When segmentation helps
Implicative structure and morph boundaries in the Navajo verb

Sacha Beniamine, U. Paris Diderot (sbeniamine@linguist.univ-paris-diderot.fr)
Olivier Bonami, U. Paris Diderot (olivier.bonami@linguist.univ-paris-diderot.fr)
Joyce McDonough, U. of Rochester (joyce.mcdonough@rochester.edu)

Recent work in Word and Paradigm morphology argues that the implicative structure of paradigms is expressed in terms of relations between surface words, and that studying the structure of paradigms in terms of sub-word units is misleading if not outright impossible (Ackerman et al, 2009; Blevins, 2006, 2016; Bonami & Beniamine, 2016). The argument typically rests on the observation that a word can only be segmented in the context of its paradigmatic alternatives, and that different aspects of the paradigm lead to different segmentations for the same word.

This line of argumentation amounts to a claim about the empirical properties of some inflection systems. It is thus entirely possible that systems differ in this respect. In this presentation we show that there are systems where a uniform segmentation is possible and helpful to addressing implicative structure. Interestingly though, the segments that are identified lack the properties of classical morphemes.

1. The Navajo verbal system

The Athabaskan languages represent a classic example of polysynthetic morphology. In polysynthetic languages word internal structure lacks the transparency of more agglutinative types; that is, the proposed morphemes are not easily separable or identifiable. In a commonly adopted model of the highly complex and synthetic Navajo (Athabaskan) verb, the verb consists of a series of prefixes attached to a rightmost and prominent stem. An extensive 'position class' template provides a prosthesis for ordering among these prefixes. Lists of the morphemes for each posited position are deconstructed from fully inflected word forms, and are understood to be sound-meaning pairs. Extensive rewrite rules are needed to recompose forms from the elements of this model.

There are many problems with this position class template, apart from its dependence on reverse engineering to derive the morphemes. Importantly, the template focuses exclusively on the morphemes and their relative order. In focusing thus, it misses many important generalizations about the structure of the verbal complex and, particularly, the relationship of words to each other, including the existence of conjugation patterns, inflectional paradigms, and the internal inflections of the stem itself. One alternate approach is to determine the actual forms in the complex that speakers may identify and use in word formation and in the organization of their lexicons. Based on the work of Young and Morgan (1980, 1987, 1992) (principle reference grammars of Navajo), McDonough and colleagues (1990, 1999, 2000, 2003, 2012, 2015; McDonough and Wood, 2008) have conducted a study of phonetic, phonotactic and phonological patterns in the verb across the Athabaskan languages and provided a consilience of arguments for the existence of two separable and independent but interdependent elements in the verb itself, identifying a minimal or 'core verb' obligatorily comprised of these elements, which carry the minimal morphosyntactic specification of a well formed verb. These are a *Mode* (*M*) element inflected for person and number, expressing the principle conjugational patterns of the verb, in penult position, and a monosyllabic *Stem* element (*S*), consisting of the 'classifier' (valence) plus 'stem shape' as the final and most prominent form in the verbal complex.

The final syllable in the verbal complex, the *Stem*, is phonetically prominent. The inflected *Mode* element on the other hand represents what Young and Morgan refer to as the *Base Paradigms*, a set of 4+ basic conjugations that all verbs are inflected in. This inflected *Mode* element may take a set of prefixal morphemes that serve to build up a very rich set of morpho-syntactic and -semantic meanings. These two elements, M(ode) and S(tem), represent independent yet inter-dependent dimensions of paradigmatic variation in the word.

(1) bidishne' ‘I break it off (by pounding on it).’ (d181)
bid -ish -ł-ne’
RED IPFV.1SG VL-stem.IPFV
Prefix Mode Stem

In Young and Morgan's dictionary, each fully inflected verb form is given in 5 principal parts carrying distinct TMA values, and exemplified below for the verb BIDISHNE’. Somewhat confusingly, these five TMA values are also called ‘modes’. Each verb involves a characteristic paradigm of mode elements and a characteristic paradigm of stem shapes. the Imperfective 1SG serves as the citation form.

TMA value	Mode	Stem	Surface form
Imperfective	bidish	(ł)ne’	bidishne'
Repetitive	bińdísh	(ł)niih	bińdíshniih
Perfective	bidíí	łne’	bidííłne'
Future	bidideesh	(ł)niił	bidideeshniił
Optative	bidósh	(ł)ne’	bidóshne'

Table 1 — Principal parts for the verb BIDISHNE’

The forms select each other to produce a rich set of aspectual meaning; these forms do not exhaust the possible combinations but act as principle components. Thus, although demonstrably independent, the inter-dependence of the M and S elements create word meaning; meaning is not compositional, but resides in the patterns of combinations of the elements in the verb. In this paper in particular we focus on the interdependence of the shapes of Stem and the Mode forms.

2. Data

Young & Morgan (1987) document the paradigms of Navaho verb in remarkable detail. Their dictionary presents information on inflectional paradigms in two guises. First, verb entries provide five principal parts for each lexeme, corresponding to five main TMA combinations. Table 2 exemplifies some entries. These are in the 1SG except for impersonal verbs which are given in the 3SG. The dictionary contains 5073 such entries.

Imperfective	Repetitive	Perfective	Future	Optative	Translation
niishkaał	náníishkał	niishkaal	dínéeshkał	nooshkaał	support by pushing on
'ałtániishkaał	'ałtánániishkał	'ałtániiłkaal	'ałtádínéeshkał	'ałtánooshkaał	chop or split lengthwise
niish'eeł	náníish'oł	nii'éél	dínéesh'oł	noosh'eeł	dissolve
niishdóóh	náníishdoh	niiłdoii	dínéeshdoh	nooshdóóh	heat
yiłdóóh	náłdoh	yiłdoii	doołdoh	wółdóóh	dry up during summer
yiłhę́ę́sh	náłhęsh	yíłhęęzh	doołhęsh	wółhę́ę́sh	move through the air (mushy matter)

Table 2 - Sample principal part series from Young & Morgan

Second, the dictionary contains person-number-TMA paradigms for all prefixes+mode combinations. For instance, paradigms for *niish-*, *'ałtániish-*, and *yił* , the three combinations exemplified above, are provided. A few hundred such paradigms are provided. Full paradigms for all lexemes can easily be deduced from verb entries and prefix+mode paradigms, since no sandhi phenomena or other unpredictable alternations occur at the mode-stem boundary (unlike what happens at the prefix-mode boundary). Unfortunately though, the prefix+mode paradigms are not available in digital form yet. Hence we leave the examination of full paradigms for a future study. For the present study we constructed a dataset on the basis of a digital version of Young & Morgan, which allowed easy semi-automatic tabulation of the five principal parts from each verb entry. After normalisation, error corrections, and exclusion of defective lexemes, we were left with a set of 1418 five cell (sub)paradigms. IPA transcriptions were deduced automatically for the orthographic forms, and mode-stem boundaries were introduced semi-automatically, by identifying and segmenting all consonant clusters occurring between the penultimate and final vowels. Table 3 exemplifies the results of these processes on the sample in Table 2.

Lexeme	Imperfective	Repetitive	Perfective	Future	Optative
NIISHKAAL	niːʃ+k͡xaːɬ	nániːʃ+k͡xaɬ	niːʃ+k͡xaːl	tínéeʃ+k͡xaɬ	noːʃ+k͡xaːɬ
'ALTANIISHKAAL	ʔaɬt͡xániːʃ+k͡xaːɬ	ʔaɬt͡xánániːʃ+k͡xaɬ	ʔaɬt͡xániː+ɬk͡xaːl	ʔaɬt͡xátínéeʃ+k͡xaɬ	ʔaɬt͡xánoːʃ+k͡xaːɬ
NIISH'EEL	niːʃ+ʔeːɬ	nániːʃ+ʔoɬ	niː+ʔéːl	tínéeʃ+ʔoɬ	noːʃ+ʔeːɬ
NIISHDOOH	niːʃ+tóːh	nániːʃ+toh	niː+ɬtoiː	tínéeʃ+toh	noːʃ+tóːh
YILDOOH	ji+ɬtóːh	ná+ɬtoh	ji+ɬtoiː	toː+ɬtoh	wó+ɬtóːh
YILHÉÉSH	ji+ɬhę́ːʃ	ná+ɬhęʃ	ni+ɬhęːʒ	toː+ɬhęʃ	wó+ɬhę́ːʃ

Table 3 - IPA version of the sample from Table 1, with mode-stem boundaries

3. Analysis

Building on previous work by Ackerman et al. (2009) and Ackerman and Malouf (2013), Bonami and Beniamine (2016) define implicative entropy as a way of assessing the predictability of one paradigm cell from any other collection of paradigm cells. Unlike previous attempts at using entropy to address predictibility, Bonami and Beniamine's algorithm does not presuppose a preexisting inflectional classification; hence it is readily applicable to new languages, as long as a large number of raw paradigms of surface forms is available.

We hence set out to apply Bonami and Beniamine's algorithm to the dataset derived from Young and Morgan (1987). When doing so, however, an immediate concern arose. At the heart of the algorithm is a generic method for inferring patterns of alternation from raw data. For the entropy calculations to be meaningful, it is essential that the patterns be as accurate as possible. Following Beniamine (2017), we assess the accuracy of a set of patterns by examining how well patterns extracted from a training set containting a random 90% of the data are used to predict the inflectional behavior of a test set consisting of the remaining 10% (using 10-fold cross-validation). Perfect accuracy is not expected (since the test set may contain inflectional behaviors not found in the rest of the system), but Beniamine's test cases lead us to expect accuracies in the 0.6-0.95 range. As it happens, the results were terribly bad: the accuracy of the patterns inferred from the raw dataset is only 0.28.

There is a rather direct explanation for this situation. As discussed earlier, the Mode and the Stem constitute independent dimensions of variation in the Navajo verb. This results in a combinatory explosion, where patterns of alternations between full forms involve a combination of a pattern of alternation for the mode and a pattern of alternation for the stem. Hence the test set is bound to contain many patterns not exemplified in the train set.

This combinatory explosion is a consequence of not taking advantage of the fact that every surface form of a verb can readily be segmented at the Mode-Stem boundary, as the stem coincides with the word's last syllable, a fact that speakers of the language are certainly attuned to. To assess the usefulness of this segmentation, we evaluated the very same algorithm on two datasets consisting of only the pre-stem material and only the stem. As indicated in Table 3, the results are then much more satisfactory, and stand in the range observed for other languages.

Dataset	Average accuracy	Average # of patterns
full words	0.33	537
pre-stem material	0.79	87
stems	0.75	102

Table 3 — Accuracy of pattern inference (10-fold cross-validation)

This result indicates that segmentability is a useful feature of the Navajo conjugation system. The problem faced by the pattern inference algorithm discussed above certainly is a problem also faced by speakers learning the language: if they did not rely on segmentation, they would be bound to make a spectacular number of errors.

Once patterns of alternation for stem and pre-stem material have been computed, we are now in a position to assess the implicative entropy of the system. We computed independently the implicative entropy of the system of stems, the system of pre-stem sequences, and the combination of both (taking the pattern relating two full words as the combination of a pre-stem pattern and a stem pattern). The results confirm that, despite its strong intricacies, the level of predictability of the Navajo conjugation system is manageable. Moreover, full words are more predictive than their parts: despite being partly independent, pre-stem sequences are partly predictive of stems, and vice-versa. This indicates that, while the identification of inflectional patterns in the Navajo verb require a segmentation, full words still are the best unit of prediction in such a system.

Dataset	Entropy
full words	0.33
pre-stem material	0.53
stems	0.62

Table 4 — Average unary implicative entropy

Selected References

Ackerman, Farrell, Jim Blevins, and Rob Malouf. 2009. Parts and wholes: implicative patterns in inflectional paradigms. In Jim Blevins and Juliette Blevins (eds.), *Analogy in Grammar*, pp. 54–82. Oxford: Oxford University Press.

Ackerman, Farrell, and Rob Malouf. 2013. Morphological organization: the low conditional entropy conjecture. *Language* 89: 429-464.

Beniamine, Sacha. 2017. Une approche universelle pour l'abstraction automatique d'alternances morphophonologiques. *Proceedings of TALN 2017*.

Blevins, Jim. 2006. Word-based morphology. *Journal of Linguistics* 42: 531–573.

Blevins, Jim. 2016. *Word and Paradigm Morphology*. Oxford: Oxford University Press.

Bonami, Olivier, and Sacha Beniamine. 2016. Joint Predictiveness in inflectional paradigms. *Word Structure* 9.2: 156-182.

Iskarous, K., J. M. McDonough, and D. H Whalen. 2012. A gestural account of velar contrast: the back fricatives in Navajo. *Laboratory Phonology* 3.1: 195-210.

McDonough, Joyce. 2003. *The Navajo Sound System*. Kluwer Academic Press.

McDonough, Joyce M. and V Wood. 2008. The stop contrasts of the Athabaskan languages. *Journal of Phonetics*. 36.3: 427-449.

Young, Robert W. and William Morgan. 1987. *The Navajo Language: A Grammar and Colloquial Dictionary*