

LowResourceEval-2019: a shared task on morphological analysis for low-resource languages

Klyachko E. L. (elenaklyachko@gmail.com)

Institute of Linguistics, HSE, Moscow, Russia

Sorokin A. A. (alexey.sorokin@list.ru)

Moscow State University, Moscow, Russia

Moscow Institute of Physics and Technology, Dolgoprudny, Russia

Krizhanovskaya N. B. (nataly@krc.karelia.ru)

Krizhanovsky A. A. (andrew.krizhanovsky@gmail.com)

Institute of Applied Mathematical Research of KarRC of RAS, Petrozavodsk, Russia

Ryazanskaya G. M. (galka1999@gmail.com)

Center for Language and Brain, HSE, Moscow, Russia

The paper describes the results of the first shared task on morphological analysis for the languages of Russia, namely, Evenki, Karelian, Selkup, and Veps. For the languages in question, only small-sized corpora are available. The tasks include morphological analysis, word form generation and morpheme segmentation. Four teams participated in the shared task. Most of them use machine-learning approaches, outperforming the existing rule-based ones. The article describes the datasets prepared for the shared tasks and contains analysis of the participants' solutions. Language corpora having different formats were transformed into CONLL-U format. The universal format makes the datasets comparable to other language corpora and facilitates using them in other NLP tasks.

Key words: morphological analysis, morpheme segmentation, minority languages, low-resource languages

LowResourceEval-2019: дорожка по морфологическому анализу для малоресурсных языков

Клячко Е. Л. (elenaklyachko@gmail.com)

Институт языкознания РАН, НИУ ВШЭ, Москва, Россия

Сорокин А. А. (alexey.sorokin@list.ru)

Московский Государственный Университет им. М. В. Ломоносова, Москва, Россия

Московский физико-технический институт, Долгопрудный, Россия

Крижановская Н. Б. (nataly@krc.karelia.ru)

Крижановский А. А. (andrew.krizhanovsky@gmail.com)

Институт прикладных математических исследований КарНЦ РАН, Петрозаводск, Россия

Рязанская Г. М. (galka1999@gmail.com)

Центр языка и мозга, НИУ ВШЭ, Москва, Россия

В статье описывается первое соревнование, посвященное морфологическому анализу малоресурсных языков России, а именно: эвенкийского, карельского, селькупского и вепсского. Указанные языки располагают корпусами небольшого размера. Соревнование включало в себя автоматическое определение морфологических признаков, деление на морфемы, а также синтез словоформ. В статье описываются корпуса, специально подготовленные для соревнования, а также анализируются методы, использованные его участниками. Наилучшие результаты показали модели, основанные на нейронных сетях.

Ключевые слова: морфологический анализ, морфемная сегментация, малые языки, малоресурсные языки

1 Introduction

According to the 2010 Census [1], more than 250 languages from 14 language families are spoken in Russia. About 100 of them are minority languages. It is worth noting that even non-minority languages, such as Yakut (Sakha), are considered vulnerable. For most languages of Russia, apart from Russian, digital resources either do not exist or are relatively scarce.

The shared task¹ was held from January to March, 2019. The aims of the shared task were as following:

1. to facilitate and stimulate the development of corpora and linguistic tools for minor languages:
 - (a) One of the results produced by the shared task are text corpora which are uniformly tagged and accessible online.
 - (b) The participants are obliged to share the resulting systems.
2. to inspire better communication between the communities of field linguists and NLP researchers;
3. to figure out how modern methods of morphological analysis, tagging, segmentation, and synthesis cope with sparse training data, the lack of standard language and large rate of dialectal varieties.

2 Related work

We present a short survey of corpora for minor languages of Russia. A detailed survey of Russian minority language corpora and morphology tools as of 2016 can be found in [2]. However, more corpora have been developed since then. Therefore we suppose that the topic should be revisited.

2.1 Corpora

Most corpus resources are created by language activists and are based on digitalized books and other printed materials. Some examples are the corpora created by The Finno-Ugric Laboratory for Support of the Electronic Representation of Regional Languages², The digital portal of Selkup language³ etc.

On the other hand, field data collected during linguistic expeditions is often transformed into corpora. These corpora are usually created by universities such as the corpora published at HSE Linghub⁴, VepKar⁵ at the Karelian Research Center, the Siberian-Lang language data⁶ collected by MSU and The Institute of Linguistics (Russian Academy of Sciences) and many other projects. Furthermore, some projects also leverage old field data, digitizing it. For example, in INEL project⁷, field data for Selkup and now extinct Kamassian language have been digitized and processed.

¹https://lowresource-lang-eval.github.io/content/shared_tasks/morpho2019.html

²<http://fu-lab.ru/>

³<http://selkup.org/>

⁴<https://linghub.ru/>

⁵<http://dictorpus.krc.karelia.ru/en>

⁶<http://siberian-lang.srcc.msu.ru/>

⁷<https://inel.corpora.uni-hamburg.de/>

Field data consists mostly of oral texts. Therefore, corpus materials often contain non-standard varieties of a language and demonstrate remarkable dialectal and sociolinguistic features. However, the high level of variation makes it challenging for automatic processing.

Table 1 shows the resources for the languages of Russia, which are available online as language corpora with search facilities. We do not include published books of interlinearized texts although they can be used as a source for future corpora.

Table 1: Morphological resources for languages of Russia

| Language | Tokens | Parallel languages | Markup | License |
|----------------------------------|------------|------------------------|--------------------------|-------------|
| Abaza ⁸ | 32 796 | Russian | no tags | NA |
| Avar ⁹ | 2 300 000 | - | tags, no disambiguation | NA |
| Adyghe ¹⁰ | 7 760 000 | Russian | tags, no disambiguation | NA |
| Archi ¹¹ | 58 816 | - | tags, not detailed | NA |
| Bagvalal ¹² | 5 819 | Russian | tags with disambiguation | NA |
| Bashkir ¹³ | 20 584 199 | - | tags, no disambiguation | NA |
| Beserman Udmurt ¹⁴ | 65 000 | Russian | tags with disambiguation | CC BY 4.0 |
| Buryat ¹⁵ | 2 200 000 | - | tags, no disambiguation | NA |
| Chukchi ¹⁶ | 6393 | English, Russian | tags with disambiguation | NA |
| Chuvash ¹⁷ | 1 147 215 | Russian (partially) | no tags | NA |
| Crimean Tatar ¹⁸ | 56 752 | - | no tags | |
| Dargwa ¹⁹ | 48 957 | Russian | tags with disambiguation | NA |
| Erzya ²⁰ | 3 130 000 | partially (Russian) | tags, no disambiguation | CC BY 4.0 |
| Evenki ²¹ | 121 286 | Russian (partially) | tags, no disambiguation | own license |
| Evenki ²² | 25 000 | Russian | tags with disambiguation | NA |
| Godoberi ²³ | 872 | English | tags with disambiguation | NA |
| Kalmyk ²⁴ | 858 235 | - | tags, no disambiguation | NA |

⁸https://linghub.ru/abaza_rus_corpus/search

⁹http://web-corpora.net/AvarCorpus/search/?interface_language=en

¹⁰<http://adyghe.web-corpora.net>

¹¹http://web-corpora.net/ArchiCorpus/search/index.php?interface_language=en

¹²http://web-corpora.net/BagvalalCorpus/search/?interface_language=en

¹³<http://bashcorpus.ru/bashcorpus/>

¹⁴<http://beserman.ru/corpus/search>

¹⁵http://web-corpora.net/BuryatCorpus/search/?interface_language=en

¹⁶<http://chuklang.ru/corpus>

¹⁷<http://corpus.chv.su>

¹⁸<https://korpus.sk/QIRIM>

¹⁹http://web-corpora.net/SanzhiDargwaCorpus/search/?interface_language=en

²⁰<http://erzya.web-corpora.net>

²¹<http://corpora.iea.ras.ru/corpora/news.php?tag=6>

²²<http://gisly.net/corpus>

²³http://web-corpora.net/GodoberiCorpus/search/?interface_language=en

²⁴http://web-corpora.net/KalmykCorpus/search/?interface_language=en

| | | | | |
|-------------------------------|-------------|---------------------------------|------------------------------------|------------------------|
| Karelian ²⁵ | 66 350 | Russian | tags with disambiguation (partial) | CC BY 4.0 |
| Khakas ²⁶ | 285 000 | Russian | tags, no disambiguation | NA |
| Khanty ²⁷ | 161 224 | Finnish, English, Russian | tags with disambiguation (partial) | CLARIN RES |
| Komi- Zyrian ²⁸ | 54 076 811 | - | tags, no disambiguation | NA |
| Mansi ²⁹ | 961 936 | - | tags (not detailed) | NA |
| Nenets ³⁰ | 125 421 | Russian (partially) | no tags | own license |
| Nenets ³¹ | 148 348 | Russian | no tags | CLARIN RES |
| Ossetic ³² | 12 000 000 | - | tags, no disambiguation | NA |
| Romani ³³ | 720 000 | - | tags, no disambiguation | NA |
| Selkup ³⁴ | 18 763 | English, German, Russian | tags with disambiguation | CC BY- NC-SA 4.0 |
| Shor ³⁵ | 262 153 | Russian (partially) | 1 | own license |
| Tatar ³⁶ | 180 000 000 | - | tags, no disambiguation | NA |
| Udmurt ³⁷ | 7 300 000 | - | tags, no disambiguation | NA |
| Veps ³⁸ | 46 666 | Russian | tags with disambiguation (partial) | CC BY 4.0 |
| Yiddish ³⁹ | 4 895 707 | - | tags, no disambiguation | NA |

2.2 Other shared tasks on low-resource evaluation

The main Shared Task concerned with morphological tagging is the well-known CoNLL Shared Task on parsing from raw data to Universal Dependencies [12]⁴⁰, [11]⁴¹. Though the main goal of this competition is evaluation of dependency parsers, it also deals with morphological analysis since morphological tags are used as features for further syntactic processing. The task included 82 corpora of different size for 57 languages in its 2017 edition, with the size of the corpora ranging from several hundred words to more than 1 mln. The shared task organizers name 9 treebanks as small (they contain from 4K to 20K words) and 9 as low-resource. The size of low-resource treebanks is less than 1000 words. These two

²⁵<http://dictorpus.krc.karelia.ru/en>

²⁶<http://khakas.altai.ca.ru>

²⁷<https://kitwiki.csc.fi/twiki/bin/view/FinCLARIN/KielipankkiAineistotKhantyUHLCS>

²⁸<http://komicorpora.ru>

²⁹<http://digital-mansi.com/corpus>

³⁰<http://corpora.iea.ras.ru/corpora>

³¹<http://www.ling.helsinki.fi/uhlcs/metadata/corpus-metadata/uralic-lgs/samoyedic-lgs/nenets>

³²http://corpus.ossetic-studies.org/search/index.php?interface_language=en

³³http://web-corpora.net/RomaniCorpus/search/?interface_language=en

³⁴<https://corpora.uni-hamburg.de/hzsk/en/islandora/object/spoken-corpus%3Aaselkup-0.1>

³⁵<http://corpora.iea.ras.ru/corpora/news.php?tag=3&period=>

³⁶<http://tugantel.tatar>

³⁷http://web-corpora.net/UdmurtCorpus/search/?interface_language=en

³⁸<http://dictorpus.krc.karelia.ru/en>

³⁹<http://web-corpora.net/YNC/search>

⁴⁰<http://universaldependencies.org/conll17/>

⁴¹<http://universaldependencies.org/conll18/>

categories of treebanks differ dramatically in terms of tagger performance: while the average accuracy of morphological tagging was 82% for small treebanks, the quality for low-resource language was only 25%. Therefore, the datasets used in our Shared Task better fit to “small” category than to the low-resource one. However, for most small treebanks in UD one can also learn from data either for a closely related language, for example, Finnish for North Sámi (sme_giella), or even a different corpus for the same language (Latin la_proiel corpus with 272K words for la_perseus corpus with 18K words). That is not the case for two main languages of our Shared Task, Evenki and Selkup, while for Veps and Karelian one can use Finnish or Estonian as an additional source.

The only known competition on morpheme segmentation was MorphoChallenge Shared Task ⁴² held from 2005 to 2010. The amount of labeled training data in its edition was rather small (about 1700 word types), however, the organizers provided an additional word list, which included several hundred thousands of unsegmented words since morpheme segmentation was usually treated as minimally supervised or semi-supervised problem. In recent studies on supervised morpheme segmentation, such as [6], the training dataset usually did not exceed 2000 words, though the amount of segmented data in our competition was even greater than in analogous studies.

The main Shared Task on morphological inflection, the Sigmorphon Shared Task [4] specially provided three types of training datasets: low-resource (100 words), middle (1000 words) and large (up to 10000 words).

3 Shared task description

The task consists of three tracks which are described below. Evaluation scripts can be found in our Github repository⁴³. The participants were allowed to use any external dataset. However, they were required to publish their solutions into open-source. It was done to accelerate NLP tool development for minor languages. The participants could provide several solutions.

3.1 Morphological analysis

The morphological analysis task was to produce lemmata, part-of-speech tags and morphological features for tokenized sentences. Training data was annotated using CONLL-U format⁴⁴ also used in Universal Dependencies project. We extended the annotation for the corpora without morphological disambiguation: in case a word had several analyses in corpus, we listed them all on consecutive lines. Words lacking morphological analysis in the corpora were annotated with distinguished UNKN tag. An example of the markup for Karelian can be found below:

```
13 julkaistuja UNKN _ _ _ _ _ _ _  
14 kirjoja kirja NOUN _ Number=Plur|Case=Par _ _ _ _
```

The following metrics were evaluated:

1. the fraction of word forms with correct lemmata;
2. the fraction of sentences where all word forms have correct lemmata;
3. the fraction of word forms with correct part-of-speech tags;

⁴²<http://morpho.aalto.fi/events/morphochallenge/>

⁴³https://github.com/lowresource-lang-eval/morphology_scripts/tree/master/evaluation

⁴⁴<https://universaldependencies.org/format.html>

4. the fraction of sentences where all word forms have correct part-of-speech tags;
5. precision, recall, and F1 score of predicted morphological features calculated according to:

$$\begin{aligned}
 P &= \frac{TP}{TP + FP}, \\
 R &= \frac{TP}{TP + FN}, \\
 F1 &= \frac{2PR}{P + R} = \frac{TP}{TP + 0.5(FP + FN)},
 \end{aligned}$$

where TP is the number of true positives (correct morphological features), FP is the number of false positives (incorrectly assigned morphological features) and FN is the number of false negatives (missed morphological features).

3.2 Morpheme segmentation

The training data consisted of tokenized sentences with each word split into morphemes. The task was to train a model which could produce morpheme segmentation for unknown words, too. Model quality was evaluated similarly to MorphoChallenge⁴⁵, i. e. we calculated boundary precision (P), recall (R), and F1 score using traditional formulas, where true positives are correct boundaries, false positives are incorrectly predicted boundaries and false negatives are missed boundaries.

3.3 Morpheme synthesis

The training data was the same as in the morphological analysis task. The participants were to generate word forms, given lemmata, part-of-speech tags, and other morphological tags. The following measures were calculated:

1. the fraction of word forms which were absolutely correct;
2. average Levenshtein distance between the word forms generated by the participant and the correct word forms. If several are possible, the closest one is used.

4 Evaluation datasets

The following datasets were kindly provided by their creators:

1. Evenki: mainly oral texts recorded in 1998–2016 during fieldwork trips by Olga Kazakevich et al. Has morphological information as well as morpheme segmentation [14];
2. Selkup: oral texts recorded by A. I. Kuzmina in 1962–1977, processed and annotated within the INEL project. Has morphological information as well as morpheme segmentation [3];
3. Veps and Karelian corpus developed within the VepKar project (described in more detail below).

⁴⁵<http://morpho.aalto.fi/events/morphochallenge2005/evaluation.shtml>

It is worth noting that the corpora have been created by linguists and are based on fieldwork data with detailed manual markup. This is unusual for ordinary computational linguistics corpora for major languages, which are usually based on written sources and are therefore more standardized and balanced.

All the corpora had different formats and incompatible markup standards. This would make it hard for the participants to use them. Furthermore, our aim was to allow the participants to combine data from different corpora, using transfer learning or other methods. Therefore, the corpora were converted into the morphological CONLL-U format, used in the Universal Dependencies (UD) treebanks⁴⁶. Moreover, using the standard format makes the resources of the languages in question accessible to researchers all over the world. It makes it possible to include them into the community of "living" and "major" languages (such as Russian and English), which are available to researchers all over the world for processing and building computational models.

On the one hand, the morphological annotation of UD is quite scarce due to the principles of its construction (new tags are only added after the treebanks are added to the project). As a result, we had to exclude many morphological tags. In some linguists' opinion, the resulting narrowed format deprived the language data of essential linguistic information. However, we regarded the narrowing as a necessary trade-off. In addition, the necessity to reformat the corpora made us reanalyze some complex cases and find mistakes in the analysis.

The complex export process is described below in greater detail.

4.1 Export of the Evenki corpus to CONLL format

The Evenki corpus data consisted of EAF⁴⁷ format files. Some texts used in the corpora were originally manually annotated interlinearized texts, lacking lemmata and POS tags. Determining those was the most difficult part of the corpus transformation process, and involved manual work. For instance, the Evenki corpus contained word forms like *oldomotto:wer* ('in order to catch fish'), with the derivative suffix *-mo* ('hunt') attached to the nominal *oldo* ('fish') stem. When preparing the data, we turned this combination of morphemes into a single lemma, namely *oldomo*.

4.2 Export of the Selkup corpus to CONLL format

In contrast with the Evenki corpus, the INEL Selkup corpus contained the necessary data. The difficulty of the transformation process consisted in the mapping between the rich and detailed corpus markup and the CONLL format. It was also troublesome to determine the lemma. Our criterion for lemma determination was to combine the stem and the derivative affixes but not the inflectional ones. Thanks to the help of experts, we could distinguish between the two sets of affixes. For example, we considered some aspectual affixes to be inflectional for Evenki, according to the grammars. In contrast with it, Selkup aspectual morphemes were considered to be derivative. For example, *kurol'na* was segmented as *kurol'(INCH)-na(CO.3SG.S)*. Therefore, the first two morphemes were considered to constitute the lemma *kurol'*.

⁴⁶<https://universaldependencies.org/>

⁴⁷<https://tla.mpi.nl/tla-news/documentation-of-eaf-elan-annotation-format/>

4.3 Export of the VepKar corpus to CONLL format

4.3.1 Languages and dialects of the VepKar corpus

For the shared task, the VepKar developers have presented texts in Veps language and in three main supradialects of Karelian language.

VepKar contains a variety of dialects and subdialects of the Karelian language (see fig. 1). The scheme is based on [13], [7].

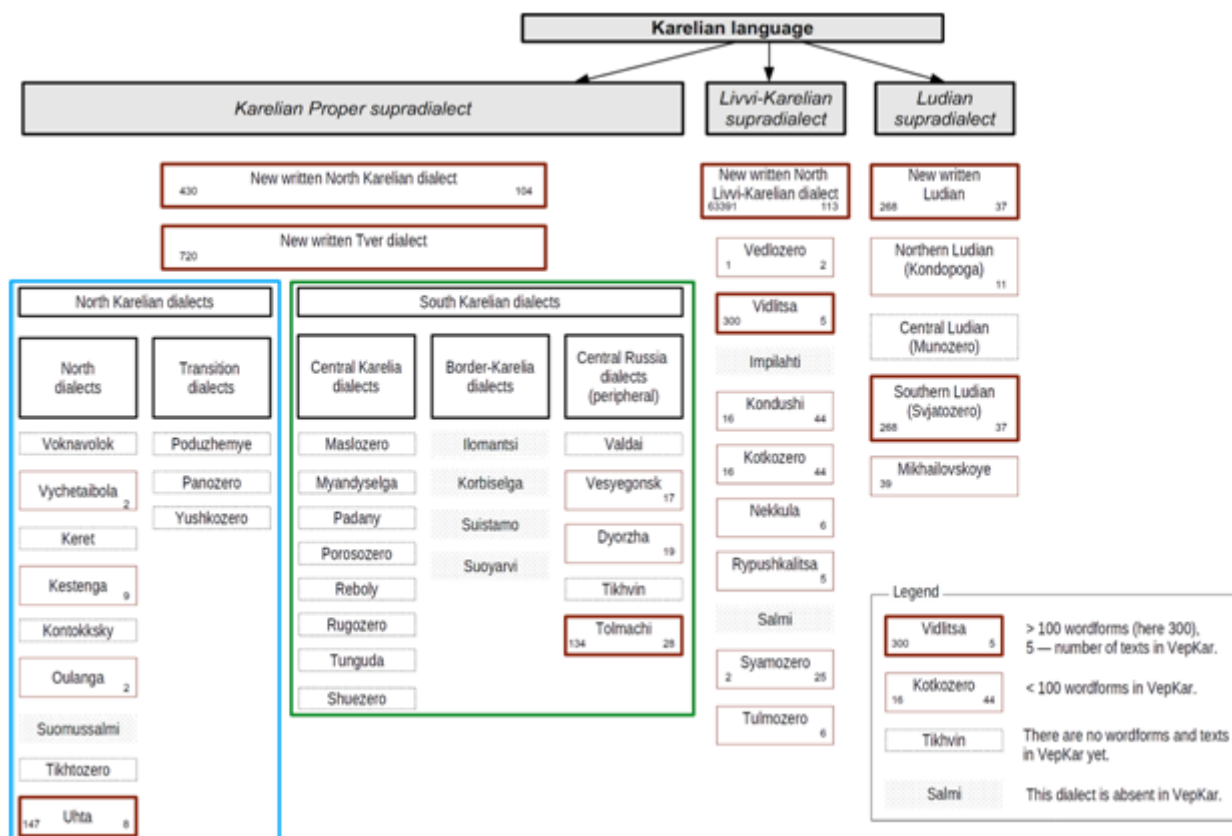


Figure 1: Scheme of dialects of the Karelian language, the number of wordforms (left) and the number of texts (right) in these dialects in the VepKar corpus

There are three written Karelian standard languages. This is due to several reasons. Native Karelian speakers live on a rather vast territory. For several centuries, the language has been influenced by the neighboring Veps, Finnish, and, of course, Russian. The lexical and phonetic systems were the ones most influenced from the outside. This influence gave rise to the three supradialects. Therefore, the corpus uses a separate Karelian dictionary for each supradialect. As of February, 2019 the statistics for the corpus were as following:

1. Olonets Karelian or Livvi (17 thousand lemmata);
2. Ludic Karelian (500 lemmata);
3. Karelian Proper (100 lemmata).

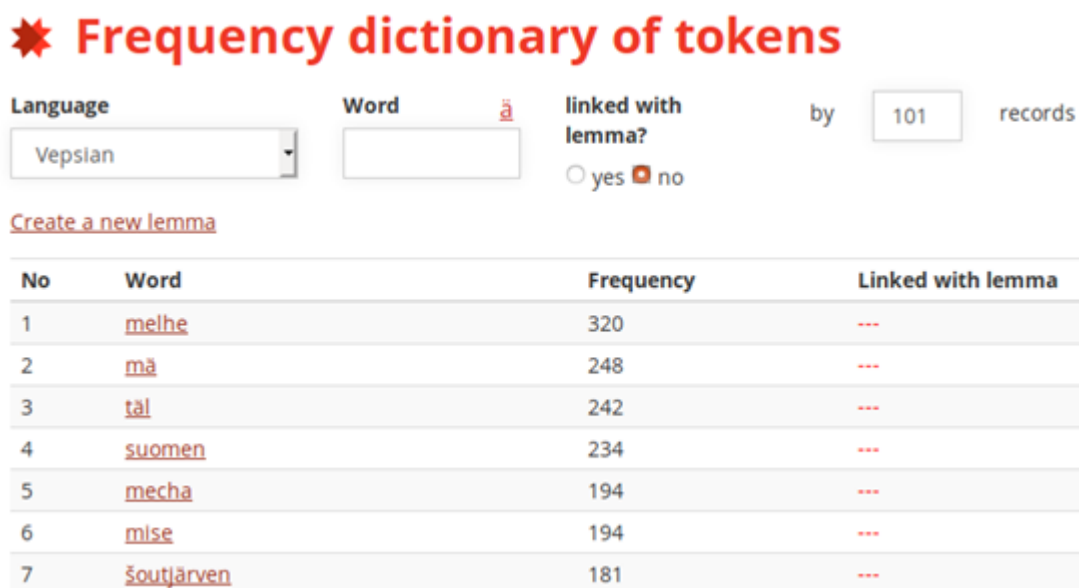
Therefore, three export data sets in CONLL format have been generated, one for each dialect.

4.3.2 The most frequent tokens list

To assist lexicographers in their manual markup process, a generator of the most frequent tokens list (words from the corpus texts) was developed. It is available at the VepKar website⁴⁸. Using the radio button “does this word exist in the dictionary?”, one can get a list of the most frequent tokens that do not have dictionary entries in the VepKar dictionary (see fig. 2).

This list allows the corpus editors to add the most frequent word forms to the corpus dictionary first. Processing primarily the most frequent word forms accelerates the morphological markup of most corpus texts. In the dictionary, two options are possible:

- There is a lemma and there is no word form.
- There is neither word form nor lemma.



★ Frequency dictionary of tokens

Language: Word: linked with lemma? yes no by records

[Create a new lemma](#)

| No | Word | Frequency | Linked with lemma |
|----|----------------------------|-----------|-------------------|
| 1 | melhe | 320 | --- |
| 2 | mä | 248 | --- |
| 3 | täi | 242 | --- |
| 4 | suomen | 234 | --- |
| 5 | mecha | 194 | --- |
| 6 | mise | 194 | --- |
| 7 | šoutjärven | 181 | --- |

Figure 2: The first most frequent tokens of Veps texts in the VepKar corpus which are absent in the VepKar dictionary, with links to usage examples and frequencies in the corpus

In the second case, while creating the lemma, the editor can also add the other word forms. This makes text markup possible if the lemma is found in texts in other grammatical forms.

4.3.3 VepKar development

Participation in this competition was the driving force for the development of the VepKar corpus. To export the data from the VepKar corpus to CONLL, the corpus structure had to be refined significantly. The following features were added to the morphological properties of a lemma:

- for nouns: animacy (“Animacy” in Universal features);
- pluralia tantum (Number=Plur);
- for verbs: transitivity (Subcat);

⁴⁸http://dictorpus.krc.karelia.ru/ru/corpus/word/freq_dict

- for numerals: type of numeral (NumType: quantitative, collective, ordinal, fractional);
- for pronouns: type of pronoun (PronType);
- for adjectives and adverbs: degree of comparison (Degree);
- for adverbs: type of adverb (AdvType).

Initially, VepKar had one part of speech to designate conjunctions. According to the Universal POS tags, the VepKar conjunctions were divided into subordinating and coordinating.

4.3.4 The exporting process

While exporting VepKar data to CONLL format, the following conventions were accepted:

1. For an unknown lemma, write UNKN to the LEMMA column and an underscore to the remaining columns.
2. Write each pair of LEMMA + UPOS on separate lines.
3. Export prepositions (PREP) and postpositions (POSTP) from VepKar corpus to ADP in CONLL. Features PREP or POSTP are indicated in the XPOS field.
4. In a multilingual corpus one file is generated for one language.
5. CONLL-style comments are used for adding sentence identifiers.

4.3.5 Corpus data not included in the CONLL export

In the VepKar corpus there are data that were not exported to CONLL: predicatives (23 lemmas in Olonets Karelian) and phraseological units.

4.4 Dataset statistics

Table 2 summarizes some statistical features of the datasets:

Table 2: Dataset statistics

| Part | Language | Sentences | Words | POSeS | Tags ⁴⁹ | Rare(3) tags ⁵⁰ | Rare(10) tags ⁵¹ | Full tags ⁵² | Rare ⁵³ full tags % |
|-------|----------------------|-----------|---------|-------|--------------------|-------------------------------|--------------------------------|----------------------------|---|
| Train | Evenki | 5 527 | 26 926 | 12 | 55 0 | 2 | 714 | 100 | |
| Test | Evenki | 548 | 2 819 | 12 | 53 0 | 1 | 270 | 98 | |
| Train | Selkup | 2 394 | 13 436 | 12 | 34 2 | 3 | 218 | 98 | |
| Test | Selkup | 425 | 2 426 | 12 | 30 1 | 1 | 109 | 95 | |
| Train | Veps | 38 793 | 357 811 | 13 | 47 0 | 4 | 147 | 99 | |
| Test | Veps | 2 163 | 19 376 | 13 | 42 0 | 1 | 86 | 100 | |
| Train | Karelian (proper) | 7 048 | 68 296 | 11 | 34 4 | 6 | 66 | 100 | |

⁴⁹Pairs of feature=value

⁵⁰occurring less than 3 times in the train set

⁵¹occurring less than 10 times in the train set

⁵²combinations of tags

⁵³less than 10% of all words

| | | | | | | | | | |
|-------|----------------------|-------|--------|----|----|---|---|----|-----|
| Test | Karelian (proper) | 919 | 8 640 | 9 | 30 | 1 | 3 | 44 | 100 |
| Train | Karelian (Ludic) | 1 711 | 15 805 | 12 | 27 | 5 | 5 | 29 | 97 |
| Test | Karelian (Ludic) | 204 | 1 968 | 11 | 22 | 0 | 0 | 19 | 95 |
| Train | Karelian (Livvi) | 6 213 | 57 093 | 13 | 23 | 4 | 6 | 26 | 88 |
| Test | Karelian (Livvi) | 745 | 7 206 | 13 | 17 | 0 | 1 | 19 | 84 |

91.2 One can see from this table that the concept of "full tags", i.e. sets of morphological features, does not seem to work well for agglutinating languages due to the huge number of combinations.

5 Participants and results

5.1 System description.

In the morphological analysis track, three teams took part. The morpheme segmentation and word form generation tasks were less popular with only one team participating in each of them.

MSU-DeepPavlov team [9] utilized recurrent neural networks on word level. The embeddings of words were obtained using convolutional networks with highway layer on the top, closely following [5] and [8]. This team demonstrated the highest scores on Evenki and Selkup datasets on all metrics and on Veps dataset on part-of-speech prediction⁵⁴.

The second team, **drovoseq** used BertBiLSTMattnNMT encoder-decoder architecture to decode the optimal sequence of morphological tags. The third one, **SPBUMorph** used a variant of Markov models to evaluate the probability of a tag given the word.

For lemmatization, the winning **MSU-DeepPavlov** team used a neural network to predict the pattern of the transformation between the surface word and its initial form, while **drovoseq** used encoder-decoder architecture.

The only submitted morpheme segmentor used the model similar to [10], which reduced the task of morpheme segmentation to sequence labeling.

5.2 Results

The results are shown in Tables 3, 4 and 5:

Table 3: Morphological analysis: results

| Team | # | Language | % of cor- rect lem- mata for wf | % of cor- rect lem- mata for sen- tences | % of cor- rect POS'es for wf | % of cor- rect POS'es for sen- tences | feature pre- ci- sion | feature recall | feature F2 |
|------|---|----------|---|---|---|---|--------------------------------|-------------------|---------------|
|------|---|----------|---|---|---|---|--------------------------------|-------------------|---------------|

⁵⁴It did not participate on other subtasks.

| | | | | | | | | | |
|--------------------|---|----------------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| drovoseq | 1 | Evenki | 0,8617 | 0,7550 | 0,8811 | 0,8086 | 0,8112 | 0,7993 | 0,8052 |
| drovoseq | 1 | Karelian (proper) | 0,9971 | 0,9869 | 0,9909 | 0,9603 | 0,9539 | 0,9373 | 0,9455 |
| drovoseq | 1 | Karelian (Ludic) | 0,9959 | 0,9828 | 0,9726 | 0,8897 | 0,9356 | 0,8769 | 0,9053 |
| drovoseq | 1 | Karelian (Livvi) | 0,9629 | 0,8631 | 0,8969 | 0,7168 | 0,8471 | 0,7985 | 0,8221 |
| drovoseq | 1 | Selkup | 0,8780 | 0,7647 | 0,8343 | 0,7529 | 0,8014 | 0,7713 | 0,7861 |
| drovoseq | 1 | Veps | 0,9761 | 0,9087 | 0,9572 | 0,8534 | 0,6691 | 0,4666 | 0,5498 |
| drovoseq | 2 | Evenki | 0,8710 | 0,7620 | 0,9075 | 0,8201 | 0,8156 | 0,8217 | 0,8187 |
| drovoseq | 2 | Karelian (proper) | 0,9977 | 0,9889 | 0,9992 | 0,9956 | 0,4045 | 0,1776 | 0,2468 |
| drovoseq | 2 | Karelian (Ludic) | 0,9970 | 0,9851 | 0,9959 | 0,9777 | 0,5962 | 0,3570 | 0,4466 |
| drovoseq | 2 | Karelian (Livvi) | 0,9797 | 0,9108 | 0,9781 | 0,9074 | 0,6751 | 0,4489 | 0,5392 |
| drovoseq | 2 | Selkup | 0,8941 | 0,7759 | 0,8586 | 0,7354 | 0,8029 | 0,8026 | 0,8028 |
| drovoseq | 2 | Veps | 0,9875 | 0,9480 | 0,9938 | 0,9753 | 0,4873 | 0,2678 | 0,3457 |
| MSU- DeepPavlov | 1 | Evenki | 0,8838 | 0,7914 | 0,9122 | 0,8421 | 0,8805 | 0,8809 | 0,8807 |
| MSU- DeepPavlov | 1 | Selkup | 0,9031 | 0,8035 | 0,8957 | 0,7965 | 0,9095 | 0,9082 | 0,9089 |
| MSU- DeepPavlov | 1 | Veps | 0,3003 | 0,5146 | 0,9943 | 0,9769 | 0,5471 | 0,8073 | 0,6522 |
| SPBUMorph | 1 | Evenki | 0,7125 | 0,2857 | 0,7222 | 0,3099 | 0,1503 | 0,3692 | 0,2137 |
| SPBUMorph | 1 | Karelian (proper) | 0,9992 | 0,9913 | 0,9994 | 0,9935 | 0,7028 | 0,9172 | 0,7958 |
| SPBUMorph | 1 | Karelian (Ludic) | 0,9975 | 0,9706 | 0,9959 | 0,9608 | 0,5692 | 0,8674 | 0,6873 |
| SPBUMorph | 1 | Karelian (Livvi) | 0,9653 | 0,7369 | 0,9460 | 0,6148 | 0,4742 | 0,7064 | 0,5674 |
| SPBUMorph | 1 | Selkup | 0,6834 | 0,2000 | 0,6818 | 0,2447 | 0,1400 | 0,3147 | 0,1938 |
| SPBUMorph | 1 | Veps | 0,9839 | 0,8798 | 0,9899 | 0,9177 | 0,5471 | 0,8073 | 0,6522 |

Table 4: Morpheme segmentation: results

| Team | # | Language | Precision | Recall | F1 | % of totally correct wordforms |
|------------|---|----------|-----------|--------|--------|--------------------------------|
| deeppavlov | 1 | Evenki | 0,9774 | 0,9783 | 0,9779 | 0,9317 |
| deeppavlov | 1 | Selkup | 0,9538 | 0,9551 | 0,9544 | 0,8640 |

Table 5: Word form generation: results

| Team | # | Language | Totally correct | Averaged Levenshtein distance |
|----------|---|----------|-----------------|-------------------------------|
| SAG_TEAM | 1 | Evenki | 0,5325 | 1,2585 |
| SAG_TEAM | 1 | Selkup | 0,5076 | 1,1621 |

6 Discussion

For the time being, the languages under consideration do not have robust rule-based parsers, therefore the only source of comparison is the annotation of the test set. We also notice that participants reported errors and discrepancies in the annotation during training phase. Although we fixed most of them after the discussion, this could potentially influence the systems' efficiency.

First, we would like to note that the topmost system achieved significantly high scores on tasks of morphological analysis, lemmatization and morpheme segmentation, which is comparable to scores of state-of-the-art systems on other datasets of similar size.

6.1 Morphological analysis results

It is interesting that the systems made similar mistakes. It certainly has to do with the limitations of the data itself, its relatively low amount and scarcity. However, the percentage of errors differs significantly between the systems, which implies that different models require different amounts of labeled data to be trained on.

6.1.1 Lemmatization

Lemmatization errors can be grouped as following:

1. Rare lemmata: e. g., the Evenki *jaja* 'to chant shamanic songs' can only be found in few texts.
2. "Non-standard" lemmata: the oral texts in a minor language naturally contain a lot of loanwords. These loanwords, especially recent ones, are often different phonetically from the basic words. They seem to present troubles for all systems. E. g., the Evenki *kirest* 'cross' < Russian *krest* has *st* consonant cluster, which is not typical for an Evenki word. Another example is *penšianerka* 'pensionnaire (woman)' < Russian *penšianerka*. This word with its initial "p" sound is not typical for the language. Furthermore, its ending corresponds with the *-rkV* suffix. Not surprisingly, most systems judged *-rka* to be a suffix in this word. Similarly, the systems split the Selkup word *poshalusta* 'please' < Russian *pozhalusta*, separating the ending *sta*. It would be interesting to check if the results could improve if the systems accounted for Russian loanwords.
3. Short lemmata. The systems seemed to prefer long roots over short ones. As a result, word forms with one-letter roots are processed incorrectly. For example, *e* 'negative verb' or *i*: 'enter' presented a trouble for the systems. On the other hand, in some cases, the lemmata were standard and quite wide-spread. However, their ending in a letter which itself constituted a wide-spread suffix caused the systems to incorrectly split the lemma. In the Evenki data, this is true for *l*, *n* or *t*.
4. Morphophonological phenomena were difficult to follow for the systems. E. g., in *uguchak-ker* 'reindeer-RFL.PL' the *-ker* part is a surface realization of the *-wer* morpheme after *k*. Similar *kw* -> *kk* alternations can be found in the training data. However, the systems could not grasp this alternation.

6.1.2 Determining POS

As regards the POS determination, the errors show that the systems could not reliably distinguish between nominal and verbal categories. One could expect the systems to confuse

nouns and *adjectives* but not *nouns* and *verbs*. However, it is the *VERB* category which was most often confused with the *NOUN* category. This behavior contradicts the naive linguistic assumptions. However, it can be justified by the fact that in agglutinating languages, verbs and nouns often have similar sets of affixes (e. g., possessive suffixes of nouns versus verbal personal suffixes).

Interestingly, both on the Evenki and Selkup dataset the quality of POS detection was comparable or even lower than the quality of morphological features detection. It contrasts the traditional ratio between the quality of recovery for POS tagging and morphological features: usually it is much easier to recover correct parts of speech than to restore all features. For example, during CoNLL2018 evaluation campaign [11], best average POS accuracy was **90.9%**, while the accuracy of features was only **87.59%**. Naturally, for Russian it is much easier to detect whether a word is a noun or a verb than to discriminate between, for example, accusative and nominative cases. Probably, this unusual performance can be explained by the abundancy of informal speech in the dataset, which is relatively “unconnected” in comparison to more formal sources of most UD treebanks. This implies that basic contextual clues (such as word order) prove too weak to predict part-of-speech labels. The corpora contain phrases with slips, repetitions, discourse markers, e. g.: *Wot amakalwi Ekondaduk bal= ekun kergentin, kergentin Ekondaduk* (So my grandfathers from Ekonda SLIP well, their family, their family from Ekonda)

6.2 Morpheme segmentation

The primary causes of the morpheme segmentation errors were the following:

1. Non-standard and borrowed lemmata: as with the morphological analysis task, loanwords cause problems, with the systems splitting them incorrectly. On the other hand, loanwords with native suffixes such as *telogrejka-t* ‘coat(<Russian>-INSTR)’
2. Suffix combinations versus complex suffixes: interestingly, the **MSU-DeepPavlov** system sometimes splits a complex suffix into parts, e. g. *d’eli* versus *d’e-li*. Actually, the etymology of the suffix supports the claim that historically it could have been made of these basic parts. However, in the synchronous view, we cannot split the suffix.

6.3 Word form generation

In the word form generation task, most errors were due to the vowel harmony and consonant alternation phenomena. Vowel harmony means that there are different forms of the same affix depending on the vowels in the stem. E. g., *d’aja-* requires *a* in some affixes. However, the system suggests *e*, which is not correct. It is worth noting that these phenomena are hard to grasp even in detailed grammatical descriptions. There is much variation in dialectal data. Sometimes the training data and gold standard data contradict the “normal” rules, so the results are not surprising.

However, some errors cannot be justified by the data complexity as the resulting letter clusters are highly improbable and cannot be found in the training data.

7 Conclusion

In this paper, we present the results of the First Shared Task on morphology for low-resource languages. As a result of the shared task, several datasets in the CONLL format were prepared, for the first time for the languages in question. The participating teams created new morphological analysis tools for the languages which lack modern NLP technology tools.

The comparison of results showed the vitality of modern neural approach when applied to low-resource datasets collected by field linguists. We also explored the limitations of the systems, which can help improve them.

8 Acknowledgements

The work of Elena Klyachko was partially supported by a grant of the Russian Science Foundation, Project 17-18-01649.

The work of Natalia Krizhanovskaya and Andrew Krizhanovsky was supported by a grant of the Russian Foundation for Basic Research, Project 18-012-00117.

We are grateful to Svetlana Toldova, Artyom Sorokin and Karina Mishchenkova for their advice, and their help with evaluation scripts and datasets.

References

- [1] Russian census 2010. http://www.gks.ru/free_doc/new_site/perepis2010/croc/perepis/_itogi1612.htm.
- [2] Timofey Arkhangel'skiy and Maria Medvedeva. Developing morphologically annotated corpora for minority languages of russia. In *CLiF*, 2016.
- [3] Maria Brykina, Svetlana Orlova, and Beáta Wagner-Nagy. INEL selkup corpus. Version 0.1. In Beáta Wagner-Nagy, Alexandre Arkhipov, Anne Ferger, Daniel Jettka, and Timm Lehmborg, editors, *The INEL corpora of indigenous Northern Eurasian languages*, volume 1. HZSK Hamburg, 2018.
- [4] Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D McCarthy, Katharina Kann, Sebastian Mielke, Garrett Nicolai, Miikka Silfverberg, et al. The conll-sigmorphon 2018 shared task: Universal morphological reinflection. *arXiv preprint arXiv:1810.07125*, 2018.
- [5] Georg Heigold, Guenter Neumann, and Josef van Genabith. An extensive empirical evaluation of character-based morphological tagging for 14 languages. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 505–513, 2017.
- [6] Katharina Kann, Manuel Mager, Ivan Meza-Ruiz, and Hinrich Schütze. Fortification of neural morphological segmentation models for polysynthetic minimal-resource languages. *arXiv preprint arXiv:1804.06024*, 2018.
- [7] I. Novak, M. Penttonen, A. Ruuskanen, and L. Siilin. Karel'skiy yazyk v grammatikakh. sravnitel'noe issledovanie foneticheskoy i morfologicheskoy sistem. page 22, 2019.
- [8] Alexey Sorokin. Improving neural morphological tagging using language models. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue"*, 2018.
- [9] Alexey Sorokin. Morphological parsing of low-resource languages. In *Dialogue: International Conference on Computational Linguistics*, 2019. To appear.

- [10] Alexey Sorokin and Anastasia Kravtsova. Deep convolutional networks for supervised morpheme segmentation of russian language. In *Conference on Artificial Intelligence and Natural Language*, pages 3–10. Springer, 2018.
- [11] Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. Conll 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, 2018.
- [12] Daniel Zeman, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, et al. Conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19. Association for Computational Linguistics, 2017.
- [13] П. М. Зайков. *Глагол в карельском языке*. Изд-во ПетрГУ, 2000.
- [14] О. А. Казакевич and Е. Л. Клячко. Создание мультимедийного аннотированного корпуса текстов как исследовательская процедура. In *Труды Международной конференции «Корпусная лингвистика – 2013»*, ISBN 978-5-8465-1335, pages 292–300. Изд-во СПбГУ Санкт-Петербург, 2013.