Extending adaptor grammars to learn phonological alternations

Canaan Breiss
University of California, Los Angeles
cbreiss@ucla.edu

Colin Wilson
Johns Hopkins University
colin@cogsci.jhu.edu

1 Overview

Recent advances in unsupervised learning of linguistic structure have demonstrated the feasibility of inferring latent morphological parses from an unannotated corpus given transparent underlying-to-surface mappings (ex., Adaptor Grammars (AGs); (Johnson et al., 2007; Johnson and Goldwater, 2009), as well as in learning predictable phonological transformations from prespecified underlying morphemes to a range of surface allomorphs via a stochastic edit distance algorithm (Cotterell et al., 2015). In this paper we introduce a nonparametric Bayesian model which builds on the morpheme-segmentation success of AGs, and incorporates the ability to learn predictable phonological transformations of underlying forms to their surface allomorphs via the interaction of markedness and faithfulness principles, inspired by generative phonology. The unsupervised nature of this model (that is, no semantic information about the words being segmented is provided) is relevant not only computationally but also psychologically, as it mirrors developmental findings (Kim, 2015) that young infants segment and cluster morphemes based solely on phonetic and distributional similarity. The model also incorporates many of the other cognitive restrictions infants during the initial period of morphophonological learning in an effort to make the model maximally realistic, and thus eventually useful in making quantitative predictions about the early stages of morphophonological acquisition that can be experimentally investigated. We evaluate the model on a novel dataset consisting of a complex system of allomorphy in Acehnese, an understudied Indonesian language.

2 Model design

The model takes the general structure of a (relatively shallow) AG with rewrite rules Word $\rightarrow$ Morph(s), Morph $\rightarrow$ Phoneme(s). The model differs, however, in that it considers whether a possible novel morpheme could be derived from an existing item in the lexicon via a phonological transformation (at a cost), as well as reused directly (if it exactly matches a lexical item) or generated anew. The parameterization of the penalty for non-identity transformations is informed by research demonstrating that infant and adult learners prefer phonetically-minimal alternations (ex., White (2013), cf. (Steriade, 2009) on the P-Map hypothesis more broadly), and that speakers are sensitive to the segment-to-segment transitional probabilities (cf. Vitevitch and Luce (2004)) of their native language(s). Thus, the probability of a novel morpheme being a transformation of an existing one is equal to the probability of the source morpheme in question being reused (as in a standard AG) multiplied by the penalty associated with a specific segment-to-segment mapping, operationalized as the number of phonological feature values by which the input and output segments differ ("faithfulness" to the input). This quantity is then multiplied by the probability of the surface string created through the unfaithful mapping, as calculated from the surface-distribution of phonemes in the unsegmented corpus (corresponding to a penalty for the "markedness" of the surface form), and the morpheme-length parameter λ. The faithfulness penalties on segment-to-segment transformations was equal to twice the featural edit distance between the two segments, and penalties for surface forms were calculated via segmental trigram probabilities of the corpus.

Proceedings of the Society for Computation in Linguistics (SCiL) 2020, pages 480-483.
New Orleans, Louisiana, January 2-5, 2020

2.1 Implementation

Unless otherwise noted below, the model was initialized with words parsed as monomorphemic roots, following the phonological acquisition literature which shows infants store unanalyzed chunks of their input during early learning (Ngon et al., 2013). Inference for all parameters was carried out via Gibbs sampling; the hyperparameters α and β, as well as the length penalty λ on morpheme lengths, were sampled using the slice-sampling technique from Neal et al. (2003), as implemented in Johnson and Goldwater (2009).

3 Data

We tested the model on a group of morphophonological alternations observed affecting labial-initial prefixes in Acehnese (Malayo-Polynesian, 3.5 million speakers, primarily in Indonesia). Two Acehnese verbal prefixes *peu-* /pɯ-/ and *meu-* /mɯ-/ exhibit allomorphy when prefixed to a base which begins with a labial consonant ({/p, b, m, w/}), surfacing as to [pu-] and [mu-] respectively with the back high unrounded vowel having undergone the phonological process *rounding*. A second process, *spirantization*, applies to the *peu-* prefix when the base to which it is attached begins with a labial consonant and is also polysyllabic, changing the initial consonant of the prefix from /p/ to [s], as in /pɯ-maŋat/ → [sɯmaŋat]. Further, spirantization *bleeds* rounding when the conditioning environments overlap, appearing to "apply" beforehand and so removing the environment (the labial onset of the prefix) which would have triggered rounding: /pɯ-maŋat/ → [sɯmaŋat], *[sumaŋat] (Durie, 1985). Thus, summarizing the data pattern, we find: /pɯ-/ → {[pɯ-, pu-, sɯ-]}, /mɯ-/ → {[mɯ-, mu-]}.

The use of Acehnese in evaluating the model is relevant for two reasons. First, there has been no known computational work on the language, nor even detailed quantitative study of the languages morphophonology. Therefore, the phenomena explored here (idealized based on corpus data gathered as part of Breiss, in prep.) provide a novel perspective on which to test traditionally English-centric tests of unsupervised learning of linguistic structure. Secondly, the specifics of the morphophonological alternations in the Acehnese data are typologically unusual, exhibiting processes which are both phonetically-motivated (rounding in the context of two labial segments) as well as phonetically arbitrary (spirantization). Prior research has shown that speakers may be biased towards learning and/or generalizing phonetically-natural patterns or processes more than phonetically-arbitrary ones; therefore, the trade-off in productivity between lexical listing and phonological derivation of allomorphs instantiated in the model can be used to make testable, quantitative predictions about human behavior.

	None	Half	All
Segmentation only			
Morpheme	1 / 0.45 / 0.62	1/ 0.87 / 0.93	1 / 1 / 1
Boundary	0.76 / 0.09 / 0.17	1 / 0.47 / 0.64	1 / 1 / 1
Source	(n/a)	(n/a)	(n/a)
Allomorphy only			
Morpheme	(n/a)	(n/a)	(n/a)
Boundary	(n/a)	(n/a)	(n/a)
Source	100% / 100 %	100% / 100 %	100% / 100 %
Both			
Morpheme	1 / 0.45 / 0.62	1 / 0.88 / 0.93	1 / 0.99 / 0.99
Boundary	0.76 / 0.09 / 0.17	1 / 0.51 / 0.67	1 / 0.98 / 0.99
Source	100% / 85%	100% / 70%	100% / 100 %

Figure 1: Evaluation statistics; each cell displays Precision / Recall / F-score for that combination of model settings and data.

4 Evaluation

F-score for identifying polymorphemic words, morpheme boundary F-score, and the percentage of surface allomorphs were derived from the correct underlying form (prefix and root) were calculated. We test each of the methods on a dataset consisting solely of polymorphemic words, a dataset with bare roots for 50% of the polymorphemic words, and a dataset with bare roots for all of the polymorphemic words (referred to as Zero, Half, and All respectively). Results are presented in 1, where each cell lists Precision / Recall / F-score.

4.1 Segmentation only

The first test is whether, under ideal conditions, the model correctly parses the data into its surface allomorphs. Disabling the option to consider non-faithful lexical reuse, the model is able to perform moderately well on segmenting the corpus. Since the Zero setting did not discover any segmentation

with words initialized as unanalyzed roots, random initialization was used for this condition only.

4.2 Allomorphy only

The phonological corollary to the morphological segmentation question is whether, under ideal conditions, the model can collapse the allomorphs of each morpheme into a single underlying representation. For this test, we gave the morphemic parse of each of the words in training, and then allowed the model to be informed by the faithfulness penalties as it discovered the most likely division between lexicalization and derivation for each of the allomorphs.

4.3 Simultaneous morphological segmentation and phonological abstraction over allomorphs

We test the model in a more realistic situation by asking it to discover the correct segmentation as well as the correct phonological alternations, and find that neither task is impaired when performed jointly with the other (in fact, in certain cases the performance is marginally improved; we take this as a suggestion that further scaling up of the model and dataset may give rise to more robust synergies; cf. Johnson (2008)).

5 Future work

While the model as presented here represents a significant step towards integrating insights from the developmental literature with computational methods of learning of linguistic structure from unlabelled data, it is hardly an adequate or complete model of early morphophonological acquisition. We see three main fronts along which the model can be improved: robustness to (more) naturalistic data, greater flexibility in non-faithful transformations to handle epenthesis and deletion phenomena, and the more robust integration of phonological principles to yield interpretable constraint-based grammars as part of the model yield.

In terms of data realism, the model can be improved so as to handle noisier, larger datasets: while the model does well given at least *some* bound-free pairs as evidence, not all languages allow roots to surface bare; thus improving the willingness of the model to consider morphological decomposition even in the absence of minimal pairs of bound-affixed forms is essential.

The linguistic validity of the range of hypotheses that the model considers can be enhanced by allowing it to consider strings of varying lengths as possible sources for non-faithful transformations. As implemented, the model only considers non-identity lexical sources for novel morphemes which are of the same length (in phonemes) as the novel morpheme under consideration. However, natural languages frequently exhibit deletion or epenthesis processes as part of morphophonological alternations (ex., the allomorphy of the English plural; /-z/ $\rightarrow$ {[-z, -s, -əz]}).

Two further improvements to the way that the model handles markedness and faithfulness penalties will allow the trained model to yield a grammar of weighted constraints, in addition to a lexicon and morphological parse, which can be compared to those which are the subject of analysis in other areas of generative phonology. On the faithfulness front, future experimental work can ground the specific penalties associated with non-identity transformations in data from confusability matrices, as in White (2017). These findings can be incorporated into the model by treating the phonetic distance between non-faithful mappings of segments as the mean of a Gaussian prior over possible penalties, rather than an absolute penalty itself. This will allow the model to deviate from the phonetically-informed priors in the face of compelling language-specific evidence for phonetically-unnatural alternations, mirroring the experimental findings of Wilson (2006).

The phonotactic markedness penalty given to surface forms can be enhanced by incorporating the ability to learn language-specific, feature-based phonotactic constraints from the already-segmented lexicon. This is motivated by the work ofHayes and Wilson (2008); Becker et al. (2011); Kager and Pater (2012); Hayes and White (2013); Rasin and Katzir (2016), among others, which shows that adult speakers internalize only a subset of available statistical generalizations latent in the data, informed by the statistics of the language and possibly prior grammatical knowledge. This constraint-based markedness penalty would replace the current phoneme trigram penalty over surface forms.

References

Michael Becker, Nihan Ketrez, and Andrew Nevins. 2011. The surfeit of the stimulus: Analytic biases

filter lexical statistics in turkish laryngeal alternations. *Language*, 87(1):84–125.

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

Mark Durie. 1985. *A grammar of Acehnese on the basis of a dialect of North Aceh*, volume 111. Mark Durie.

Bruce Hayes and James White. 2013. Phonological naturalness and phonotactic learning. *Linguistic Inquiry*, 44(1):45–75.

Bruce Hayes and Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic inquiry*, 39(3):379–440.

Mark Johnson. 2008. Using adaptor grammars to identify synergies in the unsupervised acquisition of linguistic structure. In *Proceedings of ACL-08: HLT*, pages 398–406.

Mark Johnson and Sharon Goldwater. 2009. Improving nonparameteric bayesian inference: experiments on unsupervised word segmentation with adaptor grammars. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 317–325. Association for Computational Linguistics.

Mark Johnson, Thomas L Griffiths, and Sharon Goldwater. 2007. Adaptor grammars: A framework for specifying compositional nonparametric bayesian models. In *Advances in neural information processing systems*, pages 641–648.

René Kager and Joe Pater. 2012. Phonotactics as phonology: Knowledge of a complex restriction in dutch. *Phonology*, 29(1):81–111.

Yun Jung Kim. 2015. *6-month-olds' segmentation and representation of morphologically complex words*. Ph.D. thesis, UCLA.

Radford M Neal et al. 2003. Slice sampling. *The annals of statistics*, 31(3):705–767.

Céline Ngon, Andrew Martin, Emmanuel Dupoux, Dominique Cabrol, Michel Dutat, and Sharon Peperkamp. 2013. (non) words,(non) words,(non) words: evidence for a protolexicon during the first year of life. *Developmental Science*, 16(1):24–34.

Ezer Rasin and Roni Katzir. 2016. On evaluation metrics in optimality theory. *Linguistic Inquiry*, 47(2):235–282.

Donca Steriade. 2009. The phonology of perceptibility effects: The p-map and its consequences for constraint organization. the nature of the word: essays in honor of paul kiparsky, ed. by kristin hanson and sharon inkelas.

Michael S Vitevitch and Paul A Luce. 2004. A web-based interface to calculate phonotactic probability for words and nonwords in english. *Behavior Research Methods, Instruments, & Computers*, 36(3):481–487.

James White. 2017. Accounting for the learnability of saltation in phonological theory: A maximum entropy model with a p-map bias. *Language*, 93(1):1–36.

James Clifford White. 2013. *Bias in phonological learning: Evidence from saltation*. Ph.D. thesis, UCLA.

Colin Wilson. 2006. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive science*, 30(5):945–982.