Cheyenne Finite State Transducer

Matt Luettgen
Indiana University, Department of Linguistics
1020 East Kirkwood Ave
Bloomington, IN 47405
mrluettg@iu.edu

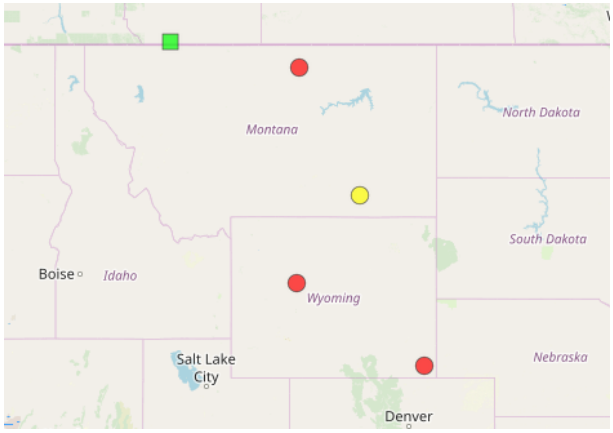


Figure 1: Cheyenne (Yellow) and the surrounding Algonquian languages: Blackfoot (green) and Plains Cree (red). (Hammarström, 2018)

Abstract

This paper describes a finite state transducer made to describe the Cheyenne Language. It provides background about the Cheyenne Language and about Finite State Transducers, describes the issues of creating an finite state transducer for Cheyenne, then

1 Introduction

Cheyenne is a language of the Algonquian language family spoken in Montana. Its polysynthetic nature causes analyzing words without knowledge of morphology to be exceptionally difficult. The purpose of the project is to create a Finite State Transducer for Cheyenne in order to generate and analyze verb forms. This is helpful in downstream applications such as search engines, machine translation, and spell checkers.

2 About the Cheyenne Language

The Cheyenne Language is spoken by about 1,200 speakers living in the Northern Cheyenne Indian Reservation in Montana and by a smaller population in Oklahoma. (Fisher, 2017) UNESCO'S *Atlas of the World's Language's in danger* ranks Cheyenne as "definitely endangered" (Christopher Moseley, 2010). Furthermore, the average age of the youngest Cheyenne speakers is about fifty (Fisher, 2017). Cheyenne is a member of the Algonquian language family, meaning that it has relations to languages such as Cree, Ojibwe, and Blackfoot (Fisher, 2017). Cheyenne does not occupy a particular subgroup of the Algonquian subgroup, though multilingualism most likely lent to similarities between surrounding languages (Valentine, 2001). Making Cheyenne distinct from most other Algonquian languages is that it has a full tonal system, consisting of neutral, high, mid, low, hanging-low and raised-high pitches, and vowels may become voiceless as well. (Leman, 2011).

3 Cheyenne Morphology

3.1 Nouns

Nouns are simpler than verbs, though still feature many morphological elements.

3.1.1 Animacy

All Cheyenne nouns fall into two categories of animacy. Most people, animals, trees, and spirits are animate nouns, while most objects fall are inanimate (Leman, 2011). However, there is no predictable pattern to determine animacy, as many some linguists compare animacy to gender in Indo-European languages (Curriculum, 2002). For example, "henene" (tomato) is animate, while "mene" (berry) is inanimate, despite both being edible plants (Leman, 2011).

3.1.2 Plural

There are four different plural morphemes: -e and -o for animate nouns and -ėstse and -ôtse for inanimate nouns. Cheyenne nouns are similar to many

Singular	Abstract	Plural	Gloss
kosa	kosán	kósáne	Sheep (an.)
hesta	hestah	hestahótse	Heart (inan.)

Table 1: Singular, Abstract, and Plural Noun Examples

french words in that the final consonant is not pronounced in the singular but is still present in the plural form. However, Cheyenne orthography does not write this final consonant in the singular (this caused complications later). In describing a system to retrieve the plural form of the noun, Leman posits an abstract form for every noun which is necessary to find the obviative, plural, and several possesive suffixes.

3.1.3 Possessives

Possessive affixes take the form of a prefix if the possessor singular or a circumfix if the possessor is plural. For example “sémo” means “boat”, nàsémo means “my boat,” and “nàsémónáne” means “our (exclusive) boat.” The plural suffix comes after the possessive suffix, so “our (exclusive) boats” is pronounced “nàsémonanótse.”

3.2 Verbs

Algonquian languages are known for their complex polysynthetic verbs.

3.2.1 Structure

There are four verb categories in Cheyenne which depend on both animacy and and transitivity. Intransitive verbs must include a prefix indicating the subject while transitive verbs must contain both a subject-prefix and a object-suffix. Leman writes that the basic structure of a verb is prefix-(tense)-(directional)-(preverb)-root-(medial)-final (Leman, 2011). Though he neglects to write a full analysis of a verb that contains all of these affixes, a simple example is the word “ná-nèh-vóóm-o” can be analyzed as “I[prefix] far-past[tense] see[root] him[final].”

3.2.2 Negative

Verbs also inflect for a negative circumfix, -saa- -he. To negate the above word, one would say “ná-saa-nèh-vóóm-o-he.”

3.2.3 Other affixes

Cheyenne includes many other verb affixes that are far beyond the scope of this project. Leman organizes these into three orders which include sixteen modes (Leman, 2011). Although the project is not meant to include all of these orders and modes, interesting affixes can be found here such as the dubitative form. For example, “mó-ná-nemenė-hehe” means ”I must have sung,” as if questioning the certainty of the situation.

3.3 Obviative

The obviative allows Cheyenne speakers to say a sentence about two different things and then continue to speak about them without referencing them directly. For example, a Cheyenne speaker may say “Aénohe émévévo vóoheho.” This sentence translates to “Hawk-PROX eat-PROX-OBV Rabbit-OBV.” The speaker may continue to say “Étàhpeta,” meaning, “(the hawk) is big-PROX,” or they could say “Étàhpetáho,” meaning, “(the rabbit) is big-OBV.” Only the inflection needs to be said to understand whether the hawk or rabbit is big.

4 About Finite State Transducers

Finite State Transducers provide a means to generate and analyze all of the forms of the noun. Such transducers have may greatly help in the implementation of downstream applications such as machine translation, spell checkers, and search engines. In highly polysynthetic languages such as Cheyenne and the rest of the Algonquian languages, Finite State Transducers become essential for these sorts of applications. With such applications, it is hoped that the speakers of these languages will have these technologies at home and not have to rely on the English equivalents, thus helping the survival of native languages at home. Finite state transducers use two different formalisms, TWOL and LEXC.

4.1 LEXC

LEXC files allow the implementation for the generation of all of the possible different inflections of a single word. LEXC keep track of both the inflections of the word and the letters that make up the affixes themselves in a parallel structure. For a simple English example, a line generated from a LEXC FST in can be “like<past>:like>ed” indicating that the lemma is the word ”like” which is inflected with the past tense, generating like>ed in the surface form.

4.2 TWOL

The other formalism used in FSTs is a TWOL, which handles the phonological changes caused by the affixes appended to the stem in the LEXC file.

TWOL code contains phonological rules that are not sequential, but parallel. This means that instead of having to declare which rules apply before which, the transducer compiles all of them at the same time (Karttunen, 1993). An example of a TWOL rule may look like... This rule means to deleted the e in the context of e followed by a morpheme boundary followed by d. This will cause the appropriate form "liked"

5 Goals

The goal of this project is to start foundational work in establishing a Finite State Transducer for the Cheyenne Language. As this project was undertaken by one inexperienced undergrad over the course of six weeks, creating a transducer to encompass the entirety of the Cheyenne Language would be quite unrealistic. In fact, the minimum viable project, which is what the transducer ended up being coded for, only aimed to accomplish simpler noun morphology and basic intransitive verb morphology.

6 Methodology

6.1 Corpus

The Cheyenne Bible was used as a corpus for this project (Society). It was not a full bible, though contained most of the commonly used books such as The bible included 57,227 words total. Words were turned into lower case and any punctuation such as periods, parenthesis, and verse numbers were deleted.

6.2 Assembling Lexicon from Dictionary

First, words were extracted from the Cheyenne Dictionary-English dictionary found online (Fisher, 2017), and were sorted by part of speech via a python script. Plurals and obviative forms of the noun were extracted as well, as these tend to be irregular. Many Biblical names were found in the corpus and not in the dictionary. These were extracted with a python script that searched for words with letters not found in Cheyenne and excluded words with letters unique to Cheyenne. Table 2 presents the size of the data set used, organized by word category.

POS	Frequency
Verb: Animate Intransitive	7689
Verb: Animate Transitive	2830
Noun: Animate	2761
Noun: Inanimate	1952
Verb: Inanimate Intransitive	1735
Verb: Inanimate Transitive	1449
Noun (Obviative)	1218
Noun (Plural)	1217
Final-VAI	931
Particle	572
Preverb	539
Initial	525
Final-VII	180
Final-VIT	114
Medial Body Part	102
Voice	90
Suffix	87
Singular	74
Medial	39
Final-NI	22
Prefix	21
Final-NA	12
Conjunct	8
Theme	5
Total:	24466

Table 2: ...

6.3 Implementing the FST

6.4 Challenges

6.5 Limitations in Grammar

To my knowledge, Leman's "A Refrence Grammar of the Cheyenne Language" is the only grammar ever made for Cheyenne. Although it always effectively gives examples of verb forms, the underlying phonetic structures, analysis of long words with many morphemes, and grammatical explanations are often left unexplained. A new grammar would provide much help in creating a transducer.

6.5.1 Limited Access to Tones

Tone proved significant problems in implementing an FST. Typical Cheyenne orthography only writes a high tones, though occasionally writes mid tones. Often when looking through the forms of a noun, a high would appear out of nowhere, breaking all perceived rules, leaving the suspicion that unrepresented tone sandhi happens underneath the orthography. Leman describes several pitch rules that affect orthography but depend on tones unwritten in orthography, such as a high tone turning into a low tone between a high tone and a low tone.

6.5.2 Limited Access to Plural Nouns

The previously mentioned abstract forms of the noun that Leman posited had no entries in the dictionary. These forms are essential to find any form of the noun that includes a suffix, such as the plural form, the obviative, or the plural possessives. In the dictionary, many entries for the plural form of the noun were entered, and thus a python script was written to extract the abstract form of the noun from the plural by omitting plural morphemes. Although the dictionary included a total of 4713 noun entries, only 1217 abstract forms could be found by subtracting plural morphemes.

6.5.3 Lack of Explanation for the Obviative

Leman left little information in regards to how to form the obviative form of a noun from the singular or stem form. Obviation is very common in Algonquian languages, leaving a huge gap in the transducer. Only 1218 obviative nouns were found in the corpus.

6.6 Accomplishments in LEXC

For the LEXC file, I managed to code a way to create Leman's "abstract" morphological forms from the plural forms through python code, then implement these abstract forms into the LEXC file. Possessive prefixes were also accomplished for nouns. Furthermore, to increase the final coverage, I also implemented names and a way to inflect names into the obviative. Intransitive verbs of both animate and inanimate types were also all implemented for person, plurality, and negativity.

6.7 Accomplishments in TWOL

Here are a few of the TWOL rules that were implemented.

6.7.1 Deletion of final consonants

The key way to derive the singular of a noun from Leman's abstract form is to delete the final consonant. For example, the singular of "he'kon" is he'ko."

6.7.2 Neutralization of a high tone at the end of the word.

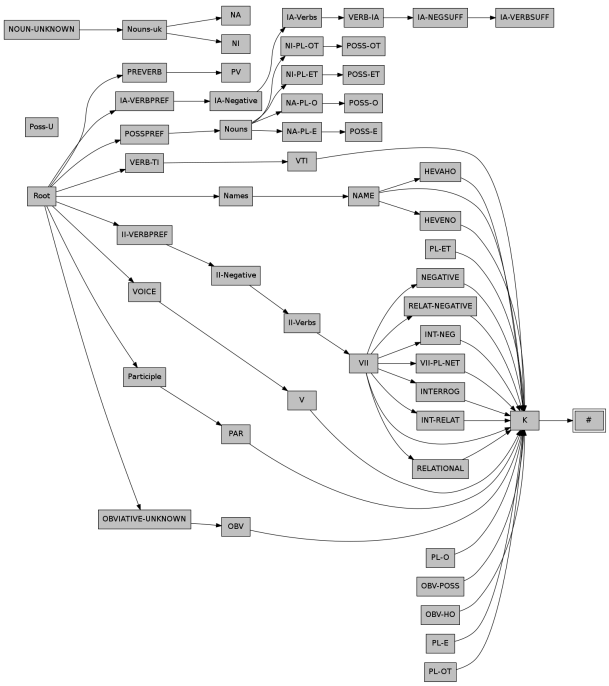
As part of natural tone patterns, all high tones are neutralized at the end of a word in Cheyenne. This was important in Leman's abstract forms. For example, the singular of "kosán" is "kosa," where the rule only applies after the n is deleted. This had to be accounted for in the code.

6.8 Teleporting Neutralized High Tone

The neutralization of a high tone in the previous rule may cause a another vowel that follows a high tone vowel to also gain a high tone. For example, when the abstract "kokohéaxán" forms the singular "kokóhéáxa," the "a" gains a high tone. Notice how this change is a result of three rules: the ending consonant is deleted, the ending high vowel is neutralized, then the "a" gains the high tone. Because of the nature of TWOL, all of these sequential rules had to be implemented non-sequentially, which is an accomplishment in itself.

6.8.1 Devoicing of consonants between certain consonants at certain positions.

Initially, a very complicated rule was implemented to account for vowels becoming voiceless before certain consonants, however, after the rule was written, it was found to decrease the coverage, and was thus deleted. It is clear that there are very complicated phonological rules happening here, and I leave it for future projects to accomplish.



A graphic of the LEXC file produced.

6.8.2 Repetition of Vowel with glottal stop after vowel.

If two vowels end a word, a glottal stop with the second vowel is appended to the end of the word. This happened often in generating the singular from the abstract, such as “xaón” to “xao’o” (the rule applies after n is deleted). This rule frequently applied when the animate plural affixes, -o and -e, were appended to an abstract form, such as nákohe-o to nákóheo’o.

7 Results

The FST was calculated on the Cheyenne Bible. The FST surpassed the goal of 40% to reach 46% naive coverage. The code is available at `https://github.com/mrluettg/cheyenne_fst`

8 Suggestions for Future Projects

- Investigate the obviative suffix for nouns in Cheyenne. This is an essential part of the language and a transducer that neglects it will have a gaping hole in it. Looking at other Algonquian languages might provide hints. In Nishnaabemwin, the form of the obviative directly corresponds to the plural form (Valentine, 2001). Arapaho is the same (Cowell and Moss, 2008).

- If anyone were continue the project beyond this point, I would suggest continuing to develop the the complex Cheyenne verb system. The second largest word category in the dictionary are transitive verbs, and they mostly lay untouched by the FST.

9 Conclusion

Although the Finite State Transducer did not perform well, this was really just meant to be a start of the project. To my knowledge, this has been the first time anyone has attempted to implement a Finite State Transducer for the Cheyenne Language. Finite State Transducers help in implementing downstream computational linguistics applications which speakers may be able to use in everyday life, preventing the loss of the language.

References

ed. Christopher Moseley. 2010. *Atlas of the World’s Languages in Danger*. Unesco.

Andrew Cowell and Alonzo Moss. 2008. *The Arapaho Language*. The University Press of Colorado, Boulder, Colorado.

The Ontario Curriculum. 2002. *Native Languages: Ojibwe and Cree*. Ontario Ministry of Education, Toronto Ontario.

Louise Fisher. 2017. *Cheyenne Dictionary*. Chief Dull Knife College, Lame Deer, Montana.

Robert Haspelmath Martin Hammarström, Harald Forkel. 2018.

Lauri Karttunen. 1993. *Finite-State Constraints*. Xerox Palo Alto Research Center, Center for the Study of Language and Information, Stanford University.

Wayne Leman. 2011. *A Reference Grammar of the Cheyenne Language*. Lulu Press, Busby, Montana.

The Bible Society. *Cheyenne Bible Scripture*.

J. Randolph Valentine. 2001. *Nishaabemwin Reference Grammar*.