Rule-based learning of phonological optionality and opacity

Ezer Rasin, Iddo Berger, Nur Lan, and Roni Katzir; MIT and Tel Aviv University

Summary. Optionality and opacity pose obvious challenges for the child learning the phonology of their ambient language: a process to be acquired loses support because of environments in which it was supposed to apply but didn't (both optionality and counterfeeding opacity) or in which it wasn't supposed to apply but did (counterbleeding opacity). Not surprisingly, no learners in the literature can handle optionality or opacity distributionally, from unanalyzed input data alone. Children, however, do manage to acquire both optionality and opacity in a variety of languages (Dell 1981, McCarthy 2007). This talk shows how optional processes and opaque interactions (including both counterfeeding and counterbleeding) can be acquired using the principle of Minimum Description Length (MDL; Solomonoff 1964, Rissanen 1978). Specifically, we use an adaptation of Rasin & Katzir 2016's MDL learner (originally used for OT phonology) to rule-based phonology and show how it applies to various cases of optionality and opacity. We then show simulations on artificial-language data illustrating the mechanization of the idea.

The MDL Principle. MDL is an evaluation criterion that balances two competing factors: the simplicity of the grammar ($|G|$; similarly to 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).

(1) 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'

We start by showing how MDL allows for the simultaneous induction of a segmented lexicon and phonological rules. **Segmentation**: (1) can allow the learner to discover the morphological 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 (1), 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**: (1) will also enable the learner to acquire various phonological processes. For example, if the language just discussed also has a process of progressive voicing assimilation across morpheme boundaries, the surface forms will seem to involve twice the actual number of suffixes, with one variant following voiceless stops and another elsewhere. Using (1) the learner will reject a naive encoding of this kind in favor of one where there is just one variant for each suffix, along with a rule of voicing assimilation. Likewise, if the language has a process that epenthesizes an [a] between coronals, a naive lexicon will store the epenthetic vowel underlyingly, but (1) will prefer an overall simpler grammar in which an epenthesis rule is added (thus adding some complexity) but its additional cost is more than offset by the savings obtained through not storing /a/ between coronals.

For each of the phenomena above, one can imagine a specialized learner that might acquire the relevant pattern. For example, approaches based on transitional probabilities have been proposed for the task of segmentation. And the distributional learner of Calamaro & Jarosz 2015 has been offered for tasks such as voicing assimilation. One advantage of the MDL metric in (1) is that it can acquire all these patterns in a simple, unified way. Another advantage, which we will capitalize on here, is that it can acquire both optional phonological processes and opaque interactions, neither of which can, to our knowledge, be captured using task-specific approaches from the literature. For example, if voicing assimilation is optional, we will occasionally find forms such as [dokza] (from /dok/+/za/), where the suffix starts with a voiced [z] despite appearing after a voiceless stop. Similarly for opacity: if voicing assimilation precedes epenthesis, we will find forms such as [dokasa] (from /dok/+/za/), where the suffix starts with a voiceless [s] despite appearing after a vowel (rather than after a voiceless stop). Forms of this kind look like counterexamples to the assimilation rule and will prevent a learner such as that of Calamaro & Jarosz from acquiring it. As we now show, MDL succeeds in learning both optionality and opaque interactions of this kind.

Simulation I: Optionality. This dataset shows a pattern modeled after optional word-final post-obstruent L(iquid)-deletion in French (Dell 1981). The learner needs to generalize beyond the data and conclude that for each pair like [tab]-[tabl] there is a single UR, and that a rule of L-deletion optionally applies.

But the learner must not overgeneralize and should restrict L-deletion to only apply word-finally after obstruents, despite the added complexity of specifying the restricted environment in the description of the rule. The data presented to the learner in the present simulation consisted of 19 words, including 4 collapsible pairs. The lexicon of the initial state in (2) was identical to the data and the rule set was empty. Search was performed using Simulated Annealing (Kirkpatrick et al. 1983) and converged on a hypothesis with the correct rule (including the specification of both context and optionality) and a lexicon in which each pair such as [tab]-[tabl] is represented with a single UR /tabl/ (3).

(2) Initial state. Rules=$\emptyset$. Lexicon={tab, tabl, lub, lubl, tap, tapl, rud, rudl, parl, birl, dail, lid, lup, puard, rlad, ruap, tbir, tid, tpul}. Description length: $|G_{initial}| + |D{:}G_{initial}| = 7,740 + 4,750 = 12,490$.

(3) Final state. Rules={$[+liquid] \rightarrow \emptyset / [-son]\_\_(optional)$}. Lexicon={tabl, lubl, tapl, rudl, parl, birl, tpul, tbir, tid, ruap, dail, rlad, puard, lup, lid}. $|G_{final}| + |D{:}G_{final}| = 6,493 + 4,200 = 10,693$.

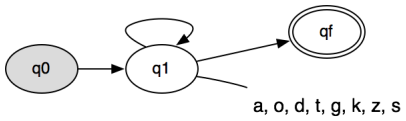
Simulation II: Opacity. We illustrate the use of MDL for opacity using a dataset involving counterbleeding along the lines described above (roughly modeled after English epenthesis and voicing assimilation with reversed order of the two processes; more directly following a similar opaque interaction in some varieties of English and Armenian, as in Vaux 2016, and in Iraqi Arabic, as in Kiparski 2000). With the opaque rule ordering (assimilation before epenthesis), the UR /kat-zoka/ undergoes assimilation (|kat-soka|) and then epenthesis ([katasoka]); with the opposite, transparent ordering (epenthesis before assimilation, as in English), the UR /kat-zoka/ would only undergo epenthesis (*[katazoka]).

For this simulation, the learner was presented with 48 words generated by creating all combinations of 12 stems (e.g., dog, kat) with 4 suffixes (e.g., $\emptyset$, -zoka) and applying voicing assimilation and a-epenthesis between coronals, in this order. Search was again conducted using Simulated Annealing. Given an initial grammar with no rules and with a lexicon that concatenates segments arbitrarily ($G_{initial}$ below), the learner successfully performed segmentation and converged on the expected lexicon and rules (G_{final} below).

$G_{initial} =$

RULES: $\emptyset$

LEX:



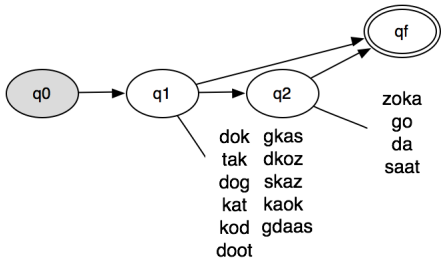
Length : $|G_{initial}| + |D{:}G_{initial}| = 97 + 35,300 = 35,397$

$G_{final} =$

RULES: $R_1 : [+cons] \rightarrow [-voice] / [-voice]\_\_$

$R_2 : \emptyset \rightarrow [+low] / [+coronal]\_\_[+coronal]$

LEX:



Length : $|G_{final}| + |D{:}G_{final}| = 402 + 7,175 = 7,585$

Implications. Our results show that optionality and opacity can be learned from distributional evidence alone. They provide further support for the MDL metric, which is very general and is not designed with optionality or opacity (or even with phonology) in mind. The same general metric that supports e.g., segmentation allows us to handle the challenge of acquiring optional phonological processes and opaque phonological interactions. We further note that the current approach allows us to quantify the difficulty of acquiring different phenomena along two separate dimensions: (a) precisely how much data are needed for acquisition given specific representational assumptions (see also Hsu & Chater 2010); and (b) how hard the search is given a fixed amount of data.