any "unattestable" patterns). The recent surge of interest in ME as a model of linguistic knowledge has been mainly driven by considerations of learnability and by the framework's ability to closely match empirical data, in particular quantitative data,

[5] The consequent UR /DT/ has two harmonically bounded candidates [TD] and [DD]. As suggested by Colin Wilson (p.c.), the candidate condition could also be satisfied by deleting such candidates. Indeed, that maneuver turns out to suffice to revive the universal $(\mathsf{D\#}, \mathsf{T\#}) \rightarrow (\mathsf{DT}, \mathsf{TT})$. However, this is not a general cure for the overgeneration problem: deleting harmonically bounded candidates sometimes helps, but it may also hurt. Consider the implication $(\mathsf{DT}, \mathsf{DT}) \rightarrow (\mathsf{TD}, \mathsf{TD})$ in (18f): preserving voicing in coda implies preserving it in onset. This implication survives in ME relative to the original candidate set (18). In this case, deleting the harmonically bounded candidates would leave the antecedent UR /DT/ with two candidates, but the consequent UR /TD/ with three, as only DT is harmonically bounded. This would violate the candidate condition, leading to the loss of the implicational universal $(\mathsf{DT}, \mathsf{DT}) \rightarrow (\mathsf{TD}, \mathsf{TD})$ in ME.

[6] We are assuming that in mappings like $(\mathsf{D\#}, \mathsf{DT})$ where the input and the output differ in the number of segments, the word-final coda corresponds to the coda in the cluster.

especially in the domains of phonotactics and phonological variation. This work typically has the goal of showing that ME is rich enough to avoid undergeneration by being able to model phenomena where competing frameworks fall empirically short.

In this paper, we have started to investigate the complementary question of whether ME is also able to avoid overgeneration. This requires a theory-independent way of measuring typological strength. We have advocated the theory's T-order as a measure of typological strength and applied it across three frameworks: Optimality Theory (OT), Harmonic Grammar (HG), and Maximum Entropy Grammar (ME). T-orders have the advantage of being familiar to practicing linguists through Greenberg's (1963) seminal work on implicational universals. They also have the distinct advantage of being applicable across incommensurable frameworks, both categorical and probabilistic.

We have seen that ME has non-trivial T-orders, but compared to OT and HG, they are relatively sparse and sometimes linguistically counterintuitive. The fact that many reasonable implicational universals fail under ME suggests that the theory overgenerates, at least in the two phonological examples we have examined. More generally, our results serve as a reminder that linguistic analyses should be evaluated in terms of both descriptive fit and explanatory depth. A good linguistic theory succeeds on both fronts: we want a flexible theory that best fits the data, but we also want an informative theory that excludes unnatural patterns and derives the correct implicational universals.

References

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

Anttila, Arto & Young-mee Yu Cho (1998). Variation and change in optimality theory. *Lingua* 104, 31–56.

Anttila, Arto & Giorgio Magri (2018a). Does Maximum Entropy overgenerate? Manuscript (Stanford, CNRS).

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

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

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.

Boyd, Stephen & Lieven Vandenberghe (2004). *Convex Optimization*. Cambridge University 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.

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 & Colin Wilson (2008). A Maximum Entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39, 379–440.

Kiparsky, Paul (1993). An OT perspective on phonological variation, URL `http://www.stanford.edu/~kiparsky/Papers/nwave94`. Handout (Stanford).

Legendre, Géraldine, Yoshiro Miyata & Paul Smolensky (1990). Harmonic Grammar: A formal multi-level connectionist theory of linguistic well-formedness: Theoretical foundations. Gernsbacher, Morton Ann & Sharon J. Derry (eds.), *Annual conference of the Cognitive Science Society 12*, Lawrence Erlbaum, Mahwah, NJ, 388–395.

Lombardi, Linda (1999). Positional faithfulness and voicing assimilation in Optimality Theory. *Natural Language and Linguistic Theory* 17, 267–302.

Magri, Giorgio (2013). HG has no computational advantages over OT: towards a new toolkit for computational OT. *Linguistic Inquiry* 44.4, 569–609.

Marshall, A., I. Olin & B. Arnold (2010). *Inequalities: Theory of majorization and its applications*. Springer Series in Statistics, Springer.

Prince, Alan & Paul Smolensky (2004). *Optimality Theory: Constraint Interaction in generative grammar*. Blackwell, Oxford.

Smith, Brian W. & Joe Pater (2017). French schwa and gradient cumulativity. Manuscript (University of California, Berkeley and University of Massachusetts, Amherst).

Smolensky, Paul & Géraldine Legendre (2006). *The Harmonic Mind*. MIT Press, Cambridge, MA.

Staubs, Robert, Michael Becker, Christopher Potts, Patrick Pratt, John J. McCarthy & Joe Pater (2010). OT-Help 2.0. Software package (University of Massachussetts, Amherst).

Wiese, Richard (1996). *The Phonology of German*. Oxford University Press, Oxford.

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