

Learning abstract URs from distributional evidence

Ezer Rasin and Roni Katzir; MIT and Tel Aviv University; rasin@mit.edu, trifilij@mit.edu

Summary: Human learners have been argued to infer URs that are sometimes different from their corresponding surface representation (SR) even without being forced to do so by an alternation. McCarthy (2005) proposes the Free Ride Principle (FRP), which allows the learner to extend nonidentical mappings in alternating forms to non-alternating forms. Recently, Rasin & Katzir (2016) proposed Minimum Description Length (MDL) as an evaluation metric for OT. The present work shows how MDL supports the induction of abstract, nonidentical URs from distributional evidence alone, both in cases that have motivated the FRP and in cases that have been argued to involve nonidentical URs in the absence of supporting alternations (and are thus outside the scope of the FRP). As a result, the FRP is redundant.

The Free Ride Principle: McCarthy (2005) discusses evidence from languages like Choctaw, Japanese, Rotuman, and Sanskrit, in which some non-alternating URs are claimed to be distinct from their SRs. These languages pose a challenge to a family of proposals regarding learning in phonology, *paradigm-based lexicon learners*, which posit URs that are identical to SRs outside of alternations (Tesar 2006, 2009, 2014; Apoussidou 2007; Merchant 2008; and Akers 2012). McCarthy augments paradigm-based learners with a principle which allows the learner to extend nonidentical mappings in alternating forms to non-alternating forms. In particular, a learner that infers the mapping /A/ → [B] based on alternations will try to derive every surface [B] from an underlying /A/. Example (Sanskrit coalescence): based on evidence from alternations as in (1), the learner will infer the mapping /ai/ → [e:] and will try to derive all surface instances of [e:], including non-alternating ones as in (2), from /ai/; positing an identical UR /be:/ for [be:] leads paradigm-based learners into the subset problem of language acquisition, as they converge on an overly faithful constraint ranking (*DIPH ≫ ... ≫ ID[VH] ≫ *MID) that overgenerates short mid vowels, which are absent in the language (e.g., *[be] from /be/). Positing unfaithful URs as in (2) yields a more restrictive ranking (*DIPH ≫ ... ≫ *MID ≫ ID[VH]) which bans mid vowels unless they are derived from a diphthong. McCarthy argues that there is no general solution for the problem that admits identical URs and works for a variety of cases similar to Sanskrit.

(1) /tava indra/ → tave:ndra ‘for you, Indra (voc.)’

(2) SR: [be:] UR: /bai/

The FRP is not enough: Alderete & Tesar (2002) note that stress patterns in several languages (e.g., Mohawk, Selayarese, Yimas) require the acquisition of URs that are not identical to the SR. Significantly, A&T argue that in these cases there are no alternations to support the induction of a nonidentical UR. Language *B* below illustrates:

Lg. A	Lexical stress	pákat	pakát	pákit	pakít	píkat	pikát	píkit	pikít
Lg. B	Stress-epenthesis		pakát	pákit	pakít		pikát	píkit	pikít
Lg. C	Final stress		pakát		pakít		pikát		pikít

A and *C* are straightforward: *A* has lexical stress and *C* final stress; in both the URs can be identical to the SRs. *B* is more problematic: it has final stress in general, but when the final vowel is [i], stress can be penultimate. A standard analysis would take *B* to have final stress, treating unstressed final [i] as epenthetic. But this means that the relevant URs do not have the final [i] appearing on the surface, and there is no alternation to provide the crucial evidence for non-identity. Paradigm-based learners that assume identical URs thus converge on an overly inclusive analysis of *B* as a lexical-stress language (e.g., predicting that an impossible *[pákat] would be a possible word). In this case, since there is no paradigmatic evidence for the unfaithful mapping, the FRP will be of no help.

Minimum Description Length: Rasin & Katzir (2016) proposed MDL as an evaluation metric for OT and showed that it can be used to learn URs (sometimes different from their SRs), con-

straint rankings, and the constraints themselves (both markedness and faithfulness constraints). Following MDL, the learner attempts to minimize the overall description of the data, measured in bits. The overall description is broken down into G , the encoding of the grammar (which, for OT, includes both the lexicon and the constraints), and $D : G$, the description of the data D given the grammar. MDL is a general principle and is not designed with URs (or even phonology) in mind. The same general metric that supports e.g., morphological segmentation and constraint induction allows us to handle the challenge of acquiring nonidentical URs. MDL will prefer a nonidentical UR whenever this leads to a shortening of the overall description of the data. For example, in cases of epenthesis, avoiding the statement of the epenthetic segment in the lexicon will shorten the lexicon and justify acquiring the appropriate constraints and their ranking for epenthesis. MDL can also lead to a nonidentical UR if a certain segment can be eliminated altogether from the alphabet of the lexicon: a shorter alphabet means that each remaining element requires fewer bits to encode, thus leading to a shorter statement of the grammar. This amounts to the acquisition of a morpheme-structure constraint.

Illustration: Sanskrit. To see why the correct grammar for Sanskrit is the one with the shortest description length, consider first the alternation in (1). A grammar without coalescence will have to store morphologically-complex forms like [tave:ndra] as undecomposed URs (/tave:ndra/) alongside individual items like /tava/ and /indra/. Learning a ranking that enforces coalescence would allow eliminating forms like /tave:ndra/ from the lexicon and would provide a compressional benefit. With coalescence in place, the MDL learner can eliminate underlying /e:/ by limiting the alphabet (an instance of a morpheme-structure constraint). Both the URs /bai/ and /be:/ would yield the surface [be:]; if /ai/ and /e:/ are equally costly (say, because two segments cost as much as a long segment), an underlying /bai/ for the non-alternating [be:] would be as good as /be:/; and choosing /ai/ over /e:/ for *every* surface [e:] would allow banning /e:/ from the alphabet and would lead to further compression in the lexicon. Similarly, banning the short /e/ from the alphabet would lead to compression in the lexicon and would block *[be].

Simulation: Stress-epenthesis interaction. The dataset for the simulation summarized here is an artificial language modelled after Yimas stress-epenthesis interaction, and is a mirror-image of Language B above: stress in bisyllabic words is penultimate but can be ultimate if the first vowel is [i]. The MDL learner of Rasin & Katzir 2016 arrives at a short lexicon where stress and the relevant instances of [i] are absent and the grammar inserts them in the right positions. The removal of stress from the alphabet achieves further compression and ensures that ungrammatical outputs (e.g., *[pakát]) are blocked.

(3) **Data:** {*tí, púk, kátu, kúit, píkat, tipú, kipúk*}

(4) **Initial state:** $\Sigma = \{t, p, k, a, i, u, '\}$ Lex: {*tí, púk, kátu, kúit, píkat, tipú, kipúk*}

CON: FAITH \gg DEP \gg MAINLEFT \gg *CC \gg HEADDEP (simplified)

(5) **Final state:** $\Sigma = \{t, p, k, a, i, u\}$ (No stress marking in the lexicon) Lex: {*ti, puk, katu, kúit, píkat, tpu, kpuk*}

CON: HEADDEP \gg *CC \gg DEP \gg MAINLEFT \gg FAITH

Implications: Our results show that there is no need for a specialized mechanism such as the FRP in order to account for unfaithful URs. The same general MDL metric that has been shown to acquire other aspects of morpho-phonology correctly guides the learner to unfaithful URs in appropriate cases. Examples such as those of A&T, where unfaithful URs are posited in the absence of supporting alternations (and are thus outside the reach of the FRP), provide support for this approach to learning. In acquiring the appropriate unfaithful URs, the MDL learner relies on the ability to restrict the alphabet of the lexicon. That is, the learner adopts a morpheme-structure constraint and violates the principle of Richness of the Base (ROTB). Assuming this is indeed the correct approach to learning, we thus have an indirect argument in favor of MSCs and against ROTB.