A finite-state morphological transducer for Kyrgyz

Jonathan North Washington, Mirlan Ipasov, Francis M. Tyers

Departments of Linguistics and Central Eurasian Studies
Indiana University
Bloomington, IN 47405 (USA)
jonwashi@indiana.edu

Computer Engineering and Mathematics Department
International Ataturk Alatoo University
Bishkek (Kyrgyzstan)
mipasov@gmail.com

Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant
E-03071 Alacant (Spain)
ftyers@dlsi.ua.es

Abstract

This paper describes the development of a free/open-source finite-state morphological transducer for Kyrgyz. The transducer has been developed for morphological generation for use within a prototype Turkish→Kyrgyz machine translation system, but has also been extensively tested for analysis. The finite-state toolkit used for the work was the Helsinki Finite-State Toolkit (HFST). The paper describes some issues in Kyrgyz morphology, the development of the tool, some linguistic issues encountered and how they were dealt with, and which issues are left to resolve. An evaluation is presented which shows that the transducer has medium-level coverage, between 82% and 87% on two freely available corpora of Kyrgyz, and high precision and recall over a manually verified test set.
Keywords: kyrgyz, morphology, transducer

1. Introduction

This paper describes the development of a morphological transducer oriented for the task of machine translation for the Kyrgyz language using the free/open-source platform HFST. The transducer was developed under the auspices of the Apertium (Forcada et al., 2011) project for use in a machine translation system from Turkish to Kyrgyz.

The paper is split into five main parts. First a background section gives some details about Kyrgyz and the toolkit used. Subsequent sections describe individual issues encountered with the morphotactics and the morphophonology. Finally, some evaluation results are given and future work outlined.

2. Background

2.1. The Kyrgyz language

Kyrgyz (written ‹кыргыз тили› or ‹قىرعىز تىلى›, pronounced [qɯɾɣɯz tilʲ]) alternatively written "Kirghiz" or "Kirgiz") is a Turkic language spoken in Kyrgyzstan, China, Tajikistan, and Uzbekistan. Its classification within Turkic remains problematic—it appears to alternatively belong to the Kypchak (Northwestern) branch and to the South Siberian (Northeastern) branch. The Turkic varieties phonetically and phonologically most similar to Kyrgyz are the southern dialects of Altay, though Kyrgyz shows strong parallels to Kazakh that these varieties lack, especially in its Talas dialects. In southern varieties of Kyrgyz there are also many similarities to Uzbek that other dialects lack.

Kyrgyz is spoken mostly in Kyrgyzstan where it has official status as the national language. Many Kyrgyz speakers in Kyrgyzstan are bilingual in Russian and/or Uzbek, and make up a majority of the population of the country. There are other sizable Kyrgyz-speaking communities outside of Kyrgyzstan, most notably in China (where the Kyrgyz are an officially recognised minority), Tajikistan, and Uzbekistan. Current estimates of the number of speakers range from 3 million to 4 million.[1] Not all ethnic Kyrgyz speak the language, and not all competent speakers are ethnic Kyrgyz, but there is a very strong correspondence between ethnic identity and knowledge of the language.[2]

The understanding of Kyrgyz grammar employed to construct the transducer for this project was gained from the Kyrgyz knowledge of two of our authors. Mirlan Ipasov is a native speaker of Kyrgyz, and Jonathan Washington is a theoretical and descriptive linguist fluent in Kyrgyz and knowledgeable about Turkic languages. Some grammar sources were consulted, such as Hebert & Poppe (1963), Үсөналиев & Өмүралиев (2003), Кудайбергенов et al. (1980), Somfai Kara (2003), and Imart (1981), but they were largely not relied on due to their approaches to Kyrgyz grammar. Some dictionaries were also consulted, including Жумакунова (2005) and Юдахин (1957 and 1965).

2.2. Morphological transducers

The objective of a morphological transducer is twofold. Firstly to take surface forms (e.g., алдым) and generate all possible lexical forms, and secondly to take lexical forms (e.g., ал<v><tv><ifi><p1><sg>, алд<n><px1sg><nom>, etc.) and generate one or more surface forms.

The project is designed based on the Helsinki Finite State Toolkit (Lindén et al., 2011) which is a free/open-source reimplementation of the Xerox finite-state toolchain, popular in the field of morphological analysis. It implements both the **lexc** formalism for defining lexicons, and the **twol** and **xfst** formalisms for modeling morphophonological rules. It also supports other finite state transducer formalisms such as **sfst**. This toolkit has been chosen as it – or the equivalent XFST – has been widely used for other Turkic languages, such

[1] Based on figures from Lewis (2009) and Central Intelligence Agency (2009)

[2] This interpretation of the situation is supported by the experiences of the authors with the language, and is common knowledge in Kyrgyzstan.

as Turkish (Çöltekin, 2010), Crimean Tatar (Altintas, 2001), and Turkmen (Tantuğ et al., 2006), and is available under a free/open-source licence.

3. Description

The tagset consists of 127 separate tags, 19 covering the main parts of speech (noun, verb, adjective, adverb, postposition, etc.) and 108 covering morphological subcategorisation for e.g. case, number, person, possession, transitivity, tense-aspect-mood, etc. The tags are represented as multicharacter symbols, between less than '<' and greater than '>' symbols. The tagset is quite extensive and still not entirely stabilised, as such a full listing is not included here. However, the tags are listed in the source code of the transducer,[3] along with comments describing their usage.

4. Morphotactics

4.1. Morphological and orthographic words

A typical tokenisation strategy is to take white space to be the delimiter between 'words'. For analysis and generation, however, there are exceptions to this. Some morphophonological processes work across the 'whitespace' boundary, and some clitics which are written next to the previous word without a whitespace can be considered syntactically separate units (words) and follow standard morphophonologial processes.

The first exception does not apply to Kyrgyz, as unlike other Turkic languages (including Tatar, Chuvash, and Kazakh), although it has morphophonological processes that work across the 'whitespace' boundary, these are not represented in the orthography.

Regarding the second exception, some clitics in Kyrgyz can be considered separate words, but are written together with the previous word. For example, the question word 'бы'; the focus clitic 'чы'; copula suffixes 'мин', etc.; and the progressive auxiliary 'жат' with certain verbs.

The analyser is designed to be used with the longest-match left-to-right (LRLM) tokenisation strategy, as described in Garrido-Alenda et al. (2002). Thus when more than one 'word' needs to be output, they are joined with the + symbol, which is reserved for joining two 'words' in one analysis. For example, [өнүктүрөбүзбү ?] '*will we develop it?*' is analysed as `өнүк<v><tv><caus><aor><p1><pl>+бы<qst>`, which is two words: [өнүк] (verb, transitive, causative, aorist, first person, plural) and [бы] (question clitic). Likewise, [келатсаң] '*if you come*' `кел<v><iv><prt_impf>+жат<vaux><gna_cnd><p2><sg>` is analysed as two words: [кел] (verb, intransitive, imperfect participle), and [жат] (auxiliary, conditional verbal adverb, second person, singular).

4.2. Irregular negatives of finite verb forms

One of the morphotactic challenges met in defining a finite-state transducer for Kyrgyz is that many finite verb forms have "irregular" negative forms. While the paradigms are completely regular, the negative morphotactics are not regularly derived from the affirmative forms. There are also several different "regular" patterns of alternation. Listed in table 1 are a couple of examples of a few finite verb paradigms with their negative forms.

Since the general verbal negation morpheme in Kyrgyz tends to be /BA/ (as it is in all non-finite verb forms), we treated forms with /BA/ (the last two in the table, for example) as regular (assuming other aspects of their morphology did not change). We then created two different sets of continuation lexica for finite verb forms—one for regular finite verb forms, and one for irregular finite verb forms. The continuation lexicon for regular finite verbal forms points to two continuation lexica: one of the regular verb suffixes (such as -/(I)птIp/ and -/E/, which in turn point to the appropriate continuation lexica for their person endings), and one of the negative -/BA/, which in turn points to the regular verb suffix continuation lexicon. The continuation lexicon for irregular finite verb forms directly contains affirmative and negative morphology which each point to the appropriate personal suffix continuation lexica.

4.3. Irregular [noun + possessive + case] forms

There are a number of other morphotactic issues involving "irregular" forms that have been dealt with in a similar way to the negative finite verb forms. One such issue involves nominal possessive morphology when followed by case suffixes.

Nouns may be followed by possessive suffixes before any case suffixes. This relates the noun to a preceding noun or pronoun in the genitive case. However, when both possessive morphology and case morphology occur after a noun, there is some irregularity in the system. Table 2 summarises some of the forms. Forms that do not result from simple concatenation of the possession and case endings are highlighted in bold as being irregular.

There are two rules that can immediately deal with some of these forms: optional[5] loss of /D/ in the ablative suffix after 1st person singular, 2nd person singular, and 3rd person possession suffixes, and mandatory loss of /G/ in the dative suffix in the same situations. Since these are rules specific to these morphological forms and do not apply generally in Kyrgyz at a phonological level, they were implemented directly in the interaction of the continuation lexica for these possessive suffixes and ablative and dative case suffixes.

However, instead of doing complicated splitting of the case continuation lexica following the third person possessive suffix to sometimes insert /н/, we decided to proceed under the premise that /н/ was instead underlying in this suffix (and a couple of others!) and got deleted in the nominative, accusative, and genitive. To accomplish this, a {n} archiphoneme was added to the 3rd person possessive suffix, creating an underlying form of {S}{I}{n}. Phonological rules were implemented in twolc which deleted {n} when followed by nothing (for nominative), another set of rules that deleted the accusative {N}{I} after {n} and a morpheme

[3] https://apertium.svn.sourceforge.net/svnroot/apertium/branches/apertium-kir/

[4] P1–P6 refer to different sets of personal suffixes; the terminology and specific numbers are based on Hebert & Poppe (1963, 29).

[5] Optional rules are dealt with by making a non-symmetrical transducer by way of marking lines in the lexc to be included in only one direction of the transducer; in this case, since the optional loss of /D/ is the default form, we do not generate forms where the /D/ remains, but do analyse them.

Table 1: Examples of different affirmative / negative alternations in finite verb forms

tense/aspect	ending + person series	ending + person series
recent eyewitness past	-/DI/ + P4[4]	-/GAн жок/ + P3
non-recent past	-/GAн/ + P3	-/GAн эмес/ + P3
non-recent evidential past	-/GAн экен/ + P3	-/GAн эмес экен/ + P3
past habitual	-/чU/ + P3	-/чU эмес/ + P3
recent evidential past	-/(I)птIр/ + P3	-/BAптIр/ + P3
habitual/future	-/E/ + P6	-/BAй/ + P6

Table 2: Combinations of possessive suffixes with case suffixes

case	morphology	1st person singular	2nd person sing.	3rd person	1st person plural
nom	—	-(I)м	-(I)ң	-(S)I	-(I)бIз
acc	-NI	-(I)мдI	-(I)ңдI	**-(S)Iн**	-(I)бIздI
gen	-NIн	-(I)мдIн	-(I)ңдIн	-(S)IнIн	-(I)бIздIн
loc	-DA	-(I)мдA	-(I)ңдA	**-(S)IндA**	-(I)бIздA
abl	-DAн	-(I)мдAн, **-(I)мAн**	-(I)ңдAн, **-(I)ңAн**	**-(S)IнAн**	-(I)бIздAн
dat	-GA	**-(I)мA**	**-(I)ңA**	**-(S)IнA**	-(I)бIзгA

boundary, and a rule that deleted {n} when followed by a morpheme boundary and the genitive {N}{I}н.[6]
While the phonological rules are not necessarily as closely tied to an accurate morphological analysis of what is going on as they could be, these few rules allowed fewer irregular continuation lexica to be created—most immediately by allowing the ablative and dative continuation lexica for 3rd person to behave similarly to that of 1st and 2nd person singular, and avoiding separate irregular continuation lexica for the other cases. The direct result of this approach was a much more concise description of the morphology in the lexicon file.

5. Morphophonology

Current morphological analysers of European languages are based on the orthography of the words, even where this may make it more difficult to write morphophonological rules. This has the advantage that in order to use the morphological analyser to analyse text (as opposed to using it as a tool to study phonology), no up/down conversion between the orthography and the transcription used in the analyser is necessary, avoiding possibilities of misconversion.
In the case of Kyrgyz, dealing with the orthographical forms directly further simplifies some aspects of the morphophonology, since Kyrgyz orthography reflects a somewhat simplified version of the phonology: it ignores processes that the orthographies of many other Turkic do not, such as the phonemic distinction between velar and uvular consonants as well as sandhi voicing effects. However, other aspects are made more complicated, such as й+vowel combinations (see §5.6.).

5.1. Vowel harmony

In Kyrgyz, there are two basic archiphonemes: a low vowel, {A}, and a high vowel, {I}. The high vowel takes the backness and rounding of the preceding vowel, resulting in the values shown in table 3.

[6]Because a general possessive suffix that can follow personal possessive suffixes can also behave like the genitive in this respect, the rule is actually more general.

Table 3: Vowel harmony for archiphoneme {I}

after	result	after	result
и	и	ы	ы
ү	ү	у	у
е	и	а	ы
ө	ү	о	у

The low vowel {A} also takes the backness and rounding of the preceding vowel, with the exception of when it occurs after /y/—n this case, it has an unrounded variant, as depicted in table 4.

Table 4: Vowel harmony for archiphoneme {A}

after	result	after	result
и	е	ы	а
ү	ө	у	а
е	е	а	а
ө	ө	о	о

There were other vowel archiphonemes that had to be implemented in this project. For example, {U} occurs in the past habitual suffix -/чU/ and the general gerund/infinitive -/U:/, but nowhere else in the language. Also, {E} was created for use in the habitual/future suffix; it has surface forms identical to those of {A}, except after vowels where it surfaces as [й]. Despite the fact that {E} is very similar to {A}, a separate but very similar twolc rule had to be created. Repeated content could have been reduced by using cascading rules instead of two-level rules, but this would have caused other complications, such as finding the correct rule ordering. An alternative possibility would have been to have a single intermediate level so that a rule like "{E}→{A} after vowels" with a de-

fault surface form of [й] could've been implemented.

5.2. Voicing assimilation

In Kyrgyz, there are two basic processes affecting the realisation of consonants when they are adjacent to other consonants: voicing assimilation and desonorisation.

Voicing assimilation in Kyrgyz involves the agreement of voicing of two consonants across a syllable boundary. An example involves the locative and dative suffixes, as shown in table 5.

Table 5: Examples comparing voicing and devoicing in Kyrgyz

underlying	surface	gloss
/алма-DA/	[алмада]	'*apple*–LOC'
/каз-DA/	[казда]	'*goose*–LOC'
/баш-DA/	[башта]	'*head*–LOC'

underlying	surface	gloss
/алма-GA/	[алмага]	'*apple*–DAT'
/каз-GA/	[казга]	'*goose*–DAT'
/баш-GA/	[башка]	'*head*–DAT'

Here, the underlying /D/ is always realised as [д] when after a voiced segment (including a vowel), but is realised as [т] after an unvoiced consonant, while underlying /G/ is always realised as [г][7] after a voiced segment, but is realised as [к][8] after voiceless consonants.

Voicing assimilation was dealt with simply by creating a single twol rule that transformed any relevant archiphoneme ({B}, {G}, {D}, {L}, and {N}) to a voiceless stop (either [п], [к], or [т]) after a voiceless consonant. The set of voiceless consonants was also defined in twolc so that rules sensitive to this set would be easier to write.

5.3. Desonorisation

Desonorisation (Washington, 2010) happens to /N/ when it follows any consonant, and to /L/ when it follows a consonant of equal or lower sonority (i.e., /л/, nasals (/м/, /н/, /ң/), and obstruents, but not /й/, /р/, or vowels). The resulting surface consonant is [д] for both /N/ and /L/, unless it follows a voiceless consonant, in which case it surfaces as [т]. An example includes the accusative and plural suffixes, as in table 6.

The orthography makes idenifying voiced and voiceless pairs straightforward, since e.g., words borrowed from Russian which end in ‹в› are treated in spoken Kyrgyz as unvoiced (and indeed have unvoiced surface forms syllable-finally), but are treated as voiced in the orthography.

Desonorisation was dealt with by a rule that changes {L} to [д] following a series of "low-sonority" consonants appropriately defined earlier in the twol file and a syllable boundary. The {N} desonorisation rule changes {N} to [д] following a series containing all voiced consonants, followed by a syllable boundary. Both sets were defined to exclude voiceless consonants that would instead trigger devoicing (§5.2.).

Table 6: Examples comparing desonorisation of {N} and {L}

underlying	surface	gloss
/алма-NI/	[алманы]	'*apple*–ACC'
/сыр-NI/	[сырды]	'*secret*–ACC'
/каз-NI/	[казды]	'*goose*–ACC'
/баш-NI/	[башты]	'*head*–ACC'

underlying	surface	gloss
/алма-LAp/	[алмалар]	'*apple*–PL'
/сыр-LAp/	[сырлар]	'*secret*–PL'
/каз-LAp/	[каздар]	'*goose*–PL'
/баш-LAp/	[баштар]	'*head*–PL'

5.4. Lenition

In Kyrgyz, stem-final [voiceless] labial and dorsal consonants voice when a suffix beginning with a vowel follows them. The implementation of this in twol is very straightforward; our rules state that /п/ and /к/ become /б/ and /г/ preceding a morpheme boundary and when between two surface vowels with optional intervening non-surfacing phonemes.

5.5. Nouns ending in /рн/, etc.

Kyrgyz has a number of nouns that underlyingly end with a consonant cluster consisting of two liquids, but which are split by an epenthetic vowel syllable-finally, as shown in table 7. However, when a vowel-initial suffix follows the noun, as in table 8, the epenthetic vowel is absent, and rules of consonant unfaithfulness involving consonant clusters (e.g., desonorisation) become relevant as well.

Table 7: Examples of /рн/-final nouns surfacing with epenthetic vowel

underlying	surface	gloss
/мурн/	[мурун]	'*nose*'
/мурн-LAp/	[мурундар]	'*nose*–PL'

underlying	surface	gloss
/орн/	[орун]	'*place*'
/орн-LAp/	[орундар]	'*places*–PL'

This was dealt with in a somewhat non-linguistic way: a default epenthetic vowel {y} was defined and inserted into the words by default (see Fig. 1.

[7] ‹г› is pronounced [ɣ] in the context of front vowels and [ʁ] in the context of back vowels; it also has realisations of [g] and [ɢ] after nasals.

[8] ‹к› is realised as [q] in the context of back vowels and [k] in the presence of front vowels.

Table 8: Examples of /рн/-final nouns surfacing with no epenthetic vowel

underlying	surface	gloss
/мурн-(S)I/	[мурду]	'*his/her/its nose*'
/мурн-Ім-DAн/	[мурдуман]	'*from my nose*'

underlying	surface	gloss
/орн-(S)I/	[орду]	'*his/her/its place*'
/орн-Ім-DAн/	[ордуман]	'*from my place*'

Figure 1: Example of use of epenthetic vowel {y}.

орун:ор%{у%}н N-INFL ; ! "place"
мурун:мур%{у%}н N-INFL ; ! "nose"

The vowel was then removed by a twol rule when a vowel followed. To realise the epenthetic vowel, a rule resembling the {I} vowel harmony rule was created—the primary difference between these rules is that the {y} harmony rule only acts when something other than a vowel (i.e., a consonant or a word boundary) follows the following consonant. While {y} will always behave like {I}, there is unfortunately no way to make this happen without replicating the context of the {I} rule. This is due to restrictions in two-level morphology.

5.6. й+vowel letters

In Kyrgyz, there are a series of letters which represent /й/ plus a vowel: /я, е, ё, ю/ represent /йа, йэ, йо, йу/ respectively. However, /э/ is also represented as /е/ when short and after a consonant (i.e., it is only represented as /э/ when long—/ээ/—or when word-initially). These "yoticised" vowels proved difficult to work with. For example, /бой+(S)I/ '*length*', given normal vowel harmony rules, would surface as [бойу]. However, the correct orthographic form is [бою]. Distinct and separate vowel harmony rules had to be created for all vowel archiphonemes in post-vocalic contexts, e.g. here to turn {I} into [ю] after /й/. An additional rule had to delete the underlying /й/ before yoticised vowel letters (so that e.g. [бойю] was not output). Because there are two levels to the phonology in twolc and it is otherwise not linear, attention had to be paid to the fact that e.g., the distinct vowel harmony rules for post-/й/ contexts were aware that the /й/ would not appear on the second [surface] level.

6. Statistics

The morphotactic lexicon contains a total of 135 continuation lexica, modelling the morphotactics. The phonological rules file contains 47 rules. Table 9 gives the approximate number of stems in each main word class. The numbers are approximate as some words may have two entries for one stem (for example if a set of forms is irregular).

Table 9: Breakdown of number of stems in the lexicon by major part of speech. Minor categories, such as particle, modal and copula are left out.

Part of speech	Number of stems
Noun	4,972
Verb	1,231
Adjective	944
Proper noun	796
Adverb	295
Numeral	63
Conjunction	58
Postposition	51
Pronoun	29
Determiner	27
Total:	8,466

7. Evaluation

We have evaluated the analyser in two ways. The first is by calculating the naïve coverage and mean ambiguity on two large freely available corpora of Kyrgyz. The second is by performing an evaluation of precision and recall on a smaller, hand-validated test set. The revision of the transducer evaluated was `r36739`.

7.1. Coverage and mean ambiguity

To calculate the naïve coverage[9] of the analyser, two corpora were used. The first corpus was a database dump of the Kyrgyz Wikipedia, dated 2011-09-23,[10] which was processed with the programs `aq-wikicrp` to extract sentences. The second was a corpus generated from the archives of Radio Free Europe / Radio Liberty (RFE/RL)'s Kyrgyz service, azattyk.org. The RFE/RL corpus was built using a script that scraped the archives for all articles from 2010, which include articles on a wide variety of topics, from sports to politics and from culture to current events.
Both corpora were split into 10 equal parts, and coverage was calculated over each part separately in order to calculate the standard deviation of the mean. As can be seen from table 10, running the analyser on the RFE/RL corpus gives a higher coverage. This is most likely because the text is more homogenous, and there is less non-Kyrgyz (e.g., Russian) text.
The column 'mean ambiguity' gives the average number of analyses for each surface form encountered when analysing this corpus.

7.2. 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

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

[10] The exact name of the dump was `kywiki-20110923-pages-articles.xml`.

Table 10: Naïve coverage and mean ambiguity of the analyser over two test corpora.

Corpus	Tokens	Known	Naïve coverage (%)	Mean ambiguity
Kyrgyz Wikipedia	329,524	270,668	82.1 ± 3.2	2.35
RFE/RL Kyrgyzstan	4,112,558	3,614,193	87.9 ± 1.2	2.43

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 1,000 unique surface forms at random from the RFE/RL corpus, and checked that they were valid words in Kyrgyz and correctly spelt. Where a word was incorrectly spelt or deemed not to be a form used in Kyrgyz, it was discarded and a new random word selected.

This list of surface forms was then analysed with the current 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 1,000 surface forms with their analyses. The list is publically available.[11]

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 11.

Table 11: Precision and recall of the analyser over the 1,000 word test set.

Precision	Recall
97.32%	94.56%

That the precision is so high is not surprising: the transducer is rule-based and the test set is the corrected output of the analyser. In order to get a more accurate figure for precision, it would be necessary to have several linguists review the gold standard.

Our evaluation of the transducer brought to light some systematic ways that we could improve the transducer.

Low recall is due to forms which do not receive any analyses; in all cases this was due to stems missing in the lexicon. Out of the 1,000 words, 91 did not receive any analyses. Of these, 52 forms were proper names, 28 were nouns, 6 were adjectives, and the remaining 5 verbs. This would suggest that in terms of adding new stems, it would be a good idea to look at increasing both the number of proper names and nouns.

Low precision was due largely to incorrect stem categorisation. Proofing our stem categorisation would increase precision significantly. Besides this, several issues with the transducer were brought to light. One such issue is that mono-syllabic verbs ending in vowels (e.g., [же] '*eat*', [жуу] '*wash*', [оо] '*tilt*') do not take a gerund/infinitive in -/U:/, but are only used with the morphologically, semantically, and syntactically similar -/(I)ш/ suffix. To prevent these forms from being output, a different version of the current verbal continuation lexicon should be made to capture the morphological difference—i.e., forms of these verbs that are marked as `<ger>` should be output with the -/(I)ш/ suffix instead of the -/U:/ suffix, and analysed likewise.

A similar issue the gold standard developed for evaluation showed us was that not all adjectives can be used used as adverbs; the transducer currently assumes they can, and every item in the adjective lexicon currently also has an adverbial reading. While usually only morphologically complex (i.e., derived) adjectives are restricted from being used as adverbs, our initial strategy for fixing this will involve making a separate "no adverb reading" lexicon for adjectives that cannot be used as adverbs.

8. Future work

The major work to be done is increasing the size of the lexicon. While a good level of coverage has been achieved with only 8,466 stems, real-word, or production morphological analysers have tens of thousands of stems, even for morphologically-rich languages like Kyrgyz.

Once good coverage has been achieved with a morphological analyser, the next logical step is to start work on morphological and syntactic disambiguation. As can be seen from the figures for mean ambiguity, there is a lot of work that can be done on disambiguation.

Also, despite the good coverage, there are still a number of grammatical forms that have not been implemented into the transducer. For example, one such form is historically the negative of the eyewitness past tense, but is now used (mostly with a question clitic) in a certain type of modal expression;[12] this cannot be tagged as the negative of the eyewitness past tense, since a suppletive form is tagged that way. Other shortcomings of the transducer are that it is not set up to deal with affixes on acronyms/abbreviations or numerals/digits (e.g., [АКШныын] '*The USA's*', [100гө] '*up to 100*') or capitalisation changes on lemmas that occur in different word classes (e.g., [Финландия] '*Finland*', but [финландиялык] '*Finnish*').

[11] https://apertium.svn.sourceforge.net/svnroot/apertium/branches/apertium-kir/eval/ref.1000.txt

[12] The semantics are a sort of expression of surprise at others not having noticed something; e.g., [Идиштерди жуубадымбы !] '*Did I not just wash the dishes !?*'

9. Conclusion

We have presented a morphological analyser and generator for Kyrgyz based on finite-state transducer technology. The transducer has medium-level coverage, of between 82-87%, and high precision and recall.

It is the hope of the authors that the work on this transducer will lead the way for work on free/open-source tagset-compatible transducers for other Turkic languages. Indeed, transducers for Kazakh, Tatar, Bashqort, and Chuvash which are currently under development have benefited from work done on this transducer: many aspects have been based on the same general approach, and a number of phonological rules from the Kyrgyz transducer have served as the basis of rules in the transducers of these other languages.

Acknowledgements

We would like to thank: the Google Summer of Code 2011, which supported the development of the Turkish→Kyrgyz MT system this transducer was designed for; Tolgonay Kubatova for additional native-speaker insight on Kyrgyz and all the ways she's supported the authors and this project; and the anonymous reviewers for comments on how to improve the manuscript. This work has been partially funded by Spanish Ministerio de Ciencia e Innovación through project TIN2009-14009-C02-01.

References

Altintas, K. (2001). A morphological analyser for Crimean Tatar. *Proceedings of Turkish Artificial Intelligence and Neural Network Conference*.

Central Intelligence Agency (2009). *The World Factbook 2009*. Washington, DC: Central Intelligence Agency. https://www.cia.gov/library/publications/the-world-factbook/index.html.

Forcada, M.L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J.A., Sánchez-Martínez, F., Ramírez-Sánchez, G., & Tyers, F.M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 24(1), pp. 1–18.

Garrido-Alenda, Alicia, Forcada, Mikel L., & Carrasco, Rafael C. (2002). Incremental construction and maintenance of morphological analysers based on augmented letter transducers. In *Proceedings of TMI 2002 (Theoretical and Methodological Issues in Machine Translation, Keihanna/Kyoto, Japan)*. pp. 53–62.

Hebert, R.J. & Poppe, N. (1963). *Kirghiz Manual*, vol. 33 of *Uralic and Altaic Series*. The Hague: Mouton & Co.

Imart, G. (1981). *Le Kirghiz (Turk d'Asie Centrale Soviétique): Description d'une langue de littérisation récente*. Aix-en-Provence: L'université de Provence.

Lewis, M.P. (Ed.) (2009). *Ethnologue: Languages of the World*. Dallas, Tex.: SIL International, sixteenth edn. Online version: http://www.ethnologue.com/.

Lindén, K., Axelson, E., Hardwick, S., Pirinen, T., & Silfverberg, M. (2011). Hfst—framework for compiling and applying morphologies. *Systems and Frameworks for Computational Morphology*.

Somfai Kara, D. (2003). *Kyrgyz*. München: Lincom Europa.

Tantuğ, A.C., Adalı, E., & Oflazer, K. (2006). Computer analysis of Turkmen language morphology. *Advances in natural language processing, proceedings (Lecture notes in artificial intelligence)*, pp. 186–193.

Washington, J.N. (2010). Sonority-based affix unfaithfulness in Turkic languages. Master's thesis, Univerity of Washington.

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.

Жумакунова, Г. (2005). *Түркчө-Кыргызча Сөздүк*. Бишкек: Кыргыз-Түрк Манас университети.

Кудайбергенов, С., Турсунов, А., & Сыдыков, Ж. (Eds.) (1980). *Кыргыз адабий тилинин грамматикасы*. Фрунзе: Илим.

Юдахин, К.К. (1957). *Орусча-кыргызча сөздүк*. Москва: Государственное издательсвто иностранных и национальных словарей.

Юдахин, К.К. (1965). *Кыргызча-орусча сөздүк*. Москва: «Советская Энциклопедия» басмасы.

Үсөналиев, С. & Өмүралиев, Б. (2003). *Азыркы кыргыз тилинин таблицалары (Фонетика, морфология жана синтаксис)*. Бишкек: АРХИ.