Phonology **32** (2015) 353–383. © Cambridge University Press 2016
doi:10.1017/S0952675715000251

*Long words in maximum entropy phonotactic grammars**

Robert Daland
University of California, Los Angeles

A phonotactic grammar assigns a well-formedness score to all possible surface forms. This paper considers whether phonotactic grammars should be probabilistic, and gives several arguments that they need to be. Hayes & Wilson (2008) demonstrate the promise of a maximum entropy Harmonic Grammar as a probabilistic phonotactic grammar. This paper points out a theoretical issue with maxent phonotactic grammars: they are not guaranteed to assign a well-defined probability distribution, because sequences that contain arbitrary repetitions of unmarked sequences may be underpenalised. The paper motivates a solution to this issue: include a *STRUCT constraint. A mathematical proof of necessary and sufficient conditions to avoid the underpenalisation problem are given in online supplementary materials.

Supercalifragilisticexpialidocious! Even though
the sound of it is something quite atrocious.
Mary Poppins (1964)

1 Introduction

Supercalifragilisticexpialidocious is a made-up word, uttered by the eponymous character of the Disney musical *Mary Poppins* when asked to describe winning a horse race. Although there are many unusual things about this occurrence, it clearly highlights two important facts: listeners occasionally encounter words they have not heard before, and there

* E-mail: R.DALAND@GMAIL.COM.

I wish to acknowledge the entire LSCP (*Laboratoire des Sciences Cognitives*, ENS/Paris), especially Benjamin Börschinger and Abdel Fourtassi, who collaborated with me on the project that led to the insights on this paper, Mark Johnson, who pointed out that maxent should work for Σ^*, and Alex Cristia, Sharon Peperkamp and Emmanuel Dupoux for inviting me to the LSCP. I also wish to acknowledge Colin Wilson and Maria Gouskova for useful discussion of the issue, and the editors of this journal for advice.

Three theorems which formalise the central contributions of this paper are discussed in the online supplementary materials, available at http://www.journals.cambridge.org/issue_Phonology/Vol32No03.

is no upper bound on how long such novel words can be. Therefore, it must be the case that listeners' speech-processing systems are equipped to recognise and process such novel words without any special advance warning. As this paper will argue, this fact has important implications for the architecture and formal character of the PHONOTACTIC GRAMMAR – a hypothesised mental component that evaluates the well-formedness of possible words.

The first part of this paper aims to characterise the properties and architecture that a phonotactic grammar should have, in terms of both the empirical and theoretical needs it serves. One reason to hypothesise such a component was pointed out by Chomsky & Halle (1965) – speakers have tacit knowledge of which sequences could be possible words of their language. Subsequent research has implicated well-formedness in a variety of other language behaviours (e.g. Mattys & Jusczyk 2001, Hay *et al*. 2003, Edwards *et al*. 2004, Storkel *et al*. 2006, Coady & Evans 2008). The empirical finding that well-formedness is gradient (i.e. it has many degrees, rather than being purely a categorical distinction between licit and illicit) can be easily accommodated by modelling the phonotactic grammar as a function that maps possible words to scalar well-formedness values (see Theorem 1 in the online supplementary materials). However, in order to *link* phonotactic well-formedness with speech perception and other observable behaviours, something stronger is required: the common 'language' of probability theory. This paper offers the novel argument that the need for a probabilistic phonotactic grammar falls out from the consensus position that word recognition is probabilistic, together with the commonsense observation that listeners are able to recognise new words like *supercalifragilisticexpialidocious*.

The paper then turns to one particular formalism recently proposed in the literature, which describes the class of MAXENT PHONOTACTIC GRAMMARS (Hayes & Wilson 2008). This approach stands out from competitors in that there is a formal proof of learnability in the face of variable input, which makes it an especially attractive formalism for theoretical development. The paper demonstrates that words like *supercalifragilisticexpialidocious* pose a theoretical problem for maxent phonotactic grammars. In a nutshell, the problem is that there is an infinite number of possible words, but only a finite amount of probability to go around. One way to handle this is to impose an arbitrary finite upper bound on possible word length, so that no probability is assigned to any word over some length k. This is the solution adopted by Hayes & Wilson (2008) in the software implementation released with their paper. They chose a default upper limit of ten segments, meaning that their software assigns a probability of 0 to *supercalifragilisticexpialidocious*. It is of course possible to raise the upper limit past the length of *supercalifragilisticexpialidocious*, but this is not a principled solution to the problem. The fundamental issue is that any sufficiently patient and creative speaker of English can craft a grammatical non-word of arbitrary length.

The opposite problem arises if the finite upper bound is simply removed, without doing anything else. In that case, it may sometimes happen that the grammar fails to sufficiently penalise one or more subsequences, which can be strung together to make arbitrarily long sequences. In fact, as demonstrated in detail in the body of this paper, the Wargamay grammar that Hayes & Wilson report assigns a 'perfect' well-formedness score to any word of the form ba^n, where b = ['bamba], a = [ˌbamba] and n means any number of repetitions (e.g. ['bamba], ['bamba,bamba], ['bamba,bamba,bamba], etc.). Since this theory is supposed to calculate the probability of a form directly from its well-formedness score, it should assign *equal* (and comparatively high) probability to all such words. But there is no way to assign the same positive probability to an infinite number of distinct items without violating the axioms of probability. This shows that it is logically possible – and may happen in practice – that arbitrarily long sequences are *underpenalised*.

Hayes & Wilson's choice of the hard upper limit on non-word length involves sacrificing a few non-words (actually, infinitely many) in order to 'save' the short, high-probability ones. However, careful inspection of the underlying theory on which Hayes & Wilson's approach is based does not actually enforce a hard upper limit (Eisner 2002, Riggle 2004). It is just that one runs the risk of underpenalisation when words of unbounded length are not expressly disallowed.

This paper shows that there is a way to steer between these two outcomes. There is a principled way to assign a proper probability distribution over possible non-words without imposing an arbitrary hard upper limit on word length. The essential problem is that the number of possible words is exponential in their length; the solution is to impose an equal or greater length penalty. Within the context of maxent phonotactic grammars, this outcome is naturally achieved by incorporating a *Struct constraint, which incurs one violation for each overt segment. Theorem 2 in the online supplementary materials gives sufficient conditions to avoid the underpenalisation problem: the weight of the *Struct constraint must have a magnitude greater than $ln\ |\Sigma|$ (where $|\Sigma|$ is the cardinality of the segmental alphabet). Theorem 3 gives an exact test to determine whether a given maxent phonotactic grammar exhibits the underpenalisation problem. Finally, the paper motivates the conjecture that underpenalisation cannot arise as a result of the weight-setting or constraint-induction steps in Hayes & Wilson's software. There are two practical implications of this work for researchers: (i) researchers using the Hayes & Wilson (2008) learner should always include a *Struct constraint in the grammar, with an initial weight greater than the natural log of the size of the alphabet; (ii) researchers who seek to implement a maxent phonotactic grammar should include the underpenalisation test, or force the grammars to include a *Struct constraint as above. In short, the second part of this paper may be thought of as offering a kind of 'safety check' for maxent phonotactic grammars.

https://doi.org/10.1017/S0952675715000251 Published online by Cambridge University Press

2 The properties of phonotactic grammars

2.1 Minimal requirements

A phonotactic grammar is a hypothesised mental component which assesses possible words. This general formulation was already explicit in Chomsky & Halle (1965), who pointed out that English speakers distinguish not only between known words and non-words (e.g. *brick vs. blick*), but also between non-words which could be English words and those which could not (e.g. *blick vs. bnick*). As will be shown in this section, what is needed to formalise this concept is an explicit representation of the set of logically possible words and a set of (ordered) well-formedness values. The phonotactic grammar is then defined as a function mapping from possible words to well-formedness values.

One natural property to assume is that the phonotactic grammar must be able to compare any pair of possible words. This assumption is motivated by the fact that human well-formedness judgement tasks often take the form of determining which of two non-words sounds more like a possible word of the listener's language (e.g. Coleman & Pierrehumbert 1997, Daland *et al*. 2011). Furthermore, it is reasonable to assume that relative well-formedness is TRANSITIVE; for example if a listener indicates that *lbick* is less well-formed than *bnick*, and also that *bnick* is less well-formed than *blick*, then the listener would agree that *lbick* is less well-formed than *blick*. Another property comes from the theory of markedness: it is impossible for a sequence of forms to continually increase in well-formedness. That is, adding a finite amount of material may cause a form to become more well-formed, but one cannot keep increasing well-formedness by adding more and more. Formally, this means that every set of forms must contain a maximally well-formed item. Finally, it is conventional to assume that possible words are strings over an alphabet Σ^*. Theorem 1 shows that any grammar which is consistent with these four assumptions can be represented by a SCORE FUNCTION $H : \Sigma^* \rightarrow \Re^-$, where $\Re^-$ is the set of non-positive real numbers. The relative well-formedness of pairs of words is then indicated with the natural order $\leq$; e.g. $H(bnick) \leq H(blick)$ means that *bnick* is equally well-formed as or less well-formed than *blick*. Under this formulation, the highest possible well-formedness value is 0, and it should be assigned to the maximally well-formed item(s) (see also Smolensky & Legendre 2006: chs 9, 10, 23).

This formulation can be illustrated with reference to Chomsky & Halle's proposal. They envisioned a categorical division between ill-formed and well-formed items. If well-formed items are assigned a score of 0, then ill-formed items should be assigned a score of −1, yielding a mapping $H : \Sigma^* \rightarrow \{-1, 0\}$. Then the relative well-formedness of *bnick vs. blick* is indicated by the inequality $H(bnick) = -1 \leq 0 = H(blick)$.[1] Note that

[1] Of course, the same well-formedness relationships can be modelled in many other ways. For example, one could define a score function that maps to non-*negative*

https://doi.org/10.1017/S0952675715000251 Published online by Cambridge University Press

because this model only allows a binary distinction in well-formedness, it predicts that all ill-formed items are *equally* ill-formed, and all well-formed items are *equally* well-formed. For example, it predicts $H(lbick) = H(bnick)$; $H(ba) = H(supercalifragilisticexpialidocious)$.

As Hayes & Wilson (2008) point out, a binary distinction in well-formedness is not adequate for the empirical data:

> In the particular domain of phonotactics, gradient intuitions are pervasive: they have been found in every experiment that allowed participants to rate forms on a scale … Gradience is also found in the frequency of 'repairs' (such as excrescent vowel insertion) participants make when asked to utter illegal nonce forms … Gradient intuitions can be found even among forms that satisfy the categorical phonotactics of the language, but contain rare sequences … Thus, we consider the ability to model gradient intuitions to be an important criterion for evaluating phonotactic models.

In principle, it is straightforward to account for gradience by extending the range of possible well-formedness values. For example, one might wish to distinguish three values of well-formedness. In that case, unambiguously ill-formed items like *bnick* would be mapped to −2, marginal items like *bwick* or *voig* would be mapped to −1 and well-formed items like *blick* would be mapped to 0. Since the scale is unrestricted, in principle it allows for an arbitrary number of distinctions, including countably infinite distinctions.

To recapitulate, the empirical data and standard theoretical assumptions jointly entail that every phonotactic grammar can be represented by a score function $H : \Sigma^* \rightarrow \Re^-$. A score function does not intrinsically define a probability distribution, but it does provide the ability to account for gradience up to arbitrary degrees of precision. That is, the need to account for gradience can be met without invoking a PROBABILISTIC model of phonotactics.

2.2 Arguments for a probabilistic model of phonotactics

The rationale given by Chomsky & Halle was simply to distinguish well-formed from ill-formed items.[2] However, subsequent research has

real numbers, $\Psi : \Sigma^* \rightarrow \Re^+$, by $\Psi(\omega) = -H(\omega)$. However, this would have the undesirable consequence that $H(a) \leq H(b)$ counterintuitively means that *a* is *more* well-formed than *b*. Alternatively, one could make the range non-negative, and also preserve the semantics of ≤ by defining $\Psi(\omega) = H(\omega) + min\{H(\omega)\}$. This will work when the grammar makes only a finite number of well-formedness distinctions, but not in the general case. Intuitively speaking, one can generally make a non-word worse by adding more ill-formed subparts to it. For example, [lr] is ill-formed, but many phonologists would agree that [lrtb] is worse. By defining score functions so that they map to non-positive reals, we obtain exactly the formulation that is used in maximum entropy models, i.e. a unified treatment.

[2] I am grateful to Ellen Kaisse for pointing out that Chomsky & Halle (1968: 417) did in fact consider a gradient model of phonotactics.

https://doi.org/10.1017/S0952675715000251 Published online by Cambridge University Press

turned up many other uses for phonotactic knowledge. For example, Coleman & Pierrehumbert (1997) found that a word-form's average well-formedness score (from human judgements) was linearly related to its log probability (according to a probabilistic phonotactic model they proposed in the same paper). A linear relationship between log probability (according to the model) and aggregate well-formedness judgments has since been replicated several times (e.g. Coetzee & Kawahara 2011, Daland *et al*. 2011).

Besides the admittedly metalinguistic task of acceptability judgments, phonotactic well-formedness has been implicated in various aspects of speech perception and speech production. For example, Davidson, Wilson and colleagues have made fairly explicit proposals for a Bayesian model of native and non-native speech perception that integrates acoustics and phonotactics (Davidson & Shaw 2012, Wilson & Davidson 2013, Chodroff & Wilson 2014, Wilson *et al*. 2014). In Chodroff & Wilson (2014), the phonotactic grammar plays the role of a 'prior model', which calculates the likelihood of encountering e.g. word-initial [bl] *vs*. word-initial [bəl]; this information is integrated into the acoustic model, which determines how well [bl] matches the observed acoustic sequence *vs*. how well [bəl] does. As in all Bayesian models, the ability to integrate this information depends on being able to express it in the common language of probability distributions.

The final argument that will be made here for a probabilistic phonotactic grammar comes from word recognition. The need for a probabilistic phonotactic grammar in adults is formally entailed by the following assumptions: (i) word recognition is probabilistic, and (ii) listeners sometimes encounter novel words. The assumption that word recognition is probabilistic represents a consensus position in modern theories of word recognition. For example, Norris & McQueen (2008) present a Bayesian model of word recognition called Shortlist B, and model an impressive array of behavioural data. Even models of word recognition which do not use overt probabilities in their update equations, such as the spreading activation model of TRACE (McClelland & Elman 1986), at least implicitly represent a probabilistic formulation in which the probability of a form is proportional to the exponential of its activation level (1986: 21). It is possible to find either explicit description of lexical probabilities or 'word-activation level' equations which correspond very closely to them (usually with activation being linearly related to log probability) in every model of word recognition I am aware of, including MATCHECK (Baayen & Schreuder 2000), automatic speech-recognition systems (e.g. Scharenborg *et al*. 2005) and the word-segmentation models pioneered by Goldwater and colleagues (e.g. Elsner *et al*. 2013).

Something special must occur when a novel word is actually encountered. To see the problem, consider what must occur when a model is presented with a novel word in context, as in (1).

(1) [aɪlaɪktəˈtov]
| | |
I like [təˈtov]

The way that these models process normal sentences is by multiplying the probabilities of all the entries. If the model already had a lexical entry *tatove*, than the probability calculation would be something like (2).

(2) $Pr(I\ like\ tatove \mid [aɪlaɪktəˈtov]) \propto$
$$Pr(I) \times Pr(like) \times Pr(tatove) \times$$
$$Pr(I \rightarrow aɪ) \times Pr(like \rightarrow laɪk) \times Pr(tatove \rightarrow təˈtov)$$

(Actually, the calculation would normally be done in the log domain, as a sum over log probabilities, rather than a product over regular probabilities.) However, since the word *tatove* does not have a preexisting lexical representation, the prior probability assigned to this would normally be 0. In that case, the probability of the parse in (2) would have to be 0 as well, which means it would be literally impossible to recognise the novel form *tatove*. Norris & McQueen (2008: 366) say as much: 'unknown words require special treatment … The model has to consider the hypothesis that the input is not a known word'.

The problem of unseen items has been extensively studied (e.g. Baayen 2001), and the standard approach is to reserve some probability mass, $Pr(\omega_{new})$, for previously unseen items. In the case of (2), this would amount to replacing the $Pr(tatove)$ term with a term $Pr(\omega_{new})$, reflecting the total probability of encountering *any* new word in this context. However, this value does not faithfully represent the event that occurred: it was not just that some previously unseen word was encountered, but also that the previously unseen word has the form *tatove*. The probability of the event should be decomposed into two parts: the probability $Pr(\omega_{new})$ that some previously unseen item has occurred, as well as the probability $Pr(\omega_{new} \rightarrow təˈtov)$ that an item has the form *tatove*, given that it is new. The revised computation is given in (3).

(3) $Pr(I\ like\ tatove \mid [aɪlaɪktəˈtov]) \propto$
$$Pr(I) \times Pr(like) \times Pr(\omega_{new}) \times$$
$$Pr(I \rightarrow aɪ) \times Pr(like \rightarrow laɪk) \times Pr(\omega_{new} \rightarrow təˈtov)$$

Note that this latter term is precisely what a probabilistic phonotactic grammar is supposed to calculate. Furthermore, this equation indicates a natural meaning for the probabilities output by a phonotactic grammar: $Pr(\omega_{new} \rightarrow təˈtov)$ represents the probability that the next nonce word a speaker encounters will have the form [təˈtov]. More generally, then, the formal architecture of word recognition entails a mental component which assigns to each possible word ω a probability, representing the probability that the next novel word the listener encounters

https://doi.org/10.1017/S0952675715000251 Published online by Cambridge University Press

will have the form ω. This point has not been forcibly made in the literature on word recognition, presumably because existing models focus mainly on modelling psycholinguistic data pertaining to the recognition of existing words.

In summary, probabilistic phonotactic grammars are needed to link phonotactic well-formedness with observable speech behaviours such as speech perception and speech production. This is particularly clear in the case of word recognition. The consensus position that word recognition is a probabilistic process and the fact that listeners encounter novel words jointly entail the need for a mental component which assigns a probability to every logically possible word. This is exactly what a probabilistic phonotactic grammar is. Formally, a probabilistic phonotactic grammar can be defined as a phonotactic grammar which assigns probabilities. That is, $\Phi : \Sigma^* \rightarrow \Re^+$ (where $\Re^+$ is the set of non-negative real numbers) is a probabilistic phonotactic grammar if and only if Φ is a phonotactic grammar, and the sum of the probabilities that Φ assigns over every string in Σ^* totals 1. This definition of probabilistic phonotactic grammars is quite standard in the literature. In fact, the only novel contribution of this section is the argument from the theory of word recognition that a probabilistic phonotactic model is needed – and the interpretation it provides for the probabilities that such a model assigns.

3 Maximum entropy Harmonic Grammar

While §2 discussed the formal properties that a phonotactic grammar should have, §3.1 discusses the properties that we desire of a phonological formalism in general. §3.2 argues that maximum entropy Harmonic Grammar (maxent HG) is at least as good as all current competitors on these desiderata, and is superior specifically in terms of learnability and analytic tractability. Finally, §3.3 discusses how Hayes & Wilson (2008) adapt maxent HG to handle phonotactic learning; the resulting class of models will be called maxent phonotactic grammars, to distinguish them from maxent HGs more generally.

3.1 Desiderata

Within generative phonology, it is standard to regard the core goal of a phonological formalism as computing a mapping from lexical representations to surface representations. For example, one way to account for the pronunciation of English *kisses* is that the lexical representation is the concatenation of the stem /kɪs/ and plural marker /z/, with the grammar computing an epenthesis operation /kɪs+z/ → [kɪsɪz]. Besides the core phenomenon of straightforward categorical mappings like these, there are several properties we might want a phonological formalism to exhibit, as shown in (4).

https://doi.org/10.1017/S0952675715000251 Published online by Cambridge University Press

(4) a. *gradience:* as discussed above
b. *variation:* non-determinism in phonological mapping, e.g. *t*-deletion
c. *learning and learnability:* various aspects of how a grammar/model can be identified from a finite amount of data, and the kinds of data that are available during language acquisition
d. *statistical inference:* the ability to draw inferences about the grammar that generated a given set of data, such as whether it justifies hypothesising a constraint against complex codas
e. *analytic tractability:* the ease with which and the extent to which a formalism's properties can be investigated analytically; for example whether the log-odds of two outputs be calculated without reference to other candidates

All of these properties either enable a formalism to account for data or enable analysts to use the formalism in order to reason about human linguistic knowledge and behaviour.

3.2 Properties of maximum entropy Harmonic Grammar

Harmonic Grammar (Legendre *et al*. 1990, Smolensky & Legendre 2006) is a constraint-based formalism which was the historical precursor to Optimality Theory (McCarthy & Prince 1993, Prince & Smolensky 1993), and which crucially differs in having constraints that are WEIGHTED rather than strictly ranked. Maxent HG (Goldwater & Johnson 2003, Jäger 2007, Hayes & Wilson 2008) refers to the natural extension of Harmonic Grammar to a log-linear (maximum entropy) model (Jaynes 1983, Jelinek 1997: ch. 13, Manning & Schütze 1999). Thus maxent HG capitalises on both the theoretically advantageous properties of constraint-based phonology and the statistically sound learning and inference properties of log-linear models.

An HG grammar contains a set of constraints $\{C_k\}_{k=1\ldots K}$ and an associated set of weights $\{w_k\}_{k=1\ldots K}$. Each constraint C_k is a function which maps input–output pairs to a violation value. The HARMONY of an input–output pair $(\mathbf{x}, \mathbf{y})$ is defined as the weighted sum of its constraint violations, as in (5).

(5) $H(\mathbf{x}, \mathbf{y}) = \sum_k w_k \times C_k(\mathbf{x}, \mathbf{y})$

Because violating a constraint is worse than not violating a constraint, more constraint violations should result in a lower harmony. This can be achieved in several notationally distinct but mathematically equivalent ways. This paper will use the convention that constraints assign non-negative values (violation counts), while all weights are non-positive.

The CANDIDATES of an input $\mathbf{x}$ consist of all the input–output pairs for which $\mathbf{x}$ is the input. Formally, this may be indicated by a relation

https://doi.org/10.1017/S0952675715000251 Published online by Cambridge University Press

R (i.e. **x***R***y** means that **y** is an output candidate for input **x**), though in practice it is most common to simply list the candidates in a tableau, as in (6).

(6)

/tra/	C_1: *Complex $w_1 = -25$	C_2: Max(C) $w_2 = -20$	C_3: Dep(V) $w_3 = -5$	H
a. tra	1			−25
b. ta		1		−20
c. ra		1		−20
☞ d. tera			1	−5

Harmonic Grammar tableaux are similar to OT tableaux, but differ in that the constraint weights are explicitly indicated, and include an additional column to indicate the harmony values. The candidate with the greatest harmony is deemed the winner. (6) illustrates a grammar in which ill-formed onset clusters are repaired by vowel epenthesis (rather than cluster simplification).

Maxent HG is a proper extension of Harmonic Grammar. The definition of harmony in (5) remains the same, but in addition, maxent HG defines a function that returns WELL-FORMEDNESS VALUES, as in (7).

(7) $\Phi(\mathbf{x}, \mathbf{y}) = e^{H(\mathbf{x}, \mathbf{y})}$

This value represents the well-formedness of the $\mathbf{x} \rightarrow \mathbf{y}$ mapping. The well-formedness value is similar to a probability, except that it is not required that the sum over all possible items must add up to 1. The PARTITION FUNCTION for an input **x**, $Z(\mathbf{x})$ is defined as the sum of the well-formedness values for the candidates for **x**, as in (8).

(8) $Z(\mathbf{x}) = \sum_{\mathbf{x}R\mathbf{z}} \Phi(\mathbf{x}, \mathbf{z})$

As long as the partition function for an input **x** is finite, it is possible to define the conditional distribution $Pr(\mathbf{y} \mid \mathbf{x})$ as in (9).

(9) $Pr(\mathbf{y} \mid \mathbf{x}) = \Phi(\mathbf{x}, \mathbf{y}) / Z(\mathbf{x})$

Thus the probability of a string is directly proportional to its well-formedness value if the partition function is finite, and is not defined if the partition function is not finite. (The well-formedness value is well-defined whether the partition function is finite or not.) Therefore, a key question – *the* key question addressed in §4 and §5 – is under what circumstances the partition function is finite.

The workings of a maxent HG grammar can also be illustrated by a tableau. (10) repeats tableau (6), augmented with Φ and *Pr* values.

https://doi.org/10.1017/S0952675715000251 Published online by Cambridge University Press

(10)

/tra/	C_1: *Complex $w_1 = -25$	C_2: Max(C) $w_2 = -20$	C_3: Dep(V) $w_3 = -5$	H	Φ	Pr
a. tra	1			-25	e^{-25}	2·06e-9
b. ta		1		-20	e^{-20}	3·05e-7
c. ra		1		-20	e^{-20}	3·05e-7
d. tera			1	-5	e^{-5}	0·9999994

Just as in (6), *Complex and Max(C) have much greater weight than Dep (V). The result is that complex onsets are essentially always repaired by vowel epenthesis. Note that the formalism does assign a marginal degree of probability to other candidates; however these probabilities are so low as to be indistinguishable from the rate of speech errors. In fact, Goldrick & Daland (2009) propose that speech errors can be modelled grammatically in just this way, using the closely related formalism of Noisy Harmonic Grammar. The remainder of this section discusses how maxent HG satisfies the desiderata outlined earlier; the discussion is brief, as many of these points have been covered in more detail by Pater and colleagues in their programme of theory comparison (Pater 2008, to appear, Coetzee & Pater 2011, Boersma & Pater to appear).

3.2.1 *Gradience*. In the context of phonotactics, gradience refers to the fact that listeners distinguish degrees of relative well-formedness even among phonotactic configurations that are generally judged well-formed or ill-formed. For example, there are no English words which begin with *rt* or *bn*, so both are ill-formed, but English listeners reliably prefer the latter over the former when forced to choose, meaning that *rt* is more ill-formed than *bn* (Daland *et al*. 2011). In maxent HG, this kind of gradience is straightforwardly predicted by 'ganging' of 'sonority-regulating constraints' (for detailed exposition see Hayes 2011). For example, consider two constraints on onset clusters: *#[−syll][−son] punishes all consonant–obstruent onset clusters (sonority plateaus and sonority falls), while #[+son][−son] punishes sonorant–obstruent onset clusters (sonority falls). Now *bn* violates *#[−syll][−son], while *rt* violates both constraints. Since constraint violations are additive in Harmonic Grammar, both violations count against the probability of *rt*, while only one violation counts against the probability of *bn*. More generally, when form **x** contains a strict superset of violations of **y** (i.e. **y** harmonically bounds **x**), any maxent HG grammar must assign lower probability to **x** than **y**. This is how maxent HG enforces gradience; analogous mechanisms apply in most competitor theories.

3.2.2 *Variation*. Like other probabilistic approaches, maxent HG is naturally suited to modelling variation. As exemplified in (10) above, essentially categorical mappings can be obtained by making the harmony difference be large between the best candidate and the next-best

https://doi.org/10.1017/S0952675715000251 Published online by Cambridge University Press

competitor. Since maxent HG allows for real-valued weights, the harmony difference can be controlled to an arbitrary degree of precision, allowing for close fits between observed and expected probabilities. However, maxent HG retains many of the restrictive properties of constraint-based grammars; for example, as noted in §3.3, maxent HG must always assign lower probability to **x** than **y** if **y** harmonically bounds **x**.

3.2.3 *Learning properties*. It is particularly in its ability to learn in the face of variation that maxent HG stands above competitor theories. The error-driven constraint demotion algorithms employed in classic OT learning (Tesar & Smolensky 1998) will not generally converge in the face of variable data. Proposals for handling variation include constraint strata (Anttila 1997), Stochastic Optimality Theory (Boersma & Hayes 2001), maxent HG (Goldwater & Johnson 2003, Hayes & Wilson 2008) and Noisy Harmonic Grammar (Goldrick & Daland 2009, Boersma & Pater to appear). Pater (2008) demonstrates that when Boersma & Hayes' (2001) Gradual Learning Algorithm was used with Stochastic Optimality Theory, it could fail to learn even a categorical pattern, owing to miscalibration of the constraint rewards/penalties (see also Magri 2012). With respect to Noisy Harmonic Grammar, Boersma & Pater (to appear) prove that 'for any nonvarying target language, Harmonic Grammar learners are guaranteed to converge to an appropriate grammar, if they receive complete information about the structure of the learning data. We also prove convergence when the learner incorporates evaluation noise'. That is, a formal proof has been given for (Noisy) Harmonic Grammar learners using the perceptron learning rule (also known as stochastic gradient ascent) when the learning data do *not* include variation (or hidden structure).[3] However, to date it has not been proven that Noisy Harmonic Grammars will converge to a relatively stable grammar when presented with language data containing variation. In contrast, it has been known for some time that log-linear models have a convex parameter space, which means that convergence to an optimal grammar is guaranteed even if the language to be learned does contain variation, provided again that the surface forms do not contain hidden structure (Della Pietra *et al*. 1997).

3.2.4 *Analytic tractability*. A case in point is when the analyst wishes only to compare two candidates out of some larger set; in this case

[3] Hidden structure is normally conceived of as structure that is present in the lexical-surface mapping, but not observable to a listener. For example, the assignment of accent/stress is often accounted for using metrical structure such as feet, which are not directly observable (see Tesar *et al*. 2003, Tesar & Prince 2003, Hayes 2004, Merchant & Tesar 2008, Jarosz 2013, Bowers 2014). For example, *banana* [bəˈnænə] has medial stress, but in theories with foot structure, this is consistent with two different output candidates, the trochaic parse [bə(ˈnænə)] and the iambic one [(bəˈnæ)nə]. Perceptual evidence alone does not determine a unique input–output mapping in many cases, thereby failing to meet the assumptions of most learning algorithms with convergence proofs.

https://doi.org/10.1017/S0952675715000251 Published online by Cambridge University Press

maxent HG has the elegant property that the log-odds of two candidates is the difference between their harmonies. This means that two candidates can be compared using only their own constraint violations and the associated weights. In contrast, Noisy Harmonic Grammar requires computationally intensive simulations to determine candidate probabilities (cf. Goldrick & Daland 2009: Table I), so that the log-odds of two candidates depends not only on their own harmonies, but also on those of all the other candidates. Relatedly, because the probability of a candidate is uniquely determined by its constraint violations in maxent HG, it is straightforward to define a LIKELIHOOD FUNCTION, which expresses the probability of the data, given the constraint weights. In combination with the convexity property mentioned in §3.2, this renders maxent HG amenable to efficient numerical methods for the constraint-weighting operations (provided that the partition function can be calculated).

In summary, maxent HG is equal or superior to other phonological formalisms on all five desiderata identified here.

3.3 Maxent phonotactic grammars

As noted earlier, a phonotactic grammar differs from the normal conception of a phonological grammar. In a 'normal' phonological grammar, we find (11).

(11) a. The grammar should pair every lexical representation /**x**/ with one or more surface representation [**y**], e.g. by indicating the probability that /**x**/ maps to [**y**].
b. In practice, the candidate set for a given input /**x**/ is finite and small: in theory, the candidates from GEN form an infinite set, but in published papers, it is customary to include around 2–6 candidates in any tableau (for discussion and critique see Bane & Riggle 2012).
c. The well-formedness of an (/**x**/, [**y**]) pair is determined by both markedness and faithfulness constraints.

However, in phonotactic grammars, we have (12).

(12) a. There is no input, and so there is nothing to be faithful to.
b. The grammar is supposed to 'score' every string in Σ^*, a countably infinite set.

Thus a phonotactic grammar is quite different from a 'normal' phonological grammar. This section shows that phonotactic grammars fall out as a special type of phonological grammar, if the perspective offered earlier on word recognition is adopted. The class of maxent phonotactic grammars can then be defined by applying this perspective to maxent HG.

Recall that, in order to handle novel words, models of word recognition need a 'special' item ω_{new} ('<new word>'), and the phonotactic grammar should distribute the probability of this item across all possible strings.

https://doi.org/10.1017/S0952675715000251 Published online by Cambridge University Press

This insight allows us to cast a phonotactic grammar as a phonological grammar in which (i) ω_{new} is the sole lexical representation; (ii) the set of candidates is the set of all possible words Σ^*. In the remainder of this section, this perspective is deployed in the context of maxent HG grammars; this is the formal underpinning for what we have been calling maxent phonotactic grammars.

Equation (9) is the fundamental equation by which maxent HG assigns probability to a candidate. In (13), subscripts have been added to explicitly represent a particular set of constraints **C** and their associated weights **w**.

(13) $Pr_{\mathbf{C},\mathbf{w}}(\mathbf{y} \mid \mathbf{x}) = \Phi_{\mathbf{C},\mathbf{w}}(\mathbf{x}, \mathbf{y}) / Z_{\mathbf{C},\mathbf{w}}(\mathbf{x})$

Since there is only a single, constant lexical representation (ω_{new}) in phonotactic grammars, it can be suppressed. Moreover, since the candidates are possible words, they are represented as ω. Substituting in this way gives (14), which defines the behaviour of maxent phonotactic grammars.

(14) $Pr_{\mathbf{C},\mathbf{w}}(\omega_i) = \Phi_{\mathbf{C},\mathbf{w}}(\omega_i) / Z_{\mathbf{C},\mathbf{w}}$
where $\Phi_{\mathbf{C},\mathbf{w}}(\omega_i) = exp(\sum_k w_k \times C_k(\omega_i))$
and $Z_{\mathbf{C},\mathbf{w}} = \sum_{\sigma \in \Sigma^*} \Phi_{\mathbf{C},\mathbf{w}}(\omega_i)$

Hayes & Wilson (2008) define PHONOTACTIC LEARNING as the process of identifying the markedness constraints that are active in distinguishing possible words from impossible ones, and inferring their weights.[4] They then propose a theory for accomplishing phonotactic learning using maxent phonotactic grammars. The core of the idea is to regard the existing lexicon $\Lambda = \{\omega_1, \omega_2, \ldots, \omega_n\}$ as a sample from the phonotactic grammar having constraints $\mathbf{C} = \{C_k\}_{k=1\ldots m}$ and weights $\mathbf{w} = \{w_k\}_{k=1\ldots m}$. The probability of the lexicon is the product of the probabilities of generating each word in it, where the probability of a word-form ω_i is calculated according to (14). Hayes & Wilson (2008) propose that the 'best' grammar (**C, w**) can be found using *maximum a posteriori* estimation, incorporating a 'prior' so that the learned grammar reflects a compromise between making the grammar simple and fitting the data well.

There are two technical obstacles to learning in a completely general way. The first is that the partition function $Z_{\mathbf{C},\mathbf{w}}$ is defined over an infinite string

[4] Note that Tesar and colleagues use the term phonotactic learning with a related but distinct meaning. According to this, phonotactic learning pertains to learning the pattern of contrast and licensing from the surface forms of a language (e.g. Merchant & Tesar 2008). This differs from the sense of phonotactic learning used here in that it involves learning a partial ordering over both markedness and faithfulness constraints. An example may serve to illustrate the contrast. In tableau (6) above, highly prioritised *COMPLEX and MAX drive the mapping /tra/ → [tera]. If the Hayes & Wilson learner were exposed to surface data from this language, it would learn that #CC sequences do not occur (formalised as a very strong weight on *COMPLEX). However, a Tesarian learner would learn something stronger: there is no vowel ~ zero contrast in the environment #C__C. This crucially entails that at least one faithfulness constraint (MAX, DEP) is prioritised lower than *COMPLEX. The Hayes & Wilson learner cannot learn this, because it does not include faithfulness constraints.

https://doi.org/10.1017/S0952675715000251 Published online by Cambridge University Press

set, so it cannot be calculated exactly by brute force. The second obstacle is that there may be billions or even trillions of logically possible markedness constraints, and it is not computationally feasible to include every one. Hayes & Wilson (2008) propose to solve the latter problem with an iterative procedure, in which the learner begins with an 'empty' grammar, and then considers constraints one by one, retaining those which increase probability sufficiently. The order in which constraints are tried is governed by heuristics, about which more will be said later. For now, this paper focuses on the first problem, calculating the partition function.

Hayes & Wilson's method for calculating the partition function makes use of finite-state Optimality Theory (Eisner 2002, Riggle 2004, 2009). Each constraint in the grammar is coded as a simple weighted finite-state machine, and the grammar is constructed by 'intersecting' all of the constraints (for exposition see Riggle 2004). This results in a very large finite-state machine, each of whose states represents a violation vector corresponding to a (sometimes infinite) class of strings. The expected violation counts for each constraint in the grammar can be summed using Eisner's 'expectation semi-ring'. From these expected counts it is straightforward to compute the partition function.

As an implementational matter, Hayes & Wilson (2008: 389) choose to impose a finite upper bound on the string set that they sum the partition function over:

> Instead of calculating expected values exactly, we approximate them by examining only the strings in [Σ^*] that are no longer than the longest string in the learning data D. This is a finite – albeit exponentially large – subset of [Σ^*].

In other words, they impose an 'artificial' upper limit on how long possible words can be, assigning a probability of 0 by fiat to all strings over this limit. However, careful inspection of the original Eisner (2002) reference reveals that the string set does not need to be finite. Rather, the requirement is that the partition function have a finite value, or equivalently, that the accumulated weight of all possible paths through the machine be finite. This condition is trivially satisfied when there is a finite upper bound on string length, since the finite sum of finite values is finite. It is also possible to have a finite partition function over an infinite string set. However, it is not guaranteed. This is not just a theoretical problem with maxent phonotactic grammars, but also one that occurs in practice, as the next section demonstrates.

4 Underpenalisation, or the infinity problem

4.1 Precedents

Infinity has beguiled and bedevilled linguists from the earliest days of generative grammar. In an early and influential article, Chomsky (1956: 115) addresses the issue of whether sentences are bounded:

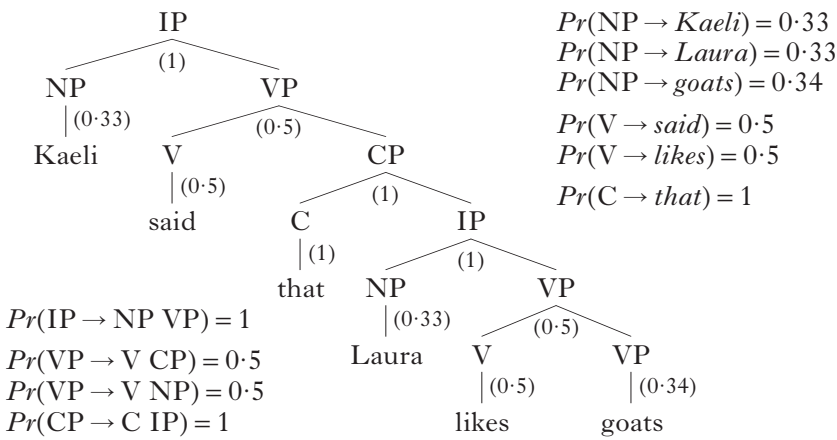
https://doi.org/10.1017/S0952675715000251 Published online by Cambridge University Press

> We might avoid this consequence by an arbitrary decree that there is a finite upper limit to sentence length in English. This would serve no useful purpose, however. The point is that there are processes of sentence formation that this elementary model for language is intrinsically incapable of handling. If no finite limit is set for the operation of these processes, we can prove the literal inapplicability of this model. If the processes have a limit, then the construction of a finite-state grammar will not be literally impossible (since a list is a trivial finite-state grammar), but this grammar will be so complex as to be of little use or interest.

Though Chomsky was primarily concerned with the syntactic organisation of sentences rather than the generation and recognition of new words, the issue is essentially the same. In some languages, it is in principle possible to construct an acceptable item that is longer than any given finite bound. Theories which simply stipulate a finite upper bound are insufficiently expressive.

The issues discussed in this paper have a close parallel in the history of the development of probabilistic context-free grammars (PCFGs). Intuitively, a PCFG may be thought of as a grammar that generates trees with probabilities. An example tree and the generating grammar are shown in (15). A PCFG consists of a list of category labels, some terminal labels and a list of rewrite rules with associated probabilities (for in-depth coverage see Chi & Geman 1998, for applications in laboratory phonology Coleman & Pierrehumbert 1997 and for a reader-friendly description Daland *et al*. 2011).

(15)



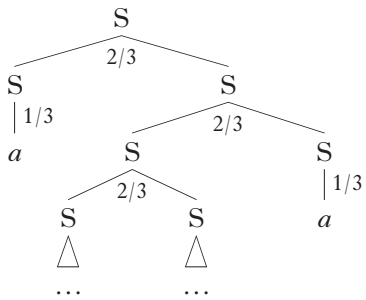
The probability of a derivation is simply the product of the probabilities of all the rewrite rules used in the derivation. For example, the probability of the derivation in (15) is $1 \times 0{\cdot}33 \times 0{\cdot}5 \times 0{\cdot}5 \times 1 \times 1 \times 1 \times 0{\cdot}33 \times 0{\cdot}5 \times 0{\cdot}5 \times 0{\cdot}34 = 0{\cdot}002314125$. The probability of a string is the sum of the probabilities of all derivations that yield the string. Since derivation (15) is the only

https://doi.org/10.1017/S0952675715000251 Published online by Cambridge University Press

one which can yield *Kaeli said that Laura likes goats*, the probability the PCFG assigns to this string is also 0·002314125.

As early as 1963, it was noted that probabilistic branching processes might have a non-zero probability of not terminating (Harris 1963). An example is given in (16).

(16)



$Pr(S \rightarrow S\ S) = 2/3$

$Pr(S \rightarrow a) = 1/3$

In the case of (16), it is more likely than not that a given S will expand to more than one S daughter. This means the probability that every S will terminate in a finite number of *a*'s is less than 1. Put another way, some of the probability is 'wasted' on infinite trees (which cannot be observed). When the sum over all observable trees is less than 1, the probability distribution is called INCONSISTENT.

More than a decade later, Grenander (1976) gave a mathematical characterisation of when PCFGs assign a consistent distribution. In essence, every label needs to derive less than one copy of itself (and other self-deriving labels) on average. For example, in the grammar in (16), a given S is expected to give rise to 4/3 immediate S daughters ($2 \times (2/3) + 0 \times (1/3) = 4/3$), meaning that the number of S's is expected to grow on average with every rule application. More formally, Grenander constructs a matrix whose (i, j) entry indicates the probability that category label j will expand into category label i. The PCFG is consistent if and only if a value called the PRINCIPAL EIGENVALUE is less than or equal to 1. A very similar condition is shown for maxent phonotactic grammars in Theorem 3.

More than two decades after Grenander's work, Chi & Geman (1998) proved that if a PCFG is trained on finite (observable) data according to the principle of maximum likelihood, the consistency condition will always be met. The reason for this is relatively simple: in all of the observable data, each symbol eventually leads to less than one copy of a self-deriving symbol (otherwise it would give rise to an infinite tree, which cannot be observed). Maximum-likelihood training causes the estimated likelihood of a rule expansion to match the observed rate of symbol expansion. In other words, the training data does include infinite trees, so the PCFG which is learned approximates this property.

The problem of learning maxent phonotactic grammars is conceptually similar. The first theoretical contribution of this paper is to point out, as Harris (1963) does for PCFGs, that there is at least the potential for an

https://doi.org/10.1017/S0952675715000251 Published online by Cambridge University Press

infinity problem with maxent phonotactic grammars. The next theoretical contribution is to give necessary and sufficient conditions for avoiding the infinity problem with maxent phonotactic grammars, as Grenander (1976) does for PCFGs. The final theoretical contribution is to motivate the conjecture that the infinity problem will not arise when maxent phonotactic grammars are trained on observable samples, provided that they are initialised with a *STRUCT constraint that is given sufficient weight; this is analogous to the proof that Chi & Geman (1998) offer for PCFGs.

4.2 Examples

This section gives three examples of maxent phonotactic grammars in which the partition function is not finite, with increasing degrees of phonological realism.

The first example is the simplest one, and is included because it illustrates the nature of the problem so simply. Suppose the alphabet consists of a single segment; $\Sigma = \{a\}$. Then Σ^* consists of strings of the form a^n, where $n \geq 0$. Further, let the grammar be 'empty' (the initial state that Hayes & Wilson 2008 employ in their software implementation). It follows that the weighted sum of the constraint violations of every form a^n is 0, $H[a^n] = 0$ for all n. Then the well-formedness measure of each such a^n is $exp(H[a^n]) = 1$. The partition function is the sum of the well-formedness values over all such strings, $Z = \sum_{n \geq 0} exp(H[a^n]) = \sum_{n \geq 0} 1$. Clearly, this sum is not finite.

The reader may rightly protest that no natural phonology includes an inventory with only one phoneme. But adding more phonemes makes the infinity problem worse, not better. To see this, consider an alphabet with just one more symbol, $\Sigma = \{C, V\}$, but with the same 'empty' grammar as in the previous paragraph. Let us define Z_k as the 'partial' partition function, summing over all strings of length k or less. Clearly, $Z_k \rightarrow Z$, as $k \rightarrow \infty$. Moreover, Z_k is equal to the number of such strings (because the well-formedness measure of each such string is 1). Then calculating Z_k is a simple matter of counting strings. The number of strings in Z_k can be counted by dividing them into sets by their length, i.e. the set of strings of length 0, the set of strings of length 1, the set of strings of length 2, etc. We use the notation Σ^j for a set of string of length exactly j; $|\Sigma^j|$ is the cardinality of this set. For example, $Z_2 = |\Sigma^0| + |\Sigma^1| + |\Sigma^2|$. Because $|\Sigma^j| = |\Sigma|^j$, it follows that Z_k is a geometric series with common ratio $|\Sigma|$. This makes perfect sense – from every string of length j, we can create exactly $|\Sigma|$ strings of length $j+1$, by simply adding one segment from Σ. Thus the total number of strings up to a certain length j should be exponential in j. From this it follows that $Z_k \rightarrow \infty$, so the partition function is not finite.

The reader may think that these artificial examples have little to do with real phonology, so the infinity problem would not occur if a maxent HG were provided with 'real' constraints and natural language data. But that is false, as illustrated by the final example, the Wargamay grammar that Hayes & Wilson discuss extensively in their paper. With some careful

https://doi.org/10.1017/S0952675715000251 Published online by Cambridge University Press

inspection, it is possible to show that [ˈbamba] is harmonically 'perfect' according to this grammar, i.e. is assigned a harmony of 0. Moreover, the possible words [ˈbamba,bamba] and [ˈbamba,bamba,bamba] are also harmonically perfect. The relevant constraints needed to verify this claim are distributed over several pages and tables, each dealing with orthogonal aspects of the grammar (Hayes & Wilson 2008: 414, 416, 417, 419, 432–432). First, the 'onset' consonant [b] is unpenalised both in absolute word-initial position and intervocalically. Second, [mb] is among the 40 medial consonant clusters described as licit both by Hayes & Wilson and by their reference grammar. Third, the authors explicitly state (2008: 418) that items of the form (ˈσσ,σσ,σσ) incur no pure metrical penalties; they do not list any constraints that penalise longer sequences which satisfy the initial main stress trochaic footing system. Finally, there are numerous constraints regulating the local co-occurrence of [+high], [−back], [+stressed] and [−stressed) vowels, but none specifically or generically targets a stressed [a] followed by [mb] and unstressed [a], or unstressed [a] followed by [b] and stressed [a]. Therefore, every string consisting of initial [ˈbamba] and followed by 0 or more repetitions of [,bamba] is unpenalised by this grammar. Then every string of the form ba^n is assigned a harmony of 0, where b = [ˈbamba] and a = [,bamba]. This example is therefore of the same character as the highly idealised one presented above: $H[ba^n] = 0$ for all $n \geq 0$, and so the partition function is not finite. This is in spite of the fact that the alphabet and the markedness constraints are derived directly from a natural language phonology.

The essential problem in these three examples is UNDERPENALISATION. Because Σ^* contains all strings over Σ, it contains some which consist of arbitrary repetitions of relatively well-formed subsequences. In the absence of any specific constraints (the CV example above), all subsequences are well-formed. But even in the presence of many constraints (the Wargamay example above), the infinity problem arises if there is even *one* well-formed subsequence that may repeat indefinitely. However, the CV example above also provides some indication of a solution. If the number of strings is exponential in their length, perhaps some penalty can be applied which scales with the length. The next section formalises this idea.

5 The solution to the underpenalisation problem

As demonstrated in the previous section, underpenalisation leads to problems in maxent phonotactic grammars over Σ^*. Relatively well-formed subsequences can be concatenated to yield relatively well-formed sequences of arbitrary length – just as *supercalifragilisticexpialidocious* consists of numerous well-formed subparts. One solution to this problem is the one adopted by Hayes & Wilson (2008) – the imposition of an arbitrary maximum word length. However, this solution is empirically inadequate and theoretically undesirable, for the reasons discussed at the beginning

https://doi.org/10.1017/S0952675715000251 Published online by Cambridge University Press

of this article. Luckily, it turns out that there is a principled solution to underpenalisation – one can ensure that long sequences are appropriately penalised, by incorporating a simple penalty that scales with the length of a segment.

Prince & Smolensky (1993) discuss a constraint they call *STRUCT, which incurs one violation mark for each overt segment in an output form. Thus, the number of *STRUCT violations of a string ω is simply the length of the string $|\omega|$. The total contribution of the *STRUCT violations to ω's harmony is $w_{*\text{STRUCT}} \times |\omega|$, since each violation has a weight of $w_{*\text{STRUCT}}$. The effect is to reduce the well-formedness measure of ω by $exp(w_{*\text{STRUCT}} \times |\omega|)$. This penalty is geometric in the length of $|\omega|$, which can be seen by rearranging terms following the laws of exponents: $exp(w_{*\text{STRUCT}} \times |\omega|) = exp(w_{*\text{STRUCT}})^{|w|} = p^{|w|}$ (where $p = exp(w_{*\text{STRUCT}})$). Theorem 2 shows that the partition function is guaranteed to be finite if $w_{*\text{STRUCT}} < -ln\ |\Sigma|$, or equivalently, if $p < 1/|\Sigma|$. Intuitively, this makes sense: the underpenalisation problem occurs because the number of possible strings grows geometrically as a function of length, with common ratio $|\Sigma|$. Underpenalisation can be eliminated by imposing a length penalty, which reduces the probability 'more' for each extra segment, i.e. by any factor $p < 1/|\Sigma|$.

In practical terms, this means that to avoid the underpenalisation problem, researchers should always include a *STRUCT constraint in maxent phonotactic grammars, and should give this constraint an initial weight satisfying the inequality $|w_{*\text{STRUCT}}| > ln\ |\Sigma|$. Note that Hayes & Wilson's software uses positive weight values that are later negated, while other formulations may just use negative weight values. The next subsection discusses why Hayes & Wilson's software never includes the *STRUCT constraint in the grammar it finds, and then reports several published examples where *STRUCT has proved useful in maxent phonotactic grammars.

5.1 *STRUCT in maxent phonotactic grammars

Recall that Hayes & Wilson (2008)'s learner includes a constraint-induction procedure, which is distinct from the 'weight-setting' procedure. Weight-setting is a learning process in which, *given* some constraints, the best constraint weights are found (where 'best' is defined as maximising the likelihood of observing the lexicon that was actually input to the model). Constraint induction is the process by which the model searches a large space of possible constraints.

In Hayes & Wilson's learner, the constraint-induction process includes two search biases – constraints are searched in order of accuracy, and among constraints that are roughly equivalent in accuracy, more 'general' constraints are considered earlier (for details see Hayes & Wilson 2008: 394). The default is for the search process to terminate after a certain number of constraints have been included in the grammar (e.g. 100). Clearly, constraints cannot be considered for inclusion in the

https://doi.org/10.1017/S0952675715000251 Published online by Cambridge University Press

grammar if the search process terminates before it reaches them. Since the search order is determined by accuracy, the practical effect of the 'accuracy bias' is to ensure that constraints with poor accuracy are never considered.

Hayes & Wilson (2008) define accuracy as the ratio of observed violations of the constraint in the training lexicon to the number that is expected according to the preexisting grammar. That is, it favours constraints which do not penalise the training data, rather than ones which distinguish training items from untrained items. Since *STRUCT penalises every segment in the training data, its accuracy value will be among the worst possible. Therefore, Hayes & Wilson's software will never consider this constraint 'early' in the constraint-search process. In practice, the software might examine several thousand constraints from a pool of several hundred millions. The constraint-induction process will always terminate before it even reaches the *STRUCT constraint, and this fact is independent of whether the *STRUCT constraint is actually useful. There are numerous reasons to think it is.

Wilson & Obdeyn (2009) report on a variant of the Hayes & Wilson software that differs primarily in the constraint-induction process. Rather than searching for constraints according to accuracy, Wilson & Obdeyn's model uses a criterion they call GAIN. The GAIN of a constraint is defined as the change in the log-likelihood of the training data when the constraint is added to the current grammar. Wilson & Obdeyn's learner induces constraints by checking a pool of constraints, and adding whichever one has the highest GAIN. Colin Wilson (personal communication) indicates that he is unable to recall a single instance in which the grammar resulting from running this model did not contain *STRUCT.

It seems likely that the accuracy-first search bias contributes more generally to an empirical problem with the Hayes & Wilson learner in particular. Hayes & White (2013) document a phenomenon they call 'accidentally true constraints'. This concerns the tendency of the Hayes & Wilson learner to find constraints that happen to be unviolated in the training data, but which appear not to figure in human grammars of the same language. The authors show this by selecting ten 'natural' constraints and ten 'unnatural' ones that receive moderate weights (−3 to −5) from an English phonotactic grammar, where naturalness is defined on the basis of phonetic and/or typological support. They then conducted a non-word acceptability judgement task with a set of forms that violate either the 'natural' or 'unnatural' constraints, finding that the 'natural' constraints appear to be much more strongly represented in the English speakers' judgements. While the article's title suggests that they interpret these results in terms of (phonetic) naturalness, they acknowledge that they may instead be a statistical artefact of the constraint-induction process in the Hayes & Wilson learner. Specifically, Hayes & White's note 23 states:

> [The Wilson & Obdeyn] learner uses the principle of 'gain' (Della Pietra *et al*. 1997: 1) to select constraints. In a preliminary examination of this

modified system, we found that the constraints it selects were indeed more general and less idiosyncratic than those chosen by the Hayes/Wilson (2008) learner; more specifically, it learned none of our 10 Unnatural constraints. However, it may still be learning *different* ones: for instance, it posited a constraint banning stressed lax vowels before coronal codas as well as a constraint against word-final sonorants. Our testing was tentative, owing to memory limitations, and more serious evaluation of the revised model awaits further research.

The *Struct constraint is one example of a 'more generic' constraint. It may explain the variation that prompted Hayes & Wilson's search procedure to induce 'accidentally true' constraints.

Another situation in which the *Struct constraint is useful is in the generation of artificial lexicons. Daland *et al*. (2014) describe a series of simulations in which the phonotactic structure of artificial languages was manipulated to study the effect on word segmentation. Specifically, artificial maxent phonotactic grammars were chosen so as to instantiate varying degrees of restrictiveness in syllable structure (e.g. only CV *vs*. CV(C)). Artificial lexicons were generated by regarding the lexicon as a sample from the probability distribution generated by the grammar. The *Struct constraint proved necessary for controlling word length during the generation process. More specifically, PhoMEnt, an open source publicly available maxent phonotactic grammar implementation, was used, which currently has a finite upper bound of only six segments on string length.[5] In the absence of *Struct, it turned out that nearly all of the sampled words had six segments, the maximum. The reason for this is very straightforward – the number of phonotactically acceptable words with six segments is many times greater than the number of phonotactically acceptable words of shorter lengths. If all acceptable words are assigned the same well-formedness value, more long words will be chosen, simply because they make up a greater proportion of the pool of candidates. *Struct was therefore included and its weight manipulated in integer steps to control the average length of the sampled lexicons. For most grammars, the best segmentation was found when *Struct was assigned a weight of −3, which is in excellent agreement with Theorem 2: Daland *et al*. (2014) use a segmental alphabet $\sum$ consisting of 18 segments, so $-ln\ |\sum| = -ln\ 18 = -2{\cdot}89$.

5.2 Necessary *vs*. sufficient conditions

Theorem 2 gives sufficient conditions to guarantee a finite partition function (and therefore a proper probability distribution) in maxent phonotactic grammars. That is, if $w_{*\text{Struct}} < -ln\ |\Sigma|$, then a proper probability distribution is guaranteed. However, it is possible for a maxent phonotactic grammar to have a finite partition function even when this condition is not met. For example, the onset grammar that is reported in Hayes &

[5] https://github.com/rdaland/PhoMEnt.

https://doi.org/10.1017/S0952675715000251 Published online by Cambridge University Press

Wilson (2008) has a finite partition function. The key is that the onset grammar strictly penalises plateaus and decreases in sonority. So a sequence like *trtrtr* contains two *rt* sonority falls, which are punished by a variety of sonority-regulating constraints (see also Hayes 2011). More generally, the sequence $(tr)^n$ contains $n-1$ sonority falls, and is punished for each of them. In this case, the number of violations grows with the sequence length for any sequence in the grammar. The underpenalisation problem does not arise, even though this grammar does not include a *STRUCT constraint. It would be desirable to be able to characterise not only sufficient, but also necessary, conditions to avoid underpenalisation. Theorem 3 does exactly that. The final result is an eigenvalue condition similar to the one given by Grenander (1976) for consistency of PCFGs. As the proof is rather technical, its intuition is explained in the online supplementary materials.

5.3 Conjecture: finiteness is guaranteed when training on finite samples

Just as Chi & Geman (1998) proved that a PCFG derived from maximum likelihood estimation will be consistent, this section offers a rationale for the conjecture that a maxent phonotactic grammar trained on a lexicon will avoid underpenalisation as long as a *STRUCT constraint is included with sufficient weight in the initial state.

As shown by Della Pietra *et al*. (1997), the likelihood of the training data is maximised with respect to a given weight w_i when the number of expected violations of C_i (over the entire candidate set) equals the number of observed violations of C_i (over the training set). We can obtain a more intuitive version of this equation by dividing both sides by the number of words in the training set. The total number of violations of *STRUCT in the training data is the total number of segments in all words; dividing this by the number of such words gives us the average length of a word in the training set. The expected number of violations of *STRUCT in the candidate set is simply the number of words in the training set multiplied by the average number of *STRUCT violations over the candidate set; dividing this by the number of words in the training set gives us the expected average length of a word according to the current grammar.

Suppose that the grammar is initialised so that the partition function is finite (or, in other words, the expected length of a word is finite). At this initialisation stage, the weight of *STRUCT might either be too low or too high. If it is too low, the expected average word length is too long, and the weight-setting procedure will simply increase the weight of *STRUCT until the expected average word length matches the observed average word length. If it is too high, the expected average word length is too short, and the weight-setting procedure will simply decrease the weight of *STRUCT until the expected average word length matches the observed average word length. As long as the weight-setting algorithm is initialised

https://doi.org/10.1017/S0952675715000251 Published online by Cambridge University Press

with reasonable starting weights, there should be no chance of 'overshoot', which would make the weight of *STRUCT so low that the partition function ceases to be finite. Adding new constraints to the grammar cannot decrease the penalty assigned to any string (since constraints only punish, never reward), so adding a constraint to a grammar with a finite partition function will not cause the partition function to become non-finite.

In short, the training data will always exhibit the property that it has a observable and finite average length. Maximum likelihood estimation with maxent phonotactic grammars will tend to cause the weight of *STRUCT to settle at whatever value makes the expected average word length match the observed average word length. As long as the prior does not force the weight of *STRUCT to be too low, underpenalisation should be avoided with the standard maximum *a posteriori* training.

6 Discussion and conclusions

6.1 Summary of major points

This article considers the properties that a phonotactic grammar should have. It began by discussing the motivations for a phonotactic grammar: the need to assess the well-formedness of potential words is motivated by the fact that humans can do so in well-formedness judgement tasks, and the need to link well-formedness with other cognitive modules, notably speech perception and word recognition. Jointly, these properties call for a probabilistic model of phonotactic knowledge.

An especially promising grammatical formalism is maxent HG, which is a stochastic variant of Optimality Theory, with weighted rather than ranked constraints; it stands out from other stochastic approaches particularly in formal guarantees of optimal learning of languages with variation. Maxent HG can be and has been adapted for modelling phonotactic grammars; the result is called maxent phonotactic grammars here to distinguish them from regular maxent HG grammars.

Like other classes of probabilistic models such as PCFGs, maxent phonotactic grammars require special care when it comes to potentially unbounded sequences. The problem, described here as underpenalisation, is that relatively well-formed subsequences may be concatenated together to form arbitrarily long sequences with no ill-formed subparts. The word *supercalifragilisticexpialidocious* is an example of just such a possible word. In general, the number of such words may grow exponentially with word length. If these words are all assigned a 'perfect' harmony score of 0, the grammar's partition function will not be finite, meaning that the probabilities of individual strings will not be well-defined. One solution to this problem that has been used in the literature is to impose a strict upper limit, such as ten segments, on the length of words to be considered. But this approach is empirically inadequate and theoretically undesirable.

After documenting the underpenalisation issue, this paper has proposed a solution: the *STRUCT constraint. Formal proofs are given of sufficient and necessary conditions which allow a researcher to steer between a hard upper bound on word length and underpenalisation. Provided that a *STRUCT constraint is included and initialised with sufficient weight, a maxent phonotactic grammar may allow for possible words of unbounded length, but penalise them enough to avoid an ill-defined probability distribution (by keeping the partition function finite). The practical implications for researchers are as follows: (i) when using the Hayes & Wilson phonotactic learner or other maxent phonotactic grammar software, always include a *STRUCT constraint and set its initial weight to be greater than $ln\,|\Sigma|$; (ii) if developing one's own maxent phonotactic grammar implementation, either force the grammars to always include a *STRUCT constraint, or build in the test described in Theorem 3.

The remainder of the discussion briefly considers two issues.

6.2 Comments on arguments against *STRUCT

Gouskova (2003) constitutes a concerted attack upon constraints whose sole purpose is to enforce economy effects, i.e. the general linguistic preference for less structure. The constraint *STRUCT is a canonical example – without exception, it prefers fewer segments. Gouskova's critique is couched within the framework of what might be called 'classic OT' (e.g. Prince & Smolensky 1993), in which a categorical grammar regulates input–output mappings with constraints that are totally ranked, rather than weighted. It is argued in this section that Gouskova's critique does not apply to *STRUCT in the maxent HG phonotactic learning setting.

Gouskova considers two types of argument against this kind of economy constraint. First, she proposes that analyses which include such constraints do not actually need it; in support, she gives an impressive array of examples that have been analysed with a *STRUCT-like constraint (e.g. *V, which penalises all vowels), and then provides an alternative analysis in which the economy effect falls out from more specific, independently motivated constraints. Gouskova's second argument is that the inclusion of *STRUCT-like constraints in a grammar leads to improper predictions. The general character of this argument can be illustrated with the most extreme case: if *STRUCT were the top-ranked constraint in an OT grammar, it would predict the null output to be the winning candidate for every input; in other words, no-one would ever include any overt material in their utterances (i.e. speak). Gouskova considers several less extreme cases that are like this in logical character, and argues that several typologically predicted languages do not occur. She concludes that constraints like *STRUCT are both unnecessary and actually harmful, so they should be excluded from the theory.

The most fundamental point to make is that Gouskova is working in a different framework, with different properties and assumptions. In classic OT, there is an input–output relationship that is regulated by

constraints ranked with a total order. In this framework, faithfulness constraints achieve the work of preventing unboundedly long repetitions of a subsequence in surface forms. High-ranking markedness constraints may require multiple epenthesis operations (e.g. vowel prothesis to repair an onset cluster and glottal stop prothesis to repair the resulting onsetless vowel, as in /ktub/ → [ʔuktub]). But in these cases, each application permanently resolves the violation that triggered it (it could not be optimal otherwise). The number of such violations must be finite (because underlying forms are only finitely long, and there are only finitely many constraints that could be violated). Therefore, high-ranking markedness constraints can trigger at most a finite number of epenthetic repairs. In the phonotactic learning setting, there is no underlying representation to be faithful to, as the goal is to spell out the well-formedness of potential words (which have not yet been learned). Therefore, faithfulness constraints cannot eliminate unboundedly long sequences. Since markedness constraints are the only constraints in phonotactic learning, they are the ones which must rule out possible words of an arbitrary length (cf. *supercalifragilisticexpialidocious*). Gouskova's essential point is that *STRUCT is not needed in classic OT because faithfulness constraints derive economy effects; this point does not hold for phonotactic learning, where there is no faithfulness.

As the field now possesses a much better theoretical understanding of phonological acquisition than it did at the time of Gouskova's (2003) work, it is worth thinking through the acquisition scenario from the child's point of view. It seems reasonable to suppose that the child is initially exposed to a language in which input from caretakers contains segmental content, and moreover that comparatively short words have a higher type frequency, at least during the early stages of phonological acquisition (Hayes 2004). According to the theory described here, it is statistically inappropriate for the child to begin with an 'empty' phonotactic grammar, as such a grammar does not assign a well-defined probability distribution over the string-space. However, the child might begin with the next best thing, a phonotactic grammar which merely states that longer words are worse – a prior bias which is empirically supported by the input (specifically, one in which $w_{*\text{STRUCT}} < -ln\ |\sum|$). As the child begins to recognise, encode and learn word-forms, their phonotactic grammar elaborates. Sequences which systematically fail to occur in the input acquire their own more specific markedness constraints, while sequences which do occur drive down the weight of *STRUCT and drive up the weight of these more specific constraints. After a large number of word-forms are learned, the grammar will have developed to a considerable extent; the work that was formerly done by *STRUCT will be mostly replaced by more specific and more accurate constraints. This will presumably leave *STRUCT with a very low weight relative to inviolable markedness constraints and other

constraints that drive active alternations. As the child begins to acquire faithfulness constraints, the weights of remaining markedness constraints may move slightly, but *STRUCT might be left with a low weight – just enough to match the observed average word length with the expected average word length.

6.3 Implications for maxent HG: the consequences of constraint omission

A reviewer raises the issue of whether the underpenalisation problem potentially arises with all maxent HG grammars, or only with maxent phonotactic grammars. This question is related to the discussion in the preceding section, and the answer is that there may be an underpenalisation problem even with 'normal' maxent HG.

As Gouskova (2003) points out, economy effects in classical OT are normally derived by faithfulness constraints. For example, a single instance of vowel epenthesis may be motivated by a high-ranking markedness constraint, such as repair of a triconsonantal cluster, illustrated in (17) by the choice of candidate (b) over (a).

(17)

/aptka/	*CCC	DEP(V)	DEP(C)
a. aptka	*!		
☞ b. apitka		*	
c. apiitka		**!	
d. apitika		**!	
e. apitkani		**!	*
…			

In general, multiple epentheses will not repair the original markedness violation better than a single epenthesis operation (c)–(e), but they do incur additional DEP violations. Of course, if the DEP constraints were omitted from (17), there would be nothing to stop the speaker from choosing one of the 'wrong' candidates (c)–(e). And according to Prince & Smolensky (1993), GEN presents the speaker with an infinite number of wrong candidates. Even if it is at the very bottom of the constraint hierarchy, DEP plays an important role in weeding out uneconomical candidates.

In maxent HG, constraints are weighted rather than strictly ranked. Tableau (18) recapitulates (17), but cast in the framework of maxent HG, with a weight of 0 assigned to both DEP constraints.

https://doi.org/10.1017/S0952675715000251 Published online by Cambridge University Press

(18)

/aptka/	C_1: *CCC $w_1 = -5$	C_2: Dep(V) $w_2 = 0$	C_3: Dep(C) $w_3 = 0$	H	Φ	*Pr*
a. aptka	1			−5	e^{-5}	<*undef*>
b. apitka		1		0	1	<*undef*>
c. apiitka		2		0	1	<*undef*>
d. apitika		2		0	1	<*undef*>
e. apitkani		2	1	0	1	<*undef*>
…				0	…	<*undef*>

Since unbounded amounts of epenthesis are unpenalised, there is a countably infinite number of candidates which satisfy the important markedness constraint equally well. Underpenalisation may also occur in this circumstance. Fortunately, the same intuitions which govern underpenalisation in maxent phonotactic grammars apply in this more general setting. Underpenalisation cannot occur if sufficient weight is allocated to Dep and/or Struct constraints. It is straightforward to adapt Theorem 2 to obtain a guaranteed finite partition function over an infinite candidate set in this case (but a formal proof is not given). In effect, this point is merely a formalisation of something that is already conventional wisdom in the field: the candidate set should not contain SRs that both differ from the UR and are underpenalised for being so.

6.4 Conclusion

In summary, it has long been a tenet of constraint-based phonology that the candidate set is infinite (Prince & Smolensky 1993). This has not caused practical problems for normal working phonologists, as typical analyses only consider a finite subset of all possible candidates. However, this paper has argued in some detail that the infinite candidate set can and should be considered in the case of phonotactic grammars. The essence of the argument is that speakers can and do produce novel words of arbitrary length, that our speech-processing systems have to be equipped to handle this and that a probabilistic phonotactic grammar plays a central role in doing so. Maximum entropy Harmonic Grammar is a particularly promising phonological formalism, owing to its attractive learning guarantees, and it can be seamlessly adapted for phonotactic grammars, as demonstrated here and in Hayes & Wilson (2008). However, existing implementations of maxent phonotactic grammars have imposed a hard upper limit on the length of possible words, which is undesirably restrictive. Without such a hard upper limit, there is the opposite risk: the infinite candidate set may contain sequences consisting of arbitrarily many relatively well-formed subsequences, which might be underpenalised. In this case, a maxent phonotactic grammar can still characterise the relative well-formedness of words, but it cannot assign a well-defined probability distribution. Besides documenting this problem, this paper has proposed a solution: include the constraint *Struct. Theorems 2 and 3 give necessary and

https://doi.org/10.1017/S0952675715000251 Published online by Cambridge University Press

sufficient conditions for maxent phonotactic grammars to avoid underpenalisation. The body of the text gives motivation for the conjecture that an underpenalisation problem will not occur provided that the grammar is initialised with a *STRUCT constraint of sufficient weight, and the default weight-setting algorithm (maximising the likelihood of the training data) is followed. In plain language, maxent HG is an excellent tool for phonological analysis, but this paper points out there is a subset of the parameter space where the tool does not function as intended; this paper also shows how to avoid that regime, and conjectures that the bad regime will be avoided in normal usage so long as grammars are initialised with a *STRUCT constraint of sufficient weight.

REFERENCES

Anttila, Arto (1997). Deriving variation from grammar. In Frans Hinskens, Roeland van Hout & W. Leo Wetzels (eds.) *Variation, change and phonological theory*. Amsterdam & Philadelphia: Benjamins. 35–68.

Baayen, R. Harald (2001). *Word frequency distributions*. Dordrecht: Kluwer.

Baayen, R. Harald & Robert Schreuder (2000). Towards a psycholinguistic computational model for morphological parsing. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* **358**. 1281–1293.

Bane, Max & Jason Riggle (2012). Consequences of candidate omission. *LI* **43**. 695–706.

Boersma, Paul & Bruce Hayes (2001). Empirical tests of the Gradual Learning Algorithm. *LI* **32**. 45–86.

Boersma, Paul & Joe Pater (to appear). Convergence properties of a Gradual Learning Algorithm for Harmonic Grammar. In McCarthy & Pater (to appear).

Bowers, Dustin (2014). Balancing leveling and composite URs. Paper presented at Phonology 2014, MIT.

Chi, Zhiyi & Stuart Geman (1998). Estimation of probabilistic context-free grammars. *Computational Linguistics* **24**. 299–305.

Chodroff, Eleanor & Colin Wilson (2014). Phonetic vs. phonological factors in coronal-to-dorsal perceptual assimilation. Paper presented at LabPhon 14: the 14th Conference on Laboratory Phonology, Tokyo.

Chomsky, Noam (1956). Three models for the description of language. *IRE Transactions on Information Theory* **2:3**. 113–124.

Chomsky, Noam & Morris Halle (1965). Some controversial questions in phonological theory. *JL* **1**. 97–138.

Chomsky, Noam & Morris Halle (1968). *The sound pattern of English*. New York: Harper & Row.

Coady, Jeffry A. & Julia L. Evans (2008). Uses and interpretations of non-word repetition tasks in children with and without specific language impairments (SLI). *International Journal of Language Communication Disorders* **43**. 1–40.

Coetzee, Andries W. & Shigeto Kawahara (2013). Frequency biases in phonological variation. *NLLT* **31**. 47–89.

Coetzee, Andries W. & Joe Pater (2011). The place of variation in phonological theory. In John Goldsmith, Jason Riggle & Alan Yu (eds.) *The handbook of phonological theory*. 2nd edn. Malden, Mass. & Oxford: Wiley-Blackwell. 401–431.

Coleman, John & Janet B. Pierrehumbert (1997). Stochastic phonological grammars and acceptability. In John Coleman (ed.) *Proceedings of the 3rd Meeting of the*

ACL Special Interest Group in Computational Phonology. Somerset, NJ: Association for Computational Linguistics. 49–56.
Daland, Robert, Benjamin Börschinger & Abdellah Fourtassi (2014). On lexical phonotactics and segmentability. Paper presented at LabPhon 14: the 14th Conference on Laboratory Phonology, Tokyo.
Daland, Robert, Bruce Hayes, James White, Marc Garellek, Andrea Davis & Ingrid Norrmann (2011). Explaining sonority projection effects. *Phonology* **28**. 197–234.
Davidson, Lisa & Jason A. Shaw (2012). Sources of illusion in consonant cluster perception. *JPh* **40**. 234–248.
Della Pietra, Stephen, Vincent J. Della Pietra & John D. Lafferty (1997). Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **19**. 380–393.
Edwards, Jan, Mary E. Beckman & Benjamin Munson (2004). The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition. *Journal of Speech, Language, and Hearing Research* **47**. 421–436.
Eisner, Jason (2002). Parameter estimation for probabilistic finite-state transducers. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. 1–8.
Elsner, Micha, Sharon Goldwater, Naomi Feldman & Frank Wood (2013). A joint learning model of word segmentation, lexical acquisition, and phonetic variability. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 42–54.
Goldrick, Matthew & Robert Daland (2009). Linking speech errors and phonological grammars: insights from Harmonic Grammar networks. *Phonology* **26**. 147–185.
Goldwater, Sharon & Mark Johnson (2003). Learning OT constraint rankings using a Maximum Entropy model. In Jennifer Spenador, Anders Eriksson & Östen Dahl (eds.) *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*. Stockholm: Stockholm University. 111–120.
Gouskova, Maria (2003). *Deriving economy: syncope in Optimality Theory*. PhD dissertation, University of Massachusetts, Amherst.
Grenander, Ulf (1976). *Pattern synthesis*. New York: Springer.
Harris, Theodore E. (1963). *The theory of branching processes*. Berlin: Springer.
Hay, Jennifer, Janet B. Pierrehumbert & Mary E. Beckman (2003). Speech perception, well-formedness and the statistics of the lexicon. In John Local, Richard Ogden & Rosalind Temple (eds.) *Phonetic interpretation: papers in laboratory phonology* VI. Cambridge: Cambridge University Press. 58–74.
Hayes, Bruce (2004). Phonological acquisition in Optimality Theory: the early stages. In René Kager, Joe Pater & Wim Zonneveld (eds.) *Constraints in phonological acquisition*. Cambridge: Cambridge University Press. 158–203.
Hayes, Bruce (2011). Interpreting sonority-projection experiments: the role of phonotactic modeling. In Wai-Sum Lee & Eric Zee (eds.) *Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong 2011*. Hong Kong: University of Hong Kong. 835–838.
Hayes, Bruce & James White (2013). Phonological naturalness and phonotactic learning. *LI* **44**. 45–75.
Hayes, Bruce & Colin Wilson (2008). A maximum entropy model of phonotactics and phonotactic learning. *LI* **39**. 379–440.
Jäger, Gerhard (2007). Maximum entropy models and Stochastic Optimality Theory. In Annie Zaenen, Jane Simpson, Tracy Holloway King, Jane Grimshaw, Joan Maling & Chris Manning (eds.) *Architectures, rules, and preferences: variations on themes by Joan W. Bresnan*. Stanford: CSLI. 467–479.
Jarosz, Gaja (2013). Learning with hidden structure in Optimality Theory and Harmonic Grammar: beyond Robust Interpretive Parsing. *Phonology* **30**. 27–71.

Jaynes, E. T. (1983). *Papers on probability, statistics, and statistical physics*. Edited by R. D. Rosenkrantz. Dordrecht: Kluwer.
Jelinek, Frederick (1997). *Statistical methods for speech recognition*. Cambridge, Mass.: MIT Press.
Legendre, Géraldine, Yoshiro Miyata & Paul Smolensky (1990). Harmonic Grammar: a formal multi-level connectionist theory of linguistic well-formedness: an application. In *Proceedings of the 12th Annual Conference of the Cognitive Science Society*. Hillsdale: Erlbaum. 884–891.
McCarthy, John J. & Joe Pater (eds.) (to appear). *Harmonic Grammar and Harmonic Serialism*. London: Equinox.
McCarthy, John J. & Alan Prince (1993). *Prosodic morphology I: constraint interaction and satisfaction*. Ms, University of Massachusetts, Amherst & Rutgers University.
McClelland, James L. & Jeffrey L. Elman (1986). The TRACE model of speech perception. *Cognitive Psychology* **18**. 1–86.
Magri, Giorgio (2012). Convergence of error-driven ranking algorithms. *Phonology* **29**. 213–269.
Manning, Christopher D. & Hinrich Schütze (1999). *Foundations of statistical natural language processing*. Cambridge, Mass: MIT Press.
Mattys, Sven L. & Peter W. Jusczyk (2001). Do infants segments words or recurring contiguous patterns? *Journal of Experimental Psychology: Human Perception and Performance* **27**. 644–645.
Merchant, Nazarré & Bruce Tesar (2008). Learning underlying forms by searching restricted lexical subspaces. *CLS* **41:2**. 33–47.
Norris, Dennis & James M. McQueen (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological Review* **115**. 357–395.
Pater, Joe (2008). Gradual learning and convergence. *LI* **39**. 334–345.
Pater, Joe (to appear). Universal Grammar with weighted constraints. In McCarthy & Pater (to appear).
Prince, Alan & Paul Smolensky (1993). *Optimality Theory: constraint interaction in generative grammar*. Ms, Rutgers University & University of Colorado, Boulder. Published 2004, Malden, Mass. & Oxford: Blackwell.
Riggle, Jason (2004). *Generation, recognition, and learning in finite-state Optimality Theory*. PhD dissertation, University of California, Los Angeles.
Riggle, Jason (2009). Violation semirings in Optimality Theory. *Research on Language and Computation* **7**. 1–12.
Scharenborg, Odette, Dennis Norris, Louis ten Bosch & James M. McQueen (2005). How should a speech recognizer work? *Cognitive Science* **29**. 867–918.
Smolensky, Paul & Géraldine Legendre (eds.) (2006). *The harmonic mind: from neural computation to optimality-theoretic grammar*. 2 vols. Cambridge, Mass.: MIT Press.
Storkel, Holly L., Jonna Armbrüster & Tiffany P. Hogan (2006). Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research* **49**. 1175–1192.
Tesar, Bruce, John Alderete, Graham Horwood, Nazarré Merchant, Koichi Nishitani & Alan Prince (2003). Surgery in language learning. *WCCFL* **22**. 477–490.
Tesar, Bruce & Alan Prince (2003). Using phonotactics to learn phonological alternations. *CLS* **39:2**. 209–237.
Tesar, Bruce & Paul Smolensky (1998). Learnability in Optimality Theory. *LI* **29**. 229–268.
Wilson, Colin & Lisa Davidson (2013). Bayesian analysis of non-native cluster production. *NELS* **40**. 265–278.
Wilson, Colin, Lisa Davidson & Sean Martin (2014). Effects of acoustic–phonetic detail on cross-language speech production. *Journal of Memory and Language* **77**. 1–24.
Wilson, Colin & Marieke Obdeyn (2009). Simplifying subsidiary theory: statistical evidence from Arabic, Muna, Shona, and Wargamay. Ms, Johns Hopkins University.