Development of a finite-state model for morphological processing of Tuvan

Washington J. N., Bayyr-oool A., Salchak A., Tyers F. M.

This paper describes the development of a free/open-source finite-state morphological transducer for Tuvan, a Turkic language spoken in and around the Tuvan Republic in Russia. The finite-state toolkit used for the work is the Helsinki Finite-State Toolkit (HFST); we use the \textit{lexc} formalism for modelling the morphotactics and \textit{twol} formalism for modelling morphophonological alternations. We describe how the development of a transducer can provide new insight into grammatical generalisations, as the transducer functions as a testable model of the language’s morphology. Based on this, we add to the existing literature on Tuvan morphology a novel description of the morphological combinatorics of quasi-derivational morphemes in Tuvan, as well as some previously undescribed morphophonological phenomena. An evaluation is presented which shows that the transducer has a reasonable coverage—around 93%—on freely-available corpora of the language, and high precision—over 99%—on a manually verified test set.

Keywords: Tuvan, morphological analysis, finite-state transducers
1. **Introduction**

This paper describes the development of a morphological transducer for Tuvan. The paper is laid out as follows: section 2 gives a short introduction to Tuvan and section 3 describes some prior work on computational linguistics for Tuvan. Then section 4 documents how a number of issues related to morphotactics (section 4.4) and morphophonology (section 4.5) were implemented. An evaluation of the transducer is provided in section 5, and section 6 outlines future work related to the transducer.

2. **Language**

Tuvan (demonym [tuβɑ]) is the largest member of the Sayan branch of Turkic languages. It is an official language of the Tuva Republic (in Southern Siberia, within the Russian Federation, see figure 1), and is also spoken in the surrounding areas. Russia’s 2010 census [Росстат 2011] recorded over 250,000 Tuvan speakers, and Lewis et al. [2015] report about 27,000 speakers in Mongolia and about 2,400 in China. Many Tuvan speakers also know Russian, Mongolian, or Chinese, depending on which country they are from.

Like other Turkic languages, Tuvan exhibits a rich system of agglutinating morphology, replete with productive and idiosyncratic morphotactics and morphophonology. There have been a number of grammars written for Tuvan, including a large academy grammar

---

1 This paper is a significantly revised and expanded version of Tyers et al. [2016].
in Russian [Исхаков, Пальмбах 1961], and a grammar sketch in English [Anderson, Harrison 1999].

Figure 1: Location of the Tuva Republic

3. Prior work

Very little work has been done on computational linguistics for Tuvan; even basic resources are lacking. Of the two publications on computational linguistics, we find one paper on proposing a tagset for the Tuvan National Corpus [Bayyr-ool, Voinov 2012], and one Bachelor’s thesis on Tuvan–English statistical machine translation [Killackey 2013]. The analyser presented in this paper does not follow the tagset designed by Bayyr-ool, Voinov [2012], and instead uses a pan-Turkic tagset being adopted by the Apertium project.² It is worth noting however that our tagset is a superset of the tagset of Bayyr-ool, Voinov [2012]—that is, it makes more distinctions rather than fewer distinctions, and as such, conversion from our tagset to theirs would be feasible.

4. Development

The development of the transducer is documented in this section; specifically, we provide background on the tools used (section 4.1), information on how the transducer is implemented

² http://www.apertium.org
through a combination of morphotactics and morphophonology (section 4.2), an overview of how tokenisation is performed (section 4.3), various morphotactic issues and how they were dealt with (section 4.4), various morphophonological issues and how they were dealt with (section 4.5), the structure of the lexicon (section 4.6), and an overview of the development cycle and how its structure can yield previously undocumented grammatical generalisations about Tuvan (section 4.7).

4.1 Background

The transducer described in this paper is designed based on the Helsinki Finite State Toolkit [Linden et al. 2011], which is popular in the field of morphological analysis. It implements both the lexc formalism for defining lexicons, and the twol and xfst formalisms for modelling morphophonological rules. This toolkit has been chosen because it has been widely used for other Turkic languages, such as Turkish [Çöltekin 2010], Kyrgyz [Washington et al. 2012], Kazakh, Tatar, and Kumyk [Washington et al. 2014], and is available under a free/open-source licence.

4.2 Finite-state transducers using HFST

A finite-state transducer is a formal way to map forms and analyses to one another. For example, номнарымга ‘to my books’ would receive the analysis ном<n><pl><px1sg><dat>. The transducer accepts the form as input and outputs the analysis, and vice-versa.

When used for modelling natural-language morphology, a finite-state transducer is a directed graph where the arcs encode relations between input symbols and output symbols. These symbols may be letters, linguistic tags or archiphonemes. Analysing or generating a form involves traversing the graph from left to right, while reading a symbol and outputting its corresponding symbol.

The graph in figure 2 represents a finite-state transducer that maps grammatical combinations of the plural, possessive, and case forms of the nouns өг ‘yurt’, аът ‘horse’, and ном ‘book’ to their

---

3 The tags used here mean noun, plural, first person singular possessive, and dative, respectively. See appendix A for the complete tagset used in this transducer.
**Figure 2:** A directed graph depicting a finite-state transducer containing several Tuvan nouns, along with plural, possessive, and case morphology. Symbols from the analysis and from the form are separated by the : character. The graph is reversible and can be used for both analysis and generation. The current configuration conventionally represents generation if reading from left to right.
analyses. The morphotactics of this are represented in figure 3 in the lexc formalism.

In lexc, lexica are defined and may be directed to other lexica. For example, the Noun lexicon in figure 3 contains a list of noun stems, each of which is followed by the N1 lexicon. The N1 lexicon in turn adds the the tag <n>, and is followed by the SUBST lexicon. The SUBST lexicon adds two paths, one directly to N-INFL-COMMON, and one which adds the <pl> tag and suffix first.

The transducer generated from lexc code, depicted in figure 3, leaves “archiphonemes” and other symbols in the output. For example, following the graph in figure 2 maps the analysis ном<n><pl>-<px1sg><dat> to the form ном-{l}{A}р>{i}м>{G}{A}, not to ном-нарымга. To process these archiphonemes and symbols, another transducer, consisting of morphophonological rules coded in twol, is compose-intersected with the transducer generated from lexc code.

In this way, the morphotactics and morphophonology are coded separately, in lexc and twol, respectively. Issues dealt with in the implementation of each will be discussed in sections 4.4 and 4.5.

4.3 Tokenisation

This analyser performs tokenisation on the basis of a left-to-right longest match algorithm as described in Garrido-Alenda et al. [2002]. Simple tokens such as хувискаал ‘revolution’ are maintained as a single token, and their lemma and morphological analysis is returned. Multiword units such as соңгу чүк ‘north’ and ачы-дузазынга ‘for their help’ are combined into a single token. Abbreviations and numerals which bear case, such as АКШ-че ‘to the USA’ and 90%-зунга ‘to 90%’ are analysed as a single token, as are verb forms written with space like өөренирмен, the first-person singular aorist of өөрен- ‘to study’.

In some cases a single token is split into two tokens, as with the third person evidential aorist copula suffix; e.g., өеде-дир ‘it seems he/she/they is/are at home’ is tokenised as өр<n><loc>+э<cop>-<aor><evid><p3><sg>.

Furthermore, two input tokens may result in three output tokens, e.g. өедел in e.g., кайы өедел ‘which house is/are he/she/they in?’ is tokenised as өр<n><loc>+э<cop><aor><p3><sg>+ыл<qst>.
LEXICON CASES
%<nom%>: CLIT-COP ;
%<gen%>:>%{N%}%{I%}%{h} # ;
%<acc%>:>%{N%}%{I%} # ;
%<dat%>:>%{G%}%{A%} # ;
%<loc%>:>%{D%}%{A%} CLIT-COP ;
%<abl%>:>%{D%}%{A%}n # ;
%<all%>:>%{J%}e # ;
%<all%>:>%{D%}%{I%}%{A%} # ; ! Dir/LR

LEXICON POSSESSION
%<px1sg%>:>%{i%}м CASES ;
%<px2sg%>:>%{i%}ң CASES ;
%<px3sp%>:>%{z}%{I%}%{n} CASES ;
%<px1pl%>:>%{i%}в%{I%}c CASES ;
%<px2pl%>:>%{i%}н%{A%}p CASES ;

LEXICON N-INFL-COMMON
CASES ;
POSESSION ;

LEXICON SUBST
N-INFL-COMMON ;
%<pl%>:>%{L%}%{A%}p N-INFL-COMMON ;

LEXICON N1
%<n%>%<attr%>: # ;
%<n%>: SUBST ;

LEXICON Nouns
өг:өг N1 ; ! "фота"
аът:аът N1 ; ! "лошадь"
ном:ном N1 ; ! "книга"

Figure 3: A noun lexicon in lexc format containing three stems, and some of the morphology the follows it (lexica are arranged in the order in which they are called). The escape character % is needed for certain symbols to be read by the compiler and can be ignored by human readers, any text on a line following ! is a comment (used for glosses and meta-code), : separates the two sides of the transducer (analysis and form), and # indicates the end of a path. Tags are enclosed in <...>, archiphonemes are enclosed in {...}, and > is used to indicate a morpheme boundary.
4.4 Morphotactics

Tuvan morphotactics, like that of other Turkic languages, is characterised by a concatenative suffixing morphology, with a large number of inflectional and derivational morphemes.

4.4.1 Nominal

The nominal morphotactics, used for modelling the inflection of nouns and substantivised adjectives, is essentially identical to that in use in previous transducers for Turkic languages [Washington et al. 2014, 2012]. One difference in Tuvan compared to Kypchak Turkic is the presence of two allative morphemes, -Je and -DIвA. These were added in the case lexicon alongside the other case morphemes.

4.4.2 Verbal

While a substantial amount of the nominal morphotactics used in the Tuvan transducer were able to be copied from Kypchak transducers, Tuvan verbal morphology is quite different from that of Kypchak, so the verbal morphotactics for the Tuvan transducer had to be written entirely from scratch. We based the verbal morphotactics on the system described in Anderson, Harrison [1999]. This grammar describes the use of many morphemes, including a set of quasi-derivational morphemes, but does not include a description of their combinatorics; to our knowledge there is no existing description of the combinatorics of Tuvan verbal quasi-derivational and inflectional morphemes. So, we developed a model using field-work techniques.

We learned that a series of quasi-derivational affixes can immediately follow the verb stem, in turn followed by inflectional suffixes. Figure 4 describes a preliminary model of how the quasi-derivational morphemes can be combined. (The inflectional suffixes which follow each “group” of quasi-derivational affixes are summarised later in Table 1.)

The quasi-derivational affixes identified in Tuvan are not true derivational morphemes.4 They appear to be almost entirely

---

4 Another level of quasi-derivational morphemes exists, which for the purposes of this paper simply form new stems: passive, causative, and cooperative. These affixes are not nearly as productive as the ones described here, but they still
productive, and do not form new parts of speech. However, the
types of verbal morphology that may follow are not the same for
each group. The affixes presented in Figure 4 are outlined below:

-\textit{kCA}: Desiderative, expressing a desire to do something. \textit{Мен чагаба бижкспе тур мен.} ‘\textit{I want} to write a letter.’\textsuperscript{5}

-\textit{BAстA}: Cessative, expressing “to stop doing something”. \textit{Мен ол номну номчу вастай бердим.} ‘\textit{I stopped} reading that book.’

-\textit{GIлA}: Iterative, expressing “to do something a little bit.” \textit{Канданга номнардан номчу тул ма!} “Make Kandan read a little bit from the books.”

-\textit{BA}: Negative, expressing one way to negate verbs. \textit{Мен ол номну номчувадым.} ‘\textit{I did not} read that book.’

-\textit{Iвut}: Perfective, having a number of different uses, for example “to do something for a short while” and “to do something to completion”. \textit{Мен ырлантар мен.} ‘I’m going to sing for a bit.’

There are two basic types of inflectional affix used with
verbs in Tuvan: ones that create finite verb forms and ones that
create non-finite verb forms. Traditional grammars of Tuvan
concede that there is some overlap between these classes (i.e., some

\textsuperscript{5} We draw a number of examples from Anderson, Harrison [1999].
morphemes can create both finite and non-finite forms). The traditional classification of non-finite forms centres around two Russian terms: “причастие” (often translated as *participle*) and “деепричастие” (often translated as *adverbial participle*, *converb*, or *gerund*). Translations for these terms vary, but they refer to verb forms that are attributive, and subordinate, respectively.

Non-finite forms may be further divided based on a more nuanced understanding of their syntactic function. The non-finite verbal morphemes create verb forms that can function substantivally, attributively, adverbially, and as dependent on an auxiliary. We refer to these forms, respectively, as verbal nouns, verbal adjectives, verbal adverbs, and participles.\(^6\) The various inflectional affixes presented in Table 1 can belong to one or more of these categories. The morphology which may follow an inflectional affix is determined in part by the category or categories to which it belongs. The following list explains each functional category in detail, and Table 1 presents the various inflectional affixes and the functional categories each one can belong to.

**Finite:** Finite verb forms function as independent clauses, and are hence the only form of verbs that can form their own predicate without depending on a copula or another verb form. All finite forms in Tuvan take person and number agreement (to agree with the subject), but are not the only verb forms that may.

**Non-finite:** Non-finite forms form dependent clauses; that is, they rely on another word form to be integrated into an independent clause.

**Participle:** These are verb forms that act as a single predicate when combined with an auxiliary verb. Participles form the root of a verb phrase, and are used in the creation of

\(^6\) While we understand that these terms may be unconventional, they represent a convenient, principled way to sub-divide non-finite forms. Note that while the terms are structured like “*verbal noun*”, we do not consider the forms to be (in this example) nouns, but instead (here) substantivised verbs. We also recognise that this tentatively used term “participle” is standardly used to refer to verbal forms that are dependents of nouns, and may lead to confusion.
“serial verb constructions”.

Since participles in Tuvan form part of the same predicate as the auxiliaries upon which they depend, the auxiliary, and not the participle, takes person/number agreement. The general tag used for participle forms is <prc>. Сут ижип тур мен. ‘I am drinking milk.’

**Substantive (verbal noun):** Verbal nouns are forms of verbs that allow one to use a verb phrase as a noun phrase (i.e., substantively), for example, as a complement clause or subject of another verb. They may take person/number agreement in the form of nominal possession suffixes, and may further serve in adverbial roles with the addition of certain case morphology. The general tag used for verbal noun forms is <ger> (since verbal nouns are sometimes referred to as gerunds). Оон ындыг дүрөн чоруи барганы бисти элдепсийдирген. ‘That he left so quickly surprised us.’ (lit., ‘His so quickly away going us surprised’).

**Attributive (verbal adjective):** Verbal adjectives are forms of verbs that allow one to use a verb phrase as an adjectival phrase (i.e., attributively). They sometimes may further be substantivised, in which case they take a limited set of nominal morphology, but otherwise they do not normally have further morphology. The general tag used for verbal adjective forms is <gpr> (for Russian глагольное прилагательное). Бир дугаар келген кижиини көрдүм. ‘I saw the person who came first.’

**Adverbial (verbal adverb):** Verbal adverbs are forms of verbs that allow one to use a verb phrase as an adjunct to another verb phrase. The conditional verbal adverb agrees in person and number with its subject, but otherwise the verbal adverb clause does not agree with its subject, which it may or may not share with the main verb. The general tag used for verbal adverbs forms is <gna> (for Russian глагольное наречие). Кызыл чоруу

---

7 These are also referred to as “auxiliary verb constructions” [Bayyr-oool, Voinov 2012].
опращ, оруука хой кордуу. ‘While going to Kyzyl, I saw a sheep in the road.’

For an example of how to read Figure 4 and Table 1, consider the following word: чурттаксавас мен ‘I would not like to live’. The stem is чуртта- ‘live’, and this is followed by the quasi-derivational desiderative morpheme -ксA-, which is in turn followed by the negative morpheme -BA-. After the negative morpheme we look up the inflectional group following the combination -ксA-BA- in figure 4, which is group vi, and find that the next suffix is -c which is the negative allomorph of the aorist in Table 1. This is then followed by мен, which is the first person singular finite agreement. The final analysis of this form is then чуртта<v><iv><des><neg><aor><p1><sg>.

4.5 Morphophonology

Using HFST, morphophonology is mostly dealt with by assigning special segments in the morphotactics (lexc) which are used as the source, target, and/or part of the conditioning environment for morphophonological (twol) rules. We refer to the special segments as “archiphonemes” due to the fact that their use largely corresponds to traditionally defined archiphonemes; we use uppercase archiphonemes for symbols that are output as a range of symbols, and lowercase archiphonemes for symbols that are never or only sometimes output. Currently there are 61 twol rules in the transducer, totaling nearly 400 lines of code (not counting commented or empty lines).

The morphophonology of Tuvan is in many ways quite similar to that of other Turkic languages, with phenomena such as voicing assimilation across morpheme boundaries, front/back vowel harmony, phonologically conditioned alternations between certain allomorphs that cannot be explained purely by the phonology of the language, phonologically conditioned epenthesis, and consonant desonorisation. There are a number of alternations that are purely

8 As one anonymous reviewer pointed out, conditionals can also be used as the main verb in sentences such as Дүргең-не келирler болза! ‘If only they would come quickly!’ It remains to be determined whether this qualifies as a finite use of the form or not.
<table>
<thead>
<tr>
<th>Affix</th>
<th>Trad. class</th>
<th>Group</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>form</td>
<td>meaning</td>
<td>I</td>
</tr>
<tr>
<td>-DI</td>
<td>PAST$_1$</td>
<td>фин.</td>
<td>+</td>
</tr>
<tr>
<td>-JIк</td>
<td>RES</td>
<td>фин.</td>
<td>+</td>
</tr>
<tr>
<td>-GАй</td>
<td>OPT</td>
<td>фин.</td>
<td>+</td>
</tr>
<tr>
<td>-ZA</td>
<td>COND$^8$</td>
<td>—</td>
<td>+</td>
</tr>
<tr>
<td>-Gие</td>
<td>LIM</td>
<td>—</td>
<td>+</td>
</tr>
<tr>
<td>-Ар/(I)p/-с</td>
<td>AOR</td>
<td>при., фин.</td>
<td>+</td>
</tr>
<tr>
<td>-ГАн</td>
<td>PAST$_2$</td>
<td>при., фин.</td>
<td>+</td>
</tr>
<tr>
<td>-ГАлАк</td>
<td>UNACMPL</td>
<td>при., фин.</td>
<td>+</td>
</tr>
<tr>
<td>-GІ deg</td>
<td>IRRE</td>
<td>при., фин.</td>
<td>+</td>
</tr>
<tr>
<td>-ВліпАн</td>
<td>DUR</td>
<td>дееpl., фин.</td>
<td>+</td>
</tr>
<tr>
<td>-GАн</td>
<td>PAST$_3$</td>
<td>дeeepl.</td>
<td>+</td>
</tr>
<tr>
<td>-(I)n</td>
<td>PERF</td>
<td>дeeepl.</td>
<td>+</td>
</tr>
<tr>
<td>-E</td>
<td>IMPF</td>
<td>дeeepl.</td>
<td>+</td>
</tr>
<tr>
<td>-GАлA</td>
<td>SINCE</td>
<td>дeeepl.</td>
<td>+</td>
</tr>
</tbody>
</table>

Table 1: Inflectional affix possibilities after given combinations of quasi-derivational morphemes. The groups correspond to the inflectional groups after a given combination of quasi-derivational morphemes (see Figure 4). The type corresponds to the syntactic function of the form in a given group. The traditional classification (trad. class) corresponds to finite (фин.), ‘деепричастие’ (дeeepl.), ‘причастие’ (при.), or some combination thereof. Here ADVL type corresponds to tags with the prefix <gna_>, SUBS corresponds to tags with the prefix <ger_>, ATTR corresponds to tags with the prefix <gpr_> and PRC corresponds to tags with the prefix <prc_>.
due to orthographic convention (such as 〈я〉 standing in for what would otherwise be 〈йа〉) and complications due to the presence of many Russian borrowings, which are quite frequently left in their original orthography. Because of the similarities of these issues to those encountered in the development of transducers for other Turkic languages (especially those with Cyrillic orthographies), the specific strategies used in previous Turkic transducers to deal with these issues were largely able to be applied in the development of the Tuvan transducer.

A number of challenges specific to Tuvan were dealt with, including its particular instantiation of irregular noun forms containing possession and case (section 4.5.1), treatment of certain types of Russian loanwords in terms of vowel harmony (section 4.5.2), a nuanced process (or set of processes) of velar deletion (section 4.5.3), and a range of phonological changes that occur during epenthesis (section 4.5.4).

### 4.5.1 Combination of third-person possession and case forms

One common challenge presented by the nominal morphology of many Turkic languages is the existence of an 〈н〉 in certain combinations of the third-person possessive morpheme and case morphemes. Table 2 presents the case forms of Tuvan теве ‘camel’, as well as the case forms of тевези ‘his/her/their camel’.

<table>
<thead>
<tr>
<th>case</th>
<th>‘camel’</th>
<th>‘camel-Poss.3’</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOM</td>
<td>теве</td>
<td>тевези</td>
</tr>
<tr>
<td>GEN</td>
<td>тевениң</td>
<td>тевезинин</td>
</tr>
<tr>
<td>DAT</td>
<td>тевеге</td>
<td>тевезинге</td>
</tr>
<tr>
<td>ACC</td>
<td>тевени</td>
<td>тевезин</td>
</tr>
<tr>
<td>LOC</td>
<td>теведе</td>
<td>тевезинде</td>
</tr>
<tr>
<td>ABL</td>
<td>теведен</td>
<td>тевезинден</td>
</tr>
<tr>
<td>ALL</td>
<td>тевеже</td>
<td>тевезинче</td>
</tr>
</tbody>
</table>

**Table 2:** The case forms of теве ‘camel’ and тевези ‘his/her/their camel’, with “irregular” combinations of the possessive marker and case suffixes highlighted.

As seen in the table, the combination of the third-person
possessive suffix and most of the case suffixes are not as might be predicted. If a special ‹н›—which is deleted word-finally and before the accusative suffix—is assumed to be underlying in the third-person possessive form, then all forms aside from the deletion of the vowel in the accusative form surface as expected. This is, in fact, how this was implemented: the morphotactic form of the possessive suffix is \{z\}{I}{n}, and morphophonological rules delete the \{n\} in certain environments and let it surface as ‹н› in others. Other morphophonological rules are conditioned by \{n\}, including the deletion of the vowel in the accusative suffix when combined with the third person possessive suffix.

4.5.2 Russian loanwords and vowel harmony

In Tuvan, there are processes of both front-back vowel harmony and rounding vowel harmony, whereby the backness and/or roundedness of an affix vowel is determined by that of the previous vowel. The vowels of Tuvan are presented in table 3.

<table>
<thead>
<tr>
<th></th>
<th>front back</th>
<th>archiphoneme</th>
</tr>
</thead>
<tbody>
<tr>
<td>unrounded</td>
<td>unrounded</td>
<td></td>
</tr>
<tr>
<td>high</td>
<td>и</td>
<td>{I}</td>
</tr>
<tr>
<td>low</td>
<td>е</td>
<td>{A}</td>
</tr>
<tr>
<td>rounded</td>
<td>у</td>
<td></td>
</tr>
<tr>
<td>back</td>
<td>ы</td>
<td></td>
</tr>
<tr>
<td></td>
<td>а</td>
<td></td>
</tr>
<tr>
<td>rounded</td>
<td>о</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: The vowels of Tuvan by phonological category, presented in Tuvan orthography, along with the Apertium-internal archiphoneme conventions for high- and low-harmonising vowels.

While harmonising high vowels (represented in the morphotactics by the archiphoneme \{I\}) acquire their backness and roundness from the previous vowel, low affix vowels that undergo vowel harmony (represented in the morphotactics by the archiphoneme \{A\}) are always unrounded, and acquire only their backness from the previous vowel. For example, /хол-{N}\{I\}Н/ ‘hand-GEN’ is realised as холдун and /бе-{N}\{I\}Н/ ‘mare-GEN’ is realised as бениң, while /хол-{G}\{A\}/ ‘hand-DAT’ is realised as холга and /бе-{G}\{A\}/ ‘mare-DAT’ is realised as беге.9 In some Russian

9 For a more detailed account of Tuvan vowel harmony, see Anderson, Harrison [1999: 4–6].
loanwords in Tuvan, however, both types of harmonising vowel harmonise as front and unrounded, despite the previous vowel being back and sometimes rounded. Specifically, harmonising affixes immediately following words that end in «бль», such as ансамбль ‘ensemble’ and рубль ‘rouble’, are always front and unrounded. As an example, compare the forms of медаль ‘medal’ and руль ‘steering wheel’ (whose suffix vowels harmonise as expected) in table 4 to the corresponding forms of ансамбль ‘ensemble’ and рубль ‘rouble’ (whose suffix vowels harmonise as front unrounded).

<table>
<thead>
<tr>
<th>stem</th>
<th>V</th>
<th>C</th>
<th>dative</th>
<th>genitive</th>
</tr>
</thead>
<tbody>
<tr>
<td>медаль</td>
<td>а</td>
<td>ль</td>
<td>медальга</td>
<td>медальдын</td>
</tr>
<tr>
<td>ансамбль</td>
<td>а</td>
<td>бль</td>
<td>ансамбльге</td>
<td>ансамбльдын</td>
</tr>
<tr>
<td>руль</td>
<td>у</td>
<td>ль</td>
<td>рульга</td>
<td>рульдын</td>
</tr>
<tr>
<td>рубль</td>
<td>у</td>
<td>бль</td>
<td>рубльге</td>
<td>рубльдын</td>
</tr>
</tbody>
</table>

Table 4: A comparison of the result of back and rounding vowel harmony of both {A} (in the dative suffix) and {I} (in the genitive suffix) in stems ending in both ль and бль.

The fact that the harmonised vowel is always front and unrounded following stems in «бль» regardless of the preceding vowel is presumably related to a pronounced—but unwritten—intrusive vowel that occurs between «б» and «ль» in the bare stem forms. However, since no vowel is intrusive in forms with a following vowel (e.g., ансамбли, рубли), this phenomenon provides an interesting case of either phonological opacity or paradigm levelling—an analysis of which is beyond the scope of the present paper. The implementation of this phenomenon into the transducer, as shown in figure 5, involved creating a two rule specific to stems in «бль», as well as exceptions to the normal vowel harmony rules matching the same environment.

While the rules were written to apply to any consonant cluster ending in «ль», it is not clear whether this prediction holds. Further investigation is required to determined what other clusters participate in this process. Furthermore, the application of this exceptional phonology appears to be variable, as examples are
Figure 5: A general rule for vowel harmony with exceptions for stems ending in бль (emphasised in black), and an additional rule to harmonise as front unrounded. The rules are simplified somewhat from the actual code for purposes of demonstration.

attested that behave as would be predicted if this process were absent in the language. In order to analyse such forms, the transducer can have multiple entries in the lexicon: one with a symbol that is used to block the rule, and one without. To ensure that only the form with the exceptional phonology (the correct form in the literary language) is generated, lexicon entries which are marked with Dir/LR (in a comment after the entry) are not included in the transducer compiled for generation.

To our knowledge, this aspect of Tuvan morphophonology has not been documented elsewhere.

4.5.3 Velar deletion

Descriptions of Tuvan morphophonology, including Anderson, Harrison [1999:22–23] and Исхаков, Пальымбах [1961:117–118], have documented a widespread and productive process of stem-final velar deletion in Tuvan. In short, this process results in the voicing of asString intervocalically at the end of monosyllabic stems (e.g., /өк+{I}/ ‘glottis–Poss.3’ → [өгү]), the deletion of asString intervocalically at the end of multisyllabic stems (e.g., /инек+{I}/ ‘cow–Poss.3’ → [инээ]), and the deletion of asString intervocalically at the end of stems of any length (e.g., /өг+{I}/ ‘yurt–Poss.3’ → [өө]). In addition to two rules that deal with these specific lenition phenomena, rules (along with exceptions to other rules) had to be implemented to create the long monophthongs that result from a consonant being lost between two potentially different vowels.
In addition to these rules, it was found that the velar nasal \( ʰ \) also deletes intervocally in stem-final position in some (but not most) words in Tuvan (e.g., /со\( ʰ \)+{I}/ ‘end–Poss.3’ → [соо]). To account for this, the rule for \( \varrho \) deletion was expanded to apply to \( ʰ \). Stems where \( ʰ \) is not deleted were marked with a special archiphoneme, which is normally used for loanwords,\(^{10}\) and an exception to the environment for this expanded rule was created so that it did not apply to these stems. The resulting set of rules is provided in figure 6.

\[
\text{"Intervocalic voiced velar deletion"}
\text{Cx:0 } \iff \text{:Vow/:0* } [\%>:\text{:Vow }]/:0* ;
\text{except}
\text{:Vow } [\%\{\varrho\}]:\text{:Vow }]/:0* ;
\text{where Cx In (Г ʰ)} ;
\]

\[
\text{"Intervocalic voiceless velar deletion"}
\text{k:0 } \iff \text{:Vow/:0* } [\%>:\text{:Vow }]/:0* ;
\text{except}
\text{.#: [ ( :Cns* ) ( :Vow* ) :Vow }]/:0 [\%>:\text{:Vow }]/:0* ;
\]

Figure 6: The rules that deal with intervocalic velar deletion, with the exception that blocks deletion in stems where \( ʰ \) does not delete emphasised in black. The exception in the “voiceless velar deletion” rule is the environment where voicing of \( \kappa \) occurs in monosyllabic stems. The rules are somewhat simplified from the actual code.

4.5.4 Phonological changes during epenthesis

Like most Turkic languages, Tuvan has a small number of stems which receive an epenthetic vowel between the last two consonants when a vowel doesn’t follow. The epenthetic vowel is always high, and harmonises in frontness and roundness to the previous vowel of the stem, itself becoming the vowel to which following vowels harmonise. In addition to the presence or absence of the vowel, the consonants on either side of it may witness various alternations based on their prosodic position (e.g., syllable-final versus intervocalic) or proximity to other segments (e.g., whether a voiceless consonant precedes it or a voiced consonant or vowel precedes it). Some examples are illustrated in table 5. Besides

\(^{10}\) Since most loanwords which create challenges for the morphophonology entered Tuvan from Russian during the Soviet period, this archiphoneme is represented using the Unicode hammer and sickle symbol, \( \text{U+262D} \), or ☭.
simple epenthesis, processes of intervocalic voicing, desonorisation, fortition, and nasal assimilation are all found. Because writing a rule to change an empty space into a character is dangerous in twol, a placeholder “archiphoneme” character {y} was used that either surfaces as zero or as a harmonised epenthetic vowel. Some example \texttt{lexc} entries containing this character are shown in the table. Rules were implemented in twol to harmonise the vowel, “combine” it with ŋ to form ю if the previous vowel was rounded, and deal with the various consonant issues (most of which are more generally active in the language).

<table>
<thead>
<tr>
<th>gloss</th>
<th>citation</th>
<th>UR</th>
<th>\texttt{lexc} entry</th>
<th>before V</th>
</tr>
</thead>
<tbody>
<tr>
<td>neck</td>
<td>мойун</td>
<td>/мойн/</td>
<td>мой{у}н</td>
<td>мойну</td>
</tr>
<tr>
<td>boil</td>
<td>хайын-</td>
<td>/хайн/</td>
<td>хай{у}н</td>
<td>хайыр</td>
</tr>
<tr>
<td>distribute</td>
<td>тывыс-</td>
<td>/тыпс/</td>
<td>тып{у}с</td>
<td>тыпсыр</td>
</tr>
<tr>
<td>hand over</td>
<td>тудус-</td>
<td>/тутс/</td>
<td>тут{у}с</td>
<td>тутсур</td>
</tr>
<tr>
<td>show</td>
<td>көзул-</td>
<td>/көсл/</td>
<td>көс{у}л</td>
<td>костур</td>
</tr>
<tr>
<td>swim</td>
<td>эжин-</td>
<td>/эшн/</td>
<td>эш{у}н</td>
<td>эштирг</td>
</tr>
<tr>
<td>take out</td>
<td>ужул-</td>
<td>/ушл/</td>
<td>уш{у}л</td>
<td>уштур</td>
</tr>
<tr>
<td>be enough</td>
<td>чедиш-</td>
<td>/четш/</td>
<td>чет{у}ш</td>
<td>четчири</td>
</tr>
<tr>
<td>take part</td>
<td>кириш-</td>
<td>/кирш/</td>
<td>кир{у}ш</td>
<td>киржир</td>
</tr>
<tr>
<td>distract</td>
<td>куюс-</td>
<td>/куйс/</td>
<td>куй{у}с</td>
<td>куйзур</td>
</tr>
<tr>
<td>beg</td>
<td>чалын-</td>
<td>/чалн/</td>
<td>чал{у}н</td>
<td>чаныры</td>
</tr>
<tr>
<td>wake up</td>
<td>одун-</td>
<td>/отн/</td>
<td>от{у}н</td>
<td>оттур</td>
</tr>
</tbody>
</table>

Table 5: Some examples of words with epenthetic vowels. Presented are the citation form (with epenthesis), a proposed underlying representation (UR), the entry used in the lexicon file (\texttt{lexc}), and a form of the stem with following vowel-initial morphology. For purposes of comparison with the citation form and UR, the stems have been highlighted in bold in the forms with a following vowel.

4.6 Lexicon

The lexicon was compiled semi-automatically. Words were added to the lexicon by frequency, based on frequency lists from the corpora described in section 5.1. In order to determine the part of speech, the Russian description in the Tuvan–Russian dictionary...
by Тенишев [1968] was used. Table 6 gives the total number of stems in the dictionary by part of speech.

<table>
<thead>
<tr>
<th>Part of speech</th>
<th>Tag</th>
<th>Number of stems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>&lt;n&gt;</td>
<td>4,226</td>
</tr>
<tr>
<td>Proper noun</td>
<td>&lt;np&gt;</td>
<td>4,217</td>
</tr>
<tr>
<td>Adjective</td>
<td>&lt;adj&gt;</td>
<td>1,603</td>
</tr>
<tr>
<td>Verb</td>
<td>&lt;v&gt;</td>
<td>1,064</td>
</tr>
<tr>
<td>Adverb</td>
<td>&lt;adv&gt;</td>
<td>136</td>
</tr>
<tr>
<td>Numeral</td>
<td>&lt;num&gt;</td>
<td>85</td>
</tr>
<tr>
<td>Conjunction</td>
<td>&lt;cnj*&gt;</td>
<td>70</td>
</tr>
<tr>
<td>Postposition</td>
<td>&lt;post&gt;</td>
<td>28</td>
</tr>
<tr>
<td>Pronoun</td>
<td>&lt;prn&gt;</td>
<td>35</td>
</tr>
<tr>
<td>Determiner</td>
<td>&lt;det&gt;</td>
<td>26</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td></td>
<td><strong>11,490</strong></td>
</tr>
</tbody>
</table>

Table 6: Number of stems in each of the main categories of the transducer lexicon.

4.7 Further expansion

The transducer is developed incrementally through the cycle depicted in Figure 7.

After initial development based on preliminary generalisations about Tuvan, the development cycle consists of testing the transducer by using it to analyse corpora and examining the forms that are not analysed (sorted by frequency). Then the transducer is adjusted to analyse previously unanalysed forms, either through adding stems to the transducer or updating the morphology (lexc) or phonology (twol) components of the transducer. After the transducer is adjusted, it is recompiled and tested to ensure that it now analyses the previously unanalysed forms; the corpora are also reanalysed to determine whether the change resulted in better overall coverage, or caused previously analysed forms to now not be analysed.

Identifying forms not analysed in a large corpus and implementing morphotactic and morphophonological solutions in the
Figure 7: The development cycle for the Tuvan transducer. The lexicon is constructed semi-automatically, either from an existing wordlist or by identifying candidate stems in a corpus and checking them. The phonological rules are developed entirely manually, but with the assistance of the corpus to find examples of different phenomena.
transducer to analyse them can result in grammatical generalisations about the language that were not previously documented, such as some that were described throughout this section.

5. Evaluation

We have evaluated the morphological analyser in three ways: naïve coverage (section 5.1), precision and recall (section 5.2), and a qualitative evaluation (section 5.3).

5.1 Naïve coverage

Naïve coverage refers to the percentage of surface forms in a given corpus that receive at least one morphological analysis. Forms counted by this measure may have other analyses which are not delivered by the transducer.

The naïve coverage of the morphological analyser was calculated over five freely available corpora in Tuvan, each representing a different domain. From the encyclopaedic domain we have selected the Tuvan Wikipedia.11 From the news domain, the archives of the Tuvan daily Шын.12 For the religious domain we have used the Tuvan translation of the New Testament.13 The two additional domains were literature14 and folklore.15 Size of and naïve coverage over each corpus is presented in table 7.

5.2 Precision and recall

Precision and recall are measures of the average accuracy of analyses provided by a morphological transducer.

Precision, or the likeliness of an analysis presented by the transducer to be correct, was calculated as the number of analyses found in both the transducer’s output and the gold standard, divided by the total number of analyses output by the transducer.

12 Content from http://shyn.ru/ up to April of 2015.
14 From the books Ш. Д. Куулар (2010) Баглааш (Кызык: Тываның ном ундурер чери) and С. Сарыг-оол (2008) Аңгыр-оолдуң Тоожузу (Кызык)
Table 7: Corpora used for naïve coverage tests. Note that tokens here is defined by the morphological analyser and includes single words, punctuation, and numerals.

Recall, or the likeliness for a correct analysis of a form to be in the transducer, was calculated as the total number of analyses found in both the transducer and the gold standard, divided by the number of analyses found in the transducer plus the number of analyses found in the gold standard but not in the transducer.

This list of surface forms was then analysed with the most recent version of the analyser, and each analysis was checked. Where an analysis was erroneous, it was removed; where an analysis was missing, it was added. This process gave us a ‘gold standard’ morphologically analysed word list of 1,425 forms. The list is publically available in Apertium’s SVN repository. We then took the same list of surface forms and ran them through the morphological analyser once more. Precision and recall were calculated as described above. The results for precision and recall are presented in table 8.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Tokens</th>
<th>Coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>News</td>
<td>1,539,459</td>
<td>95.73</td>
</tr>
<tr>
<td>Religion</td>
<td>746,124</td>
<td>93.84</td>
</tr>
<tr>
<td>Literature</td>
<td>297,830</td>
<td>91.96</td>
</tr>
<tr>
<td>Encyclopaedic</td>
<td>276,547</td>
<td>90.86</td>
</tr>
<tr>
<td>Folklore</td>
<td>27,902</td>
<td>91.57</td>
</tr>
<tr>
<td>Average</td>
<td>–</td>
<td>92.79</td>
</tr>
</tbody>
</table>

Table 8: Precision and recall over all tokens and only known tokens. Known tokens are those tokens for which the stem exists in the lexicon.
5.3 Qualitative

Along with calculating the precision and recall, we also performed a qualitative evaluation using the gold standard data. We looked at each word where an error was found, and categorised the error into five types: missing stem, incorrect categorisation, bad morphotactics, bad phonology, and other. The other category included Russian words not used in Tuvan, spelling mistakes, and tokenisation errors. These errors are summarised in Table 9.

<table>
<thead>
<tr>
<th>Error type</th>
<th>Count</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing stem</td>
<td>364</td>
<td>78.8</td>
</tr>
<tr>
<td>Incorrect categorisation</td>
<td>6</td>
<td>1.3</td>
</tr>
<tr>
<td>Bad morphotactics</td>
<td>19</td>
<td>4.1</td>
</tr>
<tr>
<td>Bad phonology</td>
<td>8</td>
<td>1.7</td>
</tr>
<tr>
<td>Other</td>
<td>65</td>
<td>14.1</td>
</tr>
<tr>
<td>Total:</td>
<td>462</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 9: Error categorisation from the gold standard.

An example of bad phonology would be the word оюнун ‘game-3sg-acc’. The morphotactic representation (before morphophonology is applied) is ой{у}{н}>{I}>{N}>{I}, which is currently rendered as *ойнун. Normally, epenthesis (conversion of {у} to an output vowel, instead of resulting in no output) would not occur in this sort of environment in Tuvan, but in this particular form it seems to be required. Additionally, because the orthography of Tuvan almost always renders a ‹йу› sequence as ‹ю›, the relevant two rules would need to specify that epenthesis, in this case, occurs by way of an input ‹й› surfacing as ‹ю›, and the archiphoneme for epenthetic vowels not being output. These problems add an additional layer of complication that has yet to be resolved.

An example of inadequate morphotactics would be the personal and demonstrative pronoun ол ‘he/she/it, this’. This pronoun can take possessive suffixes, but the current paradigms in the transducer only allow for case suffixes after personal and demonstrative pronouns. Another example would be the derivational suffix -ла,
which when applied to proper nouns produces a verb which means ‘to go to X’, e.g. москвалать ‘go to Moscow’.

In terms of categorisation, we found errors in both phonological and morphological categorisation. One example of a phonologically miscategorised stem would be that proper nouns loaned via Russian, e.g. Париж ‘Paris’, need to be added to a special lexicon for borrowed stems to ensure that their final “voiced” consonants are treated as voiceless. The correct dative would be Парижке ‘to Paris’, but we currently generate *Парижге. We also found morphological categorisation errors where verbs were incorrectly categorised for aorist, e.g. they were categorised to take -{I}р instead of -{A}р.

Around a third of all missing stems were noun stems, and another third were verb stems, while the remaining third were made up of proper nouns and adjectives, with one modal word, one adverb, and two interjections found.

6. Future work

The analyser we have presented here forms part of a family of computational morphological descriptions for Turkic languages. We are actively working with the Universal Dependency project to express our annotation scheme in a way compatible with their objectives. Figure 8 provides an example of several aspects of a Universal Dependency analysis for Tuvan; for more information on the application of the annotation scheme to a Turkic language, see Tyers, Washington [2015].

Spellcheckers may be derived from this transducer fairly easily. Spellcheckers for Microsoft Word™ and Firefox (both under Windows) are currently available, and instructions are available for compiling a spellchecker for LibreOffice from the transducer’s source code. However, documentation and ready-to-use installers are not yet available in Tuvan, so they are not fully accessible to the Tuvan language community.

There is a clear need to increase the size of the lexicon: in the evaluation, nearly 80% of all errors were caused by missing stems.

16 http://apertium.projectjj.com/spellers/
17 http://wiki.apertium.org/wiki/Using_Apertium_spellers_with_LibreOffice-Voikko_on_Debian
Figure 8: A dependency tree for a sentence in Tuvan, meaning ‘Figures worsened when we stopped conducting business with the district’s constituents.’, based on the guidelines from the Universal Dependencies project.
The few remaining issues in morphotactics, morphophonology, and incorrect categorisation can be fixed relatively easily. One approach to dealing with the missing stems is to add them by using a guesser—that is, to extract possible stems from corpora by their affixes and then manually check them before adding. Another possibility would be to incorporate the guesser directly into the morphological analyser (see for example Lindén [2009]), although at the expense of accuracy.

Morphological analysis is a vital part of any natural language processing pipeline for Turkic languages. However, as tokens often receive more than one analysis (in the case of Tuvan it is somewhere in the region of 2.4 analyses per token on average), there is a need for working on disambiguation—that is, selecting the most appropriate analysis in context. We intend to adapt work done on Kazakh by Assylbekov et al. [2016] to Tuvan.

7. Conclusions

We have presented, to our knowledge, the first ever published morphological analyser for Tuvan. The analyser is free and open-source, meaning that it can be used and extended by anyone interested. The analyser has a high precision, over 99%, and fairly high coverage, over 90%, on a range of available corpora. We have outlined the development of the transducer, addressing specific issues encountered, and have demonstrated how previously undocumented grammatical generalisations about Tuvan were discovered through this type of development process. The analyser is currently used to provide morphological analyses for an online corpus of Tuvan, and we intend to use it for annotating the Tuvan National Corpus.

Acknowledgements

We would like to thank Aldynaj Saryglar for her help in developing a prototype version of the transducer. Thanks also to Vitaly Voinov for thoughtful discussions and to the anonymous reviewers for their extremely helpful comments and suggestions.
A. Tagset

This appendix presents list of the grammatical tags used in the transducer. The tag names are fairly idiosyncratic, with some being based on English terms, some on Russian terms, and some on Catalan terms. This is as a result of being from a multilingual project. Conversion from this tagset to another one would be fairly straightforward.

<abbr> abbreviation
<abl> ablative case
<acc> accusative case
<adj> adjective
<adv> adverb
<advl> averbial
<al> other (proper names)
<all> allative
<ant> anthroponym
<aor> aorist
<aapos> apostrophe
<attr> attributive
<caus> causative
<cess> cessative
<cm> comma
<cnjadv> adverbial conjunction
<cnjcco> coördinating conjunction
<cnjsub> subordinating conjunction
<coo> surname
<coll> collective numeral
<coop> coöperative form
<cop> copula
<dat> dative
<def> definite
<dem> demonstrative
<des> desiderative
<det> determiner
<du> dual
<emph> emphatic
<evid> evidential
<f> feminine
<gen> genitive
<ger_aor> aorist gerund
<ger_past> past gerund
<ger_perf> perfective gerund
References


Jonathan North Washington  
Linguistics Department  
Swarthmore College  
Swarthmore, PA 19081  
jonathan.washington@swarthmore.edu

Aziyana Bayyr-ool  
Institute of Philology  
Russian Academy of Sciences  
Novosibirsk  
azikoa@mail.ru

Aelita Salchak  
Dept. of Tuvan Philology  
Tuvan State University  
Kyzyl  
aelita_74@mail.ru

Francis M. Tyers  
HSL-fakultehta  
UiT Norgga árktalaš universitehta  
N-9019 Romsa  
francis.tyers@uit.no

Родной язык, 1(4) 2016