A learnability argument for constraints on underlying representations*

Ezer Rasin and Roni Katzir

September 12, 2017

1 Where are phonological generalizations captured?

As noted by Halle (1962) and Chomsky and Halle (1965), speakers judge some nonce forms as nonexistent but possible – that is, as *accidental gaps* – and other nonce forms as nonexistent and impossible – that is, as *systematic gaps*. In English, for example, forms such as [k^hæg] and [æk] are accidental gaps, while *[kæg] and *[æk^h] are systematic gaps. Capturing this distinction in speakers' judgments is a central task of phonological theory, and it involves answering two questions. First, how is the distinction between the two kinds of gaps represented? And second, since the judgments of speakers regarding nonce forms differ between languages, how is the relevant knowledge acquired? In what follows, we point out a dependence between the two questions: on certain assumptions about the learning process, the phonological representations must allow for language-specific constraints on underlying representations (CURs), along the lines of the morpheme-structure constraints of early generative phonology.

To set the stage for our argument, let us briefly review the two main views in the literature on the representations behind phonological well-formedness judgments. Early generative approaches relied on a combination of two factors: constraints on underlying representations in the lexicon;[1] and phonological rules. In the example above, and assuming a highly simplified version of the aspiration pattern in English, an early generative account might use a CUR such as (1) and a phonological rule such as (2) as the basis for capturing the distribution of aspiration in English:[2]

(1) CUR IN ENGLISH: No aspirated stops in the lexicon

[*] Acknowldgments: To be added.

[1] Halle (1959, 1962) proposed to capture the relevant generalizations through rules that apply to URs. Stanley (1967) argued that these should be constraints rather than rules. In the generative tradition these became known as morpheme-structure constraints. We use CURs as a cover term for rules or constraints of this kind.

[2] Here and below we use *h* as shorthand for *spread glottis*. In the characterization assumed here, stops are aspirated exactly when they are voiceless and appear prevocalically, but this is a clear oversimplification. A more adequate characterization, following Kahn 1976, is that voiceless stops are aspirated in an initial position within a (stressed or initial) syllable. To further simplify notation, we will omit [−*voice*] from the statements of phonological processes below. As far as we can tell, the simplifications we adopt are orthogonal to the argument we will present.

(2) $[-\textit{voice}, -\textit{cont}] \rightarrow [+h] \,/\, \_\_[-\textit{cons}]$

(1) ensures that stops will be unaspirated underlyingly, while (2) ensures that they will become aspirated in exactly the right environment. Their combination handles the distinction between the accidentally missing [k^{h}æg] and [æk] on the one hand and the systematically missing *[kæg] and *[æk^h] on the other, on the assumptions that accidental gaps are those forms that can be derived by a new UR and without changing the rest of the grammar and that systematic gaps are those forms that would require a change to the rest of the grammar. The accidentally missing [k^{h}æg] and [æk] could be added to English with the URs /kæg/ and /æk/; the aspirating rule in (2) would then turn the former into its surface form. For *[kæg] and *[æk^h] the situation is different. Since (1) prohibits the storing of aspiration in the lexicon of English, aspiration must follow from rule application; but the aspiration rule in (2) does not apply to non-prevocalic stops, which leaves no way to derive *[æk^h]. For *[kæg], on the other hand, obligatory aspiration necessarily ensures that this surface form cannot appear. Both gaps are thus correctly treated as systematic.

CURs, then, offer one way in which patterns such as aspiration can be captured. A different way to capture the same pattern forgoes CURs and relies on phonological rules alone. For example, instead of stating that stops are unaspirated by default using a CUR, we could accomplish the same by a rule such as (3) below, which makes stops unaspirated regardless of their underlying specification or of their environment:

(3) $[-\textit{cont}] \rightarrow [-h]$

If (3) is ordered before (2), any UR would first have its stops made unaspirated, after which its prevocalic voiceless stops will be made aspirated. This would make the URs /kæg/ and /k^{h}æg/ surface as [k^{h}æg], while the URs /æk/ and /æk^h/ will surface as [æk]. The systematically missing *[kæg] and *[æk^h] will correctly be predicted to be impossible to derive.

We thus have two different ways to represent the distinction between accidental and systematic gaps. The first involves a combination of CURs and of phonological processes, and the second relies on phonological processes alone. The former approach was the one favored in early generative phonology: while the architecture assumed at the time allowed for both kinds of analysis, the one using CURs was taken to be preferred by the simplicity metric. The latter approach has been adopted within Optimality Theory (OT; Prince and Smolensky, 1993), where a representational principle, *Richness of the Base*, prevents CURs from being stated:[3]

(4) Richness of the Base (ROTB; Prince and Smolensky 1993, p. 191,Smolensky 1996, p. 3):

   a. All systematic language variation is in the ranking of the constraints.

[3] The discussion above uses phonological rules, but both approaches can just as easily be stated using OT constraints (which will be the representation used in section 2). Stated in terms of constraints, the first approach would combine the CUR in (1) with a constraint ranking such as $^*[-\textit{cont}, -h][-\textit{cons}] \gg \textsc{Ident}(h)$, while the second approach would avoid (1) and instead add a mid-ranking constraint banning aspirated stops, as in $^*[-\textit{cont}, -h][-\textit{cons}] \gg ^* [+h] \gg \textsc{Ident}(h)$. The question of whether to use CURs is thus separate from the choice between rules and constraints, and we will focus exclusively on the former in what follows.

b. In particular, there are no language-specific CURs.

Clearly, the two representational choices for how to handle the aspiration pattern are meaningfully different. For example, the use of CURs distributes the knowledge of patterns like aspiration between two distinct components of the grammar – CURs versus phonological rules or constraints – while ROTB leads to a unitary treatment of such patterns. This difference can lead to different ways in which various phenomena can be accounted for – for example, in loanword adaptation – but to date it has been hard to find empirical arguments for one view or the other. Below, we will show how considerations of learnability can be brought to bear on the choice.[4]

Turning to the question of how the relevant knowledge is acquired, our focus will be on a general approach to learning according to which the child attempts to make inductive inferences by balancing the simplicity of the grammar (or its prior probability) against its fit to the data (or the likelihood of the data). A preference for simplicity favors general grammars that do not overfit the data. However, simplicity on its own, as in the evaluation metric of early generative grammar, leads to grammars that are overly general. By balancing simplicity against tightness of fit, the learner can hope to find an intermediate level of generalization that is appropriate given the data.

Our argument, which we present in section 2, can be followed at the informal level of balancing the simplicity of grammar against tightness of fit. As far as we can tell, nothing in our discussion will depend on the specifics of how this balancing is formalized. For concreteness, however, we will frame our discussion of balanced generalization in terms of one particular formalization: namely, the principle of Minimum Description Length (MDL; Solomonoff 1964, Rissanen 1978), which we now briefly sketch, along with references for details and further discussion. For this sketch, it will be convenient to think of both grammars and their encoding of the data as sitting in computer memory according to a given encoding scheme. Using $|\cdot|$ to notate length, we can write $|G|$ for the length of the grammar G as measured in bits. The encoding of the data D using G will be written $D : G$, and the tightness of fit will be the length of this encoding, $|D : G|$. Using this notation, MDL can be stated as follows:[5]

(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'

The balancing of economy and tightness of fit has made MDL – and the closely related Bayesian approach to learning – helpful across a range of grammar induction tasks, in works such as Horning 1969, Berwick 1982, Ellison 1994, Rissanen and Ristad 1994, Stolcke 1994, Grünwald 1996, de Marcken 1996, Brent 1999, Clark 2001,

[4] In certain cases, CURs in rule-based phonology seem to conspire with phonological rules to enforce what looks like a unitary surface effect (see Chomsky and Halle 1968, p. 382). This fragmentation of explanation – the *duplication problem* of Kenstowicz and Kisseberth (1977, p. 136) – has been taken to be a challenge to rule-based phonology and an argument in favor of OT. Below we will argue that CURs should be brought back into phonological theory. In the cases we will discuss, however, we will see no obvious instantiation of the duplication problem, and we will not attempt to re-evaluate the problem here.

[5] Here and below the grammar G will be taken to be not just the phonological rules and their ordering (or the constraints and their ranking) but also the lexicon. Thus, by saying that a grammar G generates the data D, we mean that every string in D can be derived as a licit surface form from some UR in the lexicon and the relevant rules (or constraints).

Goldsmith 2001, and Dowman 2007, among others. Recently, this approach to learning has been used to provide learners for both constraint-based phonology (Rasin and Katzir 2016) and rule-based phonology (Rasin et al. 2017) that acquire a lexicon, the phonological processes involved, and their interactions, all from distributional evidence alone.

The MDL view predicts that the child will invest in grammatical statements only when the cost of the investment (in terms of increase in $|G|$) will be offset by the increase in tightness of fit to the data (in terms of decrease in $|D : G|$). Applied to the case of aspiration in English, the fact that *[kæg] and *[æk^h] are judged as ill-formed teaches us that the investment, through the relevant statements in G, in ruling out these forms has been justified by the shortening of $D : G$. The acceptability of [k^{h}æg] and [æk], on the other hand, teaches us that the benefits for $|D : G|$ in ruling out these forms are too small to justify an investment in the requisite grammatical statements.

In section 2 we use the MDL view of learning to argue that of the two representational choices above, only the one that supports CURs (and avoids ROTB) allows the child to acquire the knowledge of patterns such as aspiration in English. As we will note, a CUR such as (1) can shorten the encoding of the lexicon, thus offering a benefit in terms of MDL. A child who rejects ROTB will therefore be motivated to acquire such a CUR and will then correctly reject aspiration in 'elsewhere' environments. If the child follows ROTB, on the other hand, the pressure for simple grammars will prevent the knowledge about 'elsewhere' environments – which now must be a statement along the lines of (3) rather than a CUR – from being acquired: the pressure for simplicity leads to the removal of predictable information, such as aspiration in English, from the lexicon; and once we remove such information, there is no longer any reason for the learner to invest in a statement such as (3). But in the absence of such a statement, speakers would be predicted to judge a novel form such as *[æk^h] as acceptable, contrary to fact.[6]

Before proceeding, we note that the view of the child as inductive learner is not the only view on phonological learning in the literature. There is a prominent alternative according to which the child attempts to find the most restrictive grammar consistent with the data. On this view, the child starts with a maximally restrictive hypothesis about the world (typically assuming a finite number of innate markedness constraints penalizing various surface patterns); this hypothesis is gradually relaxed, with individual prohibitions being eliminated or demoted, in the face of conflicting evidence. Representative proposals of this kind include an initial ranking of markedness over faithfulness (M≫F; Smolensky 1996, Tesar and Smolensky 1998) from which a search for a consistent ranking begins, as well as a more sustained bias for M≫F (Hayes 2004, Prince and Tesar 2004) throughout the search for a consistent ranking. On this view, the child hypothesizes in advance that they should ban *[kæg] and *[æk^h], and since English provides no counter evidence, they never need to change their mind. What the

[6] The idea, central to the present work that that a general evaluation metric can help choose between competing representational possibilities, was pursued in several works in earlier generative grammar such as Halle 1962, 1978, Baker 1979, and Dell 1981. More recently, Katzir 2014 and Piantadosi et al. 2016 have raised the possibility of using MDL and Bayesian reasoning to evaluate competing architectures, thus returning to the kind of architecture comparison envisioned in early generative grammar but with a better supported approach to learning than the one assumed at the time.

acceptability of $[\text{k}^h\text{æg}]$ and [æk] teaches us, on this view, is that UG simply neglects to provide the means to ban these forms (without banning attested forms). Had it done so, the absence of $[\text{k}^h\text{æg}]$ and [æk] from the child's input data would have allowed the child to maintain the more restrictive hypothesis that these forms are impossible.

We will not argue directly in this paper against the view of the child as a maximally restrictive consistency seeker. We note that the inductive-learning approach that we rely on here has provided full distributional learners for both rule-based phonology and constraint-based phonology, while the restrictive consistency seeking approach has not yet done so. More broadly, we note that the inductive-learning approach, balancing the simplicity of hypotheses against their fit to the data, ties in with results about human generalization in a variety of tasks, such as word learning (Xu and Tenenbaum 2007) and visual scene chunking (Orbán et al. 2008). If humans indeed use this way of learning across different domains, it will be reasonable to consider their use of the same in phonology. We are not aware of similar considerations for restrictive consistency seeking. In view of this, it is at least interesting to examine the architectural implications of the inductive reasoning approach, which is what we propose to do in this paper.

2 MDL and the argument against ROTB

For concreteness, we present our argument for CURs and against ROTB within the framework of constraint-based phonology (though, as mentioned above, our argument is separate from the question of rules versus constraints). We will also stay with the example of aspiration in English where possible (though the argument for CURs can be made using any of a wide variety of patterns from different languages, and on a couple of occasions we will refer to such patterns).

We start in section 2.1 by presenting the argument within the setting in Rasin and Katzir 2016 in which the learner needs to acquire the constraints themselves rather than just their ranking and in which aspiration is a separate, costly segment rather than a feature. We then proceed, in section 2.2, to discuss the argument in the context of other representational choices. In section 2.3 we address a possible concern with the argument against ROTB: that phonotactic learning might allow the child to succeed even without CURs.

2.1 A naive encoding: constraints acquired; aspiration as a separate segment

Suppose that the child, using MDL, needs to acquire the constraints, with each additional constraint costing a positive number of bits, and suppose further that aspiration is a separate segment, ^{h}, that likewise costs a positive number of bits. Consider first the situation of the child on the assumption that ROTB holds. Since instances of aspiration are costly to store in the lexicon, it will be preferable in terms of MDL to invest in a markedness constraint that aspirates stops prevocalically (e.g., $^*[-cont, -h][-cons]$) and then store all forms as unaspirated. Adding the markedness constraint will cost a few bits, but this cost will be outweighed by the savings from not having to store any instances of aspiration in the lexicon. By following this reasoning, the child has successfully learned that prevocalic stops are systematically aspirated in the language. For example, the child will now correctly rule out forms such as *[kæg], with an unaspirated prevocalic [k].

Unfortunately for ROTB, this is also the extent of the child's acquisition of the aspiration pattern: the child will not learn to block forms such as *[æk^h], with aspiration in 'elsewhere' environments. The reason is that, for an ROTB child, such forms must be blocked through the input-output mapping, for example through $^*[+h]$ or a similar markedness constraint that penalizes aspiration. And there is simply no MDL justification for acquiring this kind of constraint. Recall that all forms are already stored in the lexicon as unaspirated. Consequently, a constraint that penalizes aspiration will be of no use in deriving the observed forms in the input data, and the cost of adding such a constraint to the grammar will not be justified. An ROTB child, then, will become an adult who knows only half of the aspiration pattern in English: they will correctly rule out *[kæg] but fail to rule out *[æk^h].[7]

If ROTB does not hold, the learning process can succeed in full. The first step is similar to the one with ROTB: the child will invest in a constraint like $^*[-cont, -h][-cons]$ and then store all forms as unaspirated, which will allow the child to correctly rule out forms like *[kæg]. But with the possibility of stating CURs, a crucial second step becomes available. The first step involved the extensional removal of all instances of aspiration from the lexicon. The child can now conclude that this was no accident, and that aspiration should be eliminated *intensionally* from the very alphabet in which the lexicon is written. That is, the child can reach the following conclusion (restating (1) above for the case of aspiration as a separate segment):

(6) CUR in English: Aspiration is not in the alphabet of the lexicon, $^h \notin \Sigma$

Let us first see why (6) is justified in terms of MDL. All things being equal, removing a possible segment from the underlying inventory makes the inventory itself smaller (if the size of the inventory is measured as part of $|G|$) and makes it slightly easier to specify the remaining segments, some of which may now cost fewer bits than before. Consequently, the lexicon will now be encoded with fewer bits, thus providing MDL justification for adopting (6). We can now see why, in a world that allows CURs, the child can go beyond what was possible under ROTB and acquire the second part of the aspiration pattern. The reason is that with a CUR like (6), the child will now correctly rule out surface forms like *[æk^h]: /æk^h/ is now no longer a possible UR; and this UR is the only potential source for this putative surface form. In other words, the impossibility of even stating aspiration in the lexicon, with its MDL justification, means that the learner has correctly learned to block illicit aspiration.

[7] The reasoning above makes no reference to alternations, but it is possible in principle that paradigms will allow the learner to acquire phonological knowledge that is otherwise hard or impossible to obtain. In the present case, one could imagine that alternations would provide MDL justification for learning that stops are unaspirated in 'elsewhere' environments, which, in turn, would allow the learner to acquire the full aspiration pattern without abandoning ROTB. As far as we can tell, however, the alternations that are actually available in English do not provide an MDL learner with the relevant information, either with the present representations or with those discussed below, and we will therefore continue to ignore their role in what follows.

We conclude that, under the representational choice of aspiration as a separate segment and constraints that need to be acquired, the ability to state CURs allows for the full pattern of aspiration to be acquired, while the adoption of ROTB leads to a failure in learning half of the pattern.

2.2 Other representational choices

The learnability argument against ROTB extends to some other representational possibilities that UG might make available. In the present subsection, we briefly review some such choices. We will try to show that on any reasonable representational choice, a version of the argument against ROTB can be made. And while there are representational choices that do not support the argument against ROTB, those choices do not seem plausible. This means that if one wishes to defend ROTB on the basis of one of these representations, one faces the task of defending what otherwise looks like a problematic representation.

In all the cases that we will consider, rejecting ROTB and allowing for CURs will resolve the problem and allow the child to acquire the full pattern of aspiration. This will follow from the same reasoning as for the choice of aspiration as a separate segment in section 2.1, and to keep the presentation simple, we will refrain from restating this point for each representational choice below.

2.2.1 Constraints given in advance

If the constraints are not acquired but rather given to the learner in advance, as is commonly assumed in the OT literature, a slightly more complex situation arises. We now turn to this case, summarizing the argument in Rasin and Katzir 2015 that an MDL learner would still need to abandon ROTB and adopt CURs.[8] Suppose that the learner is given the two relevant markedness constraints: $^*[-cont, -h][-cons]$, which penalizes unaspirated prevocalic stops; and $^*h$, which penalizes aspiration in general.

As in our discussion earlier, the constraint $^*[-cont, -h][-cons]$ poses no special problem for an MDL learner following ROTB. In the present setting – as in the earlier one – ranking this markedness constraint above the relevant faithfulness constraints will serve the compressional purpose of enabling the elimination of aspiration from all URs. As for $^*h$, the learner is now assumed to be given this constraint in advance; differently from the case of a learner that needs to acquire the constraints, the presence of $^*h$ will no longer incur costs in the present setting. However, the constraint still offers no compressional advantage. Consequently, the learner will not benefit from ranking this constraint above any faithfulness constraints, such as MAX(^{h}), that prevent underlying aspiration from being deleted. We would thus expect speakers to vary in the relative ranking of $^*h$ and MAX(^{h}). But this means, on ROTB, that speakers of English should differ in whether they accept forms such as *[æk^h] as possible, contrary to fact. In other words, for an MDL learner following ROTB that is given the constraints in advance, the problem lies not with the possibility of attaining the appropriate constraint ranking

[8] Here we discuss what seems to us the most straightforward version of the relevant constraints. See Rasin and Katzir 2015 for discussion of some other possibilities for the constraints.

but rather with ensuring that this ranking is attained systematically, for all speakers, and not just occasionally.

Before concluding our discussion of constraints that are given in advance, we note that the problematic ranking of $\textsc{Max}(^{h}) \gg {}^{*}\ h$ can be avoided – and with it the challenge to ROTB under the current representational choice – if the learner has a general preference for M≫F. As discussed in the introduction, variants of such a preference have been used within consistency-seeking approaches, which are not considered in this paper, to increase the restrictiveness of the grammars arrived at. Within inductive learning approaches, such as the MDL one discussed here, a preference for M≫F can be implemented in various ways, either as a separate principle or through the representations themselves. While possible, this move amounts to a commitment to a rather specific implementation of inductive learning that is not, as far as we are aware, supported empirically and would be made for the sole purpose of preserving ROTB. In view of this consideration, we will set aside the possibility of incorporating M≫F into an inductive learner in what follows but note that this matter should be revisited should independent evidence for M≫F emerge.

2.2.2 Binary aspiration feature with globally fixed costs

The argument against ROTB can also be made if, instead of a separate segment, aspiration is a simple binary feature, $+/-h$, with fixed costs for $+h$ and $-h$. However, the exact form of the argument depends on which of the two values is costlier, if either. We will consider the three different possibilities in turn.[9]

Consider first a representation in which $Cost(+h) > Cost(-h)$. As with the representation of aspiration as a separate segment, compression will lead the learner to remove all instances of aspiration from the lexicon – in the current representation, by storing $-h$ throughout – which in turn will motivate a constraint such as $^{*}[-cont, -h][-cons]$ for aspirating prevocalically but will not support a constraint against aspiration elsewhere. Just as with aspiration as a separate segment, then, an ROTB learner will fail to reject nonce words such as *[æk^{h}].

Consider now the possibility that $Cost(+h) = Cost(-h)$. In this case, compression cannot learn either part of the pattern in the absence of CURs: with fixed, equal costs for $+h$ and $-h$, compression will favor not learning any aspiration-related markedness constraints and simply storing URs that are identical to their corresponding surface forms in terms of aspiration. An ROTB learner will therefore fail to reject both *[æk^{h}] and *[kæg].

Suppose now that $Cost(+h) < Cost(-h)$. In this case, the learner will store $+h$ throughout the lexicon and acquire constraints that will ban aspiration in 'elsewhere' (roughly, non-prevocalic) environments. The precise constraints and their implications, however, can vary, depending on the complexity of describing the 'elsewhere' environments.

Suppose first that these environments are easy to characterize. In that case, a learner following ROTB will acquire a single constraint banning aspiration in those environ-

[9] Here we return to framing the discussion in terms of constraints that are acquired, rather than given in advance, which will allow us to avoid discussing different populations with different judgments. As we just saw, the general shape of the argument is very similar under the two choices.

ments – say, $^*[-cont, +h][+cons]$ (no aspiration before a consonant) – and be done. In particular, $^*[-cont, -h][-cons]$ will not be learned, and the target grammar will fail to ban impossible nonce words such as *[kæg]. This is the mirror image of the problem for ROTB in the previous setting, and again the ability to violate ROTB and remove elements from the alphabet of the lexicon will allow the learner to acquire the full pattern.

On the other hand, if the 'elsewhere' environment is difficult to describe, it will be costly to state a constraint that bans aspiration directly in these contexts, and it will make more sense for the child to learn to ban aspiration in general and allow it only prevocalically. That is, it will acquire a low-ranking $^*h$ to prevent underlying $+h$ from surfacing faithfully and a high-ranking $^*[-cont, -h][-cons]$ to ensure that prevocalic aspiration surfaces correctly. On this scenario, the learner will have correctly acquired the full pattern of aspiration without requiring a CUR, thus allowing ROTB to be maintained.

While this scenario provides a way to learn aspiration without CURs, it suffers from two problems: it relies on a cost assignment that seems to go against the typological pattern of markedness, and it makes problematic predictions about the learnability of various patterns with the same profile as aspiration. First, since aspiration is the typologically marked option, anyone who wishes to support ROTB based on this scenario would need to explain how to reconcile the typological markedness pattern (aspiration marked, lack of aspiration unmarked) with the cost assignment required for the argument (aspiration cheap, lack of aspiration costly). The second problem with $Cost(+h) < Cost(-h)$ is more troubling from the perspective of ROTB: the very asymmetry that allows the pattern of English aspiration to be acquired without CURs makes other imaginable patterns of aspiration unlearnable. Consider, for example, a language that has the mirror image of the pattern of aspiration in English, with stops that are unaspirated in some specific environments – say, before consonants – but are aspirated elsewhere. Given the present cost assignment, an ROTB learner will store all stops as $+h$ and then acquire a markedness constraint enforcing stops to surface without aspiration in the relevant environment. The same reasoning used earlier will prevent the learner from acquiring a constraint that enforces aspiration elsewhere (in this case, since all stops are already stored as $[+h]$ in the lexicon), which will result in an inability to rule out nonce forms without aspiration in 'elsewhere' environments.

Are such patterns of de-aspiration learnable? We suspect they are, though we are not familiar with typologically attested cases (or artificial-language learning experiments) that would confirm this. Either way, though, there are other, well-documented patterns where each value of a feature is the 'elsewhere' case in some language, thus allowing us to complete the argument against ROTB with binary features and global costs, regardless of which value (if either) is made costlier by UG. We illustrate using the feature [anterior]. In Dutch, the anterior strident [s] is realized as [ʃ] precisely before [j], and [ʃ] is banned elsewhere (see Booij 1995); in Bengali, the default sibilant is the non-anterior [ʃ], and [s] occurs only in word-initial consonant clusters and word-medially before dental stops (see Evers et al. 1998). Given the reasoning that we have now used repeatedly, ROTB using an asymmetric cost assignment cannot succeed on both Dutch and Bengali: assigning a greater cost to [-anterior] over [+anterior] will fail on the Dutch pattern, while assigning a greater cost to [+anterior] over [-anterior] will fail on Bengali; and equal costs will fail on both. On our current assumptions, then, succeeding in learning both Dutch and Bengali requires abandoning ROTB and using CURs.

2.2.3 Contextualized costs

Of course, there are many other cost assignment possibilities to consider in addition to the fixed, uniform ones outlined above. For example, suppose that UG makes the cost of $+h$ lower than that of $-h$ prevocalically and higher than it in other environments. In the absence of the ability to state CURs, this cost assignment will make both kinds of markedness constraints unlearnable by an MDL learner for the same reasons as discussed above.

For the opposite weighting scheme, with the cost of $+h$ higher than that of $-h$ prevocalically and lower than it in other environments, things are different. This scheme will allow both kinds of markedness constraints to be learned by an MDL learner, regardless of CURs. As with $Cost(+h) < Cost(-h)$, however, this scheme faces a couple of challenges. First, its cost assignment goes directly against the typological pattern of markedness. And second, it makes a suspicious prediction about the learnability of other aspiration patterns: in this case, it predicts that it is impossible for an MDL learner following ROTB to acquire the aspiration pattern in a language that has the exact opposite aspiration pattern from that of English, with non-aspiration prevocalically and aspiration elsewhere. To maintain ROTB, one would need to defend this cost scheme in the face of these challenges.

2.2.4 Underspecification

A more complex situation than with binary features arises if UG allows for underspecification, with $0h$ alongside $+h$ and $-h$, and if the encoding of $0h$ is shorter than that of $+/-h$.[10] In this case, compression will militate against storing the specified values $+/-h$ and in favor of storing $0h$ throughout, with markedness constraints ensuring that the correct aspiration values surface. Note that the markedness constraints will include not just the ones for aspirating prevocalically and against aspiration elsewhere but also a constraint against the surfacing of an unspecified value. This constraint, call it $*C^{0h}$, will need to outrank the faithfulness constraint Ident(^{h}) if it is to be useful. The other two markedness constraints, on the other hand, are equally useful if they outrank Ident(^{h}) or are outranked by it; and given that the present approach to learning does not rely on a principle such as $M \gg F$, such rankings will sometimes be acquired. The following shows the evaluation of the UR $/k^{0h}at^{0h}/$ using a constraint ranking in which faithfulness outranks the two relevant markedness constraints:[11]

[10] If the cost of $0h$ is not lower than that of the two specified values, there is no particular need to further consider underspecification: the reasoning used for binary features already derives the failure of ROTB and the need for CURs.

[11] The lowest-ranked constraint in the tableau is used as shorthand for whatever markedness constraints penalize aspiration in non-prevocalic environments.

(7)

	$/k^{0h}at^{0h}/$	$*C^{0h}$	Ident(^{h})	$*C^{-h}V$	$*C^{+h}/\neg(\_V)$
	$k^{0h}at^{0h}$	*!*			
	$k^{0h}at^{+h}$	*!	*		*
	$k^{0h}at^{-h}$	*!	*		
	$k^{+h}at^{0h}$	*!	*		
	$k^{+h}at^{+h}$		**		*!
☞	$k^{+h}at^{-h}$		**		
	$k^{-h}at^{0h}$	*!	*		
	$k^{-h}at^{+h}$		**	*!	*
	$k^{-h}at^{-h}$		**	*!	

The problem here is that if faithfulness outranks one or both of the relevant markedness constraints, as it does in (7), an inappropriately aspirated nonce form can be accommodated. That is, in addition to those English speakers who reject inappropriately aspirated nonce forms, we would incorrectly predict that some other English speakers would reject $k^{+h}at^{+h}$ but accept $k^{-h}at^{-h}$; that others would reject $k^{-h}at^{-h}$ but accept $k^{+h}at^{+h}$; while a fourth population would accept both $k^{+h}at^{+h}$ and $k^{-h}at^{-h}$.

2.3 Remaining concern: phonotactic learning

The above illustrated the argument against ROTB using the pattern of aspiration in English. Similar patterns abound, and the same argument could be made with other instances. In the present section we discuss a potential concern with the argument. We assumed that acquiring the pattern of aspiration is part of full phonological learning, where what is acquired is not just the constraint ranking (possibly along with the constraints themselves) but also the lexicon. It is conceivable, however, that the relevant constraint ranking is established at an early stage of purely phonotactic learning without reference to the lexicon (perhaps along the lines of Hayes and Wilson 2008's Maximum Entropy learner); this ranking could then be passed along to a second, more complete phase of phonological learning, which could proceed without requiring CURs.[12]

While aspiration in English might be acquired using a phonotactic learner, possibly without requiring phonological learning to abandon ROTB, for other patterns such an account seems less plausible. In particular, a phonotactic first step will run into difficulties with any pattern where the 'elsewhere' part is obscured and cannot be directly observed from surface forms alone. Opaque interactions are a case in point. For example, McCarthy (2007) discusses a case of opacity in Bedouin Arabic that allows us to replicate the argument against ROTB in a setting that is considerably more challenging for phonotactic learning than the pattern of aspiration in English. Bedouin Arabic has a process that palatalizes velar consonants before front vowels. It also has a process that deletes high vowels in certain environments. Palatalization precedes deletion, which results in certain velars surfacing as palatalized because of an underlying /i/ that then

[12] Such an approach would need to ensure that in the second, phonological stage, the constraint(s) banning aspiration in 'elsewhere' environments will not end up being ranked below the relevant faithfulness constraints. This is an instance of the challenge for ROTB in the case of constraints given in advance, and as discussed earlier it is not a trivial one. For the purposes of the present section, however, we will assume that there is a principled way to prevent the phonological-learning stage from ranking the markedness constraints for 'elsewhere' environments too low.

deletes, an instance of counterbleeding opacity (e.g., /haːkim-iːn/ [haːk^{j}m-iːn] ‘ruling(m.pl.)’). McCarthy provides evidence that these processes are active in speakers’ grammars and are thus learned.

Our argument from aspiration can be restated for the palatalization process. In particular, an ROTB learner who needs to induce the constraints in the phonology will fail to acquire the part of the pattern that says that velars should not be palatalized in ‘elsewhere’ environments (here, not preceding a front vowel). The reasoning is familiar by now: velars will be stored as unpalatalized in ‘elsewhere’ environments (unless, implausibly, they are cheaper to store as palatalized), which will give the child no reason to invest in a markedness constraint that penalizes palatalization; as an adult, they will then fail to rule out palatalized velars in ‘elsewhere’ environments.

Because of the opacity involved in this case, however, it is hard to see how an earlier phase of phonotactic learning might help: surface velars in ‘elsewhere’ environments appear sometimes as palatalized (when a following underlying /i/ was deleted) and sometimes as unpalatalized; consequently, a phonotactic constraint banning palatalized velars in ‘elsewhere’ environments will clash with the surface pattern and will most likely not be induced. Penalizing palatalization must be left for the lexicon-aware stage of full phonological learning, where, as we just saw, an ROTB learner will fail.

3 Summary

We started this paper by asking how the distinction between accidental and systematic gaps is represented and how the relevant knowledge is acquired. We pointed out a connection between the two questions: assuming that phonological knowledge is acquired using MDL (or a similar inductive approach), and across several different representational choices, we must allow CURs to be stated as part of the grammar if we wish to account for speakers’ ability to distinguish between the two kinds of gap. The general shape of the argument was this. The pressure for simple grammars, which is part of the MDL criterion, leads to the removal of predictable information, such as aspiration in our main example, from the lexicon. Once this information is eliminated, there is no longer any reason for the learner to invest in a statement – whether a rule or a constraint – that bans the appearance of this material. But in the absence of such a statement, speakers would be predicted to accept nonce forms in which this material appears in an inappropriate environment, contrary to fact. The solution, then, is to abandon ROTB and allow the learner to eliminate predictable material not just from the lexicon but, using a CUR, from the very alphabet in which the lexicon is written. We saw that various instantiations of this argument can be made across a range of representational choices. In a few cases, we saw that ROTB could be maintained if particular representations are adopted: for example, constraints given in advance along with an implementation within the representational costs of M$\gg$F; or contextualized costs in which aspiration is expensive prevocalically and cheap elsewhere. This served to highlight the challenge for ROTB: to defend this principle, on the assumption of inductive learning, one would need to provide support for specific, seemingly unnatural representations.

References

Baker, Carl L. 1979. Syntactic theory and the projection problem. *Linguistic Inquiry* 10:533–581.

Berwick, Robert C. 1982. Locality principles and the acquisition of syntactic knowledge. Doctoral Dissertation, MIT, Cambridge, MA.

Booij, Geert. 1995. *The phonology of Dutch*. Oxford: Clarendon Press.

Brent, Michael. 1999. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Computational Linguistics* 34:71–105.

Chomsky, Noam, and Morris Halle. 1965. Some controversial questions in phonological theory. *Journal of Linguistics* 1:97–138.

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

Clark, Alexander. 2001. Unsupervised language acquisition: Theory and practice. Doctoral Dissertation, University of Sussex.

Dell, François. 1981. On the learnability of optional phonological rules. *Linguistic Inquiry* 12:31–37.

Dowman, Mike. 2007. Minimum description length as a solution to the problem of generalization in syntactic theory. Ms., University of Tokyo, Under review.

Ellison, Timothy Mark. 1994. The machine learning of phonological structure. Doctoral Dissertation, University of Western Australia.

Evers, Vincent, Henning Reetz, and Aditi Lahiri. 1998. Crosslinguistic acoustic categorization of sibilants independent of phonological status. *Journal of Phonetics* 345–370.

Goldsmith, John. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics* 27:153–198.

Grünwald, Peter. 1996. A minimum description length approach to grammar inference. In *Connectionist, statistical and symbolic approaches to learning for natural language processing*, ed. G. S. S. Wermter and E. Riloff, Springer Lecture Notes in Artificial Intelligence, 203–216. Springer.

Halle, Morris. 1959. *The sound pattern of Russian*. Walter de Gruyter.

Halle, Morris. 1962. Phonology in generative grammar. *Word* 18:54–72.

Halle, Morris. 1978. Knowledge unlearned and untaught: What speakers know about the sounds of their language. In *Linguistic theory and psychological reality*, ed. Morris Halle, Joan Bresnan, and George A. Miller, 294–303. MIT Press.

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, UK: Cambridge University Press.

Hayes, Bruce, and Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39:379–440.

Horning, James. 1969. A study of grammatical inference. Doctoral Dissertation, Stanford.

Kahn, Daniel. 1976. Syllable-based generalizations in English phonology. Doctoral Dissertation, Massachusetts Institute of Technology, Cambride, MA.

Katzir, Roni. 2014. A cognitively plausible model for grammar induction. *Journal of Language Modelling* 2:213–248.

Kenstowicz, Michael, and Charles Kisseberth. 1977. *Topics in phonological theory*. Academic Press.

de Marcken, Carl. 1996. Unsupervised language acquisition. Doctoral Dissertation, MIT, Cambridge, MA.

McCarthy, John J. 2007. Derivations and levels of representation. In *The cambridge handbook of phonology*, ed. Paul de Lacy, 99–117. Cambridge University Press.

Orbán, Gergő, József Fiser, Richard N Aslin, and Máté Lengyel. 2008. Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences* 105:2745–2750.

Piantadosi, Steven T., Joshua B. Tenenbaum, and Noah D. Goodman. 2016. The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological review* 123:392–424.

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

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

Rasin, Ezer, Iddo Berger, Nur Lan, and Roni Katzir. 2017. Learning rule-based morpho-phonology. Ms., MIT and Tel Aviv University, April 2017.

Rasin, Ezer, and Roni Katzir. 2015. Compression-based learning for OT is incompatible with Richness of the Base. In *Proceedings of NELS 45*, ed. Thuy Bui and Deniz Özyıldız, volume 2, 267–274.

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

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

Rissanen, Jorma, and Eric Sven Ristad. 1994. Language acquisition in the MDL framework. In *Language computations: DIMACS Workshop on Human Language, March 20-22, 1992*, 149. Amer Mathematical Society.

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.

Stanley, Richard. 1967. Redundancy rules in phonology. *Language* 43:393–436.

Stolcke, Andreas. 1994. Bayesian learning of probabilistic language models. Doctoral Dissertation, University of California at Berkeley, Berkeley, California.

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

Xu, F., and J.B. Tenenbaum. 2007. Word learning as Bayesian inference. *Psychological review* 114:245–272.