

Approaching explanatory adequacy in phonology using Minimum Description Length

Ezer Rasin, Iddo Berger, Nur Lan, Itamar Shefi, and Roni Katzir
Tel Aviv University

ABSTRACT

A linguistic theory reaches explanatory adequacy if it arrives at a linguistically-appropriate grammar based on the kind of input available to children. In phonology, we assume that children can succeed even when the input consists of surface evidence alone, with no corrections or explicit paradigmatic information – that is, in learning from *distributional evidence*. We take the grammar to include both a lexicon of underlying representations and a mapping from the lexicon to surface forms. Moreover, this mapping should be able to express optionality and opacity, among other textbook patterns. This learning challenge has not yet been addressed in the literature. We argue that the principle of Minimum Description Length (MDL) offers the right kind of guidance to the learner – favoring generalizations that are neither overly general nor overly specific – and can help the learner overcome the learning challenge. We illustrate with an implemented MDL learner that succeeds in learning various linguistically-relevant patterns from small corpora.

Keywords:
learning,
phonology,
opacity,
optionality,
Minimum
Description Length

As part of language acquisition, the child needs to acquire many different aspects of the morpho-phonology of their language. If the child is learning English, for example, they will need to learn that in ‘cats’, pronounced [k^hæts], the aspiration of the initial [k] and the voicelessness of the final [s] are no accident: in English, voiceless stops such as [k] are always aspirated in this position (roughly, syllable-initially in a stressed syllable), and the expression of the plural morpheme is always the voiceless [s] after a voiceless stop such as [t]. Thus, the child will need to learn that imaginable forms such as [kæts] or [k^hætz] are not possible in the language. These pieces of knowledge come from a very large – possibly unbounded – set of possible choices that languages can make and that children must be able to acquire. Moreover, children are capable of acquiring at least some linguistic knowledge of this kind from distributional cues alone, without access to analyzed forms or paradigms and without negative evidence. The result is a nontrivial learning task that is challenging even in relatively simple cases such as deterministic, surface-true phonotactics (as in the aspiration pattern of English) or alternations providing useful information (such as the voicing pattern concerning the /z/ suffix in English). The learning challenge is even more pronounced in cases of optional phonological processes and of opaque interactions of phonological processes. A theory that addresses this challenge can be said to have reached explanatory adequacy (Chomsky 1965). To date, no general solution to this challenge has been provided in the literature.

In this paper, we propose a response to the learning challenge in terms of a certain kind of simplicity metric. The simplicity metric will follow the principle of Minimum Description Length (MDL; Rissanen 1978), which incorporates both the idea of grammar simplicity (as in the evaluation metric of early generative phonology) and that of restrictiveness (or how easy it is for the grammar to capture the data). The representational framework that we use for our discussion will be that of rule-based phonology, which offers a particularly direct handle on the representation of both optionality and opacity. We wish to emphasize, however, that our focus in this paper is the learning approach – namely, the MDL metric – and how it guides the learner given a rep-

representational framework rather than the representational framework itself. In order to illustrate how the MDL metric can guide the learner toward appropriate hypotheses, we present several simulations that start with a small corpus of unanalyzed surface forms – generated from artificial grammars based on morpho-phonological patterns in various languages – and arrive at a full grammar including a lexicon of underlying representations (URs), a morphological segmentation of forms into morphemes and their attachment possibilities, and different kinds of phonological rules (both obligatory and optional) and their ordering (including both transparent and opaque interactions). While it might seem that these different aspects of morpho-phonological knowledge call for a fragmented learning approach, with specialized learners for the different sub-tasks, we will show how the MDL evaluation metric allows all of them to be acquired in a unified way.

We start, in Section 2, by reviewing the challenge of explanatory adequacy in phonology. In Section 3, we present the MDL metric in the context of rule-based phonology and specify a concrete set of representations for phonological grammars and their MDL costs. In Section 4, we present proof-of-concept learning simulations with optionality, rule interaction (including opacity), and interdependent phonology and morphology. Section 5 discusses previous work on learning in phonology and its relation to the goals of this paper. Section 6 concludes the paper.

EXPLANATORY ADEQUACY IN PHONOLOGY

2

An explanatorily adequate linguistic theory accounts for how the child arrives at a descriptively-adequate grammar based on the primary linguistic data (Chomsky 1965, pp. 25–27). The present paper focuses on this learning challenge in phonology. In Section 3, we argue that combining a suitable theory of phonological representations with the general principle of MDL goes beyond all other proposals in the literature in terms of approaching the goal of explanatory adequacy. Before that,

in the present section, we briefly outline certain aspects of explanatory adequacy in phonology that will be important for evaluating our claim below.

First, we follow Calamaro and Jarosz (2015) in assuming that children can acquire significant aspects of phonological knowledge from distributional evidence alone (that is, from surface forms alone, without systematic negative evidence, direct information about underlying representations, or other kinds of assistance). To be sure, children are also exposed to a great deal of other information, including contextual cues as to the meanings of words. Calamaro and Jarosz's (2015) assumption, which we adopt here, is simply that children can succeed in phonological learning even when such additional information is not present. Some support for this view comes from experimental work that provides evidence for children's ability to acquire key aspects of morpho-phonology, including segmentation (Saffran *et al.* 1996), allomorphy (Gerken *et al.* 2005), and phonological alternations (White *et al.* 2008), all from distributional evidence. We note, in addition, that non-distributional information such as morpheme meanings is more limited in its ability to assist phonological learning than is often assumed in the phonological learning literature. A common assumption made in the literature is that semantic information can teach the learner about the existence of phonological processes. On this common view, when the learner encounters two morphemes with different phonological surface forms that have exactly the same meaning, the learner knows that a phonological process is responsible for the surface difference between them. Semantics is therefore assumed to take the learner a long way towards learning the phonological grammar. We believe that this view overestimates the utility of semantics for the learner because it mistakenly ignores the possibility that two morphemes with the same meaning are not related through phonological processes: namely, it ignores the possibility of suppletion, where two semantically identical forms are stored separately in the lexicon, without being derived from a common lexical entry through any phonological process. Since nobody tells the learner when suppletion is involved, the learner has to figure out the existence of phonological processes itself. We assume that an explanatorily adequate theory needs to account for this aspect of learning as well. However, a more complete characterization of the evidence that children base

their learning on, both in lab settings and during acquisition, awaits further work.

Second, we assume that children can acquire their phonological knowledge even in the face of nontrivial dependencies between morphological segmentation and phonological processes, and we assume that underlying representations may be abstract, in the sense of differing from surface forms even in the absence of conclusive evidence from alternations. Moreover, we take the phonological knowledge that children attain to involve various textbook properties such as opacity and optionality. We discuss each of these aspects of phonological knowledge and learning in turn.

Dependencies between morphological segmentation and phonological processes exist in many affixes and alternations across languages. Vowel harmony in Turkish provides a particularly clear illustration. 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 vowel harmony applies within such forms, they can undo it and reason that *ler* and *lar* might be underlyingly identical (and similarly for *in* and *ın*). This, in turn can guide the child toward the correct morphological segmentation of the forms:

		‘rope’	‘girl’
(1)	Plural	<i>ip-ler</i>	<i>kız-lar</i>
	Genitive	<i>ip-in</i>	<i>kız-ın</i>

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 *ın*), which can guide the child toward a discovery of vowel harmony. 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.

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, conservatively, that abstractness is possible, illustrating with a schematic example, based on an example from Alderete and Tesar (2002), which was in turn modeled after the interaction of stress and epenthesis in Yimas. In this example, stress in bisyllabic words is generally initial, but there are some words, in all of which the first vowel is [i], where stress falls on the second syllable. The following table, showing three possible (and different) words and one impossible form, illustrates:

	Initial vowel = i	Initial vowel = a
(2) 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.

Finally, the acquired phonological knowledge should capture speakers' intuitions not just in simple cases but also in more complex patterns, of which we focus here on two: optionality and opacity. An example of optionality is the process of liquid deletion in French, analyzed by Dell (1981) and discussed in some detail below, which allows a word-final liquid to optionally delete in certain environments (as in [tabl]~[tab] for 'table'). An example of opacity is the counter-feeding interaction between nasal deletion and cluster simplification in Catalan (Mascaró 1976). As the following illustrates, word-final nasals sometimes delete 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 ~ kuzín-s 'cousin.SG ~ cousin.PL'
 kəlén ~ kəlént-ə 'hot.MASC ~ hot.FEM'

To summarize, we take the following to be requirements of any theory that achieves explanatory adequacy in the domain of phonology. It should allow for learning from distributional evidence alone. It

should support the joint learning of morphological segmentation and phonological processes and the learning of abstract URs. And it should handle complex patterns such as optionality and opacity. To be sure, this is just a starting point; we certainly do not wish to suggest that these requirements are all there is to learning in phonology. However, we do believe that it is a meaningful starting point that is relevant for the evaluation of any theory that aims at explanatory adequacy in phonology.

In Sections 3 and 4 below we show that the MDL principle, when coupled with a suitable representational framework (for concreteness, we will use rule-based phonology), favors hypotheses that seem appropriate with respect to the different aspects of the learning challenge considered here. This makes MDL a promising candidate for the child's learning criterion. In Section 5 we argue that other approaches in the literature on learning in phonology have yet to address central aspects of the learning challenge.

THE PRESENT WORK

3

The current section presents the assumptions behind our learning model. One general assumption that we make is that the child chooses between competing grammars using some kind of evaluation metric. We start, in Section 3.1, by considering two evaluation metrics from the literature – the evaluation metric of the *Sound Pattern of English* (SPE; Chomsky and Halle 1968, p. 334), which aims for grammar economy, and the subset principle, which aims for restrictiveness – in the context of acquiring a single optional phonological rule. We will see that in order to acquire the relevant rule, the child cannot follow grammar economy alone or restrictiveness alone but must instead balance between the two. This balancing of economy and restrictiveness is the essence of the MDL evaluation metric, and while we motivate it here using one simple rule, the very same metric will serve as a good guide for learning whole (though at present artificial) phonological grammars, including the lexicon, the morphological segmentation of forms into stems and affixes, a variety of phonological rules, and both transparent and opaque rule interactions. In order to use the MDL

evaluation metric as a part of an actual phonological learner, we need to adopt explicit representations for phonological grammars. We do this in Section 3.2, where we present the concrete representations we assume and the costs they induce in terms of MDL. Section 3.3 presents a search procedure that will allow us to turn the MDL metric into a full learner, and while our focus in this paper is the MDL metric rather than the full learner, it is through reporting simulations with the learner that we will be able to best illustrate the kind of guidance provided by MDL (in Section 4).

3.1

The MDL criterion

French has an optional process of liquid-deletion word-finally following an obstruent (Dell, 1981). The French-learning child, then, might be exposed to surface forms such as [tabl] and [tab] for ‘table’ and [katr] and [kat] for ‘four’ (but only [gar] and not *[ga] for ‘train station’, since its liquid does not appear in the right environment for deletion). Suppose that the child uses a simplicity metric such as the one in SPE, which optimizes grammar economy. Restricting our attention here and below to grammars that are licensed by Universal Grammar (UG) and using $|G|$ to notate the length of a grammar G , we can state this metric as follows:¹

- (5) SPE EVALUATION METRIC: If G and G' can both generate the data D , and if $|G| < |G'|$, prefer G to G'

To see how we can use (5), we need to be precise about how $|\cdot|$ is measured. Anticipating our discussion below, it will be convenient to think of grammars as sitting in computer memory according to a given encoding scheme – a scheme that is provided by UG – with $|G|$ being the number of bits taken up by G . In Section 3.2 we will present the details of one specific encoding scheme and show how $|G|$ is measured within it. For now, however, we will set aside such details as we build toward the MDL criterion.

¹ Here and below the grammar G will be taken to be not just the phonological rules and their ordering but also the lexicon. Thus, by saying that a grammar G generates the data D , we mean that every string in D can be derived as a licit surface form from some UR in the lexicon and the ordered phonological rules.

Early on, the child will store a separate UR for each surface form of the alternating pairs: both /tabl/ and /tab/ for ‘table’; both /katr/ and /kat/ for ‘four’; both /arbr/ and /arb/ for ‘tree’; and so on (along with a single /gar/ for ‘train station’). After seeing a few additional alternating pairs of this kind, however, (5) will lead the child to conclude that for each such pair there is just one UR – /tabl/ for ‘table’, /katr/ for ‘four’, /arbr/ for ‘tree’, and so on – and that an optional phonological rule such as the following applies (where *L* stands for *liquid*):²

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

The rule in (6) adds complexity to the grammar, but this complexity is more than offset by the savings obtained by the elimination of all the *L*-less forms from the lexicon. Consequently, the overall size of the grammar is shorter using (6), and (5) will favor the new grammar.

As mentioned above, however, the actual process of *L*-deletion in French is somewhat more specific than (6) suggests: *L* may be deleted, but only in certain contexts. A more appropriate rule is the following, in which *L*-deletion is restricted to word-final environments following an obstruent:

(7) $L \rightarrow \emptyset$ /[-son]__# (optional)

And unfortunately, as pointed out by Dell (1981), a child using (5) will fail to acquire the appropriate context for the application of the rule. That is, the child will prefer (6) to the more appropriate (7). This is so since (a) both a grammar *G* using the unrestricted (6) and a grammar *G'* using the restricted (7) can generate the data; and (b) *G* is shorter than *G'* (since specifying the context in (7) adds to the grammar’s length). By the SPE evaluation metric in (5), the child will prefer *G* to *G'*, which is the wrong result. For example, a child using *G* will

²An even simpler grammar is one in which the lexicon includes just one, empty UR and in which any segment can be inserted by an optional rule. Such a grammar would be an extreme example of a very simple but wildly overgenerating grammar, and we could have used it instead of (6) to illustrate the perils of minimizing $|G|$ alone in our discussion below. In the interest of keeping the presentation focused on deletion processes, however, we set this grammar aside and start from (6).

erroneously rule in *L*-deleted forms such as *[ga] for /gar/.³ Moreover, the child will never recover from this error: since the child sees only positive evidence, they will never be forced to leave the simpler but overly inclusive *G*.

The problem is quite general, as discussed by Braine (1971) and Baker (1979), and goes well beyond phonology: a child guided solely by a preference for grammar economy, as in the SPE evaluation metric in (5), will fail to learn the contexts for optional rules. Just as in the example of optional *L*-deletion, a grammar *G* in which an optional rule *R* has no context will generally be both simpler and more inclusive than a minimal variant *G'* in which the optional rule does have a context. If *G'* is the correct grammar, both grammars will be able to generate the input data: *G'* since it is the correct grammar, and *G* since its *language* – that is, the set of all licit forms according to the lexicon and rules of *G* – is a superset of the language of *G'*. By (5), then, the child will incorrectly prefer the simpler *G* to *G'* and – since the child will not receive negative evidence – will never recover from this error.

One solution to this predicament – the one advocated by Dell (1981) and adopted in much later work – is to change the evaluation metric from one that favors simple grammars to one that favors restrictive ones, where restrictiveness is captured in terms of subsethood: *G* is more restrictive than *G'* if its language is a subset of the language of *G'*.⁴ This solution, also known as the *subset principle* (Berwick

³In fact, as mentioned in footnote 2, a preference for grammar economy will lead the learner to even more extreme solutions if left unchecked. In particular, consider a grammar (as in footnote 2) that has an optional epenthesis rule for each segment that appears in the data and a lexicon that consists only of the empty string. Such a grammar can generate the data and is extremely short to state. Unless it is blocked by some other principle, this grammar will be preferred by (5) to both *G* and *G'*.

⁴Other ways of cashing out the informal idea of restrictiveness have been proposed in the literature. Within Optimality Theory (Prince and Smolensky 1993), for example, restrictiveness is often interpreted as subsethood not of the languages of the original grammars *G* and *G'* but rather of the languages of variants of *G* and *G'* in which the lexicon is replaced with the set Σ^* of all possible strings over the alphabet Σ in which the lexicon is written (see Smolensky 1996). The MDL metric, which we will present and argue for below, implements restrictive-

1985; Wexler and Manzini 1987; Hale and Reiss 2003, 2008), directs the learner to never choose a grammar for a superset language when a grammar for a proper subset is compatible with the data:⁵

- (8) SUBSET EVALUATION METRIC: If G and G' can both generate the data D , and if the language of G is a proper subset of the language of G' , prefer G to G'

A child following (8) will always choose from among the grammars sanctioned by UG and whose language is compatible with the data a grammar whose language is minimal in terms of subsethood. Such a child will therefore avoid the overgeneralization problem. In the case of optional L -deletion in French, the grammar with the unrestricted (6) generates a language that is a strict superset of the one with the restricted (7), and both grammars generate the data D ; consequently, the unrestricted (6) will be rejected and the restricted (6) chosen, which is the correct result.

While choosing correctly between (6) and (7), the subset principle gives rise to a problem of undergeneralization – the mirror image of the overgeneralization problem of the SPE simplicity metric – and does not offer a general solution for learning. To see the problem in the case of French L -deletion, consider the situation of a learner who has heard a surface form such as [sabl] but, accidentally, has not yet heard its L -elided variant [sab] (both for the UR /sabl/ ‘sand’). If the learner has heard sufficiently many other pairs differing only in whether they have a final liquid, we would expect them to adopt (7), even if for /sabl/ only one member of the pair has been observed so far. That is, we would like the learner to generalize beyond the data in this case. But if the learner follows the subset principle, this will not be possible: with (7), the language will include also the L -deleted form [sab], which makes the language a strict superset of the language of a grammar that does not generate [sab]. One example of such an

ness in yet another way, by comparing how easy it is to specify the actual input data using G and G' : if the data can be more easily specified using G than using G' , then G is the more restrictive grammar of the two.

⁵As Baker (1979) notes, Braine’s (1971) alternative to the SPE evaluation metric, while stated in procedural terms, has a similar effect to a restrictiveness metric.

overly restrictive grammar is one without any deletion rules and with a lexicon that has separate URs for each of the L -variants that have been seen in the input data. For a learner that follows the subset principle, the only way to avoid such an overly restrictive grammar is if it is not licensed by UG. On most theories of UG, however, a memorizing and overly specific grammar is perfectly capable of being represented. Consequently, the learner will fail to choose the correct and more permissive (7). In other words, as long as UG makes available overly restrictive grammars, a single accidental gap is enough to prevent a learner following the subset principle from making what seems like a reasonable generalization.

We have seen that minimizing $|G|$, as in the SPE evaluation metric, makes the child generalize; when left unchecked, however, it leads to overgeneralization. Meanwhile, restrictiveness (as in the subset principle) protects from overgeneralization, but on its own prevents useful generalizations. It seems sensible, then, to try to balance the two principles against each other: look for a grammar that is both reasonably small and reasonably restrictive. This is exactly the idea behind Minimal Description Length (MDL; Rissanen 1978), which we will adopt here.⁶ To make it work, however, we need to specify how we quantify both grammar size and restrictiveness and how the two are balanced. The insight of MDL – building on the work of Solomonoff (1964a,b), Kolmogorov (1965), and Chaitin (1966) – is that we can think of restrictiveness as another simplicity criterion and combine it naturally with grammar economy. As above, for grammar economy we will consider G as sitting in computer memory according to a given encoding – as specified by UG – and measure $|G|$ in terms of how many bits the storage of G takes up. Restrictiveness, meanwhile, will be thought of in terms of how simple it is to describe the data, D , given the grammar, G . We will use the notation $D : G$, somewhat loosely, for the shortest description of D given G (loosely because there might be multiple such shortest descriptions), and we will notate the length of the shortest description of D given G as $|D : G|$.⁷ To see how $|D : G|$

⁶ See also the closely related idea of Minimal Message Length of Wallace and Boulton (1968).

⁷ In what follows, we will consider D to be the actual data sequence that the learner is exposed to. Consequently, $D : G$ will be the description of those

is measured given a grammar G , consider again the case of optional L -deletion. Suppose that the learner has acquired a lexicon with the single UR /tabl/ and an optional rule such as (6) or (7). To describe an instance of the surface form [tabl] or the surface form [tab], we need to first specify the UR /tabl/ and then specify whether L -deletion has applied (for [tab]) or not (for [tabl]). Specifying the UR /tabl/ involves a choice from among the URs. In general, the greater the number of URs from which we choose, the longer the specification of the UR we have selected. A convenient way of specifying such choices – and one that will allow us to directly balance the length of $D : G$ against that of the grammar G – is using bits. A single bit encodes one binary choice, and as the number of bits grows, the number of choices that can be stated grows (exponentially) with it. For example, if there are just two possible URs, we can specify the choice using one bit. With four URs in the lexicon, we now need about two bits to specify each choice. And so on.⁸ The optional L -deletion rule requires the further specification of whether it applied or not, which can be stated as one additional bit (perhaps 0 to specify that the rule did not apply and 1 to specify that it did). These specifications for the different surface forms in the input data D are accumulated to provide the complete $D : G$, the encoding of the specific input data D given the grammar G .

actual input tokens given the grammar. This choice is made for concreteness and in order to keep the presentation simple. A different possibility would be to abstract away from individual tokens and consider only the types – that is, the distinct surface forms – rather than the tokens. It is also possible to define the restrictiveness factor $|D : G|$ in terms of a combined measure of types and tokens. We will not attempt to investigate these choices and their implications for learning within this paper (see Goldwater *et al.* 2006, Endress and Hauser 2011, and Yang 2016 for relevant discussion).

⁸Exactly how many bits are needed for each choice will depend on the specific grammar G , relative to which the choices are made. In Section 3.2 we show how $D : G$ is stated relative to the grammars presented in that section. For similar considerations regarding the measurement of $|G|$ and $|D : G|$ in bits but within constraint-based phonology see Rasin and Katzir 2016. We further note that the number of bits used for a given choice point need not be uniform. In general, the optimal cost of each choice x in bits will be $-\lg P(x)$ (that is, minus the logarithm base two of the probability of x). A fixed number of bits per choice point is optimal only if the probability distribution at each choice point is uniform.

We can now see how the motivation for restricting the context for optional L -deletion can be stated in terms of simplicity. If L -deletion were not optional – if it always applied or if it never applied – the final bit would have been unnecessary for the specification of the relevant surface forms: selecting a UR would have fully determined the surface form. For URs like /tabl/ and /katr/, L -deletion is optional, and the extra bit of the appropriate rule cannot be avoided. But for /gar/ L -deletion never applies, so paying an extra bit for each occurrence is an unnecessary expense. The unrestricted (6) forces us to pay this unnecessary expense: the optional rule is applicable whenever a UR is chosen that contains liquids (and for each occurrence of a liquid within such a UR), including URs such as /gar/ that do not allow for L -deletion, so a bit specifying whether the rule applies is always required, leading to $D : G$ that is longer than needed. The more restrictive (7), on the other hand, makes us pay the extra bit only when an appropriate UR such as /tabl/ is chosen but not when /gar/ is chosen. Consequently, (7) leads to a shorter $D : G$.

Having recast the notion of restrictiveness in terms of simplicity (specifically, the simplicity of $D : G$), we can directly combine it with simplicity of grammar: instead of minimizing $|G|$ alone, as in the SPE evaluation metric, we can now minimize the sum of the two quantities, $|G| + |D : G|$, thus balancing between the goal of a simple, general grammar and a restrictive one.

- (9) 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'

Combining grammar economy with restrictiveness in terms of the subset principle as stated in (8) is a nontrivial challenge. Combining it with the reformulation of restrictiveness in terms of $|D : G|$, on the other hand, is straightforward, as (9) shows. Moreover, the MDL quantity $|G| + |D : G|$ has a direct interpretation in terms of quantities that are arguably available to the learner, as discussed in Katzir 2014 and Rasin and Katzir 2020. Grammars are stored in memory according to the specifications provided by UG, and $|G|$ is therefore simply the amount of memory required to store G using this specification. As for $|D : G|$, any given grammar G considered by the learner and compatible with D can presumably be used to parse D , and if this parse is stored in memory, its storage space is $|D : G|$. This makes $|G| + |D : G|$

nothing more than the overall storage space used for keeping G and its (shortest) parse of D in memory. This makes MDL a natural evaluation criterion that uses only quantities that are available to the learner with minimal stipulation beyond what is already needed to represent grammars and use them to parse the data.⁹

Let us now return to the L -deletion example and see how MDL leads to an adequate level of generalization. As discussed above, storing a single UR for pairs like [tabl]/[tab] and [katr]/[kat] will shorten $|G|$ sufficiently (given a large enough number of such pairs) to justify adding an optional rule of L -deletion to G , just as with the SPE evaluation metric. As for the precise form of the rule, the simultaneous consideration of both $|G|$ and $|D : G|$, as in (9), will mean that the more complex rule in (7) will eventually be chosen over the unrestricted (6), despite its increased $|G|$. The reason is that after sufficiently many instances of words like [gar] have been encountered, the savings in terms of $|D : G|$ obtained with (7) – since no bit will need to be spent when a UR such as /gar/ is chosen – will more than outweigh the increase in $|G|$. Figure 1 illustrates. The MDL metric in (9) thus allows the child to generalize but protects them from overgeneralizing.

Note that, differently from the case of restrictiveness-only (as in the subset principle), the MDL metric has the means to generalize beyond the data even in the face of certain gaps in the input. Consider again the situation of a learner who has heard the form [sabl] but has not (yet) heard its L -deleted variant [sab]. We saw earlier how this kind of gap in the input data will prevent a restrictiveness-only learner from generalizing correctly. For an MDL learner (that is, a learner that relies on the MDL metric to choose between hypotheses), the added restrictiveness of ruling out [sab] is weighed against the added complexity in stating a grammar that does that while still accounting for

⁹A reviewer suggests combining $|G|$ not with $|D : G|$ but rather with $|L(G)|$, the cardinality of the language of G . We note, however, that using $|L(G)|$ as a proxy for restrictiveness will only be useful when the language of the target grammar is finite, and this assumption is problematic even within morpho-phonology due the possibility of unbounded processes of affixation. And even if the languages under consideration are assumed to be finite, computing $|L(G)|$ strikes us as significantly more challenging than using $|D : G|$, a quantity that as just discussed is presumably already available to the learner.

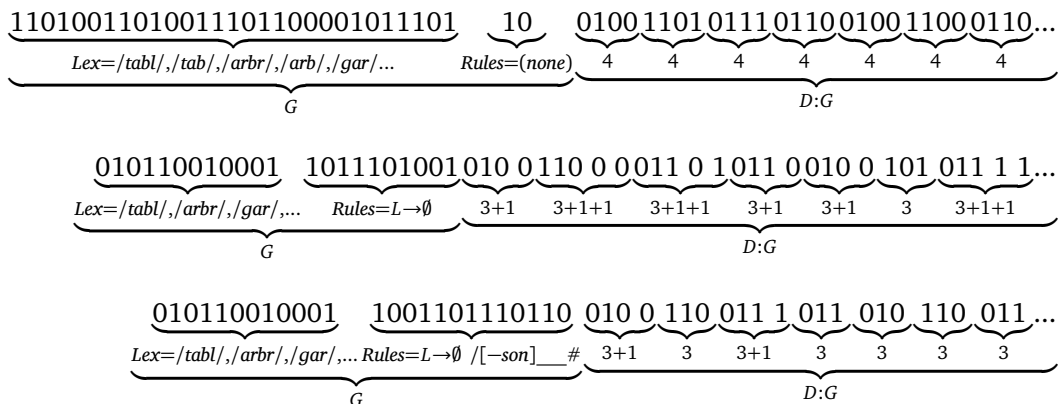


Figure 1: Schematic illustration of three hypotheses. (The order of URs in the lexicon and of tokens in $D : G$ are unrelated.) Introducing a naive lexicon (*top*), in which [tabl] and [tab] have distinct URs results in a complex grammar. Capturing optional L -deletion with (6) allows the grammar to 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 (7), allows us to limit the extra bit to just those URs that require it (*bottom*): $/\text{tabl}/$ but not $/\text{gar}/$

both [tabl] and [tab]. In the present case, a grammar that rules out [sab] will be quite complex: it might dispense with L -deletion and resort to memorizing each observed surface form using a separate UR; or it might state a highly involved rule (or system of rules) that license L -deletion in those forms where both variants of a pair has been observed. Either way, the result will be a complex grammar that does not justify the minimal savings obtained by not having to specify whether L -deletion has applied for the single occasion when the UR $/\text{sabl}/$ was chosen. (This is very different from the case of [gar], where preventing inappropriate L -deletion involved only a slight increase in grammar size, and where there were sufficiently many relevant instances of L in non-deleting environments to justify the added complexity.) Consequently, the accidental gap arising from seeing an occurrence of [sabl] without an instance of [sab] will not prevent the MDL learner from keeping the rule of L -deletion in (9), thus generalizing beyond the data, which seems to be the correct result.

Suppose now that the learner sees not just one instance of [sabl] but rather many instances, still without any instance of [sab]. In this case, the absence of [sab] will start looking less like an accident of the specific data sequence seen so far and more like a systematic fact of French that needs to be captured. The MDL learner allows us to make this intuition precise: with sufficiently many occurrences of [sabl], the extra bit that is needed to state for each occurrence that /sabl/ does not undergo optional *L*-deletion results in an increase to $|D : G|$ that is big enough to justify blocking *L*-deletion for this UR. How exactly *L*-deletion is blocked will depend on the representations available to the learner. For example, if these representations offer a general way to mark exceptions to rules, the learner might choose to mark /sabl/ as an exception to *L*-deletion. If such a method is not available, the learner might choose to block *L*-deletion in a more *ad hoc* way. For example, the learner might decide to add a special segment at the end of the UR (e.g., storing the relevant UR as /sablx/), thus preventing the *L* under consideration from appearing in the right context for deletion, along with a rule that deletes that special segment and is ordered after *L*-deletion.

Before proceeding, we note that in the discussion above we assumed that the input to the learner is a sequence of surface forms of words in isolation. If further information is available to the learner, such as the order of words in sentences or representations of scenes in which words are uttered, the decision of the learner regarding which forms to collapse using phonological rules can change. For example, a learner considering a small portion of the English lexicon containing ‘spare’, ‘pear’, ‘spit’, ‘pit’, ‘stick’, ‘tick’, and similar pairs might mistakenly collapse these pairs with the aid of an optional rule of [s]-deletion before [p] word-initially. By considering not just words in isolation but also the linguistic and extra-linguistic contexts in which they appear, however, an MDL learner will be justified in moving to a more complex grammar that does not collapse the relevant pairs but rather represents them using distinct URs in the lexicon.

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

and Clark (2001), among others.¹⁰ Recently, Rasin and Katzir (2016) have used MDL to show how phonological grammars can be acquired distributionally within constraint-based phonology, and Rasin and Katzir (2018, 2020) have discussed the acquisition of abstract URs using MDL. The present work extends this approach, using rule-based phonology as a concrete representational framework. In particular, we will show how the same MDL metric that supported the correct generalization in the case of the optional rule of *L*-deletion in French, as discussed above, will support the acquisition of whole phonological grammars, including the lexicon, the segmentation of forms into stems and affixes, a variety of phonological rules, and both transparent and opaque rule interactions. The simulations illustrating the use of MDL for the acquisition of phonological grammars – at present, using small corpora generated from artificial grammars – will be presented in Section 4. Before that, in the remainder of the present section, we describe the phonological representations that we assume, in order to make explicit their contribution to the MDL score, and we describe the search procedure we use to traverse the space of possible grammars.

3.2

Representations

As is standard, we assume that segments in phonological rules are represented not atomically but as feature bundles.¹¹ For convenience, each simulation below works with a feature table that makes distinctions that are relevant to the phenomenon at hand, but we remain agnostic here as to whether learners start with a large innate table or acquire language-specific tables at an earlier stage. To illustrate, the feature table in Table 1 will be used for those simulations that are based on English.

¹⁰MDL and Bayesian grammar induction are almost equivalent. There are some differences, such as MDL's use of the shortest encoding of D given G , which corresponds to the maximal probability of a parse of D given G , while Bayesian learning marginalizes over all parses. As far as we can tell, however, such differences are irrelevant to the examples discussed here, and we will treat MDL and Bayesian inference as essentially the same for the purposes of this paper.

¹¹In principle, the same holds also for the lexicon, though in the implementation reported here, the representation of segments in the lexicon does not explicitly use feature bundles.

Table 1: Feature table

	<i>cons</i>	<i>voice</i>	<i>cont</i>	<i>coronal</i>	<i>low</i>	<i>high</i>	<i>back</i>	<i>son</i>	<i>lateral</i>	<i>labial</i>	<i>strident</i>
d	+	+	+	-	-	-	-	-	-	-	-
t	+	-	+	-	-	-	-	-	-	-	-
z	+	+	+	+	-	-	-	-	-	-	+
s	+	-	+	+	-	-	-	-	-	-	+
g	+	+	-	-	-	-	-	-	-	-	-
k	+	-	-	-	-	-	-	-	-	-	-
b	+	+	-	-	-	-	-	-	-	+	-
p	+	-	-	-	-	-	-	-	-	+	-
m	+	+	-	-	-	-	-	+	-	+	-
n	+	+	+	-	-	-	-	+	-	-	-
r	+	+	+	+	-	-	-	+	-	-	-
l	+	+	+	+	-	-	-	+	+	-	-
a	-	+	+	+	+	-	+	+	-	-	-
o	-	+	+	+	-	-	+	+	-	-	-
e	-	+	+	+	-	-	-	+	-	-	-
i	-	+	+	+	-	+	-	+	-	-	-
u	-	+	+	+	-	+	+	+	-	-	-

Phonological rules

3.2.1

Feature bundles based on feature tables such as the one in Table 1 are used to state the phonological rules. The general form of rules is as follows, where A, B are feature bundles or \emptyset ; X, Y are (possibly empty) sequences of feature bundles; and *optional?* is a boolean variable specifying whether the rule is obligatory or optional (Figure 2).

$$\underbrace{A}_{\text{focus}} \rightarrow \underbrace{B}_{\text{change}} / \underbrace{X}_{\text{left context}} \text{ — } \underbrace{Y}_{\text{right context}} \text{ (optional?)}$$

Figure 2:
Rule format

The following, for example, is an optional phonological rule of vowel harmony that fronts a vowel before another front vowel when the two are separated by arbitrarily many consonants, stated in textbook notation in (10a) and in string notation (more convenient for the purposes of the conversion to bits below, and using various delimiters, marked with # with certain subscripts and discussed shortly) in (10b).

(10) Vowel harmony rule

a. Textbook notation

$$[-cons] \rightarrow [-back] / _ [+cons]^* \begin{bmatrix} -cons \\ -back \end{bmatrix} \text{ (optional)}$$

b. String notation

$$-cons\#_{rc} -back\#_{rc}\#_{rc} +cons*\#_b -cons\#_f -back\#_{rc} 1\#_{rc}$$

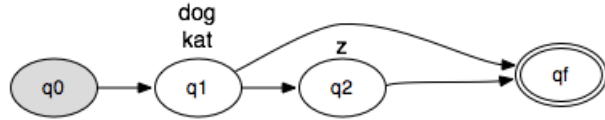
As discussed informally in Section 3.1 above, determining both $|G|$ and $|D : G|$ for purposes of MDL is done in bits, where each bit represents a single binary choice. In the simple representations that we use in this paper, all possible outcomes at any particular choice point (whether binary or otherwise) are treated as equally easy to encode. For purposes of presentation, we will first discuss a particularly simple representation in which at any given choice point, the different outcomes are not just equally easy on average to encode but actually have fixed, equal length codes. This will allow us to discuss the various encodings in terms of fixed conversion tables in which if there are n possible outcomes, each will be assigned a code whose length in bits is $\lceil \lg n \rceil$ (that is, the logarithm base two of n , rounded up to the closest integer). In our actual simulations, presented in Section 4, we will deviate from the encoding presented below by allowing non-integral code lengths, taking $\lg n$ rather than $\lceil \lg n \rceil$ as the code length for an n -ary choice point.¹²

Within the simplified representational framework just described, determining the length in bits of a single phonological rule for the purposes of MDL is done by using a conversion table that states the codes for the possible elements within phonological rules. An example of a possible conversion table appears in Table 2.

¹²The reason for this change is that the encoding used in the current section, using $\lceil \lg n \rceil$, is highly sensitive to changes in which the number of outcomes at a given choice point crosses a power of 2 (which is where $\lceil \lg n \rceil$ changes). By taking $\lg n$ instead of $\lceil \lg n \rceil$, this unhelpful sensitivity to powers of 2 is avoided. On the other hand, using conversion tables with fixed code lengths, corresponding to $\lceil \lg n \rceil$, allows us to keep the presentation considerably simpler than if we had to discuss $\lg n$ in terms of code lengths. We therefore keep the presentationally simpler $\lceil \lg n \rceil$ for the current section and the more robust $\lg n$ for the actual simulations.

listed in the emission table for specific states, and the possible combinations are defined by state transitions. A simple example is provided in Figure 3.

Figure 3:
An HMM representation
of a lexicon



The HMM in Figure 3 defines a lexicon with two kinds of morphemes: the stems /dog/ and /kat/, and the optional suffix /z/. As with rules, description length is not calculated directly for the standard, graphical notation of the HMM but rather for a bit-string form. As before, we start with an intermediate string representation for the HMM, as presented in Figure 4 (derived from the concatenation of the string representations for the different states, as listed in Table 3; the delimiter $\#_s$ marks the end of the list of outgoing edges from a state and $\#_w$ marks the end of each emitted word; another $\#_w$ is added at the end of each state). Within the simplified representational framework described earlier, we convert the string to a bit-string using a conversion table, as in Table 4. As before, all choices at a given point are uniform, with the same code length for all possible selections at that point ($\lceil \lg n \rceil$ if there are n possible choices). As discussed above, the actual simulations presented in Section 4 use $\lg n$ rather than $\lceil \lg n \rceil$ as the code length.

Table 3:
String representations
of HMM states

State	Encoding string
q_0	$q_0q_1\#_s\#_w$
q_1	$q_1q_2q_f\#_s\text{dog}\#_w\text{kat}\#_w\#_w$
q_2	$q_2q_f\#_s\text{z}\#_w\#_w$

Figure 4:
String representation
of an HMM

$q_0q_1\#_s\#_w\#_wq_1q_2q_f\#_s\text{dog}\#_w\text{kat}\#_w\#_wq_2q_f\#_s\text{z}\#_w\#_w$

State	Code	Segment	Code
$\#_s$	000	$\#_w$	0000
q_0	001	a	0001
q_1	010	k	0010
q_2	011	d	0011
q_f	100

Table 4:
Conversion table for HMM

Data given the grammar

3.2.3

Turning to the encoding of the data given the grammar, $D : G$, recall that the generation of a surface form involves concatenating several morphemes in a specific order and applying a sequence of phonological rules. Given the grammar as described above, specifying a surface form will therefore involve: (a) specifying the sequence of morphemes (as a sequence of choices within the lexicon, repeatedly stating the code for a morpheme according to the table in the current state followed by the code to make the transition to the next state); and (b) specifying the code for each application of an optional rule. Note that obligatory rules do not require any statement to make them apply.

Given a surface form, we need to determine the best way to derive it from the grammar in terms of code length. A naive approach to this parsing task would be to try all the ways to generate a surface form from the grammar. Even with simple grammars, however, this approach can be unfeasible. Instead, we compile the lexicon and the rules into a weighted finite-state transducer (FST) that allows us to obtain the best derivation using dynamic programming. The compilation of the rules relies on Kaplan and Kay (1994), and the FST is created by combining the rules with the HMM representing the lexicon using transducer composition.

Let us illustrate the encoding of best derivations in the case of the form $[k^h\text{æts}]$ – actually, of the simpler $[k\text{æts}]$ – using the FSTs for two simple grammars. First, consider the FST in Figure 5, which corresponds to a grammar with the lexicon in Figure 6 and no phonological rules. Using this FST, encoding the word $[k^h\text{æts}]/[k\text{æts}]$ requires 16 bits. The initial transition from q_0 to q_1 is deterministic and costs zero bits. After that, each of the four segments costs four bits: three bits to specify the segment itself (since there are eight outgoing edges

Figure 5:
Naive FST

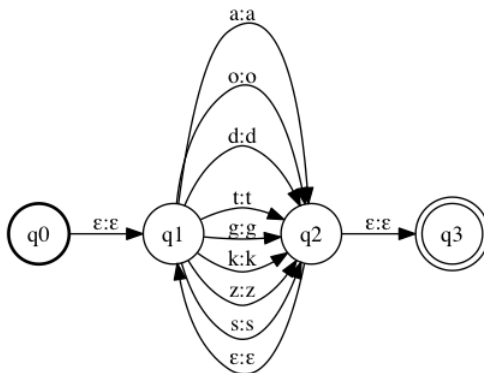


Figure 6:
Lexicon corresponding
to the naive FST

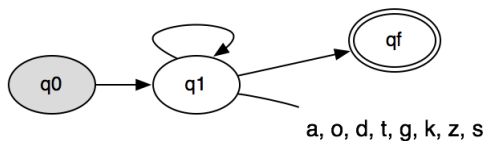


Figure 7:
Encoding of a surface form
using the naive FST

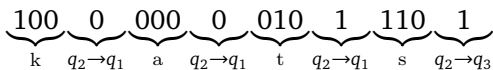


Table 5:
Conversion table
for naive FST

State q ₀		State q ₁		State q ₂	
Arc	Code	Arc	Code	Arc	Code
(-,q ₁)	ε	(a,q ₂)	000	(-,q ₁)	0
		(o,q ₂)	001	(-,q ₃)	1
		(t,q ₂)	010		
		(d,q ₂)	011		
			

from q₁) followed by one bit to specify the transition from q₂ (loop back to q₁ or proceed to q₃). The encoding, using the conversion table in Table 5, is in Figure 7.¹³

¹³Specifying [k^hæts] requires handling the aspiration of the initial segment. Since the relevant rule is obligatory, the same number of bits is required as for [kæts], though the FST is slightly more complex.

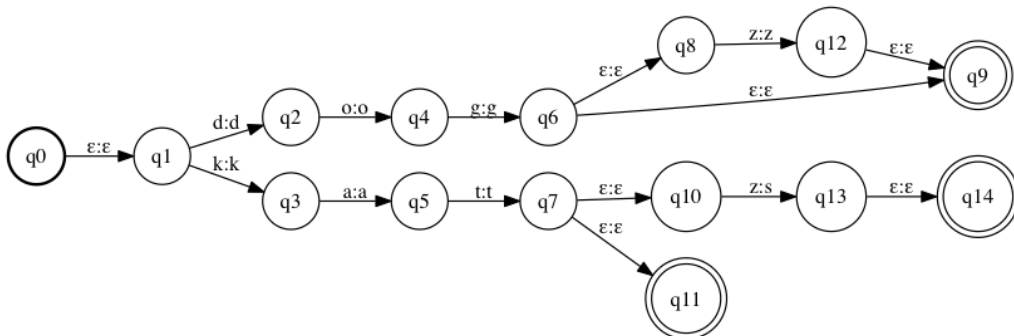


Figure 8: A more complex FST

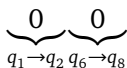


Figure 9: Encoding of a surface form using the more complex FST

Consider now the more complex FST in Figure 8, which corresponds to a grammar with the lexicon in Figure 3 and the English voicing assimilation rule. This FST corresponds to a more restrictive grammar: differently from the simpler FST in Figure 5, the present FST can only generate a handful of surface forms. Consequently, the present FST offers a shorter $D : G$. Specifically, since specifying $[k^h \text{æts}] / [k \text{æts}]$ requires making only two choices in the FST, both of them binary, it allows us to encode the relevant string using only 2 bits, as in Figure 9.

Search

3.3

Above we saw how encoding length, $|G| + |D : G|$, is derived for any specific hypothesis G . In order to use it for learning, the learner can search through the space of possible hypotheses provided by UG and look for a hypothesis that minimizes encoding length. We do not wish to make any claims about the search that the human learner might perform: our only claim in this paper concerns the MDL evaluation metric as a promising guide in comparing hypotheses. However, in order to show how this metric can guide the learner not just in the minimal comparisons discussed above but also when the learner faces a large space of possible hypotheses, we must combine the metric with

some search procedure. Since the hypothesis space is big – infinitely so in principle – an exhaustive search is out of the question, and a less naive option must be used. For concreteness, we adopt a genetic algorithm (GA), a general strategy that supports searching through complicated spaces that involve multiple local optima (Holland 1975).

The search starts with a random population of hypotheses that are generated by randomly selecting a lexicon and a set of ordered rules for each hypothesis. Individual hypotheses are selected for the next generation based on their fitness. The fitness of a hypothesis G equals $|G|+|D : G|$, the encoding length derived for it. Once a set of hypotheses is selected for the next generation, each pair of hypotheses is crossed-over to produce two offspring which replace their parents, and each offspring undergoes a random mutation to either its lexicon or its rule set. The simulation ends after a specified number of generations. The fittest hypothesis in the last generation is reported below as the final grammar.¹⁴

4

SIMULATIONS

The present section provides several simulations in which the MDL learner described in Section 3 is faced with unanalyzed data exhibiting various linguistically-relevant patterns.¹⁵ We are not able to test the learner on real-life corpora at this point: both the size of the relevant part of the search space and the time it takes to parse each hypothesis during the search grow rapidly with the size and complexity of the corpus. Instead, we provide a proof-of-concept demonstration, using small datasets generated by artificial grammars that incorporate phonologically interesting dependencies. We return to this matter in Section 6. To simulate a larger corpus, we multiply $|D : G|$ by 10 in the simulations reported below (the effect is similar to presenting the learner with each word 10 times). The one exception to the multiplication of $|D : G|$ by 10 is the simulations in Section 4.1 for which we use

¹⁴For a detailed discussion of the search procedure see Lan (2018).

¹⁵The code for the simulations is available at https://github.com/taucompling/morphophonology_spe.

different multipliers, as discussed below. Also with the exception of Section 4.1, each simulation allowed for between 1 and 5 states in the HMM, between 0 and 5 phonological rules, and between 0 and 2 feature bundles in both the left context and the right context of each rule.

Section 4.1 illustrates our learner's acquisition of optionality, using a dataset based on the case of optional French *L*-deletion discussed above. Section 4.2 uses a dataset based on /-z/-affixation in English to illustrate the joint acquisition of affixation and phonological processes. Section 4.3 extends the results of Section 4.2 by showing how the learner can acquire two rules and their ordering in the case of transparent rule interaction. Section 4.3 modifies the English-based dataset to one that involves counterbleeding opacity and shows that the MDL learner succeeds in this case as well. Section 4.5 shows that the MDL learner succeeds on a case of counterfeeding opacity modeled after the interaction of two processes in Catalan.

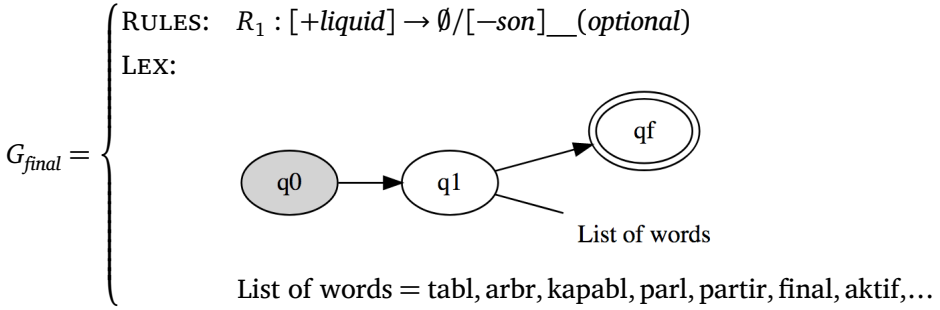
Optionality

4.1

The first dataset shows a pattern modeled after French *L*-deletion (Dell, 1981) and is designed to test the learner on the problem of restricted optionality. As discussed in Section 3.1, the challenge for the learner is to strike the right balance between economy and restrictiveness. The learner needs to generalize beyond the data and conclude that for each pair like [tab]–[tabl] there is a single UR, and that a rule of *L*-deletion optionally applies. But the learner must not overgeneralize and should restrict *L*-deletion to only apply after obstruents, despite the added complexity of specifying the restricted environment in the description of the rule.

The data presented to the learner in the present simulation consisted of 91 words, including 33 collapsible pairs (since the task in our simulations is the acquisition of a grammar from distributional evidence alone, from the learner's perspective the data are an unstructured sequence of surface forms: the learner does not know that surface forms like [tab] and [tabl] are related in any way). A sample of the data is given in (12).

(12) tab, tabl, arb, arbr, kapab, kapabl, parl, partir, final, aktif, ...



Description length: $|G_{final}| + |D : G_{final}| = 29,100.4 + 30,153.8 = 59,254.3$

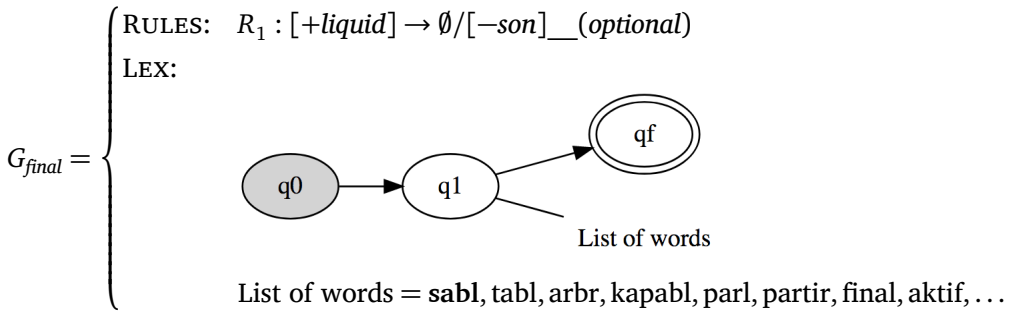
Figure 10: Final grammar for the French optionality simulation. The grammar includes the restricted *L*-deletion rule and forms like /tabl/ without their *L*-deleted counterparts (like /tab/). Here and below all scores are rounded to the first decimal place

The parameters for the present simulation were different from those for the other simulations reported in this paper (and mentioned above). In the present simulation, the encoding length of the data given the grammar was multiplied by 50, and the encoding length of the HMM was multiplied by 20. The simulation also allowed only one state in the HMM, between 0 and 2 phonological rules, and up to one feature vector in the left context and in the right context of each rule. We tried running the simulation also with the usual parameters, but the search did not converge. At present, we are not sure whether this is because the search was difficult in this case or because of something more significant.

The learner induced the correct optional rule and converged on the target lexicon (Figure 10). Compared to the final (correct) grammar, the over-generating hypothesis has a shorter grammar but a longer $D : G$, leading to an overall longer description:

- (13) a. Correct Hypothesis:
- $R_1 : [+liquid] \rightarrow \emptyset / [-son] _ _ (\text{optional})$
 - Description length:
 $|G| + |D : G| = 29,100.4 + 30,153.8 = 59,254.3$
- b. Over-generating Hypothesis:
- $R_1 : [+liquid] \rightarrow \emptyset / _ _ (\text{optional})$
 - Description length:
 $|G| + |D : G| = 29,092.9 + 32,853.8 = 61,946.7$

In Section 3.1 we discussed the undergeneralization problem for restrictiveness-only learning principles like the subset principle. We mentioned a scenario in which a learner has heard a surface form such as [sabl] but, accidentally, has not yet heard its *L*-elided variant [sab]. We noted that, while we would expect the human learner to generalize and learn *L*-deletion in the face of a single accidental gap, the subset principle predicts that *L*-deletion would be avoided. The MDL principle, on the other hand, predicts generalization. We ran another simulation of French using a variant of the corpus in (12) in which [sabl] was added without its *L*-elided variant [sab]. As expected, the learner generalized correctly and converged on the hypothesis in Figure 11 which includes the *L*-deletion rule and a variant of the lexicon that also contains /sabl/.



Description length: $|G_{final}| + |D : G_{final}| = 29,517.5 + 30,610.1 = 60,127.6$

Figure 11: Final grammar for a variant of the French-optionality simulation with an occurrence of [sabl] in the data but no occurrences of [sab]. The grammar includes the *L*-deletion rule which can generate the unattested [sab] as an output of /sabl/

Joint learning of morphology and phonology

4.2

Our next simulation demonstrates the learner’s ability to perform joint learning of morphology and a single phonological rule. Other works in the literature that perform joint learning of this kind include Naradowsky and Goldwater (2009) and (in a framework of constraint-based phonology) Rasin and Katzir (2016). After establishing this baseline,

we will proceed, in the following sections, to the joint learning of morphology and rule interaction, a task that, as discussed in Section 5, has not been accomplished in previous work. In the present simulation, the learner’s tasks are to decompose the unanalyzed surface forms into a lexicon of underlying morphemes and to learn the relevant phonological rule.

Our example is modeled after English voicing assimilation where, as discussed in Section 1, the suffix /z/ becomes voiceless following a voiceless consonant. The learner was presented with 250 words generated by creating all combinations of 25 verbal stems with 10 suffixes (including the null suffix) and applying voicing assimilation.¹⁶ A sample of the data is provided in (14).

stem\suffix	∅	-z	-ing	-er	...
rent	rent	rents	renting	renter	
(14) kontrol	kontrol	kontrolz	kontrolling	kontroler	
glu	glu	gluz	gluing	gluer	
...					

The simulation converged on the grammar in Figure 12, which contains the correct rule and segmented lexicon. Given this grammar, generating a surface form requires first choosing a stem (out of 25 stems, at a cost of $\lg 25$ bits), then choosing a suffix (out of 10 suffixes, at a cost of $\lg 10$ bits), which makes a total of $\lg 25 + \lg 10 \approx 7.96$ bits for encoding each surface form. For comparison, consider the minimally-different alternative hypothesis in (15) that fails to learn the voicing-assimilation rule and stores both -z and -s as suffixes without collapsing them into a single UR. The hypothesis in (15) has a slightly smaller $|G|$: it stores an additional suffix in the lexicon (-s) but saves some space by omitting the rule. On the other hand, (15) over-generates. Any stem can be suffixed by either -z or -s regardless of the voicing of its final consonant. Thus, for example, both [rents] and [rentz] can be generated from the stem /rent/. This over-generation

¹⁶When attached to verbs, as in our simulation, the suffix /z/ marks the 3rd person singular in present tense. Since at present we do not model part-of-speech categories, our presentation of voicing assimilation will not distinguish this suffix from the nominal plural marker /z/.

set of forms. In terms of $|D : G|$, encoding each surface form given the memorizing hypothesis would require choosing one out of 250 words in the lexicon at a cost of $\lg 250$ bits. Since $\lg 250 = \lg 25 + \lg 10$, this cost is identical to the cost given the target hypothesis. Despite the tie in the value for $|D : G|$, the target hypothesis wins due to its strictly smaller $|G|$. In a more realistic setting, the corpus will typically contain gaps, which would give the memorizing hypothesis an advantage in terms of $|D : G|$. For example, if five stem + suffix combinations (e.g., [kontrol-er]) are missing from the corpus, encoding a surface form given the memorizing hypothesis would cost $\lg 245$ bits, compared to an unchanged cost of $\lg 250$ for the target hypothesis (which can generate the five unattested combinations). As the data D grows, this wastefulness of the target hypothesis in terms of $|D : G|$ would accumulate and at some point outweigh the savings in the lexicon obtained by segmenting D . To estimate the effect of an increase in D , we created a variant of the data in (14) by omitting five words chosen at random, and we calculated different values for $|G| + |D : G|$ while varying the multiplier for $|D : G|$. We found that when the multiplier for $|D : G|$ exceeds 1,039, the target hypothesis loses to the memorizing hypothesis in terms of the combined $|G| + |D : G|$. We re-ran the simulation several times with the gapped corpus using each of the following multipliers for $|D : G|$: 10, 100, 1,000, 10,000, and 100,000. The simulation converged on the target hypothesis in Figure 12 in all cases. At least for the cases of the multipliers 10,000 and 100,000, this means that the simulation converged on a sub-optimal hypothesis. Since this is an accident of the search procedure, whose modeling is not our focus in this paper (as mentioned in Section 3.3), we leave attempts to optimize the results with larger multipliers to a separate occasion.

4.3

Rule ordering

Rule-based phonology accounts for the interaction of phonological processes through rule ordering. In English, as we have seen, voicing assimilation devoices the suffix $/-z/$ when preceded by a voiceless obstruent. Epenthesis inserts the vowel [ɪ] between two sibilants (as in [glæsɪz], ‘glasses’). To derive forms such as [glæsɪz], where voicing

assimilation does not apply and the suffix remains voiced, epenthesis is ordered before assimilation. When epenthesis applies to the UR /glæs-z/, it *bleeds* assimilation by disrupting the adjacency between the suffix and the preceding consonant, rendering assimilation inapplicable. The opposite ordering would have derived the incorrect form *[glæsis], as demonstrated in (16):

- (16) a. Good: epenthesis before assimilation

	/glæs-z/
Epenthesis	glæsɪz
Assimilation	–
	[glæsɪz]

- b. Bad: assimilation before epenthesis

	/glæs-z/
Assimilation	glæss
Epenthesis	glæsis
	*[glæsis]

Our next dataset was generated by an artificial grammar modeled after the interaction of voicing assimilation and epenthesis in English. The learner was presented with 250 words generated by creating the same combinations of stems and suffixes as in the previous section and applying epenthesis (17a) and voicing assimilation (17b), in this order. A sample of the data is provided in (18). The learner converged on the expected lexicon and on the two rules – epenthesis (R_1) and assimilation (R_2) – and their correct ordering (Figure 13).

- (17) Rules

- a. Rule 1: [i]-epenthesis between stridents
- b. Rule 2: Progressive assimilation with [–voice] spreading to an adjacent segment

	stem\suffix	∅	-z	-ing	-er	...
	rent	rent	rents	renting	renter	
(18)	klaimb	klaimb	klaimbz	klaimbing	klaimber	
	kros	kros	krosiz	krosing	kroser	
	...					

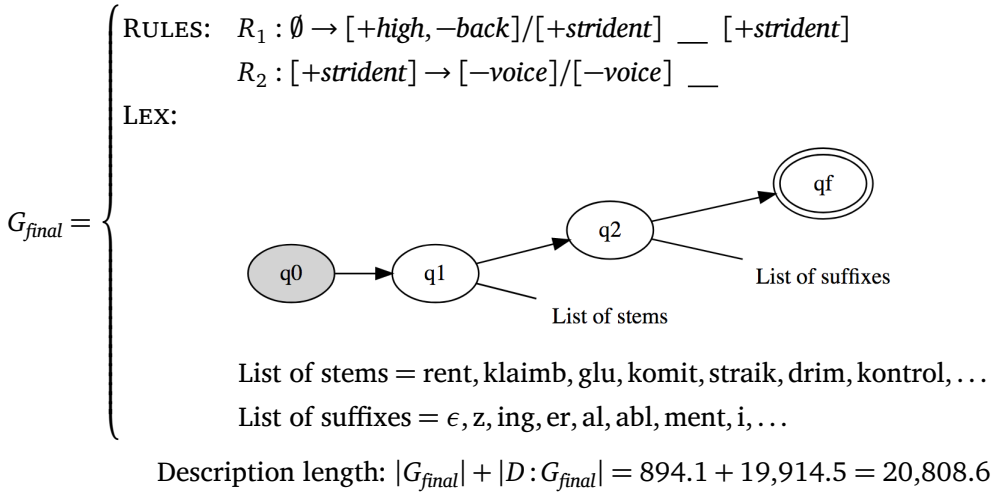


Figure 13: Final grammar for the rule-ordering simulation. The grammar includes epenthesis and voicing assimilation, in this order, and a segmented lexicon

4.4

Counterbleeding opacity

The term *opacity* is used to describe rules whose effect is obscured on the surface, often because of an interaction with another rule (Kiparsky 1971, Baković 2011). One type of opacity called *counterbleeding* in the literature results when a rule R_2 removes the conditions for the application of another rule R_1 which has applied earlier in the derivation. R_1 is opaque since its environment of application is missing on the surface.

Our next dataset was designed to test the learner on the problem of counterbleeding opacity. We used two rules modeled after English epenthesis and voicing assimilation and changed the order such that assimilation was ordered first:

(19) Rules

- a. Rule 1: Progressive assimilation with $[-voice]$ spreading to an adjacent segment
- b. Rule 2: $[i]$ -epenthesis between stridents

The result is that feature spreading takes place even between segments that are separated by an epenthetic vowel on the surface.

Examples of natural languages that reportedly show a similar interaction between feature spreading and epenthesis are some varieties of English and Armenian, as reported in Vaux (2016), and Iraqi Arabic, as reported in Kiparsky (2000, citing Erwin, 1963).

As shown in (20), the opposite rule ordering would lead to the wrong result. Given the correct order, epenthesis applies after assimilation, rendering assimilation opaque: the first consonant of the suffix undergoes assimilation but is preceded by the epenthetic vowel on the surface.

(20) Voicing assimilation crucially precedes epenthesis

a. Good: assimilation before epenthesis

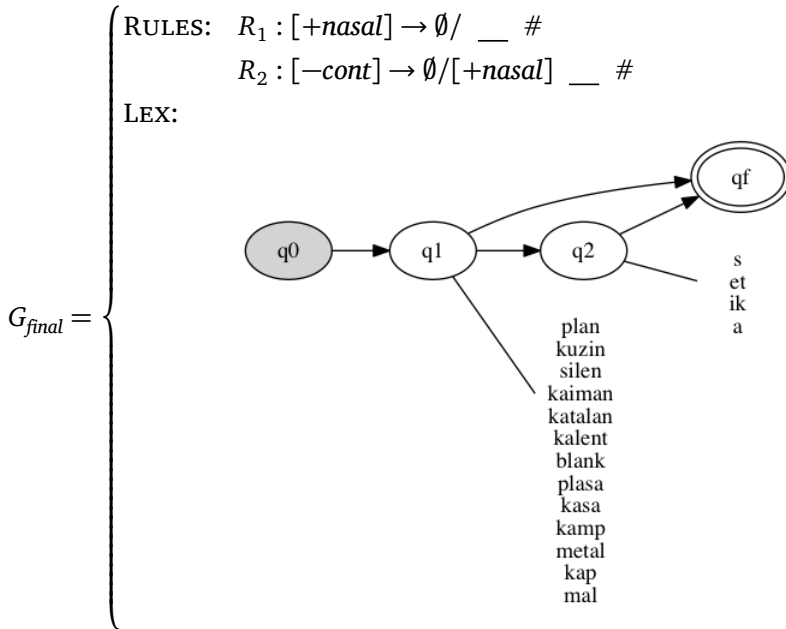
	/glæs-z/
Assimilation	glæss
Epenthesis	glæsis
	[glæsis]

b. Bad: epenthesis before assimilation

	/glæs-z/
Epenthesis	glæsɪz
Assimilation	–
	*[glæsɪz]

For this simulation, the dataset was generated by taking the same combinations of 25 stems and 10 suffixes as before and applying voicing assimilation and epenthesis, in this order. A sample of the data is provided in (21). The learner converged on the expected lexicon and on the two rules – assimilation (R_1) and epenthesis (R_2) – and their correct ordering (Figure 14).

stem\suffix	∅	-z	-ing	-er	...
rent	rent	rents	renting	renter	
kontrol	kontrol	kontrolz	kontrolling	kontroler	
kros	kros	krosis	krosing	kroser	
...					



Description length: $|G_{final}| + |D : G_{final}| = 1093.9 + 14,563.1 = 15,657.1$

Figure 15: Final grammar for the counterfeeding opacity simulation. The grammar includes final-nasal deletion and cluster simplification (in this order) and a segmented lexicon

(23) Rules

- a. Rule 1: Delete a nasal word-finally
- b. Rule 2: Delete a word-final stop following a nasal

stem\suffix	\emptyset	-s	-et	...
(24) kalent	kalen	kalents	kalentet	
kuzin	kuzi	kuzins	kuzinet	
...				

The learner converged on a segmented lexicon and on the two rules – final-nasal deletion (R_1) and cluster simplification (R_2) – and their correct ordering, as in Figure 15. There was one difference between the final result and the grammar used to generate the corpus. The rule of cluster simplification induced by the learner deletes stops

in a broader environment: after any non-continuant consonant rather than only after nasals. Since all word-final consonant-stop clusters in our corpus were nasal-stop clusters, multiple left contexts for cluster simplification were consistent with the data, including left contexts that specify nasal consonants ([+nasal]), any non-continuants ([−cont]), or any consonants ([+cons]). The statements of these three left contexts are equally simple under our current representations, so the learner is expected to choose between them arbitrarily given this corpus.

5

EXPLANATORY ADEQUACY
AND 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 instances of opacity. By accomplishing all of these tasks, the learner goes beyond previous work in terms of its ability to address the challenge of explanatory adequacy discussed in Section 2: arriving at a descriptively-adequate grammar based on primary linguistic data.

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 achieving explanatory adequacy. This is because previous learners either do not work with what we take to be the primary linguistic data (e.g., by assuming that the child is given direct information about URs) or because they do not arrive at a full phonological grammar (e.g., by not acquiring opacity). To make the comparison easier, we will focus on five components of the learning challenge: learning from distributional evidence alone, learning segmentation simultaneously with phonology, learning opacity, learning optionality, and learning

Table 6: Some prominent proposals from past work on learning in phonology and their ability to address five learning challenges

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)			

abstract URs. Each of the learners we discuss fails on at least one of those components, as summarized in Table 6 (and as discussed in the rest of this section).

We first consider constraint reranking algorithms (row 1 in Table 6), 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 do not address 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 Table 6). 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 sup-

porting evidence from alternations at all, as in Alderete and Tesar's (2002) 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 Table 6). 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 – both allophony and alternations – in a fully distributional way (row 4 in Table 6). Since their learner is closer to our goals than the previous learners are, we discuss it here in more detail. The proposal extends the allophonic learner of Peperkamp *et al.* (2006). Peperkamp *et al.* detect maximally dissimilar contexts as hints for allophonic distribution. For example, [æ] and [ǣ] are allophones in English, and the contexts that they can appear in are very different: [ǣ] can only appear before a nasal consonant, while [æ] can only appear elsewhere. Peperkamp *et al.* provide a statistical score that identifies such dissimilarities in the contexts in which two segments can appear; when two segments have highly dissimilar contexts, they are considered to be potential allophones.¹⁷ Calamaro and Jarosz (2015) look to extend Peperkamp *et al.*'s (2006) model beyond allophony, in order to account for neutralization processes. The challenge, given Peperkamp *et al.*'s dissim-

¹⁷This raises well-known issues with phonemics, such as the fact that, in English, [h] and [ŋ] are in complementary distribution but are not phonemically related. And indeed, Peperkamp *et al.* encounter many false positives (a problem that is exacerbated by the fact that their model does not require full complementary distribution). Echoing early structuralist proposals, they propose that complementarity should be combined with requirements of phonological similarity. As discussed by Chomsky (1964, p. 85), such requirements do not resolve the problem for phonemic analysis.

ilarity score, is that neutralization involves segments whose possible contexts may have a significant overlap. Consider, for example, a language like Dutch that has final devoicing. In such a language, [t] and [d] might contrast everywhere except for the context __#; a global score of contextual dissimilarity will consequently treat [t] and [d] as quite similar and fail to relate them to one another. In order to overcome this challenge, Calamaro and Jarosz consider contextualized distributional dissimilarity: for a given context X_Y and two potential alternants A and B , they compute a dissimilarity score for the triple $\langle X_Y, A, B \rangle$ by comparing the probability of the context X_Y given A and given B . These dissimilarity scores are summed for the context and for the featural change over all pairs A and B that have that change, thus allowing for generalization in terms of the change. A further extension introduces generalization over contexts (subject to two special conditions). In terms of comparison with the present proposal, Calamaro and Jarosz's model faces two challenges that, as far as we can tell, are 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 to Calamaro and Jarosz's model 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's (2006, 2009) Maximum Likelihood OT learner and Riggle's (2006) Lexicon Entropy OT learner (rows 5 and 6 in Table 6). 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 overgeneralization and undergeneralization problems of the kinds discussed earlier in Section 3. 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 those proposed by Goldwater and Johnson (2004), Goldsmith (2006), Naradowsky and Goldwater (2009), and 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. The first three learn rule-based morpho-phonology, while the fourth learns constraint-based phonology.¹⁸ Goldwater and Johnson's (2004) algorithm starts with a morphological analysis based on Goldsmith's (2001) MDL-based learner and then searches for phonological rules that lead to an improved grammar, where the improvement criterion is Bayesian. Goldsmith's (2006) learner follows a similar path but uses MDL also for the task of phonological learning. Naradowsky and Goldwater's (2009) learner is a variant of Goldwater and Johnson's (2004) learner with joint learning of morphology and phonology, thus addressing (similarly to the present learner) the interdependency of phonology and morphology. As originally presented, all three learners can acquire rules only at morpheme boundaries and generalize only with respect to X_Y and not with respect to A and B .¹⁹ They are also aimed at obligatory rules and do not handle rule interaction. Rasin and Katzir (2016) propose an MDL-based learner for Optimality Theory that can learn the URs, constraint ranking, and also the constraints themselves, from distributional evidence alone. That 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.

6

DISCUSSION

We argued that the MDL metric can adequately guide the child in choosing between competing hypotheses while learning phonology.

¹⁸Naradowsky and Goldwater (2009) target orthographic rules rather than phonology, but the difference is immaterial. Other balanced learners proposed in the literature, which are not fully distributional, include Cotterell *et al.* (2015) and Ellis and O'Donnell (2017).

¹⁹By limiting the kinds of rule that can be learned, these learners are similar to the procedural rule-based learners of Johnson (1984), Albright and Hayes (2002, 2003), and Simpson (2010).

We illustrated this with an implemented MDL-based learner for the unsupervised learning of rule-based morpho-phonological grammars. The generality of the MDL metric has allowed the learner to simultaneously perform morphological segmentation and acquire complete grammars, including URs and ordered rules, and including transparent and opaque rule interactions, as well as optional rules. By doing that, the learner is the first learner we know of that acquires opacity and optionality – basic textbook patterns that any theory of learning will have to address – from distributional evidence alone.²⁰ More generally, the learner goes beyond the phonological learning literature – including both rule-based and constraint-based learners – in its ability to address the challenge of explanatory adequacy. Previous proposals have not gone as far because they either rely on richer input data than children require or do not return a full, 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 systematic 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.

While the present work goes beyond the literature in terms of the challenge of explanatory adequacy in phonology, the simulation results we presented use corpora that are smaller than corpora used by some previous learners. In this respect the present work is in line with Chomsky's view (Chomsky 1965, p. 26), which prioritizes the comparison of learning theories based on their success on explanatory adequacy rather than on their ability to apply to large datasets:

“Clearly, it would be utopian to expect to achieve explanatory adequacy on a large scale in the present state of linguistics. Never-

²⁰To be clear, the ability of the learner to acquire opacity does not necessarily rely on its use of a rule-based formalism. For example, as noted by Baković (2011), rule-based phonology does not necessarily offer a uniform improvement over Optimality Theory in terms of its account of known opaque patterns. Since the MDL metric is general, it could in principle support the acquisition of opaque patterns using a variety of formalisms, as long as these formalisms are capable of representing these patterns.

theless, considerations of explanatory adequacy are often critical for advancing linguistic theory. Gross coverage of a large mass of data can often be attained by conflicting theories; for precisely this reason it is not, in itself, an achievement of any particular theoretical interest or importance.”

Still, an investigation of how well the MDL metric can extend to larger, more realistic corpora remains an important task that the present work has not addressed. A central part of this task is a study of the optimization procedure to see where it adequately navigates the highly complex search space and where it fails. The present work, with its focus on the MDL metric rather than the search barely starts to probe the behavior of the optimization procedure. We have to leave the examination of this question to future work.

As mentioned in Section 3.1, the simple and very general MDL metric compares hypotheses in terms of two readily available quantities: the storage space required for the current grammar and the storage space required for the current grammar’s best parse of the grammar. It has been argued recently that this approach has cognitive plausibility as a null hypothesis for language learning in humans and that it offers a reasonable framework for the comparison of different representational choices in terms of predictions about learning (see Katzir 2014, Katzir *et al.* 2020, and Rasin and Katzir 2020). From an empirical perspective, Pycha *et al.* (2003) have provided evidence that simplicity plays a central role in the acquisition of phonological rules.²¹ If correct, the present work is a step toward a cognitively plausible learner for rule-based morpho-phonology, and its predictions can be compared with those of MDL or Bayesian learners for other representation choices such as Rasin and Katzir’s (2016) MDL learner for constraint-based phonology. We leave the investigation of such predictions for future work.

²¹ See also Moreton and Pater (2012a,b) for simplicity in phonological learning (though see Moreton *et al.* 2017 for an argument that phonotactic and concept learning are guided by something closer to a Maximum Entropy model rather than by simplicity), and see Goodman *et al.* (2008) and Orbán *et al.* (2008), among others, for empirical evidence for balanced learning elsewhere in cognition.

REFERENCES

- Adam ALBRIGHT 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*, pp. 58–69, Association for Computational Linguistics.
- Adam ALBRIGHT and Bruce HAYES (2003), Rules vs. analogy in English past tenses: a computational/experimental study, *Cognition*, 90(2):119–161, doi:http://dx.doi.org/10.1016/S0010-0277(03)00146-X.
- John ALDERETE 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.
- Carl L. BAKER (1979), Syntactic theory and the projection problem, *Linguistic Inquiry*, 10(4):533–581.
- Eric BAKOVIĆ (2011), Opacity and ordering, in *The Handbook of Phonological Theory, Second Edition*, pp. 40–67, Wiley-Blackwell.
- Robert C. BERWICK (1982), *Locality principles and the acquisition of syntactic knowledge*, Ph.D. thesis, MIT, Cambridge, MA.
- Robert C. BERWICK (1985), *The acquisition of syntactic knowledge*, MIT Press, Cambridge, Massachusetts.
- Paul BOERSMA and Bruce HAYES (2001), Empirical Tests of the Gradual Learning Algorithm, *Linguistic Inquiry*, 32:45–86.
- Martin D. S. BRAINE (1971), On Two Types of Models of the Internalization of Grammars, in D. J. SLOBIN, editor, *The Ontogenesis of Grammar*, pp. 153–186, Academic Press.
- Michael BRENT (1999), An efficient, probabilistically sound algorithm for segmentation and word discovery, *Computational Linguistics*, 34(1–3):71–105.
- Shira CALAMARO and Gaja JAROSZ (2015), Learning General Phonological Rules From Distributional Information: A Computational Model, *Cognitive Science*, 39(3):647–666, doi:10.1111/cogs.12167.
- Gregory J. CHAITIN (1966), On the Length of Programs for Computing Finite Binary Sequences, *Journal of the ACM*, 13:547–569.
- Noam CHOMSKY (1964), *Current issues in linguistic theory*, Mouton & Company.
- Noam CHOMSKY (1965), *Aspects of the theory of syntax*, MIT Press, Cambridge, MA.
- Noam CHOMSKY and Morris HALLE (1968), *The Sound Pattern of English*, Harper and Row Publishers, New York.

- Alexander CLARK (2001), *Unsupervised Language Acquisition: Theory and Practice*, Ph.D. thesis, University of Sussex.
- Ryan COTTERELL, 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.
- Carl DE MARCKEN (1996), *Unsupervised Language Acquisition*, Ph.D. thesis, MIT, Cambridge, MA.
- François DELL (1981), On the learnability of optional phonological rules, *Linguistic Inquiry*, 12(1):31–37.
- Kevin ELLIS and Timothy O’DONNELL (2017), Inducing phonological rules: Perspectives from Bayesian program learning, Presented at the MIT Workshop on Simplicity in Grammar Learning.
- Timothy Mark ELLISON (1994), *The machine learning of phonological structure*, Ph.D. thesis, University of Western Australia.
- Ansgar D. ENDRESS and Marc D. HAUSER (2011), The influence of type and token frequency on the acquisition of affixation patterns: Implications for language processing., *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(1):77–95.
- Wallace M. ERWIN (1963), *A short reference grammar of Iraqi Arabic*, Georgetown University Press.
- Louann GERKEN, Rachel WILSON, and William LEWIS (2005), Infants can use distributional cues to form syntactic categories, *Journal of Child Language*, 32(2):249–268.
- John GOLDSMITH (2001), Unsupervised Learning of the Morphology of a Natural Language, *Computational Linguistics*, 27(2):153–198.
- John GOLDSMITH (2006), An Algorithm for the Unsupervised Learning of Morphology, *Natural Language Engineering*, 12(3):1–19.
- Sharon. GOLDWATER, Thomas L. GRIFFITHS, and Mark JOHNSON (2006), Interpolating between types and tokens by estimating power-law generators, *Advances in neural information processing systems*, 18:459.
- Sharon GOLDWATER and Mark JOHNSON (2004), Priors in Bayesian Learning of Phonological Rules, in *7th Annual Meeting of the ACL Special Interest Group on Computational Phonology*, pp. 35–42.
- N.D. GOODMAN, J.B. TENENBAUM, J. FELDMAN, and T.L. GRIFFITHS (2008), A Rational Analysis of Rule-Based Concept Learning, *Cognitive Science*, 32(1):108–154.
- Peter GRÜNWARD (1996), A Minimum Description Length Approach to Grammar Inference, in Stefan WERMTER, Ellen RILOFF, and Gabriele SCHELER, editors, *Connectionist, Statistical and Symbolic Approaches to Learning for Natural*

Language Processing, Springer Lecture Notes in Artificial Intelligence, pp. 203–216, Springer.

Mark HALE and Charles REISS (2003), The Subset Principle in phonology: why the tabula can't be rasa, *Journal of Linguistics*, 39:219–244.

Mark HALE and Charles REISS (2008), *The phonological enterprise*, Oxford University Press.

John H. HOLLAND (1975), *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence.*, University of Michigan Press.

James HORNING (1969), *A Study of Grammatical Inference*, Ph.D. thesis, Stanford University.

Gaja JAROSZ (2006), *Rich Lexicons and Restrictive Grammars – Maximum Likelihood Learning in Optimality Theory*, Ph.D. thesis, Johns Hopkins University, Baltimore, Maryland.

Gaja JAROSZ (2009), Restrictiveness in Phonological Grammar and Lexicon Learning, in *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, volume 43, pp. 125–139, Chicago Linguistic Society.

Mark JOHNSON (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*, pp. 344–347.

Ronald M. KAPLAN and Martin KAY (1994), Regular Models of Phonological Rule Systems, *Computational Linguistics*, 20(3):331–378.

Roni KATZIR (2014), A Cognitively Plausible Model for Grammar Induction, *Journal of Language Modelling*, 2(2):213–248.

Roni KATZIR, Nur LAN, and Noa PELED (2020), A note on the representation and learning of quantificational determiners, in Michael FRANKE *et al.*, editors, *Proceedings of Sinn und Bedeutung 24*, volume 1, pp. 392–410.

Paul KIPARSKY (1971), Historical linguistics, in W. O. DINGWALL, editor, *A Survey of Linguistic Science*, pp. 576–642, University of Maryland Linguistics Program, College Park.

Paul KIPARSKY (2000), Opacity and Cyclicity, *The Linguistic Review*, 17(2–4):351–366, doi:10.1515/tlir.2000.17.2-4.351.

Andrei Nikolaevic KOLMOGOROV (1965), Three Approaches to the Quantitative Definition of Information, *Problems of Information Transmission (Problemy Peredachi Informatsii)*, 1:1–7, republished as Kolmogorov (1968).

Andrei Nikolaevic KOLMOGOROV (1968), Three Approaches to the Quantitative Definition of Information, *International Journal of Computer Mathematics*, 2:157–168.

- Martin KRÄMER (2012), *Underlying Representations*, Cambridge University Press, Cambridge, UK.
- Nur LAN (2018), *Learning morpho-phonology using the Minimum Description Length Principle and a Genetic Algorithm*, Master's thesis, Tel Aviv University.
- Joan MASCARÓ (1976), *Catalan Phonology and the Phonological Cycle*, Ph.D. thesis, MIT.
- John J. MCCARTHY (2005), Taking a free ride in morphophonemic learning, *Catalan Journal of Linguistics*, 4:19–56.
- Elliott MORETON and Joe PATER (2012a), Structure and Substance in Artificial-phonology Learning, Part I: Structure, *Language and Linguistics Compass*, 6(11):686–701.
- Elliott MORETON and Joe PATER (2012b), Structure and Substance in Artificial-Phonology Learning, Part II: Substance, *Language and Linguistics Compass*, 6(11):702–718.
- Elliott MORETON, Joe PATER, and Katya PERTSOVA (2017), Phonological Concept Learning, *Cognitive Science*, 41(1):4–69.
- Jason NARADOWSKY and Sharon GOLDWATER (2009), Improving Morphology Induction by Learning Spelling Rules, in *IJCAI*, pp. 1531–1536.
- Aleksei NAZAROV and Joe PATER (2017), Learning opacity in Stratal Maximum Entropy Grammar, *Phonology*, 34(2):299–324.
- Andrew NEVINS and Bert VAUX (2007), Underlying representations that do not minimize grammatical violations, in Sylvia BLAHO, Patrik BYE, and Martin KRÄMER, editors, *Freedom of analysis?*, pp. 35–61, Mouton de Gruyter.
- Charlie O'HARA (2017), How abstract is more abstract? Learning abstract underlying representations, *Phonology*, 34(2):325–345.
- Gergő ORBÁN, József FISER, Richard N. ASLIN, and Máté LENGYEL (2008), Bayesian learning of visual chunks by human observers, *Proceedings of the National Academy of Sciences*, 105(7):2745–2750.
- Sharon PEPPERKAMP, Rozenn Le CALVEZ, Jean-Pierre NADAL, and Emmanuel DUPOUX (2006), The acquisition of allophonic rules: Statistical learning with linguistic constraints, *Cognition*, 101(3):B31–B41, doi:<http://dx.doi.org/10.1016/j.cognition.2005.10.006>.
- Alan PRINCE and Paul SMOLENSKY (1993), *Optimality Theory: Constraint Interaction in Generative Grammar*, Technical report, Rutgers University, Center for Cognitive Science.
- Alan PRINCE and Bruce TESAR (2004), Learning phonotactic distributions, in René KAGER, Joe PATER, and Wim ZONNEVELD, editors, *Constraints in phonological acquisition*, pp. 245–291, Cambridge University Press.

- Anne PYCHA, Pawel NOWAK, Eurie SHIN, and Ryan SHOSTED (2003), Phonological rule-learning and its implications for a theory of vowel harmony, in *Proceedings of the 22nd West Coast Conference on Formal Linguistics*, pp. 101–114, Cascadilla Press, Somerville, MA.
- Ezer RASIN and Roni KATZIR (2016), On evaluation metrics in Optimality Theory, *Linguistic Inquiry*, 47(2):235–282, doi:10.1162/ling_a_00210.
- Ezer RASIN and Roni KATZIR (2018), Learning abstract underlining representations from distributional evidence, in S. HUCKLEBRIDGE and M. NELSON, editors, *Proceedings of NELS 48*, pp. 283–290, doi:10.1017/S0022226720000146.
- Ezer RASIN and Roni KATZIR (2020), A Conditional Learnability Argument for Constraints on Underlying Representations, *Journal of Linguistics*, 56(4):745–773.
- Jason RIGGLE (2006), Using entropy to learn OT grammars from surface forms alone, in *Proceedings of the 25th West Coast Conference on Formal Linguistics*, pp. 346–353.
- Jorma RISSANEN (1978), Modeling by Shortest Data Description, *Automatica*, 14:465–471.
- Jorma RISSANEN and Eric Sven RISTAD (1994), Language Acquisition in the MDL Framework, in *Language computations: DIMACS Workshop on Human Language, March 20-22, 1992*, p. 149, Amer Mathematical Society.
- Jenny R. SAFFRAN, Elissa L. NEWPORT, and Richard N. ASLIN (1996), Statistical learning by 8-month old infants, *Science*, 274:1926–1928.
- Marc SIMPSON (2010), *From alternations to ordered rules: A system for learning Derivational Phonology*, Master’s thesis, Concordia University, Montreal.
- Paul SMOLENSKY (1996), The initial state and ‘richness of the base’ in Optimality Theory, Technical Report JHU-CogSci-96-4, Johns Hopkins University.
- Ray J. SOLOMONOFF (1964a), A formal theory of inductive inference, part I, *Information and Control*, 7(1):1–22.
- Ray J. SOLOMONOFF (1964b), A formal theory of inductive inference, part II, *Information and Control*, 7(2):224–254.
- Andreas STOLCKE (1994), *Bayesian Learning of Probabilistic Language Models*, Ph.D. thesis, University of California at Berkeley, Berkeley, California.
- Bruce TESAR (1995), *Computational optimality theory*, Ph.D. thesis, University of Colorado.
- Bruce TESAR (2014), *Output-Driven Phonology*, Cambridge University Press.
- Bruce TESAR and Paul SMOLENSKY (1998), Learnability in Optimality Theory, *Linguistic Inquiry*, 29(2):229–268.

Bert VAUX (2016), Can epenthesis counterbleed assimilation?, talk presented at NAPhC 9, Concordia University, May 7–8, 2016.

Christopher S. WALLACE and David M. BOULTON (1968), An Information Measure for Classification, *Computer Journal*, 11(2):185–194.

Kenneth WEXLER and Rita M. MANZINI (1987), Parameters and Learnability in Binding Theory, in Thomas ROEPER and Edwin WILLIAMS, editors, *Parameter Setting*, pp. 41–76, D. Reidel Publishing Company, Dordrecht, The Netherlands.

Katherine S. WHITE, Sharon PEPPERKAMP, Cecilia KIRK, and James L. MORGAN (2008), Rapid acquisition of phonological alternations by infants, *Cognition*, 107(1):238–265.

Charles YANG (2016), *The Price of Linguistic Productivity: How Children Learn to Break the Rules of Language*, MIT Press.

Ezer Rasin

Ⓘ 0000-0001-8980-5566
rasin@tauex.tau.ac.il

Iddo Berger

Ⓘ 0000-0003-1117-1166
iddoberger@gmail.com

Nur Lan

Ⓘ 0000-0003-0712-4236
nurlan@mail.tau.ac.il

Itamar Shefi

Ⓘ 0000-0001-7534-3006
itamarshefi@gmail.com

Department of Linguistics
Tel Aviv University
Tel Aviv, Israel 6997801

Roni Katzir


Ⓘ 0000-0002-0241-1896
rkatzir@tauex.tau.ac.il

Department of Linguistics
and Sagol School of Neuroscience
Tel Aviv University
Tel Aviv, Israel 6997801

Ezer Rasin, Iddo Berger, Nur Lan, Itamar Shefi, and Roni Katzir (2021), *Approaching explanatory adequacy in phonology using Minimum Description Length*, *Journal of Language Modelling*, 9(1):17–66

doi <https://dx.doi.org/10.15398/jlm.v9i1.266>

This work is licensed under the *Creative Commons Attribution 4.0 Public License*.

cc  <http://creativecommons.org/licenses/by/4.0/>