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

Washington J. N., Bayyr-ool 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 `lexc` formalism for modelling the morphotactics and `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

Статья посвящена разработке конечного морфологического анализатора тувинского языка, одного из тюркских языков, носители которого проживают в Республике Тыва в России и за ее пределами. Анализатор находится в открытом доступе. Конечный морфологический анализатор используется в программном обеспечении Helsinki Finite-State Toolkit (HFST); для моделирования морфотактики применяется формальный язык `lexc`, а для моделирования морфонологических чередований - формальный язык `twol`. Показано, как разработка анализатора может способствовать новому пониманию грамматических обобщений, как в опциях самого анализатора, так и в модели языковой морфологии,

подвергаемой проверке. Анализатор позволяет добавить к существующим моделям тувинской морфологии описание морфологической комбинаторики квазидеривационных морфем тувинского языка, а также впервые описать некоторые морфонологические явления. Представленный результат показывает, что анализатор справляется со своими задачами на 93% объема тестовой выборки, находящейся в открытом доступе, и высокая точность проявляется на более чем 99% тестовой выборки, проверенной вручную.

Ключевые слова: тувинский язык, морфологические анализаторы, конечные автоматы

1. Introduction

This paper describes the development of a morphological transducer for Tuvan.[1] 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 [tɯβɑ]) 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>`.[3] 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.

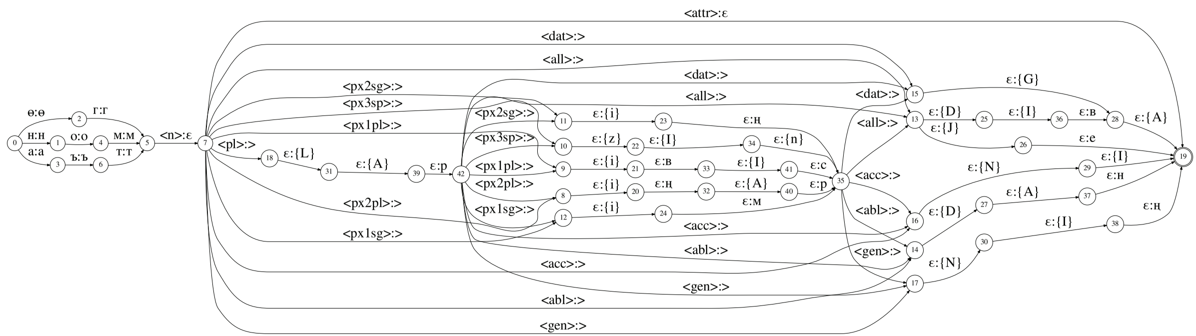

Figure 2: A directed graph depicting a finite-state transducer containing several Tuvan nouns, along with plural, possessive, and case morphology. Symbols from the analysis and from the form are separated by the : character. The graph is reversible and can be used for both analysis and generation. The 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 чук* '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 *өөренир мен*, 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., *өгде-дир* 'it seems he/she/they is/are at home' is tokenised as өг<n><loc>+э<cop>-<aor><evid><p3><sg>.

Furthermore, two input tokens may result in three output tokens, e.g. *өгдел* in e.g., *кайы өгдел* 'which house is/are he/she/they in?' is tokenised as өг<n><loc>+э<cop><aor><p3><sg>+ыл<qst>.

```
LEXICON CASES

%<nom%>: CLIT-COP ;
%<gen%>:%>%{N%}{I%}ң # ;
%<acc%>:%>%{N%}%{I%} # ;
%<dat%>:%>%{G%}%{A%} # ;
%<loc%>:%>%{D%}%{A%} CLIT-COP ;
%<abl%>:%>%{D%}%{A%}н # ;
%<all%>:%>%{J%}е # ;
%<all%>:%>%{D%}%{I%}в%{A%} # ; ! Dir/LR

LEXICON POSSESSION

%<px1sg%>:%>%{i%}м CASES ;
%<px2sg%>:%>%{i%}ң CASES ;
%<px3sp%>:%>%{z%}%{I%}%{n%} CASES ;
%<px1pl%>:%>%{i%}в%{I%}с CASES ;
%<px2pl%>:%>%{i%}ң%{A%}р CASES ;

LEXICON N-INFL-COMMON

CASES ;
POSSESSION ;

LEXICON SUBST

N-INFL-COMMON ;
%<pl%>:%>%{L%}%{A%}р N-INFL-COMMON ;

LEXICON N1

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

LEXICON Nouns

өг:өг N1 ; ! ”юрта”
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

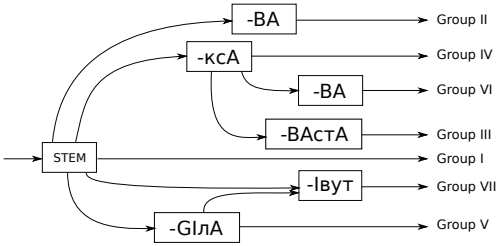

Figure 4: Preliminary verb morphotactics for inflectional and quasi-derivational affixes. The inflectional affix groups are described in Table 1.

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:

-*ксA*: Desiderative, expressing a desire to do something. *Мен чагаа бижи***кс***еп тур мен*. 'I **want** to write a letter.'[5]

-*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***ва***дым*. 'I **did not** read that book.'

-*Івут*: Perfective, having a number of different uses, for example "to do something for a short while" and "to do something to completion". *Мен ырла***пт***ар 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

probably do not constitute true derivation.

[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".[7] 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>. *Сүт* **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). *Ооң ындыг дурген чоруй* **барганы** *бисти элдепсиндирген*. '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** *кижини 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* **чоруп**

[7] These are also referred to as "auxiliary verb constructions" [Bayyr-ool, Voinov 2012].

ораш, *орукка хой көрдум*. '**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 *-ксА-*, which is in turn followed by the negative morpheme *-ВА-*. After the negative morpheme we look up the inflectional group following the combination *-ксА-ВА-* in figure 4, which is group VI, and find that the next suffix is *-с* 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 *Дүргөн-не келирлер болза!* 'If only they would come quickly!'. It remains to be determined whether this qualifies as a finite use of the form or not.

Affix		Trad. class	Group							Type				
form	meaning		I	II	III	IV	V	VI	VII	FIN	SUBS	ATTR	ADVL	PRC
-DI	PAST$_1$	фин.	+	+	+	+	+	+	+	+				
-JIк	RES	фин.	+	+					+	+				
-GAй	OPT	фин.	+	+		+	+		+	+				
-ZA	COND[8]	—	+	+		+	+	+	+				+	+
-GIже	LIM	—	+	+							+		+	
-Ар/-(I)р/-с	AOR	при., фин.	+	(+)		+	+		+	+	+	+		
-GAн	PAST$_2$	при., фин.	+	+	+	+	+	+	+	+	+	+		
-GAлAк	UNACMPL	при., фин.	+	+						+		+		
-GI дег	IRRE	при., фин.	+	+					+	+		+	+	
-BIшAAн	DUR	дееп., фин.	+			+	+			+			+	
-GAш	PAST$_3$	дееп.	+			+	+		+	+			+	
-(I)п	PERF	дееп.	+	(+)		+	+						+	+
-E	IMPF	дееп.	+		+	+	+							+
-GAлA	SINCE	дееп.	+										+	

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' (дееп.), '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 *тевe* ‘camel’, as well as the case forms of *тевези* ‘his/her/their camel’.

case	‘camel’	‘camel-Poss.3’
NOM	тевe	тевези
GEN	тевениң	тевезиниң
DAT	тевеге	тевез**инге**
ACC	тевени	тевез**ин**
LOC	теведе	тевез**инде**
ABL	теведен	тевез**инден**
ALL	тевеже	тевез**инче**

Table 2: The case forms of *тевe* ‘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.

	front		back		archiphoneme
	unrounded	rounded	unrounded	rounded	
high	и	ү	ы	у	`{I}`
low	е	ө	а	о	`{A}`

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}`ң/ ‘hand-Gen’ is realised as *холдуң* and /бе-`{N}{I}`ң/ ‘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).

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 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 `twol` 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

```
"{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 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, Пальмбах [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 ‹к› intervocalically at the end of monosyllabic stems (e.g., /өк+{I}/ ‘glottis–Poss.3’ → [өгү]), the deletion of ‹к› intervocalically at the end of multisyllabic stems (e.g., /инек+{I}/ ‘cow–Poss.3’ → [инээ]), and the deletion of ‹г› intervocalically at the end of stems of any length (e.g., /өг+{I}/ ‘yurt–Poss.3’ → [өө]). In addition to `twol` 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 ‹ң› also deletes intervocalically in stem-final position in some (but not most) words in Tuvan (e.g., /соң+{I}/ ‘end–Poss.3’ → [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,[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.

```
"Intervocalic voiced velar deletion"
Cx:0 <=> :Vow/:0* _ [ %>: :Vow ]/:0* ;
     except
         :Vow _ [ %{☭%}: :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 6: The rules that deal with intervocalic velar deletion, with the exception that blocks deletion in stems where ң does not delete emphasised in black. The exception in the “voiceless velar deletion” rule is the environment where voicing of ‹к› 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, 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 `lexc` entries containing this character are shown in the table. Rules were implemented in `twol` to harmonise the vowel, "combine" it with й 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).

gloss	citation	UR	`lexc` entry	before V
neck	моюн	/мойн/	мой{y}н	**мойн**у
boil	хайын-	/хайн/	хай{y}н	**хайн**ыр
distribute	тывыс-	/тыпс/	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	**ушт**ур
be enough	чедиш-	/четш/	чет{y}ш	**четч**ир
take part	кириш-	/кирш/	кир{y}ш	**киржи**р
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 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 (`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.

Part of speech	Tag	Number of stems
Noun	<n>	4,226
Proper noun	<np>	4,217
Adjective	<adj>	1,603
Verb	<v>	1,064
Adverb	<adv>	136
Numeral	<num>	85
Conjunction	<cnj*>	70
Postposition	<post>	28
Pronoun	<prn>	35
Determiner	<det>	26
Total:		11,490

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

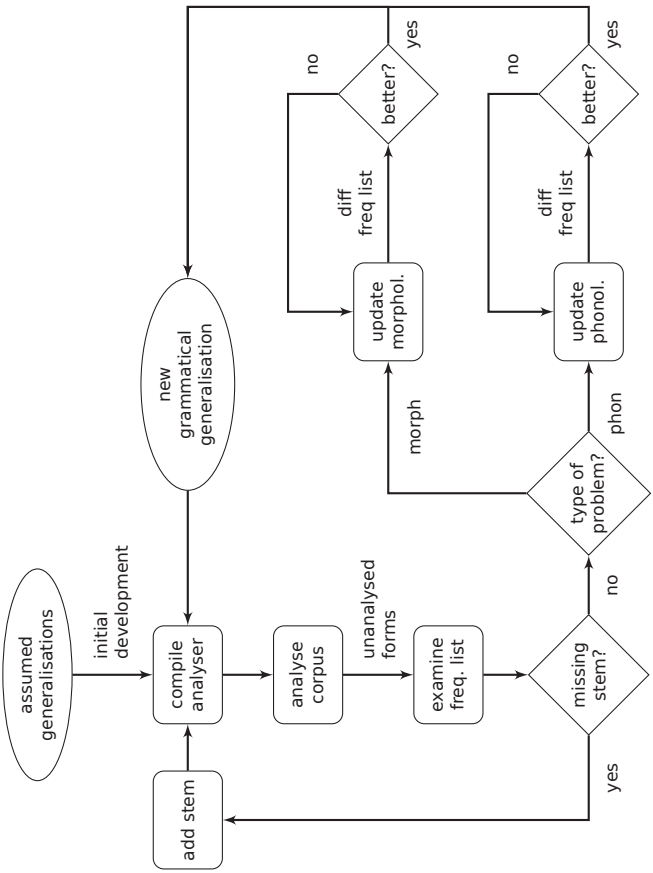

Figure 7: The development cycle for the Tuvan transducer. The lexicon is constructed semi-automatically, either from an existing wordlist or by identifying candidate stems in a corpus and checking them. The phonological rules are developed entirely manually, but with the assistance of the corpus to find 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 literature[14] 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.

[11] A dump of `https://tyv.wikipedia.org/` from April of 2015.

[12] Content from `http://shyn.ru/` up to April of 2015.

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

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

[15] 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.

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 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.

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

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.

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

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`{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 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 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. *москвала* ‘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}p` instead of `-{A}p`.

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,[16] and instructions are available for compiling a spellchecker for LibreOffice from the transducer’s source code.[17] 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`

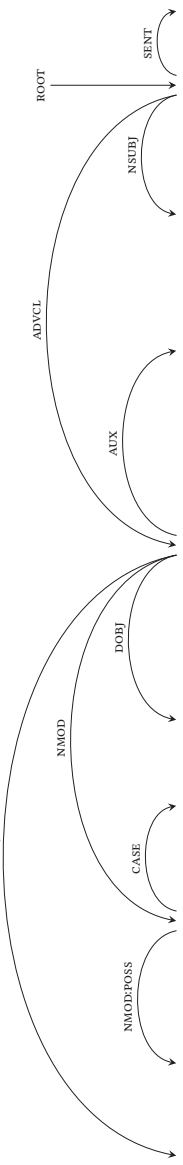

Figure 8: A dependency tree for a sentence in Tuvan, meaning ‘Figures worsened when we stopped conducting business with the district’s constituents.’, based on the guidelines from the Universal Dependencies project.

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,[18] 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.

[18] http://gtweb.uit.no/tyv_korp/

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
<apos>	apostrophe
<attr>	attributive
<caus>	causative
<cess>	cessative
<cm>	comma
<cnjadv>	adverbial conjunction
<cnjcoo>	coördinating conjunction
<cnjsub>	subordinating conjunction
<cog>	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

<gna_after>	verbal adverb ‘since’
<gna_cond>	conditional verbal adverb
<gna_lim>	verbal adverb ‘until’
<gna_mod>	modal verbal adverb
<gna_past>	past verbal adverb
<gna_perf>	perfective verbal adverb
<gna_still>	durative verbal adverb
<gna_unac>	unaccomplished verbal adverb
<gpr_aor>	aorist verbal adverb
<gpr_like>	verbal adjective ‘like’
<gpr_past>	past verbal adjective
<gpr_perf>	perfect verbal adjective
<guio>	hyphen
<ifi>	recent past
<ij>	interjection
<imp>	imperative
<ind>	indefinite
<iter>	iterative
<itg>	interrogative
<iv>	intransitive verb
<loc>	locative
<lpar>	left parenthesis
<lquot>	left quote
<m>	masculine
<mf>	masculine/feminine
<mod>	modal word
<n>	noun
<neg>	negative
<nom>	nominative
<np>	proper noun
<num>	number
<opt>	optative
<ord>	ordinal
<p1>	first person
<p2>	second person
<p3>	third person
<pass>	passive
<past>	past
<pat>	patronymic
<percent>	percentage
<perf>	perfect
<pers>	personal
<pl>	plural
<pol>	polite

<post>	postposition
<prc_aor>	aorist participle
<prc_cond>	conditional participle
<prc_impf>	imperfective participle
<prc_perf>	perfective participle
<prn>	pronoun
<px1pl>	1st person plural possessive
<px1sg>	1st person singular possessive
<px2pl>	2nd person plural possessive
<px2sg>	2nd person singular possessive
<px3pl>	3rd person plural possessive
<px3sp>	3rd person singular/plural possessive
<qnt>	quantifier
<qst>	question marker
<quot>	quote mark
<ref>	reflexive
<resu>	resultative
<rpar>	right parenthesis
<rquot>	right quote
<sent>	sentence marker
<sg>	singular
<subst>	substantive
<TD>	transitivity undetermined
<top>	toponym
<tv>	transitive verb
<unac>	unaccomplished
<unk>	unknown
<v>	verb
<vaux>	auxiliary

References

Anderson G., *Harrison K. D.* (1999). Tyvan. Lincom Europa.

Assylbekov Z., *Washington J. N.*, *Tyers F. M.*, *Nurkas A.*, *Sundetova A.*, *Karibayeva A.*, *Abduali B.*, *Amirova D.* (2016). A free/open-source hybrid morphological disambiguation tool for Kazakh. // *Proceedings of the 1st International Workshop on Turkic Computational Linguistics*.

Bayyr-ool A., *Voinov V.* (2012). Designing a tagset for annotating the Tuvan National Corpus. // *International Journal of Language Studies* 6.4, pp. 1–24.

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

Garrido-Alenda A., *Forcada M. L.*, *Carrasco R. C.* (2002). Incremental construction and maintenance of morphological analysers based on augmented letter transducers. // *Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation*, pp. 53–62.

Killackey R. (2013). Statistical Machine Translation from English to Tuvan. Unpublished B.A. Thesis. Swarthmore College.

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

Lindén K. (2009). Guessers for Finite-State Transducer Lexicons. // *Computational Linguistics and Intelligent Text Processing 10th International Conference, CICLing 2009*, pp. 158–169.

Linden K., *Silfverberg M.*, *Axelson E.*, *Hardwick S.*, *Pirinen T.* (2011). HFST—Framework for Compiling and Applying Morphologies. // *Systems and Frameworks for Computational Morphology*. Ed. by *C. Mahlow*, *M. Pietrowski*. Vol. 100. Communications in Computer and Information Science, pp. 67–85.

Tyers F. M., *Washington J. N.* (2015). Towards a free/open-source dependency treebank for Kazakh. // *Proceedings of the 3rd International Conference on Turkic Languages Processing*, 276–289.

Tyers F., *Bayyr-ool A.*, *Salchak A.*, *Washington J.* (2016). A Finite-state Morphological Analyser for Tuvan. // *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Ed. by *N. Calzolari*, *K. Choukri*, *T. Declerck*, *S. Goggi*, *M. Grobelnik*, *B. Maegaard*, *J. Mariani*, *H. Mazo*, *A. Moreno*, *J. Odijk*, *S. Piperidis*. Portorož, Slovenia: European Language Resources Association (ELRA).

Washington J. N., *Ipasov I. S.*, *Tyers F. M.* (2014). Finite-state morphological transducers for three Kypchak languages. // *Proceedings of the 9th Conference on Language Resources and Evaluation, LREC2014.*

Washington J. N., *Ipasov M.*, *Tyers F. M.* (2012). A finite-state morphological analyser for Kyrgyz. // *Proceedings of the 8th Conference on Language Resources and Evaluation, LREC2012.*

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

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

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

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