Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2581–2589
Marseille, 11–16 May 2020

A Finite-State Morphological Analyser for Evenki

Anna Zueva, Anastasia Kuznetsova, Francis M. Tyers
School of Linguistics, Department of Linguistics, Department of Linguistics
Higher School of Economics, Indiana University, Indiana University
anna.zueva.v@gmail.com, anakuzne@iu.edu, ftyers@iu.edu

Abstract

It has been widely admitted that morphological analysis is an important step in automated text processing for morphologically rich languages. Evenki is a language with rich morphology, therefore a morphological analyser is highly desirable for processing Evenki texts and developing applications for Evenki. Although two morphological analysers for Evenki have already been developed, they are able to analyse less than a half of the available Evenki corpora. The aim of this paper is to create a new morphological analyser for Evenki. It is implemented using the Helsinki Finite-State Transducer toolkit (HFST). The `lexc` formalism is used to specify the morphotactic rules, which define the valid orderings of morphemes in a word. Morphophonological alternations and orthographic rules are described using the `twol` formalism. The lexicon is extracted from available machine-readable dictionaries. Since a part of the corpora belongs to texts in Evenki dialects, a version of the analyser with relaxed rules is developed for processing dialectal features. We evaluate the analyser on available Evenki corpora and estimate precision, recall, and F-score. We obtain coverage scores of between 61% and 87% on the available Evenki corpora.

Keywords: evenki, fst, morphology

1. Introduction

Morphological analysis is an essential part of natural language processing (NLP), especially for languages with rich morphology. Morphological analysers detect the structure of a word form and return a lemma and corresponding grammatical tags. Apart from morphological analysers, there are morphological generators, which use lemmas and grammatical tags to produce a particular word form. Morphological analysers and generators are widely implemented in various NLP applications such as corpus annotation, information retrieval, machine translation, speech recognition, speech synthesis, and proofing tools.

Evenki is an endangered Tungusic language spoken in Russia (see Figure 1), China and Mongolia. According to Ethnologue it currently has 15,800 speakers in total[1]. Since Evenki has rich morphology, a morphological analyser is extremely useful for processing Evenki texts. Without a morphological module, less than one third of the tokens in Evenki texts can be found in dictionaries and therefore can get part-of-speech tag or translation. Table 1 shows the comparison[2] between the number of tokens in Evenki and English (as a language with little inflection), which can get analysis without morphological processing.

Although two morphological analysers for Evenki have already been developed (see Section 2.2.), they are only able to produce analyses for at most 58% of Evenki corpora in the literary language[3] and 44% of dialectal Evenki corpora.

The aim of this work is to create a new morphological analyser for the Evenki language, which can be used in further NLP applications. Since, like most languages, Evenki is under resourced and there is a lack of annotated texts for the Evenki literary language, we have chosen a rule-based approach. The analyser is implemented using Helsinki Finite-State Toolkit (Lindén et al., 2011), which provides a free/open-source framework for compiling and applying linguistic descriptions. The `lexc` formalism is used to specify the morphotactic rules, which define the valid orderings of morphemes in a word. Orthographic rules and morphophonological alternations are described using the `twol` formalism (Koskenniemi, 1983).

Figure 1: Map of dialects of Evenki within the Russian Federation. Map by N. A. Mamontova, based on Vasilevich (1948). The dialects are grouped into northern, in slanted stripes, southern in horizontal stripes and eastern in grey.

Evaluation includes measuring coverage and mean ambiguity on the available corpora for Evenki as well as calculating precision and recall using annotated texts in the task of morphological tag assignment.

The key contribution in this paper is improved treatment of dialectal forms, which much of the existing text is written in. In addition, it is important for processing speech and informal text produced by native speakers, for example on social

[1] `https://www.ethnologue.com/language/evn` (retrieved 2020-03-06).

[2] Evenki corpora and Evenki dictionaries are described in Section 2.1. and Section 3.5. respectively. The English corpus was collected from the English treebanks of the Universal Dependencies project. The English dictionary was obtained from the Apertium repository for English (Forcada et al., 2011).

[3] We use the term *literary language* to refer to the standardised variety of Evenki.

	Evenki	English
Tokens found in dictionaries	73,586	353,867
All tokens	247,300	437,556
Percentage	29.8%	80.9%

Table 1: The number of tokens in Evenki and English corpora which can get their analysis without morphological processing.

media.[4]

2. Existing resources

In this section we describe the available Evenki corpora, which are used for the development and evaluation of the analyser, and existing (unpublished) analysers for Evenki.

2.1. Available corpora

Newspaper corpus The collection of texts in the Evenki literary language[5] consists of 464 texts from the newspaper *Эвенкийская жизнь* 'Evenki life', which amount to about 222,266 words. These texts do not contain the sign of vowel length, which is marked in Evenki orthography using a macron. For example, the word *улукӣ* 'squirrel' (normatively spelt with a macron over the long *ӣ* /iː/) is written as *улуки* without the macron. Another feature of the texts is that Evenki nasal /ŋ/ is represented with the combination of letters *нг* instead of a single Evenki letter *ŋ*. For example, a 3rd-person singular pronoun can be written as *нунган* instead of the normative spelling *нуŋан*.

Linguistic Corpora at IEA RAS Evenki texts,[6] which are available at the website of *Linguistic Corpora at the Institute for Ethnology and Anthropology of the Russian Academy of Sciences* (IEA RAS), were collected within the project *The Further Development and Filling of Digital Corpora in Siberian Minority Languages (Nenets, Teleut, Shor and Evenki)* and kindly provided by Kirill Shakhovtsov, the project leader of the Linguistic Corpora at IEA RAS.
The corpus contains 106 texts, only one of which is marked as Evenki in the literary language. However, in addition, we will take texts from the Newspaper corpus to be Evenki in the literary language. Other texts belong to different dialects. Some of the texts, mainly collected by G. M. Vasilevich in 1931–1960, contain letters which are not used in modern Evenki alphabet. These texts were converted to modern orthography using the rules provided with the morphological analyser at Linguistic Corpora IEA RAS. For instance, the rules cover replacing *ı* with *ы* and the combinations of *йа*, *йо*, *йу*, *йэ* with the corresponding letters *я*, *ё*, *ю*, *е*, as they are written in modern orthography (Afanasieva, 2007).

Siberian Lang The corpora from the *Siberian Lang*[7] project also contain Evenki texts in local dialects. The texts are provided in two formats: enumerated sentences in Cyrillic (1),

(1) *ēкуна-вал вāмтымил, нимокилдӯвэр борӣунакалду!*

and enumerated transcribed sentences with segmentation, morphological annotation and translation in Russian (see Figure 2).
The texts were automatically extracted from the website. The corresponding Cyrillic spelling of the word was found for each transcription using the numbers of transcribed sentences and sentences in Cyrillic.
The collected data and the scripts which were created for mapping the transcription to Cyrillic words and segmentation are available at the repository with the developed analyser.[8] Most Evenki examples in this paper are taken from the Siberian Lang corpora. If an example is from the other source, the source is specified.

2.2. Existing morphological analysers

This section describes two existing morphological analysers for Evenki, `evenkiMorph` and the Evenki analyser from IEA RAS.

evenkiMorph [9] is a morphological analyser, which was developed by means of a free/open-source finite-state toolkit `foma` (Hulden, 2009).
It is able to produce analyses for the open classes of nouns, adjectives, verbs, adverbs, and numerals. Function words are not included. The lexicon contains 17,500 words with Russian translation. The `lexc` file consists of 40 continuation lexica.[10] Phonological rewrite rules consist of 16 rules that cover most common morphological alternations, such as consonant assimilation and *i*-epenthesis.
The transducer processes words in an *ASCII phonetic string* format. Words in Cyrillic can be converted to the required format using a Python script which is provided with the analyser. An example of the analysis for a noun *бултаван* with a root *булта* 'hunt', an accusative marker *-ва*, and a 3rd-person singular possessive affix *-н* is given in Table 2. The wordform *бултаван* receives two analyses, since the transducer contains two lexical forms with different Russian translations for the noun *булта*: `bulta_охота` '~hunt' and `bulta_промысел` '~foraging'.
The evaluation of `evenkiMorph` on the available corpora shows higher results for a corpus with Evenki texts in the literary language and lower results for texts in dialects.

The Evenki morphological analyser at IEA RAS , which is available at the Linguistic corpora IEA RAS, was also developed by means of a finite-state toolkit `foma` (Hulden, 2009). There are about 58 sequential rewrite rules,

[4] Two reviewers note that Evenki has a very small speaker population, but we note that sites like `http://www.indigenoustweets.com/` record many instances of social-media interaction in indigenous languages between much smaller language communities.

[5] `https://drive.google.com/open?id=1he2q6RncA_NKHPIJjSzlkK-2qgEFTiCG`

[6] `http://corpora.iea.ras.ru/corpora/texts.php`

[7] `http://minlang.srcc.msu.ru/`

[8] `https://github.com/zu-ann/apertium-evn`

[9] `https://github.com/gisly/evenkiMorph`

[10] A continuation lexicon is a set of morphemes used for modelling morphotactics. These are linked together in a graph-like structure. See Beesley and Karttunen (2003) for details.

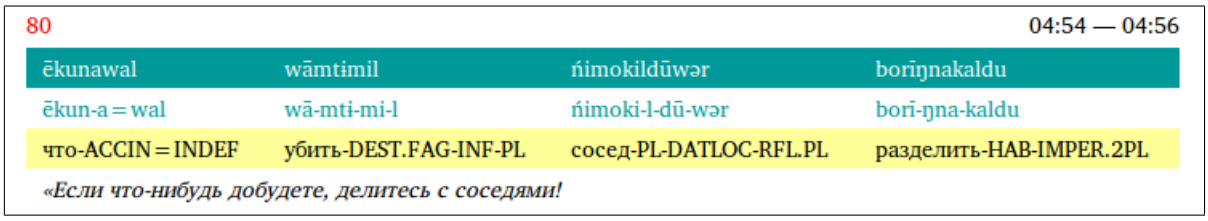

Figure 2: Example of a transcribed Evenki sentence with segmentation, morphological annotation and translation in Russian from the Siberian Lang corpora. The sentence translates as "If you get something, share it with your neighbours".

Cyrillic:	бултаван
	↓
ASCII string:	bultawan
	↓
Analyses:	bulta_охота+N+Sg+Acc+Ps3Sg
	bulta_промысел+N+Sg+Acc+Ps3Sg

Table 2: Example of output of `evenkiMorph` for the word *бултаван* 'hunt-ACC.SG3'

Cyrillic:	бултаван
	↓
Analysis:	бултаван
	охота[N]:сn-SG-ACC-POSS.3SG

Table 3: Example output from the IEA RAS analyser for the word *бултаван* 'hunt-ACC.SG3'

90 continuation lexica, and approximately 820 words in lexicon. In comparison to `evenkiMorph`, the analyser processes words in Cyrillic, contains function words and includes more derivational affixes.

An example of the analysis is given in Table 3. The analyser returns the wordform with the Russian translation and grammatical tags. `cn` stands for *common noun*.

3. Development

This section describes the development of the transducer we created, paying particular attention to nominal and verbal inflection and phonological processes.

3.1. Notes on orthography

Modern Evenki orthography is based on the Russian Cyrillic alphabet with an addition of a letter *ŋ* for the nasal /ŋ/.

Letters *д* and *н* represent both /d/, /n/ and palatal /dʲ/, /nʲ/[11], which can be distinguished by a subsequent vowel letter. Following Russian conventions, vowel letters *я*, *е*, *ё*, *ю* are written after palatal instead of *а*, *э*, *о*, *у*. For example, *дяв* /dʲɑβ/ 'boat' and *даға* /dɑgɑ/ (Afanasieva, 2010).

However, it is not clear in all cases by the spelling which phoneme the letter indicates. Vowels /e:/ and /ə:/ cannot be distinguished after *м*, *д* and *н*, as, for example, *н* in *нэ̄* can denote /n/ (/e:/ is written as *э̄*) and /nʲ/ (/ə:/ is written as *э̄*).

[11]Transcriptions in grammars are given using different notation. In this paper the transcription is written following Afanasieva (2010) and de Boer (1996).

3.2. Nominal inflection

Evenki nouns are inflected for number, case, and optionally for possession. The morpheme ordering of an Evenki noun form[12] is given in (2).

(2) *stem - derivational affix(es) - indirect possession - number - case - alienable possession - personal / reflexive possession - clitic*

Alternations in case and possession affixes can be mainly described using `twol`. Apart from assimilation, additional rules were required for the description of *i*-epenthesis. A conjunctive vowel *-u-* /i/ is inserted before number, case and possession markers if a stem ends in a consonant. `{i}` is used to indicate the conjunctive vowel and by default is realised as *u*.

Rules for number, case and possession affixes are also used for other nominal categories (for instance, adjectives and pronouns) which may inflect similarly. Adjectives inside the noun phrase do not agree with the noun, but may be used with nominal suffixes in case of ellipsis.

3.3. Verbal morphology

Bulatova (2002) defines the following ordering of Evenki verbal affixes:

(3) *stem - derivational affix - evaluation - aspect - tense/mood - agreement*

The analyser at IEA RAS follows this ordering and allows the verb to have only one aspectual marker. In addition, affixes of non-finite forms take the same slot as evaluative affixes in the analyser, therefore non-finite forms with evaluative or aspectual markers cannot be analysed. As a result, the transducer does not cover all possible verbal forms, for example, a participle with two aspectual markers from the Siberian Lang corpora *уӈку-ли-де-ча* /uŋku-li-dʲə-tʃɑ/ 'pour-INCH-IPFV-PANT' does not receive an analysis.

In contrast, `evenkiMorph` has two slots for aspectual markers: `VAspect` and `VSubaspect`. This realisation corresponds to the description of Evenki aspectual markers in Bulatova and Grenoble (1999). Bulatova and Grenoble divide aspectual markers into two primary aspects (*imperfective* and *perfective*, the latter is marked by a zero morpheme and is in binary opposition to the imperfective) and eight subaspects. They underline that a verbal stem in Evenki can have

[12]Indirect possession was added according to Bulatova and Grenoble (1999) and the Siberian Lang corpora.

more than one sub-aspect affix. `evenkiMorph` allows finite and non-finite forms to have an aspect marker followed by any number of sub-aspect affixes. However, the wordforms with an aspect marker after a sub-aspect marker (the same example is relevant: *уŋку-ли-де-ча* /uŋku-li-dʲə-tʃɑ/ 'pour-INCH-IPFV-PANT'), cannot be analysed. As for evaluative markers, they are not specified in `evenkiMorph`.

Nedjalkov (1997) describes the following ordering of Evenki verbal affixes:

(4) *stem - derivational affix - valency - voice - modality - aspect - evaluation - aspect - tense / non-indicative moods / non-finite forms - agreement - similarity*

A maximal morphemic chain of Evenki verb, which is presented in Nedjalkov (1997) and shown in (5)[13], includes 19 positions. All the positions, except for positions 17 *tense/non-indicative moods/non-finite forms* and 18 *agreement*, are optional.

(5) Verb stem (stem forming affix) -

1	dispersive	*-ктА-*	11	iterative	*-вАт-*
2	causative	*-вкАн-*	12	quick action	*-мАлчА-*
3	sociative	*-лды-*	13	evaluation	*-кАкут-*, *-кАт-*, *-влА-*, *-мАтыт-*, *-ма-*
4	reciprocal	*-мАт-*	14	ingressive	*-л-*
5	passive	*-в(у)-*	15	imperfective	*-дЯ-*
6	directive	*-нА-*	16	habitual	*-ӈна-*
7	conative	*-ссА-*	17	tense / non-indicative moods / non-finite forms	
8	semelfactive	*-син-*	18	agreement (personal / reflexive)	
9	desiderative	*-му-*	19	similarity marker	*-гАчин*
10	continuous	*-т-*			

However, positional variability of suffixes in (5) is possible. For example, in the Siberian Lang corpora 14 wordforms have the inchoative (ingressive in Nedjalkov (1997)) marker *-л-* before the imperfective *-дЯ-*, while 10 wordforms have imperfective before inchoative: *и-л-де-вкӣ* 'enter-INCH-IPFV-PHAB' vs. *дєву-де-ли-вкӣ* 'eat-IPFV-INCH-PHAB'. Verbal affixes, which are not adjacent in the ordering (5), can also exchange places: *ā-hӣн-му-т-та-н* 'sleep-INCEP-CAUS-DISTR-NFUT-3SG' vs. *нā-в-ни-рэ-н=кэ* 'hit-CAUS-INCEP-NFUT-3SG=FOC'.

In order to clarify, if using the description in Nedjalkov (1997) allows to significantly improve the quality of the analyser, four test analysers are evaluated on the available corpora. One of the analysers is designed to include free ordering of the listed verbal affixes, so that we can calculate how much wordforms in the ordering (5) it does not support in comparison to the free ordering. The analysers differ only in the possible combinations of markers between the positions 1 and 17, each analyser contains the same number of verbal affixes. Irregular forms of non-future tense and participles, except for the negative auxiliary э-, are not included in the test analysers.

1 Analyser IEA RAS follows the description of Bulatova (2002) and the transducer at IEA RAS. The slots for affixes are strictly determined and two affixes from the same slot cannot occur in a word.

2 Analyser evenkiMorph is based on the implementation of `evenkiMorph`. Finite and non-finite forms are allowed to have an aspect marker followed by any number of sub-aspect markers.

3 Analyser Nedjalkov represents the morpheme ordering, which was suggested in Nedjalkov (1997) and is shown in (5). Verbal affixes, which are present in other test analysers, but are not described in (5), are added to the analyser according to the ordering in (4) and Nedjalkov (1990).

4 Analyser Free allows a free combination of affixes which can appear before the position 17 (tense / non-indicative moods / non-finite forms) in (5).

The results are shown in Table 4. Analyser IEA RAS, based on the transducer at IEA RAS, gives lower scores in comparison to others analysers. Consequently, the final transducer should produce analysis for combinations of aspectual markers and for non-finite forms with aspectual and evaluative markers. Analyser *evenkiMorph*, which does not allow the imperfective marker *-дЯ-* to attach to a sub-aspect marker, receives less impressive results, therefore this feature is also required.

The other analysers do not have a significant difference in results. *Analyser Nedjalkov* gives less analyses per word, so that the analyses are less ambiguous. At the same time, it is not able to analyse words, where verbal suffixes do not directly follow the ordering in (5). The results in Table 4 show that these cases are not very rare.

In constrast, *Analyser Free* can produce analysis for wordforms with different ordering of affixes. However, the mean ambiguity also increases, as wordforms receive more possible analyses. For the final transducer the structure of *Analyser Free* is chosen in order to reduce the number of wordforms without analysis. If less ambiguity is required, *Analyser Nedjalkov* can be used.

	Coverage (Mean ambiguity)			Evaluation on texts		
	News.	Sib. Lang	IEA RAS	P	R	F-score
1	76.7 (1.7)	59.4 (1.2)	58.6 (1.1)	27.5	42.1	33.3
2	77.6 (1.8)	61.0 (1.3)	60.4 (1.3)	29.9	47.8	36.7
3	77.7 (1.8)	61.1 (1.3)	60.4 (1.3)	**30.1**	48.2	**37.0**
4	**77.8** (1.9)	**61.0** (1.3)	**60.5** (1.3)	30.0	**48.3**	**37.0**

Table 4: Evaluation of the test analysers (1..4) with different morpheme orderings.

3.4. Vowel harmony

Evenki suffixes follow vowel harmony rules. The vowel in a suffix is chosen depending on the stem vowel. Most of the suffixes have three vocalic variants (Nedjalkov, 1997): *-а* /a/, *-э* /ə/ and *-о* /o/ (e.g. non-future tense marker *-ра*, *-рэ*, *-ро*). Suffixes may also have only one vowel variant, when the suffix vowel belongs to neutral vowels *u* /i/, *ū* /i:/, *ӯ* /u/ or *ȳ* /u:/[14].

[13] А and Я represent the existence of vowel harmony variants.

[14] These vowels are regarded as neutral in the Evenki literary language, however, for example, Vasilevich describes /i/, /i:/ and /ɨ/, /ɨ:/, /ʉ/, /ʉ:/ and /u/, /u:/, which follow rules of vowel harmony (Vasilevich, 1940; Vasilevich, 1948)

Stem vowels	Suffix vowels
If the short vowel is in the first syllable:	
1) а, я	а, ā, я, я̄
2) о, ё (w/ neut. vowel between stem and suffix)	
3) э, е	э, э̄, е, е̄
4) о, ё	о, ё
After any syllable with the long vowel:	
5) ā, я̄	э, е, ā, я̄
6) э̄ (word-initial and after *m*, *д*)	
7) е̄ (in the first syllable after consonants)	а, ā, я, я̄
8) о̄, ё̄	
9) э̄, е̄	э, э̄, е, е̄

Table 5: The correspondence between stem and suffix vowel letters, based on Vasilevich (1948)

Neutral vowels, unlike other vowels, can appear in a word with vowels of any other group.

According to the description of Evenki vowel harmony in Bulatova and Grenoble (1999), vowel harmony rules can be defined as follows:

1. stems with vowels /ɑ/, /ɑ:/, /e:/, /o/, /o:/ attach suffixes with /ɑ/, /ɑ:/;
2. stems with vowels /ə/, /ə:/ attach suffixes with /ə/, /ə:/;
3. stems with only neutral vowels *u* /i/, *ū* /i:/, *y* /u/ and *ȳ* /u:/ attach suffixes with /ɑ/, /ɑ:/ or with /ə/, /ə:/.

The correspondences of the stem and suffix vowel letters are shown in Table 5 (Vasilevich, 1948)[15]. Vowel letters are given instead of vowel phonemes, since different letters can represent the same sound (see 3.1.). Vasilevich (1948) describes separately the cases for short and long vowels. The first part of Table 5 contains information about suffix vowel letters that can follow the corresponding short vowel letters. The second part indicates possible suffix vowel letters after the letters which represent long vowels.

Neutral vowels are not included in the table, since the following suffix can have either а, ā, я, я̄ or э, э̄, е, е̄. When the stem contains only neutral vowels, the choice of the suffix vowel is not clear, therefore the information about the correct vowel variant should be stored with the word entry in the morphological analyser. Symbols `{а}` and `{э}` are used for this purpose: they trigger *а*, *ā*, *я*, *я̄* or э, *э̄*, *е*, *е̄* in suffixes respectively. The information about the correct vowel variant was collected from Russian-Evenki dictionary (Boldyrev, 2000), where the wordform with the correct affix vowel is given either in accusative definite (*ȳ* 'scraper', *ȳ-вэ* 'scraper-ACCD') or in non-future tense 3rd-person singular (*хус-мȳ* 'cut', *хус-ма-н* 'cut-NFUT-3SG').

For words in lexicon, which contain only neutral vowels and are not mentioned in the dictionary (Boldyrev, 2000), symbols `{а}` and `{э}` were added according to the most frequent wordform in corpora with the corresponding stem.

The rules from Table 5 were converted to the `twol` format. Four sets of vowel letters were created:

[15]In Vasilevich (1948) only suffix vowels in line 5 are divided into short and long, while others are not marked for vowel length. In Table 5 long suffix vowels are included.

	Coverage (Mean ambiguity)		
	Newspaper	Siberian Lang	IEA RAS
1) original rules	75.22 (1.9)	59.95 (1.3)	57.64 (1.2)
2) neutral vowels	77.63 (2.0)	61.11 (1.3)	59.73 (1.3)
3) without (4)	77.49 (2.0)	61.16 (1.3)	59.55 (1.3)
4) without (5)	77.23 (1.9)	60.50 (1.3)	59.63 (1.3)
5) without (6)	**78.11** (2.0)	**61.74** (1.4)	**60.59** (1.4)
6) preceding syllable	**78.11** (2.0)	**61.74** (1.4)	**60.59** (1.4)

Table 6: Coverage evaluation of different `twol` vowel harmony rules, due to orthographical issues and variations in description.

	Evaluation on texts		
	precision	recall	*F*-score
1 original rules	28.96	46.35	35.65
2 with neutral vowels	29.48	47.91	36.50
3 without (4)	29.25	47.68	36.25
4 without (5)	29.66	47.33	36.47
5 without (6)	29.49	48.13	36.57
6 preceding syllable	**29.75**	**48.80**	**36.96**

Table 7: Evaluation metrics of different `twol` vowel harmony rules using standard metrics: *precision*, *recall* and *F-score*.

- `VowA` *а*, *я*
- `VowE` *э*, *е*
- `VowO` *о*, *ё*
- `VowNeutral` *и*, *ы*, *у*, *ю*

An archiphoneme `{A}`, which can be realised as *а*, *э* or *о*, is mostly used for vowel harmony description. At the same time, due to the Evenki orthography, some affixes require vowel letters *я*, *е*, *ё* instead of *а*, *э*, *о*, for instance, the imperfective marker *-дЯ-*, as the preceding consonant is *д* /dʲ/, and a marker of accusative indefinite is *-Я-*, since after vowels it is a combination of /j/ with a vowel. For these cases an archiphoneme `{Я}` was created: it follows the rules for `{A}` with an addition of specific constraits that deal with orthographic issues.

The analyser, which is implemented using the described rules of vowel harmony in Table 5, produces the results shown in Table 6 (row 1).

However, the current rules are not able to process words, similar to *бира* 'river' (*бира-ва* 'river-ACCD'), where the first vowel is neutral, but the subsequent root vowel can give information about vowel variants in suffixes. If the left context in line (7) of Table 5 is changed, so that the first non-neutral vowel has an influence on vowel harmony, more wordforms receive analysis and evaluation results increase: Table 6 (row 2).

Line (6) in Table 5 describes the cases, when the letter э̄ represents vowel /e:/, and therefore *а*, *ā*, *я*, *я̄* are required in suffixes. However, in the same positions this letter can denote vowel /ə:/ and attach different suffixes. If the rule for word-initial э̄ in line (6) is removed, evaluation results in Table 6 (row 3) do not increase, as well as if the rule for *е̄* in line (7) is removed, see Table 6 (row 4). In contrast, deleting the rule for э̄ after `[ :д | :т ] :0*` in line (6) improves the

results in Table 6 (row 5). This can be explained by different frequency of /ə:/ and /e:/.
Konstantinova and Lebedeva (1979) describe vowel harmony for short vowels regarding the preceding syllable instead of the first syllable. Changing the rules do not influence greatly the results, see Table 6 (row 6). Results for evaluation with standard metrics such as *precision*, *recall* and *F-score* are shown at Table 7.
The description of vowel harmony in Evenki grammars by Nedjalkov (1997) and Bulatova and Grenoble (1999) is given only using vowel sounds without listing possible sequences of vowel letters. The differences from the Table 5 are listed below:

- Nedjalkov (1997)
  - stems with vowels /ʊ:/, /ɪ:/ attach suffixes with /ə/
  - stems with vowel /ə:/ attach suffixes with /ɑ/
- Bulatova and Grenoble (1999)
  - stems with final vowels /o/, /o:/ attach suffixes with /o/, /o:/

Evaluation results with changes, according to Nedjalkov (1997) and Bulatova and Grenoble (1999), are given in Table 8. In comparison to the previous results, the scores did not improve. For the final analyser the `twol` file, which produced results in Table 6 (row 6), was chosen.

3.5. Lexicon

Both dictionaries and grammars were used to fill the lexicon, which amounts to approximately 34,000 words. Verbs, nouns, adjectives and adverbs were extracted from dictionaries, while words that belong to other parts of speech were collected from grammars.
To the best of our knowledge, there are no available Evenki dictionaries which can give grammatical information about the words. Therefore words were extracted from Evenki-Russian and Russian-Evenki dictionaries. Evenki-Russian (Boldyrev, 2000) and Russian-Evenki (Vasilevich, 2005) dictionaries are available in digital format at the website *Evengus*[16], which contains training materials for learning Evenki. An OCR version of Evenki-Russian dictionary by A. N. Myreeva (Myreeva et al., 2004) is provided at *Evenkiteka*[17], an online library with books in the Evenki language and about it. The dictionary includes also information about dialectical variants of words. Dialectical variants were collected and processed in the same way as other words, since corpora *Siberian Lang* and *IEA RAS* contain texts in dialects.
Unlike other parts of speech, verbs have special notation in dictionaries, so that the part of speech can be easily determined. They contain either an affix *-мū ō-мū* 'to do' or end with a hyphen, indicating a bare stem *ō-*. In the latter case the stems can also belong to other parts of speech, and for this reason the extracted stems were proofread, however, they needed only a small number of corrections.

[16] `http://www.evengus.ru/prilozheniya/lingvo/`
[17] `http://evenkiteka.ru/`

In order to determine the part of speech of other words, Russian translations were used. If a translation was a phrase, then the head of the phrase was chosen. POS-tagging and dependency parsing was produced using Russian-SynTagRus, a UDPipe model (Straka and Straková, 2017) for Russian. In most cases nouns and adjectives were defined correctly, however, some adjectives, adverbs, participles and converbs were also determined as nouns, so the resulting lists of words were checked and corrected.
As for the drawbacks of the OCR version of the dictionary (Myreeva et al., 2004), not all symbols were recognized, so the process of extracting words and their definitions required also fixing unrecognized symbols. Apart from that, many words in the OCR version do not have vowel length marks in comparison to the original dictionary in `pdf` format. In order to reduce the number of words with missing vowel length mark, the words without a macron were removed if they had an equivalent with vowel length mark in other dictionaries (Boldyrev, 2000; Vasilevich, 2005).
The dictionaries and therefore the `lexc` file contain a lot of stems with derivational affixes. Accordingly, most of derivational affixes are not included in the basic version of the analyser. In this case adding derivation increases the mean ambiguity much than the coverage (see Section 4.1.): extra analyses are produced both for the stems with derivational affixes, which are already stored in the `lexc` file, and for the wordforms which do not have a derivational affix, but are homographous to a stem with it.
A version of the analyser with derivational affixes was created based on the description in Nedjalkov (1997). Deverbal verbal affixes and evaluative affixes are included also in the basic version of the analyser, due to the morpheme ordering, described in Section 3.3., and low frequency in dictionaries respectively.

3.6. Loan words

There are several differences in inflection of Russian loan words in comparison to Evenki words, therefore additional rules are required. Russian loan words ending in *и* /n/ take case and possession markers like the Evenki words ending in *и* /n/, but attach the plural marker *л* /l/ instead of *р* /r/: *банан-ыл* 'banana-PL' (the example from the Newspaper corpus). A separate continuation class was added for processing these words. Russian loan words, which end in consonants, except for nasal, inflect like Evenki words ending in voiceless consonants: *город-му* 'city-DATLOC' (Konstantinova and Lebedeva, 1979). A special sign {Ҳ} was added to mark loan words, so as in `twol` they were processed as Evenki words, which end in voiceless consonants.
Suffix vowels in loan words are chosen following vowel harmony rules (Konstantinova and Lebedeva, 1979; Afanasieva, 2007). Stressed vowels in Russian loan words are regarded as long vowels, therefore in `lexc` file stressed vowels in Russian loan words received a vowel length mark. The information about the stress in a word was obtained using a Python library `rupo`[18] for poem analysis and generation, which also provides methods for getting information about stress in Russian words.

[18] `https://github.com/IlyaGusev/rupo`

	Coverage (Mean ambiguity)			Evaluation on texts		
	Newspaper	Siberian Lang	IEA RAS	precision	recall	F-score
/ʊ:/, /ɪ:/	**77.49** (1.9)	**61.63** (1.4)	59.60 (1.3)	29.53	**48.40**	**36.68**
/ə:/	77.38 (1.9)	61.33 (1.3)	59.01 (1.3)	**29.55**	47.94	36.57
/o:/	77.48 (2.0)	61.03 (1.3)	**59.69** (1.3)	29.43	48.07	36.51

Table 8: Evaluation of different `twol` vowel harmony rules, due to variations in descriptions in Nedjalkov (1997), Bulatova and Grenoble (1999).

If a loan word contains a mixed set of vowels, then *а*, *а̄*, *я*, *я̄* are written in the suffix, so the loan words, which satisfied this condition, were marked with *{а}*. Loan words with only neutral vowels take suffixes with *э*, *э̄*, *е*, *е̄* therefore the sign *{э}* was added to such loan words in a `lexc` file.

3.7. Spellrelax

In order to cover common typographical variance, an additional `spellrelax` transducer was introduced. Marking vowel length is not obligatory in Evenki (for example, see Section 2.1.), therefore words without a macron, which indicates vowel length, should be also treated as words with a macron. `spellrelax` contains two mappings for a macron: `¯:¯` and `¯:0`. This means that a macron in a surface form (a macron before colon) produced by the analyser can correspond either to a macron in a surface form in text (a macron after colon) or to zero (zero after colon), so that a surface form in text may not have a macron at this position. Two vowels, *и* /i/ and *у* /u/ have unicode precomposed forms, *ӣ* and *ӯ*. These may also be substituted.

Further mappings were added for letters *ӯ*, which can be substituted with *н*, and *в*, which is sometimes written in the Siberian Lang corpora as *ф*. In addition, `spellrelax` supports mapping *h* to *x*, since /h/ can be represented using a letter *x* or a letter *h*: the word *хугун* /hugun/ 'steam (from the breath)' (Boldyrev, 2000) can be written as *hугун* (Myreeva et al., 2004).

As it was mentioned in Section 2.1., both a letter *ӯ* and a combination of letters *нг* may indicate Evenki nasal /ŋ/. The sequence *нг* cannot be just replaced with *ӯ*, since there are Evenki words such as *илинго-* 'to stop (suddenly)' or loan words, for example, *английскай* 'English', where this combination should be treated as /ng/. A rule for mapping *нг* to *ӯ* was also added to `spellrelax`.

3.8. Dialectal features

The Evenki people are dispersed throughout the large territory, and there are many Evenki dialects. The dialects are grouped into northern, southern and eastern (Figure 1). Evenki dialects do not have many differences in grammar, but "differ greatly, as far as phonetics and vocabulary are concerned" (Nedjalkov, 1997). In order to improve the analysis of texts in Evenki dialects, a relaxed version of the analyser was created.

The variation /s/ ∼ /h/ is the most characteristic difference in phonetics (Nedjalkov, 1997):

- /s/ in a word-initial position in eastern and southern dialects corresponds to /h/ in northern dialects: *сулаки* — *хулаки* 'fox';
- words with intervocal /s/ in southern dialects have intervocal /h/ in eastern and northern dialects: *эси* — *эхи* 'now';
- some words begin with /h/ in all dialects: *хава* 'work'.

	Newspaper	Siberian Lang	IEA RAS
ороrво	**193**	**7**	**11**
ороrвэ	1	4	0

	Newspaper		Siberian Lang
дуниэлдулэ	**22**	*никэрэн-дэ̄*	2
дуниэлдуло	8	*никэрэн-до̄*	4
дуниэлдула	1	*никэрэн-да̄*	**5**

Table 9: Example of words with different vowel variants and their frequency in corpora.

Literary Evenki is based on the southern dialects with /s/ in both intervocalic and word-initial positions, therefore rules, which allow mapping *x* and *h* to *c* in the beginning of the word and between vowels, were added to `spellrelax`.

In addition, southern dialects are also divided into hissing and hushing, which can have /ʃ/ instead of /s/. Mapping of *ш* to *c* was also added to `spellrelax`.

Another modification, which allows to improve the analysis of texts in dialects, is disregarding vowel harmony, due to dialectal specifities in phonetics (Vasilevich, 1948). For instance: apart from the literary form *оро-р-во* 'deer-PL-ACCD', a word form *оро-р-вэ* with *э* instead of *о* can also be found. Three vowel variants are possible as well: the annotation 'get.ready-NFUT-3SG=FOC' corresponds to *никэ-рэ-н-да̄*, *никэ-рэ-н-до̄*, and *никэ-рэ-н-дэ̄* in the Siberian Lang corpora, forms *дуннэ-л-дулэ*, *дуннэ-л-дуло*, and *дуннэ-л-дула* 'land-PL-LOCALL' can be found for *дуннэ* 'land' in plural locative-allative in the Newspaper corpus. The frequency of the mentioned wordforms in corpora is given in Table 9.

In the relaxed transducer archiphonemes `{А}` and `{Я}` correspond to all possible vowels in a particular position, so that all the examples above are analysed.

A special archiphoneme `{Ч}`, which can realise as *ч*, *с*, *ш*, *щ*, was created for the affixes of past tense *-чА-* and participle of anteriority *-чА-*. For example, in the Siberian Lang corpora (the number of tokens found is given in brackets) the annotation 'be-PST-3SG' belongs to: *би-чо̄-н* (53), *би-со̄-н* (33), *би-сё̄-н* (13), *би-чэ-н* (13), *би-чо-н* (6), *би-сё-н* (5), *би-чэ̄-н* (2), *би-са̄-н* (1), *би-шё̄-н* (1), *би-що̄-н* (1). Similarly, the annotation 'do-PANT' corresponds to: *о-са* (5), *о-ча* (4), *о̄-са* (1), *о̄-ся* (1), *о-ся* (1).

Another archiphoneme `{В}`, which can realise as *в*, *б* and *п*, was created for *в* /β/ in affixes, as can be found in the cases

	Coverage (Mean ambiguity)			
	News.	Sib. Lang	IEA RAS	Average
bare analyser	78.4 (1.9)	61.7 (1.3)	61.1 (1.3)	67.1 (1.5)
+ D	81.5 (2.7)	63.6 (1.9)	63.4 (2.1)	69.5 (2.2)
+ SR	84.1 (2.3)	75.9 (1.8)	68.2 (1.7)	76.0 (1.9)
+ D and SR	**87.7** (3.5)	**78.8** (2.9)	**71.8** (2.7)	**79.4** (3.0)

Table 10: Coverage and mean ambiguity of the developed analysers. D refers to derivations and SR refers to spellrelax.

	Precision	*Recall*	*F-score*
bare analyser	30.58	49.45	37.79
+ D	26.27	52.53	35.03
+ SR	**35.52**	67.30	**46.50**
+ D and SR	29.56	**70.26**	41.61

Table 11: Precision, recall and F-score between the tags from annotation and the tags from the analyser. D refers to derivations and SR refers to spellrelax.

of regressive assimilation (6, 7) and progressive assimilation (8, 9).

(6) оро-р-во deer-PL-ACCD
оро-р-бо deer-PL-ACCD

(7) хунāт-ви girl-ACCD.REFL
хунāт-пи girl-ACCD.REFL

(8) hӣ-в-де-ми extinguish-PASS-IPFV-INF
ичэ-б-де-рэ-н see-PASS-IPFV-NFUT-3SG

(9) хокори-в-са-тин lose-PASS-PST-3PL
тага-п-сā trap-PASS-PANT

4. Evaluation

4.1. Coverage and mean ambiguity

The evaluation results of the final analyser and the versions with relaxed rules and extended derivation are shown in Table 10.

The current analyser achieves higher coverage results all the Evenki corpora in comparison to other Evenki analysers, which were described in Section 2.2.. Adding derivation and relaxing the rules allows to further increase the number of words, which receive analysis, but also influences the mean ambiguity.

4.2. Evaluation using annotated texts

Table 11 shows precision, recall and F-score in assigning morphological tags. To the authors' knowledge the only available morphologically annotated Evenki data set *Siberian Lang* corpus used for evaluation consists of the text recorded in the course of field work. The inconsistency between the tags from the corpus and our analyser is mainly solved by creating mapping rules. Morphological tag assignment task implies the comparison between the 'gold' tags from the aforementioned corpus with the tags assigned by morphological analyser which can be ambiguous. Low precision score may be explained by this unresolved ambiguity when the analyser produces more forms than the corpus has for the particular lexeme.

	bare analyser	analyser relax + derivation
vowel harmony	27	0
dialectal features	18	12
incorrect spelling	6	5
missing stem	3	11
missing morphotactics	3	10
vowel length	1	12

Table 12: Error analysis (a random sample of 50 incorrect outputs for each analyser).

In addition, a random sample of 100 wordforms with incorrect or missing tags was collected and analysed. First 50 tokens were taken from the output of the analyser which does not include extended derivation or relaxed rules, and the second 50 tokens belonged to the relaxed transducer with extended derivation. These analysers were chosen, as they allow to see, how the error types change after implementing relaxed rules and extended derivation and what errors this implementation does not cover.

Errors were categorised into 6 types: missing stem, missing morphotactics, vowel length, vowel harmony, dialectal features, incorrect spelling.

The comparison of the errors is shown in Table 12. Some wordforms combined errors of different types, therefore they were included in the numbers of all corresponding error types. For instance, an example of a missing stem and a dialectal feature is *восток-тикӣ=кун* 'east-ALL=FOC' with an allative marker *-тикӣ-* instead of *-тыкӣ-*. The errors marked as dialectal features include variations of affixes or stems, the literary form of which is stored in the analyser. Another example of this error is *исё-т-чо-ӈнэ-Ø-м* 'see-DUR-IPFV-HAB-NFUT-1SG' with missing stem. The dictionaries (Vasilevich, 2005; Myreeva et al., 2004; Boldyrev, 2000) do not contain a stem *исё-*, but according to the translation *ичэ-* can be found.

Errors with vowel harmony are the most common for the first analyser, as the correspondences of stem and suffix vowels letters in some wordforms, for example, *тавӯ-ӈнэ-рэ* 'gather-HAB-NFUT-3PL' or *дэрэ-лӧ-тын* 'face-LOCALL-PS3PL' cannot be described using the standard rules (Section 3.4.). These errors also include the cases when the stem contains only neutral vowels and either do not have special symbols `{a}`, `{ə}` for vowel harmony or these symbols do not trigger the required harmony. For instance, *буру-рак-ин* 'fall-CVCOND-PS3SG' does not receive analysis, since it has `{ə}` in the underlying form following the example *буру-рэ-н* 'fall-NFUT-3SG' in the dictionary (Boldyrev, 2007). The second analyser is able to process different combinations of stem and suffix vowels, therefore it does not have any errors of this type.

An example of an incorrect spelling is *буру-кэл=дэ̄* 'fall-IMPER.2SG=FOC', which is written as *бурукэлдэ̄* without a hyphen for separating a clitic *дэ̄*. The current transducers are

able to process only clitics separated with a hyphen (following Evenki orthography) in order not to allow processing affixes as clitics. Vowel length errors occur, when a vowel length mark is present in a wordform, but is missing in the corresponding stem or suffix in `lexc`. For instance, the root *hāpɐu* in *hāpɐu-л* 'devil-PL' is stored in `lexc` as *xapɐū*, since this spelling is in Boldyrev (2000) and Vasilevich (2005).[19]

5. Concluding remarks

This paper has presented a new morphological analyser for the Evenki language. The analyser gives higher coverage than the existing Evenki analysers. In addition to a basic analyser, the versions for processing texts in Evenki dialects and extended derivation have already been developed.

As for the future work, one of the possible directions is updating the lexicon. Since available machine-readable dictionaries do not contain grammatical information, the lexicon collection was implemented relying on the translations. However, having information about grammatical features of the words, for example, transitivity will allow to reduce the ambiguity. Alternatively, as can be seen in Table 12, most of the errors for the relaxed transducer with extended derivation were caused by missing morphotactics, vowel length marks and stems, therefore future work at this issues is also required. Apart from the changes mentioned above, the HFST transducer can be extended to include weights, which represent probabilities of different analyses. Assigning the weights can further improve the quality of the analyser.

The developed morphological analyser can be a useful aid for analysing and lemmatising Evenki words as well as for generating and segmenting wordforms. It can be used for many practical purposes, e.g. for automating the process of text annotation, generating wordforms, creating proofing tools, information retrieval and for machine translation.

6. Acknowledgements

We would like to thank Elena Klyachko for her valuable comments on the Evenki grammar and the anonymous reviewers for their extensive and constructive feedback.

7. Bibliographical References

Afanasieva, E. F. (2007). *Orfograficheskij slovar evenkijskogo yazyka. Uchebnoe posobie dlya uchashhixsya 5–8 klassov obshheobrazovatelnoj shkoly*. Drofa. [Evenki orthographic dictionary. Educational material for students of 5–8 grades of general education school].

Afanasieva, E. F. (2010). To the question about principles of the classification of the Evenk consonant phonemes. *Vestnik Zabajkalskogo gosudarstvennogo universiteta (Bulletin of ZabGu)*, (6).

Beesley, K. R. and Karttunen, L. (2003). *Finite-state morphology: Xerox tools and techniques*. CSLI, Stanford.

Boldyrev, B. V. (2000). *Evenkijsko-russkij slovar*. Novosibirsk: Academic Publishing House Geo Ltd. [Evenki-Russian dictionary].

[19] Although the word *hāpɐu* can be found in Myreeva (2004), the OCR version contains *hapɐu* without vowel length mark. *hapɐu* was not included in `lexc`, since *xapɐū* with the similar translation and vowel length mark was retrieved from other dictionaries, see Section 3.5.

Boldyrev, B. V. (2007). *Morfologija evenkijskogo jazyka [The Morphology of the Evenki Language]*. Novosibirsk: Nauka Publ.

Bulatova, N. and Grenoble, L. A. (1999). *Evenki*, volume Languages of the world: Materials (Vol. 141). Lincom Europa.

Bulatova, N. (2002). *Evenkijskij yazyk v tabliczax: Uchebnoe posobie dlya evenkijskix shkol, pedagogicheskix kolledzhej, vuzov*. Drofa. [The Evenki language in tables: Educational material for Evenki schools, pedagogical colleges, higher educational institutions].

de Boer, E. (1996). Present state of the study of Evenki vowel harmony. In *Proceedings of the 38th Permanent International Altaistic Conference (PIAC)*, pages 121–133, 01.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Hulden, M. (2009). Foma: a finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*, pages 29–32. Association for Computational Linguistics.

Konstantinova, O. A. and Lebedeva, E. P. (1979). *Evenkijskij yazyk*. Gosuchpedgiz. [The Evenki language].

Koskenniemi, K. (1983). *Two-level morphology: A general computational model for word-form recognition and production*, volume 11. University of Helsinki, Department of General Linguistics Helsinki.

Lindén, K., Axelson, E., Hardwick, S., Pirinen, T. A., and Silfverberg, M. (2011). HFST–framework for compiling and applying morphologies. In *International Workshop on Systems and Frameworks for Computational Morphology*, pages 67–85. Springer.

Myreeva, A., Andreeva, T., Bagaeva, P., Varlamova, G., and Kudrina, N. (2004). *Evenkijsko-russkij slovar [Evenki-Russian dictionary]*. Novosibirsk: Nauka Publ.

Nedjalkov, I. (1990). *Glagolnye kategorii v evenkijskom yazyke (zalog i vid) [Verbal categories in the Evenki language (voice and aspect)]*. Leningrad.

Nedjalkov, I. (1997). *Evenki. Descriptive grammar*. London: Routledge.

Straka, M. and Straková, J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics.

Vasilevich, G. M. (1940). *Ocherk grammatiki evenkijskogo (tungusskogo) yazyka*. Uchpedgiz. [A sketch of the grammar of the Evenki (Tungus) language].

Vasilevich, G. M. (1948). *Ocherki dialektov evenkijskogo (tungusskogo) jazyka*. Uchpedgiz. [Sketches of dialects of Evenki (Tungus)].

Vasilevich, G. M. (2005). *Russko-evenkijskij slovar*. Saint Petersburg: Prosveshcheniye. [Russian-Evenki dictionary].