Generalizing patterns in Instrumented Item-and-Pattern Morphology

Sarah Beniamine & Olivier Bonami
Université Paris Diderot, Laboratoire de linguistique formelle

Recent years have seen a rise of interest in *Intrumented Item-and-Pattern morphology* (IIP), an approach to morphology charaterized by two main features:

Item and Pattern Morphology is modeled directly in terms of patterns of alternation relating surface words (Blevins, 2006; Ackerman, Blevins, and Malouf, 2009), rather than in terms of combinations of morphemes (Item and Arrangement) or of application of processes to stems (Item and Process).

Instrumented Morphological descriptions are based on computational implementations systematizing analytic procedures, applied to large datasets, typically inflected lexica (Albright, 2002; Stump and Finkel, 2013; Bonami and Boyé, 2014; Bonami and Luís, 2014)

The main success of this approach has been in providing a quantitative take on the value of implicative relations (Wurzel, 1989). Taking implicative relations between surface forms at face value shows that the apparent complexity of inflection system often has little consequences for speakers, because large and intricate paradigms are easy to deal with if they are highly predictable (see Ackerman and Malouf's (2013) *low conditional entropy conjecture*). It also shows the untenability of Albright's (2002) *single base hypothesis*, because simultaneous knowledge of multiple forms in a paradigm radically improves predictibility (Bonami and Beniamine, 2015); finally, it provides a clear motivation for the distribution of paradigmatic gaps in highly unpredictable cells (Sims, 2015). Complementarily to this line of work, IIPM has been used to extract new descriptive generalizations (Bonami and Boyé, 2014; Bonami and Luís, 2014) and inflectional classifications (Beniamine, Bonami, and Sagot, 2015).

At this point in the development of IIP, an important analytic and theoretical bottleneck is constituted by the lack of satisfactory systematic and cross-linguistically applicable strategies for the inference of patterns from raw data (Bonami, 2014). Previous work has relied either on hand-coded patterns (Ackerman, Blevins, and Malouf, 2009; Ackerman and Malouf, 2013) or on algorithms that presuppose some typological characteristics of the system at hand: purely suffixal morphology (Bonami and Boyé, 2014; Sims, 2015), absence of infixation or internal alternations (Albright, 2002), absence of prefixation (Bonami and Beniamine, 2015). The goal of the present work is to test a general purpose algorithm with no such strong bias on datasets whose morphophonological complexity goes beyond that of the Indo-European and Uralic languages examined in previous work.

Specifically, we elaborate on a strategy already hinted at by Albright and Hayes, 2006. Given a pair of paradigm cells, we first find, for each lexeme, the set of alignements that minimize the edit distance between the two forms, where the cost of substitution is weighted by the phonological similarity between segments (Frisch, Pierrehumbert, and Broe, 2004). Second, we infer the set of patterns relevant to the system by Minimal Generalization over the set of alignments. Finally, we choose the optimal pattern for each lexeme by optimizing on both the *accuracy* of the patterns (what proportion of the lexemes to which the pattern is applicable actually instantiate that pattern) and their *coverage* (to what proportion of the lexicon is the pattern applicable).

As a simple illustration of the virtues of that strategy, consider the toy example of a lexeme in an imaginary language with two forms *ba* and *baba*. There are three optimal alignments, as indicated in Table 1; these correspond to a prefixing, suffixing, or infixing pattern. Which of these patterns is the right one cannot be determined locally but only on the basis of examination of the rest of the language. Specifically, in the context of the data in Table 2, Pattern (i) is clearly optimal, with perfect reliability and perfect coverage.

We apply this approach to Zenzontepec Chatino data collected by Eric Campbell (Campbell, 2014) as provided by the *Oto-Manguean Inflectional Class Database* (Feist and Palancar, 2016). As Table 3 illustrates, Chatino conjugation involves both prefixation and tone alternations (see exemples (iv), (v), (vii), (viii)). Our algorithm correctly infers complex, non-affixal patterns. , e.g. example (iv) is characterized by a pattern $\ulcorner \text{kwi}^{1}\_^{2}\_ \rightleftharpoons \text{te}^{0}\_^{1}\_/\text{n\_}$`[-dist, +ant]`*`[-high]_?`$\urcorner$ relating the completive and the progressive.

	Alignment				Pattern
	b	a	b	a	
(i)	_	_	b	a	$\epsilon \rightleftharpoons$ ba/ _ab
(ii)	b	a	_	_	$\epsilon \rightleftharpoons$ ba/ ab_
(iii)	b	_	_	a	$\epsilon \rightleftharpoons$ ab/ b_a

Table 1: Alignments minimizing edit distance between *ba* and *baba*

SG	PL
to	bato
ri	bari
su	basu
ne	bane
ba	baba

Table 2: A toy dataset supporting a prefixing view of the relation between *ba* and *baba*

Class	CPL	POT	HAB	PROG	Translation
(i)	nkasesu	kisesu	ntisesu	ntesesu	'turn'
(ii)	nka^{1}ra^{2}	ku^{1}ra^{2}	ntu^{1}ra^{2}	nte^{1}ra^{2}	'hit'
(iii)	nkatehę^{1}	tyehę^{1}	ntyehę^{1}	ntetehę^{1}	'have diarrhea'
(iv)	nkwi^{1}so^{2}ʔ	kiso^{1}ʔ	ntiso^{1}ʔ	nteso^{1}ʔ	'pick'
(v)	nkuhna^{2}	kihna^{1}	ntihna^{1}	nte^{1}hna^{2}	'flee'
(vi)	nkutyehna^{1}	tyehna^{1}	ntyehna^{1}	ntetyehna^{1}	'start'
(vii)	nkya^{2}na^{1}	chana	nchana	nteya^{2}na^{1}	'wilt'
(viii)	ke^{2}ʔ	ka^{1}ke^{2}ʔ	nti^{1}ke^{2}ʔ	nchake^{1}ʔ	'cook'
(xix)	yaku	kaku	ntaku	nchaku	'eat'

Table 3: Sample of Zenzontepec Chatino conjugation classes (Feist and Palancar, 2016)

To evaluate the relevance of the patterns inferred by our algorithm, we proceed by 10-fold cross-validation of the prediction of which pattern is instantiated by which verb. What is of interest to us is not the raw accuracy, but the comparison of the accuracy attained using different pattern inference algorithms.[1], As the following Table shows, we observe no increase of accuracy due to the use of the new algorithm in the case of French. This is as expected, since the algorithm used by Bonami and Beniamine (2015) was tailored to address suffixing systems such as French. For Chatino on the other hand, we see a dramatic increase in accuracy. This suggests that the present proposal is on the right track and captures much of the structure of Zenzotepec Chatino conjugation patterns.

Algorithm	French	Z. Chatino
Non-prefixing algorithm	0.94	0.27
Current algorithm	0.94	0.62

Table 4: Accuracy of prediction of patterns using different pattern inference algorithms

In the talk we will illustrate how the present strategy can be used to study both the implicative structure and the inflectional classification of Zenzontepec Chatino and other Oto-Manguean languages.

[1]Overall, the accuracy for Chatino is much lower than for French. This might be due to two independent reasons. First, the French dataset we used (Bonami, Caron, and Plancq, 2014) is an order of magnitude larger than the Chatino dataset: only very frequent verbs are documented for Chatino, and frequent verbs tend to show more irregularity. We thus expect that accuracy would rise substantially if the dataset was similar in size to the one used for French. Second, it is quite possible that the two systems contrast in predictibility, and that this is reflected in the accuracy difference.

References

Ackerman, Farrell, James P. Blevins, and Robert Malouf (2009). "Parts and wholes: implicative patterns in inflectional paradigms." In: *Analogy in Grammar*. Ed. by James P. Blevins and Juliette Blevins. Oxford: Oxford University Press, pp. 54–82.

Ackerman, Farrell and Robert Malouf (2013). "Morphological organization: the low conditional entropy conjecture." In: *Language* 89, pp. 429–464.

Albright, Adam C. (2002). "The Identification of Bases in Morphological Paradigms." PhD thesis. University of California, Los Angeles.

Albright, Adam and Bruce Hayes (2006). "Modeling productivity with the Gradual Learning Algorithm: the problem of accidentally exceptionless generalizations." In: *Gradience in Grammar: Generative Perspectives*. Ed. by Gisbert Fanselow et al. Oxford: Oxford University Press, pp. 185–204.

Beniamine, Sarah, Olivier Bonami, and Benoît Sagot (2015). "Information-theoretic inflectional classification." In: *First International Quantitative Morphology Meeting*. Belgrade.

Blevins, James P. (2006). "Word-based morphology." In: *Journal of Linguistics* 42, pp. 531–573.

Bonami, Olivier (2014). "La structure fine des paradigmes de flexion." Mémoire d'habilitation, Université Paris Diderot.

Bonami, Olivier and Sarah Beniamine (2015). "Implicative structure and joint predictiveness." In: *Word Structure and Word Usage. Proceedings of the NetWordS Final Conference*. Ed. by Vito Pirelli, Claudia Marzi, and Marcello Ferro.

Bonami, Olivier and Gilles Boyé (2014). "De formes en thèmes." In: *Foisonnements morphologiques. Etudes en hommage à Françoise Kerleroux*. Ed. by Florence Villoing, Sarah Leroy, and Sophie David. Presses Universitaires de Paris Ouest, pp. 17–45.

Bonami, Olivier, Gauthier Caron, and Clément Plancq (2014). "Construction d'un lexique flexionnel phonétisé libre du français." In: *Actes du quatrième Congrès Mondial de Linguistique Française*. Ed. by Franck Neveu et al., pp. 2583–2596.

Bonami, Olivier and Ana R. Luís (2014). "Sur la morphologie implicative dans la conjugaison du portugais : une étude quantitative." In: *Morphologie flexionnelle et dialectologie romane. Typologie(s) et modélisation(s)*. Ed. by Jean-Léonard Léonard. Mémoires de la Société de Linguistique de Paris 22. Leuven: Peeters, pp. 111–151.

Campbell, Eric (2014). "Aspects of the phonology and morphology of Zenzontepec Chatino, a Zapotecan language of Oaxaca, Mexico." PhD thesis. University of Texas at Austin.

Feist, Timothy and Enrique L. Palancar (2016). *Oto-Manguean Inflectional Class Database*. Tech. rep. University of Surrey. DOI: http://dx.doi.org/10.15126/SMG.28/1.

Frisch, Stefan A., Janet B. Pierrehumbert, and Michael B. Broe (2004). "Similarity avoidance and the OCP." In: *Natural Language and Linguistic Theory* 22, pp. 179–228.

Sims, Andrea (2015). *Inflectional defectiveness*. Cambridge: Cambridge University Press.

Stump, Gregory T. and Raphael Finkel (2013). *Morphological Typology: From Word to Paradigm*. Cambridge: Cambridge University Press.

Wurzel, Wolfgang Ulrich (1989). *Inflectional Morphology and Naturalness*. Dordrecht: Kluwer.