A unified approach to several learning challenges in phonology

Ezer Rasin, Itamar Shefi, and Roni Katzir; Leipzig University and Tel Aviv University

Summary. The child acquiring the morpho-phonology of their language needs to acquire a large amount of knowledge from distributional evidence alone (i.e., from unanalyzed surface forms, with no corrections, explicit URs, or paradigmatic information). This task raises nontrivial challenges even in the case of simple alternations (e.g., final devoicing), but there are several aspects of phonological knowledge that make phonological induction particularly difficult, beyond the challenges posed by simple alternations. We discuss four seemingly unrelated learning challenges in phonology and provide new simulation results showing how an approach to learning based on the principle of Minimum Description Length (MDL; Solomonoff 1964, Rissanen 1978) addresses all of them (as well as their interactions) in a unified way.

Challenges. (1) Unbounded dependencies. Vowel Harmony (VH) often applies across several intervening consonants. This poses a problem for phonological learners that are limited to small, local contexts of fixed size. Unsupervised learners that can capture long-distance dependencies, such as Hayes & Wilson 2008 or Heinz 2010, are phonotactic learners that are not yet integrated within a full morpho-phonological learner. Moreover, as noted by Hayes & Wilson 2008:402, languages with many disharmonic roots like Turkish (e.g., backness mismatch in mezar(-lar), hotel(-ler)) pose a problem for attempts to acquire VH using a phonotactic learner. **(2) Dependencies between phonology and segmentation.** VH also poses a difficulty for proposals that split morpho-phonological learning into morphological segmentation and phonological induction. If phonological induction applies first, learning will be hampered by the fact that VH often applies only across morphemes, which in this scenario are not yet available to the learner, while morpheme-internally vowels can be disharmonic. If segmentation is acquired first, there remains the nontrivial task of unifying the different surface forms and positing an appropriate phonological process, a challenging task if the child does not have access to, e.g., paradigms. **(3) Abstract URs.** Alderete & Tesar (2002) (see also McCarthy 2005) note that stress patterns in several languages (e.g., Mohawk, Selayarese, Yimas) require the acquisition of URs that are not identical to the SR, and argue that in these cases there are no alternations to support the induction of a nonidentical UR. These languages pose a challenge to a family of learning proposals which posit URs that are identical to SRs outside of alternations (e.g., Tesar 2014). **(4) Opacity.** Opaque interactions pose a learning challenge by obscuring the form of a phonological process and its environment of application. In Catalan, for example (Mascaró 1976), word-final nasals obligatory delete in some environments (kuzí ∼ kuzín-s); a second process deletes word-final stops after a nasal (e.g., kəlén ∼ kəlént-ə). Crucially, nasal deletion does not apply if the relevant nasal is word-final due to a following stop that deleted, which can be captured by ordering nasal deletion before stop deletion (counterfeeding opacity). The learning challenge comes from surface forms such as [kəlén] that can confuse a naive attempt to learn that word-final nasals delete.

The MDL Principle. MDL is an evaluation criterion that balances two competing factors: the simplicity of the grammar ($|G|$; as in the evaluation metric of SPE); and the tightness of fit of the grammar to the data ($|D : G|$, the length of the encoding of the data D given G; similarly to the subset principle):

(5) MDL EVALUATION METRIC: If G and G' can both generate the data D, and if $|G| + |D : G| < |G'| + |D : G'|$, prefer G to G'

Before addressing (1)–(4), we first show how MDL supports the induction of a segmented lexicon and phonological rules using a toy example. **Segmentation**: (5) allows the learner to discover the segmentation of words into stems and affixes (de Marcken 1996, Goldsmith 2001). If the surface forms are generated from, e.g., 8 different stems (e.g., /dok/, /kab/, etc.) and

4 different suffixes (e.g., /za/, /ti/, etc.), a naive lexicon for the language will include all the different $8 \times 4 = 32$ surface forms. By (5), the learner will prefer a simpler grammar (shorter $|G|$, while $|D : G|$ remains the same) in which the stems and the suffixes are stored separately, with only $8 + 4 = 12$ different entries (which, in addition, are shorter than those in the naive encoding). **Phonology**: (5) also enables the learner to acquire phonological processes (e.g., Goldwater & Johnson 2004, and Rasin et al. 2018). If the language just discussed also has a process of progressive voicing assimilation across morphemes, the surface forms will seem to involve twice the actual number of suffixes (e.g., [sa] after voiceless stops, [za] elsewhere). Using (5) the learner will reject a naive encoding of this kind given sufficiently many suffixes (since the storage of pairs of surface forms for each suffix is costly) in favor of one where there is just one variant for each suffix, along with a rule of voicing assimilation (since the savings obtained by storing just one form for each suffix outweigh the costs of adding the relevant phonological rule). **A unified solution to the four challenges using MDL.** MDL works directly with the linguistics representations, which in our case use rewrite rules with the possibility of variable-length marking using the equivalent of C_0 from SPE. This enables the MDL learner to acquire an appropriate VH rule that applies across an unbounded number of consonants, thus addressing (1). Under MDL, segmentation and phonology are induced in a unified way, so they can be learned jointly, thus addressing (2). MDL resolves (3) since abstract URs discussed in the literature support a shorter encoding of the phonological statements and the lexicon (sometimes due to simplification of the lexicon's alphabet). MDL also supports the induction of non-surface-true processes, making (4) possible. We present simulations showing each of (1)–(4) as well as some nontrivial interactions. We illustrate here for VH and opacity.

Simulation #1: Turkish. The dataset was modeled after front-back VH in Turkish. The learner needs to learn both a lexicon of URs and VH. The data were generated by taking all combinations of 10 monosyllabic Turkish nouns (e.g., kent, jıl) and 8 suffixes (e.g., -ler/-lar, -in/-ın) and applying VH. Words were presented as unsegmented strings, without any morphology (e.g., [kentler], [jıllar]). The hypothesis space consisted of grammars with a lexicon (represented as a Hidden Markov Model) and a set of ordered rules. Search was performed using a Genetic Algorithm (Holland 1975) and converged on a hypothesis with a VH rule that applies in all appropriate places (matching vowels in the data were all separated by $[-back]$ segments) and a lexicon in which each pair (e.g., -ler/-lar) is represented with a single UR:

Final state: Rule: $[+syll] \rightarrow [+back] \, / \begin{bmatrix} +back \\ +cont \end{bmatrix} [-back]^{*} \_\_$ (obligatory)

Lexicon: Stems = {kent, jıl, güz, tuz . . . }; Suffixes = {ler, in, ten, siz, . . . }

Simulation #2: Catalan counterfeeding. The dataset was generated by taking all combinations of 13 stems and 5 suffixes from Catalan and applying final-nasal deletion and stop deletion, in this order. Using the same setting as before, the words were again presented to the learner without any morphology. The learner correctly segmented the data and induced the rules and their ordering (example derivations: /kuzin/ → [kuzi], /kalent/ → [kalen]):

Final state: Rules: 1) $[+nasal] \rightarrow \emptyset / \_\_\#$

2) $[-cont] \rightarrow \emptyset / [-cont] \_\_\#$

Lexicon: Stems = {kuzin, kalent, blank, kasa, . . . }; Suffixes = {s, et, ik, a, . . . }

Implications. Our results show that morpho-phonological patterns involving (1)–(4), which seem to pose particular challenges to learning, can be learned from distributional evidence alone. They provide further support for the MDL metric, which is very general and is not designed with any of the particular patterns (or even with phonology) in mind: the same general metric that supports segmentation alone in simple cases allows us to handle straightforwardly the specific challenges discussed above.