Allomorph discovery as a basis for learning alternations

Bruce Hayes
Department of Linguistics, UCLA

A well-known difficulty in the learning of morphophonemic alternations is the "vicious circle" created by the need to learn simultaneously both the grammar and the underlying forms. As Tesar (2014:§6.2) points out, the combination leads to a truly vast search space. Might there be a way to "break into the circle", obtaining early information that could ease the search? I suggest that there is, and it takes the form of finding the allomorphs of the alternating morphemes before the phonology is known.

Consider the invented data below. How do we know that 'wolf-ACC' is [ŋexe+xa] and not *[ŋexex+a]?

(1)

[kuŋanpa]	'turtle-NOM.'	[ruxiŋpa]	'dove-NOM.'	[tuɸærpa]	'fox-NOM.'
[kuŋanta]	'turtle-DAT.'	[ruxiŋta]	'dove-DAT.'	[tuɸærta]	'fox-DAT.'
[kuŋanka]	'turtle-ACC.'	[ruxiŋka]	'dove-ACC.'	[tuɸærka]	'fox-ACC.'
[piθoɸa]	'dog-NOM.'	[ŋexeɸa]	'wolf-NOM.'		
[piθoθa]	'dog-DAT.'	[ŋexeθa]	'wolf-DAT.'		
[piθoxa]	'dog-ACC.'	[ŋexexa]	'wolf-ACC.'		

One strategy is to detect that the language has a process of intervocalic spirantization, /ptk/ → [ɸθx] / V__V, so that [-xa] is simply the surface reflex of underlying /-ka/, observed unaltered when attached to consonant stems (top row). But this is the very strategy that, applied to harder problems, transports us into vast search spaces.

To take a different approach, observe that [ŋexe-xa] is compatible with a deeply sensible set of allomorphs for the data as a whole:

(2) Stems: [kuŋan], [ruxiŋ], [tuɸær], [piθo], [ŋexe]
Suffixes: [-pa ~ -ɸa], [-ta ~ -θa], [-ka ~ -xa]

The stems do not alternate, and the suffix allomorphs are closely similar, differing only in continuancy. The incorrect parse [ŋexex+a], to the contrary, leads us inexorably to a highly variegated allomorph list: putative "[ŋexex]" will have to alternate with some combination of [ŋexe], [ŋexeɸ], and/or [ŋexeθ] (C ~ ∅ alternation, or place alternation), and things will only get worse when we try to incorporate the remaining morphemes into the analysis.

I put forth the hypothesis that good morpheme parses reveal themselves even if we don't yet understand the phonology behind them. I have tested this hypothesis by implementing a supervised-learning system that inputs labeled paradigm data such as (1) and outputs a parse in which every segment is assigned to a morpheme. For example, for (1) the system assigns to the input {[ŋexexa], 'wolf'$_1$, ACC$_2$} the output parse {[ŋ$_1$e$_1$x$_1$e$_1$x$_2$a$_2$], 'wolf'$_1$, ACC$_2$}. The architecture of my system is that of a maxent OT grammar, with GEN encompassing all possible morphemic

Proceedings of the Society for Computation in Linguistics (SCiL) 2018, pages 215-216.
Salt Lake City, Utah, January 4-7, 2018

affiliations of all segments (including discontinuous ones, to handle metathesis and infixation). The constraints include widely-assumed principles of phonology and morphology:

(3) a. FAITHFULNESS: Prefer parses in which the allomorphs of a morpheme resemble one other (Kiparsky 1982, Benua 1995, Steriade 2000).
b. CONTIGUITY: Prefer parses in which the segments of an allomorph are adjacent (McCarthy and Prince 1995).
c. VARIEGATION: Prefer parses in which no single phoneme dominates beginning or endings of stems.

The output of the system is designated as the candidate assigned the highest probability; it is taken to be correct if it matches a handcrafted morphemic assignment provided by the author. The weights of the system's constraints were set to enable it to parse into allomorphs about 20 data sets, among them problem sets taken from Kenstowicz and Kisseberth (1979). The system is not perfect but does show substantial success in locating correct morphemic parses.

Once learned, a valid set of allomorphs can be of great help in the learning of alternations and underlying forms, the topic of a flourishing research literature (Tesar et al. 2003, Jarosz 2006, Apoussidou 2007, Merchant 2008, Pater et al. 2012, etc.). My suggestion is to employ string alignment based on phonetic similarity to establish correspondence between allomorphs, as in the following Polish example:

(4)
ʒ	w	u	p			'crib'
|	|	|	|			
ʒ	w	o	b	+	i	'crib-PLUR.

Free recombination of nonidentical segments (Tesar 2014) then determines a strictly limited candidate set for underlying forms; for Polish 'crib' this would be {/ʒwup, ʒwub, ʒwop, ʒwob/}. I believe that for purposes of locating underlying forms, a complete set of alignments like (4), suitably concatenated into word-candidates, provides a huge amount of useful information. From it can be extracted: (a) a full list of observed alternations, informing us about the ranking of Faithfulness constraints; (b) a sort of mini-GEN that necessarily includes the correct surface form as well as all the other candidates that are derivable from the candidate UR's under the set observed alternations. In the full version of this paper I demonstrate the usefulness of this information by using it to derive the correct UR's and rankings (of an externally-provided constraint set) for several of the Kenstowicz/Kisseberth problems.

In future work, the model must be scaled up so that it can project new paradigm members for "wug" stems (Albright and Hayes 2001; Cotterell et al. 2016, 2017). The key, I think, will be to employ more fine-grained Faithfulness constraints, weighting them in a way biased toward generality. Once the model has achieved wug testing capacity in this way, it will be testable against human intuitions.

Lastly, we should remember that not all alternation is due to underlying forms and phonology: we also need to discover how lexically-listed allomorphs get distributed, following phonological, morphosyntactic, and lexical constraints. The existence of a reliable allomorph set would, I believe, be essential in this task.