*J. Linguistics* (2020), 1–29. © The Author(s) 2020.
Published by Cambridge University Press
doi:10.1017/S0022226720000146

# A conditional learnability argument for constraints on underlying representations[1]

EZER RASIN
*Leipzig University*

RONI KATZIR
*Tel Aviv University*

(Received 22 April 2019; revised 2 April 2020)

We explore the implications of a particular approach to learning for an architectural question in phonology. The learning approach follows the principle of Minimum Description Length (MDL), which has recently been used for learning in both constraint-based and rule-based phonology. The architectural question on which we focus is whether the grammar allows language-specific statements to be made at the level of the lexicon, as was assumed in early generative phonology, or whether such statements are prohibited, as is commonly assumed within more recent work. We show that under MDL, the architectural question has real empirical implications: across a range of seemingly natural representational schemes, an ability to make language-specific statements about the lexicon is needed to ensure the learnability of an important aspect of phonological knowledge.

KEYWORDS: learning, minimum description length, phonology, richness of the base, theory comparison

## 1. INTRODUCTION

In this paper, we explore the implications of a particular approach to learning for an architectural question in phonology. The learning approach follows the principle of Minimum Description Length (MDL; Rissanen 1978), which has recently been used for learning in both constraint-based and rule-based phonology. We briefly present MDL in Section 2 and note some of the properties that make it a promising approach in our view. However, we will not attempt to argue here against other approaches to learning. Rather, in the spirit of Halle (1978), Baker

[1] We thank Adam Albright, Naomi Feldman, Morris Halle, Michael Kenstowicz, Roger Levy, Giorgio Magri, Donca Steriade, and the audiences at MIT, CNRS, UC San Diego, Leipzig University, Uppsala University, and Tel Aviv University, as well as two anonymous *Journal of Linguistics* reviewers. For help with the Dutch data, we thank Loes Koring, Marc van Oostendorp, and Coppe van Urk. For help with the Bengali data, we thank Neil Banerjee and Ishani Guha.

(1979), Dell (1981), and others, we choose one reasonable learning framework and investigate its consequences for linguistic theory.[2]

The architectural question on which we focus our investigation is whether the grammar allows language-specific statements to be made at the level of the lexicon, as was assumed in early generative phonology, or whether such statements are prohibited, as is commonly assumed within more recent work. While this question has been often bundled with the choice between rule-based and constraint-based approaches, the two are separate; and as we note in Section 3, the architectural question remained difficult to probe empirically and was largely left as a matter of theoretical taste. Our main contribution here, presented in Section 4, is showing that under MDL, the architectural question has real empirical implications. Specifically, across a range of seemingly natural representational schemes discussed below, the ability to make language-specific statements about the lexicon is needed to ensure the learnability of an important aspect of phonological knowledge. And while these consequences are complex – and the settling of the question will clearly require further work – we believe that the very fact that such consequences exist is significant.

## 2. MDL LEARNING

MDL balances the simplicity of the grammar against the length of the grammar's encoding of the data. We describe this balancing in some detail immediately below, but in a nutshell the idea is as follows. A preference for simplicity favors general grammars. On its own, however, as in the evaluation metric of early generative grammar (e.g., Chomsky & Halle 1968), it leads to overly general grammars that offer a poor fit to the data. On the other hand, a preference for grammars that encode the data compactly favors grammars that fit the data well. On its own, however, it leads to overfitting grammars that are too restrictive. By balancing simplicity against compactness of encoding of the data, the learner can hope to find an intermediate level of generalization that is appropriate given the data.

### 2.1 *A brief overview of MDL*

For our sketch of how MDL accomplishes this balance, 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. We will write $|G|$ for the length of the grammar $G$ as measured in bits. The encoding of the data $D$ using $G$ will be written as $D : G$, and the tightness of fit will be the length of this encoding, which we will write as $|D : G|$. Using this notation, MDL can be stated as follows:[3]

[2] Recent work in this vein includes Piantadosi et al. 2016 and Pearl et al. 2017.

[3] 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

(1) MDL EVALUATION METRIC: If $G$ and $G'$ can both generate the data $D$, and if $|G| + |D : G| < |G'| + |D : G'|$, prefer $G$ to $G'$

The balancing of grammar 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 & Ristad (1994), Stolcke (1994), Grünwald (1996), de Marcken (1996), Brent (1999), Clark (2001), and Goldsmith (2001), among others.

In the context of language learning, MDL can be thought of as follows. Universal Grammar provides a template for specifying grammars. In a sense, it is a programming language in which various programs – equivalently, grammars – can be written. The programming language says exactly how programs are written, and this specification is provided ahead of the actual data that are encountered and independently of them. $|G|$, from this perspective, is simply the storage space required for representing the grammar/program $G$ according to the specifications of the programming language. (What can be stated in a grammar and how much it costs can of course vary greatly between programming languages. One programming language, for example, might allow context-sensitive rewrite rules to be stated and might make vowels costlier than consonants. Another programming language will allow markedness and faithfulness constraints to be stated and will make consonants costlier than vowels. And so on.) While grammars are stated in terms of an *a priori* programming language, they also interact with the particular data $D$ that the child happens to see. In particular, a grammar can be used to parse $D$, which can be thought of as a sequence of instructions to $G$ that result in the generation of $D$. This sequence of instructions is what we referred to above as $D : G$. The MDL metric says that the best grammar $G$ given the data $D$ is the one that minimizes the overall storage space of the grammar itself (stated according to the innate programming language) and of the sequence of instructions through which $G$ generates $D$. In the remainder of the present section, we will try to make the above discussion concrete by going through two relevant linguistic examples.

### 2.2 *Two examples: learning alternations and phonotactics using MDL*

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 $|D : G|$). As an illustration, consider the pattern of optional L(iquid)-deletion in French, discussed by Dell (1981) and Rasin et al. (2018a). In French, L-deletion applies optionally in certain contexts – specifically, word-finally and following obstruents. For example, the French-learning child might hear both [tabl] and [tab]

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).

'table' and both [arbr] and [arb] 'tree' (but only [gar] 'train station'). With some simplification, the relevant process can be described as follows:

(2) $L \rightarrow \emptyset$ /[–*son*]__# (optional)

A learner minimizing $|G|$ will favor collapsing pairs such as [tabl] and [tab] onto a single UR (/tabl/) and deriving the two surface forms using a process of optional $L$-deletion. This is so since, with sufficiently many alternating surface pairs, the savings obtained by storing just one UR for each will outweigh the costs of adding the relevant rule to the grammar. To illustrate the trade-off between grammatical statements and lexicon size, consider the two competing grammars in (4): G1 has simply memorized the data without making any rule-based generalization, while G2 generates pairs like [tabl] and [tab] from a single UR /tabl/ using the optional L-deletion rule.

(3) (a) G1 (costlier, memorizing):

- Lexicon: /tabl/, /tab/, /arbr/, /arb/, /gar/, . . .
- Rules: ∅

(b) G2 (cheaper, correct):

- Lexicon: /tabl/, /arbr/, /gar/, . . .
- Rules: $L \rightarrow \emptyset$ /[–*son*]__# (optional)

There are various conceivable ways of measuring grammar size. In this paper, we will assume – in the spirit of the evaluation metric of early generative grammar – that grammar size grows with the number of symbols used to state URs in the lexicon as well as the processes (either rules or constraints). So, for example, the cost of the L-deletion rule of G2 above would be the cost of the six symbols $L$, ∅, −, *son*, __, and #, and the cost of storing of URs like/tab/ in the lexicon would be the cost of the relevant three symbols (one symbol for each segment). More precisely, within MDL, grammar size is measured in terms of bits. We can think of the calculation of the size of the grammar as follows. We first obtain a string representation of the grammar. For G1 and G2 above, the string representations would be along the lines of the following strings (with various delimiters separating URs, rules, and the different components of the grammar; the three dots are placeholders for the rest of the lexicon):

(4) String representations of G1 and G2 (illustration)

(a) G1 = tabl#tab#arbr#arb#gar#(· · · )##

(b) G2 = tabl#arbr#gar#(· · · )#L→ ∅/-SON__#Yes$_{(optional)}$##

String representations are then converted into bit strings, according to an encoding scheme that assigns a binary code to each symbol that can be used in writing grammar-string representations. For example, if the segment *t* is assigned the code 0010 and the segment *a* is assigned the code 1011, the beginning of the

bit strings for both grammars would start with the bits 0010 1011. Rasin et al. (2018a) discuss specific encoding schemes, but here we will not commit to one encoding scheme in order to keep the discussion general. What will matter for the discussion is the assumption that the number of bits required to specify a symbol grows with the number of symbols the choice is made from. For example, if the possible set of symbols for writing URs contains 7 symbols including *t*, the cost of specifying *t* in the lexicon would be, say, 3 bits, but if the possible set of symbols for writing URs contains 10 symbols including *t*, the cost of specifying the choice of *t* from among that set would be higher, say, 4 bits. In other words, a larger alphabet for the lexicon results in a higher cost for each underlying segment, a point that will play a central role in our discussion in Section 4. After the string representation of a grammar is converted to a bit string, the size of the grammar is taken to be the length of the resulting bit string. In our current French example, the bit string for G1 will be longer than the bit string for G2, provided that the data include enough collapsible pairs such as [tabl] and [tab], which justifies the addition of the L-deletion rule to G2.

This kind of balancing within the $|G|$ component – adding a phonological process to the grammar despite its length, since the addition makes it possible to attain savings within the lexicon that outweigh the added complexity – is familiar from early work in generative phonology (in particular, Halle 1962). It will also play a central role in our discussion in Section 4. However, as noted by Dell (1981), if the learner *only* minimizes $|G|$ (as in the evaluation metric of early generative phonology but differently from MDL, where it is balanced by $|D : G|$), the learner will overgeneralize, preferring the incorrect but shorter (5), which permits L-deletion regardless of context, to the correct but more complex (2).

(5) $L \rightarrow \emptyset$ (optional)

Minimizing the second component of the MDL metric, $|D : G|$, helps address this problem by ensuring that the grammar is restrictive. As mentioned above, $D : G$ lists a sequence of instructions for $G$ that result in the generation of $D$. In the case of rule-based phonology, this sequence includes the instruction to pick a specific UR from the lexicon, along with instructions for whether any applicable optional rule applies. A grammar that is more restrictive will allow the input data to be specified using fewer instructions than a less restrictive one and will thus typically result in a shorter $D : G$. For example, an instruction to make a binary choice (such as applying or not applying an optional rule whose structural description is met) will cost 1 bit. In the present case, a grammar using the less restrictive (5) will require a specification within $D : G$ for each underlying liquid stating whether it is deleted or not. Thus, for example, encoding the surface form [gar] using (5) will require first specifying the choice of the UR /gar/ from the lexicon and then – since (5) optionally deletes liquids anywhere – paying an additional bit to specify that optional deletion does not apply to the final segment. With the more restrictive (2), on the other hand, only liquids that are in the appropriate context (word-finally and following an obstruent) require such

a specification. This means that under (2), encoding the surface form [gar] does not require an extra bit for specifying that deletion does not apply to the final segment, leading to a shorter $D : G$ (see Rasin et al. 2018a for a more detailed discussion of measuring $|D : G|$).

Like minimizing $|G|$ alone, minimizing $|D : G|$ alone is problematic. Specifically, while helpful in avoiding overly general grammars, minimizing $|D : G|$ alone gives rise to a problem of *under*-generalization: it leads the learner to overfit the data, regardless of how complex the grammar becomes. For example, if the learner has heard only [sabl] but not yet its $L$-elided variant [sab] (both for the UR /sabl/ 'sand'), we might expect them to generalize beyond the data and accept the L-deleted [sab]. A learner minimizing $|D : G|$ alone, however, will incorrectly prefer to treat [sab] as ungrammatical: by doing so, when the UR /sabl/ is selected, no further statement as to whether the final $L$ is deleted will be needed, which in turn leads to a shorter $D : G$.

By minimizing the sum of $|G|$ and $|D : G|$, MDL aims at an intermediate level of generalization: the $|G|$ component biases the learner toward relatively general grammars, while the $|D : G|$ component protects the learner from overly general grammars. In the case of French $L$-deletion, MDL will avoid the excesses of both $|G|$ minimization (which leads to free $L$-deletion, as in the short (5)) and $|D : G|$ minimization (which leads the overfitting memorization of [sabl] as non-alternating). Instead, it will lead to the correct (2): this grammar is only slightly more complex (and less general) than (5) but fits the data much better; and it fits the data only slightly less well than a grammar that rules out [sab] but is much less complex than it. The three grammars and their MDL costs are illustrated schematically in Figure 1.

In addition to supporting the learning of patterns involving alternations, like L-deletion in French, MDL can also ensure the learning of phonotactic patterns. Here, the trade-off between the size of the lexicon and the size of the component where processes are specified also plays an important role. This trade-off will be particularly important in Section 4, where we discuss the possible choices and their implications in detail. For now, though, let us illustrate this with a well-known example from English, due to Chomsky & Halle (1965) and originally stated in a learning framework that is different from MDL (specifically, the simplicity metric, where only $|G|$ is minimized). We wish to emphasize that we use Chomsky & Halle's case only as an informal motivational example and will present the case only in very rough outline, leaving important questions unaddressed.

As noted by Halle (1962) and Chomsky & 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, [blɪk] is an accidental gap while [bnɪk] is a systematic gap. With respect to [bnɪk], it is generally true that if an English word begins with a stop, the following consonant cannot be nasal: it must be a liquid. The lexicon of English can be simplified by removing the specification of the [liquid] feature

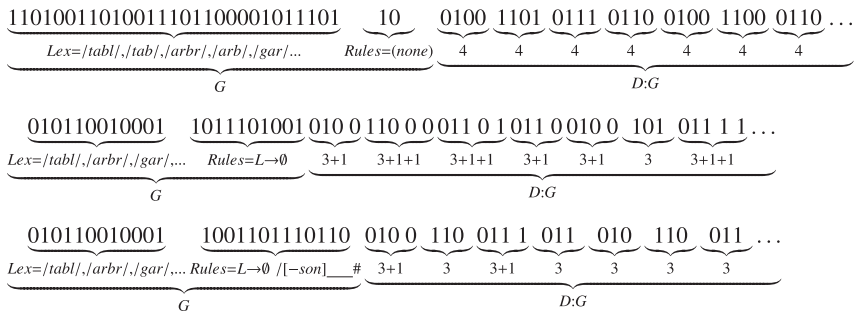


*Figure 1*

Schematic illustration of three hypotheses. (The order of URs in the lexicon and of tokens in $D : G$ are unrelated.) Introducing a naive lexicon (*top*), in which [tabl] and [tab] have distinct URs results in a complex grammar. Capturing optional $L$-deletion with (5) allows the grammar to be simplified (*middle*): the complexity of the rule is outweighed by the savings of eliminating unnecessary URs. Moreover, since there are now fewer URs than with the naive lexicon, each UR can be specified more succinctly. However, an additional bit is needed for specifying the actual surface form of each occurrence of $L$ in a UR (for each surface token of that UR). Finally, restricting the context of $L$-deletion, using (2), allows us to limit the extra bit to just those URs that require it (*bottom*): e.g., /tabl/ but not /gar/.

from many consonants:[4] for example, the word /brɪk/ could be represented without a specification that /r/ is a liquid. Instead, this feature could be specified using the following rule:

(6) [+*consonantal*] → [+*liquid*] / # [+*stop*] __

Adding (6) to the grammar will make the rule component more complex, but this added complexity will be easily offset by the savings in the lexicon, which are obtained by eliminating the [liquid] specification from the second consonant of every English word starting with a stop-consonant cluster. A learner that minimizes both the size of the lexicon and the size of the rules as part of $|G|$ will add (6) to the grammar and correctly treat [bnɪk] as ungrammatical.[5] For words like [blɪk], as Chomsky and Halle note, the situation is different, since excluding such words by means of a rule would require a rule whose complexity outweighs the savings in the lexicon. For instance, since *brick* is an English word, a rule that excludes [blɪk] will have to be very specific and ensure that a liquid is never [l] in this specific environment (while leaving other consonants unaffected):

[4] A more detailed and more current account might replace the atomic feature *liquid* by other, more standard features, and should consider exactly what savings would be obtained by removing those alternative features given the rest of the available set of features and the dependencies between them.

[5] Note that $|D : G|$ remains constant, since the addition of the rule does not affect the instructions for generating words like [brɪk]. Both with and without the rule, the instructions are to choose the UR for [brɪk] from the lexicon and to apply the rules deterministically.

(7) [+*consonantal*] → [–*lateral*] / # b __ ik

Rule (7) is quite complex, but it only allows eliminating a single feature from the lexicon: the [lateral] feature of the second consonant in the word *brick*. By minimizing both the size of the lexicon and the size of the rules, a learner that minimizes $|G|$ will not add (7) to the grammar and will correctly treat [blɪk] as an accidental gap. Note that in this example there are two phonotactic patterns in the data that could in principle be captured by the grammar: the absence of word-initial sequences like #bn and the absence of the phonotactic configuration #blɪk#. With a $|G|$-minimizing learner, only the phonotactic pattern that leads to an improvement in terms of $|G|$ – namely, the ban on #bn – is captured by the grammar.

### 2.3 *Motivation for testing the predictions of MDL*

MDL is certainly not the only learning framework that has been proposed for phonology.[6] As mentioned above, our goal in this paper is not to argue against alternative learning approaches. Rather, we will briefly explain (in the remainder of this section) why we find MDL promising, so as to motivate our exploration below of its implications for the representational question of language-specific statements in the lexicon.

Our first reason for finding MDL promising is that it has 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 & 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.[7] The MDL learner of Rasin et al. (2018a) and Rasin et al. (2018b) 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. The learner has been shown to handle a range of linguistically relevant phenomena such as opacity

[6] In particular, there is a prominent alternative, usually stated within constraint-based phonology, according to which the child attempts to find the most restrictive grammar consistent with the data. On one approach that adopts 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 in the face of conflicting evidence, with individual prohibitions being eliminated or demoted. Representative proposals of this approach include an initial ranking of markedness over faithfulness ($M \gg F$; Smolensky 1996, Tesar & Smolensky 1998) from which a search for a consistent ranking begins, as well as a more sustained bias for $M \gg F$ (Hayes 2004, Prince & Tesar 2004) throughout the search for a consistent ranking. A different approach that adopts this view is that of Jarosz (2006, 2009), who defines a general measure of restrictiveness (specifically, the likelihood of the data given the grammar, which is closely related to the $|D : G|$ component of MDL but without balancing it with the $|G|$ component) and performs a search for a hypothesis that maximizes this measure.

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

(including both counterfeeding and counterbleeding), optionality, and the long-distance dependencies of vowel harmony.

Our second reason to focus on MDL is that it has been supported empirically by a range of lab experiments on generalization. In a variety of learning tasks in the lab, ranging from word learning (Xu & Tenenbaum 2007) through causal reasoning (Sobel et al. 2004) to sensorimotor control (Körding & 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.

Our third reason to focus on MDL is methodological. As mentioned above, we can think of MDL in terms of an innately given programming language in which grammars ($G$'s) are stated, grammars that can then be used to parse the input data $D$. As noted in Katzir 2014, this perspective highlights how much of the basis for MDL learning is available to the child without further stipulation. The grammars are real cognitive objects: they are stored according to the specifications of the programming language and take up space. Similarly, if the child remembers $D$ as parsed by $G$ – the sequence of instructions for generating $D$ using $G$ that constitutes an understanding of $D$ using $G$ – then they also store $D : G$ in memory, and that also takes up space. The overall storage space for the two specifications, $|G| + |D : G|$ is exactly the quantity used by MDL. All that is needed is a way to compare neighboring grammars in terms of this quantity and search for the one that minimizes it. If the above is indeed correct, MDL constitutes a simple, non-stipulative approach to learning, which is then a sensible starting point to an investigation of learning (and can of course be modified when empirical reasons are encountered to add stipulations).

## 3. Where are phonological generalizations captured?

The architectural question we will investigate from the perspective of MDL learnability is whether the grammar includes language-specific statements about the lexicon: early generative phonology assumed that such statements were possible, while more recent work mostly disallows them. As a reminder of where such statements might matter, we briefly repeat the distinction between accidental and systematic gaps and then review two central approaches to the capturing of this distinction.

Recall that as noted by Halle (1962) and Chomsky & 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, as we saw above for an example from English phonotactics. Here we focus on a different example, from Dutch, which will serve us in the discussion to follow. In Dutch, 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.[8] 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 the assumption of MDL 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;[9] and phonological rules. In the example above, an early generative account might use a CUR such as (8) and a phonological rule such as (9) as the basis for capturing the distribution of stridents in Dutch:[10]

(8) CUR IN DUTCH: No ʃ in the lexicon

(9) [+*strident*] → ʃ / __ j

(8) ensures that stridents will be alveolar underlyingly, while (9) ensures that they will become palatalized in exactly the right environment. The combination of (8) and (9) 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 /osjər/; the palatalizing rule in (9) would then turn

[8] 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.

[9] Halle (1959, 1962) proposed capturing 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, though in recent years the use of the term morpheme-structure constraints has been sometimes extended to refer to morpheme-structure generalizations that hold on the surface (see, e.g., Rose & Walker 2004 and Ozburn & Kochetov 2018). To make it clear that we refer to grammatical statements that restrict underlying forms, we will use CURs as a cover term for rules or constraints of this kind.

[10] 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 [ʒ]. As far as we can tell, this simplification is orthogonal to the argument we will present.

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 (9) using *anterior* would be [+*strident*] → [–*anterior*] / __ j. To simplify the presentation, we will use symbols (like [s] and [ʃ]) instead of features whenever possible.

the latter into its surface form. For *[ɔʃtər] and *[osjər], the situation is different. Since (8) prohibits the storing of /ʃ/ in the lexicon of Dutch, [ʃ] must follow from rule application; but the palatalization rule in (9) 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 (10) below, which makes stridents alveolar regardless of their underlying specification or of their environment:

(10) [+*strident*] → s

If (10) is ordered before (9), 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 & Smolensky 1993), where a representational principle, *Richness of the Base*, prevents CURs from being stated:[11]

(11) Richness of the Base (ROTB; Prince & 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.

In terms of the perspective discussed above of UG as an innate programming language, ROTB entails that programs cannot change the costs of storage in the lexicon: encoding costs for URs are universal.

[11] 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 4) or even more neutrally, using mapping statements as in Tesar 2014. Stated in terms of constraints, the first approach would combine the CUR in (8) with a constraint ranking such as *sj ≫ IDENT[ANT], while the second approach would avoid (8) 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.

Clearly, the two representational choices for handling 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 (for relevant discussion, see, e.g., Smolensky 1996, Vaysman 2002, Peperkamp & Dupoux 2003, Booij 2011). Below, we will show that for an MDL learner, the representational choice of CURs vs. ROTB has meaningful implications.

## 4. The MDL-learnability implications of ROTB

We now turn to our investigation of the consequences of ROTB for learnability if MDL-based learning is assumed. Our discussion will center on a problem that arises when a phonological grammar captures aspects of the input data as an accident of the lexicon while speakers' knowledge appears to be systematic. We will refer to this as the *Extensional Restriction Problem* (ERP) and state it as follows:

(12) Extensional Restriction Problem. A grammar *G* suffers from the Extensional Restriction Problem for an element *x* in context *C* if both of the following hold:
   (a) Speakers reject instances of *x* in *C* in nonce words.
   (b) In *G*, the only way to avoid instances of *x* from surfacing in context *C* is extensionally, by not listing them in *C* in the lexicon.

Concretely, and using the case of the distribution of stridents in Dutch, *x* might be [ʃ], *C* might be positions that do not precede [j], and *G* might be a grammar that allows /ʃ/ to be written freely in the lexicon and mapped faithfully to the surface but in which /ʃ/ simply happens not to be written in any position that does not precede /j/. Since Dutch speakers systematically reject occurrences of [ʃ] in positions that do not precede [j], such a *G* will suffer from the ERP for [ʃ] in non-pre-[j] positions. The reason that such a *G* is problematic is that a novel form with *x* in *C*, which speakers reject (e.g., *[ɔʃtər] in Dutch), can incorrectly be accommodated in *G* by adding a suitable UR to the lexicon (e.g., /ɔʃtər/): this is possible since *G* only avoids the relevant surface forms extensionally, through an accident of the lexicon, rather than through a systematic statement of the grammar such as a markedness constraint *ʃ ranked sufficiently high.[12]

[12] 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 4.3 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).

As we will see, on the assumption of MDL learning, ROTB gives rise to the ERP for some $x$ and $C$ across a range of possible representational assumptions (the identity of $x$ and $C$ varies across these representational assumptions). CURs, on the other hand, avoid the problem.[13]

We will point out two possible responses to this predicament. The first response is to abandon ROTB and admit CURs, which, as just mentioned (and as we will show in greater detail below), 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, a more imaginative way to maintain ROTB is to use underspecification in the storage of URs (see Archangeli 1988, Steriade 1995, and Inkelas 1995, among others). 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. (For OT, as opposed to rule-based phonology, underspecification will not help maintain ROTB.)

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: with MDL as the learning criterion, 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 paper. 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 5.

In order to develop our argument, we will need to examine the MDL implications of the possible choice points under various reasonable representational

[13] Using the terminology of Chomsky (1964: p. 29), a grammar that suffers from the ERP is observationally but not descriptively adequate. On the assumption of MDL learning, this will mean that a theory that adopts ROTB lacks explanatory adequacy.

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 & Pater 2012, Pater 2014, and references therein, as well as the earlier literature on arbitrary phonological rules, e.g., Bach & 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 in Section 4.1, and introduce the complications of constraints that are given in advance only later, in Section 4.2. The configurations we discuss in Sections 4.1 and 4.2 are summarized in Figure 2. The discussion in Sections 4.1 and 4.2 assumes full specification. The possibility of underspecification has interesting implications for the learning pattern under consideration, and we turn to this possibility in Section 4.3.

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

*Figure 2*
Summary of the configurations we discuss in Sections 4.1 and 4.2. 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 $M \gg F$.

### 4.1 *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 (along the lines of our discussion in Section 1), and suppose further that /s/ and /ʃ/ each costs some fixed number of bits to store in the lexicon. Using the perspective of an *a priori* programming language discussed above, we can say that the programming language determines costs for /s/ and /ʃ/ that do not depend on the position of the segment in a given UR and that hold in any grammar that does not make further statements about the format in which URs are stored (statements that are possible if CURs are allowed but are not possible under ROTB).

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. In this subsection, we assume that features are fully specified in the lexicon.

### 4.1.1 *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/). Assume that the child encounters Dutch words such as [ɛɪs] 'ice cream' and [ɛiʃjə] 'a small ice cream'. We first consider the situation of the child on the assumption that ROTB holds and show that in this case MDL chooses a grammar that suffers from the ERP as stated in (12).

Since MDL prefers an economical lexicon, it will push the child toward segmenting words like [ɛɪs] and [ɛiʃjə] into morphemes and toward storing repeating morphemes using a single UR in the lexicon (e.g., either /ɛis/ or /ɛiʃ/ for 'ice cream'). Since instances of /ʃ/ are costly to store in the lexicon, it will be preferable in terms of MDL to store all stridents as /s/ (e.g., prefer /ɛis/ to /ɛiʃ/ for 'ice cream') and invest in constraints that trigger palatalization of /s/ before /j/ (e.g., the markedness constraint *sj, ranked above the appropriate faithfulness constraints). Adding the constraints will cost a few bits, but this cost will be outweighed by the savings from not having to store multiple entries for a single morpheme and from not having to store any instances of the costlier /ʃ/ in the lexicon. Two competing hypotheses are illustrated in (13), with G2 being preferred by a ROTB child that follows MDL. In these hypotheses, the lexicon is broken down to the alphabet (used to write URs in the lexicon) and the URs. By choosing G2 over G1, the child has successfully learned that pre-palatal stridents are systematically palatalized in the language.[14] For example, the child will now correctly rule out forms such as *[osjər], with an alveolar [s] before [j].

[14] The presentation of successive grammars considered by the learner is made for convenience only. The implications of MDL that we discuss in this paper depend only on which grammar is the global optimum, not on the order in which grammars are evaluated.

(13) Available first step for a ROTB child with MDL

(a) G1 (costlier, incorrect): instances of the costly /ʃ/ in URs, faithful input-output mapping

- Lexicon: Alphabet = /s/, /ʃ/, . . . ; URs = /ɛɪs/, /ɛɪʃjə/, . . .
- Constraints: FAITH[15]

(b) G2 (cheaper, partly correct): segmented lexicon, absence of /ʃ/ in URs, constraints forcing palatalization of /s/ before /j/

- Lexicon: Alphabet = /s/, /ʃ/, . . . ; URs = /ɛɪs/, /jə/, . . .
- Constraints: *sj ≫ IDENT[ANT] ≫ . . .

Unfortunately for ROTB, G2 is also the extent of the child's acquisition of the pattern under MDL. In particular, the child will not learn to block forms such as *[ɔʃtər], with [ʃ] in 'elsewhere' environments. The reason is that, for a ROTB child, such forms must be blocked through the input–output mapping, for example through *ʃ or a similar markedness constraint that penalizes [ʃ], as in G3 in (14) below. And there is simply no MDL justification for acquiring this kind of constraint. Under G2, all stridents are already stored in the lexicon as /s/ and are mapped faithfully to the surface. Consequently, a markedness constraint such as *ʃ in G3 will be of no use in deriving the observed forms in the input data from the lexicon, and the cost of adding such a constraint to the grammar will not be justified. A ROTB child, then, will become an adult who knows only half of the distributional pattern of stridents in Dutch: having learned G2, 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). In other words, a ROTB child will arrive at a grammar, G2, that suffers from the ERP for [ʃ] in non-pre-[j] positions. Note that markedness constraints such as *ʃ in the present case are analogous to rule (7) (which blocked [blɪk], as discussed in Section 1), which had no benefit in terms of MDL and thus was not added to the grammar, leaving a phonotactic pattern in the data as accidental.

(14) G3 (correct but costlier than G2, will not be chosen by MDL): like G2, but with an additional constraint banning [ʃ] from 'elsewhere' environments

- Lexicon: Alphabet = /s/, /ʃ/, . . . ; URs =/ɛɪs/, /jə/, . . .
- Constraints: *sj ≫ *ʃ ≫ IDENT[ANT] ≫ . . .

If ROTB does not hold and CURs are allowed, the learning process can succeed in full and arrive at a grammar that does not suffer from the ERP. The first step – choosing G2 – is similar to the one with ROTB: the child will invest in a constraint

[15] See Rasin & Katzir 2016 for why faithfulness constraints will be acquired by an MDL learner.

like *sj and then store all stridents as /s/; again, the result 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 (8) above):

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

Let us first see why (15) is justified in terms of MDL. All things being equal, removing a possible segment from the underlying inventory makes the lexicon easier to encode. This is because of the assumption, mentioned in Section 1, that specifying a choice from a smaller set requires fewer bits. If the alphabet of Dutch includes /ʃ/, every segment in the lexicon (e.g., ɛ, ɪ, and s in the UR /ɛɪs/) will have to be specified from among the alphabet that includes /ʃ/. If, however, the alphabet of Dutch does not include /ʃ/, specifying the same segments (e.g., ɛ, ɪ, and s in the UR /ɛɪs/) will be made from among the smaller alphabet and require fewer bits. As a result, the entire lexicon will be cheaper to encode with a smaller alphabet, thus providing MDL justification for adopting (15). The resulting grammar, now with (15), is given in (16).

(16) G4 (correct and cheaper than G2, will be chosen by MDL): like G2, but underlying /ʃ/ impossible

- Lexicon: Alphabet = /s/, . . . (no /ʃ/); URs = /ɛɪs/, /jə/, . . .
- Constraints: *sj ≫ IDENT[ANT] ≫ . . .

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 grammar like G4, that has a CUR like (15), the child will now correctly rule out surface forms like *[ɔʃtər]:/ɔʃtər/ is now no longer a possible UR; and given the constraints that have been induced, no other UR is a 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.

Note that while the CUR in (15) and a markedness constraint like *ʃ are similar in their complexity and their role in the adult grammar, they have a different status from the perspective of MDL, which causes an MDL learner to acquire the CUR but not the markedness constraint. The key difference from the perspective of MDL is that the CUR allows for a shorter specification of the lexicon. A surface constraint, on the other hand, only affects the input–output mapping and does not contribute to savings in the lexicon. In this case, and in many others, this difference will result in a CUR being acquired (if UG allows it to be represented) while a mapping statement such as a markedness constraint will not be acquired, even when the two are very similar.

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.

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

Suppose now that *Cost*(/ʃ/) < *Cost*(/s/). In this case, the Dutch-learning child will be able to avoid the ERP even with ROTB. They can do so by storing /ʃ/ throughout the lexicon and acquiring constraints that will ban [ʃ] in 'elsewhere' (that is, non-pre-palatal) environments: before a vowel, before a non-palatal consonant, word-finally, and so on. In this case, it will be costly to state constraints that ban [ʃ] directly in all of 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, the child 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 avoiding the ERP and allowing ROTB to be maintained.[16]

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 – a pattern with stridents that are alveolar in some specific environments but are palatalized elsewhere – unlearnable on the assumption of ROTB. 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 [t̪uʃka], with an [ʃ] before a velar consonant, and [t̪ust̪a], with an [s] before a dental consonant, are both accidental gaps, while the nonce form *[t̪uska], with an [s] before a velar consonant, is a systematic gap.[17] An appropriate constraint ranking for the Bengali pattern (ignoring both optionality and word-initial clusters) that does not rely on CURs would be the following:

[16] Of course, the precise constraints and their implications can vary, depending on the complexity of describing the 'elsewhere' environments. Suppose that 'elsewhere' environments were easy to characterize. In that case, a learner following ROTB would acquire a single markedness constraint banning [ʃ] in those environments – say, *ʃT (no [ʃ] before a stop) – and be done. In particular, *sj would not be learned, and the target grammar would 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 – here with a grammar that suffers from the ERP for [s] in pre-[j] positions – and again the ability to violate ROTB and remove elements from the alphabet of the lexicon would allow the learner to acquire the full pattern.

[17] We consulted two native speakers of Bengali who confirmed that the forms [t̪uʃka] and [t̪ust̪a] are judged as acceptable, whereas *[t̪uska] is judged as unacceptable. 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 and the other accepted it as well-formed. This variability is consistent with our argument. For example, it could be that the process applies obligatorily in one dialect and optionally in another, and our argument is independent of whether the process applies obligatorily or optionally.

(17) 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. In particular, it will lead a ROTB child to a grammar that suffers from the ERP for [s] in pre-[j] positions. Given the present cost assignment, a ROTB child 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 (here, a CUR that removes /s/ from the alphabet of the lexicon).

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

Consider now the possibility that *Cost*(/ʃ/) = *Cost*(/s/). In this case, the outcome depends on whether information from alternations (like [ɛis]~[ɛiʃjə] in Dutch) is available and can be used by the learner, but either way the result with ROTB will be a grammar that suffers from the ERP.

If for whatever reason the learner cannot make use of alternations, MDL 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/, MDL 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. A ROTB child will consequently fail to reject both *[ɔʃtər] and *[ɔsjər] – that is, a ROTB child will arrive at a grammar that suffers from the ERP for both [ʃ] in non-pre-[j] positions and [s] in pre-[j] positions.

If, on the other hand, the learner can make use of alternations, then in the present case of equal costs for /s/ and /ʃ/, such 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 [ɛis] and [ɛiʃjə] 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 MDL purpose and thus will not be learned; in this case, a ROTB child 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 MDL purpose and thus will not be learned; in this case, a ROTB child will fail to reject *[osjər].

### 4.1.4 *Contextualized costs for /s/ and /ʃ/*

The problem for ROTB extends to some other representational possibilities that UG might make available. For example, suppose that UG makes the cost 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 the full Dutch pattern unlearnable by an MDL learner for the same reasons as discussed above for the case of identical costs. If the learner cannot make use of alternations, it 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; and as before, even with alternations there will be no reason for the learner to learn to ban [ʃ] in 'elsewhere' environments.

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,[18] the contextualized cost assignment will have the same effect as *Cost*(/ʃ/) < *Cost*(/s/): a ROTB child 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 MDL learning that avoids the ERP, assuming that the constraints are not given in advance.

## 4.2 *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 (keeping our other assumptions from Section 4.1 unchanged), building on the argument in Rasin & Katzir 2015 that, unless a preference for markedness over faithfulness is incorporated, an MDL learner

[18] Ferguson & 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.

would still need to abandon ROTB and adopt CURs to avoid the ERP. 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 Section 4.1.1, 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 MDL 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 MDL 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 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.[19] 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 $M \gg F$ becomes relevant.[20] In the settings discussed in Section 4.1, with binary features and acquired constraints, $M \gg F$ 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, $M \gg F$ 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 $M \gg F$ addresses this challenge.[21] The combination of $M \gg F$ 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

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

[20] As discussed in footnote 5, variants of such a preference have been used within other learning approaches in the literature, 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 $M \gg F$ can similarly be implemented, most straightforwardly through the cost scheme for the statement of rankings.

[21] 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 $M \gg F$ (but otherwise remains a problem for a ROTB child). Note that the constraint *s, which will be given in advance as well, will be properly ranked in Dutch given $M \gg F$ without causing trouble. The ranking will be *sj ≫ *ʃ ≫ *s ≫ IDENT[ANT]: *s can in principle be ranked below or above IDENT[ANT], but $M \gg F$ will lead to the former being ranked higher. And ranking *s higher than *ʃ will result in a ranking that bans any [s] on the surface. Since this ranking would not be able to generate the Dutch data, it would not be considered by the learner.

part of the initial state). We now turn to a less stipulative response available within rule-based phonology.

### 4.3 *Special case: rule-based phonology with underspecification*

So far the discussion in this section assumed that features are fully specified. This assumption contributed to the fact that a 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 language-specific 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 $M \gg F$). A third response suggests itself if underspecification is allowed (for general discussion of underspecification in generative phonology, see Archangeli 1988 and Steriade 1995). If storing an underspecified value for anteriority is less costly than either of the specified values, the learner might prefer storing all stridents unfaithfully as underspecified for anteriority and invest in the markedness constraints *sj and *ʃ, along with a high-ranking markedness constraint 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 initially appear to be the case: 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. As before, abandoning ROTB and adopting language-specific CURs allows the learner to succeed straightforwardly.

The problem with the use of underspecification by an MDL learner on the assumption of 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 child might have avoided positing such a constraint (depending on feature costs): in analogy with our discussion of Dutch and Bengali, an incorrect grammar such as the following $G_0$, which stores voicing faithfully (and which has no need for *ɣ), could have been optimal.

(18) $G_0$ (complex; incorrect)
  (a) Lexicon: [k], [g], and [x] are stored faithfully
  (b) Constraint ranking: FAITH

With underspecification, faithful storage of this 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 grammar $G_1$ in (19). Here we use [0*voice*] to refer to an underspecified value for voicing, which could mean either that a feature value for voicing is not listed in the lexicon at all or that [0*voice*] acts like a third value that is listed and could be referred to by the grammar. The constraint *[0*voice*] penalizes underspecified values on either of these two options.

(19) $G_1$ (simpler than $G_0$; correct)
  (a) Lexicon: [x] is stored as [0*voice*][22]; [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:

(20) $G_2$ (simpler than $G_1$; 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

[22] As pointed out by a reviewer, some (but not all) previous work on underspecification has assumed that underspecification is not possible in the absence of observable surface alternations (e.g., Inkelas 1995 and Hale & Reiss 2008). Our discussion of underspecification assumes that learning follows MDL. If UG makes underspecified representations available in principle, then given MDL, underspecification is possible even without evidence from alternations, since underspecification can make the lexicon cheaper to encode even in the absence of alternations.

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, a ROTB child 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 an innate *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 $G'_0$, $G'_1$, and $G'_2$ serve as counterparts of $G_0$, $G_1$, and $G_2$:

(21) $G'_0$ (complex; incorrect)
  (a) Lexicon: [k], [g], and [x] are stored faithfully
  (b) Rules: (none)

(22) $G'_1$ (simpler than $G'_0$; correct)
  (a) Lexicon: [x] is stored as [0*voice*]; [k], [g] are stored faithfully
  (b) Rule: [*velar*, +*cont.*] → [–*voice*]

(23) $G'_2$ (simpler than $G'_1$; 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 way to avoid the problem: the simple but incorrect $G'_2$ crucially relies on referring to underspecified values within a rule, while the more complex but correct $G'_1$ does not. (This contrasts with the case of the constraint-based $G_1$ and $G_2$ above, neither of which required referring to underspecified values in their constraints: for those grammars, *[0*voice*] simply penalizes feature vectors that are underspecified for voicing.) So 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 (21) will be learned at all), then the incorrect $G'_2$ will not be statable, and 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):

(24) 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 (24), CURs are not necessary for an MDL learner to acquire systematic gaps (as in the Dutch, Bengali, and German patterns above) even if knowledge of the pattern is not given to the learner in advance.

## 5. Discussion

We examined the implications of MDL learning for phonological architecture, focusing on the distinction between accidental and systematic gaps. We observed that 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 gaps on the assumption of MDL learning. 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. If the representational condition of ROTB holds, there will generally be some element $x$ and context $C$ such that the absence of $x$ in $C$ on the surface is captured by simply not listing $x$ in $C$ in the lexicon. In such cases, it will serve no MDL purpose to invest in a mapping statement (such as a markedness constraint) to eliminate $x$ in $C$. But then, given ROTB, it will be possible to accommodate a form with an inappropriate $x$ in $C$ by adding a new UR with $x$ in $C$ to the lexicon. Given ROTB, $x$ can be written there, and given the absence of a relevant mapping statement, $x$ will surface faithfully. In other words, MDL will select a grammar that suffers from the ERP for $x$ in $C$. If ROTB is abandoned, on the other hand, the ERP will be avoided. It will be possible to acquire a CUR that eliminates $x$ in $C$ from the lexicon. This will serve the MDL purpose of making the lexicon less costly to encode, and it will prevent $x$ from being written in context $C$.

The solution, assuming MDL learning, 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 language-specific 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 the relevant segments 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 on the assumption of MDL learning, 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).[23] 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.[24] Since language-specific constraints have occasionally been proposed in the literature (see, e.g., Kager & Pater 2012, Pater 2014, and references therein) and since the empirical 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 paper.

Looking past the specific question of CURs versus ROTB, this paper illustrates a way in which a general learning criterion can help evaluate competing representational possibilities. The idea is not new. 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.[25] More recently, Katzir (2014), Piantadosi et al. (2016), and Katzir et al. (2020) 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 paper offers a concrete application of this idea to an actual architectural question.

[23] 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 simply happens to be true of actual speakers.

[24] As pointed out by a reviewer, defending markedness over faithfulness and constraints given in advance also means that OT cannot be substance-free as in Blaho 2008.

[25] 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. 2018b 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, the simplicity argument by Halle (1962, 1978) for feature-based representations stands.

## REFERENCES


Anderson, Stephen R. 1981. Why phonology isn't 'natural'. *Linguistic Inquiry* 12.4, 493–539.

Archangeli, Diana. 1988. Aspects of underspecification theory. *Phonology* 5.2, 183–207.

Bach, Emmon & Robert T. Harms. 1972. How do languages get crazy rules? In Robert P. Stockwell & Ronald K. S. Macaulay (eds.), *Linguistic change and generative theory*, 1–21. Bloomington: Indiana University Press.

Baker, Carl L. 1979. Syntactic theory and the projection problem. *Linguistic Inquiry* 10.4, 533–581.

Berwick, Robert C. 1982. *Locality principles and the acquisition of syntactic knowledge*. Ph.D. dissertation, MIT.

Blaho, Sylvia. 2008. *The syntax of phonology: A radically substance-free approach*. Ph.D. dissertation, Universitetet i Tromsø.

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

Booij, Geert. 2011. Morpheme structure constraints. In M. van Oostendorp, C. J. Ewen, E. Hume & K. Rice (eds.), *The Blackwell companion to phonology*, 3299–3333. Oxford: Blackwell-Wiley.

Brent, Michael. 1999. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Computational Linguistics* 34.1–3, 71–105.

Chomsky, Noam. 1964. *Current issues in linguistic theory*. The Hague: Mouton.

Chomsky, Noam & Morris Halle. 1965. Some controversial questions in phonological theory. *Journal of Linguistics* 1.2, 97–138.

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

Clark, Alexander. 2001. *Unsupervised language acquisition: Theory and practice*. Ph.D. dissertation, University of Sussex.

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

Dell, François. 1981. On the learnability of optional phonological rules. *Linguistic Inquiry* 12.1, 31–37.

Ellison, Timothy Mark. 1994. *The machine learning of phonological structure*. Ph.D. dissertation, University of Western Australia.

Evers, Vincent, Henning Reetz & Aditi Lahiri. 1998. Crosslinguistic acoustic categorization of sibilants independent of phonological status. *Journal of Phonetics* 26, 345–370.

Ferguson, Charles A. & Munier Chowdhury. 1960. The phonemes of Bengali. *Language* 36.1, 22–59.

Goldsmith, John. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics* 27.2, 153–198.

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

Hale, Mark & Charles Reiss. 2008. *The phonological enterprise*. Oxford & New York: Oxford University Press.

Halle, Morris. 1959. *The sound pattern of Russian*. The Hague: Mouton.

Halle, Morris. 1962. Phonology in generative grammar. *Word* 18.1/2, 54–72.

Halle, Morris. 1978. Knowledge unlearned and untaught: What speakers know about the sounds of their language. In Morris Halle, Joan Bresnan & George A. Miller (eds.), *Linguistic theory and psychological reality*. Cambridge, MA: MIT Press.

Hayes, Bruce. 2004. Phonological acquisition in Optimality Theory: The early stages. In René Kager, Joe Pater & Wim Zonneveld (eds.), *Constraints in phonological acquisition*, 158–203. Cambridge, UK: Cambridge University Press.

Horning, James. 1969. *A study of grammatical inference*. Ph.D. dissertation, Stanford.

Inkelas, Sharon. 1995. The consequences of optimization for underspecification. In Jill Beckman (ed.), *Proceedings of the North East Linguistic Society 25* (NELS 25), 287–302. Amherst, MA: GLSA.

Jarosz, Gaja. 2006. *Rich lexicons and restrictive grammars – Maximum likelihood learning in optimality theory*. Ph.D. dissertation, Johns Hopkins University.

Jarosz, Gaja. 2009. Restrictiveness in phonological grammar and lexicon learning. *Proceedings from the Annual Meeting of the Chicago Lnguistic Society*, vol. 43, 125–139. Chicago: Chicago Linguistic Society.

Kager, René & Joe Pater. 2012. Phonotactics as phonology: Knowledge of a complex restriction in Dutch. *Phonology* 29.1, 81–111.
Katzir, Roni. 2014. A cognitively plausible model for grammar induction. *Journal of Language Modelling* 2.2, 213–248.
Katzir, Roni, Nur Lan & Noa Peled. 2020. A note on the representation and learning of quantificational determiners. *Proceedings of Sinn und Bedeutung 24*, To appear.
Körding, Konrad P. & Daniel M. Wolpert. 2006. Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences* 10.7, 319–326.
de Marcken, Carl. 1996. *Unsupervised language acquisition*. Ph.D. dissertation, MIT.
Orbán, Gergő, József Fiser, Richard N. Aslin & Máté Lengyel. 2008. Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences* 105.7, 2745–2750.
Ozburn, Avery & Alexei Kochetov. 2018. Ejective harmony in Lezgian. *Phonology* 35.3, 407–440.
Pater, Joe. 2014. Canadian raising with language-specific weighted constraints. *Language* 90.1, 230–240.
Pearl, Lisa, Timothy Ho & Zephyr Detrano. 2017. An argument from acquisition: Comparing English metrical stress representations by how learnable they are from child-directed speech. *Language Acquisition* 24.4, 307–342.
Peperkamp, Sharon & Emmanuel Dupoux. 2003. Reinterpreting loanword adaptations: the role of perception. *Proceedings of the 15th international congress of phonetic sciences*, vol. 367, 370. Barcelona: Causal Productions.
Piantadosi, Steven T., Joshua B. Tenenbaum & Noah D. Goodman. 2016. The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological Review* 123.4, 392–424; doi:10.1037/a0039980.
Prince, Alan & Paul Smolensky. 1993. Optimality theory: Constraint interaction in generative grammar. Technical Report Rutgers University, Center for Cognitive Science.
Prince, Alan & Bruce Tesar. 2004. Learning phonotactic distributions. In René Kager, Joe Pater & Wim Zonneveld (eds.), *Constraints in phonological acquisition*, 245–291. Cambridge: Cambridge University Press.
Rasin, Ezer, Iddo Berger, Nur Lan & Roni Katzir. 2018a. Learning phonological optionality and opacity from distributional evidence. In Sherry Hucklebridge & Max Nelson (eds.), *Proceedings of the North East Linguistic Society 48* (NELS 48), 269–282. Amherst, MA: GLSA.
Rasin, Ezer, Iddo Berger, Nur Lan & Roni Katzir. 2018b. Learning rule-based morpho-phonology. Ms., MIT and Tel Aviv University: http://ling.auf.net/lingbuzz/003665.
Rasin, Ezer & Roni Katzir. 2015. Compression-based learning for OT is incompatible with Richness of the Base. In Thuy Bui & Deniz Özyıldız (eds.), *Proceedings of the North East Linguistic Society 45* (NELS 45), vol. 2, 267–274. Amherst, MA: GLSA.
Rasin, Ezer & Roni Katzir. 2016. On evaluation metrics in Optimality Theory. *Linguistic Inquiry* 47.2, 235–282; doi:10.1162/ling_a_00210.
Rissanen, Jorma. 1978. Modeling by shortest data description. *Automatica* 14, 465–471.
Rissanen, Jorma & Eric Sven Ristad. 1994. Language acquisition in the MDL framework. *Language computations: Dimacs workshop on human language, march 20–22, 1992*, 149. Philadelphia: American Mathematical Society.
Rose, Sharon & Rachel Walker. 2004. A typology of consonant agreement as correspondence. *Language* 80, 475–531.
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 & Alison Gopnik. 2004. Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science* 28.3, 303–333.
Stanley, Richard. 1967. Redundancy rules in phonology. *Language* 43.2, 393–436.
Steriade, Donca. 1995. Underspecification and markedness. In John Goldsmith (ed.), *The handbook of phonological theory*, 114–174. Oxford: Blackwell.
Stolcke, Andreas. 1994. *Bayesian learning of probabilistic language models*. Ph.D. dissertation, University of California at Berkeley.
Tesar, Bruce. 2014. *Output-driven phonology*. Cambridge, UK: Cambridge University Press.
Tesar, Bruce & Paul Smolensky. 1998. Learnability in optimality theory. *Linguistic Inquiry* 29.2, 229–268.
Vaysman, Olga. 2002. Against richness of the base: evidence from Nganasan. *Annual Meeting of the Berkeley Linguistics Society*, vol. 281, 327–338. Berkeley, CA: Berkeley Linguistics Society.
Xu, F. & J. B. Tenenbaum. 2007. Word learning as Bayesian inference. *Psychological Review* 114.2, 245–272.

*Authors' addresses:* *(Rasin)*
*Leipzig University, Institut für Linguistik, Beethovenstr. 15, 04107, Germany*
*ezer.rasin@uni-leipzig.de*

*(Katzir)*
*Tel Aviv University, Department of Linguistics and Sagol School of Neuroscience, Tel Aviv 6997801, Israel*
*rkatzir@tauex.tau.ac.il*