Learning rule-based morpho-phonology

Ezer Rasin, Iddo Berger, Nur Lan, and Roni Katzir

January 31, 2018

1 Introduction

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^{h}æts], the aspiration of the initial [k] and the voicelessness of the final [s] are no accident: voiceless stops such as [k] are always aspirated in this position (roughly, syllable-initially in a stressed syllable) in English, 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^{h}ætz] are not possible in the language. These pieces of knowledge come from a very large – most likely unbounded – set of possible choices that languages can make and that children must be able to acquire. Moreover, the child acquires this knowledge 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. To date, no general solution to this challenge has been provided in the literature.

In this paper we will show how a certain kind of simplicity metric can address the learning challenge, supporting the distributional learning of a morpho-phonological grammar that handles optionality and opacity. 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 SPE) and that of restrictiveness (as in the subset principle). The representational framework will be that of rule-based phonology, which offers a particularly direct handle on the representation of both optionality and opacity. The resulting learner will 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 forms, a morphological segmentation of forms into morphemes and their attachment possibilities, and different kinds of phonological rules 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 us to learn all of them in a unified way.

We start, in section 2, by presenting the MDL metric in the context of rule-based phonology and by specifying our representations and their MDL costs. In section 3 we present proof-of-concept learning simulations with optionality, rule interaction (including opacity), and interdependent phonology and morphology. To keep the presentation simple, the discussion in the first part of the paper sets aside a variety of proposals in the literature and focuses entirely on two kinds of learning biases – grammar simplicity and tightness of fit – and their combination within the MDL metric. Section 4 discusses previous work on the learning of rule-based morpho-phonology more broadly. Section 5 concludes.

2 The present work

The current section presents the assumptions behind our learning model. We start, in section 2.1, 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 the subset principle alone but must instead balance between the two. This balancing of economy and the subset principle is the essence of the MDL evaluation metric, and while we motivate it here using one simple rule, the very same metric will allow us to learn whole 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 turn the MDL evaluation metric into an actual phonological learner, we need to make explicit our representations. We do this in section 2.2, where we present the concrete representations we assume and the costs they induce in terms of MDL. Section 2.3 completes the description of our learner by presenting the search procedure that we use to find a grammar that yields a good MDL score.

2.1 The MDL criterion

French has an optional process of liquid-deletion word-finally following an obstruent (Dell, 1981). The French-learning child, then, will be exposed to surface forms such as [tabl] and [tab] for 'table' and [arbr] and [arb] for 'tree' (but only [parl] and neither *[par] nor *[pal] for 'speak', since neither liquid appears in the right environment). Suppose that the child uses a simplicity metric such as the one in SPE, which optimizes grammar economy:[1]

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

[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.

We use $|\cdot|$ to notate length, and to see how we can use (1) 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, with $|G|$ the number of bits taken up by G. In section 2.2 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.

Early on, the child will store a separate UR for each surface form of the alternating pairs: both /tabl/ and /tab/ for 'table'; both /arbr/ and /arb/ for 'tree'; and so on (along with a single /parl/ for 'speak'). After seeing a few additional alternating pairs of this kind, however, (1) will lead the child to conclude that for each such pair there is just one UR – /tabl/ for 'table', /arbr/ for 'tree', and so on – and that an optional phonological rule such as the following applies (where *L* stands for *liquid*):

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

The rule in (2) 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 (2), and (1) will favor the new grammar.

As mentioned above, however, the actual process of *L*-deletion in French is somewhat more specific than (2) 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:

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

And unfortunately, as pointed out by Dell (1981), a child using (1) will fail to acquire the appropriate context for the application of the rule. That is, the child will choose (2) rather than the more appropriate (3). This is so since a grammar G using the unrestricted (2) and a grammar G' using the restricted (3) both can generate the data (by allowing both surface forms to be derived from the single UR) and G is shorter than G' (since specifying the context in (3) adds to the grammar's length). By the SPE evaluation metric in (1), the child will prefer G to G', which is the wrong result. For example, the child will erroneously rule in *L*-deleted forms such as *[par] for /parl/. 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 (1), 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 is a superset of that of G'. By (1), 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 the set of all licit surface forms according to the lexicon and rules of G is a subset of the set of all licit surface forms for G'. This solution, also known as the *subset principle* (Berwick 1985; Wexler and Manzini 1987), directs the learner to never choose a superset language when a proper subset is compatible with the data:[2]

(4) 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 (4) will always choose a minimal language compatible with the data and will thus avoid the overgeneralization problem. In the case of optional *L*-deletion in French, the grammar with the unrestricted (2) generates a language that is a strict superset of the one with the restricted (3); consequently, the unrestricted (2) will be rejected and the restricted (2) chosen, which is the correct result.

While choosing correctly between (2) and (3), 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 pairs differing only in whether they have a final liquid, we would expect them to adopt (3), even if for /sabl/ only one member of the pair has been observed so far. But if the learner is following the subset principle, this will not be possible: with (3), the language will include also the *L*-deleted form [sab], which makes the language a strict superset of the language of a grammar 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. In other words, a single accidental gap is enough to prevent a learner following the subset principle from generalizing at all.

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, the subset principle protects from overgeneralization, but on its own prevents any generalization at all. It seems reasonable, then, to try to balance the two principles against each other: look for a grammar with a relatively small G that generates a relatively small language. This is exactly the idea behind Minimal Description Length (MDL; Rissanen 1978), which we will adopt here.[3] To make it work, however, we need to be more precise about how we quantify both grammar size and the subset principle. The insight of MDL – building on the work of Solomonoff (1964), Kolmogorov (1965), and Chaitin (1966) – is that we can think of restrictiveness as another simplicity criterion and combine it naturally with grammar economy. In particular, restrictiveness can be thought of in terms of how simple it is to tell the story of the data, D, given the

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

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

grammar, G, a story that we will notate as $D : G$. 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 (2) or (3). 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.[4] And so on. 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.

We can now see how optionality can be cashed out 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 /arbr/, L-deletion is optional, and the extra bit of the appropriate rule cannot be avoided. But for /parl/ L-deletion never applies, so paying an extra bit for each occurrence is an unnecessary expense. The unrestricted (2) forces us to pay this unnecessary expense: the optional rule is applicable whenever an L-final UR is chosen, including URs such as /parl/ that do not allow for L-deletion, so a bit specifying whether the rule applies is always needed, leading to $D : G$ that is longer than needed. The more restricted (3), on the other hand, makes us pay the extra bit only when an appropriate UR – /tabl/ or /arbr/ – is chosen but not when /parl/ is chosen. Consequently, (3) leads to a shorter $D : G$.

Having recast the subset principle – or the idea of restrictiveness – in terms of simplicity (specifically, the simplicity of the story of $D : G$), we can immediately see how we can combine this idea 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.

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

In our L-deletion example, storing a single UR for pairs like [tabl] and [tab] or [arbr] or [arb] will shorten $|G|$ sufficiently to justify adding an optional rule of L-

[4]Exactly how many bits are needed for each choice will depend on the specific grammar G, relative to which the choices are made. In section 2.2 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.

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 (5), will mean that the more complex rule in (3) will eventually be chosen over the unrestricted (2), despite its increased $|G|$. The reason is that after sufficiently many instances of $[\mathrm{parl}]$ have been encountered, the savings in terms of $|D : G|$ obtained with (3) – since no bit will need to be spent when a UR such as $/\mathrm{parl}/$ is chosen – will more than outweigh the increase in $|G|$. Figure 1 illustrates. The MDL metric in (5) thus allows the child to generalize but protects them from overgeneralizing.[5]

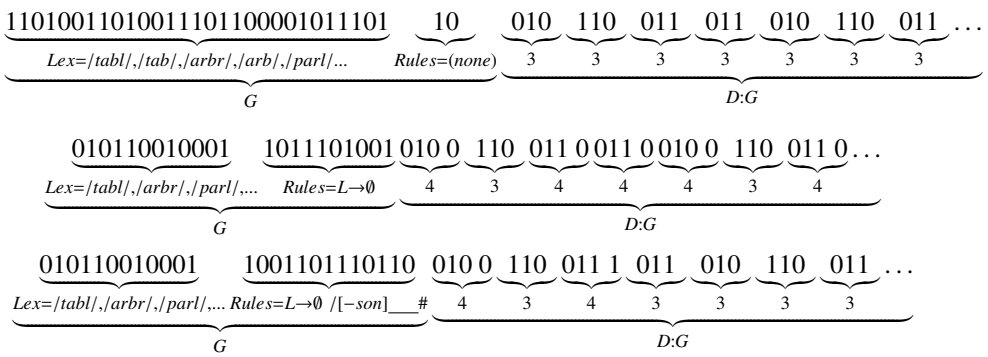


Figure 1: Schematic illustration of three hypotheses. Introducing a naive lexicon (*top*), in which $[\mathrm{tabl}]$ and $[\mathrm{tab}]$ have distinct URs results in a complex grammar. Capturing optional *L*-deletion with (2) allows the grammar to be simplified (*middle*): the complexity of the rule is outweighed by the savings of eliminating unnecessary URs. However, an additional bit is needed for specifying the actual surface form of each *L*-final UR. Finally, restricting the context of *L*-deletion, using (3), allows us to limit the extra bit to just those URs that require it (*bottom*): $/\mathrm{tabl}/$ but not $/\mathrm{parl}/$.

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), Clark (2001), Goldsmith (2001), and Dowman (2007), among others. Recently, Rasin and Katzir (2016) have used MDL to show how complete phonological grammars can be acquired distributionally within constraint-based phonology. The present work shows how the same can be done within rule-based phonology. In particular, we

[5]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 collapse these pairs with the aid of an optional rule of $[\mathrm{s}]$-deletion before $[\mathrm{p}]$ 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.

will show how the same MDL metric that supported the correct generalization in the case of the optional rule of *L*-deletion in French 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 will be presented in section 3. 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.

2.2 Representations

As is standard, we assume that segments, both in phonological rules and in the lexicon, are represented not atomically but as feature bundles. 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 Figure 2 will be used for those simulations that are based on English.

	cons	*voice*	*coronal*	*cont*	*low*
d	+	+	+	-	-
t	+	-	+	-	-
g	+	+	-	-	-
k	+	-	-	-	-
z	+	+	+	+	-
s	+	-	+	+	-
a	-	+	-	+	+
o	-	+	-	+	-

Figure 2: Feature table

2.2.1 Phonological rules

Feature bundles based on the feature table in Figure 2 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 3).

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

Figure 3: 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 arbitrar-

ily many consonants, stated in textbook notation in (6a) and in string notation (more convenient for the purposes of the conversion to bits below) in (6b).

(6) Vowel harmony rule

a. Textbook notation

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

b. String notation

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

Determining the length of the rule for the purposes of MDL is done using a conversion table that states the codes for the possible elements within phonological rules. An example of a possible conversion table appears in Figure 4. The representation scheme we use here treats all possible outcomes at any particular choice point as equally easy to encode. For the conversion table, this means that if there are n possible elements that can appear within a rule, each will be assigned a code of length $\lceil \lg n \rceil$ bits.

Symbol	Code	Symbol	Code
$\#_f$ (feature)	0000	cons	0110
$\#_b$ (bundle)	0001	voice	0111
$\#_{rc}$ (rule component)	0010	velar	1000
+	0011	back	1001
-	0100	...	...
*	0101	...	...

Figure 4: Conversion table for rules

Using the conversion table in Figure 4, we can now encode the phonological rule of vowel harmony (in (6) above) by converting each element in the string representation in (6b) into bits according to Figure 4 and concatenating the codes. To ensure unique readability, we use various delimiters to mark the end of the description of features, feature bundles, and the rule's components. The following is the result, and its length is 73 bits:

(7) Vowel harmony rule (bit representation):

$$\underbrace{0100}_{-}\,\underbrace{0110}_{cons}\,\underbrace{0010}_{\#_{rc}}\,\underbrace{0100}_{-}\,\underbrace{1001}_{back}\,\underbrace{0010}_{\#_{rc}}\,\underbrace{0010}_{\#_{rc}}\,\underbrace{0011}_{+}\,\underbrace{0110}_{cons}\,\underbrace{0101}_{*}\,\underbrace{0001}_{\#_b}$$

$$\underbrace{0100}_{-}\,\underbrace{0110}_{cons}\,\underbrace{0000}_{\#_f}\,\underbrace{0100}_{-}\,\underbrace{1001}_{back}\,\underbrace{0010}_{\#_{rc}}\,\underbrace{1}_{1}\,\underbrace{0010}_{\#_{rc}}$$

A phonological rule system is a sequence of phonological rules. Since each rule ends with the code for optionality followed by $\#_{rc}$, we can specify a phonological rule system by concatenating the encodings of the individual rules while maintaining unique readability with no further delimiters. The ordering of the rules is the order in which they are specified, from left to right. At the end of the entire rule system another $\#_{rc}$ is added.

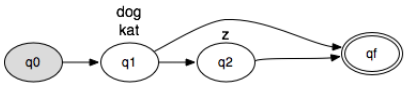


Figure 5: An HMM representation of a lexicon

2.2.2 Lexicon

The lexicon contains the URs of all the possible morphemes. Since morphemes combine selectively and in specific orders, some information about morpheme combinations must be encoded. We encode this information using Hidden Markov Models (HMMs), where morphemes are listed in the emission table for specific states, and the possible combinations are defined by state transitions. A simple example is provided in Figure 5.

The HMM in Figure 5 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 7 (derived from the concatenation of the string representations for the different states, as listed in Figure 6; 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 end of each state). We then convert the string to a bit-string using a conversion table, as in Figure 8. As before, all choices at a given point are uniform, with the same code length for all possible selections at that point.

state	encoding string
q_0	$q_0 q_1 \#_S \#_w$
q_1	$q_1 q_2 q_f \#_S \text{dog} \#_w \text{kat} \#_w \#_w$
q_2	$q_2 q_f \#_S \text{z} \#_w \#_w$

Figure 6: String representations of HMM states

$$q_0 q_1 \#_S \#_w \#_w q_1 q_2 q_f \#_S \text{dog} \#_w \text{kat} \#_w \#_w q_2 q_f \#_S \text{z} \#_w \#_w$$

Figure 7: String representation of an HMM

2.2.3 Data given the grammar

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

State	Code
$\#_S$	000
q_0	001
q_1	010
q_2	011
q_f	100

Segment	Code
$\#_w$	0000
a	0001
k	0010
d	0011
...	...

Figure 8: Conversion table for HMM

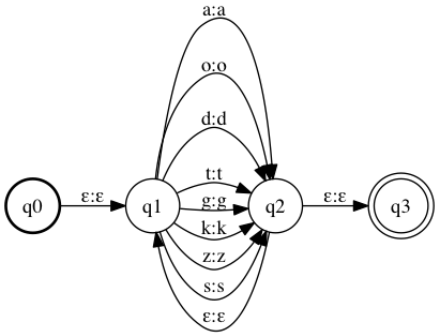


Figure 9: Naive FST

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.

Our goal, given a surface form, is 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 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).

Let us illustrate the encoding of best derivations in the case of the form $[\mathrm{k}^h\mathrm{æts}]$ – actually, of the simpler $[\mathrm{kæts}]$ – using the FSTs for two simple grammars. First, consider the FST in Figure 9, which corresponds to a grammar with the lexicon in Figure 10 and no phonological rules. Using this FST, encoding the word $[\mathrm{k}^h\mathrm{æts}]/[\mathrm{kæts}]$ 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 from q_1) followed by one bit to specify the transition from q_2 (loop back to q_1 or proceed to q_3). The encoding, using the conversion table in Figure 12, is in Figure 11.[6]

Consider now the more complex FST in Figure 13, which corresponds to a grammar with the lexicon in Figure 5 and the English voicing assimilation rule. This FST

[6]Specifying $[\mathrm{k}^h\mathrm{æts}]$ requires handling the aspiration of the initial segment. Since the relevant rule is obligatory, the same number of bits is required as for $[\mathrm{kæts}]$, though the FST is slightly more complex.

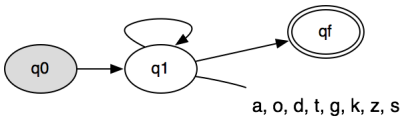


Figure 10: Lexicon corresponding to the naive FST

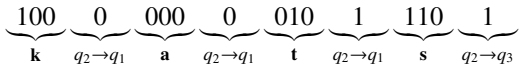


Figure 11: Encoding of a surface form using the naive FST

corresponds to a more restrictive grammar: differently from the simpler FST in Figure 9, the present FST can only generate a handful of surface forms. Consequently, the present FST offers a shorter *D*:*G*. Specifically, since specifying $[\mathrm{k}^h\mathrm{æts}]/[\mathrm{kæts}]$ 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 14.

2.3 Search

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 and look for a hypothesis that minimizes encoding length. 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. 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.

3 Simulations

The present section provides several simulations in which the MDL learner described in section 2 is faced with unanalyzed data exhibiting various linguistically-relevant patterns. 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

State q_0	
Arc	Code
(-,q_1)	ϵ

State q_1	
Arc	Code
(a,q_2)	000
(o,q_2)	001
(t,q_2)	010
(d,q_2)	011
...	...

State q_2	
Arc	Code
(-,q_1)	0
(-,q_3)	1

Figure 12: Conversion table for naive FST

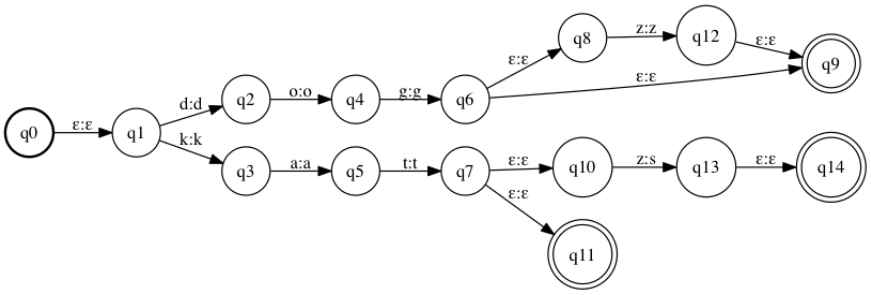


Figure 13: A more complex FST

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. 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). Section 3.1 illustrates our learner's acquisition of optionality, using a dataset based on the case of optional French *L*-deletion discussed above. Section 3.2 uses a dataset based on plural /-z/-affixation in English to illustrate the joint acquisition of affixation and phonological processes. Section 3.3 extends the results of section 3.2 by showing how the learner can acquire two rules and their ordering in the case of transparent rule interaction. Section 3.3 modifies the English-based dataset to one that involves counterbleeding opacity and shows that the MDL learner succeeds in this case as well. Section 3.5 shows that the MDL learner succeeds on a case of counterfeeding opacity modeled after the interaction of two processes in Catalan.

3.1 Optionality

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 2.1, 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

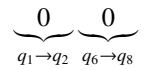
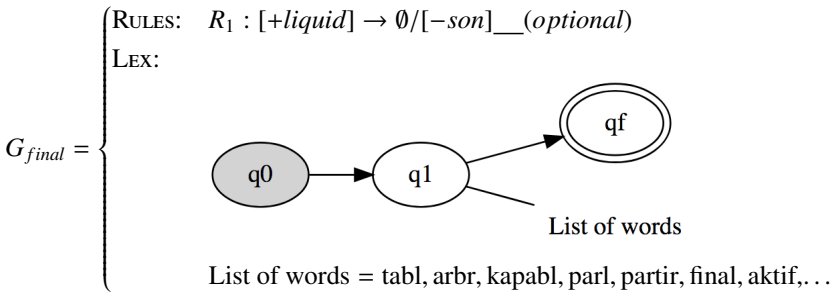


Figure 14: Encoding of a surface form using the more complex FST

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 104 words, including 36 collapsible pairs (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 (8). Encoding length of the data given the grammar was multiplied by 50 and the encoding length of the HMM was multiplied by 20.[7]

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



Description length: $|G_{final}| + |D{:}G_{final}| = 40,113 + 40,200 = 80,313$

Figure 15: 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/).

The learner induced the correct optional rule and converged on the target lexicon (Figure 15). 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:

(9) a. Correct Hypothesis:

[7]The French simulation uses other parameters than all other simulations (where the encoding length of the data given the grammar was multiplied by 10 and the encoding length of the HMM was not multiplied by any factor). In the case of French, the search with the usual parameters 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.

- R_1 : [+*liquid*] → ∅/[−*son*]__ (optional)
- Description length: $|G| + |D{:}G| = 40,113 + 40,200 = 80,313$

b. Over-generating Hypothesis:

- R_1 : [+*liquid*] → ∅/__ (optional)
- Description length: $|G| + |D{:}G| = 40,105 + 43,200 = 83,305$

3.2 Joint learning of morphology and phonology

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 4, 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 rule.

Our example is modeled after English voicing assimilation (where, as discussed in section 1, the plural suffix /z/ devoices following a voiceless obstruent). The learner was presented with 220 words generated by creating all combinations of 22 stems with 10 suffixes (including the null suffix) and applying voicing assimilation. A sample of the data is provided in (10).

(10)

stem\suffix	∅	-go	-saat	⋯
kat	kat	katko	katsaat	
dog	dog	doggo	dogsaat	
⋯				

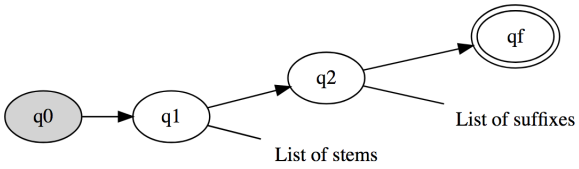
The simulation converged on the grammar in Figure 16, which contains the correct rule and segmented lexicon. Given this grammar, generating a surface form requires first choosing a stem (out of 22 stems, at a cost of 5 bits), then choosing a suffix (out of 10 suffixes, including the null suffix, at a cost of 4 bits), which makes a total of 9 bits for encoding each surface form given the final grammar. Compare this to an alternative hypothesis that memorizes the data and stores all 220 surface forms without learning any rule. On the memorizing hypothesis, generating a surface form amounts to choosing one word out of 220 words, at a cost of just 8 bits per surface form. $|D{:}G|$ is thus higher on the final grammar than on the memorizing hypothesis, but the final grammar is ultimately better since the higher $|D{:}G|$ is offset by the savings in $|G|$: since the final lexicon stores one instance of each stem and suffix – a total of 32 morphemes – it is significantly simpler than the lexicon of the memorizing hypothesis that would store 220 words, most of which consist of multiple morphemes.[8] Finally, the assimilation rule

[8]The advantage that the memorizing hypothesis has over the target hypothesis in terms of $|D : G|$ means that with a large enough corpus, this component of the MDL metric will dominate the sum $|G| + |D : G|$ and make the memorizing hypothesis win despite its larger $|G|$. This can be seen as a cause for concern, but we note that the advantage of the memorizing hypothesis in terms of $|D : G|$ in the present case is an artifact of the naive encoding scheme that we have adopted here. On this scheme, each choice point (such as selecting a particular stem from among the set of possible stems) involves a fixed, uniform – and integer –

$G_{final} =$ {

RULES: $R_1 : [+cons] \rightarrow [-voice] / [-voice]$ __

LEX:



List of stems = kat, dog, oktado, skad, daot, gkas, toot, . . .

List of suffixes = ϵ, go, saat, zoka, oad, dot, kook, kazka, . . .

Description length: $|G_{final}| + |D{:}G_{final}| = 702 + 19,800 = 20,502$

Figure 16: Final grammar for the joint learning simulation. The grammar includes the voicing assimilation rule and a segmented lexicon with URs like /-go/ from which both surface [-go] and [-ko] can be derived.

adds complexity to the set of rules, but it allows collapsing pairs of morphemes (like [-go] and [-ko]) that differ minimally on the surface into a single underlying morpheme. Compared to a hypothesis that does not learn the rule and duplicates morphemes in the lexicon (like [-go] and [-ko]), this move decreases both $|G|$ – since it allows storing fewer items in the lexicon – and $|D{:}G|$ – since having fewer morphemes means that there are fewer choices to make in specifying a surface form.

3.3 Rule Ordering

Rule-based phonology accounts for the interaction of phonological processes through rule ordering. In English, voicing assimilation devoices the plural morpheme /-z/ when preceded by a voiceless obstruent (as in [kʰæts] ,'cats', but not in [dɔgz], 'dogs'). 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 plural morpheme remains voiced, epenthesis can be ordered before assimilation. When epenthesis applies to the UR /glæs-z/, it disrupts the adjacency between the plural morpheme and the preceding consonant, rendering assimilation inapplicable. The opposite ordering would have derived the incorrect form *[glæsɪs], as demonstrated in (11):

code length. For the present case, 22 stems meant 5 bits per stem, which is wasteful (since 5 bits can mark a choice between 32 different items and not just 22, for which $\lg 22 \approx 4.459$ bits would suffice), and similar remarks apply to the choice of suffix. This wastefulness accumulates and leads to the larger $|D : G|$ of the target hypothesis. With a less naive encoding, the code length can approximate the entropic $-\lg P(\cdot)$, which means that with the present corpus each of the two hypotheses will encode a combination of stem and suffix using the same average number of bits ($\lg \frac{1}{220} \approx 7.78$) and have the same $|D : G|$. This, in turn, means that the target hypothesis will remain shorter than the memorizing hypothesis even as D grows indefinitely.

(11) a. Good: epenthesis before assimilation

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

b. Bad: assimilation before epenthesis

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

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 220 words generated by creating the same combinations of stems and suffixes as in the previous section and applying epenthesis (12a) and voicing assimilation (12b), in this order. A sample of the data is provided in (13). The learner converged on the expected lexicon and on the two rules – epenthesis (R_2) and assimilation (R_3) – and their correct ordering (Figure 17).[9] The final grammar has an interesting property that we would like to mention. The pressure for grammar economy pushed the learner to remove instances of inter-coronal low vowels from within morphemes in the lexicon, despite the lack of evidence from alternations that the relevant vowels are epenthetic (for example, [ogtad] was stored as /ogtd/). We take this to be a positive outcome, given the evidence discussed in the literature that learners sometimes posit non-identical URs for non-alternating forms (see especially McCarthy 2005).[10]

(12) Rules

a. Rule 1: Low-vowel epenthesis between coronals

b. Rule 2: Progressive assimilation of [−*voice*] (to an adjacent segment)

(13)

stem\suffix	-go	-zoka	-saat	⋯
dog	doggo	dogzoka	dogsaat	
kat	katko	katazoka	katasaat	
dok	dokko	doksoka	doksaat	
⋯				

[9] As shown in Figure (17), the search added a redundant epenthesis rule to the grammar (R_1) which has no effect on the phonological mapping. As far as we have been able to establish, this is an artifact of the mapping from actual grammars to the finite-state transducers that we use for parsing the corpus (as described in section 2.2.3). Given the usefulness of the transducers for efficient parsing, we have preferred to use them despite the slight deviation from the correct MDL measure that they lead to.

[10] Another property of the final lexicon is that the 22 stems are split between two states (8 stems in q_1 and 14 stems in q_2). This is another artifact of the naive encoding scheme we have adopted. On the split lexicon, encoding the choice of a stem costs 1 bit to first specify the choice of q_1 or q_2 and then a variable number of additional bits to specify the stem – either 3 bits (for each stem in q_1) or 4 bits (for each stem in q_2), making a total of either 4 bits or 5 bits for each stem. This split thus offers a minor saving in terms of $|D{:}G|$ over a lexicon in which all 22 stems are stored in a single state, which would uniformly require 5 bits (for a choice of one element out of a set of 22) to specify the choice of each stem.

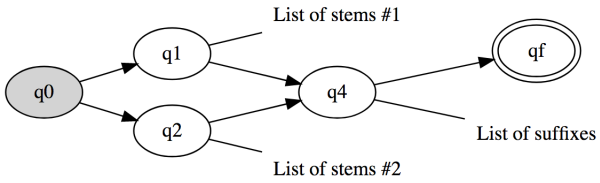
$G_{final} =$

RULES: $R_1 : \emptyset \rightarrow [+low]/[+coronal]$ __ $[+coronal, +voice]$

$R_2 : \emptyset \rightarrow [+low]/[+coronal]$ __ $[+coronal]$

$R_3 : [+cons] \rightarrow [-voice]/[-voice]$ __

LEX:



List of stems #1 = dog, skad, daot, gkas, . . .

List of stems #2 = kat, oktdo, toot, . . .

List of suffixes = ϵ, go, saat, zoka, oad, dot, kook, kazka, . . .

Description length: $|G_{final}| + |D{:}G_{final}| = 814 + 18,280 = 19,094$

Figure 17: Final grammar for the rule-ordering simulation. The grammar includes epenthesis and voicing assimilation (given in bold), in this order, and a segmented lexicon, as well as a redundant epenthesis rule that has no effect on the phonological mapping.

3.4 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 environment 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:

(14) Rules

a. Rule 1: Progressive assimilation of $[-voice]$ (to an adjacent segment)

b. Rule 2: Low-vowel epenthesis between coronals

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 (15), 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.

(15) Voicing assimilation crucially precedes epenthesis

a. Good: assimilation before epenthesis

	/kat-zoka/
Assimilation	katsoka
Epenthesis	katasoka
	[katasoka]

b. Bad: epenthesis before assimilation

	/kat-zoka/
Epenthesis	katazoka
Assimilation	-
	*[katazoka]

For this simulation, the dataset was generated by taking the same combinations of 22 stems and 10 suffixes as before and applying voicing assimilation and epenthesis, in this order.[11] A sample of the data is provided in (16). The learner converged on the expected lexicon and on the two rules – assimilation (R_1) and epenthesis (R_3) – and their correct ordering (Figure 18).[12]

(16)

stem\suffix	-go	-zoka	-saat	⋯
dog	doggo	dogzoka	dogsaat	
kat	katko	katasoka	katasaat	
dok	dokko	doksoka	doksaat	
⋯				

3.5 Counterfeeding opacity

The type of opacity called *counterfeeding* in the literature results when a rule R_2 creates the environment of another rule R_1 which applies earlier in the derivation. R_1 is opaque since it does not apply even though its environment of application is met on the surface. In Catalan (Mascaró 1976), for example (and simplifying), nasals are deleted word-finally (17a) and a rule of cluster simplification deletes a stop word-finally after a nasal (17b) and creates the environment for final-nasal deletion, which does not apply on the surface in (17b).

(17) a. kuzí ~ kuzín-s ‘cousin.SG ~ cousin.PL’

b. kəlén ~ kəlént-ə ‘hot.MASC ~ hot.FEM’

[11] The corpus for this simulation is slightly different from the corpus used in previous simulations. In a handful of cases, we made small modifications to the words so as to avoid accidental patterns that would have made the search space more difficult to traverse.

[12] The final grammar contains a redundant deletion rule (R_2) that has no effect on the phonological mapping (its effect is immediately reversed by the following epenthesis rule). See footnote 9.

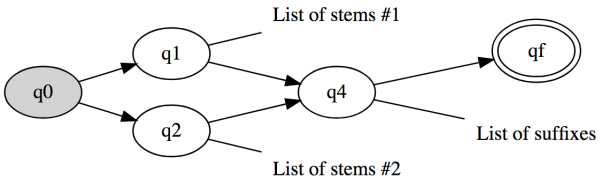
$G_{final} =$ {

RULES: $\boldsymbol{R_1 : [+cons] \rightarrow [-voice]/[-voice]\ \_\_}$

$R_2 : [+low] \rightarrow \emptyset/[+coronal]\ \_\_\ [+coronal]$

$\boldsymbol{R_3 : \emptyset \rightarrow [+low]/[+coronal]\ \_\_\ [+coronal]}$

LEX:



List of stems #1 = toot, skad, daot, . . .

List of stems #2 = dog, kat, oktado, gkas, . . .

List of suffixes = ϵ, go, saat, zoka, oad, dot, kook, kozka, . . .

Description length: $|G_{final}| + |D{:}G_{final}| = 878 + 18,400 = 19,278$

Figure 18: Final grammar for the counterbleeding opacity simulation. The grammar includes voicing assimilation and epenthesis (given in bold), in this order, and a segmented lexicon, as well as a redundant vowel deletion rule that has no effect on the phonological mapping.

Our next dataset was designed to test the learner on the problem of counterfeeding opacity. We used two rules modeled after final-nasal deletion and cluster simplification in Catalan. We generated 65 words by creating all combinations of 13 stems and 5 suffixes and applying final-nasal deletion and cluster simplification, in this order (18). The stems and suffixes were taken from a Catalan dictionary. A sample of the data is given in (19). The learner converged on the expected lexicon and on the two rules – final-nasal deletion (R_1) and cluster simplification (R_2) – and their correct ordering (Figure 19).

(18) Rules

a. Rule 1: Delete a nasal word-finally

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

(19)

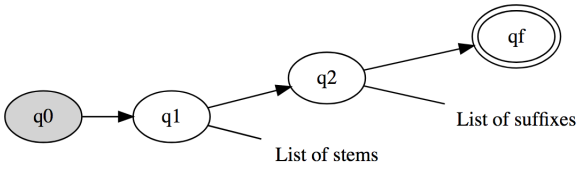
stem\suffix	∅	-s	-et	⋯
kalent	kalen	kalents	kalentet	
kuzin	kuzi	kuzins	kuzinet	
⋯				

$G_{final} =$

RULES: $R_1 : [+nasal] \rightarrow \emptyset /$ __ #

$R_2 : [-cont] \rightarrow \emptyset / [+nasal]$ __ #

LEX:



List of stems = kuzin, plan, kalent, blank, kasa, metal, kap, ...

List of suffixes = #, et#, a#, ik#, s#, ...

Description length: $|G_{final}| + |D{:}G_{final}| = 495 + 4,390 = 4,885$

Figure 19: Final grammar for the counterfeeding opacity simulation. The grammar includes final-nasal deletion and cluster simplification (in this order) and a segmented lexicon. Word boundaries are represented directly using the symbol '#'.

4 Previous work on learning rule-based 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 for a rule of the form $A \rightarrow B / X\_\_Y$ it can arrive at generalizations both in terms of the focus and the change (A and B, respectively) and in terms of the context (X and Y). And it handles both optionality and rule interaction, including instances of opacity. In this section we review past work on inducing rule-based phonology and highlight aspects of the task handled by our learner that were left open in the literature.

As we discussed in section 2.1 above, evaluation metrics that do not balance $|G|$ against $|D : G|$ have not been successful. In particular, and as discussed by Dell (1981) and others, the evaluation metric of SPE, which aimed at minimizing $|G|$, leads to overgeneralization. We further showed how a restrictiveness metric, which can be stated in terms of minimizing $|D : G|$, addresses the problem for the SPE metric but does so at the cost of failing to generalize at all. Not surprisingly, neither of these two evaluation metrics have led to actual learners.

Johnson (1984) offers the first working learner for phonological rule systems. It is particularly significant since it can handle the task of learning rule interactions, including cases of opacity. Differently from Chomsky and Halle's approach and the present one, Johnson's learner is based not on an evaluation metric that compares hypotheses given the data but rather on a procedure that obtains contexts for individual phonolog-

ical rules. In particular, when A and B alternate, Johnson's procedure examines the contexts in which A appear and those in which B appears; for the rule $A \rightarrow B / X\_\_Y$, a context $X\_\_Y$ is obtained (not necessarily uniquely) by considering what is common to all the contexts in which B appears and different from every context in which A appears. The alternating segments A and B themselves are identified with the help of morphologically analyzed paradigms, which the procedure assumes as input. The learner is thus not fully distributional. The dependence on morphological analysis to identify A and B also means that the procedure is aimed at alternations and cannot generally acquire cases of allophony that are not identifiable from alternations. It also generalizes only in terms of the context $X\_\_Y$ and provides no handle on generalizations in terms of A or B. Finally, by relying on contexts in which B appears but A does not, the procedure misses cases of optionality, which by definition involve contexts where both A and B can appear.

Johnson (1984)'s learner can be seen as the direct predecessor of the procedure-based learner for rule-based phonology proposed by Albright and Hayes (2002, 2003). Like Johnson's learner, Albright and Hayes; Albright and Hayes's learner assumes that morphological paradigms are identified in advance and are thus not fully distributional.[13] For Albright and Hayes, paradigms serve a similar role in morphology to the role they served for Johnson in phonology, namely the identification of change in an alternation, leaving the learner the task of finding the context for the change. Albright and Hayes then add a step of phonological acquisition in which the learner examines the morphological changes obtained so far and checks whether a given morphological change can apply even when superficially inappropriate by adding a phonological rule. During phonological induction, the set of possible contexts for phonological rules is provided in advance (rather than acquired) in the form of phonotactically illicit sequences. Like Johnson (1984), Albright and Hayes (2002, 2003)'s learner is aimed at alternations and cannot generally acquire cases of allophony that are not identifiable from alternations. Moreover, it does not provide a handle on generalizations in terms of A and B or on optionality, and it does not acquire rule interactions.

A different procedure-based learner was proposed by Gildea and Jurafsky (1995, 1996), who adapt Oncina et al. (1993)'s OSTIA model for the induction of certain deterministic finite-state transducers (FSTs) – specifically, subsequential FSTs – to the task of acquiring phonology.[14] OSTIA starts from an FST that faithfully maps inputs to outputs and gradually merges states in the FST while maintaining subsequentiality, and Gildea and Jurafsky enhance this process with linguistically-motivated constraints

[13] See Dunbar (2008) and Simpson (2010) for later procedure-based learners for aspects of morphophonology. Like Albright and Hayes (2002, 2003)'s learner, these proposals rely on pre-analyzed paradigmatic pairs as input to the learner and are thus not distributional.

[14] Thus, while aiming at phonological rule systems, Gildea and Jurafsky (1995, 1996) do not learn such systems directly but rather FSTs, which are a rather different kind of representation. In fact, FSTs are a computationally convenient form into which one can compile both rule-based phonology (see Kaplan and Kay 1994) and constraint-based phonology (see Frank and Satta 1998 and Riggle 2004). See Cotterell et al. (2015) for a recent learner for FSTs that, while not siding with either rule-based or constraint-based phonology is closer in spirit to the latter. We should note that Gildea and Jurafsky's goal is not the modeling of the acquisition of rule-based phonology as such but rather to investigate the role of linguistic biases in this kind of learning. In particular, they show that three quite general biases improve the acquisition of rule-based grammars within Oncina et al.'s framework.

to obtain linguistically-natural mappings of URs to surface forms. Since the procedure requires the URs to be given in advance, however, it is not distributional. Like Johnson (1984), it also generalizes entirely in terms of the context *X*__*Y* not in terms of *A* or *B*. It also has no handle on optionality (though Gildea and Jurafsky suggest that a stochastic HMM merger framework, for example along the lines of Stolcke and Omohundro 1993, might address this).[15]

Of the learners for rule-based phonology in the literature, our learner is closest to those proposed by Goldwater and Johnson (2004), Goldsmith (2006), and Naradowsky and Goldwater (2009). All three are fully distributional learners for rule-based morpho-phonology that, like Chomsky and Halle (1968), rely on an evaluation metric rather than on a procedural approach.[16] Differently from Chomsky and Halle (1968) – and similarly to the present proposal – these learners use a balanced evaluation metric that optimizes economy and restrictiveness simultaneously.[17] Goldwater and Johnson (2004)'s algorithm starts with a morphological analysis based on Goldsmith (2001)'s MDL-based learner and then searches for phonological rules that lead to an improved grammar, where the improvement criterion is Bayesian. Goldsmith (2006)'s learner follows a similar path but uses MDL also for the task of phonological learning. Naradowsky and Goldwater (2009)'s learner is a variant of Goldwater and Johnson (2004)'s learner with joint learning of morphology and phonology, thus addressing (similarly to the present learner) the interdependency of phonology and morphology. As stated, all three learners can acquire rules only at morpheme boundaries, which, as in the learners of Johnson (1984), Albright and Hayes (2002, 2003), and Simpson (2010), limits considerably the phonological rules that they learn. Like these procedural learners, the three balanced learners generalize only with respect to the context and not with respect to the change. They are also aimed at obligatory rules and do not handle rule interaction. One way of interpreting our simulations above is as showing that these limitations are not essential within this framework and that a balanced evaluation metric can support the acquisition of allophony, generalizations over both the context and the change, optionality, and rule interactions.

A final comparison for the current proposal is with the recent procedural learner of Calamaro and Jarosz (2015), which learns phonological rules – both allophony and alternations – in a fully distributional way by extending the allophonic learner of Peperkamp et al. (2006). Peperkamp et al. detect maximally dissimilar contexts as hints for allophonic distribution. For example, [æ] and [æ̃] are allophones in English, and the contexts that they can appear in are very different: [æ̃] 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 consonants in which two segments can appear; when two segments have highly dissimilar contexts, they are

[15] It is difficult to evaluate the suitability of the model to rule interaction. Gildea and Jurafsky (1995, 1996) provide an example with multiple rules, but these rules do not interact, and it remains unclear whether rule interaction (and, in particular, opacity) can be handled by their system.

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

[17] Outside of rule-based phonology, Cotterell et al. (2015) and Rasin and Katzir (2016) propose balanced learners for the acquisition of phonology, the former within a phonological framework of weighted edits and the latter within constraint-based phonology.

considered to be potential allophones.[18] Calamaro and Jarosz (2015) look to extend Peperkamp et al. (2006)'s model beyond allophony, in order to account for neutralization processes. The challenge, given Peperkamp et al.'s dissimilarity 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 $< X\_\_Y, A, B >$ 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 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.

5 Discussion

We presented an MDL-based learner for the unsupervised joint learning of lexicon, morphological segmentation, and ordered phonological rules from unanalyzed surface forms. The learner contributes to the literature on learning rule-based morphophonology, a literature that starts with Chomsky and Halle (1968) and continues with Johnson (1984), Albright and Hayes (2002, 2003), Gildea and Jurafsky (1995, 1996), Goldwater and Johnson (2004), Naradowsky and Goldwater (2009), and Calamaro and Jarosz (2015), among others. The current learner goes beyond the literature in two main respects. First, it can handle rule systems that involve not just obligatory rules but also optional ones.[19] And second, it can handle rule interaction, including cases of opacity. In handling both optionality and rule interaction the present proposal offers what to our knowledge is the first distributional learner that can acquire a full morpho-

[18] This raises all the usual 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 (even more so since they do 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.

[19] In the literature on constraint-based phonology, which we do not discuss in the present paper, the acquisition of optionality has received a fair amount of attention. See Coetzee and Pater 2011 for review and discussion. We note that proposals within this literature, such as Boersma and Hayes 2001, generally assume that the learner is given information about URs and is therefore not fully distributional.

phonological rule system with the structure proposed in the phonological literature. However, the present work has focused on small, artificial corpora that exhibit specific morpho-phonological patterns, and it remains to be seen if and how the approach can extend to larger, more realistic corpora.

The proposed learner uses the simple and very general MDL approach, in which hypotheses are compared 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 (Katzir, 2014). From an empirical perspective, Pycha et al. (2003) have provided evidence that simplicity plays a central role in the acquisition of phonological rules.[20] 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 (2016)'s MDL learner for constraint-based phonology. We leave the investigation of such predictions for future work.

References

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

Albright, Adam, and Bruce Hayes. 2003. Rules vs. analogy in English past tenses: a computational/experimental study. *Cognition* 90:119–161.

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

Baković, Eric. 2011. Opacity and ordering. In *The handbook of phonological theory, second edition*, 40–67. Wiley-Blackwell.

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

Berwick, Robert C. 1985. *The acquisition of syntactic knowledge*. Cambridge, Massachusetts: MIT Press.

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

Braine, Martin D. S. 1971. On two types of models of the internalization of grammars. In *The ontogenesis of grammar*, ed. D. J. Slobin, 153–186. Academic Press.

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

[20] 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.

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

Chaitin, Gregory J. 1966. On the length of programs for computing finite binary sequences. *Journal of the ACM* 13:547–569.

Chomsky, Noam. 1964. *Current issues in linguistic theory*. Mouton & Company.

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

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

Coetzee, Andries, and Joe Pater. 2011. The place of variation in phonological theory. In *The handbook of phonological theory*, ed. John Goldsmith, Jason Riggle, and Alan C. L. Yu, chapter 13, 401–434. Wiley-Blackwell.

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

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

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

Dunbar, Ewan. 2008. The acquisition of morphophonology under a derivational theory: A basic framework and simulation results. Master's thesis, University of Toronto.

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

Erwin, Wallace M. 1963. *A short reference grammar of Iraqi Arabic*. Georgetown University Press.

Frank, Robert, and Giorgio Satta. 1998. Optimality theory and the generative complexity of constraint violability. *Computational Linguistics* 24:307–315.

Gildea, Daniel, and Daniel Jurafsky. 1995. Automatic induction of finite state transducers for simple phonological rules. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, 9–15. Association for Computational Linguistics.

Gildea, Daniel, and Daniel Jurafsky. 1996. Learning bias and phonological-rule induction. *Computational Linguistics* 22:497–530.

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

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

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

Goodman, N.D., J.B. Tenenbaum, J. Feldman, and T.L. Griffiths. 2008. A rational analysis of rule-based concept learning. *Cognitive Science* 32:108–154.

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

Holland, John H. 1975. *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence.*. U Michigan Press.

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

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

Kaplan, Ronald M., and Martin Kay. 1994. Regular models of phonological rule systems. *Computational Linguistics* 20:331–378.

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

Kiparsky, Paul. 1971. Historical linguistics. In *A survey of linguistic science*, ed. W. O. Dingwall, 576–642. University of Maryland Linguistics Program, College Park.

Kiparsky, Paul. 2000. Opacity and cyclicity. *The Linguistic Review* 17:351–366.

Kolmogorov, Andrei Nikolaevic. 1965. Three approaches to the quantitative definition of information. *Problems of Information Transmission (Problemy Peredachi Informatsii)* 1:1–7. Republished as Kolmogorov (1968).

Kolmogorov, Andrei Nikolaevic. 1968. Three approaches to the quantitative definition of information. *International Journal of Computer Mathematics* 2:157–168.

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

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

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

Moreton, Elliott, and Joe Pater. 2012a. Structure and substance in artificial-phonology learning, part i: Structure. *Language and Linguistics Compass* 6:686–701.

Moreton, Elliott, and Joe Pater. 2012b. Structure and substance in artificial-phonology learning, part ii: Substance. *Language and Linguistics Compass* 6:702–718.

Moreton, Elliott, Joe Pater, and Katya Pertsova. 2017. Phonological concept learning. *Cognitive Science* 41:4–69.

Naradowsky, Jason, and Sharon Goldwater. 2009. Improving morphology induction by learning spelling rules. In *IJCAI*, 1531–1536.

Oncina, J., P. García, and E. Vidal. 1993. Learning subsequential transducers for pattern recognition interpretation tasks. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 15:448–458.

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

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

Pycha, Anne, 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*, volume 22, 101–114.

Somerville, MA: Cascadilla Press.
Rasin, Ezer, and Roni Katzir. 2016. On evaluation metrics in Optimality Theory. *Linguistic Inquiry* 47:235–282.
Riggle, Jason. 2004. Generation, recognition, and learning in finite state Optimality Theory. Doctoral Dissertation, UCLA, Los Angeles, CA.
Rissanen, Jorma. 1978. Modeling by shortest data description. *Automatica* 14:465–471.
Rissanen, Jorma, and Eric Sven Ristad. 1994. Language acquisition in the MDL framework. In *Language computations: DIMACS Workshop on Human Language, March 20-22, 1992*, 149. Amer Mathematical Society.
Simpson, Marc. 2010. From alternations to ordered rules: A system for learning derivational phonology. Master's thesis, Concordia University, Montreal.
Solomonoff, Ray J. 1964. A formal theory of inductive inference, parts I and II. *Information and Control* 7:1–22, 224–254.
Stolcke, Andreas. 1994. Bayesian learning of probabilistic language models. Doctoral Dissertation, University of California at Berkeley, Berkeley, California.
Stolcke, Andreas, and Stephen Omohundro. 1993. Hidden Markov Model induction by Bayesian model merging. In *Advances in neural information processing systems*.
Vaux, Bert. 2016. Can epenthesis counterbleed assimilation? Talk presented at NAPhC 9, Concordia University, May 7-8, 2016.
Wallace, Christopher S., and David M. Boulton. 1968. An information measure for classification. *Computer Journal* 11:185–194.
Wexler, Kenneth, and Rita M. Manzini. 1987. Parameters and learnability in binding theory. In *Parameter setting*, ed. Thomas Roeper and Edwin Williams, 41–76. Dordrecht, The Netherlands: D. Reidel Publishing Company.