James Pustejovsky Ben Rozonoyer
LING 190-B

Developing a Morphological Analyzer for Blackfoot Nouns

Introduction

As my final project for *The Lexicon* I furthered my ongoing work at implementing a morphological analyzer for Blackfoot nouns. Blackfoot is an agglutinative polysynthetic language with transparent morpheme boundaries and regular morphophonological processes. The phonological regularity exhibited by Blackfoot makes it amenable to modeling by FSTs because of their capacity to encode regular linguistic patterns (cf. Snoek et. al. 2014 analyzer of Plains Cree nouns [also an Algonquian language]). I take Frantz's (2017) *Blackfoot Grammar* (and *Dictionary*) as a reference for Blackfoot phonology, morphology, and standardized orthography. For a finite state framework I use the Helsinki Finite State Technology (HFST) implementation of the *lexc* language laid out in Beesley and Karttunen's *Finite State Morphology* (2003).

Basics of Blackfoot Noun Inflection

Blackfoot noun stems are bound morphemes specified for *animate* or *inanimate* gender (abbreviated *nan* and *nin*), as well as particularity. While there is some correlation between a noun's animacy and the real-world animacy of its referent (nouns for animals, persons and spirits are always animate), many nouns that don't designate an animate being can also be of animate gender, such as nouns for most metal tools (e.g. "pail", "knife", "bullet")—and gender must be memorized for each noun (much like gender in Indo-European languages).

Animate nouns may be either proximate (major third person) or obviative (minor third [4th] person)—whenever two third person nouns appear in a Blackfoot sentence, only one of them is allowed to be major third person to signify relative prominence, and the other nouns (including all possessed nouns) are demoted to minor third person. Additionally, Blackfoot has a single non-particular suffix *-i* to signal a noun that has no specific referent (e.g. as in "get me a dozen apples"); particular nouns, in contrast, are specified for both gender and number. Below is a summary of the basic Blackfoot noun endings:

non-particular/non-referring			*-i*	
particular		animate		inanimate
		3	4	
	sg	*-wa*	*-yi*	*-yi*
	pl	*-iksi*		*-istsi*

The lexical entries for the noun stems cannot be considered in isolation from the regular phonological rules that apply to them, primarily at morpheme boundaries. Non-permanent consonants are an instance of a lexically-encoded phonological trigger: a significant number of

noun stems that have an underlying stem-final *m*, *n* or *s* (represented as *M*, *N* or *S*) lose this stem-final consonant before suffixes that underlyingly start with a vowel. In particular, a non-permanent consonant is lost before the plural suffixes and the non-particular suffix, but remains in place otherwise. These respective cases provide distinct environments for consequent regular phonological rules to apply. The non-permanent status of these stem-final consonants isn't directly encoded in the *Blackfoot Dictionary* and must be inferred from the various forms; I encode this by directly capitalizing the stem-final consonant and having it trigger the corresponding NonPermanentConsonantLoss *xfst* rule.

An example to illustrate non-permanent consonants and rule-interaction is the animate stem *áwanááN* "rattle". Below are two basic phonological rules in Blackfoot (refer to Appendix B [pp. 176-179] of the *Grammar*):

(*i*) Vowel Shortening: $V_i: \rightarrow V_i$ / _+V (Rule 10)
(*ii*) Accent Spread: $V \rightarrow$ [+accent] / $V_{[+accent]}$+_ (Rule 26)

i.e., a long vowel gets shortened before another vowel, and an accented vowel causes the next vowel to be accented. We can see Accent Spread at work independently in a form such as /*matsiní+istsi*/ → [*matsiníistsi*] ("tongues"); and Vowel Shortening at work independently in /*niiítahtaa+istsi*/ → [*niiítahtaistsi*]. For stems that have an accented long stem-final vowel, applying Accent Spread first neatly provides an environment for Vowel Shortening, as in: /*kakkóó+iksi*/ –(*i*)→ *kakkóiksi* –(*ii*)→ [*kakkóiksi*]. The grammar doesn't comment on the relative ordering of non-permanent consonant loss and these two rules, but it can be inferred from the dictionary entry *awanáiksi/awanáániksi* "rattles". The first of these suggests that non-permanent consonant loss precedes both the above rules: /*áwanááN+iksi*/ –(C-Loss)→ *áwanááiksi* –(*i*)→ *áwanáiksi* –(*ii*)→ [*áwanáiksi*]; the second variant *awanáániksi* signals the presence of an alternative underlying form /*awanáán*/, without a non-permanent stem-final consonant.

The singular form of the same stem presents us with another rule, that of Semivowel Loss, formulated as {w,y} → 0 / {#,C} _ . The underlying non-permanent consonant in *áwanááN* triggers the loss of the suffix-initial glide *w*, causing /*áwanááN+wa*/ → *áwanááNa* = [*áwanáána*] "rattle". Semivowel Loss doesn't interact with either (*i*) or (*ii*), but we can tell from these two forms that it must precede non-permanent C loss, which in turn must precede (*i*) and (*ii*).

Implementing the Basics

To familiarize myself with *lexc* and *xfst* (finite-state computational analogues to a mental lexicon and finite-state phonological component) I referred to Beesley and Karttunen's invaluable *Finite State Morphology* (2003), which made me excited about the potential of finite-state methods in computational linguistic modeling.

My FST is the composition of the lexicon (*blackfoot.lexc*, containing word stems, affixes and continuations) with definitions of segment types and ordered phonological rules (*blackfoot.script*, written in *xfst*)[1]. Below is a simplified schema of the continuations needed to account for the examples presented above (every *lexc* file must start with a root):

Root → **NounStem** (*nan*, *nin*) → **AgreementSuffix**

I populate each category with the underlying forms of its lexical entries, along with a continuation for each lexical entry. I prepend suffixes (and postpend prefixes) with the special symbol ^ to encode morpheme boundaries and constrain the domain of some processes (see below). The *lexc* continuations allow the analyzer to recognize strings of underlying forms in morphologically valid order, e.g. *áwanááN+^iksi = knife+Animate.Plural*. In *blackfoot.script* I encode the allomorphy and phonological rules (some of which I will detail in the course of this report) that apply on top of these underlying strings to produce the surface forms. Hence the phonological component is encapsulated as `define Rules [ Allomorphy .o. Phonology ] ;`, and the entire program as `regex [ Lexicon .o. Rules ] ;`.

Below are the implementations of all the rules needed to account for the above examples (I am copying them here to be pedantic and avoid confusion), ordered as presented here:

SemivowelLoss:
`[ SEMI -> 0 || [.#.|CONS] _ ] ;`
NonPermanentConsonantLoss:
`[ NONPERM -> 0 || _ %^ VOW .#. ] ;`
VowelShortening:
`[ {aa} -> a, {ii} -> i, {oo} -> o, {áá} -> á, {íí} -> í, {óó} -> ó || _ %^ VOW ] ;`
AccentSpread:
`[ a -> á, i -> í, o -> ó || ACCENT _ ] ;`
IgnoreMorphemeBoundary:
`[ %^ -> 0 ] ;`
IgnoreNonPermanentConsonant:
`[ M -> m, N -> n, S -> s ] ;`

The *lexc* continuations permit both the combinations /*awanááN^wa*/ (“rattle”) and /*awanááN^iksi*/ (“rattles”). The defined rules apply as expected, and we note that IgnoreNonPermanentConsonant simply rewrites the underlying representation in lowercase. Notice that NonPermanentConsonantLoss and VowelShortening, which activate sequentially in /*awanááN^iksi*/, both have an *xfst* environment of the form `_ ^ V`. This is a natural requirement for non-permanent consonants, which only appear at morpheme edges (stem-finally). Frantz does not make reference to morpheme boundaries in his VowelShortening rule, but it becomes

[1] Please refer to my actual implementation, shared with you on github.

necessary given the presence of three consequtive vowels in some roots, such as *aaápan* "blood" (*nin*). Had we omitted morpheme boundaries in our the rule, *aaápan* would have become **aápan* morpheme-internally, but we do indeed find a form such as *aaápaistsi* that can be neatly accounted if we allow reference to morpheme boundaries in our rule application.

Another basic phonological rule is t-Affrication: t → ts / _ i. I have not seen any exceptions to this rule in the grammar, i.e., there are no cases of *t* followed directly by an *i* without affricating or "breaking". My *xfst* implementation of this rule is [t -> ts || _ %^ [i|í]] ;, making reference to morpheme boundaries similarly to VowelShortening. My original attempts at implementing this rule caused me a lot of grief because I had placed a *ts* into the list of consonants, and the transducer was treating it as a single multi-character symbol, seeking to undo a *ts* combination to a *t*+*i* every time it saw one (even when the underlying form had a *ts* and not a *ti*).

Implementing Possessive Affixes and Basic Allomorphy

Blackfoot has a set of personal prefixes and suffixes to indicate possession on nouns:

	singular	**plural**
1	*n-/nit-*	*n-/nit-...(i)nnaan*
21[2]	—	*k-/kit-...(i)nnoon*
2	*k-/kit-*	*k-/kit-...-oaawa*
3	*w-/ot-*	*w-/ot-...-oaawa*
4	*w-/ot-*	*w-/ot-...-oaawa*

These affixes are very similar to verb agreement affixes (with some caveats), and the *n* and *k* in the first and second person prefixes are ubiquitous in the Algonquian languages (cf. pronouns in John Elliott's *An Indian Grammar Begun* on Wampanoag). Note that the plural forms have the same prefixes as the singular forms, and additionally have the plural person suffixes.

At this point the grammar introduces a class of nouns called *inherently relational* nouns—nouns that are always inflected for a possessor, and are semantically relational. There are numerous animate relational nouns (*nar*), and very few inanimate relational nouns (*nir*). Some animate relational noun stems include *iksisst* "mother", *ohko* "son", *itan* (or *itán*) "daughter"; inanimate relational nouns include *inihka'sim* "name", *ookóówa* (or *ookoowa*) "house", *saayiimi* "ruthlessness". This is a neat instance of the lexical semantic relationality of these concepts being reflected in the morphology (unlike in English, where *mother* has the same template as any non-relational noun).

[2] The 21 refers to the inclusive 'we' form. This number also occasionally indicates *unspecified subject*.

While relational nouns are obligatorily possessed (i.e. necessarily take possessive affixes), non-relational nouns may also optionally be possessed, with minor derivational morphology on the noun stem. To accommodate the possessive template, the *lexc* continuations expand to:

Root → **PossessivePrefix** → **NounStem** (*nan*, *nin*) → **PossessiveSuffix** → **AgreementSuffix**

Hence we have forms such as *n-+itán+-wa* = 1+*daughter+Animate.Singular* = "my daughter", which is realized as *nitána* after SemivowelLoss; or *n-+itán+innaan+wa* = 1+*daughter*+1.*Plural+Animate.Singular* = "our daughter", realized after *nitáninnaana*. While the analyzer handles such valid sequences of four morphemes, it will also analyze an invalid morpheme sequence permitted by the *lexc* continuations, such as *n-+itán+oaawa+wa* = 1+*daughter*+2.*Plural+Animate.Singular*, even though it has an incompatible first person prefix and second person suffix. This problem is addressed in *Finite State Morphology* either by filtering or by using flag diacritics to handle long-distance dependencies (e.g. circumfixes in Arabic, similar to above affixes). Both of these methods, however, are ways around the computational limitations of FST's, which don't have a stack (as push-down automata do) to serve as a memory for long-distance constraints. To me the issue of overgenerating is of minor importance in comparison with the ability to analyze every valid form, so I ignore it for now (and may want to implement a separate filter later on that only selects the correct combinations).

I implemented some very basic allomorphy encountered in possessive affixes. The third person prefix *w-* is replaced by *m-* before stems beginning with *a*, e.g. *w+aaáhs+yi* "his elder relation" → *maaáhsi*. For the first person and first person inclusive/unspecified (21) plural suffixes *(i)nnaan* and *(i)nnoon*, the *i* is not present after stems ending in *i*, *a*, *w* or *y*; the analyzer is able to account for both *nitáninnaana = 1+itán+1P(innaan)+3S* "our daughter", and *nookóówannaani = 1+ookóówa+1P(nnaan)+IN.S* "our house".

Body parts form an interesting subgroup of inanimate nouns in Blackfoot. Most of them begin with an *m*, which is present only if no prefix precedes the stem (stem-initial nasal loss after prefixes can be generalized to most nasal-initial stems). For roots such as *móókoan* "stomach" or *motsíS* "hand", the analyzer can handle forms such as *k+móókoan+yi* → *kóókoani* = "your stomach", or *n+mo'tsíS+istsi* → *no'tsíístsi* "my hands".

Future Work and Approach

Though my current progress sets a strong baseline, much remains to be implemented. This includes deriving a possessable form from a non-relational noun, which is done by the addition of an *m* that has effects on vowel length (chapter 14); also implementing the allomorphy detailed in chapter 15—morpheme-initial variation, variable-length vowels, and morpheme-final allomorphy (such as semivowel alternation and diphthongization; I have already implemented non-permanent consonants).

Another substantial challenge involves incorporating the rich class of *adjuncts* (which often serve a modifying purpose) into the continuations for the nominal morphology. A prominent example is the stem *ómahk* "big/old", and frozen into pages of dictionary entries. A good example involving *ómahk* that would be good to account for is *kitómahkotaniksi* = *kit+ómahk+itan+iksi* "your oldest daughters" (this form involves yet another *i* ~ *o* alternation I would need to account for in the allomorphy).

I implemented phonological rules to account for the bulk of basic noun inflection, but the implementation is not exhaustive. Refer to chapter 5 and Appendix B of the grammar for a detailed description and a full compilation of rules. In this regard, my approach is data-driven rather than theoretical, and seeks to refine the implementation by accounting for every attested form. The goal is to be able to correctly analyze every noun entry and inflection in the *Blackfoot Dictionary*—please refer to this spreadsheet, which I am using to keep track of my progress (the list has omissions due to bugs in my dictionary-scraping script, and I will proof-check this later).

Another issue involves the phonemic status of pitch accent (prominence of a vowel or diphthong) in Blackfoot, and the (ir)regularity of its alternations that involve it. It is prominent enough to merit orthographic marking, and participates regularly in Accent Spread (see above); the addition of a prefix often causes the pitch raising of the vowel in the next syllable (e.g. *inaaáhs* "father-in-law" and *nitsínaaáhsiksi* "my fathers-in-law" with the stem-initial *í* accented). However, Frantz makes no indication of the phonematicity of pitch accent, and there are very few minimal pairs that differ solely by the accent of their vowels[3]. Moreover, many forms recorded in the dictionary contain seemingly unpredictable variations in pitch accent, e.g. *iihtáihkitsoohpatts**a**akio'p* "clothes dryer" vs. *iihtáihkitsoohpatts**á**akio'piksi* "clothes dryers". Attempting to account for all such forms with rules seems all but impossible and unnecessary, since simply de-accenting the forms would be a simple solution that presents no morphological ambiguity. My plan is to make a clone of the current analyzer that has all accent-related rules removed from it, and operates over de-accented versions of all the morphemes in the lexicon. Then I can create a wrapper that first de-accents any incoming form, analyzes it, and maps the component morphemes to their potentially accented counterparts in a dictionary. Forms that the analyzer currently can't handle solely due to accenting are currently marked in yellow in the spreadsheet.

Nominalizations are a prominent and sophisticated component of the *Blackfoot Grammar*, and merit a chapter of their own in the book. I do not expect to be able to account for the complexities of nominalizations, since they entail all the continuations of verbal morphology, which is too variable and rich in a polysynthetic language to be accounted for with an FST. This seems to be a good opportunity to experiment with sequence-to-sequence approaches, potentially using the "gold" rule-based output of the morphological analyzer to train a neural one, as Moeller et. al. neatly do in *A Neural Morphological Analyzer for Arapaho Verbs Learned from a Finite State Transducer* (Arapaho being another Algonquian language).

[3] The only such minimal pair I found so far is *aapáiai* "common cattail" vs. *áápaiai* "ermine/weasel"

References

University of Helsinki. 2003-2008. Helsinki Finite State Technology (HFST).

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Publications.

Donald G. Frantz. 2017. *Blackfoot Grammar,* 3 edition. University of Toronto Press.

Donald G. Frantz and Norma Jean Russell. 2017. *Blackfoot Dictionary of Stems, Roots, and Affixes*, 3 edition. University of Toronto Press.

Conor Snoek, Dorothy Thunder, Kaidi Lõo, Antti Arppe, Jordan Lachler, Sjur Moshagen, and Trond Trosterud. 2014. Modeling the noun morphology of Plains Cree. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 34–42, Baltimore, Maryland, USA. Association for Computational Linguistics.

Moeller, Sarah, et al. "A neural morphological analyzer for Arapaho verbs learned from a finite state transducer." *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*. 2018.