LEARNABILITY TO THE RESCUE OF THE QUEST FOR PHONOLOGICAL UNIVERSALS

Workshop on Language Universals,
Crecchio, 9-11 June 2017

Giorgio Magri
SFL (CNRS, University of Paris 8, UPL)
magrigrg@gmail.com

1. Position statement

Universals in phonology usually have the form of generalizations stated with the technical vocabulary of phonetics obtained by direct induction from some large linguistic database. The quest for phonological universals is thus part of the field of linguistic typology. To illustrate, Hyman (2008) looks at the UCLA Phonological Segment Inventory Database (UPSID) collected by Maddieson (1984). And he entertains and tests generalizations such as "every phonological system has coronal phonemes", stated in terms of basic phonetic notions such as place of articulation. Pericliev (2008) offers a more large scale and automatized illustration of this same approach.

Evans and Levinson (2009) submit that that this approach has failed: "for the substantial findings about universals across languages one must turn to the field of linguistic typology, which has laid bare a bewildering range of diverse languages, where the generalizations are really quite hard to extract" because "languages differ so fundamentally from one another at every level of description." Strikingly, phonologists seem to share Evans and Levinson's pessimistic assessment. Thus, Maddieson (1984) admits that "there appear to be so few absolute universals". Hyman (2008) warns that "the study of universals is fraught with difficulties". And van Oostendorp (2013) flatly concludes in his assessment of the achievements of phonology that "the quest for universals has failed." Needless to say, Hyman's conjecture about coronal phonemes was quickly shown "to bite the dust" (Blevins 2009). What has happened?

I submit that the quest for universals in phonology has failed because the methodological approach adopted by phonologists, despite its innocuous straightforwardness, is actually doomed to fail because of two serious mistakes. The first mistake of that approach is that it defines universals as invariant *substantive* properties of the typology of *languages*. Indeed, Evans and Levinson in the quote above talk about "universals across languages". And phonologists have focused on universals at the level of extensional languages such as segment inventories, stated in terms of phonetic notions. This is a mistake. We should instead focus on *formal* universals at the level of the *grammars* which generate those languages, stated in terms of whatever formalism is used to define those grammars. This intuition was very clear at the outset of generative linguistics: Chomsky (1965, section 1.5, p. 27) clearly states that "the study of linguistic universals is the study of the properties of any *generative grammar* [emphasis mine] for a natural language".

The second mistake of the current approach to phonological universals is that it assumes that the formulation of universals pertains to linguistic typology. Indeed, Evans and Levinson in the quote above submit that "one must turn to the field of linguistic typology" for universals. This is a mistake. The formulation of universals pertains instead to the mathematical theory of language learnability and algorithmic complexity. Once again, this intuition was very clear at the outset of generative linguistics: Chomsky (1965, section 1.5, p. 27) states that "the main task of linguistic theory must be to develop an account of linguistic universals that, on the one hand, will not be falsified by the actual diversity of languages and, on the other, *will be sufficiently rich and explicit to account for the rapidity and uniformity of language learning* [emphasis mine]."

Unfortunately, the generalizations entertained by phonologists have usually nothing to do with learnability: how could learnability benefit from the mandatory presence of [t]? We should instead focus on generalizations for which we can distill through mathematical theorems a provable effect on the complexity of the learning problem.

These considerations lead to the following alternative approach to universals in phonology. First, precise and mathematically explicit formulations of the various components of the language learning problem are considered. Second, tools from complexity theory and the theory of algorithms are used to individuate algorithmic bottlenecks and to prove intractability results, showing that learning is hard without substantial typological restrictions. Third, tools from machine learning are employed to formulate universal typological restrictions that provably allow those bottlenecks and computational limitations to be overcome, yielding provable learnability gains. Finally, the empirical soundness of the posited universal restrictions is assessed by the typologist against the sparse data provided by linguistic databases. Crucially, the role of the typologist is now downgraded from the main character to a supporting role who comes in only at the very end of the process. In the rest of this document, I illustrate this approach by sketching two hypothesized universals motivated by learnability considerations that I am currently working on.

2. First example: is feature interaction acyclic?

Let's agree that a phonological grammar G is a function from underlying forms (which I denote as x) to surface forms (which I denote as y). The initial, most basic formulation of the language learning problem is (1). In compliance with the generative perspective, the learner is provided with full typological information (1a.i). The linguistic data the learner is trained on consist of a finite number of underlying/surface form mappings (1a.ii). The target grammar which has allegedly generated the data ought to be consistent with those mappings. The learner is thus required to find a grammar in the typology which satisfies the consistency condition (1b).

(1) a. Given: i. a typology of phonological grammars;
ii. a set D of linguistic data consisting of pairs (x, y) of an underlying form x and a corresponding surface form y.
b. Find: a grammar G in the typology *consistent* with the data: for each pair (x, y) in D, G maps the underlying form x to the corresponding surface form y.

This formulation (1) of the problem of learning phonology is unrealistically simple for two reasons: the success condition (1b) is too weak and the training data (1a.ii) are too rich. Let me start with the former issue (I will come back to the latter issue below in section 3).

Plausibly, many grammars satisfy the consistency condition (1b). Problem (1) is thus *ill-posed* because it admits multiple solutions. The criteria for success in (1b) must be made more stringent through additional conditions. A natural additional condition is that the set L_G of surface forms phonotactically licit according to the grammar G returned by the learner be as small as possible compatibly with consistency, as stated by the additional *restrictiveness* condition (2b.ii). This is the classical *Subset Principle* (Fodor and Sakas 2005 and references therein).

(2) a. Given: i. a typology of phonological grammars;
ii. a set D of linguistic data consisting of pairs (x, y) of an underlying form x and a corresponding surface form y.
b. Find: a grammar G in the typology which is:
i. *consistent* with the data: for each pair (x, y) in the data set D, G maps the underlying form x to the corresponding surface form y;
ii. *restrictive* relative to the data: the typology contains no other grammar $\widehat{G}$ such that $L_{\widehat{G}} \subsetneq L_G$ and $\widehat{G}$ also satisfies the consistency condition (2b.i).

Unfortunately, the addition of the restrictiveness condition (2b.ii) changes drastically the complexity of the learning problem. In fact, suppose for concreteness that the typology of grammars fed to the learner in (1a.i)/(2a.i) is defined in OT (*Optimality Theory*; Prince and Smolensky 2004) terms. While the original learning problem (1) is efficiently solvable (Tesar and Smolensky

1998), the amended problem (2) turns out to be intractable (Magri 2013): there exists no efficient learner which succeeds on an arbitrary instance of problem (2).

These complexity results say that the structure provided by the bare formal architecture of OT suffices for the easier learning problem (1) but not for the more realistic and thus harder problem (2). For the latter problem, additional structure is needed, plausibly brought about through restrictions on the constraint set used to define the typology in (2a.i). What should these restrictions on viable constraint sets look like? For concreteness, let me focus on the simplest case, namely the phonology of segment inventories. Segments are described through binary phonological *features*. One of the two feature values counts as *marked* (Clements 2005), namely it is penalized by a *simple markedness constraint* corresponding to that feature. For instance, the feature [lateral] comes with the simple markedness constraint ∗[+later]. Feature interaction is modeled through *feature co-occurrence constraints* (FCCs). These are constraints of the form $*[\pm\varphi_1, \pm\varphi_2, \dots, \pm\varphi_m]$ which penalize certain combinations of values of certain features $\varphi_1, \varphi_2, \dots, \varphi_m$.

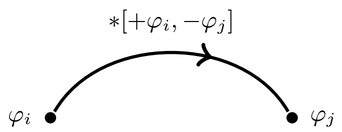
The logical space of possible FCCs is huge. Levelt and van Oostendorp (2007) suggest to restrict it by considering only two formal types of FCCs:

- *Multi-marked* FCCs target an arbitrary number of phonological features and penalize forms which have the marked value for each of the targeted features. An example is the FCC ∗[+velar, +voice], because "+" is the marked value of both features [velar] and [voice].
- *Implicational* FCCs target only two features and penalize forms which have the marked value for one of them (called the *antecedent* feature) and the unmarked value for the other (called the *consequent* feature). An example is the FCC ∗[+back, −round], because "+" is the marked value for backness (Krämer 2003) and "−" is the unmarked value for roundness.

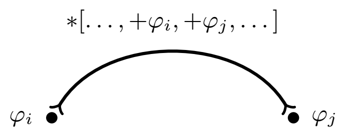
The intuition behind Levelt and van Oostendorp's restriction to just these two types of FCCs is as follows. FCCs which punish combinations of marked values are easy to deal with because they do not disagree with simple markedness constraints. For this reason, there is no need to limit their complexity: multi-marked FCCs can indeed target an arbitrary number of features. FCCs which instead punish a combination of values which includes unmarked values are difficult to deal with because they introduce disagreement among markedness constraints. For this reason, it makes sense to start from the assumption that they are as simple as possible, namely that they target a unique unmarked value and a unique marked value, yielding implicational FCCs.

From now on, let me call "+" the marked value and "−" the unmarked value. The fact that two features φ_i, φ_j interact through an agreement constraint $*[+\varphi_i, -\varphi_j]$ with antecedent feature φ_i and consequent feature φ_j can be graphically represented as in (3a). The fact that those two features φ_i, φ_j interact through a multi-marked constraint $*[\dots, +\varphi_i, +\varphi_j, \dots]$ can be graphically represented as in (3b).

(3) a.



b.



A set of FCCs thus defines a *feature interaction graph* $(\mathcal{V}, \mathcal{A}, \mathcal{E})$ as follows: the set $\mathcal{V}$ of vertices is the set of features, represented as dots in (3); the set $\mathcal{A}$ of directed edges contains an ordered pair $\langle\varphi_i, \varphi_j\rangle$ for every agreement constraint $*[+\varphi_i, -\varphi_j]$ with antecedent feature φ_i and consequent feature φ_j, as represented in (3a); finally, the set $\mathcal{E}$ of undirected hyper-edges contains an edge $\{\varphi_1, \dots, \varphi_m\}$ for every multi-marked constraint $*[+\varphi_1, \dots, +\varphi_m]$, as represented in (3b).

We want to formulate restrictive assumptions on the set of FCCs which might support learnability results and thus overcome the intractability of the unrestricted problem (2). Since a set of FCCs induces a feature graph, it makes sense to start from the assumption that the set of FCCs induces a feature graph which is as simple as possible. From a graph-theoretic perspective, the simplest graphs are those with no cycles (since the feature interaction graph is an hyper-graph, some care is needed in the definition of cyclicity; Brault-Baron 2016). These considerations lead to the following:

(4) **Acyclicity hypothesis:** the markedness constraint set contains (only) two types of FCCs: multi-marked and implicational; and the feature interaction graph induced by these FCCs has no cycles.

My current research is trying to *prove* that problem (2), although intractable in the general case, admits an efficient learner when the typology provided to the learner in (2a.i) is the OT typology defined by a set of FCCs which complies with the acyclicity hypothesis. This computational work is effectively uncovering a rich and beautiful typological structure entailed by this formal hypothesis. A subsequent stage of the project will try to validate this structure on the PHOIBLE database of segment inventories (Moran et al. 2014). If successful, this project will establish FCC's acyclicity as a phonological universal of the new formal type envisioned in section 1.

3. Second example: are chain shifts phonotactically innocuous?

Another shortcoming shared by the formulations (1) and (2) of the problem of learning phonology is that they assume that the linguistic data (1a.ii) and (2a.ii) available to the learner consist not only of surface forms but also of the corresponding underlying forms. This assumption that the underlying forms are also provided to the learner is particularly implausible for what Hayes (2004) calls the *early stage* of the acquisition of phonology, when morphology is lagging behind and the child therefore lacks the information on underlying forms crucially provided by alternations. These considerations suggest a further reformulation of the learning problem, whereby the linguistic data (5a.ii) consist of nothing more than a set of surface forms. The consistency condition (5b.i) is restated accordingly, through existential quantification over underlying forms.

(5) a. Given: i. a typology of phonological grammars;
ii. a set D of linguistic data consisting of just surface forms.
b. Find: a grammar G in the typology which is:
i. *consistent* with the data: for every surface form y in the data set D, there exists some underlying form x such that G maps x to y;
ii. *restrictive* relative to the data: the typology contains no other grammar $\widehat{G}$ such that $L_{\widehat{G}} \subsetneq L_G$ and $\widehat{G}$ also satisfies the consistency condition (5b.i).

A number of scholars in both the computational literature (e.g., Hayes 2004 and Prince and Tesar 2004) and the acquisition literature (e.g., Gnanadesikan 2004 and Pater and Barlow 2003) have proposed that the learner circumvents the lack of underlying forms in the training data (5a.ii) by assuming fully faithful underlying forms, as stated in (6). This assumption makes sense whenever there are no representational differences between underlying and surface forms (in the OT literature, this is known as Moreton's 2004b *homogeneity condition*).

(6) For each training surface form y in the training data (5a.ii), the learner assumes the corresponding fully faithful underlying form $\mathsf{x} = \mathsf{y}$.

The learner thus effectively tackles problem (7) instead of the original problem (5). They differ because the consistency condition (7a.i) as been restated so as to require a solution grammar to map each form y (construed as an underlying form) to itself (construed as a surface form).

(7) a. Given: i. a typology of phonological grammars;
ii. a set D of linguistic data consisting of just surface forms.
b. Find: a grammar G in the typology which is:
i. *consistent* with the faithful interpretation of the data: for every surface form y in the data set D, G maps y (construed as a surface form) to y itself;
ii. *restrictive* relative to the data: the typology contains no other grammar $\widehat{G}$ such that $L_{\widehat{G}} \subsetneq L_G$ and $\widehat{G}$ also satisfies the consistency condition (7b.i).

The assumption (6) of faithful underlying forms is called *computationally sound* provided each solution of the derived problem (7) also counts as a solution of the original problem (5), so that the learner can solve the original problem (5) by tackling the derived problem (7) instead. Under which conditions is the assumption (6) of faithful underlying forms computationally sound?

The original consistency condition (5b.i) says that there exists some underlying form x which is mapped by G to y. The revised consistency condition (7b.i) says that G faithfully maps y to y itself. A grammar G is called *idempotent* provided these two conditions are equivalent (Magri to appear). If every grammar in the typology (5a.i)/(7a.i) explored by the learner is idempotent, the original learning problem (5) and the derived (7) are obviously equivalent. We have thus obtained an initial sufficient condition for the computational soundness of the assumption (6) of faithful underlying forms: no harm comes from entertaining faithful underlying forms under the hypothesis that the typology explored by the learner consists of all idempotent grammars.

Unfortunately, this idempotency hypothesis is unrealistic. In fact, a grammar G is idempotent provided it displays no chain shift (8) whereby: some form x is mapped to y, whereby the original consistency condition (5b.i) holds; and yet y is not faithfully realized but rather mapped to some form z different from y, whereby the derived consistency condition (7b.i) fails.

(8) $\mathsf{x} \rightarrow \mathsf{y} \rightarrow \mathsf{z}$

But chain shifts have been widely documented in adult phonology (Łubowicz 2011; Moreton 2004a; Moreton and Smolensky 2002), child phonology (Velten 1943; Smith 1973; Macken 1980; Dinnsen and Barlow 1998; Cho and Lee 2000, 2003; Dinnsen et al. 2001; Jesney 2007), second language acquisition (Lee 2000; Jesney 2007), and delayed phonological acquisition (Dinnsen and Barlow 1998; Dinnsen et al. 2011). How can we resolve this tension between the existence of chain shifts and the assumption (6) of faithful underlying forms?

The hypothesis that all grammars in the typology are idempotent ensures that the original problem (5) and the derived problem (7) are actually *equivalent*: a grammar solves one if and only if it solves the other. Yet, that is much more than what we actually need: the computational soundness of the assumption (6) of faithful underlying forms only requires that a grammar which solves the derived problem (7) also solves the original problem (5), while the reverse implication is not needed. The problematic hypothesis that all grammars in the typology are idempotent can therefore be weakened: the typology can contain a non-idempotent grammar as long as it also contains a companion idempotent grammar which makes the same phonotactic distinctions. In other words, failures of idempotency must be phonotactically innocuous, as hypothesized in (9).

(9) **Innocuousness hypothesis:** for every non-idempotent grammar G, the typology contains a companion grammar $\widehat{G}$ such that $\widehat{G}$ is idempotent and nonetheless G and $\widehat{G}$ define the same set of phonotactically licit surface forms, namely $L_G = L_{\widehat{G}}$.

To illustrate, consider an OT ranking which yields the non-idempotent chain shift (8). To show that it is phonotactically innocuous, I need to be able to manipulate that ranking to obtain a different ranking whose corresponding mapping (10) is idempotent and makes the same phonotactic distinctions (x, y are licit; z is illicit). Under which condition is that possible?

(10) $\mathsf{x} \rightarrow \mathsf{y} \circlearrowright \quad \mathsf{z} \circlearrowright$

Recent results in the theory of OT idempotency provide some guidance for answering this question. In order to flout idempotency and thus derive a chain shift (8), the constraint set needs to contain a faithfulness constraint which fails at formal condition called the *faithfulness idempotency condition* (FIC; Magri to appear). Under mild assumptions, this FIC admits the following interpretation: it requires faithfulness constraints to measure the *phonological distance* between underlying and surface forms in compliance with a basic axiom of the notion of distance, namely the *triangle inequality*. Because of this interpretation, it comes as no surprise that all *basic* faithfulness constraints used in phonological theory sensibly satisfy the FIC. The faithfulness constraints which do fail at the FIC are instead *derived* from basic FIC-complying faithfulness constraints, thorough operations such as constraint conjunction or constraint restriction. These considerations suggest that, by replacing the derived non-FIC-complying faithfulness constraint in a ranking (11a) which yields the chain shift (8) with a corresponding basic FIC-complying faithfulness constraint, I might be able to obtain a ranking (11b) which yields the companion mapping (10).

(11) a. $\ldots \gg$ derived non-FIC-complying faithfulness constraint $\gg \ldots$
b. $\ldots \gg$ corresponding basic FIC-complying faithfulness constraint $\gg \ldots$

If this strategy can be successfully implemented for various languages reported to display chain shifts (as listed for instance in Moreton 2004a and Moreton and Smolensky 2002), it will establish the phonotactic innocuousness as a phonological universal of the new formal type envisioned in section 1.

References

Blevins, Juliette. 2009. Another universal bites the dust: Northwest Mekeo lacks coronal phonemes. *Oceanic Linguistics* 48.1:264–273.

Brault-Baron, Johann. 2016. Hypergraph acyclicity revisited. *ACM Computing Surveys (CSUR)* 49:54.

Cho, Mi-Hui, and Shinsook Lee. 2000. Ordering problems in phonological acquisition. *Studies in Phonetics, Phonology and Morphology* 6:137–157.

Cho, Mi-Hui, and Shinsook Lee. 2003. The acquisition of fricatives: chain shift cases of english and korean. *Studies in Phonetics, Phonology and Morphology* 9:485–498.

Chomsky, Noam. 1965. *Aspects of the theory of syntax*. The MIT Press.

Clements, G. N. 2005. The role of features in phonological inventories. In *Contemporary views on architecture and representations in phonological theory*, ed. Eric Raimy and Charles Cairns. Cambridge, MA: MIT Press.

Dinnsen, Daniel A., and Jessica A. Barlow. 1998. On the characterization of a chain shift in normal and delayed phonological acquisition. *Journal of Child Language* 25:61–94.

Dinnsen, Daniel A., Christopher R. Green, Judith A. Gierut, and Michelle L. Morrisette. 2011. On the anatomy of a chain shift. *Journal of Linguistics* 47.2:1–25.

Dinnsen, Daniel A., Kathleen M. O'Connor, and Judith A. Gierut. 2001. The puzzle-puddle-pickle problem and the duke-of-york gambit in acquisition. *Journal of Linguistics* 37:503–525.

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

Fodor, Janet Dean, and William Gregory Sakas. 2005. The Subset Principle in syntax: costs of compliance. *Linguistics* 41:513–569.

Gnanadesikan, Amalia E. 2004. Markedness and faithfulness constraints in child phonology. In *Constraints in phonological acquisition*, ed. René Kager, Joe Pater, and Wim Zonneveld, 73–108. Cambridge: Cambridge University Press. Circulated since 1995.

Hayes, Bruce. 2004. Phonological acquisition in Optimality Theory: The early stages. In *Constraints in phonological acquisition*, ed. René Kager, Joe Pater, and Wim Zonneveld, 158–203. Cambridge: Cambridge University Press.

Hyman, Larry. 2008. Universals in phonology. In *The linguistic review*, ed. Harry van der Hulst.

Jesney, Karen. 2007. Child chain shifts as faithfulness to input prominence. In *Proceedings of the 2nd Conference on Generative Approaches to Language Acquisition*, 188–199. Somerville, MA: Cascadilla Proceedings Project.

Krämer, Martin. 2003. *Vowel harmony and correspondence theory*. Walter de Gruyter.

Lee, Shin Sook. 2000. Chain shift in second language phonological acquisition. *Journal of the Pan-Pacific Association of Applied Linguistics* 4:175–199.

Levelt, Clara, and Marc van Oostendorp. 2007. Feature co-occurrence constraints in l1 acquisition. In *Linguistics in the netherlands 2007*, ed. Bettelou Los and Marjo van Koppen, 162–172. John Benjamins.

Łubowicz, Anna. 2011. Chain shifts. In *Companion to phonology*, ed. Marc van Oostendorp, Beth Ewen, Colin Hume, and Keren Rice, 1717–1735. Wiley-Blackwell.

Macken, Marlys. 1980. The child's lexical representation: the 'puzzle-puddle-pickle' evidence. *Journal of Linguistics* 16.1:1–17.

Maddieson, Ian. 1984. *Patterns of sounds*. Cambridge University Press.

Magri, Giorgio. 2013. The complexity of learning in OT and its implications for the acquisition of phonotactics. *Linguistic Inquiry* 44.3:433–468.

Magri, Giorgio. to appear. Idempotency in Optimality Theory. *Journal of Linguistics* .

Moran, Steven, Daniel McCloy, and Richard Wright, ed. 2014. *PHOIBLE online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. URL `http://phoible.org/`.

Moreton, Elliott. 2004a. A compendium of synchronic chain shifts. Ms., Chapel Hill, NC: University of North Carolina Chapel Hill.

Moreton, Elliott. 2004b. Non-computable functions in Optimality Theory. In *Optimality theory in phonology: A reader*, ed. John J. McCarthy, 141–163. Malden: MA: Wiley-Blackwell.

Moreton, Elliott, and Paul Smolensky. 2002. Typological consequences of local constraint conjunction. In *WCCFL 21: Proceedings of the 21st annual conference of the West Coast Conference on Formal Linguistics*, ed. L. Mikkelsen and C Potts, 306–319. Cambridge, MA: Cascadilla Press.

van Oostendorp, Marc. 2013. Phonology between theory and data. In *L'interface langage-cognition / the language-cognition interface. actes du 19e congrès international des linguistes (langue et cultures, 45)*, ed. S.R. Anderson, J. Moeschler, and F. Reboul, 289–306. Genève - Paris: Libraire Droz.

Pater, Joe, and Jessica A. Barlow. 2003. Constraint conflict in cluster reduction. *Journal of Child Language* 30:487–526.

Pericliev, Vladimir. 2008. Implicational phonological universals. *Folia Linguistica* 42:195–225.

Prince, Alan, and Paul Smolensky. 2004. *Optimality Theory: Constraint interaction in generative grammar*. Oxford: Blackwell. As Technical Report CU-CS-696-93, Department of Computer Science, University of Colorado at Boulder, and Technical Report TR-2, Rutgers Center for Cognitive Science, Rutgers University, New Brunswick, NJ, April 1993. Also available as ROA 537 version.

Prince, Alan, and Bruce Tesar. 2004. Learning phonotactic distributions. In *Constraints in phonological acquisition*, ed. R. Kager, J. Pater, and W. Zonneveld, 245–291. Cambridge University Press.

Smith, N. V. 1973. *The acquisition of phonology: a case study*. Cambridge, England: CUP.

Tesar, Bruce, and Paul Smolensky. 1998. Learnability in Optimality Theory. *Linguistic Inquiry* 29:229–268.

Velten, Harry de Veltheyme. 1943. The growth of phonemic and lexical patterns in infant language. *Language* 19:281–292.