Simultaneous learning of vowel harmony and segmentation

Ezer Rasin
Leipzig University
ezer.rasin@uni-leipzig.de

Nur Lan
Tel Aviv University
nurlan@mail.tau.ac.il

Roni Katzir
Tel Aviv University
rkatzir@post.tau.ac.il

1 Overview

Vowel harmony (VH) is a long-distance phonological pattern in which multiple vowels share one or more features, often across an unbounded number of intervening consonants which are inert with respect to the pattern. In Turkish, for example, the plural suffix can surface as [ler] or [lar] due to VH, depending on the backness of the preceding vowel, as in table 1; and the genitive suffix can surface as [in], [ɪn], [un], or [ün], depending on both the backness and rounding of the preceding vowel, again as in table 1. In these examples, the number of consonants that separate matching vowels ranges from one (as in [ip-in]) to three (as in [kurt-lar]).

	'rope'	'girl'	'village'	'wolf'
Plural	ip-ler	kɪz-lar	köj-ler	kurt-lar
Genitive	ip-in	kɪz-ɪn	köj-ün	kurd-un

Table 1: Turkish vowel harmony

VH presents special challenges to the child acquiring the morpho-phonology of their language that go beyond the general difficulty of morpho-phonological induction from distributional evidence alone – that is, from unanalyzed surface forms, with no corrections, explicit underlying representations (URs), or paradigmatic information. First, VH poses difficulties for learners who separate the segmentation task, where individual morphemes are identified, from the phonological task, where the phonological dependencies are established. Second, VH requires identifying phonological processes that apply across an unbounded number of intervening consonants. We discuss these learning challenges in more detail in section 2. Then, in section 3, we present a general learning approach, based on the principle of Minimum Description Length (MDL), which has been used for the simultaneous induction of segmentation and phonology in Rasin et al. 2018b using phonological rewrite rules as in SPE (Chomsky and Halle 1968). We explain why the MDL criterion is able to address the two learning challenges at once. In section 4, we extend the MDL learner of Rasin et al. (2018b) by allowing its phonological rewrite rules to be stated using an equivalent of the subscript-zero notation of early generative phonology (where, for example, C_0 stands for a sequence of zero or more consonants). In section 5, we show that with this minimal extension of the representations available to the learner, the learner succeeds in acquiring the segmentation and the VH pattern simultaneously from distributional evidence alone in a small corpus of Turkish words.

2 Two learning challenges posed by VH

2.1 Challenge I: Segmentation vs. phonological learning

According to various proposals in the literature, morpho-phonological learning is split into two parts: morphological segmentation and phonological induction (see, e.g., Goldwater and Johnson 2004, as well as work on morpho-phonological learning that assumes a prior segmentation stage, such as Johnson 1984, Albright and Hayes 2003, Jarosz 2006, Merchant 2008, and Tesar 2014, among others). Regardless of the order in which these tasks are performed, VH poses a difficulty. If phonological induction applies first, learning will be hampered by the fact that VH often applies only across morpheme boundaries, which in this scenario are not yet available to the learner, while morpheme-internally vowels can be disharmonic, as in (1). In particular, as noted by Hayes and Wilson (2008, p. 402), languages with abundant disharmonic roots like Turkish pose a problem for attempts to acquire VH using a phonotac-

Proceedings of the Society for Computation in Linguistics (SCiL) 2019, pages 353-357.
New York City, New York, January 3-6, 2019

tic learner.

(1) Turkish disharmonic roots (±back mismatch)

a. me̲za̲r(-lar) 'grave(s)'

b. ho̲te̲l(-ler) 'hotel(s)'

Suppose, on the other hand, that morphological segmentation applies first, and consider again the Turkish genitive suffix in table 1, which has four different surface forms. One potential difficulty in this case is that, in the absence of knowledge of VH, each form will receive much less distributional support than a non-alternating suffix. Furthermore, even if segmentation is acquired independently of VH, there remains the nontrivial task of unifying the different surface forms and positing an appropriate phonological process, a task that can be challenging on the assumption that the child uses distributional evidence alone and does not have access to information such as paradigms.

2.2 Challenge II: Long-distance dependencies

VH often applies across a sequence of several intervening consonants that are inert with respect to the pattern. This poses a problem for phonological learners like Goldwater and Johnson 2004, Calamaro and Jarosz 2015, and Rasin and Katzir 2016, that are limited to small, local contexts of fixed size and do not have a counterpart to variable-length marking such as the C_0 notation or to autosegmental tiers. Unsupervised learners that can capture long-distance dependencies, such as Heinz 2010 or Hayes and Wilson 2008, are phonotactic learners that are not yet integrated within a full morpho-phonological learner. Moreover, surface violations of harmony within roots, as mentioned above, pose a challenge to the use of phonotactic learners for VH.

3 The MDL Principle

In what follows we show how a general approach to learning, using the MDL Principle (Solomonoff 1964; Rissanen 1978), supports the acquisition of morpho-phonology that involves VH, and in particular how it addresses Challenges I and II above. MDL is an evaluation criterion that balances two competing factors: the simplicity of the grammar ($|G|$; as in the evaluation metric of SPE); and the tightness of fit of the grammar to the data ($|D : G|$, the length of the encoding of the data D given G; similarly to the subset principle):

(2) 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'

We first show how MDL allows for the induction of a segmented lexicon and phonological rules using a toy example before turning to VH in Turkish.

Segmentation: Setting aside phonology for the moment, (2) can allow the learner to discover the morphological segmentation of words into stems and affixes (de Marcken 1996; Goldsmith 2001; Rasin et al. 2018b). If the surface forms are generated from, e.g., 8 different stems (e.g., /dok/, /kab/, etc.) and 4 different suffixes (e.g., /za/, /ti/, etc.), a naive lexicon for the language will include all the different $8 \times 4 = 32$ surface forms. By (2), the learner will prefer a simpler grammar (shorter $|G|$, while $|D : G|$ remains the same) in which the stems and the suffixes are stored separately, with only $8 + 4 = 12$ different entries (which, in addition, are shorter than those in the naive encoding).

Phonology: (2) also allows the learner to acquire various phonological processes (Goldwater and Johnson 2004; Goldsmith 2006; Rasin et al. 2018b). For example, if the language just discussed also has a process of progressive voicing assimilation across morpheme boundaries (which is slightly simpler than VH but formally very similar), the surface forms will seem to involve twice the actual number of suffixes (e.g., [sa] after voiceless stops, [za] elsewhere). Using (2) the learner will reject a naive encoding of this kind given sufficiently many suffixes (since the storage of pairs of surface forms for each suffix is costly). Instead, it will favor an encoding where there is just one variant for each suffix, along with a rule of voicing assimilation (since the savings obtained by storing just one form for each suffix outweigh the costs of adding the relevant phonological rule).

The MDL metric in (2) has two advantages as far as the learning of VH is concerned. First, it acquires the segmentation and the phonology in a simple, unified way. This is central to addressing Challenge I: since segmentation and phonological induction are performed in a unified way, they can be learned jointly, thus avoiding the potentially circular dependencies between a segmentation stage and a phonological stage. Second, MDL works directly with linguistics representations (see Katzir 2014 for discussion), which in our implementation use rewrite rules with the possibility of the equivalent of C_0. This possibility allows the MDL learner to acquire an appropriate VH rule that applies across an unbounded number of intervening consonants.

4 Representations for an MDL learner

The following is a brief summary of the representations for the MDL learner that we use (see Rasin et al. 2018b for further detail and discussion).[1]

Phonological rules

The general form of rules is as in Fig. 1, 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 optional.

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

Figure 1: Rule format

To this basic rule format, which was used in Rasin et al. (2018a,b) to acquire local phonological processes, we add the possibility of representing a feature bundle with the symbol '*', which indicates arbitrarily many repetitions of the feature bundle. Table 2 illustrates the rule format with an optional phonological rule of vowel harmony that fronts a vowel before another front vowel when the two are separated by arbitrarily many consonants. The rule is stated in textbook notation, in string notation (including various delimiters to ensure unique readability), and as a bit string (obtained from the string notation using a conversion table such as the one in table 3). For purposes of MDL, the length of the rule is number of bits in the bit-string notation.

A phonological rule system is a sequence of phonological rules. (Note that we can specify a rule system by concatenating the encodings of the individual rules while maintaining unique readability.) 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 delimiter is added.

Lexicon

The lexicon contains the URs of all the possible morphemes, along with information about their possible combinations. 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, for a toy version of English with two stems and one suffix, is provided in Fig. 2.

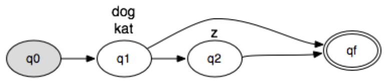

Figure 2: An HMM representation of a lexicon

The HMM in Fig. 2 defines a lexicon with two kinds of morphemes: the stems *dog* and *cat* (using a simplified transcription), 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, derived from an intermediate string representation using a conversion table (see Rasin et al. 2018b for further details).

Data given the grammar

Specifying a surface form given the grammar involves: (a) specifying the sequence of morphemes (repeatedly stating the code for a morpheme according to the table in the current state followed by the code for the transition to the next state); and (b) specifying the code for each application of an optional rule. (Obligatory rules do not require any statement to make them apply.) Given a surface form, we need to find the shortest code that derives it from the grammar. A naive approach to this parsing task would be to try all the ways to generate a surface form from the grammar, but this approach is typically infeasible. Instead, we compile the lexicon and the rules into a finite-state transducer (FST), relying on Kaplan and Kay (1994), which allows us to obtain the best derivation using dynamic programming.

5 Simulation: Turkish

The dataset shows a pattern modeled after front-back VH in Turkish. The learner's challenge is to learn both a lexicon of URs and the phonology of VH. The data consisted of 80 words, created by taking all combinations of 10 monosyllabic Turkish nouns (kent, jɪl, ek, güz, gün, kük, kalp, renk, saat, tuz) and 8 Turkish nominal suffixes (-ler/-lar, -in/-ɪn, -i/-ɪ, -lik/-lɪk, -siz/-sɪz, -

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

Notation	Representation
Textbook	$[-cons] \rightarrow [-back] / \_\_ [+cons]^* [-cons, -back]$ (optional)
String	$-cons\#_{rc} - back\#_{rc}\#_{rc} + cons * \#_b - cons\#_f - back\#_{rc}1\#_{rc}$
Bit string	$\underbrace{0100}_{-}\underbrace{0110}_{cons}\underbrace{0010}_{\#_{rc}}\underbrace{0100}_{-}\underbrace{1001}_{back}\underbrace{0010}_{\#_{rc}}\underbrace{0010}_{\#_{rc}}\underbrace{0011}_{+}\underbrace{0110}_{cons}$ $\underbrace{0101}_{*}\underbrace{0001}_{\#_b}\underbrace{0100}_{-}\underbrace{0110}_{cons}\underbrace{0000}_{\#_f}\underbrace{0100}_{-}\underbrace{1001}_{back}\underbrace{0010}_{\#_{rc}}\underbrace{1}_{1}\underbrace{0010}_{\#_{rc}}$

Table 2: A phonological rule for vowel harmony stated in textbook notation (top), string notation (middle), and as a bit string (bottom). To ensure unique readability, we use various delimiters to mark the end of the description of features, feature bundles, and the rule's components.

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

Table 3: Conversion table for phonological rules. The representation scheme used here treats all possible outcomes at any particular choice point as equally easy to encode: 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.

Grammar:

1. Rule: $[+syll] \rightarrow [+back]$
   $/ \begin{bmatrix} +back \\ +cont \end{bmatrix} [-back]^* \_\_$ (obligatory)
2. Lexicon: Stems = {kent, jɪl, güz, tuz . . . };
   Suffixes = {ler, in, ten, siz, . . . }

Figure 3: Final state

sün/-sun, -ten/-tan, -∅), and applying VH. The words were presented to the learner as unsegmented strings, without any morphological information (e.g., [kentler], [jɪllar]). Search was performed using a Genetic Algorithm (Holland 1975), as described in Lan 2018. The parameters of the search procedure were set as follows: crossover rate $= 0.2$, mutation rate $= 0.8$, overall population size $= 250,000$, number of islands $= 500$, island population size $= 500$, total generations $= 5,000$. The hypothesis space consisted of grammars with a lexicon (represented as a Hidden Markov Model) and a set of ordered rules. The search converged on the hypothesis in Fig. 3, which includes the VH rule and a segmented lexicon: the VH rule applies in all appropriate places (matching vowels in the data were all separated by $[-back]$ segments, so the induced rule is equivalent to the expected VH rule); in the lexicon, each pair (e.g., -ler/-lar) is correctly represented with a single UR.

6 Implications

Our results, while preliminary and obtained only for a very small corpus, suggest that morpho-phonological patterns involving VH can be learned using MDL from distributional evidence alone. They provide further support for the MDL metric, which is very general and is not designed with VH (or even with phonology) in mind. While VH seems to pose particular challenges to learning (and is indeed problematic for various proposals in the literature), it is straightforwardly learned using MDL: the same general metric that supports segmentation alone in simple cases – and that has been argued elsewhere to allow for the induction of phenomena such as optionality and opacity (Rasin et al. 2018a,b) and abstract URs (Rasin and Katzir 2018) – allows us to handle the challenge of jointly acquiring segmentation and a pattern of VH.

References

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

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

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

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.

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

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

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

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

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.

Lan, Nur. 2018. Learning morpho-phonology using the Minimum Description Length principle and a genetic algorithm. Master's thesis, Tel Aviv University.

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

Merchant, Nazarré Nathaniel. 2008. Discovering underlying forms: Contrast pairs and ranking. Doctoral Dissertation, Rutgers, The State University of New Jersey.

Rasin, Ezer, Iddo Berger, Nur Lan, and Roni Katzir. 2018a. Learning phonological optionality and opacity from distributional evidence. In *Proceedings of NELS 48 (to appear)*, ed. Sherry Hucklebridge and Max Nelson.

Rasin, Ezer, Iddo Berger, Nur Lan, and Roni Katzir. 2018b. Learning rule-based morphophonology. Ms., MIT and Tel Aviv University, http://ling.auf.net/lingbuzz/003665, February 2018.

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

Rasin, Ezer, and Roni Katzir. 2018. Learning abstract URs from distributional evidence. In *Proceedings of NELS 48 (to appear)*, ed. Sherry Hucklebridge and Max Nelson.

Rissanen, Jorma. 1978. Modeling by shortest data description. *Automatica* 14:465–471.

Solomonoff, Ray J. 1964. A formal theory of inductive inference, parts I and II. *Information and Control* 7:1–22, 224–254.

Tesar, Bruce. 2014. *Output-driven phonology: Theory and learning*. Cambridge University Press.