Converting a comprehensive lexical database into a computational model: The case of East Cree verb inflection

Antti Arppe
Department of Linguistics
University of Alberta
arppe@ualberta.ca

Marie-Odile Junker
School of Linguistics and Language Studies
Carleton University
marieodile.junker@carleton.ca

Delasie Torkornoo
School of Linguistics and Language Studies
Carleton University
Delasie.Torkornoo@carleton.ca

1 Introduction

In this paper we present a case study of how comprehensive, well-structured, and consistent lexical databases, one indicating the exact inflectional subtype of each word and another exhaustively listing the full paradigm for each inflectional subtype, can be quickly and reliably converted into a computational model of the finite-state transducer (FST) kind. As our example language, we will use (Northern) East Cree (Algonquian, ISO 639-3: crl), a morphologically complex Indigenous language. We will focus on modeling (Northern) East Cree verbs, as their paradigms represent the most richly inflected forms in this language.

2 Background on East Cree

East Cree is a Canadian Indigenous language spoken by over 12,000 people in nine communities situated in the James Bay region of Northern Quebec. It is still learned by children as their first language and fluently spoken in schools and in the communities overall, involved in most spheres of life as an oral language. Speakers have basic literacy in East Cree, but written communication tends to be in English or, to a lesser extent, French. The language is fairly well documented, with main resources available on a web site (www.eastcree.org) that includes multilingual dictionaries of two dialects (English, French, East Cree Northern and Southern dialects), thematic dictionaries, an interactive grammar, verb conjugation applets, oral stories, interactive lessons and exercises, a book catalogue, and other various resources like typing tools for the syllabics used, tutorials, spelling manuals, and so forth.

3 Verb structure

The verb in Northern East Cree follows the general Algonquian structure. Verbs fall into four major types according to transitivity and the animacy of the participants (intransitive with inanimate, or no, subject: II; intransitive with animate subject: AI; transitive with animate subject and inanimate object: TI; and transitive with animate subject and animate object: TA). Verbs are inflected for the person of the subject and/or object, and for modality. There are three major types of inflections, with their specific properties, known as orders: *independent*, *conjunct*, and *imperative*. As can be seen in the examples below, only the independent forms have person prefixes, while for conjunct and imperative forms person is only marked in the suffixes.

The orders can be divided into sub-paradigms according to how modality is marked in the post-verbal suffix complex. For Northern East Cree, a total of 15 distinct sub-paradigms have been identified, 7 for independent, 6 for conjunct, and 2 for imperative (Junker & MacKenzie, 2015), taking a classical *Word-and-Paradigm* approach (Blevins, 2011). In addition, verb stems can be combined with several pre-stem elements, known as preverbs, which can be divided into grammatical and lexical ones and which functionally correspond to auxiliary verbs or adverbials in English. For example, the preverb *chî(h)* indicates 'past' tense, *wî* 'want', and *nitû* 'go and V'. These are illustrated in examples (1a-c) below.

(1a) INDEPENDENT INDIRECT
chichî wî nitû mîchisunâtik
chi-chî wî nitû mîchisu-n-âtik
2-PAST WANT GO eat.AI-1/2SG-INDIRECT
'You wanted to go eating (so I was told)'

(1b) CONJUNCT DUBITATIVE PRETERITE
châ sâchihîtiwâhchipinâ
châ sâchih-îti-w-âhch-ipin-â
FUT love.TA-INV-DUB-1PL→2(PL)-PRET-DUB
'If only we could love you'

(1c) IMPERATIVE IMMEDIATE
sâchihînân
sâchih-înân
love.TA-2(PL)→1PL
'(You [sg. or pl.]) love us!'

As for orthographical conventions, grammatical and lexical preverbs are separated from the rest of the verb construction by spaces, (though this is not followed consistently for lexical preverbs, sometimes written attached to the stem). Personal prefixes (in the case of independent order forms) are attached onto the first preverb or the verb stem, as can be seen in (1a) and (3a-b). Moreover, long vowels may be indicated with a circumflex, such as <â>, used throughout the examples in this paper, or by doubling the vowel graphemes, i.e. <â> could alternatively be written as <aa>. The double-vowel notation is used for long-vowels in the computational model to be discussed below.

Morphophonology: While Northern East Cree (NEC) is fairly regularly agglutinative in its structure, there are some morpho-phonological phenomena occurring at the stem-suffix juncture, at the prefix-preverb/verb stem initial morpheme juncture, as well as within the suffix complex. For instance, a template morphology approach such as Collette (2014) presents 10 different suffix positions for the NEC verb. Furthermore, in the case of conjunct verb forms, the first syllable of the verbal complex, whether that of the first preverb or the stem, can undergo *ablaut*, known as *Initial Change* (IC), and resulting in a *changed conjunct* form. For example, the vowel *-â-* of the first syllable of the verb *mâtû* below (2a-c) changes to *-iyâ-* in the conjunct neutral form used in partial questions. Initial change of the verb stem only happens when there is no preverb before the verb stem, as preverbs can undergo initial change as well (cf. Junker, Salt & MacKenzie, 2015a).

(2a) m**â**tu-u
cry.AI-3(INDEPENDENT)
'S/he is crying'

(2b) âh m**â**tu-t
when cry.AI-3(CONJUNCT)
'When s/he is crying'

(2c) awân m**iyâ**tu-t
who IC.cry-3(CONJUNCT)
'Who is crying?'

To account for stem-suffix juncture morpho-phonological phenomena, Junker, Salt and MacKenzie (2015b) identify up to 19 stem types[1]. For example, *t-/sh*-stems alternate depending on person marking (3a-b), and *h*-stems trigger vowel *i*- lengthening (4).

(3a) t/sh-stem: *nâtâu*
chinâshin
chi-nâ**sh**-in
2-come.to.TA- DIR.2SG(SUBJ)→1SG(OBJ)
'you [sg.] come to me'

(3b) t/sh-stem: *nâtâu*
chinâtitin
chi-nâ**t**-itin
2-come.to.TA-INV.1SG(SUBJ)→2SG(OBJ)
'I come to you [sg.]'

(4) *h*-stem: *sâchihâu*
chisâchihîtin
chi-sâchih-**î**tin
2-love.TA- INV.1SG(SUBJ)→2SG(OBJ)
'I love you [sg.]'

All the inflectional information above is encoded into two databases, (1) a verb paradigm database and (2) a dictionary database. The verb paradigm database, consisting of 9,457 entries, lists exhaustive paradigms for each inflectional subtype (19 in all), plus some partial paradigms as well. That is, all basic prefix and suffix sequence combinations, indicating the person and number of subject (for all verb classes) and object (for TA verbs) as well as the various possible types of modality, are identified for each inflectional paradigm subtype and verb class (II, AI, TI, TA). Each entry in the verb paradigm database is a fully inflected verb form, which is associated with the relevant set of morphological features (Table 1). Importantly, each entry is provided with several different orthographical representations and structural partitions for different usage purposes. In particular, in a field named 'Search engine chunks', not only are all the suffixes lumped together, but this word-final segment/chunk, which we can call the *technical suffix*, also includes the stem final vowel or consonant (*h*- here), leaving behind what we call

[1] These are divided into 7 subtypes for TA verbs, 3 for II verbs, 6 for AI verbs, 1 for AI+O-verbs, and 2 for TI verbs.

a *technical stem* (*sâchi-* here), which remains invariant throughout the entire paradigm.

Table 1. ECN verb paradigm database entry representing one inflected form (selected fields)

ᒋᓴᒼᐃᑎᓵ	Word form in standard ECN syllabic spelling
chisâchihîtin	Word form in standard ECN roman spelling
chi-sâchi-hîtin	**Search Engine chunks**
chi-sâchih-îtin	Morpheme cuts for display
chi-sâchih-it-in	Morpheme breaks with underlying forms
1→2	Person (Subject→Object)
h	**Stem type**
VTA	Grammatical class
sâchihâu	Dictionary entry
01	Paradigm number

Table 2. ECN Dictionary database entry (selected fields)

ᓴᒼᐃᐧᐁᐤ	Word form in standard ECN syllabic spelling
sâchihâu	Word form in standard ECN roman spelling
h	**Stem type**
VTA	Grammatical class
sâchi	**Technical stem (regular)**
siyâ-chi-hât	changed conjunct (first syllable + rest of technical stem + endings for conjunct indicative neutral, *h* stem)
s/he loves someone	English translation

The dictionary database (15,614 entries) (Junker et al. 2012) determines the inflectional subtype for each verb. This allows for linking each verb with its entire paradigm according to a model verb for each inflectional subtype, as enumerated in the verb paradigm database. In addition, the aforementioned technical stem, in both its regular and changed (conjunct) form, is explicitly stored directly for each verb in the dictionary database. Using these technical stems and the corresponding word-final (and word-initial) technical suffix chunks from the verb paradigm database, one can generate all the inflected forms by simple concatenation, without needing any morphophonological rules. Nevertheless, one needs to bear in mind that these technical stems and word-final technical suffix chunks have no morphological reality, but are simply representations of convenience (see Junker & Stewart, 2008). Furthermore, all grammatical and lexical preverbs are also included as their own entries in the dictionary database, and we are treating initial-changed forms of preverbs as separate entries labelled as conjunct preverbs.

4 Computational modeling of the Northern East Cree verb

As our computational modeling technology, we are using Finite-State Transducers (FST) (e.g. Beesley & Karttunen 2003), well-known computational data structures that are optimized for word form analysis and generation, with a calculus for powerful manipulations. FSTs are easily portable to different operating systems and platforms, and thus can be packaged and integrated with other software applications, like providing a spell-checking functionality within a word-processor. In designing a finite-state computational model, with a fairly regularly agglutinative language such as East Cree, one has to decide whether one models morphophonological alternations at stem+affix junctures by (1) dividing stems into subtypes which are each associated with their own inflectional affix sets that can simply be glued onto the stem, or whether (2) one models such morphophonological alternations using context-based rewrite rules. Furthermore, one has to decide the extent to which one treats affix sequences by splitting these into their constituent morphemes, each associated with one morphosyntactic feature, or rather treats affixes as chunks which are associated with multiple morphosyntactic features (Arppe et al., in press). The more one splits affix sequences, the more one may need to develop and test rules for dealing with morphophonological alternations at these morpheme junctures, whereas in the case of chunking such alternations are precomposed within the chunk. In contrast, the more one uses chunks, the more one has to enumerate chunks based on the number of relevant inflectional subtypes.

While the chunking strategy is not parsimonious and compact, in our experience it results in FST source code which is nevertheless structurally quite flat and easily comprehensible for scholars who are not specialists for the language in question. Importantly, current finite-state compilers, e.g. XFST, HFST, or FOMA (Beesley and Karttunen 2003; Lindén et al. 2011; Hulden 2009), implement a minimization procedure on the finite-state model, so that

recurring realizations of string-final character sequences and associated morphological features are systematically identified and merged, resulting in the end in a relatively compact model (that in practice might not be much larger, nor structurally substantially different, than a model compiled from source code implementing maximal splitting). On the other hand, if some aspect of the chunked morpheme sequences needs to be changed, with the chunking strategy these have to be implemented in potentially quite a large number of locations.

For the Northern East Cree model, we decided to (1) split the pre-stem morphemes (personal prefixes for the independent order forms, and the regular and initial-changed forms of the grammatical and lexical preverbs), as there are very few morphophonological phenomena (initial change, epenthesis), and these are very regular. We deal with initial change by exhaustively listing the two alternative preverbs or stems (regular vs. changed); (2) entirely chunk the post-stem suffix morphemes, associating the chunks with multiple morphological feature tags; and (3) make maximal use of inflectional subtypes through using the aforementioned technical stems and post-stem word-final technical suffix chunks. Thus we will require no morphophonological rules for the stem-suffix morpheme juncture, and only two regular morphophonological rules in the pre-stem part.[2] These morphophonological rules are implemented using the TWOLC formalism within the FST framework. As to the rest, the LEXC formalism in the FST framework is used to define the concatenation of the morpheme sequences as treated above. For Independent order forms with subject (and object) person and number marked with a combination of a prefix and suffix (which can be understood to constitute a circumfix), agreement constraints between these affixes are implemented with the flag diacritic notation within the LEXC formalism.

5 Model statistics and details

The computational model currently includes stems and suffixes for AI, TI, and TA, but not for II verbs (which have the simplest paradigms). The LEXC source code for verb affixal morphology in its current form consists of 16,590 lines, of which 68 concern the pre-stem component and 16,514 the post-stem technical suffix chunks.[3] With minimization, its compilation with XFST takes 5.462 seconds with a 2 GHz Intel Core i7 processor and 8MB of RAM, resulting in a 108 kB XFST model (1,084kB with HFST).[4]

While this full enumeration of suffix chunks per each inflectional paradigm type results in a large number lines in the LEXC code, in comparison to a decompositional approach, the structure of the source code is quite flat and easy to grasp. As can be seen in Table 3 presenting the source code for the Independent Neutral Indicative suffix chunks for Animate Intransitive verbs of the *-aa* paradigm type, the suffix chunk *-aan*, which requires a first person prefix *ni-* to have been observed at the very beginning of the verb construction, indicated by the flag-diacritic `@U.person.NI@`, is associated with three morphological tags `+Indic`, `+Neu` and `+1Sg`, corresponding to the morphological features INDICATIVE, NEUTRAL and FIRST PERSON SINGULAR actor, respectively. In addition, the numeric code `+[01]` is provided, indicating the paradigm subset for Regular (Non-Relational) Independent Neutral Indicative verb forms.

Table 3. LEXC description of suffix chunk set for the Regular (Non-Relational) Independent Neutral Indicative forms for Animate Intransitive verbs of the *-aa* paradigm subtype.

```
LEXICON VAI_SUFFIX_aa_IND01
@U.person.NI@+[01]+Indic+Neu+1Sg:@U.person.NI@aan # ;
@U.person.NI@+[01]+Indic+Neu+1Pl:@U.person.NI@aanaan # ;
@U.person.KI@+[01]+Indic+Neu+2Sg:@U.person.KI@aan # ;
@U.person.KI@+[01]+Indic+Neu+21Pl:@U.person.KI@aanaaniu # ;
@U.person.KI@+[01]+Indic+Neu+2Pl:@U.person.KI@aanaawaau # ;
@U.person.NULL@+[01]+Indic+Neu+3Sg:@U.person.NULL@aau # ;
@U.person.NULL@+[01]+Indic+Neu+3Pl:@U.person.NULL@aawich # ;
@U.person.NULL@+[01]+Indic+Neu+4Sg/Pl:@U.person.NULL@aayiuh
# ;
@U.person.NULL@+[01]+Indic+Neu+XSg:@U.person.NULL@aaniuu # ;
@U.person.NULL@+[01]+Indic+Neu+XSgObv:@U.person.NULL@aanaani
wiyiu # ;
```

Example analyses provided by the FST analyzer for the forms (1a-c) are presented below in (5a-c). Grammatical and lexical preverbs are indicated with the notation `PV/…+`, and the subset of the paradigm using a notation with bracketed

[2] (i) insertion of an epenthetic *-t-* between the personal prefix and a vowel-initial stem or preverb; and (ii) assimilation of *i-* before a stem-initial *u-*.

[3] The entire source code for the (Northern) East Cree computational model presented here can be found at: `https://victorio.uit.no/langtech/trunk/startup-langs/crl/src/`

[4] This compares well with HFST models for other Algonquian languages the first author has experience of, e.g. 1,728kB for Odawa (otw) and 5,320kB for Plains Cree (crk), though on must note these two models cover also noun morphology not yet implemented in our East Cree model.

numbers, e.g. `+[05]` for Independent Indirect Neutral verb forms, `+[15]` for Conjunct Dubitative Preterite verb forms, and `+[17a]` for Immediate Imperative forms.

(5a) `chichii wii nituumiichisunaatik`
`PV/chi+PV/wii+PV/nituu+miichisuu+V+AI+Ind+[05]+Indir+Neu+2Sg`

(5b) `chaa saachihiitiwaahchipinaa`
`PV/chaa+saachihaau+V+TA+Cnj1+[15]+Dub+Prt+1Pl+2(P)lO`

(5c) `saachihiinaan`
`saachihaau+V+TA+Imp+[17a]+Imm+2(P)l+1PlO`

This almost entirely concatenative modeling strategy described above is made possible thanks to the exhaustive listing of the technical stems (both regular and changed) for each verb in the dictionary database, and the likewise comprehensive enumeration of all inflected forms for each subtype in the verb paradigm database, with one of the representations of each inflected form providing a partitioning into the technical stem and a technical suffix chunk. All the forms in the verb paradigm database have been verified in countless sessions with fluent East Cree Elders over decades.

Importantly, though the creation of the two databases has taken a substantial amount of meticulous human work and scrutiny, and while FST source code for the (relatively straight-forward) pre-stem component has been written by hand, the FST source code for the suffix component is generated in its entirety from the underlying two lexical databases, minimizing the risk for human typing error (when the underlying databases are error-free). Equally importantly, the automatic generation allows for easy generation of revised versions, if changes need to be implemented.

In terms of time required to create this this general FST architecture, the manual coding of the basic pre-stem morphology, and developing the scripts for automatically generating the post-stem FST source code has taken altogether 2 weeks of 3 people's work.

6 Conclusion

Having comprehensive, well-structured resources such as those described above, and people with appropriate programming and linguistic skills, the brute-force listing strategy presented in this paper is a surprisingly fast and efficient way of creating a finite-state computational model, to form a basis for subsequent development of practical end-user applications.

Acknowledgements

This work has been supported by funding from the Social Sciences and Humanities Research Council of Canada Partnership Development (890-2013-0047) and Insight (435-2014-1199) grants, a Carleton University FAAS research award, and Kule Institute for Advanced Study, University of Alberta, Research Cluster Grant.

References

Arppe, A., Junker M.-O. Harvey, C., and J. R. Valentine (in press). Algonquian verb paradigms. a case for systematicity and consistency. *Papers of the Algonquian Conference 47*.

Beesley, K. R. and L. Karttunen (2003). *Finite State Morphology*. CSLI Publications.

Blevins, James P. (2006) Word-based morphology. *Journal of Linguistics* 42: 531-573.

Collette, V. (2014) Description de la morphologie grammaticale du cri de l'Est (dialecte du Nord, Whapmagoostui) (unpublished doctoral thesis). Québec: Université Laval.

Hulden, M. (2009). Foma: A finite state toolkit and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, 29–32.

Junker, M.-O., MacKenzie, M., Bobbish-Salt, L., Duff, A., Salt, R., Blacksmith, A., Diamond, P., & Weistche, P. (Eds.). (2012). *The Eastern James Bay Cree Dictionary on the Web: English-Cree and Cree-English, French-Cree and Cree-French (Northern and Southern dialects)*. Retrieved from http://dictionary.eastcree.org/

Junker, M.-O. & MacKenzie, M. (2010-2015). *East Cree (Northern Dialect) Verb Conjugation* (4th ed.). Available at: http://verbn.eastcree.org/.

Junker, M.-O., Salt, L., & MacKenzie, M. (2015). *East Cree Verbs (Northern Dialect)*. [Revised and expanded from 2006 original edition] In *The Interactive East Cree Reference Grammar*. Retrieved from:
(a) [http://www.eastcree.org/cree/en/grammar/northern-dialect/verbs/cree-verb-inflection/initial-change/]
(b) [http://www.eastcree.org/cree/en/grammar/northern-dialect/verbs/cree-verb-stems/]

Junker, M-O. & Stewart, T. (2008). Building Search Engines for Algonquian Languages. In Karl S. Hele and Regna Darnell (eds*). Papers of the 39th Algonquian Conference*. London: University of Western Ontario Press, 378-411.

Lindén, K., E. Axelson, S. Hardwick, M. Silfverberg, and T. Pirinen (2011). HFST - Framework for Compiling and Applying Morphologies. Proceedings of *Second International Workshop on Systems and Frameworks for Computational Morphology (SFCM)*, 67-85.