An approximation approach to the problem of the acquisition of phonotactics in Optimality Theory

Giorgio Magri
Laboratoire de Linguistique Formelle, CNRS and University of Paris 7
magrigrg@gmail.com

Abstract

The *problem of the acquisition of phonotactics* in Optimality Theory is intractable. This paper offers a way to cope with this hardness result: the problem is reformulated as a well known integer program (the *Assignment problem* with *linear side constraints*) paving the way for the application to phonotactics of *approximation algorithms* recently developed for integer programming.

Knowledge of the *phonotactics* of a language is knowledge of its distinction between licit and illicit forms. The acquisition of phonotactics represents a distinguished and important stage of language acquisition. In fact, in carefully controlled experimental conditions, nine-month-old infants already react differently to licit and illicit sound combinations (Jusczyk et al., 1993). They thus display knowledge of phonotactics already at an early stage of language development.

Usually, the problem of the acquisition of the phonotactics of a language given a finite set of linguistic data is formalized as the problem of finding a smallest language in the typology that is consistent with the data (Berwick, 1985; Manzini and Wexler, 1987; Prince and Tesar, 2004; Hayes, 2004; Fodor and Sakas, 2005). Section 1 formulates the problem of the acquisition of phonotactics along these lines within the mainstream phonological framework of *Optimality Theory* (Prince and Smolensky, 2004; Kager, 1999).

Unfortunately, (such a formulation of) the problem of the acquisition of phonotactics in OT turns out to be intractable (NP-complete): for any attempted efficient solution algorithm, there are some instances of the problem where the algorithm fails (Magri, 2010; Magri, 2012b). This hardness result holds for the *universal* formulation of the problem, in the sense of Heinz et al. (2009): there are no restrictions on the constraint set that defines the OT typology and indeed the OT typology itself figures as an input to the problem.

There are two strategies to cope with this hardness result. One approach weakens the formulation of the problem through proper restrictions on the constraint set: certain constraint sets are implausible from a phonological perspective, and should therefore be ignored in the proper formulation of the problem (Magri, 2011; Magri, 2012c). This approach raises interesting challenges, as it requires a through investigation of the algorithmic implications of various generalizations developed by phonologists on what counts as a "plausible" OT constraint set. Another approach is to bypass this difficulty, and weaken the formulation of the problem by lowering the standard for success: we settle on an *approximate* solution, namely a "small" language rather than a smallest language. This paper paves the way for the latter approach.

I focus on the specific formulation of the problem of the acquisition of OT phonotactics developed in Prince and Tesar (2004). In Sections 2 and 3, I show that this formulation of the problem can be restated as a classical integer program, namely the *Assignment problem* with *liner side constraints* (AssignLSCsPbm). The theory of *approximation algorithms* for integer programing is a blooming field of Computer Science (Bertsimas and Weismantel, 2005). In particular, powerful approximation algorithms have been recently developed for the AssignLSCsPbm. A state-of-the-art algorithm is due to Arora et al. (2002). The integer programming formulation developed in this paper thus paves the way for a new approximation approach to the problem of modeling the acquisition of phonotactics within OT. In Magri (2012a), I report simulation results with Arora's et. al. (2002) algorithm on various instances of the problem of the acquisition of phonotactics.

Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology (SIGMORPHON2012), pages 52–61,
Montréal, Canada, June 7, 2012. ©2012 Association for Computational Linguistics

1 Formulation of the problem

1.1 Basic formulation

A *typology* in Optimality Theory (OT) is defined through a 4-tuple $\tau = (\mathcal{X}, \mathcal{Y}, Gen, \mathfrak{C})$, where $\mathcal{X}$ is the set of *underlying forms*; $\mathcal{Y}$ is the set of *candidate surface forms*; Gen is the *generating function* that pairs an underlying form $x \in \mathcal{X}$ with a set $Gen(x) \subseteq \mathcal{Y}$ of surface forms called the *candidates* for x; and $\mathfrak{C}$ is the set of n *constraints* $C_1, \ldots, C_n$. Each constraint C_i is a function that maps a pair (x, y) of an underlying form $x \in \mathcal{X}$ and a candidate $y \in Gen(x)$ into a number $C_i(x, y)$, called the corresponding *number of violations*. The constraint set is split into the subset $\mathcal{M}$ of markedness constraints and the subset $\mathcal{F}$ of *faithfulness constraints*. As the constraint set is finite and can therefore only distinguish among a finite number of forms, I can assume that the set of underlying forms $\mathcal{X}$ is finite, as well as the candidate set $Gen(x)$ for any underlying form $x \in \mathcal{X}$.

Let π be a *ranking*, namely a total order over the constraint set. I denote by $\mathsf{OT}_\pi \colon \mathcal{X} \to \mathcal{Y}$ the *OT grammar* corresponding to the ranking π, as defined in Prince and Smolensky (2004). And I denote by $\mathcal{L}(\pi)$ the language corresponding to the ranking π, namely the range of the corresponding grammar OT_π (or, more explicitly, the set of all and only those surface forms $\hat{y} \in \mathcal{Y}$ such that there exists an underlying form $x \in \mathcal{X}$ such that $\mathsf{OT}_\pi(x) = \hat{y}$). Throughout the paper, I use x for an underlying form, $\hat{y}$ for a surface form which is an intended winner, and y for a surface form which is an intended loser.

The *Problem of the acquisition of phonotactics* in OT can be stated as in (1) in its universal formulation (Berwick, 1985; Manzini and Wexler, 1987; Prince and Tesar, 2004; Hayes, 2004). We are given an OT typology as well as a finite set $P \subseteq \mathcal{X} \times \mathcal{Y}$ of linguistic data. These data consist of pairs $(x, \hat{y})$ of an underlying form $x \in \mathcal{X}$ and a corresponding intended winner form $\hat{y} \in Gen(x)$. I assume that P is *consistent*, namely that there exists at least a ranking π such that $\mathsf{OT}_\pi(x) = \hat{y}$ for every pair $(x, \hat{y}) \in P$. We are asked to return a ranking π which has two properties. First, π is *consistent*: the corresponding OT grammar maps x into $\hat{y}$ for every pair $(x, \hat{y}) \in P$. Second, π is *restrictive*: there exists no other ranking π' consistent with P too such that the language $\mathcal{L}(\pi')$ corresponding to π' is a proper subset of the language $\mathcal{L}(\pi)$ corresponding to π. A solution algorithm needs to run in time polynomial in the number of constraints $|\mathfrak{C}|$ and the numbers of forms $|\mathcal{X}|, |\mathcal{Y}|$ (recall that $\mathcal{X}$ and $\mathcal{Y}$ are finite).

(1) *given*: an OT typology $\tau = (\mathcal{X}, \mathcal{Y}, Gen, \mathfrak{C})$
and a finite set $P \subseteq \mathcal{X} \times \mathcal{Y}$ of data;
find: a ranking π s.t. $P \subseteq \mathsf{OT}_\pi$ and there is
no π' s.t. $P \subseteq \mathsf{OT}_{\pi'}$ and $\mathcal{L}(\pi') \subset \mathcal{L}(\pi)$;
time: $\max\{|\mathfrak{C}|, |\mathcal{X}|, |\mathcal{Y}|\}$.

Problem (1) is NP-complete: there exists no efficient algorithm that is able to solve any instance of the problem (Magri, 2010; Magri, 2012b).

An interesting variant of the problem (1) assumes that we are given only the surface forms but not the corresponding underlying forms. Prince and Tesar (2004) and Hayes (2004) suggest that we can circumvent this difficulty as follows. Assume that the set of underlying forms and the set of surface forms coincide, namely $\mathcal{X} = \mathcal{Y}$. Assume furthermore that the typology is *output driven* (Tesar, 2008): a surface form $\hat{y}$ belongs to the language $\mathcal{L}(\pi)$ corresponding to a ranking π iff the corresponding grammar OT_π maps that form $\hat{y}$ (construed as an underlying form) into itself (construed as a surface form), as stated in (2)

$$(2) \qquad \hat{y} \in \mathcal{L}(\pi) \iff \mathsf{OT}_\pi(\hat{y}) = \hat{y}.$$

In this case, a way to cope with the lack of the underlying forms is to assume that the underlying form corresponding to a given surface form $\hat{y}$ is the completely faithful underlying form $\hat{y}$ itself. For this reason, I stick with the formulation (1) of the problem, whereby we are provided with both surface and underlying forms.

1.2 ERC notation

Consider an underlying form $x \in \mathcal{X}$ and two different candidate forms $y, \hat{y} \in Gen(x)$, with the convention that $\hat{y}$ is the intended winner for x while y is a loser. Following Prince (2002), all the relevant information concerning the underlying/winner/loser form triplet $(x, \hat{y}, y)$ can be summarized into the corresponding *elementary ranking condition* (ERC), namely the n-tuple $\mathbf{e}$ with entries $e_1, \ldots, e_n \in \{\text{L}, e, \text{W}\}$ defined as in (3).

$$(3) \quad (x, \hat{y}, y) \quad \Longrightarrow \quad \mathbf{e} = \begin{array}{|c|c|c|c|c|}\hline e_1 & \ldots & e_i & \ldots & e_n \\ \hline\end{array}$$

$$e_i \doteq \begin{cases} \text{W} & \text{if } C_i(x, \hat{y}) < C_i(x, y) \\ \text{L} & \text{if } C_i(x, \hat{y}) > C_i(x, y) \\ e & \text{if } C_i(x, \hat{y}) = C_i(x, y) \end{cases}$$

In words, The ith entry e_i is $e_i = \text{W}$ iff constraint C_i assigns more violations to (x, y) than to $(x, \hat{y})$ and thus favors the intended winner $\hat{y}$ over the loser y; $e_i = \text{L}$ iff the opposite holds; finally, $e_i = e$ iff the constraint C_i assigns the same number of violations to the two pairs (x, y) and $(x, \hat{y})$.

A ranking π can be represented as a permutation over $\{1, \ldots, n\}$, with the understanding that $\pi(i) = j$ means that the ranking π assigns constraint C_i to the jth stratum of the ranking, with the convention that the stratum corresponding to $j = n$ (to $j = 1$) is the top (bottom) of the ranking. For every such permutation π, let $\mathbf{e}_\pi$ be the n-tuple $\mathbf{e}$ with the components reordered according to π in decreasing order, as in (4).

(4) $\mathbf{e}_\pi \doteq (e_{\pi(n)}, \ldots, e_{\pi(1)})$

The ERC $\mathbf{e}$ is *OT-consistent* with π provided the left-most component of $\mathbf{e}_\pi$ different from e is a W.

For each of the pairs $(x, \hat{y})$ in the set P given with an instance of the problem (1), consider each loser candidate $y \in Gen(x)$ different from $\hat{y}$, construct the ERC corresponding to the underlying/winner/loser form triplet $(x, \hat{y}, y)$ as in (3) and organize all these ERCs one underneath the other into an *ERC matrix* with n columns and many rows (the order of the ERCs does not matter). I denote a generic ERC matrix by $\mathbf{E}$ and I say that a ranking π is *OT-consistent* with $\mathbf{E}$ provided it is consistent with each of its ERCs. The problem of the acquisition of phonotactics in (1) can thus be equivalently restated in ERC notation as in (5).

(5) *given*: an OT typology $\tau = (\mathcal{X}, \mathcal{Y}, Gen, \mathfrak{C})$ and an ERC matrix $\mathbf{E}$;
find: a ranking π s.t. π is OT-consistent with $\mathbf{E}$ and there is no π' consistent with $\mathbf{E}$ too s.t. $\mathcal{L}(\pi') \subset \mathcal{L}(\pi)$;
time: $\max\{|\mathfrak{C}|, |\mathcal{X}|, |\mathcal{Y}|\}$.

The latter formulation of the problem is only partially stated in terms of ERC notation, as the condition $\mathcal{L}(\pi') \subset \mathcal{L}(\pi)$ still requires knowledge of the entire OT typology. This difficulty is tackled in the next Subsection.

1.3 Restrictiveness measures

Let a *restrictiveness measure* be a function μ which takes a ranking π and returns a number $\mu(\pi) \in \mathbb{N}$ that provides a relative measure of the size of the language $\mathcal{L}(\pi)$ corresponding to π, in the sense that the (strict) monotonicity property in (6) holds for any two rankings π, π'.

(6) If $\mathcal{L}(\pi') \subset \mathcal{L}(\pi)$, then $\mu(\pi') < \mu(\pi)$.

Any solution of the optimization problem (7) is a solution of the corresponding instance (5) of the problem of the acquisition of phonotactics. In fact, if π solves (7) then there cannot exist any other ranking π' consistent with the ERC matrix that corresponds to a smaller language $\mathcal{L}(\pi') \subset \mathcal{L}(\pi)$, since (6) would imply that $\mu(\pi') < \mu(\pi)$, contradicting the hypothesis that π is a solution of (7).

(7) *minimize*: $\mu(\pi)$;
subject to: π is OT-consistent with the given ERC matrix $\mathbf{E}$;
time: number of columns and rows of $\mathbf{E}$.

As problem (7) is stated completely in terms of the ERC matrix $\mathbf{E}$, the time required by a solution algorithm needs to scale just with the size of $\mathbf{E}$.

From now on, I will focus on the new formulation (7). Thus, I need a restrictiveness measure (6). Of course, not just any restrictiveness measure will do. For instance, the function (8), which pairs a ranking π with the cardinality of its language $\mathcal{L}(\pi)$, trivially satisfies (6).

(8) $\mu(\pi) \doteq |\mathcal{L}(\pi)|$.

Yet, this is not a *good* restrictiveness measure, because there seems to be no way to compute $\mu(\pi)$ without actually computing the language $\mathcal{L}(\pi)$, which requires knowledge of the entire typology.

Prince and Tesar (2004) suggest a better candidate, which is defined for any ranking π as in (9). Recall that the constraint set $\mathfrak{C} = \mathcal{F} \cup \mathcal{M}$ is split up into the subset $\mathcal{F}$ of faithfulness constraints and the subset $\mathcal{M}$ of markedness constraints. For each faithfulness constraint $F \in \mathcal{F}$, determine the number $\mu(F)$ of markedness constraints $M \in \mathcal{M}$ ranked by π *below* that faithfulness constraint, i.e. $\pi(F) > \pi(M)$. Finally, add up all these numbers $\mu(F)$ together to determine the value $\mu(\pi)$.

(9) $$\mu(\pi) \doteq \sum_{F \in \mathcal{F}} \underbrace{\Big|\{M \in \mathcal{M} \mid \pi(F) > \pi(M)\}\Big|}_{\mu(F)}$$

Is the function μ defined in (9) is a restrictiveness measure? namely, does it satisfy condition (6)? Prince and Tesarconjecture that it is, based on the following intuition. Markedness (faith-