Creating lexical resources for polysynthetic languages—the case of Arapaho

Ghazaleh Kazeminejad and **Andrew Cowell** and **Mans Hulden**
Department of Linguistics
University of Colorado
{ghazaleh.kazeminejad,james.cowell,mans.hulden}@colorado.edu

Abstract

This paper discusses the challenges in creating pedagogical and research resources for the Arapaho language. Because of the complex morphology of this language, printed resources cannot provide the typical useful functions that they can for many other languages. We discuss the challenges faced by most learners and researchers working with polysynthetic languages, of which Arapaho is an excellent example, as well as some currently implemented solutions in creating computer resources for this language, including a lexical database, a morphological parser, and a concordancer. The construction of the finite-state morphological parser which many downstream tasks rely on will be discussed in some detail.

1 Introduction

Of the approximately 6,000 languages spoken in the world today, at least half are critically endangered, and the death of a language often corresponds to the death of the culture of the speakers of that language (Minett and Wang, 2008). One of the strategies to preserve this heritage to some extent is documenting such languages through linguistic fieldwork. However, revitalization of an endangered language does not occur solely by documenting it in a book. What is required is sufficient motivation and practical means for the descendants of these cultures that help them in learning and using the language, which automatically leads to revitalization of that language.

Arapaho (ISO 639-3:arp) is one of the critically endangered languages in the Algonquian family (Cowell and Moss Sr., 2008; Lewis et al., 2009). Like all languages in the family, it is both polysynthetic and agglutinating. The learners and speakers of the language are disproportionately disadvantaged by limiting themselves to printed media in comparison with more isolating languages such as the Indo-European ones. On the other hand, considering properties of this language which will be explained in section 2.1 below, electronic resources are indispensable for providing useful dictionaries and lexicons.

In the current project, the focus has been on the lexicon and morphology of Arapaho and developing computational tools mainly to support pedagogical purposes. Using the documented grammar (Cowell and Moss Sr., 2008) and the collected corpus of Arapaho, we have constructed an online lexical database (containing audio files), a morphological parser, and a concordancer implemented in the dictionary.

Section 2.1 briefly describes the current situation of the Arapaho language, in addition to some of its prominent phonological features, along with its verbal system which is the current focus of the parser, that being the most complex part of the morphology. Section 2.2 explains the necessity of creating such online resources rather than relying on traditional printed ones. Section 3 describes the design issues and technical details of each of the developed resources and tools, and section 4 discusses the reasons behind the decisions to develop such tools for a language like Arapaho.

2 Background

2.1 Arapaho

Arapaho is an Algonquian language currently spoken in two dialects: Northern Arapaho which has less than 200 native speakers all in their late fifties in the Wind River Indian Reservation in Wyoming, and Southern Arapaho which is spoken by only a handful of people all near eighty or older in western Oklahoma (Cowell and Moss Sr., 2008). After World War II, children began to be raised speaking

Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 10–18,
Honolulu, Hawai'i, USA, March 6–7, 2017.

English rather than Arapaho. However, the Northern Arapaho have attempted to maintain their language through documentation (both written and taped) and pedagogical efforts (producing extensive curricular materials). The standard orthography for Arapaho was developed in the late 1970s.

In general, the Northern Arapaho have a positive attitude toward the language, and the tribal government spends money on preservation efforts. A large transcribed and annotated spoken corpus has been created, parts of which are available in the Endangered Languages Archive.[1] Many young people take classes and show interest in the language. However, there are various economic and pedagogical limitations, and also the learners tend to not allot sufficient time to learn the language effectively. This work has largely relied on on locally produced ad-hoc curriculum and on the Arapaho Language Project website[2].

Arapaho has no fixed word order; pragmatic factors largely determine word order. It's a highly polysynthetic language, which incorporates as much information as possible into complex verbs. Consequently, many Arapaho sentences consist only of a verb. Roughly speaking, the following is the order of elements in the verb:

1. PROCLITIC
2. PERSON MARKER PREFIX
3. TENSE/ASPECT/MODE PREVERB
4. LEXICAL PREVERB
5. VERB STEM
6. DIRECTION OF ACTION THEME
7. PERSON/NUMBER SUFFIX
8. MODAL SUFFIX

The proclitics have modal and evidential functions. The person inflections mark first, second, and third person. Preverbs indicate tense and aspect, as well as negation and content questions. The verb stem is itself typically internally complex. The theme occurs with transitive verb stems and indicates the direction of action when multiple arguments are marked on the verb (i.e., who is acting on whom). Singular and plural are marked as well. The term *mode* in Arapaho refers to markers that indicate iterative and subjunctive constructions.

There are four classes of verb stems in Arapaho: transitive verbs with animate objects (TA) or with inanimate objects (TI), for which two arguments are obligatorily marked on the verb inflectionally, and intransitive verbs with animate subjects (AI) or with inanimate subjects (II), for which one argument is marked on the verb inflectionally.

A second key feature of Arapaho inflectional morphology is the existence of four different verbal orders: affirmative order (used primarily in affirmative independent clauses), non-affirmative (used primarily in negative and interrogative independent clauses), conjunct (used primarily in subordinate clauses), and imperative (used in imperative and prohibitive commands). The inflectional morphology used with any given verb stem varies according to the particular verbal order in question.

Expressing nominal arguments is not obligatory in Arapaho as long as the referents are clearly marked on the verb. Thus one could say:[3]

(1) **ne'-nih-'ii'-cesis-oowooceineti-3i'**
that-PAST-when-begin-lower self by rope-3PL
'that's when they started lowering themselves on the rope.'

The underlying polysynthetic verb stem in this expression is **hoowooceineti-**, "lower oneself by rope," and the verb stem itself is internally complex, consisting of **hoow-oocei-n-eti-**, down-rope-TRANSITIVE-REFLEXIVE. Thus, the overall expression has nine morphemes, four in the verb stem, four as prefixes, and one as a suffix. Such an expression is not exceptional, but is in fact a fairly common Arapaho word/sentence.

In addition to its extremely complex verbal morphology, Arapaho feature a rich array of phonological processes as well. These phonological alternations include phenomena such as progressive and regressive vowel harmony with non-parallel effects and distribution, allophonic rules, vowel epenthesis, consonant deletion, vowel reduction, vowel lengthening, and consonant mutation. Arapaho also has a complex pitch accent system, with a related system of vowel syncope. One of the very pervasive morphophonological processes found in Arapaho is an initial change that serves grammat-

[1] http://elar.soas.ac.uk/deposit/0194
[2] http://www.colorado.edu/csilw/alp/

[3] The orthography used for Arapaho largely corresponds to the International Phonetic Alphabet equivalents, except the symbols y = /j/, c = /tʃ/, ' = /ʔ/ and 3 = /θ/.

ically to indicate either present ongoing or present perfect in affirmative order and conjunct iterative verbs. In such tense/aspect combinations, the verb stem undergoes an initial change: for verb stems whose first syllable's nucleus is a short vowel, the vowel is lengthened; and for verb stems whose first syllable's nucleus is a long vowel, an infix /en/ or /on/ (depending on vowel harmony) is inserted between the initial consonant and the long vowel.

2.2 Necessity of Creating Computer Resources

It is immediately obvious that using an Arapaho dictionary—attempting to find or look up **hoowooceineti-** in particular—would be an extremely difficult task for anyone without linguistic training, and the audience of these resources would not be an exception, as they are neither linguists, nor Arapaho native speakers. Locating the actual stem among the many prefixes and suffixes requires fairly advanced knowledge of the morphosyntax of the language, and the underlying stem does not appear as such in the surface word form, due to loss of the initial /h/ following a consonant, so a knowledge of morphophonemic changes would also be required to successfully find the stem. These latter changes can be much more variable than simply loss of initial /h/. If we take the stem **ni'eenow-** "like s.o.", we find that surface forms are **nii'eenow-o'** "I like him/her," but **nii'eeneb-e3en** "I like you." The initial vowel is lengthened (due to lack of a preceding prefix), the final consonant mutates from /w/ to /b/ prior to a front vowel, and the final vowel of the stem shifts due to vowel harmony. Working backwards, a user would have to go from **nii'eeneb-** to **ni'eenow-**. In many cases, a stem has a half dozen or more allostems due to these types of changes.

A language such as Arapaho produces severe problems for a producer of print dictionaries. Is one to list all the different allostems, and refer the user back to a single base stem? If so, a base listing of 10,000 verbs (very quickly obtainable for a polysynthetic language) will produce a need for perhaps 50,000+ individual entries, most of them 'empty' cross-references. In addition, for transitive verbs with animate grammatical objects, the number of potential inflections is in the dozens (**-e3en** = 1S/2S, **-e3enee** = 1S/2P, **-o'** = 1S/3S, **-ou'u** = 1S/3P, **-een** = 1P/2S, etc.[4]). Since Arapaho verbs cannot appear without an inflection in most cases, the underlying stem never actually appears in discourse. One must thus list an inflected form for the user to show actual pronunciation. But clearly, dozens of different inflected forms occur for each transitive verb with an animate object, and even intransitive verbs have five different person marking suffixes, plus number suffixes and an exclusive/inclusive distinction with 1P. To top off the problems, when the verbs are used for negations and questions, a different set of markers—primarily prefixes—are used for person and number. Thus, each transitive verb has something approaching 100 common inflections, prior to the addition of any tense, aspect, modal or other prefixes or suffixes. Listing one or two of these may not be much help to a beginning learner.

The Algonquian languages are in this regard similar to several other language families in North America, including Iroquoian, Athabaskan (and the larger Na-Dene phylum) and Inuit-Aleut. The issues raised by Arapaho for dictionary users (as well as for those attempting to examine a textual corpus for a given morpheme or verb stem) are thus highly relevant for many different languages. Similar points to the ones above have been raised in previous basic computational work for morphologically complex and polysynthetic languages (Cox et al., 2016; Gasser, 2011; Hulden and Bischoff, 2008; Rios, 2011; Snoek et al., 2014).

3 Creating Computer Resources for Arapaho

3.1 Morphological Parser

The best first-step solution for many of the typical problems faced by polysynthetic and agglutinating languages such as Arapaho is a morphological parser. Similar efforts has taken place for another Algonquian language, Plains Cree (Snoek et al., 2014). Having developed a finite-state morphological parser for a morphologically complex language, developing a spell checker (or even corrector), a lemmatizer, or e-dictionary tools would be more accessible for any language (Alegria et al., 2009; Pirinen and Hardwick, 2012). This is much more crucial for languages with a heavy use of morphology, such as the Algonquian languages. Since in heavily agglutinating languages one word contains what in more isolating languages is equal to several isolated words (or maybe even a full

[4] 1S/2P: 1SG Agent, 2PL Patient

sentence), access to a morphological analyzer for such languages is indispensable, and furthermore a prerequisite for other NLP tasks such as dependency parsing (Wagner et al., 2016).

We used the *foma* finite-state toolkit (Hulden, 2009) to construct a finite state transducer (FST)—the standard technology for producing morphological analyzers—which is bidirectional and able to simultaneously parse given surface forms and generate all possible forms for a given stem (Beesley and Karttunen, 2003). All the concatenative morphological rules as well as irregularities of the morphology of the language were taken care of using a finite-state lexicon compiler within *foma*, or *lexc*, which is a separate component in the system with a high-level declarative language for streamlining lexicon creation modeled as finite transducers (Beesley and Karttunen, 2003; Karttunen, 1993).

In the next step, the full set of Arapaho morphophonological rules were implemented as a set of morphophonological rewrite rules that perform context-conditioned sound/morpheme changes, so that the generated forms do not merely consist of a number of morphemes put together (the underlying form), but undergo the necessary alternations before the surface forms are generated. These transducers are essentially a series of phonological rewrite rules to move between an input strings (the analysis form, providing the grammatical information encoded within that form) and an output string (a surface form). For this purpose, the FSTs produce intermediate representations which are not visible after all the participating transducers have been combined together by transducer composition. Figure 1 is a schema of how our FST is designed to generate and parse word forms. Combining the lexicon transducer with the individual morphophonological rule transducers through transducer composition produces as output a single monolithic finite state transducer that can be used for both generation and parsing.

For example, **nonoohóbeen** is a surface form in Arapaho. Given to the parser, it is correctly parsed as

```
[VERB][TA][ANIMATE-OBJECT][AFFIRMATIVE]
[PRESENT][IC]noohow[1PL-EXCL-SUBJ][2SG-OBJ]
```

This simply interprets the given surface form as the verb stem **noohow** (*to see someone*) which is a transitive verb with an animate subject ([TA]) in the affirmative order ([AFFIRMATIVE]) order

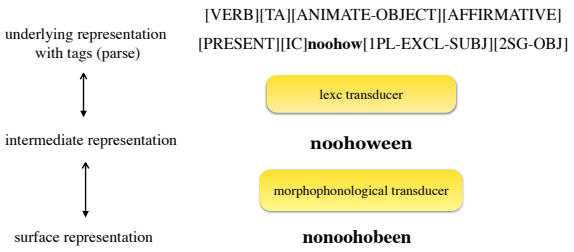

Figure 1: Composition in an FST illustrating the underlying (input) forms and the resulting surface (output) forms after mapping morpheme tags to concrete morphemes and subsequently undergoing morphophonological alternations.

and in present tense ([PRESENT]) (and therefore undergoing an initial change), with exclusive first person plural agent and second person singular patient. Note that the parse also shows an [IC] tag, which stands for Initial Change as described in section 2.1. It could be therefore translated as "We are seeing you.".

In the first step to design the parser, all the verb stems were automatically extracted from our lexical database (c.f. section 3.2), and each was flagged with its stem type (II, AI, TI, or TA) based on its part of speech in the lexical database. Some verbs were considered irregular by the grammar with regards to their inflectional markings. These were marked as irregular in a pre-processing step.

To enforce agreement constraints between non-adjacent morphemes, we use the formalism of flag diacritics within the grammar. These allow morpheme—and segments in general—to carry feature-setting and feature-unification operations and enable the FST to remember earlier paths taken, allowing for the grammar writer to enforce compatibility constraints between parts of a word (Beesley and Karttunen, 2003, p. 339).

In the next step, the lexicon compiler (*lexc*) file was designed using the *lexc* formalism to tag each stem with all its possible parses, connecting the input and output levels of the transducer with a colon (:). The part of speech and stem type were extracted from the lexical database in the preprocessing step, and the verbs receive all the possible inflections and we filter out inappropriate combinations using flag diacritics to control co-occurrence. For instance, a negative verb should be blocked from inflecting in anything but the non-affirmative. Or a past tense verb which has been prefixed by the past tense morpheme can not be

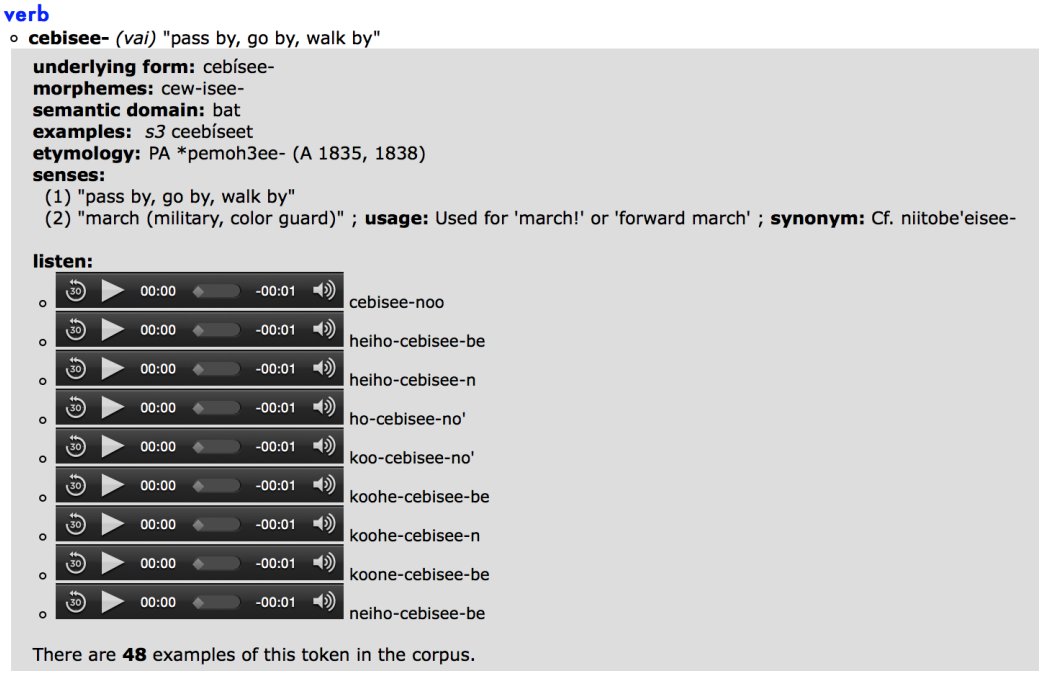

Figure 2: View of the Arapaho Dictionary.

simultaneously inflected for present tense by another morpheme.

One of the most important issues in designing the *lexc* component is the ordering of the morphemes. It may seem trivial to put the morphemes in order since the grammar book has (Cowell and Moss Sr., 2008) explicitly provided the ordering, but this still presents challenges once we get reach the specifics of each element. Person/number markers, for instance, are circumfixes for some verb orders. In order to prevent an AI (i.e. intransitive verb with animate subject) verb stem prefixed with, say, first person singular marker **ne-**, it needs to be flagged specifically with its person-number-stemType so that it is blocked from combining with random person-number suffixes.

In the final step, the *lexc* file is read into the *foma* system where all the morphophonological rewrite rules are applied to the intermediate forms created by the *lexc* file. The flag diacritics are very critical in this step as well, as we use the flag symbols as triggers for certain morphophonological changes. For example, the initial change process that marks present tense affirmative order forms is taken care of in this step. The following rewrite rule (in Xerox formalism) is one of the four rules required to apply this alternation to any word.

```
def IC1 i -> i i , o -> o o , u -> u u ,
e -> e e || "@U.IC.Yes@" \Alphabet*
"@U.StemInitial.Yes@" \Alphabet*
Consonant _ Consonant ;
```

The above rule indicates the initial change process where the present tense affirmative verbs with a short vowel in the first syllable of the stem would be lengthened. Using the `@U.IC.Yes@` flag in the conditioning environment forces the application of this vowel-lengthening rule only to those intermediate forms which contain this flag, where the vowel of the first syllable (hence `@U.StemInitial.Yes@` preceding it) is short.

The regressive vowel harmony rule (e∼o) applies to only a subset of verbal inflections. To model this, we introduce a blocking condition to the appropriate rewrite rule in form of flag diacritics, so that the rewrite rule for regressive vowel harmony doesn't apply to the person-number-stemType combinations that don't undergo vowel harmony.

The negative marker **ihoow-** poses an exception when no other morpheme precedes it. Normally in Arapaho, when a word is vowel-initial, an /h/ is inserted word-initially. For the negative marker, however, only the initial 'i' drops and the **hoow-** part remains. For instance, we expect the negative 3PL subject for the verb stem **towoti-** to be **ihoow-towoti-no**, but in fact what we get in the surface form is **hoowtowotino**. This exception is also covered using a flag in the *lexc* file that marks

negative morphemes, which is later remembered in the *foma* file to be treated appropriately.

As mentioned above, the verb stems themselves may have a complex underlying form. Some verb stems with specific endings (such as *-oti*) in their underlying form show some irregularities in their inflection. Such issues are also accounted for in the lexicon description.

All the rewrite rules are individually defined at first, and then are applied to the intermediate forms defined in the *lexc* file. Using standard transducer composition, the defined rules are combined with the lexicon description, i.e. output of the *lexc* compilation, in the appropriate order. As is often the case when developing such grammars, care needs to be taken to assure that the alternations are applied in the correct order.

We have evaluated the parser against verb tables provided in the Arapaho Language reference grammar (Cowell and Moss Sr., 2008). The current recall of the parser is 98.2%.

3.2 Lexical Database

The lexical database software (which is also implemented online) was designed to include an annotation system where authorized annotators (linguists who know the Arapaho language) annotate words elicited from the existing Arapaho corpus with its relevant linguistic features such as the underlying form, gloss, part of speech, semantic domain, etc. and update the system, followed by the work of an adjudicator who accepts, rejects, or updates the annotation before the change applies to the lexical database.

The database also contains sound files associated with some word. For verbs, this includes different inflections for a verb stem uttered by an Arapaho native speaker. The sound files can be viewed in the publicly accessible online Arapaho-English dictionary[5] that we have developed, where for each verb stem queried, multiple fields of information are displayed in addition to a simple gloss. Figure 2 shows the dictionary entry for a verb stem, which contains all the sound files related to this stem along with the transcription of each uttered word in front of the corresponding sound file.

The morphological parser, used in conjunction with both a lexical and text database, solves several problems simultaneously for users. First, we have automatically generated all possible allostems of each verb using the morphological parser (they are fully predictable), and written these into a subfield in our lexical database for /allostem. For example **hiine'etii-** 'to live' has allostems **iine'etii-**, **'iine'etii-** and **heniine'etii-**. Each allostem is linked to a parent stem. We can then direct that any lexical query to the database query both the /stem field and the /allostem field in the lexical database. If a user query involves an allostem, then the user is automatically referred (via the search function) to the base stem. Thus the user can enter any possible surface/spoken stem in the query field and find the appropriate base form, without needing to have knowledge of Arapaho morphophonemic processes.

When and if a print dictionary is eventually produced, we would likely use only the base stem as an entry. This arrangement of database fields and subfields prevents the possibility of 50,000 separate entries for 10,000 verbs; or if we choose to actually print all these allostems, it would be trivial to simply include a subheading such as "variant of…" in the gloss field to redirect the user.

3.3 Concordancer

Conversely, a user might want to find examples of a given stem in actual usage. For that purpose, we have developed a concordancer which is implemented inside the dictionary. The users have the option to specify the number of examples desired, and the dictionary will return the relevant sentence examples if the corpus contains them. The Arapaho corpus underlying this concordancer currently contains around 80,000 sentences (with an eventual goal of 100,000 sentences).

Performing lemmatization in Arapaho is not as easy as for isolating languages, and the listed base form of a stem is often not the most common alloform of the stem to occur in actual discourse. However, we have designed the concordance query function so that when a user asks for occurrences of the base stem, the search function searches simultaneously for all occurrences of allostems as well. Thus, the concordance reports back all instances of the stem in usage, without the user having to perform searches allostem by allostem. Indeed, the user does not even need to be able to predict or derive the possible allostems. Because the concordancer also reports back the

[5] https://verbs.colorado.edu/arapaho/public/view_search

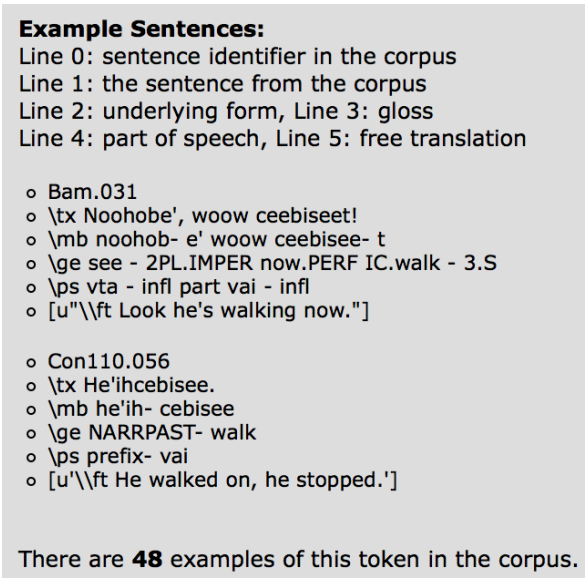

Figure 3: Example sentences retrieved through the concordance Function in the Arapaho lexical database interface.

total number of instances of the stem and its allostems in the text database, it constitutes a valuable pedagogical resource, as it allows teachers to determine the relative frequency of all verb stems in the database. Figure 3 shows the example sentences retrieved through the concordance function and displayed in the dictionary, with a guideline on top indicating how to read the lines from the corpus.

4 Discussion

One question that often arises is the following: why have we not designated the most common form of a stem as the base form, since this is normal practice in linguistics with allophones and allomorphs, and one might logically expect that this is the form users would most often search for? The reason we have not done this is because of the polysynthetic nature of Arapaho. In English, we may find a verb 'walk' which can occur in collocations such as 'walk down', 'walk up', 'walk around' and so forth. In Arapaho, all of these occur as different verb stems: **hoowusee-** 'walk down', **noh'ohusee-** 'walk up', **noo'oesee-** 'walk around' (the bound morpheme **-see-** indicates 'walk'). The result of this is that Arapaho has far more verb stems in a given large chunk of text than English does. There are dozens of different 'walk' verb stems for example. A secondary result of this is that a given verb stem will occur much less commonly across tens of thousands of lines of discourse than is the case in English. While our text corpus of (eventually) 100,000 lines (less than one million words) is not tiny in size, it certainly does not approach the several-billion-word corpora in English that one can access through resources such as The Sketch Engine (Kilgarriff et al., 2004). Thus, for any uncommon Arapaho verb stem, the combination of a relatively small overall text corpus and multiple allostems results in very low frequencies of occurrence per allostem (low single digits in many cases). As a result, chance factors can play a significant role in what is the 'most frequent' allostem. Moreover, once one starts getting **hiine'etii-** as the most common form for 'live', but some other h-initial verb stem turns out to be most common as **'iten-** (from base **hiten-** 'get, take') and another as **iisiiten-** (from base **hiisiiten-** 'grab, catch, seize'), then our morphological parsing ability would collapse. The parser is built to produce all allostems from a uniform base stem, with uniform phonology (all final **b/w** stem alternations take the **w** form as the base form, for example, and list the **b** stem as an allostem). Listing stems under the 'most common form' would destroy this uniformity.

Furthermore, in our design we decided to assign separate entries in the dictionary to some derived forms even though they include a productive morpheme. This happens only for the derived forms that occur in the corpus, and we included them in the dictionary to facilitate glossing. In addition however, productive morphemes such as the causative or benefactive morphemes sometimes produce derived forms with idiosyncratic meaning. For instance, combining **hiicoo-** 'smoke' with the causative suffix **-h** does not give a prototypical causative meaning, rather it means giving a cigarette and allowing one to smoke. There is also cognitive evidence from more recent studies (Cowell et al., 2017) suggesting that many of the morphologically complex stems are in fact part of the lexicon rather than the result of syntactic movement phenomena. So we definitely need to list such verbs in separate lexical entries.

Moreover, since our resources are primarily pedagogical and our audience are primarily beginning learners, we need to put a minimum burden on them in designing pedagogical resources. Since identifying and correctly implementing morphemes is a daunting and confusing

task for beginning learners (and more so for the learners of a polysynthetic language), including them in the dictionary as far as they occur in the corpus seems to be a reasonable idea and not a redundant task.

In summary, the morphological parser when applied to stems allows us to point the user from allostems to a main stem in the lexical database, and from a main stem to all the allostems in the text database, in a way which requires no linguistic knowledge from the user, at least in terms of morphophonemics. This ability resolves the single most problematic issue with dictionaries of polysynthetic, agglutinating languages. This is also not a functionality available in commonly-used linguistic interlinearizing software such as Toolbox[6] or FLEx[7]. As a second stage of the project, we have also applied the morphological parser to generate all possible inflected forms of each verb. When a user finds a stem in the lexical database, that person can simply request, via a single query, that all inflections of the verb (with morphophonemic changes applied) be generated in a list. The list gives both the inflected form, its linguistic labels/parse, and a gloss ('I am [verb]ing'), for users without linguistic knowledge. Thus all one hundred or so forms of the transitive verbs with animate objects can be produced automatically, in a way impossible in a print dictionary with its space limitations.

The generated inflected surface forms for all verb stems are, in the next step, going to occupy another subfield in the dictionary (say, /inflstem). Thus, a user could enter a query for a word in the search field, and the database will be directed to search all /stem fields, all /allostem fields, and all /inflstem fields for a match. This will make the search function much more powerful, since not only does the average user not have the ability to do morphophonemic analysis, but he or she may often, at least initially, not be able to recognize all inflectional prefixes and suffixes, and thus enter an inflected form into the search field rather than just a stem. This is again the common problem with trying to use a dictionary with a polysynthetic agglutinating language. And again, due to the issue of allostems as well as morphophonemics, common linguistic dictionaries and annotation software have very imperfect functionality.

[6] http://www.sil.org/computing/toolbox/
[7] http://fieldworks.sil.org/flex/

5 Future Work

The next step in this process, which we have not yet implemented, will be to extend the morphological parser so that it generates all possible temporal, aspectual and modal forms of a verb. Currently the analyzer is only able to generate and parse verb forms in the present and past tense, and perfective and present-ongoing aspect. Continuing with our example of **hiine'etii-** 'to live', past tense is **nih'iine'etiinoo** 'I lived', the future tense is **heetniine'etiinoo**, the imperfective aspect is **niine'etiinoo**, and so forth. Since all 100 different inflections of a transitive verb can surface with around a dozen different Tense-Aspect-Mood forms, plus reduplicated forms, plus forms with lexical prefixes, the numbers quickly rise to the thousands or even millions of possible forms for any base verb stem. At this point, we encounter a separate problem (which is not the focus of this paper): that multiple base stems can generate the same surface inflected stem form via a process of random convergence, and a disambiguation component thus becomes necessary. The more powerful the parser, and the farther one moves beyond the stem itself, the more likely this is to become a problem. Although relatively simple statistical methods using weighted automata in the analyzer can be used to reliably filter out improbable analyses, if enough labeled data is given for training such a model. Such a disambiguator has been implemented for Plains Cree in Arppe et al. (2017), and we assume the same model would also be applicable to Arapaho. However, it is undeniable that at this point syntax becomes the fundamental problem, and that not all disambiguation can be performed by analyzing the plausibility of a particular morpheme combination. With the availability of much more labeled data, deep learning methods also become applicable for context-sensitive disambiguation (Shen et al., 2016). This problem in general then becomes a problem of syntactic disambiguation, as in NLP applications for more isolating languages such as English. Unlike English, however, in this case the syntax is internal to a single verbal form, rather than occurring across multiple words, and at this point some equivalent of the English VerbNet system (Schuler, 2005) could be to be used, though it will have to be more like a "StemNet," combined with "Prefix/SuffixNet."

References

Inaki Alegria, Izaskun Etxeberria, Mans Hulden, and Montserrat Maritxalar. 2009. Porting Basque morphological grammars to foma, an open-source tool. In *International Workshop on Finite-State Methods and Natural Language Processing*, pages 105–113. Springer.

Antti Arppe, Katherine Schmirler, Miikka Silfverberg, Mans Hulden, and Arok Wolvengrey. 2017. What are little Cree words made of? Insights from computational modelling of the derivational structure of Plains Cree stems. In *Papers of the 48th Algonquian Conference*.

Kenneth R. Beesley and Lauri Karttunen. 2003. Finite-state morphology: Xerox tools and techniques. *CSLI, Stanford*.

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

Andrew Cowell, Gail Ramsberger, and Lise Menn. 2017. Dementia and grammar in a polysynthetic language: An Arapaho case study. *Language*, 93(1).

Christopher Cox, Mans Hulden, Miikka Silfverberg, Jordan Lachler, Sally Rice, Sjur N. Moshagen, Trond Trosterud, and Antti Arppe. 2016. Computational modeling of the verb in Dene languages—the case of Tsuut'ina. In *Dene Languages Conference*.

Michael Gasser. 2011. Computational morphology and the teaching of indigenous languages. In *Indigenous Languages of Latin America—Actas del Primer Simposio sobre Enseñanza de Lenguas Indígenas de América Latina*, pages 52–61.

Mans Hulden and Shannon T. Bischoff. 2008. An experiment in computational parsing of the Navajo verb. *Coyote Papers: Working Papers in Linguistics*.

Mans Hulden. 2009. Foma: a finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 29–32. Association for Computational Linguistics.

Lauri Karttunen. 1993. *Finite-state lexicon compiler*. Xerox Corporation. Palo Alto Research Center.

Adam Kilgarriff, Pavel Rychlý, Pavel Smrž, and David Tugwell. 2004. The Sketch Engine. In *Proceedings of the Eleventh EURALEX International Congress*, pages 105–116.

M. Paul Lewis, Gary F. Simons, and Charles D. Fennig. 2009. *Ethnologue: Languages of the world*, volume 16. SIL international Dallas, TX.

James W. Minett and William S-Y. Wang. 2008. Modelling endangered languages: The effects of bilingualism and social structure. *Lingua*, 118(1):19–45.

Tommi A. Pirinen and Sam Hardwick. 2012. Effect of language and error models on efficiency of finite-state spell-checking and correction. In *Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing (FSMNLP)*, pages 1–9, Donostia–San Sebastin, July. Association for Computational Linguistics.

Annette Rios. 2011. Spell checking an agglutinative language: Quechua. In *5th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 51–55.

Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania.

Qinlan Shen, Daniel Clothiaux, Emily Tagtow, Patrick Littell, and Chris Dyer. 2016. The role of context in neural morphological disambiguation. In *Proceedings of COLING 2016*, pages 181–191, Osaka, Japan, December.

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, June. Association for Computational Linguistics.

Irina Wagner, Andrew Cowell, and Jena D. Hwang. 2016. Applying universal dependency to the Arapaho language. In *Proceedings of LAW X—The 10th Linguistic Annotation Workshop*, pages 171–179. Association for Computational Linguistics.