A unified approach to several learning challenges in phonology[*]

Ezer Rasin,[1] Itamar Shefi[2] & Roni Katzir[2]

[1]Leipzig University, [2]Tel Aviv University

1. Introduction

Children acquiring the morpho-phonology of their language face multiple non-trivial learning challenges. In this paper we focus on several aspects of phonological learning that we take to be particularly significant: learning from distributional evidence alone, dependencies between phonology and morphology, abstract underlying representations (URs), optionality, and opacity. The remainder of the introduction reviews these aspects of phonological learning and discusses the challenges that they pose. While there is more to learning in phonology than these challenges, we take it that any theory of learning in phonology should at least handle this list. Our goal in this paper is to show how an approach to phonological learning based on the principle of Minimum Description Length (MDL; Rissanen 1978) handles all the learning challenges in a simple, unified way. To our knowledge this is the only learning approach that handles all these learning challenges.

Challenge I: Learning from distributional evidence. We take the primary linguistic data available to the child to consist of surface forms alone, without systematic negative evidence, direct information about underlying representations, or other kinds of assistance. Some support for this view comes from experimental work such as White et al. 2008, which provides evidence for children's ability to acquire phonological alternations from distributional evidence.

Challenge II: Dependencies between morphology and phonology. We assume that children can acquire their phonological knowledge even in the face of nontrivial dependencies between morphological segmentation and phonological processes. Front-back vowel harmony (VH) in Turkish illustrates this challenge. Focusing on stems such as /ip/ 'rope' and /kız/ 'girl' and on the suffixes for the genitive and the plural, the child's input might consist of surface forms such as [ipler], [kızlar], [ipin], and [kızın]. If the child already knows that VH applies within such forms, they can undo it and reason that [ler] and [lar]

[*]We thank Iddo Berger, Adam Jardine, Nur Lan, Giorgio Magri, Charlie O'Hara, Taly Rabinerson, and the audiences at Leipzig University, Tel Aviv University, Uppsala University, and NELS 50.

might be underlyingly identical (and similarly for [in] and [ɪn]). This, in turn, can guide the child toward the correct morphological segmentation of the forms:

(1)

	'rope'	'girl'
Plural	ip-ler	kɪz-lar
Genitive	ip-in	kɪz-ɪn

Similarly, if the child already knows the morphological decomposition of these forms, they can reason about the relation of [ler] and [lar] (and similarly for [in] and [ɪn]), which can guide the child toward a discovery of VH. However, if the child does not yet know either about the process of vowel harmony or about the morphological decomposition of the surface forms, they will face the challenge of discovering both despite the bidirectional dependencies between the two: without knowledge of VH, morphological segmentation seems challenging, and without knowledge of morphological segmentation, VH (which applies regularly across morpheme boundaries but not morpheme-internally) will be harder to discover.

Challenge III: Abstract URs. Abstract URs are URs that differ from their surface forms despite insufficient evidence for the discrepancy from alternations. The extent to which URs may be abstract was a matter of much debate in early generative phonology. More recently, abstractness has been argued for by Alderete and Tesar (2002), McCarthy (2005), and Nevins and Vaux (2007), among others (see also discussion in Krämer 2012). Here we will assume that abstractness is possible, illustrating with a schematic example from Alderete and Tesar (2002), based on the interaction of stress and epenthesis in Yimas. In this example, stress in bisyllabic words is initial but can be pen-initial if the first vowel is [i], as in the following table:

(2)

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

A familiar kind of analysis would posit a pattern of initial stress, where an unstressed initial [i] is always epenthetic:

(3) /pkut/ → |pkút| → [pikút]

According to Alderete and Tesar (2002), however, this generalization is acquired without support from alternations.

Challenge IV: Optionality. The acquired phonological grammar should capture speakers' knowledge not just in simple cases but also in more complex patterns, such as optionality. An example of optionality is the process of liquid deletion in French, analyzed by Dell (1981), which allows a word-final liquid to optionally delete in certain environments (as in [tabl]∼[tab] for 'table'). Optionality, as in Dell's example, poses various difficulties to the learner. One difficulty is that optionality obscures the evidence for the application of a phonological process. For example, the presence of surface forms such as [tabl] can make it hard for the learner to conclude that final liquids may delete. Another challenge concerns the environment of application: even if the learner concludes that final liquids delete, what will tell them that this process can only apply after an obstruent? See Dell 1981 and Rasin et al. 2018 for further discussion of this challenge.

Challenge V: Opacity. An example of opacity is the counterfeeding interaction between nasal deletion and cluster simplification in Catalan (Mascaró 1976), which we present here in a simplified form. As the following illustrates, word-final nasals delete in certain contexts in Catalan, as do post-nasal word-final stops, but while the latter process creates an appropriate environment for the former, cluster simplification does not lead to nasal deletion:

(4)	kuzí ∼ kuzín̲-s	‘cousin.SG ∼ cousin.PL’
	kəlén ∼ kəlént̲-ə	‘hot.MASC ∼ hot.FEM’

As with optionality, opaque interactions pose a learning challenge by obscuring the form of a phonological process and its environment of application. In the case of Catalan, the learning challenge comes from surface forms such as [kəlén] that can confuse a naive attempt to learn that word-final nasals delete.

The remainder of the paper is structured as follows. First, in section 2, we introduce the MDL principle and present a concrete learning algorithm that uses it as an evaluation metric. Then, in section 3, we present simulation results illustrating the ability of the MDL learner to address the learning challenges described above in a simple, unified way. In section 4 we discuss previous proposals from the literature and note that they have not been able to go as far in addressing these learning challenges. Section 5 concludes.

2. An MDL learner

2.1 The MDL principle

MDL is an evaluation criterion that balances two competing factors. One factor is the simplicity of the grammar. We use $|G|$ for the length of the grammar G in bits. Minimizing $|G|$, as in the evaluation metric of SPE (Chomsky and Halle 1968), prefers simple but often overly general grammars. The second factor that MDL balances is the tightness of fit of the grammar G to the data D. We use $D:G$ for the shortest encoding of D using G, and we use $|D:G|$ for the length of $D:G$ in bits. Minimizing $|D:G|$, roughly as in the proposals of Dell (1981) and Wexler and Manzini (1987), leads to grammars that fit the data well but are often overly specific. By balancing $|G|$ and $|D:G|$ against each other, MDL attempts to find a reasonably simple grammar that fits the data reasonably well. The criterion can be stated as follows:

(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'

Before discussing our concrete MDL learner, which will assume rule-based phonology, we will first explain why MDL supports the induction of a segmented lexicon and phonological rules from distributional evidence using a toy example.[1] Consider first segmentation. The MDL metric in (5) allows the learner to discover the segmentation of words into stems and affixes (see de Marcken 1996, Goldsmith 2006). If the surface forms are generated from, e.g., 8 different stems (e.g., /dok/, /kab/, etc.) and 4 different suffixes (e.g., /za/, /ti/, etc.), a naive grammar G_1 for the language will store all the different $8 \times 4 = 32$ surface forms in the lexicon (each of which costing its length times $\lg|\Sigma|$ bits, where Σ is the alphabet in which the lexicon is written and $|\Sigma|$ its cardinality). Specifying any given surface form (for the purposes of constructing $D : G_1$) will require $\lg 32 = 5$ bits. A less naive grammar, call it G_2, will store the stems and the suffixes separately. This will amount to just $8+4 = 12$ different entries (which, in addition, are shorter than those in the lexicon of $|G_1|$). Specifying any given surface form will require choosing a stem and a suffix. The former choice, from among 8 possibilities, costs $\lg 8 = 3$ bits, and the latter, from among 4 possibilities, costs $\lg 4 = 2$ bits. In total, then, specifying a surface form using G_2 will require 5 bits: the same as with G_1. Consequently, $|D : G_1| = |D : G_2|$. Since $|G_2| < |G_1|$, (5) will lead the learner to prefer G_2 to G_1, which seems intuitively correct.

MDL also enables the learner to acquire phonological processes (e.g., Goldwater and Johnson 2004, and Rasin et al. 2018). If the language just discussed also has a process of regressive voicing assimilation across morphemes, the surface forms will seem to involve twice the actual number of stems (e.g., [dog] before [za] and [dok] before [ti]). Using (5), the learner will reject a naive encoding of this kind (since the storage of two versions of each stem is costly) in favor of one where there is just one variant for each stem, along with a rule of voicing assimilation (since the savings obtained by storing just one form for each stem outweigh the costs of adding the relevant phonological rule).

In section 2.2 we will present a concrete MDL learner, and in section 3 we will present simulations for each of the challenges discussed in the introduction illustrating the ability of an implemented MDL learner to address these challenges in practice.

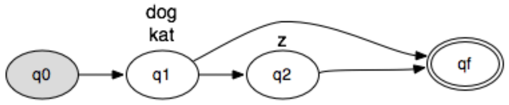
2.2 A concrete MDL learner

In this section we present the rule-based MDL learning algorithm we use for the simulations in this paper.[2] The algorithm, which was proposed by Rasin et al. (2018), has the following properties. First, the learner is assumed to be equipped with a hypothesis space consisting of grammars with a lexicon and a list of ordered rewrite rules. The lexicon is represented as a Hidden Markov Model, as in (6), and the rules are of the form given in (7).

[1] For further discussion of MDL – and the closely related Bayesian approach to learning – in the context of various grammar induction tasks, see Horning 1969, Berwick 1982, Rissanen and Ristad 1994, Stolcke 1994, de Marcken 1996, Brent 1999, and Clark 2001, among others. For discussion of possible motivations for MDL see Katzir 2014, Katzir, Lan, and Peled 2020, and Rasin and Katzir 2020.

[2] The code for the learner is available at https://github.com/taucompling/morphophonology_spe.

(6) A lexicon (represented as a Hidden Markov Model)



(7) Ordered rules
$A \rightarrow B/C\_\_D$ (optional?)

The MDL metric assigns the score $|G| + |D:G|$ to each grammar G in the hypothesis space. To measure the size of the grammar, $|G|$, we sum the size of the lexicon and the size of the list of rules. To see how rules are measured, consider the vowel harmony rule in (8). The rule, given first in textbook notation (8a), is converted into a string (8b). Then, each symbol in the string receives a cost of $\lg n$, where n is the number of symbols that can be used in writing rules. To measure the size of the lexicon, the HMM is converted into a string in a similar fashion, as in (9). Here, the cost of each phonological segment in lexical items is calculated relative to a language-specific alphabet, which is measured in addition to the HMM itself. To measure $|D:G|$, we calculate the amount of information needed to describe the derivation of every surface form given the grammar, as detailed in Rasin et al. 2018, which also discusses the details of the search for the optimal grammar using a genetic algorithm.

(8) Measuring a rule (example: vowel harmony)

a. $\begin{bmatrix}-cons\end{bmatrix} \rightarrow \begin{bmatrix}-back\end{bmatrix} / \_\_ \begin{bmatrix}+cons\end{bmatrix}^* \begin{bmatrix}-cons \\ -back\end{bmatrix}$ (optional)

b. $-cons\#_{rc} - back\#_{rc}\#_{rc} + cons * \#_{b} - cons\#_{f} - back\#_{rc}1\#_{rc}$

(9) Measuring a lexicon (example: string representation of the HMM in (6)):

$q_0 q_1 \#_S \#_w \#_w q_1 q_2 q_f \#_Sdog\#_wkat\#_w \#_w q_2 q_f \#_Sz\#_w \#_w$

3. Simulations

3.1 Turkish VH

The dataset for this simulation was modeled after front-back VH in Turkish and illustrates the challenge of learning in the face of dependencies between morphology and phonology.[3]

[3] VH illustrates a further learning challenge: that of learning unbounded dependencies. Specifically, VH often applies across several intervening consonants, thus posing a problem for phonological learners that are limited to small, local contexts of fixed size. Unsupervised learners that can capture long-distance dependencies, such as Hayes and Wilson 2008 or Heinz 2010, are phonotactic learners that are not yet integrated within a full morpho-phonological learner. Moreover, as noted by Hayes and Wilson (2008, p. 402), languages with many disharmonic roots like Turkish (e.g., backness mismatch in m$\underline{e}$z$\underline{a}$r(-lar), h$\underline{o}$t$\underline{e}$l(-ler)) pose a problem for attempts to acquire VH using a phonotactic learner. The MDL learner presented here does not assume an upper bound on the size of contexts of application of phonological rules and is therefore able to acquire VH

The learner's task was to segment the data into morphemes and acquire both a lexicon of URs and a rule of VH. The data were generated by taking all combinations of 23 Turkish nouns (e.g., kent, jɪl) and 10 suffixes (e.g., -ler/-lar, -in/-ɪn) and randomly removing 10 words from the result, a total of 220 words. Then, a rule of progressive front-back VH was applied to the words. The words were presented to the learner as unsegmented strings, without any morphology (e.g., [kentler], [jɪllar]).

The learner converged on a hypothesis with a VH rule that applies in all appropriate places and a lexicon in which each pair (e.g., -ler/-lar) is represented with a single UR:

(10) Final grammar:

•Lexicon:

–Stems = {kent, jɪl, güz, tuz ...};

–Suffixes = {ler, in, ten, siz, ...}

•Rules: $\left[+syll\right] \rightarrow \left[+back\right] / \begin{bmatrix}+back \\ +cont\end{bmatrix} \left[-syll\right]^{*}$ __ (obligatory)

•Description length: $|G| + |D{:}G| = 872.03 + 17,260.08 = 18,132.11$

To illustrate how MDL leads to the learner's success, we can compare the final grammar to alternative grammars in the hypothesis space in terms of description length. Consider, for example, a naive hypothesis that simply memorizes the data, has no phonological rules, and performs no segmentation:

(11) Naive grammar:

•Lexicon: {kentler, tuzlar, ...}

•Rules: ∅

•Description length: $|G| + |D{:}G| = 6,350.62 + 17,118.99 = 23,469.62$

In this naive hypothesis, the lexicon is highly complex, as it stores every surface form without segmentation. The size of its grammar, $|G|$, is thus significantly larger than the size of the final grammar (6,350.62 versus 872.03). Even though the native hypothesis is missing a rule, this minor saving is offset by the dramatic increase to the size of the lexicon. In terms of $|D{:}G|$, the naive hypothesis is slightly better (17,118.99 versus 17,260.08), as it can generate nothing more than the observed data (the final grammar can additionally generate the 10 morpheme combinations that have been randomly removed). Overall, the size of its lexicon makes the naive hypothesis lose decisively to the final grammar in terms of description length (23,469.62 versus 18,132.11).

Another hypothesis to consider is the hypothesis that segments each word into morphemes but does not learn the VH rule, and thus stores two variants of each suffix, a back variant and a front variant:

despite its non-locality. It can also, in principle, deal with disharmonic roots (though, in this simulation, our dataset did not include such roots; see below other simulations where non-surface-true generalizations are learned).

(12) Grammar with segmentation but without VH:

- •Lexicon:
  - –Stems = {kent, jıl, güz, tuz ...};
  - –Suffixes = {ler, lar, in, ɨn, ...}
- •Rules: ∅
- •Description length: $|G| + |D{:}G| = 938.07 + 19{,}297.28 = 20{,}235.35$

Here, both $|G|$ and $|D{:}G|$ are worse than in the final hypothesis. $|G|$ is worse (938.07 versus 872.03) because of the duplication of each suffix in the lexicon, which could have been avoided if the VH rule had been learned. $|D{:}G|$ is worse (19,297.28 versus 17,260.08) because this hypothesis can generate any combination of stems and suffixes, including combinations with backness mismaches (e.g., kentlar), and this over-generation translates into a longer description of the observed data. Since this hypothesis loses to the final grammar on each of $|G|$ and $|D{:}G|$, it also loses overall.

3.2 Abstract URs

The next simulation was designed to test the learner on the problem of abstract URs. Here we did not use the stress-epenthesis interaction described above but rather tested the learner on a simplified pattern of complementary distribution between aspirated and unaspirated stops, modeled after aspiration in English. In this simplified pattern, aspiration (represented as a separate segment [h] rather than a feature) occurs between a voiceless stop and a following vowel, but never elsewhere. This pattern by no means requires an analysis in terms of abstract URs, but a grammar with abstract URs is nevertheless what MDL will lead the child to: instances of aspiration, which can be predicted by a simple rule, will be removed from the lexicon even in the absence of direct evidence from alternations because removing information from the lexicon makes it simpler to describe. Note that the logic in this case is similar to the case of the stress-epenthesis interaction discussed above, where abstract URs could be reached by removing predictable vowels from the lexicon and inserting them through a rule of epenthesis. The aspiration pattern is thus a simple example that illustrates the ability of MDL to deal with abstract URs more generally.

The dataset for this simulation consisted of 10 words, given in (13).

(13) ithik, thatkha, thatthat, ikhak, khikthak, khathiit, ithak, thikhiat, thakhiit, khikhi

The result was the desired grammar in which aspiration is completely absent from the lexicon and is inserted by an aspiration rule. In addition, aspiration was removed from the alphabet in which the lexicon is written, meaning that aspiration cannot be used in URs at all. Given the prohibition against underlying aspiration, aspiration can only be found on the surface if it is derived from the aspiration rule, but it cannot surface elsewhere.

(14) Final grammar:

- Lexicon (no /h/): {ikak, itak, itik, katiit, kiki, kiktak, takiit, tatka, tattat, tikiat}
- Rules: $\emptyset \rightarrow \left[+asp\right] / \left[+stop\right] \,\_\_\, \left[-cons\right]$
- Description length: $|G| + |D{:}G| = 220.4 + 3,321.93 = 3,542.33$

3.3 French optionality

The next simulation was designed to test the learner on optionality in French, where liquids are optionally deleted word-finally after an obstruent. The dataset consisted of 91 words, a sample of which is given below. Note that the learner is not provided with information about lexical relatedness of surface forms such as [tabl] and [tab] and needs to discover such information as part of the learning task.

(15) tab, tabl, final, aktif, parl, lub, lubl, tap, tapl, ...

The result is the grammar in (16). The learner correctly collapsed pairs of surface forms into a single UR and learned the restricted optional liquid deletion rule.

(16) Final grammar:

- Lexicon: {tabl, final, aktif, parl, lubl, tapl, ...}
- Rules: $[+liquid] \rightarrow \emptyset / [-son] \,\_\_$ (optional)
- Description length: $|G| + |D{:}G| = 31,936.09 + 30,153.81 = 62,089.91$

To see why MDL succeeds on restricted optionality, it would be useful to compare the final grammar to an alternative where the optional liquid-deletion rule does not have a left context (this alternative grammar is otherwise identical to the final grammar):

(17) Alternative grammar:

- Lexicon: {tabl, final, aktif, parl, lubl, tapl, ...}
- Rules: $[+liquid] \rightarrow \emptyset / \_\_$ (optional)
- Description length: $|G| + |D{:}G| = 31,928.47 + 33,003.81 = 64,932.29$

In this alternative grammar, $|G|$ is slightly simpler (31,928.47 versus 31,936.09) because of the omission of the left context of the rule. However, $|D{:}G|$ is significantly larger (33,003.81 versus 30,153.81): since the context-free rule can optionally apply to any underlying liquid, the grammar can generate many additional surface forms with missing liquids (e.g., from the UR /lubl/, the grammar generates the four surface forms [lubl], [ubl], [lub], and [ub]). This over-generation translates into a longer $|D{:}G|$, which causes the alternative grammar to lose based on considerations of restrictiveness.

3.4 Counterfeeding opacity in Catalan

Our next dataset was designed to test the learner on the problem of counterfeeding opacity. We used two rules modeled after a simplified version of final-nasal deletion and cluster simplification in Catalan. We generated 65 words by creating all combinations of 13 stems and 5 suffixes (taken from a Catalan dictionary) and applying final-nasal deletion and cluster simplification, in this order (18). A sample of the data is given in (19).

(18) Rules:

a. Delete a nasal word-finally.
b. Delete a word-final stop following a nasal.

(19)

stem\suffix	$\emptyset$	-s	-et	$\cdots$
kalent	kalen	kalents	kalentet	
kuzin	kuzi	kuzins	kuzinet	
$\ldots$				

The learner converged on a segmented lexicon and on the two rules – final-nasal deletion and cluster simplification – and their correct ordering, as in (20).

(20) Final grammar:

•Lexicon:

–Stems = $\{$kuzin, kalent, blank, kasa, $\ldots\}$
–Suffixes = $\{$s, et, ik, a, $\ldots\}$

•Rules:

–$[+nasal] \to \emptyset / \_\_\#$
–$[-cont] \to \emptyset / [-nasal]\_\_\#$

•Description length: $|G| + |D{:}G| = 562.2 + 3,914.54 = 4,476.74$

4. Previous work on learning in phonology

We presented a learner that uses the MDL evaluation metric, which minimizes $|G| + |D{:}G|$, to jointly learn morphology and phonology within a rule-based framework. This learner is fully distributional, working from unanalyzed surface forms alone – without access to paradigms or negative evidence – to obtain the URs in the lexicon, the possible morphological combinations, and the ordered phonological rules. It acquires both allophonic rules and alternations and handles both optionality and rule interaction, including opacity.

In this section we review prominent proposals from past work on learning in phonology and show that they have not gone as far in terms of addressing these challenges: they either do not work with what we take to be the primary linguistic data (e.g., they assume that the child is given information about URs), or they do not acquire an important part of the phonological grammar (e.g., they cannot deal with opacity). We will focus on five

Theory ↓	Distributional evidence	Simultaneous segmentation	Opacity	Optionality	Abstract URs
1) Constraint reranking	✘	✘	?	✓	✘
2) Reranking + Free Ride	✘	✘	?	✓	✘
3) MaxEnt + OT	✘	✘	✓	?	?
4) Dist. alt. learner	✓	✘	✘	✘	✘
5) MaxLikelihood + OT	* (see discussion below)				
6) Lexicon Entropy	* (see discussion below)				

Figure 1: Some prominent proposals from past work on learning in phonology and their ability to address five learning challenges.

components of the learning challenge: learning from distributional evidence alone, learning segmentation simultaneously with phonology, learning opacity, learning optionality, and learning abstract URs. Each of the learners we discuss fails on at least one of those components, as summarized in Figure 1 (and as discussed in the rest of this section).

We first consider constraint reranking algorithms (row 1 in Figure 1), a family of learning algorithms for OT that include the proposals by Tesar (1995, 2014), Tesar and Smolensky (1998), Boersma and Hayes (2001), Prince and Tesar (2004), and much related work. These proposals assume that URs are given to the learner in advance or that the learner is exposed to surface forms already segmented into morphemes, along with the information of which surface morphemes come from the same UR. Therefore, these works fail on the challenge of learning from distributional evidence and the challenge of learning segmentation simultaneously with the phonology.

Another shortcoming of the constraint-reranking proposals just mentioned is that they assume that, in the absence of direct evidence from alternations, URs are identical to their corresponding surface forms, Hence, they do not address the challenge of learning abstract URs. An attempt to address this problem was made by McCarthy (2005), who proposed to extend constraint reranking algorithms with the Free Ride Principle, a learning principle that aims to deal with some cases of abstract URs (row 2 in Figure 1). This principle allows using information from alternations to infer non-identical URs for non-alternating forms. While addressing some cases of abstract-UR learning, McCarthy's algorithm does not offer constraint reranking algorithms a handle on cases of abstract URs where there is no supporting evidence from alternations at all, as in Alderete and Tesar (2002)'s stress-epenthesis example. See Rasin and Katzir 2018 for further discussion.

Another family of learners in the OT literature are the so-called MaxEnt learners (Goldwater and Johnson 2004, Nazarov and Pater 2017, and O'Hara 2017, among others), which rely on the principle of Maximum Entropy as an evaluation metric (row 3 in Figure 1). These learners receive morphologically-segmented surface forms, as well as information about which surface morphemes come from the same UR. Hence, like constraint reranking algorithms, they do not address the challenges of learning from distributional evidence alone and learning segmentation simultaneously with the phonology.

Similarly to the present proposal, the distributional alternation learner of Calamaro and Jarosz (2015) learns phonological rules in a fully distributional way (row 4 in Figure 1).

The proposal extends the allophonic learner of Peperkamp et al. (2006), which detects maximally dissimilar surface contexts as hints for allophonic distribution. Calamaro and Jarosz (2015) extend Peperkamp et al.'s model to account for neutralization rules which lead to the occurrence of alternating segments in overlapping contexts. To account for this challenge, Calamaro and Jarosz compute surface dissimilarity scores that are contextual (for a given context $X\_\_Y$ and two potential alternants A and B, they compute a dissimilarity score for the triple $< X\_\_Y, A, B >$ by comparing the probability of the context $X\_\_Y$ given A and given B). Calamaro and Jarosz's model faces two challenges that seem hard to address within the framework of distribution comparison that they adopt. First, their model does not handle rule orderings. This gap is particularly difficult to bridge in the case of opaque rule interactions, where surface distributions obscure the correct context for rule application. The second challenge concerns optionality. When a rule is optional, the distribution of A and B can be similar in all contexts, so a dissimilarity detector will fail to identify the rule.

Other learners close to our goals include Jarosz (2006)'s Maximum Likelihood OT learner and Riggle (2006)'s Lexicon Entropy OT learner (rows 5 and 6 in Figure 1). Both learners rely on evaluation metrics rather than on a procedural approach to acquire an OT ranking and URs. Differently from MDL, however, these evaluation metrics do not balance economy and restrictiveness and thus lead to over-generalization and under-generalization problems. These problems for Maximum Likelihood and Lexicon Entropy have been discussed in detail in Rasin and Katzir 2016.

Of the other learners proposed in the literature, our learner is closest to the rule-based learners proposed by Goldwater and Johnson (2004), Goldsmith (2006), Naradowsky and Goldwater (2009), and the OT learner by Rasin and Katzir (2016), all of which are fully distributional phonological learners that rely on the same kind of balanced evaluation metric as the present paper.[4] At present, the three rule-based learners can acquire rules only at morpheme boundaries and generalize only with respect to $X\_\_Y$ and not with respect to A and B.[5] They are also aimed at obligatory rules and do not handle rule interaction. The constraint-based learner has not yet been shown to acquire opacity. One way of interpreting our simulations above is as showing that the limitations of all these balanced distributional learners are not essential within this framework and that MDL can support the acquisition of allophony, generalizations over both the context and the change (in the case of rule-based phonology), optionality, and opacity.

5. Discussion

We presented an MDL-based learner for the unsupervised learning of rule-based morphophonological grammars. The generality of the MDL metric has allowed the learner to si-

[4] The learner of Naradowsky and Goldwater (2009) targets orthographic rules rather than phonology, but the difference is immaterial.

Other balanced learners proposed in the literature, which are not fully distributional, include those of Cotterell et al. (2015) and Ellis and O'Donnell (2017).

[5] By limiting the kinds of rule that can be learned, these learners are similar to the procedural rule-based learners of Johnson (1984) and Albright and Hayes (2002).

multaneously perform morphological segmentation and acquire complete grammars, including URs and ordered rules, and including transparent and opaque rule interactions, as well as optional rules. Previous proposals have not gone as far because they either do not work with what we take to be the primary linguistic data or do not acquire important parts of a descriptively-adequate grammar. In particular, by learning from distributional evidence alone, the learner differs from many proposals in the literature on phonological learning which assume that the learner is given paradigmatic information, information about URs, or even the URs themselves. The ability of our learner to acquire opaque rule interactions and optional rules distinguishes it from other learners that are limited to transparent process interactions or deterministic processes. To our knowledge, this makes the present learner the first distributional learner that can acquire a comprehensive morpho-phonological system with the structure proposed in the phonological literature.

References

Albright, Adam, and Bruce Hayes. 2002. Modeling English past tense intuitions with minimal generalization. In *Proceedings of the ACL-02 workshop on Morphological and phonological learning*, volume 6, 58–69. Association for Computational Linguistics.

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.

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

Boersma, Paul, and Bruce Hayes. 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32:45–86.

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

Calamaro, Shira, and Gaja Jarosz. 2015. Learning general phonological rules from distributional information: A computational model. *Cognitive Science* 39:647–666.

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

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

Cotterell, Ryan, Nanyun Peng, and Jason Eisner. 2015. Modeling word forms using latent underlying morphs and phonology. *Transactions of the Association for Computational Linguistics* 3:433–447.

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

Ellis, Kevin, and Timothy O'Donnell. 2017. Inducing phonological rules: Perspectives from Bayesian program learning. Presented at the MIT Workshop on Simplicity in Grammar Learning.

Goldsmith, John. 2006. An algorithm for the unsupervised learning of morphology. *Natural Language Engineering* 12:1–19.

Goldwater, Sharon, and Mark Johnson. 2004. Priors in Bayesian learning of phonological rules. In *7th Annual Meeting of the ACL Special Interest Group on Computational Phonology*, 35–42.

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

Heinz, Jeffrey. 2010. Learning long-distance phonotactics. *Linguistic Inquiry* 41:623–661.

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

Jarosz, Gaja. 2006. Rich lexicons and restrictive grammars – Maximum Likelihood learning in Optimality Theory. Doctoral dissertation, Johns Hopkins University, Baltimore, MD.

Johnson, Mark. 1984. A discovery procedure for certain phonological rules. In *Proceedings of 10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics*, 344–347.

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

Katzir, Roni, Nur Lan, and Noa Peled. 2020. A note on the representation and learning of quantificational determiners. In *To appear in Proceedings of Sinn und Bedeutung 24*.

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

de Marcken, Carl. 1996. Unsupervised language acquisition. Doctoral dissertation, MIT.

Mascaró, Joan. 1976. Catalan phonology and the phonological cycle. Doctoral dissertation, MIT.

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

Naradowsky, Jason, and Sharon Goldwater. 2009. Improving morphology induction by learning spelling rules. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, 1531–1536.

Nazarov, Aleksei, and Joe Pater. 2017. Learning opacity in stratal maximum entropy grammar. *Phonology* 34:299–324.

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'Hara, Charlie. 2017. How abstract is more abstract? learning abstract underlying representations. *Phonology* 34:325–345.

Peperkamp, Sharon, Rozenn Le Calvez, Jean-Pierre Nadal, and Emmanuel Dupoux. 2006. The acquisition of allophonic rules: Statistical learning with linguistic constraints. *Cognition* 101:B31–B41.

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. 2018. Learning phonological optionality and opacity from distributional evidence. In *Proceedings of NELS 48*, ed. Sherry Hucklebridge and Max Nelson, 269–282.

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

Rasin, Ezer, and Roni Katzir. 2018. Learning abstract underlying representations from distributional evidence. In *Proceedings of NELS 48*, ed. Sherry Hucklebridge and Max Nelson, 283–290.

Rasin, Ezer, and Roni Katzir. 2020. A conditional learnability argument for constraints on underlying representations. To appear in *Journal of Linguistics*.

Riggle, Jason. 2006. Using entropy to learn OT grammars from surface forms alone. In *Proceedings of the 25th West Coast Conference on Formal Linguistics*, 346–353.

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.

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

Tesar, Bruce. 1995. Computational optimality theory. Doctoral dissertation, University of Colorado.

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

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

Wexler, Kenneth, and Rita M. Manzini. 1987. Parameters and learnability in binding theory. In *Parameter setting*, ed. Thomas Roeper and Edwin Williams, 41–76. Dordrecht, The Netherlands: D. Reidel Publishing Company.

White, Katherine S., Sharon Peperkamp, Cecilia Kirk, and James L. Morgan. 2008. Rapid acquisition of phonological alternations by infants. *Cognition* 107:238–265.

Ezer Rasin, Itamar Shefi, Roni Katzir
ezer.rasin@uni-leipzig.de, itamarshefi@mail.tau.ac.il, rkatzir@tauex.tau.ac.il