Learning abstract underlying representations from distributional evidence[*]

Ezer Rasin & Roni Katzir

Massachusetts Institute of Technology and Tel Aviv University

1. Introduction

Human learners have been argued to acquire underlying representations (URs) that are sometimes different from their corresponding surface representation (SR) even without being forced to do so by an alternation. Consider, for example, the case of coalescence in Sanskrit discussed in McCarthy (2005). In Sanskrit, the underlying sequence /ai/ undergoes coalescence and surfaces as the long mid vowel [eː], as in (1). While some long mid vowels have alternants that suggest an underlying /ai/, as in (1), for some long mid vowels the learner will not encounter any alternations supporting a coalescence source, as in the simple constructed example in (2).

(1) /tava indra/ → [taveːndra] 'for you, Indra (voc.)'

(2) [beː] (non-alternating)

Given a grammar with coalescence, a non-alternating form such as (2) can be represented in the lexicon in one of two ways. It can be represented either with an identical UR (/beː/), or with the underlying sequence /ai/, as in /bai/, in which case the correct surface form will be derived by the grammar through coalescence. McCarthy (2005) argued that the correct representation within Sanskrit of a form like [beː] is the abstract UR /bai/, despite the absence of supporting evidence from an alternation in this case. In particular, he argued that learners that posit the faithful UR (e.g., the UR /beː/ in the case of the surface form [beː]) converge on an incorrect, over-generating grammar for Sanskrit.

We will use the term *abstract URs* to informally refer to URs such as /bai/ – that is, URs which contain deviations from SRs that are not directly supported by an alternation. Abstract URs were occasionally assumed within early generative phonology: as a well-known example, Chomsky & Halle's (1968) proposed representation of the English word *nightin-*

[*] We thank Adam Albright, Naomi Feldman, Michael Kenstowicz, Donca Steriade, and the audiences at MIT and NELS 48.

gale contained the voiceless velar fricative /x/ which never surfaces in English.[1] And while recent work has for the most part avoided abstractness in phonological representations, sometimes adopting principles such as Lexicon Optimization (Prince & Smolensky 1993) that prevent abstract URs from being learned altogether, arguments such as McCarthy's have re-opened the question of whether abstract URs are needed in phonology. If such arguments are correct, as we will assume in this paper, they require accepting abstract URs and addressing their representation and learning.[2]

Abstract URs pose a challenge for the child learning the phonology of their language. Assuming that the child is only exposed to distributional evidence – that is, unanalyzed surface forms, with no further assistance from paradigms or URs – how will the child infer that a UR is distinct from its corresponding SR without direct evidence from alternations?

The present paper addresses this challenge by showing how a general approach to learning allows abstract URs to be acquired from distributional evidence alone. Our approach to learning is based on the principle of Minimum Description Length (MDL; Solomonoff 1964, Rissanen 1978), which provides a simple, unified framework for the acquisition of grammatical knowledge. Recently, Rasin & Katzir (2016) have used MDL to show how complete phonological grammars can be acquired distributionally within constraint-based phonology, including URs (that are sometimes different from their SRs), constraint rankings, and the constraints themselves (both markedness and faithfulness constraints).[3] In section 2 we briefly review Rasin & Katzir's (2016) proposal and discuss its implications for the acquisition of abstract URs. We show that the MDL metric can correctly select abstract URs in the case of Sanskrit.

In section 3 we compare the MDL metric to McCarthy (2005)'s proposal for learning abstract URs. To account for cases like Sanskrit, McCarthy proposed the Free Ride Principle (FRP), which allows the learner to infer abstract URs in some cases, by extending unfaithful mappings from alternating to non-alternating forms. Given the discussion in section 2, our first conclusion is that there is no need for a specialized principle for learning abstract URs: the same general MDL metric that allows us to learn other aspects of morpho-phonology correctly guides the learner to abstract URs in appropriate cases. However, we will be able to strengthen this conclusion by examining cases of abstract-UR learning from the literature where MDL succeeds and the FRP fails. In those cases, there are no alternations at all, so the FRP is of no help, while MDL successfully acquires the relevant URs.

[1] See Kiparsky 1968, Hyman 1970, Yip 1996, Nevins & Vaux 2007, and Krämer 2012, among many others, for relevant discussion of abstract URs.

[2] In addition to the Sanskrit case, McCarthy (2005) argued that human learners select abstract URs in a range of similar cases in languages like Choctaw, Japanese, Rotuman, and Arabic. Other arguments in the literature for abstract URs include those in Alderete & Tesar 2002 and Nevins & Vaux 2007.

[3] In what follows, we will discuss MDL learning in the context of constraint-based phonology, but we do so only to keep the presentation simple and close to the discussion in the recent literature. The MDL metric itself is not tied to this or any other particular representational formalism. For the MDL learning of rule-based phonology see Rasin et al. 2018a,b.

2. MDL and its implications for the acquisition of abstract URs

The MDL metric balances two competing factors: (a) the simplicity of the grammar, $|G|$; and (b) the tightness of fit of the grammar to the data, $|D : G|$.[4] According to MDL, the best grammar is the one that minimizes the sum of $|G|$ and $|D : G|$, as stated in (3). The grammar G includes not just the constraints and their ordering but also the lexicon.

(3) 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'

As mentioned above, the MDL metric has been shown to acquire complete phonological grammars within constraint-based phonology, including URs, the constraints, and their ranking, at least in certain simple cases. While MDL is a general metric that is not designed with UR-learning (or even phonology) in mind, one of its consequences is that abstract URs should be learned whenever they lead to a shorter description length of the data:

(4) MDL IMPLICATION FOR THE ACQUISITION OF ABSTRACT URS
Abstract URs are learned when they decrease the combination of $|G|$ and $|D : G|$

We will illustrate (4) using the Sanskrit case, by explaining why the correct grammar for Sanskrit – the one with abstract URs – leads to the shortest description length of the data. We will do so in two steps: first, we will explain why coalescence would be learned to begin with. Then, we will explain why abstract URs would be preferred given that coalescence has been learned. For the first step, consider the alternation in (1), repeated here:

(5) /tava indra/ → [taveːndra] 'for you, Indra (voc.)'

A grammar without coalescence would not be able to derive SRs like [taveːndra] from a combination of individual words. Consequently, a grammar without coalescence would have to store morphologically-complex forms like [taveːndra] as undecomposed lexical items (/taveːndra/) alongside individual items like /tava/ and /indra/, as in (6a). Learning a ranking that enforces coalescence allows eliminating forms like /taveːndra/ from the lexicon, which leads to a significant simplification to the grammar, as in (6b). So long as this simplification is not offset by an increase in $|D : G|$, coalescence will be acquired.

(6) *Two preliminary grammars for Sanskrit*

a. G_1 (complex): /taveːndra/, /tava/, /indra/, . . . , no coalescence

b. G_2 (simple): /tava/, /indra/, . . . , a ranking that enforces coalescence

The acquisition of coalescence completes the first step, and, once it is in place, both the URs /bai/ and /beː/ will yield the surface non-alternating [beː]. We now turn to the second step and show how the MDL metric leads to abstract URs. Specifically, we will show why abstract URs such as /bai/ will be preferred to /beː/. To do so, we need to consider another component of the grammar: the alphabet with which URs are written, which we indicate

[4] Where G is the grammar, D the data, and $D : G$ the encoding of the data given the grammar. Both $|G|$ and $|D : G|$ are typically measured in bits.

using the symbol Σ. For the present step, we will need Σ to be modifiable through learning, which means that it can differ across languages (for example, the Σ of English might not include /x/, which would mean that URs with /x/ could not be written). Allowing Σ to vary between languages amounts to accepting language-specific constraints on URs, and we note that this is at odds with common assumptions within current theoretical phonology. Constraints on URs were assumed in early generative phonology but were abandoned in later work, and the OT principle of Richness of the Base (Smolensky 1996) explicitly prohibits language-specific constraints on URs from being stated in the grammar. However, we believe that allowing Σ to vary between languages is generally required for an MDL learner to successfully acquire the phonological knowledge that speakers have (see Rasin & Katzir 2018 for evidence for this claim based on allophonic patterns that, superficially at least, seem quite different from the Sanskrit case). As we now show, allowing Σ to vary will also lead an MDL learner to correctly posit abstract URs in cases like Sanskrit.

Consider the two candidate grammars for Sanskrit in (7). G_1 represents SRs like [beː] as the identical UR /beː/ and has the vowel eː in its alphabet. G_2 represents SRs like [beː] as the abstract UR /bai/ and does not include eː in its alphabet, which means that URs with /eː/ cannot be written as part of the lexicon of G_2: every surface [eː] must be derived through coalescence.

(7) *Two grammars for Sanskrit*

a. G_1 (complex):
- Σ = {a, i, eː, ...}
- /beː/ /tava/, /indra/ + ranking that enforces coalescence

b. G_2 (simple):
- Σ = {a, i, ...} (no eː in the lexicon)[5]
- /bai/ /tava/, /indra/ + ranking that enforces coalescence

Eliminating a segment from Σ as in G_2 decreases $|G|$ without increasing $|D : G|$, making G_2 preferable to G_1 in terms of MDL. There are two ways in which eliminating a segment from Σ can decrease $|G|$. First, if Σ is listed as part of the grammar (as is done in Rasin & Katzir 2016), its size will contribute to $|G|$, so eliminating elements from Σ will lead to a simpler grammar. Second, a smaller alphabet allows for a more compact encoding of each segment within the lexicon. It is convenient to think of this in terms of the sequences of bits that are needed to specify each segment from within Σ. If Σ has just two elements, a single bit will suffice for specifying which of them is chosen. If Σ is larger, at least some choices from within it will require longer sequences of bits. For example, if there are four segments, one can assign a unique two-bit code to each. If there are more than four elements in Σ, specifying at least some of them will require more than two bits. And so on. So eliminating elements from Σ will generally allow shorter specifications to be used for the remaining segments, with the result that the lexicon can be stated using fewer bits than with the original alphabet, which in turn means that $|G|$ is shorter with a smaller alphabet.

[5] The alphabet will exclude additional segments not used in the Sanskrit lexicon, such as the short mid vowel [e].

While eliminating eː from the alphabet, as is done in moving from G_1 to G_2, decreases $|G|$, it makes no difference with respect to $|D : G|$ for the cases under consideration here: $|D : G_1| = |D : G_2|$. This is because, in the absence of any optional processes affecting the mapping from URs to SRs, specifying an SR amounts to specifying the relevant UR within the lexicon, and this specification is equally easy in G_1 and G_2.[6] We can conclude that the enforcement of abstract URs for the relevant forms, as done in G_2, decreases the combination of $|G|$ and $|D : G|$ and thus falls out as a by product of the general MDL metric.

3. Comparison with McCarthy's (2005) Free Ride Principle

McCarthy (2005), who argued that abstract URs are required for Sanskrit and a variety of other cases, also proposed a specialized mechanism for inducing abstract URs – the Free Ride Principle (FRP) – which allows a learner to extend an unfaithful mapping from alternating to non-alternating forms:

(8) The Free Ride Principle (McCarthy 2005)
A learner that infers the mapping /A/ → [B] based on alternations will try to derive every surface [B] from an underlying /A/

Applied to Sanskrit, a learner equipped with the FRP will infer the mapping /ai/ → [eː] based on evidence from alternations as in (1), and will try to derive all surface instances of [eː], including in non-alternating ones like (2), from /ai/. If our proposal for learning abstract URs using MDL is correct, a specialized mechanism like the FRP is redundant: the same metric that allows acquiring other aspects of morpho-phonology also takes care of abstract URs. In the present section we go one step further in supporting MDL over the FRP. We discuss a case of abstract-UR learning where no alternations are involved, so the FRP is of no help, but MDL succeeds.

Alderete & Tesar (2002) suggest that stress patterns in several languages (e.g., Mohawk, Selayarese, Yimas) require the acquisition of URs that are not identical to the SR and that, significantly, in these cases there are no alternations to support the induction of a nonidentical UR. Consider the following, simplified example, modeled after Yimas as described in Alderete & Tesar 2002. In this language, stress in bisyllabic words is initial but can be pen-initial if the first vowel is [i]. The following table illustrates:

(9) *Position of stress in a language modeled after Yimas*

	Initial vowel = i	Initial vowel = a
Initial stress	píkut	pákut
Pen-initial stress	pikút	*pakút

A standard analysis of this case would take the language to have initial stress, treating unstressed initial [i] as epenthetic:

(10) /pkut/ → [pikút]

[6] See Rasin & Katzir 2016 and Rasin et al. 2018a,b for discussion of how optionality affects $D : G$.

But this means that the relevant URs do not have the final [i] that appears on the surface, and there is no alternation to provide the crucial evidence for non-identity. Alderete & Tesar show that learners that assume identical URs converge on an overly inclusive analysis of the language as a lexical-stress language (e.g., predicting that an impossible *[pakút] 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.

Differently from the FRP, the MDL metric does not rely on paradigms and can learn both the mapping and the relevant abstract URs without evidence from alternations. To see how the MDL metric can work in the case of a language like Yimas, consider the toy dataset in (11), in which stress is generally initial but can be pen-initial if the first vowel is [i], and suppose the constraints available to the child are the ones in (12).

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

(12) *Constraints (cf. Alderete & Tesar 2002)*

a. MAINLEFT: Main stress falls on the leftmost syllable

b. HEADDEP: No epenthetic vowel in a stressed syllable

c. $^{*}CC$: No sequence of two consonants

d. DEP: No epenthesis

e. FAITH: No changes between UR and SR

As a first grammar – one that the learner will need to avoid – consider (13), which has: (a) a constraint ranking in which all faithfulness constraints outrank all markedness constraints; and (b) a lexicon that stores faithful URs for all surface forms. This is an overly general grammar that treats stress as lexical and can generate an impossible form like *[pakút] from the UR /pakút/. In other words, (13) is exactly the problematic grammar that needs to be avoided. As we will now show, the MDL metric guides the learner away from this grammar and toward a correct, restrictive grammar that treats stress as initial and handles apparent exceptions such as [tipú] through epenthesis.

(13) *Overly inclusive grammar*

- $\Sigma = \{t, p, k, a, i, u, '\}$
- Lex: {*tí*, *púk*, *kátu*, *kúit*, *píkat*, *tipú*, *kipúk*}
- CON: FAITH≫DEP≫MAINLEFT≫$^{*}CC$≫HEADDEP (simplified)

There are several ways in which the overly inclusive grammar in (13) is also suboptimal in terms of MDL. First, note that all instances of unstressed initial [i] can be left out of the lexicon and inserted by the input-output mapping rather than stored as part of the relevant URs, and leaving those instances out will lead to a shorter lexicon than the faithful one in (13). If stress is represented as a distinct symbol (as in (13)), further compression can be obtained by removing all instances of stress from the URs and inserting stress, just like epenthetic [i], through the input-output mapping. That is, the pressure to reduce the size of the grammar leads to a short lexicon where stress and the relevant instances of [i] are absent from the URs and the grammar inserts them in the right positions. The result of this shortening, specified in (14), amounts to a treatment of the existing forms in the Yimas lexicon in terms of grammatical (specifically, initial) stress and is an improvement over the initial hypothesis.

(14) *Correct grammar I (original alphabet)*

- $\Sigma = \{t, p, k, a, i, u, '\}$
- Lex: {*ti*, *puk*, *katu*, *kuit*, *pikat*, *tpu*, *kpuk*}
- Con: HeadDep≫*CC≫Dep≫MainLeft≫Faith

If, as discussed above in the case of Sanskrit, it is possible to modify Σ, a further step of compression becomes available: the learner has already removed all instances of stress from the lexicon in the previous step, but now the learner can remove stress from the alphabet in which the lexicon is written. As before, eliminating elements Σ is beneficial since it allows the remaining elements to be specified more briefly, which in turn allows the entire lexicon to be written using fewer bits. With the resulting grammar, stated in (15), it is not just the case that abstract URs are acquired: non-abstract URs with stress marking cannot even be represented.

(15) *Correct grammar II (smaller alphabet)*

- $\Sigma = \{t, p, k, a, i, u\}$ (**No stress marking in the lexicon**)
- Lex: {*ti*, *puk*, *katu*, *kuit*, *pikat*, *tpu*, *kpuk*}
- Con: HeadDep≫*CC≫Dep≫MainLeft≫Faith

To verify that the reasoning just described works in practice, we ran a simulation of the MDL learner of Rasin & Katzir (2016) on the dataset in (11). The initial grammar was the overly inclusive grammar in (13) (with a description length of 335 bits) and, as expected, the learner converged on a grammar as in (15) (with a description length of 305 bits).[7]

4. Summary

We have argued that there is no need for a specialized mechanism such as the FRP in order to account for the learning of abstract URs. The same general MDL metric that supports the learning of other aspects of morpho-phonology also correctly guides the learner to abstract URs in appropriate cases. Moreover, examples such as those of Alderete & Tesar, where abstract URs are posited in the absence of supporting alternations, are acquired by MDL but fall beyond the reach of the FRP, thus providing support for the MDL approach to learning.

[7] The full details of the simulation are available at: https://raw.githubusercontent.com/taucompling/otml/master/source/logging/yimas_simulation.txt.

References

Alderete, John, & Bruce Tesar. 2002. Learning covert phonological interaction: an analysis of the problem posed by the interaction of stress and epenthesis. Technical Report RuCCS-TR-72, Rutgers Center for Cognitive Science, Piscataway, NJ.

Chomsky, Noam, & Morris Halle. 1968. *The sound pattern of English*. New York: Harper and Row Publishers.

Hyman, Larry M. 1970. How concrete is phonology? *Language* 46:58–76.

Kiparsky, Paul. 1968. How abstract is phonology? Indiana University Linguistics Club.

Krämer, Martin. 2012. *Underlying representations*. Cambridge University Press.

McCarthy, John J. 2005. Taking a free ride in morphophonemic learning. *Catalan Journal of Linguistics* 4:19–56.

Nevins, Andrew, & Bert Vaux. 2007. Underlying representations that do not minimize grammatical violations. In *Freedom of analysis?*, ed. Sylvia Blaho, Patrik Bye, & Martin Krämer, 35–61. Mouton de Gruyter.

Prince, Alan, & Paul Smolensky. 1993. Optimality theory: Constraint interaction in generative grammar. Technical report, Rutgers University, Center for Cognitive Science.

Rasin, Ezer, Iddo Berger, Nur Lan, & Roni Katzir. 2018a. Learning phonological optionality and opacity from distributional evidence. In *Proceedings of NELS 48 (to appear)*, ed. Sherry Hucklebridge & Max Nelson.

Rasin, Ezer, Iddo Berger, Nur Lan, & Roni Katzir. 2018b. Learning rule-based morphophonology. Ms., MIT and Tel Aviv University, http://ling.auf.net/lingbuzz/003665, February 2018.

Rasin, Ezer, & Roni Katzir. 2016. On evaluation metrics in Optimality Theory. *Linguistic Inquiry* 47:235–282.

Rasin, Ezer, & Roni Katzir. 2018. A conditional learnability argument for constraints on underlying representations. Ms., MIT and TAU, March 2018.

Rissanen, Jorma. 1978. Modeling by shortest data description. *Automatica* 14:465–471.

Smolensky, Paul. 1996. The initial state and 'richness of the base' in Optimality Theory. Technical Report JHU-CogSci-96-4, Johns Hopkins University.

Solomonoff, Ray J. 1964. A formal theory of inductive inference, parts I and II. *Information and Control* 7:1–22, 224–254.

Yip, Moira. 1996. Lexicon optimization in languages without alternations. *Current trends in phonology: Models and methods* 2:757–788.

Ezer Rasin, Roni Katzir
rasin@mit.edu, rkatzir@post.tau.ac.il