Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5143–5152
Marseille, 20-25 June 2022

An Expanded Finite-State Transducer for Tsuut'ina Verbs

Joshua Holden[1], **Christopher Cox**[2], **Antti Arppe**[1]
University of Alberta[1] Carleton University[2]
holden1@ualberta.ca, christopher.cox@carleton.ca, arppe@ualberta.ca

Abstract

This paper describes the expansion of a finite state transducer (FST) for the transitive verb system of Tsuut'ina (ISO 639-3: `srs`), a Dene (Athabaskan) language spoken in Alberta, Canada. Dene languages have unique templatic morphology, in which lexical, inflectional and derivational tiers are interlaced. Drawing on data from close to 9,000 verbal wordforms, the expanded model can handle a great range of common and rare argument structure types, including ditransitive and uniquely Dene object experiencer verbs. While challenges of speed remain, this expansion shows the ability of FST modelling to handle morphology of this type, and the expanded FST shows great promise for community language applications such as a morphologically informed online dictionary and word predictor, and for further FST development.

Keywords: Na-Dene, Indigenous languages, finite state transducers, morphology

1. Introduction

This paper describes the development of a finite state transducer (FST) for the inflectional system of verbs in Tsuut'ina (ISO 639-3: `srs`), a highly endangered Dene (Athabaskan) language spoken near Calgary, Alberta, Canada. Arppe et al. (2017) outlined the creation of an prototype FST for intransitive verbs in Tsuut'ina, with only a few exemplary lexemes. The current paper describes the expansion of this FST architecture with a larger set of intransitive verbs as well as to encompass transitive verbs of various argument structures, which vastly increases the range and complexity of the morphology that the FST needs to represent. Full verb conjugations are needed for language learning resources, but manually creating paradigm tables would be impractical. The FST automates this and responds to the Tsuut'ina community's desire for computer-based language tools that are typical FST applications, such as a morphologically enhanced online dictionary, a spellchecker, and predictive word suggestions on mobile devices. Through the integration of this FST with the GiellaLT infrastructure, all of these applications are readily derivable from the FST described in this paper.[1]

FSTs (Beesley and Karttunen, 2003) use bidirectional rule-based mapping between sets of items, e.g. a string of characters as input, and a morphological analysis as output, or vice versa. FSTs are ideal for this work: they have open source implementations of compilers such as Foma (Hulden, 2009) (used in this project); they are compatible with most operating systems; their computational properties and end-user applications are well established (see Moshagen et al. (2013)) and Trosterud (2004; Trosterud (2006), Arppe et al. (2016), Antonsen et al. (2013), Johnson et al. (2013) for discussions of similar language-learning applications.). As data structures they are efficient for generating rule-based verb paradigms, which is advantageous given the limited data and corpora that are available for Tsuut'ina. However, caution is needed: rule-based paradigm generation is only as accurate as the morphological model used as input. The automatic generation of paradigms of a morphologically complex language can produce thousands of potential forms, but not all such forms are grammatically and pragmatically acceptable to speakers, who cannot verify them all.

1.1. Previous FST Work for Polysynthetic, Indigenous and Non-Concatenative Morphology

FSTs have been used numerous times to generate morphologically complex wordforms in polysynthetic Indigenous languages spoken in Canada and elsewhere in North America, for example Snoek et al. (2014) and Harrigan et al. (2017) for Plains Cree, Kazeminejad et al. (2017) for Arapaho, Kazantseva et al. (2018) for Kanyen'kéha (Mohawk), Lachler et al. (2018) for Northern Haida, Bowers et al. (2017) for Odawa, and Chen and Schwartz (2018) for Yup'ik. Further afield, Antonsen et al. (2013) show that FSTs are well-suited to computer-assisted language learning (CALL) tools for Northern Saami, as does Hurskainen (2009) for Swahili.

While the above projects show that FSTs are extremely promising for developing computational models and end-user applications for First Nations languages, with their varying origins and morphological characteristics, Dene languages pose a particular challenge, as they are virtually unique in having a "templatic" morphological structure, to be described in detail in sections 1.2 and 2. This poses significant and unique challenges to building an FST (see Sections 1.2 and 2 below; see also Hulden and Bischoff (2008) for an early exploration of these questions), some of which were addressed in Arppe et al. (2017) for the intransitive verb model (see section 1.2 below), but many others of which are described in

[1] `https://giellalt.uit.no/infra/`

this article.

1.2. Basic FST Architecture for Tsuut'ina Intransitive Verbs

Arppe et al. (2017) presented the creation of the core architecture of the FST for Tsuut'ina. This subsection 1.2 will give the briefest review of features of Dene verb template that had to be modelled for the FST. Section 2 provides a more thorough overview of Tsuut'ina inflection and participant marking relevant to the current expansion. Then the core features of the FST architecture are described and the next steps needed for the expansion.

Dene verbs do not use a simple, linear lexical-derivational-inflectional concatenation; rather, the three types of prefixes are interlaced (see Figure 2). The lexical tier or "verb theme" consists of a stem on the right edge, a voice/valence prefix directly to its left, and up to three prefix positions: inner lexical, a middle "areal" slot (the areal prefix is a historical agreement prefix referring to a place or situation, but is lexicalized as part of many verb themes), and an outer lexical position. "Outer", "middle" and "inner" lexical prefixes are defined in relation to inflection sites. (This is a radical simplification that avoids more technical Athabaskan prefix terminology; Rice (2000), Hoijer (1945), and Sapir and Hoijer (1967) are basic comparative references.) These lexical zones have degrees of internal complexity, but this does not need to be accounted for in an inflectional FST, which simply must generate and accept correct verb paradigms.

The inflectional tier includes categories of viewpoint aspect, mood, as well as number and person agreement for subjects and objects. For the basic, "skeleton" FST described in Arppe et al. (2017), it was determined that three insertion points in the verb word were necessary to accurately generate intransitive inflections: 1) the outer inflectional zone directly to the right of the outer lexical prefixes (to handle the distributive plural), 2) the middle zone between the areal and the inner lexical prefixes (where third-person unspecified subject prefix *ts'i–*, and third-person plural subject *gi–* are found), and finally 3) the inner, TAMA (tense-aspect-mood-agreement) chunk zone, between the inner lexical zone and the stem-classifier combination. There are many allomorphic co-occurrence restrictions between TAMA chunks and lexical portions. See Section 2 for a full overview of this area.

Computationally, FST uses a 'chunking', or portmanteau, approach to the combination of aspect, mood and agreement prefixes to the left of the classifier, sequences called "TAMA chunks" in this paper, even though there is no true tense marking in this zone. The classifier is a single segment, one of which is the vowel *i* in Tsuut'ina; many of the viewpoint aspect prefixes are zero-marked, or marked by vowels, or by a vowel plus a glide (such as *yi-*). This whole area is often collapsed into single syllables whose internal complexity

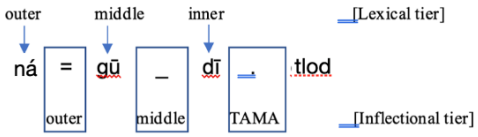

Figure 1: FST conventions for inflectional zones of *náguditlod* 's/he jumps down'

(see Rice (2001)) would be quite hard to model in a fully decomposed form. While the TAMA chunking approach has some support from linguistic studies—see Rice et al. (2002) for psycholinguistic results involving Dene Sųłiné, Young and Morgan (1987), Faltz (1998), and McDonough (2000) for Navajo, Holden (2013) for Dene Sųłiné and Leer (1999) (inter alia) for comparison across multiple Dene languages, from the perspective of FST modelling it simply made the complexity more manageable, as only the junctures between the TAMA chunk and the surrounding morphemes are visible to the model.

Because both the lexical and inflectional tiers of the morphology are discontinuous, we make use of three separate finite-state models for each of the three inflectional zones, a fourth for the lexical tier (including the stem at the right edge and the possible preceding discontinous lexical prefixes), specified with the *lexc* formalism (Beesley and Karttunen, 2003). Using finite-state operations, the three inflectional component FSTs are inserted in the appropriate slots within the lexical tier (see Arppe et al. (2017, 58, ex. 2) for the master specification).[2] The morphophonemic processes are modeled with contextual rewrite rules using *xfst* (Beesley and Karttunen, 2003), with the resultant fifth FSM then composed together with the morphological FSM. A sixth FSM is concatenated with this to link flag-diacritics with morphological feature tags. Figure 2 outlines the structure of these six interlocking constituent FSMs.

The slots for inflectional morphemes within the lexical tier are indicated by a specific notation (see Figure 1 above): "." (period) stands for the inner boundary, where the TAMA chunk is to be inserted, "_" (underscore) for the middle boundary (where object agreement and third-person subjects are located), and "=" (equal sign) for the outer boundary (where the distributive plural prefix can occur). A system of flag diacritics then filter out disallowed combinations (i.e. implementing co-occurrence restrictions).

1.3. Next Steps

Arppe et al. (2017) demonstrated that it was possible to model Dene verb morphology using this architecture,

[2]One should note that this "flat" lexical model is distinct from that developed for Upper Tanana (another Dene language), which employs a stem-based model using verb theme categories (Kari, 1979), which treats the addition of any material to the left of the stem as derivational prefixation.

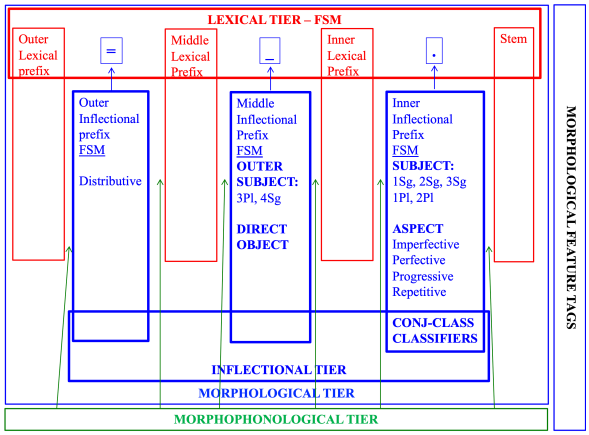

Figure 2: Core FST architecture

at least for a handful of intransitive verbs, but the true test of its robustness would come with the expansion to transitive verbs and a much wider sample of many hundreds of verbs with many new argument structures. The current expansion handles 1,557 lemmas (approximately) from the Onespot-Sapir glossary (see Section 3.1 below).[3] To describe this, the following sections will review the structure of the verb, with special attention to areas that impact the design of an FST.

2. An Overview of Participant Marking in Tsuut'ina Verbal Morphology

While the overall architecture described in the preceding section (and in Arppe et al. (2017) and Lovick et al. (2018)) is sufficient for modelling intransitive verb forms in Tsuut'ina (and Upper Tanana), in which grammatical subjects may be realized through combinations of outer, middle, and inner inflectional prefixes, it does not address the morphological realization of additional, non-subject discourse participants in more complex inflected forms. Indeed, a single inflected form such as (1) can index up to three event participants with pronominal markers appearing in several places in the verb-word (in boldface):[4]

(1) moghàninischúd-áát'a

mi- *oghà-* ***ni-*** ***nis-***
3SG.IO to 2SG.DO 1SG.SBJ:*ni*.IPFV

s- *chúd* *=i* *=át'a*
VV feed.IPFV FUT ASRT

'I am going to feed you to it.' (Doris Roan, speaker; October 23, 2012)

[3] The full source code for the Tsuut'ina FST is available at `https://github.com/giellalt/lang-srs`.

[4] In additional to the standard labels defined in the Leipzig Glossing Rules, the following abbreviations are also used in this paper: AR 'areal', ASRT 'assertion', CON 'conative', DIST 'distributive', DO 'direct object', IO 'indirect object', POT 'potential (optative)', STAT 'stative' TERM 'terminative', VV 'voice/valence marker', 3D 'distal third-person object'.

In this example, inflectional markers related to event participants appear at several points: (1) at the far left-edge of the verb-word, with an 3SG indirect object *mi-* marked before the incorporated postposition *oghà-* 'to'; (2) after that postposition, with the 2SG direct object *ni-*; and (3) immediately before the verb stem, with a portmanteau morpheme *nis-* representing the 1SG subject form of the *ni-*imperfective paradigm with an *s-* voice-valence marker. While ditransitive forms such as this in which three distinct event participants are realized morphologically are relatively rare in Tsuut'ina, pronominal marking of event participants in each of these three positions is not: verb forms containing object marking are frequent, and thus need to be modelled.

In order to extend the previously proposed FST architecture to represent non-subject event participants, then, several other linguistic facts need to be taken into consideration:

1. Object-marking patterns need to be implemented to represent both the indirect and direct object morphology seen in (1) above, which was not present in previous models that explored the representation of discontinuous morphology (Arppe et al., 2017) and common derivational paths (Lovick et al., 2018). As we note below, this sometimes involves defining an additional inflectional FSM (e.g., for indirect object morphology not present in prior models) or refining the existing inflectional FSMs to process both subject and object markers.

2. Interactions between subject and object-marking morphology need to be represented adequately in an expanded computational model. As in many other Dene languages, the forms of object markers often depend on subject person and number, with third-person objects having different realizations when acted on by third-person vs. non-third-person subjects (see Section 3.2.2). Similarly, some subject markers' positions may differ based on the presence or absence of certain object markers. This is the case with *gi-* 3PL, which typically appears as a middle prefix (2a), but can also appear at the left edge of the word when combining with *yi-* 3D in 3PL>3SG.IO contexts (2b):

(2) a. soghàgistà

si- *oghà-* *gi-* *is-*
1SG.IO to 3PL 3.SBJ:*ni*.IPFV:*s*.VV

tà
handle-animate.IPFV

'they will give something (animate) to me.'

b. giyoghàstà

gi- *yi-* *oghà-* *is-*
3PL 3D.IO to 3.SBJ:*ni*.IPFV:*s*.VV
tà
handle-animate.IPFV

'they will give something (animate) to him/her/it.'

Other inflectional markers already defined in previous, intransitive-only models of Tsuut'ina verbal morphology also require adaptation to reflect their distribution when other event participants are present. The distributive plural *dà-*, an inflectional prefix, may only appear when one of the participants is plural or impersonal. Prior models would therefore reject forms with singular subjects that contained *dà-*, whereas a revised model that incorporates additional event participants also needs to accept cases where the subject is singular, but the object is plural or impersonal (e.g., ***dà**gimiyis?í* 'I saw each and every one of them');

3. Individual verb lexemes in Tsuut'ina differ in the number of participants involved in the associated verbal event and how those participants are realized morphologically with the above subject, direct object, and indirect object markers. Section 3.2 begins by detailing the participant marking patterns in Tsuut'ina for which an expanded model must account.

   While many of these classes resemble cross-linguistically common argument marking patterns—intransitive verbs realized with prototypically subject-related inflection, or transitive verbs realized with both direct object and subject inflection—other less familiar patterns are also attested in Tsuut'ina. (See Sections 3.2.5–3.2.7 below).

An expansion of the FST model of Tsuut'ina verbal morphology that aims to include all the attested patterns of verbal participant marking thus presents a potentially valuable "stress test" of the existing finite-state architecture for Dene languages, particularly in its ability to generate and recognize a much wider range of verb forms. As we note below, applying such a model to a much larger lexical sample also raises questions about current conceptions of what constitutes a lemma in morphologically rich languages such as Tsuut'ina, and of where the boundary lies between inflectional and derivational information in Dene verbal morphology.

3. Expansion of the Inflectional Tsuut'ina FST

The current expansion is meant to generate full inflectional paradigms for a comprehensive range of verbs as represented by the Onespot-Sapir glossary (see Section 3.1).[5] This includes all transitive, oblique object and ditransitive cases in this database. Numerous morphological complexities specific to transitivity in Dene languages must be handled, along with a multiplicity of minor argument structure with distinct prefixing patterns (see below). A model of the derivational morphology is beyond the scope of the current model.

We have not followed, however, a very rigorous distinction between inflection and derivation for our FST model. In particular, several aspectual categories normally seen as derivational are included in the inflectional FST either because it made the modelling itself much easier (see discussion of the transitional or inchoative and semelfactive prefixes in Section 3.2.1 and of the conative in Section 3.2.5) or in order to make the model's output fit with Tsuut'ina community language teaching traditions and priorities (see Section 3.1's discussion of the repetitive and progressive aspects). The following section 3.1 will review the language source data used for this expansion, and Section 3.2 will address the new characteristics of FST modelling of transitive verbs and the other argument structure types.

3.1. Re-elicitation and Organization of the Onespot-Sapir Glossary

The source material was an unpublished glossary collected by Edward Sapir and John Whitney-Onespot in Tsuut'ina in 1922 that contains numerous verb paradigms. In the 1990s, Tsuut'ina community linguist and speaker Bruce Starlight and collaborator Gary Donovan started transcribing and editing Sapir and Whitney-Onespot's notebooks, occasionally expanding on incomplete paradigms or adding related words. This curation and editorial process is outlined in further detail in Starlight et al. (2016).

Paradigms were transferred into a preliminary lexical database and labelled for argument structure, transitivity, aspect, and TAMA chunk subtype by co-author Holden (sometimes in consultation with co-author Cox), and the specific allomorphic combination of the stem + lexical prefixes for each TAMA value was recorded (with some temporarily excluded because their TAMA pattern was unclear, or some other factor that required verification by speakers).

The next step was lemmatization. Normally any non-inflectional material (barring noted exceptions) required the creation of a lemma, which would stand for a group of inflected wordforms. The lemmas are thus similar to lexemes, with some caveats: we treated repetitive and progressive aspects (linguistically derivational) as inflections to make the paradigms compatible with Tsuut'ina community language education

[5] It should be noted that the current FSM has not yet implemented the optative (aka potential) mood forms due to their rarity in the Onespot-Sapir data. However, because this inflection is highly regular, we are confident in being able to include it soon.

preferences: verbs are typically taught as a "basic set" of four variants: imperfective (aka non-past), perfective (aka past), progressive, and repetitive. Taking a strict view of the inflection-derivation distinction would have broken up these sets. Some Onespot-Sapir paradigms seemed to be split between inceptive and progressive aspects (both derivational), with no obvious meaning change. Following a pedantic division of inflection and derivation would have broken up what is functionally a natural complete set into two incomplete paradigms, in contrast with community preferences, and producing some inflections that are not used or common. The third-person singular wordform was chosen as this citation form, a convention already adopted for the intransitive model (Arppe et al., 2017).

3.2. Expansions Needed for the Onespot-Sapir Glossary

The Onespot-Sapir list contains close to 9 thousand verbal wordforms with many argument structure types, of which several are new to the FST model. The *lexc* excerpt below shows the argument structures and associated flag diacritics, with all except the first two being novel to the expanded FST.

```
LEXICON Root
@U.VALENCE.IMPERSONAL@      NoDistributive;
@U.VALENCE.INTRANSITIVE@    SubjectOnly;
@U.VALENCE.TRANSITIVE@      SubjectAndDirectObject;
@U.VALENCE.OBLIQUEOBJECT@   SubjectAndObliqueObject;
@U.VALENCE.DITRANSITIVE@
                SubjectDirectObjectOrObliqueObject;
@U.VALENCE.EXPERIENCER@     ObliqueObjectOnly;
@U.VALENCE.TRANSITIONAL@    SubjectAndDirectObject;
```

Given that inflectional material can be found in several zones of the verb, expanding the range of verbs and argument structures requires an additional inflectional prefix position as well as an increase in the range of prefixes found in existing slots. The expansions could therefore be sorted by argument structure or by prefix zone. To avoid repetition, we will take a blended approach: in sections 3.2.1–3.2.4 below, we will first lay out additions to each of the FST inflectional zones as well the new zone that came with the expansion, focusing on cross-linguistically well known argument structure types, before wrapping up the section with the less common and more Dene-specific argument structures (subsections 3.2.5–3.2.7), describing their characteristics as well as the prefix zones they occupy.

3.2.1. Inner zone TAMA inflections

The point directly to the left of the stem-classifier combination is the most complex site of inflection in Dene languages, where aspect, mood and person agreement are found. Prefix sequences here are treated as cumulative morphemes informally called "TAMA chunks".[6] The most frequent morphemes here are the imperfective and perfective aspects and the optative or potential mood. (The progressive aspect, which is strictly speaking derivational in Dene languages, is also included as a value of this "category".) The TAMA chunking approach was adopted for practical reasons laid out in Section 1.2 above; for the same reason, a few surrounding lexical prefixes (the classifier to the right of the TAMA inflection, and the transitional prefix to its left) are included with the TAMA sequence. This approach to the inner prefixes was already in place for the intransitive model (minus the transitional paradigms), but the expansion saw the addition of a much wider range of imperfective and perfective allomorphs.

Dene languages have derivational situation aspect marking (see Rice (2000) for a further discussion) In dynamic, transitive transitional verbs such as (2b) above, the low tone is present with the *mi–* object prefix (e.g., *nà**gimì**nistà* 'I am setting them (*gimì-*, with low-tone *ì-*) down'), but absent with a full nominal complement (e.g., *tłích'áká nànistà* 'I am setting the dogs down' (no low tone)). To handle this, we added the specific argument structure "transitional transitive" in addition to distinct transitional TAMA chunk allomorphs.

3.2.2. Middle Zone Inflections

The "outer subject" prefixes *ts'i-* 'impersonal subject' and *gi-* 'third person plural subject' in this zone were already in the intransitive model in Arppe et al. (2017). This is the primary insertion point for the direct object prefixes, so adding these was a significant expansion to this zone. Flag diacritics for some of these values and the corresponding prefixes are exemplified in the *lexc* excerpt below.

```
LEXICON Transitive-Markers
! 1SG direct obj. (e.g., siyí?í "you saw me")
@U.OBJECTPERSON.1@
  @U.OBJECTNUMBER.SG@
  @P.PREFIX.MIDDLE@si     Outer-Subjects;
! 2SG direct obj. (e.g., niyis?í "I saw you")
@U.OBJECTPERSON.2@
  @U.OBJECTNUMBER.SG@
  @P.PREFIX.MIDDLE@ni     Outer-Subjects;
```

The decision was made to handle third-person direct objects in their own continuation lexicon, as shown in the next set of *lexc* definitions, due to a number of unique complexities, one of which being a proximal/distal third-person distinction. A distal third-person object is one acted on by another third-person actant. In these cases, the usual third-person object marker *mi–* is replaced by the distal third person object marker *yi–*. For example, in the verb *miyaà?í* 'we saw it' (*mi-* 3.DO + *yaà?í* see.1PL.PFV), *mi–* is the third-person standard or proximal object prefix. In the case of two third person actants, the verb form would be *yá?í* "s/he saw it" (*yi-* 3D + *yí?í* see.3.PFV), where *mi–* is replaced by distal third-person object marker *yi–*.

[6] Although from a strictly linguistic standpoint most Dene languages, including Tsuut'ina, arguably do not express tense in this zone, some of these morphemes are referred to with tense names (past=PFV, present=IPFV, future=POT) in Tsuut'ina language teaching, and we adopt the "fuzzy" term TAMA here.

Tsuut'ina object agreement prefixes are not required when a full nominal object is present. In practice this affects almost exclusively the third person inflections (first- and second-person pronouns are infrequent and emphatic in Tsuut'ina). For example, in the sentence *ístłí yí?í* "s/he saw a horse", *ístłí* means 'horse' and *yí?í* means 's/he saw (it)'. Another such contrast is seen in the *lexc* definitions below. Because the overt nominal *ístłí* 'horse' is present, the verb has no third-person direct object prefix. In contrast, the form *yá?í* "s/he saw it" mentioned above (with no overt direct object) the third-person distal prefix is present (*yá?í* = *yi-* 3D + *yí?í* see.3.PFV). In the *lexc* file we treated these as two allomorphs of the third-person object prefixes, one of which was zero, as shown in the *lexc* definitions reproduced below:

```
LEXICON 3SG-Direct-Objects
@U.SUBJECTNUMBER.SG@
  @U.SUBJECTPERSON.3@
  @U.DIRECTOBJECT.NOMINAL@ Filter-Transitives;
@U.SUBJECTNUMBER.SG@
  @U.SUBJECTPERSON.3@
  @U.DIRECTOBJECT.NONE@
  @P.PREFIX.MIDDLE@yi      Filter-Transitives;
```

The flag `@U.DIRECTOBJECT.NOMINAL@` tells the FST to use the zero allomorph (i.e., no prefix) when an object noun is present, while the flag `@U.DIRECTOBJECT.NONE@` is followed by `@P.PREFIX.MIDDLE@yi`, telling the FST to add the object prefix *yi–* in the absence of a direct object noun. The reflexive and reciprocal objects are handled in their own continuation lexica as well, due to a number of linguistic complexities (see *lexc* definitions below).

```
LEXICON Reflexive-Direct-Objects
@U.SUBJECTNUMBER.SG@
  @U.SUBJECTPERSON.1@
  @U.OBJECTNUMBER.SG@idi      Filter-Transitives;
@U.SUBJECTNUMBER.PL@
  @U.SUBJECTPERSON.3@
  @U.OBJECTNUMBER.PL@
  @D.GI@@P.GI.ON@igidi        Filter-Transitives;
@U.SUBJECTNUMBER.SG@
  @U.SUBJECTPERSON.4@
  @U.OBJECTNUMBER.SG@its'idi  Filter-Transitives;

LEXICON Reciprocal-Direct-Objects
@U.OBJECTNUMBER.PL@
  @U.SUBJECTNUMBER.PL@
  @U.SUBJECTPERSON.1@átłi     Filter-Transitives;
```

First of all, the reflexive prefix *idi-* is actually two prefixes, which can be interrupted by the impersonal subject *ts'i–* (*its'idi–*) or third-person plural subject *gi–* (*igidi–*). Secondly, the reflexive prefixes trigger a change in the classifier portion of the TAMA inflection chunk. Furthermore, reciprocal forms are limited to plural and impersonal subjects.

For transitive verbs, we also had to filter out subject-object combinations that were semantically implausible. The following *lexc* snippet shows part of this lexicon, which uses flag diacritics to allow only felicitous subject-object combinations such as 1SG.SBJ>2SG.OBJ to proceed:

```
LEXICON Filter-Transitives
@U.SUBJECTNUMBER.SG@
  @U.SUBJECTPERSON.1@
  @U.OBJECTNUMBER.SG@
  @U.OBJECTPERSON.2@   Filter-Ditransitives;
  (...)
```

3.2.3. Outer Zone Inflections

The outer insertion point only takes the *dà–* distributive prefix. While this inflection point was already present in the initial model, the current expansion meant a new set of rules for the distributive to account for restrictions and a wider range of possible uses.

First of all, while *dà–* usually pluralizes the subject, for transitive stems the distributive can refer to plurality of either the subject or the object. In this case, the lexicon shown below is needed, which excludes only the singular subject-singular object combination. In other transitive cases, the second line would be marked `NoDistributive` if subject and direct object are both singular.

```
LEXICON SubjectAndDirectObject
@U.SUBJECTNUMBER.PL@
  @U.OBJECTNUMBER.PL@   Distributive;
@U.SUBJECTNUMBER.SG@
  @U.OBJECTNUMBER.PL@   Distributive;
@U.SUBJECTNUMBER.PL@
  @U.OBJECTNUMBER.SG@   Distributive;
@U.SUBJECTPERSON.4@
  @U.OBJECTNUMBER.SG@   Distributive;
@U.SUBJECTNUMBER.SG@
  @U.OBJECTNUMBER.SG@   NoDistributive;
```

Furthermore, for any verbs whose object markers occur in the left-edge (historically oblique) inflection zone (see below), the distributive can refer to either subjects or (oblique or formerly oblique) objects. This is accounted for in a special lexicon for subjects with oblique objects, shown in the following *lexc* definitions.

```
LEXICON SubjectAndObliqueObject
@U.SUBJECTNUMBER.PL@
  @U.OBLIQUENUMBER.PL@   Distributive;
@U.SUBJECTNUMBER.SG@
  @U.OBLIQUENUMBER.PL@   Distributive;
@U.SUBJECTNUMBER.PL@
  @U.OBLIQUENUMBER.SG@   Distributive;
@U.SUBJECTPERSON.4@
  @U.OBLIQUENUMBER.SG@   Distributive;
@U.SUBJECTNUMBER.SG@
  @U.OBLIQUENUMBER.SG@   NoDistributive;
```

3.2.4. Left-edge Inflections

In this expanded model, a fourth inflectional zone was necessary at the left edge of the verb word. If you recall the verb structure outlined in Section 1.2 and Section 2, Dene verb themes frequently incorporate

postpositions at or toward their left edge (in the outer lexical zone). Often the postposition/preverb is lexicalized so its meaning is not transparent. The incorporation of a postposition affects the verb's valence. This is the case notably with transfer verbs where the postposition holding the recipient is the one incorporated in the verb. This was seen in example (1) above, where at the left edge the third-person object *mi–* is added to refer to the object of the feeding, while the normal direct object position of *s+chut* contains second-person object *ni–* 'you', resulting in a ditransitive verb. The verb "inherits" the incorporated postposition complement, resulting in double object marking (if it is not a phrase). The FST assigns the flag diacritic `@U.VALENCE.DITRANSITIVE@` to these cases. If the original analytical construction was a two-actant verb with an oblique object complement, the incorporation of the postposition results in a direct transitive verb where the direct object inflection occurs at the left edge, rather than the standard middle position. The FST treats these as oblique object verbs with the flag diacritic `@U.VALENCE.OBLIQUEOBJECT@`. As with the case of the direct objects, the object prefix is suppressed when an overt nominal is present, so the FST must implement a null allomorph in these cases, as shown in the lexicon below.

```
LEXICON 3SG-Oblique-Objects
@D.SUBJECTPERSON.3@
  @U.OBLIQUEOBJECT.NOMINAL@  #;
@D.SUBJECTPERSON.3@
  @U.OBLIQUEOBJECT.NONE@mi   #;
```

As with the transitive verbs, the 3rd person, reflexive and reciprocal were handled in separate lexica, because the same rules regarding distal/proximal alternations and classifier changes apply.

3.2.5. Conative Paradigms

The conative is a "situation aspect" derivation (see Rice (2000, 260–263)). At its most compositional, it contributes a meaning loosely glossable as 'attempt' or 'at' to the verb (e.g. 'shoot at' versus just 'shoot'). In other cases the conative is vestigial or it is hard to see what its exact contribution is, but it is associated with dynamic transitive verbs, as in (3), where it appears as a high tone *í–* morpheme that merges with the preceding object prefix.

(3) yízi
yi– í– zi
3D CON call.IPFV
's/he is calling him/her'

Formally the conative appears as a high tone or *í–* morpheme directly to the left of the TAMA sequence in most Dene languages, but in Tsuut'ina, unusually, it occurs between the direct object prefixes and the third-person subject prefixes, directly in the middle of the middle zone of the FST inflectional model, so the chunking approach with TAMA will not work. This became apparent with the expanded inventory of verbs from the Onespot-Sapir glossary. Because the FST architecture was not designed to break up the middle inflectional zone, a unique flag diacritic was added to conative verbs, `@U.CONATIVE.ON@`, which ensured that the conative prefix was inserted in verbs that were lexically specified to be of this type.

3.2.6. Object Experiencer Verbs

The left-edge position is also an inflection site used to generate what we termed object experiencer verbs– single-actant predicates whose object markers refer to the semantic experiencer of an event, and the subject position in the TAMA and middle positions are empty, arguably filled with a zero-marked dummy third person inflection. These often refer to states or events such as sickness or emotions, where there is no tangible agent that could be identified as acting on the (morphologically object-encoded) experiencer. In *tamíyiłił* 'he/she/it is floating' in (4), for example, the person floating is expressed by the object marker *mi-*.

(4) tamíyiłił
ta– mi– í– yi– 0- łił
on 3.DO CON 3.PROG VV float
'he/she/it is floating (in one spot)'

Morphologically there are two types of object experiencer verbs: those whose object markers occur in the standard middle slot, termed direct object experiencer verbs, and those inflected at the left edge, where the object marker derives, at last historically, from the object of an incorporated postposition, deemed oblique object experiencer verbs. (4) above is a direct object experiencer verb, while (5) below is an oblique object experiencer verb. In (4) below, the third-person plural prefixes *gimi–* refer to the people experiencing sickness.

(5) gimádàgúdìlo
gi– mi– á– dà– gú– dì– 0–
3PL 3.IO by DIST AR STAT 3.IPFV
lo
many.lie
'they are all sick'

The meaning 'sick' is produced from the lexicalized combination of the stem *-lo* 'many lie', the prefix *gú-*, and the incorporated postposition *á-* 'by'.

To implement object experiencer verbs, the FST must exclude any subject other than third person and specify the proximal third-person object for third person experiencer. (There can be no distal third person marking for these themes, as the morphological subject is suppressed.) We used the flag diacritics `@U.VALENCE.DO-EXPERIENCER@` and `@U.VALENCE.OO-EXPERIENCER@` for direct- and

oblique-object experiencer verbs, respectively, to prevent any subject person-number combinations other than third-person singular (with zero-marked imperfective allomorph) from being realized.

3.2.7. Restricted Argument Structures

Additional minor argument structures were needed for verbs which, for pragmatic or other reasons, had restrictions on non-third persons as either subjects or objects. For example, impersonal verbs (typically referring to agentless events such as weather or celestial situation) such as *gudìsghál* 'it is getting dark (outside)' can only take a dummy singular third-person, and no distributive marker. The flag diacritic `@U.VALENCE.IMPERSONAL@` was used for such verbs to prevent other person forms. There are also third-person subject-only verbs which, while not 'impersonal' in the above agentless sense, would be pragmatically odd with first or second persons. This is the case, for instance, of *taánimòsh* 'it is boiling' (barring a story with anthropomorphic water or kettles). Analogously, the transitive verb *íyin* 'to sing (it)', can only take third-person objects. To prevent first- and second-person inflections from being included in these cases, sequences of flag diacritics such as `@R.OBJECTNUMBER.SG@@R.OBJECTPERSON.3@` were added to require that only third-person singular forms were generated and recognized.

4. Size and Performance

The current extended lexicon contains altogether 1,557 lexemes, distributed among various argument structure types and subtypes as shown in Table 1. On average, these lexemes have 1.62 suppletive verb theme variants, with a median of 1.0 and a maximum of 11 theme allomorphs per lexeme (resulting from multiple allomorphic variants for the same aspectual value). This means that some suppletive forms are missing from the Onespot-Sapir glossary, and must be elicited from fluent speakers.

$n_{lexemes}$	$\overline{x}_{allomorphs}$	Argument structure
842	1.50	Intransitive
630	1.83	Transitive
39	1.15	Transitive-SubjSuppr[essed]
25	1.36	ObliqueObjectExperiencer
9	1.67	ObliqueObject
3	1.33	Transitive-Conative
2	2.00	Intransitive-SubjPl[ural]Only
2	1.00	Transitive-D[irect]Obj3SgOnly
2	1.00	DirectObjectExperiencer
1	4.00	Ditransitive
1	1.00	Intransitive-SubjSuppr[essed]
1	1.00	Intransitive-Subj3[rdPerson]Only
1557	1.62	TOTAL

Table 1: Counts of lemmas and average theme allomorphs for different argument structure types

When compiled with Foma, this entire FST is quite large, at 63.2 MB in overall size, with 1,664,399 states, 4,143,483 arcs, and more than 9 x 10^{18} paths (before pruning based on flag-diacritics). The number of verbal wordforms that this FST covers is nevertheless finite, adding up to 1,472,669 forms in total that take 15 minutes to output using the *pairs* command in Foma. In terms of speed, this expanded FST is noticeably slow in analyzing wordforms (2min 2.57s for 1000 random word-forms, on a 2020 Macbook Pro with 32GB of RAM and a 2.3 GHz Quad-Core Intel Core i7 CPU) but alarmingly snail-paced in generating the same 1000 wordforms (43min 6.27s). This is caused by the original design of the FST, where the flag-diacritics that constrain the acceptable strings are largely specified at the right edge of the FST, thus resulting during the FST lookup, reading the network from left-to-right, in the generation of a huge number of possible strings before encountering the limiting flag diacritics. In linguistic analysis, though, the wordform string strongly restricts possible analyses, resulting in a more acceptable but still slow speed. A possible fix would be to specify constraining flag-diacritics at the left edge of the FST, at the beginning of the lexical tier, which, though feasible, would involve reconfiguring the matching flag-diacritics in the three inflectional FSTs, which would then follow the constraining flag-diacritics (e.g. switching P-flags into R-flags). Another solution is to enumerate all the wordform-analysis pairs, as their number is finite, and subsequently create a wordform-based FST. We have already attempted this, producing a slightly smaller FST of 44.4 MB, with 2,350,317 states, 2,910,952 arcs, and 1,455,806 paths, without any flag-diacritics. Most importantly, this word-form based FST is many degrees of magnitude faster than the original flag-based one, analyzing 1000 random word-forms in 0.663s and generating those same word-forms in 1.411s; for 100k random wordforms, the analysis and generation speeds are 3.713s and 1min 16.46s, respectively.

5. Conclusion and next steps

We have demonstrated that a working full-scale finite-state model can be created for Tsuut'ina, and thus Dene languages in general, implementing all argument types of verbs, the morphologically most complex word class, with a comprehensive lexicon. This expansion has its performance challenges, i.e. its relatively large size and concerningly slow speed would render real-time paradigm generation by the model impractical. However, we have envisioned solutions to these challenges. Since the set of inflections is finite, we can generate the entire vocabulary that the more intricate flag-based FST specifies, and use that as a basis for a wordform-based FST that is acceptable speed-wise. Another pressing task is to elicit or uncover missing allomorphic variants of the verb themes that remained unconfirmed.

6. Acknowledgements

This work represents one part of a long-term collaboration with partners at Tsuut'ina Nation to develop

language technology in support of local language education, documentation, and revitalization initiatives. We gratefully acknowledge the leadership, support, and linguistic expertise of Bruce Starlight (Tsuut'ina Language Commissioner) and Janelle Crane-Starlight (Director, Tsuut'ina Gunaha Institute) and their offices throughout this project. This work has been funded by a Partnership Grant (895-2019-1012) from the Social Sciences and Humanities Research Council (SSHRC) of Canada.

7. Bibliographical References

Antonsen, L., Johnson, R., Trosterud, T., and Uibo, H. (2013). Generating modular grammar exercises with finite-state transducers. In *Proceedings of the second workshop on NLP for computer-assisted language learning at NODALIDA*, pages 27–38.

Arppe, A., Lachler, J., Trosterud, T., Antonsen, L., and Moshagen, S. N. M. (2016). Basic language resource kits for endangered languages: A case study of plains cree. In *CCURL 2016: Collaboration and Computing for Under-Resourced Languages: Towards an Alliance for Digital Language Diversity (LREC 2016 Workshop), Portoroz, Slovenia*. European Language Resource Association.

Arppe, A., Cox, C., Hulden, M., Lachler, J., Moshagen, S. N., Silfverberg, M., and Trosterud, T. (2017). Computational modeling of the verb in Dene languages. the case of Tsuut'ina. In *Working Papers in Athabascan Linguistics ("Red Book" series)*, Fairbanks. Alaska Native Language Center.

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

Bowers, D., Arppe, A., Lachler, J., Moshagen, S., and Trosterud, T. (2017). A morphological parser for Odawa. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 1–9.

Chen, E. and Schwartz, L. (2018). A morphological analyzer for St. Lawrence Island / Central Siberian Yupik. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Faltz, L. M. (1998). *The Navajo verb: A grammar for students and scholars*. University of New Mexico Press, Albuquerque, NM.

Harrigan, A. G., Schmirler, K., Arppe, A., Antonsen, L., Trosterud, T., and Wolvengrey, A. (2017). Learning from the computational modelling of Plains Cree verbs. *Morphology*, 27(4):565–598.

Hoijer, H. (1945). The Apachean verb, part I: Verb structure and pronominal prefixes. *International Journal of American Linguistics*, 11(4):193–203.

Holden, J. (2013). *Benasní-I Remember: Dene Sųłiné Oral Histories with Morphological Analysis*. Brill.

Hulden, M. and Bischoff, S. (2008). An experiment in computational parsing of the navajo verb. *Coyote Papers*, 16.

Hulden, M. (2009). Foma: a finite-state compiler and library. In *Proceedings of EACL*, pages 29—32, Athens. Association for Computational Linguistics.

Hurskainen, A. (2009). Intelligent computer-assisted language learning: Implementation to Swahili. *Technical reports on language technology*, 3.

Johnson, R., Antonsen, L., and Trosterud, T. (2013). Using finite state transducers for making efficient reading comprehension dictionaries. In *Proceedings of the 19th Nordic conference of computational linguistics (NODALIDA 2013)*, pages 59–71.

Kari, J. (1979). *Athabaskan Verb Theme Categories: Ahtna*, volume 2 of *Alaska Native Language Center Research Paper*. Alaska Native Language Center, Fairbanks, Alaska.

Kazantseva, A., Maracle, O. B., and Pine, A. (2018). Kawennón:nis: the Wordmaker for Kanyen'kéha. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 53–64.

Kazeminejad, G., Cowell, A., and Hulden, M. (2017). Creating lexical resources for polysynthetic languages–the case of arapaho. pages 10–18, 01.

Lachler, J., Antonsen, L., Trosterud, T., Moshagen, S., and Arppe, A. (2018). Modeling Northern Haida verb morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Leer, J. (1999). Comparative Athabaskan class notes. Unpublished manuscript CA965L1999a, Alaska Native Language Archive.

Lovick, O., Cox, C., Silfverberg, M., Arppe, A., and Hulden, M. (2018). A computational architecture for the morphology of upper tanana. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

McDonough, J. M. (2000). On a bipartite model of the Athabaskan verb. In Theodore B. Fernald et al., editors, *The Athabaskan languages: Perspectives on a Native American language family*, number 24 in Oxford Studies in Anthropological Linguistics, pages 139–166. Oxford University Press, Oxford.

Moshagen, S., Pirinen, T. A., and Trosterud, T. (2013). Building an open-source development infrastructure for language technology projects. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 343–352.

Rice, S., Libben, G., and Derwing, B. (2002). Morphological representation in an endangered, polysynthetic language. *Brain and Language*, 81:473–486.

Rice, K. (2000). *Morpheme order and semantic scope*. Cambridge University Press, Cambridge.

Rice, K. (2001). Slave (Northern Athabaskan). In Andrew Spencer et al., editors, *The Handbook of Morphology*, pages 648–689. Blackwell, Malden, MA.

Sapir, E. and Hoijer, H. (1967). *The Phonology and Morphology of the Navaho Language*, volume 50 of

University of California Publications in Linguistics. University of California Press, Berkeley, CA / Los Angeles, CA.

Snoek, C., Thunder, D., Lõo, K., Arppe, A., Lachler, J., Moshagen, S., and Trosterud, T. (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.

Starlight, B., Donovan, G., and Cox, C. (2016). From archival sources to revitalization resources: Revisiting the Tsuut'ina notebooks of John Onespot and Edward Sapir. American Philosophical Society symposium 'Translating Across Time and Space: Endangered Languages, Cultural Revitalization, and the Work of History', Philadelphia, PA, October 13–15, 2016.

Trosterud, T. (2004). Porting morphological analysis and disambiguation to new languages. In *SALTMIL Workshop at LREC 2004: First Steps in Language Documentation for Minority Languages*, pages 90–92.

Trosterud, T. (2006). Grammatically based language technology for minority languages. In *Lesser-known languages of South Asia*, pages 293–316. De Gruyter Mouton, The Hague.

Young, R. W. and Morgan, W. (1987). *The Navajo language*. University of New Mexico Press, Albuquerque, NM.