A Finite-State Morphological Analyser for Tuvan

Francis M. Tyers, Jonathan North Washington, Aziyana Bayyr-ool, Aelita Salchak
HSL-fakultehta, Depts. of Linguistics and Central Eurasian Studies, Institute of Philology, Dept. of Tuvan Philology
UiT Norgga árktalaš universitehta, Indiana University, Russian Academy of Sciences, Tuvan State University
N-9019 Romsa; Bloomington, IN 47405; Novosibirsk; Kyzyl
francis.tyers@uit.no, jonwashi@indiana.edu, azikoa@mail.ru, aelita_74@mail.ru

Abstract

This paper describes the development of free/open-source finite-state morphological transducers 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 `lexc` formalism for modelling the morphotactics and `twol` formalism for modelling morphophonological alternations. We present a novel description of the morphological combinatorics of pseudo-derivational morphemes in Tuvan. An evaluation is presented which shows that the transducer has a reasonable coverage—around 93%—on freely-available corpora of the languages, and high precision—over 99%—on a manually verified test set.

Keywords: morphological analysis, finite-state transducers, Tuvan

1. Introduction

This paper describes the development of a morphological transducer for Tuvan. The paper is laid out as follows: §2. gives a short introduction to Tuvan and §3. describes some prior work on computational linguistics for Tuvan. Then §4. documents how a number of issues related to morphotactics (§4.2.) and morphophonology (§4.3.) were dealt with. An evaluation of the transducer is provided in §5., and §6. outlines future work related to the transducer.

2. Language

Tuvan (demonym [tɯβa]) 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, repleat with productive and idiosyncratic morphotactics and morphophonology. There have been a number of grammars written for Tuvan, including a large academy grammar in Russian (Исхаков and Пальмбах, 1961), and a grammar sketch in English (Anderson and Harrison, 1999).

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 and 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 and Voinov (2012), and instead uses a pan-Turkic tagset being adopted by the Apertium project.[1] It is worth noting however that our tagset is a superset of the tagset of Bayyr-ool and Voinov (2012), that is it makes more distinctions rather than fewer distinctions, and as such conversion from our tagset to theirs would be feasible.

Figure 1: Location of the Tuva Republic

4. Development

4.1. Background

The transducer 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 as 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. 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.

[1] `http://www.apertium.org`

4.2.1. 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.

4.2.2. 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 and Harrison (1999). This grammar describes the use of many morphemes, but does not include a description of their combinatorics; to our knowledge there is no existing description of the combinatorics of Tuvan verbal pseudo-derivational and inflectional morphemes. So, we developed a model using field-work techniques. We learned that a series of pseudo-derivational affixes can immediately follow the verb stem, in turn followed by inflectional suffixes. Figure 2 describes a preliminary model of how the pseudo-derivational morphemes can be combined. The inflectional suffixes which follow each "group" of pseudo-derivational affixes are summarised later in Table 1.

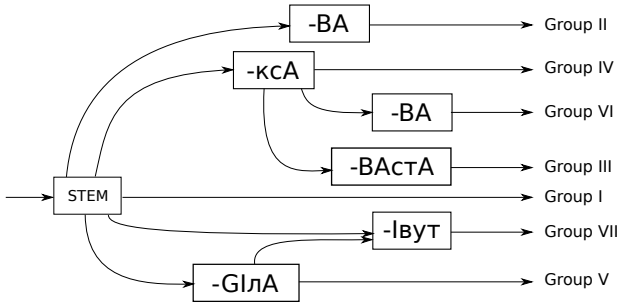

Figure 2: Preliminary verb morphotactics for inflectional and pseudo-derivational affixes. The inflectional affix groups are described in Table 1.

The pseudo-derivational affixes identified in Tuvan are not true derivational morphemes.[2] They appear to be almost entirely 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 2 are outlined below:

-ксA: Desiderative, expressing a desire to do something. *Мен чагаа бижи**ксеп** тур мен.* 'I **want** to write a letter.'

-BAстA: Cessative, expressing "to stop doing something". *Мен ол номну номчу**васта**й бердим.* 'I **stopped** reading that book.'

-GIлA: Iterative, expressing "to do something a little bit." *Канданга номнардан номчут**кула**!* "Make Kandan read **a little bit** from the books."

-BA: Negative, expressing one way to negate verbs. *Мен ол номну номчу**вадым.*** 'I **did not** read that book.'

-Iвут: Perfective, having a number of different uses, for example "to do something for a short while" and "to do something to completion".

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 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.[3] 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 it belongs 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 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 "compound verb tenses".[4] Participles in Tuvan almost never take person/number agreement. *Сүт **ижип** тур мен* 'I am **drinking** milk.'

Substantive (verbal noun): Verbal nouns are forms of verbs that allow a verb phrase to be used as a noun phrase, e.g., as a complement clause or subject of another verb. They may take person/number agreement in the form of nominal

[2] Another level of pseudo-derivational morphoemes 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 probably do not constitute true derivation.

[3] While we understand that these terms may be unconventional, they represent a convenient, principled way to sub-divide non-finite forms. Note that while they are termed e.g. *verbal nouns*, we do not consider them to be e.g. nouns, but e.g. substantivised verbs.

[4] These are also referred to as "auxiliary verb constructions".

Affix		Trad. class	Group							Type				
form	tag		I	II	III	IV	V	VI	VII	FIN	SUBS	ATTR	ADVL	PRC
-DI	PAST$_1$	fin.	+	+	+	+	+	+	+	+				
-JIк	RES	fin.	+							+				
-ZA	COND	fin.	+			+	+	+	+				+	+
-GI дег	IRRE	fin.	+										+	
-GAй	OPT	fin.	+	+	+		+	+	+	+				
-GIже	LIM	fin.	+								+		+	
-Ар	AOR	pri., fin.	+		+	+	+		+	+	+	+		
-GAн	PAST$_2$	pri., fin.	+	+	+	+	+	+	+	+	+	+		
-GAлAк	UNACMPL	pri., fin.	+	+						+		+		
-BIшAAн	DUR	deep.	+			+	+						+	
-GAш	PAST$_3$	deep.	+		+	+	+		+	+			+	
-In	PERF	deep.	+	+	+	+	+						+	+
-E	IMPF	deep.	+		+	+	+							+
-GAлA	SINCE	deep.	+										+	

Table 1: Inflectional affix possibilities after given combinations of pseudo-derivational morphemes. The groups correspond to the inflectional groups after a given combination of pseudo-derivational morpheme (see Figure 2). The type corresponds to the syntactic function of the form in a given group. The traditional classification (trad. class) corresponds to either finite (fin.), 'деепричастие' *deepričastije* (deep.), or 'причастие' *pričastije* (pri.).

possession suffixes, and certain case morphology may accord them adverbial roles. *Ооң ындыг дүрген чоруй* ***барганы*** *бисти элдепсиндирген.* 'That he left so quickly surprised us.' (lit. His so quickly away **going** surprised us).

Attributive (verbal adjective): Verbal adjectives are forms of verbs that allow a verb phrase to be used as an adjectival phrase. They sometimes may further be substantivised, in which case they take a limited set of nominal morphology, but otherwise they do normally have no further morphology. *Бир дугаар* ***келген*** *кижини көрдүм.* 'I saw the person **who came** one time.'

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; otherwise, verbal adverb clauses do not agree with subjects, which they may or may not share with a main verb. *Кызыл* ***чоруп ораш****, орукка хойну көрдүм.* '**While going to** Kyzyl, I saw a sheep in the road.'

For an example of how to read Figure 2 and Table 1, consider the following word: *чурттаксавас мен* 'I would not like to live', the stem is *чурттa-* 'live', this is followed by the pseudo-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-*, which is group VI, and find in Table 1 that the next suffix is *-с* which is the negative allomorph of the aorist, this is then followed by *мен* which is the first person singular finite agreement.

4.3. 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 `twol` rules. 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 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 the specific treatment of certain types of Russian loanwords in terms of vowel harmony, a nuanced process (or set of processes) of velar deletion, and a range of phonological changes that occur during epenthesis.

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. While harmonising high vowels (represented by the archiphoneme `{I}`) acquire their backness and roundedness from the previous vowel, low affix vowels that undergo vowel harmony (represented by the archiphoneme `{A}`) are always unrounded, and only acquire their backness from the previous vowel.[5] In some Russian loanwords in Tuvan, however, affix vowels harmonise as front and unrounded, despite the previous vowel being back and sometimes rounded. Specifically, harmonising affixes

[5]For a more detailed account of Tuvan vowel harmony, see Anderson and Harrison (1999, pp. 4–6).

immediately following words ending in ‹бль›, such as *ансамбль* ‘ensemble’ and *рубль* ‘rouble’, are always front and unrounded. Table 2 provides an example comparing forms of *медаль* ‘medal’ and *руль* ‘steering wheel’ to corresponding forms of *ансамбль* ‘ensemble’ and *рубль* ‘rouble’.

stem	V	C	dative	genitive
медаль	а	ль	медальга	медальдың
ансамбль	а	бль	ансамбльге	ансамбльдиң
руль	у	ль	рульга	рульдуң
рубль	у	бль	рубльге	рубльдиң

Table 2: A comparison of the result of back and rounding vowel harmony of both `{A}` (in the dative suffix) and `{I}` in stems ending in both ль and бль

The fact that the harmonised vowel is always front and unrounded is presumably related to a pronounced—but unwritten—epenthetic vowel that occurs between ‹б› and ‹ль› in the bare stem forms. However, since no vowel is inserted in forms with a following vowel (e.g., ансамбли, рубли), this phenomenon provides an interesting case of phonological opacity—an analysis of which is beyond the scope of the present paper. Our implementation of this phenomenon in the transducer involved creating a `twol` rule specific to stems in ‹бль›, as well as exceptions to the normal vowel harmony rules matching the same environment, as shown in figure 3. To our knowledge, this aspect of Tuvan morphophonology has not been documented elsewhere.

```
"{I} harmony"
%{I%}:Vy <=> :Vx [ :Cns* :RealCns ]/[ :0 | %- ]* _ ;
          except
             [ :BackVow :Cns* :Cns :л ь: :Cns* :RealCns ]/:0* _ ;
             [ :BackVow :Cns* :Cns :л ь:0 ]/:[ :0 - ь: ]* _ ;
          where Vx in ( ү и е э ө а о ы у я ё ю )
                Vy in ( ү и и и ү ы у ы у ы у у )
          matched ;
"{I} always front when intervening Сль"
%{I%}:и <=> [ :BackVow :Cns* :Cns :л ь: :Cns* :RealCns ]/:0* _ ;
            [ :BackVow :Cns* :Cns :л ь:0 ]/:[ :0 - ь: ]* _ ;
```

Figure 3: 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.

Descriptions of Tuvan morphophonology, including Anderson and Harrison (1999, pp. 22–23) and Исхаков and Пальмбах (1961, pp. 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 ‹к› intervocalically at the end of monosyllabic stems (e.g., /өк+`{I}`/ → [өгү]), the deletion of ‹к› intervocalically at the end of multisyllabic stems (e.g., /инек+`{I}`/ → [инээ]), and the deletion of ‹г› intervocalically at the end of stems of any length (e.g., /өг+`{I}`/ → [өө]). In addition to `twol` rules that deal with these specific deletion 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 ‹ң› also deletes intervocalically in stem-final position in some (but not most) words in Tuvan (e.g., /соң+`{I}`/ → [соо]). To account for this, the rule for ‹г› deletion was expanded to apply to ‹ң›. Stems where ‹ң› is not deleted were marked with a special archiphoneme, which is normally used for loanwords, 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 4.

```
"Intervocalic voiced velar deletion"
Cx:0 <=> :Vow/:0* _ [ %>: :Vow ]/:0* ;
     except
          :Vow _ [ %{ң%}: :Vow ]/:0* ;
     where Cx in ( г ң ) ;

"Intervocalic voiceless velar deletion"
к:0 <=> :Vow/:0* _ [ %>: :Vow ]/:0* ;
    except
        .#. [ ( :Cns* ) ( :Vow* ) :Vow ]/:0 _ [ %>: :Vow ]/:0* ;
```

Figure 4: The rules that deal with intervocalic deletion, with the exception that blocks deletion in stems where ң does not delete emphasised in black. The exeption in the voiceless deletion rule is the environment where voicing of ‹к› occurs in monosyllabic stems. The rules are somewhat simplified from the actual code.

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 of absence of a 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 3. Besides 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. The lexc entries containing this character are shown in the table. Rules to harmonise the vowel, “combine” it with й to form ю if it was rounded, and deal with the various consonant issues, were all implemented.

gloss	citation	UR	`lexc` entry	before V
front	мурун	/мурн/	мур`{y}`н	**мурн**у
neck	моюн	/мойн/	мой`{y}`н	**мойн**у
boil	хайын-	/хайн/	хай`{y}`н	**хайн**ыр
distribute	тывыс-	/тыбс/	тып`{y}`с	**тыпс**ыр
hand over	тудус-	/тутс/	тут`{y}`с	**тутс**ур
show	көзүл-	/көсл/	көс`{y}`л	**көст**үр
swim	эжин-	/эшн/	эш`{y}`н	**эшт**ир
take out	ужул-	/ушл/	уш`{y}`л	**уштур**
be enough	чедиш-	/четш/	чет`{y}`ш	**четч**ир
take part	кириш-	/кирш/	кир`{y}`ш	**киржир**
distract	куюс-	/куйс/	куй`{y}`с	**куйз**ур
beg	чалын-	/чалн/	чал`{y}`н	**чанн**ыр
wake up	одун-	/отн/	от`{y}`н	**отт**ур

Table 3: Some examples of words with epenthetic vowels. Presented are the citation form, a proposed underlying representation (UR), the entry used in the lexicon file (`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.

4.4. 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.

Part of speech	Number of stems
Noun	4226
Proper noun	4217
Adjective	1603
Verb	1064
Adverb	136
Numeral	85
Conjunction	70
Postposition	28
Pronoun	35
Determiner	26
Total:	11,490

Table 4: Number of stems in each of the main categories.

5. Evaluation

We have evaluated the morphological analysers in two ways. The first was by calculating the naïve coverage and mean ambiguity on freely available corpora. Naïve coverage refers to the percentage of surface forms in a given corpora that receive at least one morphological analysis. Forms counted by this measure may have other analyses which are not delivered by the transducer. The mean ambiguity measure was calculated as the average number of analyses returned per token in the corpus.

5.1. Corpora

We have selected corpora from five domains to be used in the evaluation of the morphological analyser. From the encyclopaedic domain we have selected the Tuvan Wikipedia.[6] From the news domain, the archives of the Tuvan daily *Шын*.[7] For the religious domain we have used the Tuvan translation of the New Testament.[8] The two additional domains were literature[9] and folklore.[10]

Domain	Tokens	Coverage (%)
News	1,539,459	95.73
Religion	746,124	93.84
Literature	297,830	91.96
Encyclopaedic	276,547	90.86
Folklore	27,902	91.57
Average	–	92.79

Table 5: Corpora used for naïve coverage tests

Table5 presents the coverage over each of these corpora, that is, the number of forms in each corpus that receives at least one analysis from the transducer. Coverage ranges from nearly 91% to nearly 96%, dependent on domain, with an average of nearly 93%.

5.2. Precision and recall

Precision and recall are measures of the average accuracy of analyses provided by a morphological transducer. Precision represents the number of the analyses given for a form that are correct. Recall is the percentage of analyses that are deemed correct for a form (by comparing against a gold standard) that are provided by the transducer.

To calculate precision and recall, it was necessary to create a hand-verified list of surface forms and their analyses. We extracted 1,500 unique surface forms at random from a Wikipedia corpus, and checked that they were valid words in the languages and correctly spelled. Where a word was incorrectly spelled or deemed not to be a form used in the language, it was discarded.

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 for each language in Apertium's SVN repository.

We then took the same list of surface forms and ran them through the morphological analyser once more. Precision was calculated as the number of analyses which were found in both the output from the morphological analyser and the gold standard, divided by the total number of analyses output by the morphological analyser.

Recall was calculated as the total number of analyses found in both the output from the morphological analyser and the gold standard, divided by the number of analyses found in the morphological analyser plus the number of analyses found in the gold standard but not in the morphological analyser.

The results for precision and recall are presented in table 6.

	Count	Precision	Recall
Known tokens	1024	0.99	0.97
All tokens	1425	0.99	0.69

Table 6: Precision & recall over all tokens and only known tokens.

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, wrong 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 7.

An example of bad phonology would be the word *оюнун* 'game.3SG.ACC'. The morphotactic representation (before morphophonology is applied) is ой{y}н>{I}>{N}{I}, which is currently rendered as **ойнун*. Normally, epenthesis (conversion of {y} to an output vowel, instead of resulting in no output) would not occur in this sort of environment

[6] https://tyv.wikipedia.org/

[7] http://shyn.ru/

[8] http://ibtrussia.org/en/ebook?id=TVN

[9] From the books Ш. Д. Куулар (2010) *Баглааш* (Кызыл: Тываның ном үндүрер чери) and С. Сарыг-оол (2008) *Аңгыр-оолдуң Тоожузу* (Кызыл:)

[10] Х. Багай-оол в кн. Тувинские народные сказки (Серия Памятники фольклора народов Сибири и Дальнего Востока). Новосибирск, 1994. С. 50–224 and Ары-Хаан: Тыва улустуң маадырлыг тоолдары, V том. Кызыл, Тываның ном үндүрер чери, 1996. 208 ар.

Error type	Count	%
Missing stem	364	78.8
Other	65	14.1
Bad morphotactics	19	4.1
Bad phonology	8	1.7
Incorrect categorisation	6	1.3
Total:	462	100

Table 7: Error categorisation from the gold standard.

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 `twol` rules would need 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 pronoun *ол* ‘this’, which can take possessive suffixes, the current paradigm only allows 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. *москвала* ‘go to Moscow’.

In terms of categorisation, we found both errors in phonological categorisation. One example would be for proper nouns loaned via Russian, e.g. *Париж* ‘Paris’, we need a special lexicon to ensure that final voiced consonants are treated as unvoiced. The correct dative would be *Парижке* ‘to Paris’, but we currently generate **Парижге*. We also found errors where verbs were incorrectly categorised for aorist, e.g. *-Ip* instead of *-Ap*.

Around a third of all missing stems were noun stems, and another third were verb stems; 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. For an example, see Tyers and Washington (2015).

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. The few remaining issues in morphotactics, morphophonology and incorrect categorisation can be fixed relatively easily.

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. In the development of the analyser, we have expanded linguistic knowledge about Tuvan, and developed strategies for difficult-to-implement grammatical patterns. The analyser has a high precision, over 99%, and fairly high coverage, over 90% on a range of available corpora. The analyser is currently used to provide morphological analyses for an online corpus of Tuvan,[11] 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 helpful comments.

References

Anderson, Gregory and K. David Harrison (1999). *Tyvan*. Lincom Europa.

Bayyr-ool, Aziyana and Vitaly Voinov (2012). “Designing a tagset for annotating the Tuvan National Corpus”. In: *International Journal of Language Studies* 6.4, pp. 1–24.

Çöltekin, Çağrı (2010). “A Freely Available Morphological Analyzer for Turkish”. In: *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC2010)*, pp. 820–827.

Killackey, Rachel (2013). “Statistical Machine Translation from English to Tuvan”. unpublished. B.A. Thesis. Swarthmore College.

Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig, eds. (2015). *Ethnologue: Languages of the World*. Eighteenth edition. Online version: `http://www.ethnologue.com`. Dallas, Texas: SIL International.

Linden, Krister, Miikka Silfverberg, Erik Axelson, Sam Hardwick, and Tommi Pirinen (2011). “HFST—Framework for Compiling and Applying Morphologies”. In: *Systems and Frameworks for Computational Morphology*. Ed. by Cerstin Mahlow and Michael Pietrowski. Vol. 100. Communications in Computer and Information Science, pp. 67–85.

Tyers, Francis Morton and Jonathan North Washington (2015). “Towards a free/open-source dependency treebank for Kazakh”. In: *Proceedings of the 3rd International Conference on Turkic Languages Processing*, pp. 276–289.

Washington, Jonathan North, Ilnar Salimzyanov Ipasov, and Francis M. Tyers (2014). “Finite-state morphological transducers for three Kypchak languages”. In: *Proceedings of the 9th Conference on Language Resources and Evaluation, LREC2014*.

Washington, Jonathan North, Mirlan Ipasov, and Francis M. Tyers (2012). “A finite-state morphological analyser for Kyrgyz”. In: *Proceedings of the 8th Conference on Language Resources and Evaluation, LREC2012*.

Исхаков, Ф. Г. and А. А. Пальмбах (1961). *Грамматика тувинского языка: Фонетика и морфология*. Москва: Издательство восточной литературы.

Тенишев, Э. Р. (1968). *Тыва-орус словарь*. Москва: Советская энциклопедия.

Федеральная служба государственной статистики российской федерации (2011). *Всероссийская перепись населения 2010 года*. Том 1.

[11] `http://gtweb.uit.no/tyv_korp/`