A conditional learnability argument for constraints on underlying representations*

Ezer Rasin and Roni Katzir

March 21, 2018

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 Dutch, for example, the distribution of the voiceless alveolar strident [s] and its palatalized variant [ʃ] is restricted such that the palatalized variant occurs precisely before the palatal glide [j] (Booij 1995). Thus, forms such as [ɔstər] and [oʃjər] are accidental gaps, while *[ɔʃtər] and *[osjər] are systematic gaps.[1] 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 learning, the phonological component must follow one of several specific representation schemes discussed below in order to ensure that the acquisition process leads to the judgments that actual speakers make.

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 (CURs) in the lexicon;[2] and phonological rules. In the example above, 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 stridents in Dutch:[3]

*Acknowledgments: To be added.

[1] Our focus in this paper is on the distribution of the stridents [s] and [ʃ]. We will set aside other systematic properties of Dutch surface forms, such as the distribution of tense vowels (like [o]) and lax vowels (like [ɔ]). As far as we can tell, this does not affect our argument.

[2] 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 conditions. We use CURs as a cover term for rules or constraints of this kind.

[3] To simplify the presentation, here and below we use *strident* to refer to the voiceless coronal stridents [s] and [ʃ] only, excluding other Dutch stridents such as [z] and [x]. As far as we can tell, this simplification is orthogonal to the argument we will present.

(1) CUR IN DUTCH: No ʃ in the lexicon

(2) [+*strident*] → ʃ / __ j

(1) ensures that stridents will be alveolar underlyingly, while (2) ensures that they will become palatalized in exactly the right environment. The combination of (1) and (2) handles the distinction between the accidentally missing [ɔstər] and [oʃjər] on the one hand and the systematically missing *[ɔʃtər] and *[osjər] on the other, on the assumption 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 [ɔstər] and [oʃjər] could be added to Dutch with the URs /ɔstər/ and /oʃjər/; the palatalizing rule in (2) would then turn the former into its surface form. For *[ɔʃtər] and *[osjər] the situation is different. Since (1) prohibits the storing of /ʃ/ in the lexicon of Dutch, [ʃ] must follow from rule application; but the palatalization rule in (2) does not apply before /t/, which leaves no way to derive *[ɔʃtər]. For *[osjər], on the other hand, obligatory palatalization 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 the distribution of stridents 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 stridents are alveolar by default using a CUR, we could accomplish the same by a rule such as (3) below, which makes stridents alveolar regardless of their underlying specification or of their environment:

(3) [+*strident*] → s

If (3) is ordered before (2), any UR would first have its stridents made alveolar ([s]), after which its pre-palatal stridents will be made palatalized ([ʃ]). This would make the URs /osjər/ and /oʃjər/ surface as [oʃjər], while the URs /ɔʃtər/ and /ɔstər/ will surface as [ostər]. The systematically missing *[ɔʃtər] and *[osjər] 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 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, CURs were taken to be preferred by the simplicity metric (for a simplicity-based argument for CURs, see Halle 1962, pp. 59–60). 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:[4]

We will assume that [s] and [ʃ] are distinguished from each other by the feature *anterior*: [s] is specified as [+*anterior*] and [ʃ] as [−*anterior*]. A different yet equivalent formulation of (2) using *anterior* would be [+*strident*] → [−*anterior*] / __ j. To simplify the presentation, we will use symbols (like s and [ʃ]) instead of features whenever possible.

[4]The discussion above uses phonological rules, but both approaches can just as easily be stated using OT constraints (which will be the main representation used in section 3) or even more neutrally, as a reviewer notes, using mapping statements as in Tesar 2014. Stated in terms of constraints, the first approach would

(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.

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

Clearly, the two representational choices for how to handle the distributional pattern of stridents are meaningfully different. For example, the use of CURs distributes the knowledge of such patterns 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.

Turning to the question of how the relevant knowledge is acquired, we will rely on a general approach to learning, following the principle of Minimum Description Length, that has much in common with the evaluation metric of early generative phonology but is quite different from much of the literature on learning within OT. The following section briefly introduces the learning approach that we will be using and motivates our choice of using it instead of familiar suggestions in the literature on learning in OT.

2 Learning

Our discussion below relies 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 3, 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

combine the CUR in (1) with a constraint ranking such as *sj ≫ Ident[ant], while the second approach would avoid (1) and instead add a mid-ranking constraint banning [ʃ], as in *sj ≫ *ʃ ≫ Ident[ant]. 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 question in what follows.

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, and Goldsmith 2001, 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, To appear) that acquire a lexicon, the phonological processes involved, and their interactions, all from distributional evidence alone.[6]

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 the distribution of stridents in Dutch, the fact that *[ɔʃtər] and *[osjər] are judged as ill-formed teaches us that the investment in ruling out these forms, through the relevant statements in G, has been justified by the shortening of $D : G$. The acceptability of [ɔstər] and [oʃjər], 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.[7]

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, which we will refer to as *restrictive consistency seeking* (RCS), according to which the child attempts to find the most restrictive grammar consistent with the data. On this view, common within OT, 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 RCS 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,

[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).

[6]As a reviewer notes, a large body of research has argued for the storage of predictable information. We cannot do justice to this literature within the scope of this squib, but as far as we can tell the storage of predictable information is fully compatible with MDL. The main claim behind MDL is that learning is based on compression, which in turn involves the identification of redundancies that can be eliminated. But it does not follow that such redundancies are in fact eliminated. If they are identified but still stored, MDL can work just as well. The proposals of de Marcken 1996, Johnson et al. 2007, and O'Donnell et al. 2009 all involve the integration of MDL/Bayesian induction with a framework for the storage of predictable information.

[7]The above assumes that the grammatical statements under consideration need to be acquired (rather than being given to the learner in advance) and that they are allowed by UG in the first place.

Prince and Tesar 2004) throughout the search for a consistent ranking. On this view, the child hypothesizes in advance that they should ban *[osjər] and *[ɔʃtər], and since Dutch provides no counter evidence, they never need to change their mind. What the acceptability of [oʃjər] and [ɔstər] teaches us, on this view, is that UG simply neglects to provide the means to ban these forms without banning attested forms.[8] Had it done so, the absence of [oʃjər] and [ɔstər] from the child's input data would have allowed the child to maintain the more restrictive hypothesis that these forms are impossible.

In the literature, the representational principle of ROTB has often been bundled together with learning models that follow RCS. This bundling does not follow logically – ROTB does not imply RCS, nor is it implied by it – but given the connection made in the actual literature, we wish to explain why we set aside RCS and instead propose to use the inductive approach to learning, as in MDL, for our probing of the choice between CURs and ROTB.

Our first reason for setting aside RCS and focusing exclusively on MDL is that, to date, only the latter has actually provided working distributional learners for phonology (that is, implemented algorithms that take raw surface data and induce a phonology and a lexicon). The MDL learner of Rasin and Katzir 2016, for example, takes unanalyzed surface forms and induces a lexicon, often with abstract URs that differ from the surface forms, along with a set of markedness and faithfulness constraints and their ranking.[9] The MDL learner of Rasin et al. 2017, To appear accomplishes a similar task but within rule-based phonology: it takes unanalyzed surface forms and induces a lexicon and a set of ordered context-sensitive rewrite rules. Despite active research into RCS in phonology over the past few decades, no currently available learner has been presented that can accomplish similar tasks, making the RCS idea promissory.

While it is conceivable that future work will provide an implemented RCS learner, the path toward such a result is quite unclear at present. For one thing, the relevant notion of restrictiveness has been hard to formulate, with concrete choices such as M≫F giving rise to problems that have been recognized in the literature (see, e.g., Hayes 2004, Prince and Tesar 2004, Tauberer 2009). The induction of the lexicon has also posed a challenge for RCS models. Such models have typically relied on notions such as Lexicon Optimization (Prince and Smolensky 1993, p. 209, Inkelas 1995, Smolensky 1996), which encourages the learner to posit URs that are identical to the corresponding surface forms unless forced to do otherwise by paradigmatic information from alternations. (Variants of this idea have been explored as well; see McCarthy 2005 and Krämer 2012, among others. In McCarthy 2005's variant, the Free Ride Principle, nonidentical URs are posited when an alternation suggests – rather than strictly force – a deviation from identity.) As argued in detail by Alderete and Tesar (2002), McCarthy (2005), Idsardi (2006), Nevins and Vaux (2007), and Krämer (2012), Lexicon Optimization (and its variants) cannot handle the abstract URs that speakers actually posit,

[8] The qualification regarding not banning attested forms is needed. UG may well provide the means to rule out [oʃjər] and [ɔstər], but if these means are blunt instruments such as *ʃ or *ɔ they will also rule out attested forms and will therefore be demoted below the relevant faithfulness constraints and will not play a role for [oʃjər] and [ɔstər]. Such accidental gaps, then, teach us about the absence of more specific constraints such as *[ʃV] (within an RCS model) but are uninformative regarding more sweeping constraints that would be demoted independently.

[9] The learner also works with a set of constraints that are given in advance, as is assumed in much work within OT.

often without any supporting alternations to force such URs. For example, Nevins and Vaux consider the case of rhotics in Spanish, which can be realized as [r] or [ɾ] word-medially but only as [r] word-initially. When induced to move a word-initial [r] to a word-medial position as part of a language game, speakers sometimes realized it as [r], in line with a faithful UR, but sometimes as [ɾ], suggesting an unfaithful UR. Crucially, Lexicon Optimization cannot attribute this deviation to any available alternations. While the induction of abstract URs posits a (currently unresolved) challenge for RCS models, MDL learners succeed in acquiring such abstract URs using nothing more than the general principle in (5), as shown in Rasin and Katzir To appear.

Our second reason to set aside RCS models and focus on MDL is that the latter has been supported empirically by a range of lab experiments on generalization, while similar support is lacking for the former. In a variety of learning tasks in the lab, ranging from word learning (Xu and Tenenbaum 2007) through causal reasoning (Sobel et al. 2004) to sensorimotor control (Körding and Wolpert 2006) and visual scene chunking (Orbán et al. 2008), among many other tasks, human subjects have been argued to balance between the specificity of a hypothesis, corresponding to $|D : G|$, and its independent plausibility, corresponding to $|G|$. If humans indeed use this way of learning across different domains, it seems sensible to consider their use of the same in phonology. We are not aware of similar considerations for RCS models.

If the technical challenges for RCS models are eventually addressed and an implemented distributional RCS learner provided, or if new lab experiments change the currently available evidence for inductive models like MDL over RCS, a reexamination of the choice between CURs and ROTB will of course be warranted. Until then, it strikes us as reasonable to examine the implications of this choice within an MDL model, to which we now turn.

3 The MDL-learnability implications of ROTB

The general shape of our MDL-based reasoning about ROTB will be as follows.[10] As we will note, ROTB will usually result in some part of the distribution of stridents being stored faithfully. For that part of the pattern, there will be no MDL motivation to invest in grammatical statements (whether rules and their ordering or constraints and their ranking), and so such statements will not be systematically acquired. Depending on the properties of the initial state, this can result in adults who do not have the knowledge of the relevant part of the pattern as part of the input-output mapping, and – again, due to ROTB – such adults will be able to accept nonce forms with the incorrect kind of strident in that part of the distribution, which does not match the judgments that actual speakers make on such forms. We will point out two possible responses to this predicament. The first response is to abandon ROTB and admit CURs, which, as

[10]For concreteness, we present most of our discussion within the framework of constraint-based phonology and refer to similar considerations within rule-based phonology only occasionally. (As mentioned above, the question of whether ROTB holds is separate from the question of rules versus constraints, though in section 3.3 we note one place in which the two choices interact.) We will also stay with the example of stridents in Dutch (though the argument for CURs can be made using any of a wide variety of patterns from different languages).

we will show, lead to the correct pattern of speaker judgments and also have a clear MDL motivation (by supporting a shorter encoding of the lexicon) and will therefore be acquired by the learner. Since the learning problem that we note is caused directly by ROTB, and since ROTB has not been particularly well supported by other evidence in the literature, the re-introduction of CURs strikes us as the most natural response. The second response is to maintain ROTB but adopt special measures to ensure the knowledge of the pattern. For example, the problem we outline obviously does not arise if the full knowledge of the pattern is given to the child in advance (by building the relevant constraints and their ranking into the initial state, as is often assumed within OT, or by doing the same with the rules and their ordering). For rule-based phonology (but not for OT), a more imaginative possibility within the second response is to allow for underspecification in the storage of URs. This choice allows a rule-based learner to store non-faithful stridents throughout, and, on certain assumptions discussed below, it ensures that the full pattern of strident distribution is acquired.

The structure of the argument and the range of responses are intricate, and in what follows we discuss both in some detail. The basic observation, however, is straightforward: ROTB leads to a learnability challenge given the data available to the child and the judgments that speakers have, and one of a small range of representational responses is called for. In the literature to date, ROTB has mostly been left as a matter of theoretical taste, but our observation shows that this need not remain the case: the range of possible responses to the learning challenge amounts to an empirical prediction of ROTB that can be tested, though we will not be able to do so within the present squib. Beyond the issue of ROTB, our argument illustrates a methodological point that was central in earlier generative phonology but has not received much attention in recent years: that a general evaluation metric for learning can yield architectural predictions about linguistic representations and help choose between competing theories of UG. We return to both the specific implications of our argument for ROTB and the general methodological point in section 4.

In order to develop our argument, we will need to examine the MDL implications of the possible choice points under various reasonable representational assumptions. These assumptions include both cases in which the constraints are given to the child in advance and cases in which they are acquired. While the former possibility has been widely assumed within the literature on OT, it will be useful to consider the latter possibility in some detail for several reasons. First, we would like to get a picture of the connection between learnability and the choice between CURs and ROTB, not just in specific configurations that have received attention in the literature but generally. As we will show, this broader examination will allow us to identify an empirical connection between assumptions that have often been bundled together in the literature without argument. Second, language-specific constraints that need to be acquired have occasionally been suggested even within the OT literature (see, e.g., Kager and Pater 2012, Pater 2014 and references therein, as well as the earlier literature on arbitrary phonological rules, e.g., Bach and Harms 1972 and Anderson 1981; of course, language-specific rules that need to be acquired were broadly assumed within earlier generative phonology). Finally, the case of constraints that need to be acquired is somewhat simpler to analyze than the case of constraints that are given in advance. It will thus be convenient presentationally to start from the simpler case, which we discuss

	Acquired constraints	Innate constraints
No $M \gg F$	CURs	CURs
$M \gg F$	CURs	

Figure 1: Summary of the configurations we discuss in sections 3.1 and 3.2 as part of the conditional argument for CURs. We consider two conditions: whether constraints are acquired or innate and whether markedness constraints preferably outrank faithfulness constraints ($M \gg F$). Cells labeled as 'CURs' correspond to configurations for which we show that ROTB fails and CURs are required for learning. The empty cell corresponds to the only configuration in which ROTB succeeds, which combines innate constraints with $M \gg F$.

in section 3.1, and introduce the complications of constraints that are given in advance only later, which is what we do in section 3.2. The configurations we discuss in sections 3.1 and 3.2 are summarized in Figure 1. In section 3.3 we discuss the special case of a rule-based system with underspecification (and certain additional assumptions), which, as mentioned above, allows a ROTB learner to acquire the full pattern of strident distribution correctly.

3.1 Constraints acquired

For the first few configurations that we will consider, 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 /s/ and /ʃ/ each costs some fixed number of bits to store in the lexicon. The exact form of the argument depends on which of the two segments is costlier, if either. We will consider the three different possibilities in turn, followed by an examination of costs that vary between contexts.

3.1.1 Globally fixed costs for /s/ and /ʃ/, $Cost(/ʃ/) > Cost(/s/)$

Suppose that the cost of storing an instance of /ʃ/ in the lexicon is greater than the cost of storing an instance of /s/, $Cost(/ʃ/) > Cost(/s/)$. Consider first the situation of the child on the assumption that ROTB holds. Since instances of /ʃ/ are costly to store in the lexicon, it will be preferable in terms of MDL to invest in a markedness constraint that triggers palatalization of /s/ before /j/ (e.g., *sj) and then store all stridents as /s/. 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 the costlier /ʃ/ in the lexicon. A faithfulness constraint ensuring that /s/ surfaces faithfully – possibly a general Faith and possibly something more specific such as Ident[ant] – will also be acquired, but it will be outranked by *sj so that the latter will have its desired effect.[11] By following this reasoning, the child has successfully learned that

[11] See Rasin and Katzir 2016 for why faithfulness constraints will be acquired by an MDL learner, quite independently of the palatalization pattern under discussion here. Faithfulness constraints ensuring that /s/ surfaces faithfully will be acquired independently of the palatalization pattern on the assumption that UG

pre-palatal stridents are systematically palatalized in the language. For example, the child will now correctly rule out forms such as *[osjər], with an alveolar [s] before [j].

Unfortunately for ROTB, this is also the extent of the child's acquisition of the pattern: the child will not learn to block forms such as *[ɔʃtər], with [ʃ] in 'elsewhere' environments. The reason is that, for an ROTB child, such forms must be blocked through the input-output mapping, for example through *ʃ or a similar markedness constraint that penalizes [ʃ]. And there is simply no MDL justification for acquiring this kind of constraint. Recall that all stridents are already stored in the lexicon as alveolar ([s]) and that by default they are mapped faithfully to the surface due to the low-ranking faithfulness constraints. Consequently, a markedness constraint such as *ʃ 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 distributional pattern of stridents in Dutch: such an adult will correctly rule out *[osjər] but fail to rule out *[ɔʃtər] (which, due to ROTB, can be stored as is and then mapped faithfully to the surface given the acquired constraints).[12]

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 *sj and then store all stridents as alveolar, which will allow the child to correctly rule out forms like *[osjər]. But with the possibility of stating CURs, a crucial second step becomes available. The first step involved the extensional removal of all instances of /ʃ/ from the lexicon. The child can now conclude that this was no accident, and that /ʃ/ 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):

(6) CUR IN DUTCH: No ʃ in the alphabet of the lexicon

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 it slightly easier to specify the remaining segments, some of which may now cost fewer bits than before.

allows writing faithfulness constraints that are at least as general as IDENT[F] – that is, constraints such as IDENT[ANT] or IDENT. We note that if there is a reason to restrict faithfulness constraints to be very specific (e.g., if IDENT-constraints can only take the form IDENT[±F]), our claim that a faithfulness constraint protecting [+ant] will be acquired independently of the palatalization pattern will no longer be true. If that turns out to be the case, at least some learners may acquire a markedness constraint such as *ʃ instead of the relevant faithfulness constraint, and if the relative costs of the two kinds of constraints ensure that all learners acquire *ʃ, the challenge for ROTB will be avoided. In the absence of reasons to assume that the only faithfulness constraints can be acquired are highly specific, we set aside this possibility and do not consider it further in what follows.

[12]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 stridents are alveolar in 'elsewhere' environments, which, in turn, would allow the learner to acquire the full distributional pattern without abandoning ROTB. As far as we can tell, however, the alternations that are actually available in Dutch do not provide an MDL learner with the relevant information and thus do not help address the challenge to ROTB, either with the present representations or with those discussed below. In section 3.1.3 we will mention one case where alternations can help learn part but not all of the Dutch pattern. There, too, the challenge to ROTB remains even if the learner can make use of alternations. Other than that case, since alternations do not help address the challenge to ROTB we will continue to ignore their role in what follows.

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 pattern of distribution of stridents. The reason is that with a CUR like (6), the child will now correctly rule out surface forms like *[ɔʃtər]: /ɔʃtər/ is now no longer a possible UR; and given the grammar that has been induced, this UR is the only potential source for this putative surface form. In other words, the impossibility of even stating /ʃ/ in the lexicon, with its MDL justification, means that the learner has correctly learned to block illicit palatalization.

We conclude that, under the representational choice of constraints that need to be acquired and of *Cost*(ʃ) > *Cost*(s), the ability to state CURs allows for the full distributional pattern of stridents to be acquired, while the adoption of ROTB leads to a failure in learning half of the pattern.

3.1.2 Globally fixed costs for /s/ and /ʃ/, *Cost*(/ʃ/) < *Cost*(/s/)

Suppose now that *Cost*(/ʃ/) < *Cost*(/s/). In this case, the learner will store /ʃ/ throughout the lexicon and acquire constraints that will ban [ʃ] in 'elsewhere' (that is, non-pre-palatal) 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 markedness constraint banning [ʃ] in those environments – say, *ʃT (no [ʃ] before a stop) – and be done.[13] In particular, *sj will not be learned, and the target grammar will fail to ban impossible nonce words such as *[osjər] (which, again, can be stored as is due to ROTB and then mapped faithfully to the surface). 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 [ʃ] directly in these contexts, and it will make more sense for the child to learn to ban [ʃ] in general and allow it only before [j]. That is, it will acquire a low-ranking *ʃ to prevent underlying /ʃ/ from surfacing faithfully and a high-ranking *sj to ensure that stridents surface correctly before [j]. On this scenario, the learner will have correctly acquired the full pattern without requiring a CUR, thus allowing ROTB to be maintained.

While this scenario provides a way to learn the distribution of stridents without CURs, it is ultimately unsuccessful because the cost assignment makes the mirror image of the Dutch pattern unlearnable – a pattern with stridents that are alveolar in some specific environments but are palatalized elsewhere. Consider Bengali, where the default sibilant is [ʃ], and [s] occurs only in word-initial consonant clusters and word-medially before dental stops (Evers et al. 1998). For example, the nonce forms [t̪uʃk̪a], with an [ʃ] before a velar consonant, and [t̪ust̪a], with an [s] before a dental consonant, were both accepted by two native speakers of Bengali we consulted and are thus

[13] As with the earlier cost scheme (as well as with the configurations below), a relevant faithfulness constraint will also be acquired and ranked below the markedness constraint.

accidental gaps, while the nonce form [t̪uska], with an [s] before a velar consonant, was rejected and is therefore a systematic gap.[14] An appropriate constraint ranking for the Bengali pattern (ignoring both optionality and word-initial clusters) would be the following:

(7) Constraint ranking for Bengali (without optionality): *ʃt̪≫*s≫IDENT[ANT]

And paralleling the discussion of the Dutch pattern with the earlier cost assignment of *Cost*(/ʃ/) > *Cost*(/s/), the present cost assignment of *Cost*(/ʃ/) < *Cost*(/s/) will prevent the full Bengali pattern from being acquired. Given the present cost assignment, an ROTB learner will store all stridents as /ʃ/ and then acquire a markedness constraint forcing stridents to surface as [s] in the relevant environments. The same reasoning used earlier will prevent the learner from acquiring a constraint that enforces [ʃ] elsewhere (in this case, since all stridents are already stored as /ʃ/ in the lexicon), which will result in an inability to rule out nonce forms with [s] in 'elsewhere' environments (e.g., before velar consonants, as in *[t̪uska]), contrary to fact. On our current assumptions that constraints are acquired and that /ʃ/ is less costly than /s/, succeeding in learning Bengali requires abandoning ROTB and using CURs.

3.1.3 Globally fixed costs for /s/ and /ʃ/, *Cost*(/ʃ/) = *Cost*(/s/)

Consider now the possibility that *Cost*(/ʃ/) = *Cost*(/s/). In this case, compression cannot learn either part of the Dutch pattern in the absence of CURs (and similarly for Bengali): with fixed, equal costs for /s/ and /ʃ/, compression will favor the storing of URs that are identical to their corresponding surface forms in terms of palatalization, along with the acquisition of the relevant faithfulness constraints that will guarantee that the stored values surface faithfully. Any markedness constraints governing palatalization will be superfluous and will therefore not be acquired. An ROTB learner will consequently fail to reject both *[ɔʃtər] and *[osjər].[15]

3.1.4 Contextualized costs for /s/ and /ʃ/

The learnability argument against ROTB extends to some other representational possibilities that UG might make available. For example, suppose that UG makes the cost

[14] The speakers' responses were variable with respect to the nonce form [t̪uʃt̪a], with an [ʃ] before a dental consonant: one speaker rejected it as ill-formed; the other speaker accepted it as well-formed, suggesting that the process that turns [ʃ] into [s] applies optionally in her grammar. That variability with respect to [t̪uʃt̪a] is consistent with our argument, which is independent of whether the process applies obligatorily or optionally.

[15] In the case of equal costs for /s/ and /ʃ/, alternations may help learn half of the Dutch pattern. The pressure for economy will push the learner to store the stem in surface pairs like *ijs* [ɛis] 'ice cream' and *ijsje* [ɛiʃjə] 'small ice cream' as a single UR – either /ɛis/ or /ɛiʃ/ – and derive the [s]~[ʃ] alternation from the input-output mapping by adding appropriate constraints to the grammar. However, on either choice of UR only half of the pattern will be learned. If the UR of the stem is /ɛis/ (and [ɛiʃjə] is derived through palatalization, using a constraint like *sj), the constraint *ʃ will serve no compressional purpose and thus will not be learned; in this case, an ROTB learner will fail to reject *[ɔʃtər]. If, on the other hand, the UR of the stem is /ɛiʃ/ (and [ɛis] is derived through de-palatalization, using a constraint like *ʃ# which penalizes word-final [ʃ]), the constraint *sj will serve no compressional purpose and thus will not be learned; in this case, an ROTB learner will fail to reject *[osjər].

of /ʃ/ lower than that of /s/ before /j/ 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 necessary for capturing the Dutch pattern unlearnable by an MDL learner for the same reasons as discussed above: as in the case of identical costs, the learner will store URs that are identical to the corresponding surface forms in terms of palatalization (with /ʃ/ before /j/ and /s/ elsewhere) and, given the faithfulness constraints, will not invest in any markedness constraints for palatalization.

For the opposite weighting scheme, with the cost of /ʃ/ higher than that of /s/ before /j/ and lower than it in other environments, things are different. This scheme will allow both kinds of markedness constraints relevant for the Dutch pattern to be learned by an MDL learner, regardless of CURs. As with $Cost(/ʃ/) < Cost(/s/)$, however, this scheme makes patterns like the one in Bengali unlearnable by an MDL learner that follows ROTB. Since neither /s/ nor /ʃ/ precedes /j/ in Bengali,[16] the contextualized cost assignment will have the same effect as $Cost(/ʃ/) < Cost(/s/)$: an ROTB learner will store all stridents as /ʃ/ in the lexicon and, given the relevant faithfulness constraints, will fail to invest in a markedness constraint such as *s; this, in turn, will lead to an inability to rule out forms with [s] in elsewhere environments (as in *[t̪uska]), contrary to fact.

This concludes our discussion of the case of constraints that need to be acquired. We have seen that across various representational choices, the ability to state CURs in the lexicon is necessary for successful learning, assuming that the constraints are not given in advance.

3.2 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, building on the argument in Rasin and Katzir 2015 that, unless a preference for markedness over faithfulness is incorporated, an MDL learner would still need to abandon ROTB and adopt CURs. Suppose that the learner is given the two relevant markedness constraints for the Dutch pattern: *sj, which penalizes alveolar pre-palatal stops; and *ʃ, which penalizes [ʃ] in general. Suppose further, as in 3.1.1, that $Cost(/ʃ/) > Cost(/s/)$.

As in the setting with acquired constraints, the constraint *sj poses no special problem for an MDL learner following ROTB. Ranking this markedness constraint above the relevant faithfulness constraints will serve the compressional purpose of enabling the elimination of /ʃ/ from all URs. As for *ʃ, 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 *ʃ 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 IDENT[ANT], that penalizes modifications of the feature *anterior*. We would thus expect

[16] Ferguson and Chowdhury (1960) suggest that the glide [j], if it exists in Bengali, is only available as the second member of a diphthong (such as /ai/). The two native speakers we have consulted confirmed that [j] never follows stridents in their dialects. One speaker reported that [j] does not exist in her dialect at all. The second speaker reported that [j] only occurs after vowels.

speakers to vary in the relative ranking of *ʃ and IDENT[ANT]. But this means, on ROTB, that speakers of Dutch should differ in whether they accept forms such as *[ɔʃtər] as possible, contrary to fact.[17] 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 but rather with ensuring that this ranking is attained systematically, for all speakers, and not just occasionally.

It is at this point that a preference for M≫F becomes relevant.[18] In the settings discussed in section 3.1 above, with binary features and acquired constraints, M≫F does not solve the problem for ROTB, and adopting CURs would be needed to ensure the learning of the distribution of stridents. With constraints that are given in advance, on the other hand, M≫F enables successful acquisition: as we just saw, the challenge in this case is not justifying the constraints (which, in the current setting, are already provided) but rather ensuring that the markedness constraints outrank the faithfulness constraints; by stipulation, a preference for M≫F addresses this challenge.[19] The combination of M≫F with constraints that are given in advance, then, is one way to preserve ROTB in the face of the learnability challenge (in effect, by giving the child knowledge of the pattern as part of the initial state). We now turn to a less stipulative response available within rule-based phonology.

3.3 Special case: rule-based phonology with underspecification

So far we have assumed that features are binary. This assumption contributed to the fact that an ROTB child would always store part of the distribution of stridents faithfully, which in turn made it superfluous to acquire that part of the distribution within the input-output mapping, thus leading to the challenge to ROTB. We saw two responses to this challenge: allowing CURs (and thereby rejecting ROTB); and endowing the child with prior knowledge of the pattern (in the shape of constraints that are given in advance combined with $M \gg F$). If underspecification is allowed, a third response suggests itself: if storing an underspecified value such as [0*ant*] is less costly than either of the specified values, the learner might prefer storing all stridents unfaithfully as [0*ant*] and invest in the markedness constraints *sj and *ʃ, along with a high-ranking markedness constraint such as *[0*F*] that blocks underspecified values from surfacing. This response still requires something like $M \gg F$ to ensure that *sj and *ʃ outrank faithfulness (since otherwise inappropriate stridents in nonce forms could be accommodated, as discussed earlier), but otherwise this seems like a way to allow ROTB to be maintained without giving full prior knowledge to the learner (since the markedness constraints are now acquired). However, as we now show, the help that underspecification offers ROTB is considerably more limited than might appear to be the case:

[17] We have consulted three native speakers of Dutch, who all rejected *[ɔʃtər].

[18] As discussed earlier, variants of such a preference have been used within RCS 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 similarly be implemented, most straightforwardly through the cost scheme for the statement of rankings.

[19] The same reasoning applies if $Cost(/ʃ/) = Cost(/s/)$ or if $Cost(/ʃ/) < Cost(/s/)$. In both cases, the problematic ranking IDENT[ANT]≫*ʃ can be avoided with a preference for M≫F (but otherwise remains a problem for an ROTB learner).

within OT, capturing certain simple cases of systematic gaps will still require both innate constraints and $M \gg F$, which means that underspecification leaves the challenge to ROTB without change; and within rule-based phonology, underspecification will enable general learning while maintaining ROTB, but only under specific assumptions.

The problem with using underspecification to succeed in learning while maintaining ROTB is that, while underspecification indeed makes a correct grammar (with underspecified URs and an investment in the requisite markedness constraints) more economical than the kinds of incorrect grammars considered earlier, it sometimes makes a new kind of incorrect grammar more economical still. As a concrete illustration, consider a case of four consonants such as the velar obstruents [k], [g], [x], and [ɣ], which are identical with respect to all features but two (*voice*, which distinguishes the voiced [g] and [ɣ] from the voiceless [k] and [x], and *continuant*, which distinguishes the continuants [x] and [ɣ] from the stops [k] and [g]). Consider now a language that has exactly three of those four consonants – for example, German, which has [k], [g] and [x], but not [ɣ]. To correctly rule out surface forms with [ɣ], the German-learning ROTB child will need to learn a high-ranking markedness constraint such as *ɣ. Earlier, with binary features, a ROTB learner would have had no motivation to posit such a constraint: in analogy with our discussion of Dutch and Bengali, an incorrect grammar storing voicing faithfully (and with no need for *ɣ) would have been optimal. With underspecification, faithful storage of that kind is no longer optimal. In particular, storing the attested [x] as underspecified for voicing in the lexicon will provide an incentive to derive the voicelessness of [x] through the input-output mapping using *ɣ, which, in turn, would correctly prevent URs with underlying /ɣ/ from surfacing faithfully, as in the following grammar G_1:

(8) G_1 (complex; correct)
   a. Lexicon: [x] is stored as [0*voice*]; [k],[g] are stored faithfully
   b. Constraint ranking: {*[0*voice*],*ɣ} ≫ IDENT[VOICE]

The challenge to ROTB with underspecification and acquired constraints is that the correct G_1 has a simpler but incorrect alternative grammar G_2 which stores not only [x] but also [k] as underspecified for voicing, and which maps only underspecified velars to voiceless:

(9) G_2 (simple; incorrect)
   a. Lexicon: [k] and [x] are stored as [0*voice*]; [g] is stored faithfully
   b. Constraint ranking: *[0*voice*] ≫ IDENT[VOICE] ≫ *[+*voice*]

G_2 is simpler than G_1 for two reasons. First, its constraint ranking is slightly simpler since it replaces the specific constraint *ɣ (= *[*velar*, +*cont.*, +*voice*]) with the more general constraint *[+*voice*]. Second, and much more significantly, its lexicon is more economical since it stores more features as underspecified than G_1 does. As opposed to G_1, the simpler G_2 does not rule out [ɣ], because the faithfulness constraint IDENT[VOICE] preserves underlying instances of /ɣ/. And crucially, the existence of [g] prevents a preference for $M \gg F$ from being helpful: a ranking with the markedness constraint *[+*voice*] above IDENT[VOICE] would rule out both [ɣ] and [g] and will thus fail to generate the data. In other words, an ROTB learner with underspecification and acquired

constraints will converge on G_2 and fail to capture the absence of [ɣ]. Avoiding this problem requires a combination of *ɣ and $M \gg F$, which in turn means that within OT, underspecification offers ROTB no learnability advantage over full specification.

The problem with using underspecification to keep ROTB can be replicated within rule-based phonology, where G'_1 and G'_2 serve as counterparts of G_1 and G_2:

(10) G'_1 (complex; correct)
   a. Lexicon: [x] is stored as [0*voice*]; [k],[g] are stored faithfully
   b. Rule: [*velar*, +*cont.*] → [−*voice*]

(11) G'_2 (simple; incorrect)
   a. Lexicon: [k] and [x] are stored as [0*voice*]; [g] is stored faithfully
   b. Rule: [*velar*, 0*voice*] → [−*voice*]

As with OT, the incorrect rule-based solution of G'_2 is more economical than the correct one in G'_1. Differently from OT, however, rule-based phonology offers a straightforward way to avoid the problem: if it is impossible to refer to underspecified values within a rule (but still cheaper to state underspecified values in the lexicon, so that the rule in (10) will be learned at all), then the correct G'_1 will be acquired. The following summarizes the two assumptions that we relied on here to preserve ROTB in the present setting (by benefiting from underspecification on the one hand and avoiding the trap of G'_2 on the other hand):

(12) Assumptions that allow ROTB to be maintained within rule-based phonology:
   a. The cost of storing an underspecified feature is strictly lower than that of storing a fully specified one
   b. Rules may not refer to underspecified feature values

Within rule-based phonology, then, and assuming the possibility of underspecification along with the specific statements in (12), CURs are not necessary to learn systematic gaps (as in the Dutch, Bengali, and now German pattern above) even if knowledge of the pattern is not given to the learner in advance.

3.4 Phonotactic learning does not help ROTB

The discussion in the previous sections illustrated the challenge for ROTB using primarily the pattern of distribution of stridents in Dutch. Similar patterns abound, and the same argument could be made with other instances. In the present section we discuss a potential concern with the challenge. We assumed that acquiring the Dutch pattern 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.[20]

We are not aware of actual implementations of the idea just sketched. Regardless of whether such an idea might handle the Dutch pattern and allow ROTB to be maintained in that specific case, there are other patterns in which an earlier phonotactic stage will be of little help and in which our reasoning regarding ROTB can be repeated, which in turn means that the challenge to ROTB will remain unchanged. 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 reasoning regarding CURs and ROTB in a setting that is considerably more challenging for phonotactic learning than the Dutch pattern. 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 deletes, an instance of counterbleeding opacity (e.g., /haːkim-iːn/ [haːkʲm-iːn] 'ruling(m.pl.)').

Our reasoning from Dutch can be restated for the Bedouin Arabic palatalization process, which in turn will allow us to restate our challenge to ROTB without the possible escape hatch of earlier phonotactic learning. Restating the discussion in section 3.1.1 as an example: if velars are costlier to store as palatalized, 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 alveolar in 'elsewhere' environments, 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 alveolar; 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 unless knowledge of the pattern is given to the learner in advance. We conclude that, even if an early phonotactic stage is available and offers a way to acquire some markedness constraints, the challenge to ROTB remains without change.

[20] Such an approach would need to ensure that in the second, phonological stage, the constraint(s) banning [ʃ] in 'elsewhere' environments will not end up being ranked below the relevant faithfulness constraints. For the purposes of the present section, 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.

4 Discussion

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 one major exception concerns the possibility that the 'elsewhere' knowledge is guaranteed to be available by some independent principle, such as the combination of constraints given in advance and a preference for markedness outranking faithfulness.

The general shape of the argument was this. A ROTB learner will usually store part of the distribution of stridents faithfully. Since these stridents then surface faithfully (whether through an independently-acquired faithfulness constraint or through the default faithfulness on a rule-based system), stating the knowledge of the relevant part of the pattern of distribution through the input-output mapping will be superfluous and will not be acquired by an MDL learner. 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 do one of the following: either (a) 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; or (b) bypass the challenge by either minimizing the learning task (for example, by providing in advance both the constraints and their preferred ranking) or by ensuring that stridents are not stored faithfully (for example, by using underspecification and rule-based phonology, along with certain additional assumptions, as discussed above).

This disjunctive conclusion might seem reassuring for ROTB: after all, the first choice within the (b) option is quite close to the view, common within OT, that all constraints are given in advance and that the markedness constraints are ranked above the faithfulness constraints unless forced otherwise. However, this conclusion also highlights the stakes for the combination of given constraints and markedness over faithfulness. In the OT literature, these assumptions are often bundled together with ROTB, but this bundling is not logically necessary: it is easy to imagine either component being true while the other is false (or that both are true or both are false). What we have shown is that there is an *empirical* dependence between them: the patterns of well-formedness that speakers show are such that, if the combination of constraints given in advance and markedness over faithfulness is false within OT, then ROTB must be abandoned (since otherwise part of the pattern becomes unlearnable).[21] Consequently, any attempt within OT to defend ROTB that does not reject MDL must involve a defense of markedness over faithfulness, along with constraints given in advance. Since language-specific constraints have occasionally been proposed in the literature (see, e.g., Kager and Pater 2012, Pater 2014 and references therein), and since the empiri-

[21] Note that things could have been different. For example, if *cost*(ʃ) > *cost*(s) and if speakers rejected [osjər] but accepted [ɔʃtər], then ROTB could have been maintained without markedness over faithfulness or even constraints that are given in advance. The observation is thus a contingent connection that happens to be true of humans.

cal support for markedness over faithfulness has been thin (though see Davidson et al. 2004), this challenge strikes us as nontrivial, though a proper assessment would clearly take us beyond the scope of the present squib.

Looking past the specific question of CURs versus ROTB, this note illustrates a way in which a general learning criterion can help evaluate competing representational possibilities. The idea is not new. Works such as Halle 1962, 1978, Baker 1979, and Dell 1981 used the simplicity metric of early generative grammar to argue for specific conclusions about representations. As noted above, however, the simplicity metric lacks a pressure for a tight fit of the data (in terms of the MDL quantity $|G| + |D : G|$, the simplicity metric minimizes $|G|$ but does not include a counterpart for $|D : G|$). Consequently, the simplicity metric leads to overly general hypotheses and is not a suitable metric for learning. From the perspective of architecture comparison, the simplicity metric often leads to incorrect conclusions about what representations are learnable.[22] 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. The present note offers a concrete application of this idea to an actual architectural question.

References

Alderete, John, and 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. ROA 543.

Anderson, Stephen R. 1981. Why phonology isn't "natural". *Linguistic Inquiry* 12:493–539.

Bach, Emmon, and Robert T Harms. 1972. How do languages get crazy rules. *Linguistic change and generative theory* 1:21.

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.

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

[22] In particular, it incorrectly suggests that restricted optionality of the kind studied in Baker 1979 and Dell 1981 is unlearnable without severe representational limitations. See Rasin et al. 2017 for discussion of this case and a comparison of learnability using the simplicity metric and using MDL. Not all uses of the simplicity metric suffer from this problem. As far as we can tell, Halle 1962, 1978's simplicity argument for feature-based representations stands.

Davidson, Lisa, Peter Jusczyk, and Paul Smolensky. 2004. The initial and final states: theoretical implications and experimental explorations of richness of the base. In *Constraints in phonological acquisition*, ed. René Kager, Joe Pater, and Wim Zonneveld, chapter 10, 321–368. Cambridge University Press.

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

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.

Ferguson, Charles A, and Munier Chowdhury. 1960. The phonemes of Bengali. *Language* 36:22–59.

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.

Idsardi, William. 2006. A Bayesian approach to loanword adaptations. Poster presented at the annual meeting of the Linguistic Society of America, Albuquerque, NM, January 2006.

Inkelas, Sharon. 1995. The consequences of optimization for underspecification. In *Proceedings of NELS 25*, ed. Jill Beckman, 287–302. GLSA.

Johnson, Mark, Thomas Griffiths, and Sharon Goldwater. 2007. Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. In *Proceedings of NIPS*.

Kager, René, and Joe Pater. 2012. Phonotactics as phonology: Knowledge of a complex restriction in dutch. *Phonology* 29:81–111.

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

Körding, Konrad P., and Daniel M. Wolpert. 2006. Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences* 10:319–326.

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

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

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

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

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

O'Donnell, Timothy J, Joshua B Tenenbaum, and Noah D Goodman. 2009. Fragment grammars: Exploring computation and reuse in language. Technical report, MIT.

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.

Pater, Joe. 2014. Canadian raising with language-specific weighted constraints. *Language* 90:230–240.

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, Iddo Berger, Nur Lan, and Roni Katzir. To appear. Rule-based learning of phonological optionality and opacity. In *Proceedings of NELS 48*.

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.

Rasin, Ezer, and Roni Katzir. To appear. Learning abstract URs from distributional evidence. In *Proceedings of NELS 48*.

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.

Sobel, David M., Joshua B. Tenenbaum, and Alison Gopnik. 2004. Children's causal inferences from indirect evidence: Backwards blocking and bayesian reasoning in preschoolers. *Cognitive science* 28:303–333.

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.

Tauberer, Joshua. 2009. Goldilocks meets the subset problem: Evaluating error driven constraint demotion (RIP/CD) for OT language acquisition. In *Proceedings of the 32nd Annual Penn Linguistics Colloquium*, volume 15 of *University of Pennsylvania Working Papers in Linguistics*, 25.

Tesar, Bruce. 2014. *Output-driven phonology*. Cambridge University Press.

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.