A finite-state morphological analyzer for a Lakota precision grammar

Christian M. Curtis

University of Washington
Department of Linguistics
UBox 352425
Seattle WA 98195-2425
cmc3c@uw.edu

Abstract

This paper reports on the design and implementation of a morphophonological analyzer for Lakota, a member of the Siouan language family. The initial motivation for this work was to support development of a precision implemented grammar for Lakota on the basis of the LinGO Grammar Matrix. A finite-state transducer (FST) was developed to adapt Lakota's complex verbal morphology into a form directly usable as input to the Grammar Matrix-derived grammar. As the FST formalism can be applied in both directions, this approach also supports generative output of correct surface forms from the implemented grammar. This article describes the approach used to model Lakota verbal morphology using finite-state methods. It also discusses the results of developing a lexicon from existing text and evaluating its application to related but novel text. The analyzer presented here, along with its companion precision grammar, explores an approach that may have application in enabling machine translation for endangered and under-resourced languages.

Keywords: Morphology, Endangered languages, Finite-state methods

1. Introduction

Lakota is a Siouan language spoken, in conjunction with its closely-related[1] Eastern Dakota (Santee-Sisseton) and Western Dakota (Yankton-Yakntonai) dialects, by approximately 25,000 speakers in the United States and Canada (Lewis, Simons & Fenning 2013). Considered "critically endangered" (Moseley, 2010), Lakota has been the target of significant revitalization efforts over the past decade, resulting in the development of new resources for learners, teachers, and speakers. The New Lakota Dictionary (Ullrich, 2011) in particular has played a significant role in bringing together a common lexicon, a grammar for learners, and a practical, phonemically-consistent orthography.[2] This orthography, referred to as Standard Lakota Orthography (SLO), is used here.

Although new texts are now being created in Lakota, creating translation of existing texts remains prohibitively labor-intensive. Machine translation provides a potential avenue for at least partial automation of the process, as well as opportunities for further developing the vitality of the language (Gasser, 2006). Because Lakota lacks a significant corpus of parallel text suitable for training of statistical machine translation (MT) systems, I explored the creation of a precision implemented grammar[3] (Curtis and McHugh, 2013) in the HPSG framework (Pollard and Sag, 1994). One of the advantages of this approach is the potential for other applications, such as computer-assisted language education (Flickinger, 2011; Schneider and McCoy, 1998), which may be especially helpful in supporting endangered language communities. To bootstrap development, the initial grammar was based on the LinGO Grammar Matrix framework (Bender et al., 2002) and its customization system (Bender et al., 2010).

While the extensions by Goodman (2013) to the Grammar Matrix customization system support many morphological rules, Lakota's verbal morphology is not strictly concatenative and so could not be modeled directly within the system. The current work aims to address this limitation by mapping Lakota surface forms to and from the grammar's lexical forms using finite-state transducer techniques. In this paper I first outline the key morphosyntactic and morphophonological phenomena of Lakota. I then describe the design and implementation of the finite state transducer. Finally, I present some evaluation results of the transducer on Lakota texts, and outline future work.

2. Lakota morphology

2.1. Affixes on the verb

The verb is the only element required to form a valid sentence in Lakota[4], and is also the most complex element. Verbs obligatorily carry from zero to two pronominal affixes (with or without overt coreferent nominals in the sentence) indicating the participants. These affixes are grouped into two basic sets: one used to represent an patientive role and one representing an agentive role (Van Valin, 1977).

Lakota verbs are divided into two principal classes: stative and active. The stative verbs generally mark the subject

[1] Ullrich (2011) reports that 73% of words are identical in all three dialects, with only 8% distinct in all three; the grammar is homogeneous.

[2] Lakota orthography has been historically contentious; Powers (1990) notes that the historical orthographies were influenced by the differing political and phonological biases of the missionaries, scholars, and linguists of the times.

[3] An early version of this grammar is available as part of CoLLAGE, `http://www.delph-in.net/matrix/language-collage/` (Bender, 2014).

[4] The grammatical analysis presented here generally follows Rood and Taylor (1996) unless otherwise noted.

	prefixing	*infixing*	*mixed*
1SG.AGT	***wa**hí*	*slol**wá**ye*	*o**wá**le*
1DU.AGT	***uŋ**hí*	*slol'**úŋ**ye*	***uŋk**óle*
	hí	*slolyÁ*	*olé*
	'come here'	*'know'*	*'look for'*

Table 1: Affixation patterns

using the patientive affixes, while the active verbs do so using the agentive affixes (and mark the transitive object with the patientive affixes). An additional, semantically-limited class of impersonal verbs take no affixes. The active verbs are also broadly grouped into three classes, partially predicted by the phonology of the stem, that exhibit distinct paradigms in the agentive affixes. A small number of verbs are partly or completely irregular.

In addition, verbs exhibit one of three affixation patterns: prefixing, infixing, and "mixed." The mixed pattern is primarily infixing, with the exception of the first person non-singular, which is prefixing. These patterns are illustrated in Table 1. The affixation pattern of a verb can frequently be predicted by its derivation, but many verbs are not synchronically analyzable or are otherwise unpredictable; the pattern is therefore generally regarded as lexically-specified.

For transitive verbs, the affixes generally occur together in the appropriate infix or prefix position and ordered with the patientive before the agentive. Transitive verbs of the mixed-position paradigm when both affixes are present are a special case: the agentive *uŋ(k)-* should be prefixed and the patientive *wičha-* infixed, which conflicts with the patientive-first ordering. In this case, the affixes are instead combined and infixed: *owíčhauŋkole* ('you (sg.) and I looked for them') instead of the expected **uŋkówičhale* .

The third person is marked by a zero affix in all instances except for the collective animate plural; the inanimate plural is marked by restricted reduplication. In addition, the plural is marked in all persons by the suffix *-pi*, and is ambiguous between patient and agent.

The affix paradigms for both stative and active verbs are summarized in Table 2.

2.2. Morphophonology

Lakota also exhibits several important morphophonological changes in the standard orthography that must be taken into consideration.

Ablaut

Ablaut is pervasive on Lakota verbs. Verbs which undergo ablaut are identified in citation form by a final uppercase *-A*. The vowel to be used is determined by what follows the verb: *-e* in sentence-final position, *-iŋ* before certain enclitics, and *-a* otherwise.

Velar fronting

The velar plosives /k k^x k'/ (*k kȟ k'*) are fronted to /tʃ $tʃ^h$ tʃ'/ (*č čh č'*) when preceded by the front vowel /i/.

Nasalization spread

Nasalization is spread to *ya*/*yi*/*yu*, *ha*/*hi*/*hu*, and *wa*/*wi*/*wu* when preceded by a nasal vowel /ã ĩ ũ/ (*aŋ iŋ uŋ*). For example, *yá* 'to go there' becomes *uŋyáŋpi* 'we go there'.

Stress

Stress in Lakota is contrastive and subject to morphologic shift. Words with second-syllable stress, when prefixed, retain stress on the (new) second syllable (*ičháǧe* → *imáčhaǧe*), while words with first-syllable stress generally retain it (*ípuze* → *ímapuze*).

3. FST implementation

This analyzer is implemented using XFST, the popular Xerox finite-state toolchain (Beesley and Karttunen, 2003). A **lexc** transducer is used to map the abstract tokens of the grammar implementation into an intermediate regular language, while an **xfst** transducer implements replacement rules to map the intermediate language to surface strings.

The finite-state formalism allows these two transducers to be composed into a single network, as well as permitting it to operate in both then analysis (look-up) and generation (look-down) directions (see Figure 1).

The upper-side, "lexical" language is designed to interface cleanly with the Grammar Matrix morphology library (Goodman, 2013) by expressing words in their lexical stem form, followed by position classes for morphosyntactic pseudo-suffixes, e.g. `olé+3SgPat+1SgAgt` ('I look for it') corresponds to the lower-side, surface form *owále* (*o-Ø-wa-lé*).

The FST lexical entries include part-of-speech information, which can be included as labels in the upper-side language. However, the HPSG grammar implementation uses typed feature structures (Copestake, 2002) for its lexical entries and ignores these labels.

3.1. Affix position

To implement the different affix positional paradigms, the **lexc** lexical entries insert the placeholder symbols '`P`' (for *Patientive*) and '`G`' (for *aGentive*) in the appropriate location on the lower-side language. '`A`' is used, as in the citation form, to represent the final vowel for verbs that undergo ablaut:

`haŋskÁ`	`PhaŋskÁ`	stative, prefixing
`ípuzA`	`íPpuzA`	stative, infixing
`slolyÁ`	`slolPGyÁ`	active/transitive, infixing

For active verbs, the lexical entry also adds a lower-side token to express the inflection class (e.g. `^WaPrefix`, `^YStem`, etc.).

The first **xfst** replace rules, in turn, substitute the P and G placeholders with the appropriate affix form, selecting based on inflection class for active verbs. Subsequent replace rules implement the morphophonological changes such as nasalization spread and velar fronting. A final set of **xfst** replace rules shift stress to the second syllable when necessary.

Number/Person		Stative or Patientive	Active Agentive		
			Class I (*wa*/*ya*)	Class II (*y*-stem	Class III (nasal)
	1	*ma-*	*wa-*	*b-*	*m-*
Singular	2	*ni-*	*ya-*	*l-*	*n-*
	3	*Ø*			
Dual	1	*uŋ(k)-*			
	1	*uŋ(k)- … -pi*			
Plural	2	*ya- … -pi*			
	3	*-pi*			
	3 (collective)	*wičha-*	*a-/wičha-*	*a-/wičha-*	*wičha-*

Table 2: Pronominal affix paradigms

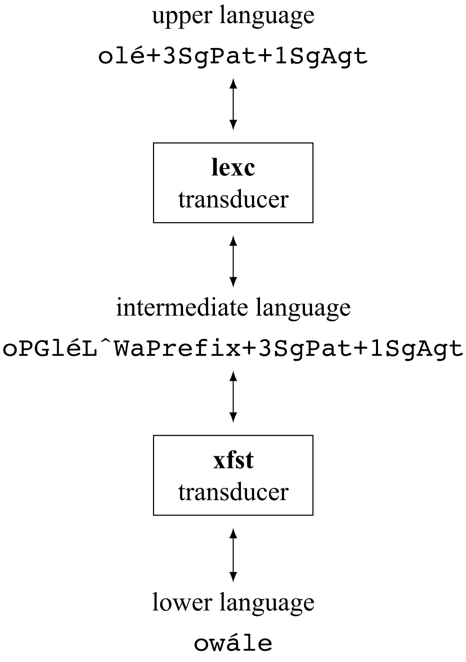

Figure 1: FST composition example

3.1.1. Mixed-position affix paradigm

Although the placeholder-replacement strategy worked well in the general case, the replacement rules were not expressive enough to resolve the true long-distance dependency of affix position on the pronominal pseudo suffix. To address this problem, I applied the unifying flag diacritics developed by Beesley and Karttunen (2003). A unifying flag diacritic takes the form `@U.Feature.Value@` and succeeds if `Feature` is either set to `Value` or unset (in which case it also sets `Feature` to `Value`); it fails otherwise, blocking that path.

For each verb of the mixed-position paradigm, I created two lexical entries in **lexc**: one with the placeholder(s) in prefix position, and one with them in infix position. The lower-side **lexc** output for each entry also included a unifying flag diacritic, either `@U.Mix.Prefix@` or `@U.Mix.Infix@` as appropriate. The lexical entries for the person/number pseudo suffixes also included flag diacritics: `@U.Mix.Prefix@` for `+1DuAgt` and `+1PlAgt`, and `@U.Mix.Infix@` otherwise. The behavior of the unifying flag diacritic ensures that only the appropriate paths can be followed.

A simplified fragment of the **lexc** lexicon illustrating the use of flag diacritics is shown in Figure 2.

4. Evaluation and future work

The initial version of the analyzer, developed primarily for testing the HPSG grammar, included lexical entries for 58 verbs, 30 nouns, and 4 articles and enclitics. In this form it produced the correct surface form for all 167 grammatically-correct cases in the precision grammar test suite.

To create a realistic test corpus for the analyzer, I transcribed 28 sentences (631 words) from stories originally collected by Deloria (1932) into the standard orthography. I extracted lexemes by cross-reference to Ullrich (2011), and added them to the **lexc** lexicon. The significantly-expanded lexicon was then applied to the source text to confirm full and correct analysis of all input words. Detailed metrics for the analyzer are shown in Table 3.

	Baseline	Expanded
Verbs	58	195
Nouns	30	79
Other	4	357
Total lexical entries	92	631
FST size	211.5K	423.9K
FST states	3160	6152

Table 3: FST metrics

I also transcribed an additional 15 sentences as held-out test cases. Without adding new lexical entries, the transducer correctly analyzed 156 of 264 words (59.1%). Although this is not an especially high accuracy, it should be noted that this was achieved purely by adding to the lexicon—no morphophonological rules were added or expanded. The failed words represented 87 distinct lexemes, over half of which (56.3%) were nouns and other types which exhibit no morphology. This suggests that significant improvements are available through simple data capture without additional analysis.

```
LEXICON Verb
  <o 0:P 0:G l é 0:L 0:%^ WaPrefix ”@U.Mix.Infix@”> PersNumAgt ;
  <0:P 0:G o l é 0:L 0:%^ WaPrefix ”@U.Mix.Prefix@”> PersNumAgt ;

LEXICON PersNumAgt
  <%+1SgAgt ”@U.Mix.Infix@”> #;
  <%+1DuAgt ”@U.Mix.Prefix@”> #;
  <%+1PlAgt ”@U.Mix.Prefix@”> #;
  <%+2PlAgt ”@U.Mix.Infix@”> #;
```

Figure 2: **lexc** lexicon fragment

Future work

As noted above, increasing the analyzer coverage may largely be achieved without significant additional analysis. With the aid of some simple tooling, new entries could be added by users without a linguistics background. This approach could be expanded to engage members of a language community in developing similar resources.

Also with respect to tooling, a major limitation of the current approach is that the morphological analyzer and HPSG grammar do not share common lexicon source files. The development of tools to produce both files from a single master file would greatly simplify developing grammars for similar morphologically-complex languages.

Finally, there are additional morphological phenomena that could be analyzed to expand coverage non-lexically. For example, reduplication is modeled here as lexically-specified, but there is some degree of productive reduplication in actual usage. There are also many productive affixes (e.g., instrumental prefixes) that could be implemented independently to support derivational analysis.

5. References

K. R. Beesley and L. Karttunen. 2003. *Finite state morphology*. CSLI Publications, Stanford, Calif.

E. M. Bender, D. Flickinger, and S. Oepen. 2002. The grammar matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In John Carroll, Nelleke Oostdijk, and Richard Sutcliffe, editors, *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, pages 8–14, Taipei, Taiwan.

E. Bender, S. Drellishak, A. Fokkens, L. Poulson, and S. Saleem. 2010. Grammar customization. *Research on Language & Computation*, 8(1):23–72.

E. Bender. 2014. Language CoLLAGE: Grammatical Description with the LinGO Grammar Matrix. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA). To appear.

A. Copestake. 2002. *Implementing Typed Feature Structure Grammars*. CSLI Publications, Stanford.

C. M. Curtis and D. S. McHugh. 2013. A precision implemented HPSG grammar for Lakota. Unpublished work.

E. C. Deloria. 1932. *Dakota Texts*. U of Nebraska Press.

D. Flickinger. 2011. Prescription and Explanation – Using an HPSG implementation to teach writing skills. Invited talk at the HPSG Conference 2010, Paris, France.

M. Gasser, 2006. *Machine Translation and the Future of Indigenous Languages*. I Congreso Internacional de las Lenguas y Literaturas Indoamericanas, Temuco, Chile, 10.

M. W. Goodman. 2013. Generation of machine-readable morphological rules from human-readable input. *Seattle: University of Washington Working Papers in Linguistics*, 30.

C. Moseley, editor. 2010. *Atlas of the World's Languages in Danger*. UNESCO Publishing, Paris, 3rd edn. edition.

C. Pollard and I.A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Studies in Contemporary Linguistics. University of Chicago Press.

W. K. Powers. 1990. Comment on the politics of orthography. *American Anthropologist*, 92(2):496–498.

D. S. Rood and A. R. Taylor. 1996. Sketch of lakhota, a siouan language. *Handbook of North American Indians*, 17:440–82.

D. Schneider and K. F. McCoy, 1998. *Recognizing syntactic errors in the writing of second language learners*, pages 1198–1204.

J. Ullrich. 2011. *New Lakota Dictionary*. Lakota Language Consortium, Bloomington, IN, 2nd ed. edition.

R. D. Van Valin. 1977. *Aspects of Lakhota Syntax: A Study of Lakhota (Teton Dakota) Syntax and Its Implications for Universal Grammar*. Ph.D. thesis, University of California.