Corpora of social media in minority Uralic languages

Timofey Arkhangelskiy
Universität Hamburg / Alexander von Humboldt Foundation
timarkh@gmail.com

Abstract

This paper presents an ongoing project aimed at creation of corpora for minority Uralic languages that contain texts posted on social media. Corpora for Udmurt and Erzya are fully functional; Moksha and Komi-Zyrian are expected to become available in late 2018; Komi-Permyak and Meadow and Hill Mari will be ready in 2019. The paper has a twofold focus. First, I describe the pipeline used to develop the corpora. Second, I explore the linguistic properties of the corpora and how they could be used in certain types of linguistic research. Apart from being generally "noisier" than edited texts in any language (e.g. in terms of higher number of out-of-vocabulary items), social media texts in these languages present additional challenges compared to similar corpora of major languages. One of them is language identification, which is impeded by frequent code switching and borrowing instances. Another is identification of sources, which cannot be performed by entirely automatic crawling. Both problems require some degree of manual intervention. Nevertheless, the resulting corpora are worth the effort. First, the language of the texts is close to the spoken register. This contrasts to most newspapers and fiction, which tend to use partially artificial standardized varieties. Second, a lot of dialectal variation is observed in these corpora, which makes them suitable for dialectological research. Finally, the social media corpora are comparable in size to the collections of other texts available in the digital form for these languages. This makes them a valuable addition to the existing resources for these languages.

Аннотация

Статьяын урал кылъёсын социальной сетьёсысь материалъя корпус лэсьтон сярысь вераськон мынэ. Удмурт но эрзя кылъёсын корпусъёс дасесь ини; мокша но коми-зырян кылъёсын корпусъёс 2018-тй арлэн пумаз ужаны кутскозы; нош коми-пермяк но мари корпусъёс 2019-тй арын дась луозы. Та статьяын кык ужпум жутэмын. Нырысь ик, мон валэктйсько, кызьы корпус лэсьтон уж радъямын. Собере корпусъёсляэсь кылтйрлык аспёртэмлыксэс эскерисько но возьматйсько, кызьы корпусэз пёртэм пумо тодосужын луэ уже кутыны. Социальной сетьёсысь текстъёс котькуд кылын «пожгес» луо, литературной текстъёсын ёошатыса (шуом, морфологической разбортэк кылем кылъёсты лыдъяно ке). Пичи кылъёсын корпусъёсты трос поллы секытгес лэсьтыны, бадзым кылъёсын ёошатыса. Нырысь ик, шуг валаны, кыче кылын гожтэмын текст, малы ке шуоно, текстлэн кылыз чем вошъяське но текстын трос асэстэм кыл кутйське. Мукетыз шуг-секыт —

текстъёсты уг луы автоматической кроулинг амалэн шедьтыны. Та шуг-секытъёсын йырин трос ужез «киын» лэсьтоно луэ. Озьы ке но, та ужлэн пайдаез вань. Нырысь ик, корпусысь кыл вераськон кыллы матын луэ. Озьы со газетын но литератураын кутӥськись кыллэсь пӧртэм луэ; отын литературной кыл ӵемысь искусственной кыллы укша. Кыкетйез, корпусын пӧртэм диалектъёс пумисько; соин ик та материалэз диалектъёсты эскерон ужын кутыны луэ. Йылпумъяса вераны кулэ, куд-ог урал кылъёсын электронной текст люкамъёс вань ини; социальной сетьёсысь материалэн корпусъёс мукет текстъёсын корпусъёслэсь ӧжыт пичигес ке но, быдӟалазыя трослы пӧртэм ӧвӧл. Озьыен, таӵе корпусъёс ӧжыт лыдъем калыкъёслэсь кылтодоссэс узырмыто.

Аннотация

В статье представлен текущий проект по созданию корпусов соцсетей на малых уральских языках. В настоящий момент готовы корпуса удмуртского и эрзянского языков; мокшанский и коми-зырянский планируется запустить в конце 2018 г., а коми-пермяцкий и марийские — в 2019 г. В работе освещены две темы. Во-первых, я описываю процедуру разработки корпусов. Во-вторых, я рассматриваю лингвистические свойства этих корпусов и то, как их можно использовать в разных видах исследований. Тексты соцсетей на любом языке в принципе более «грязные», чем стандартные (например, в смысле количества слов без морфологического разбора), однако тексты на рассматриваемых языках представляют дополнительные сложности по сравнению с аналогичными текстами на крупных языках. Одна из них — это определение языка текста, которое затрудняется многочисленными переключениями кодов и заимствованиями. Другая — это поиск таких текстов, который невозможно произвести с помощью полностью автоматического кроулинга. Обе проблемы требуют некоторого количества ручной работы. Тем не менее, полученные результаты стоят приложенных усилий. Во-первых, язык в этих корпусах близок к разговорному регистру. Этим он отличается от языка газет и литературы, где часто используется до некоторой степени искусственный стандартный вариант. Во-вторых, в корпусах наблюдается диалектная вариативность, что делает их пригодными для диалектологических исследований. Наконец, по размеру корпуса соцсетей сопоставимы с коллекциями других текстов, существующих для соответствующих языков в электронном виде. Это делает их ценным дополнением к существующим языковым ресурсам.

1 Introduction

There are seven minority Uralic languages in the Volga-Kama area and adjacent regions of Russia[1]: Komi (Zyrian, Permyak), Udmurt, Mari (Meadow, Hill), Erzya and Moksha. All these languages fall in the middle of the Uralic spectrum in terms of the number of speakers. Similarly, they all belong to the middle level of digital vitality: based on the amount of digital resources available for them, Kornai (2016) calls them digitally "borderline" languages. Their sociolinguistic situation is also rather similar; see Blokland and Hasselblatt (2003) for an overview. All of them have had intensive contact with the dominant Russian language; almost all their speakers are bilingual in Russian; the number of speakers is on the decline. Despite the fact that all of these

[1]Seven literary standards, to be more precise.

languages have some official status in the respective regions, their use in the public sphere and education is very limited.

Social media have been a target for both NLP and linguistic research for a long time now. However, the overwhelming majority of papers deal with social media texts in English or one of several other major languages. Smaller languages are severely underrepresented in this domain. There are corpora of social media texts in large Uralic languages, e.g. the Suomi24 forum corpus for Finnish (Aller Media Oy, 2014), and investigations based on them, e.g. Venekoski et al. (2016). All minority Uralic languages spoken in Russia lack such corpora.

Collecting social media corpora for the seven languages listed above is the central part of my ongoing project. There are notable differences between social media in these languages and those in major languages, which pose certain challenges for corpus development. First, they are smaller in size by several orders of magnitude. While, for example, the Edinburgh Twitter Corpus contains 2.26 billion tokens of tweets in English collected within a 2.5-month span (Petrović et al., 2010), all corpora I am dealing with do not exceed 3 million tokens despite representing an 11-year period. This scarcity of data makes every single post valuable. Another difference is ubiquitous code switching instances and Russian borrowings, which makes reliable language tagging a necessity. Yet another challenge comes from the fact that many social media users are not well acquainted with, or consciously avoid, the literary norm. On the one hand, this means that dialectal variation can be studied in Uralic social media corpora, but on the other, it makes morphological annotation more difficult.

The paper is organized as follows. In Section 2, I describe how I find, harvest and process the social media texts. In Section 3, I consider the linguistic and sociolinguistic properties of collected texts and discuss how that could be beneficial for certain kinds of research. In Section 4, I briefly describe the web interface through which the corpora are available.

2 Processing the data

2.1 Identifying and harvesting texts

A common approach to harvesting various kinds of texts from the web is to apply some kind of automatic crawling, which takes a small set of URLs as a seed and then follows the hyperlinks to find more content. Unfortunately, it is almost impossible to use this approach without adjustments for languages with small digital presence. Most links that appear in pages written in such languages lead to texts written in the dominant language (Russian in this case), and sifting through all of them to find relevant pages or fragments would require too much computational power.

In order to make text harvesting more efficient and less time-consuming, I try to make the seed as close to the comprehensive URL list as possible. Only after processing all pages from that list do I apply limited crawling. When identifying the pages for the seed list, I build upon a strategy proposed and used by Orekhov et al. (2016) for collecting and researching minority languages of Russia on the Internet, as well as on the results obtained by them. A slightly different version of the same strategy was previously used by Scannell (2007) in the Crúbadán project for similar purposes. This approach involves searching for relevant pages with a conventional search engine, using a manually compiled small set of tokens which are frequent in the relevant language, but do not exist or are very infrequent in any other language. This contrasts

to the strategy employed by the "Finno-Ugric Languages and the Internet" project (Jauhiainen et al., 2015), which relied on large-scale crawling and subsequent fully automatic filtering by language.

Out of a dozen social media services with presence in Russia, I currently limit my search to *vkontakte*[2], which is by far the most popular of them both in relevant regions and in Russia as a whole. My preliminary research shows that in major Western social media, such as Facebook or Twitter, texts in minority Uralic languages are almost non-existent. However, there is at least one other Russian resource, *odnoklassniki*[3], which seems to contain texts in these languages in quantities that may justify the effort needed to process them. *Odnoklassniki* is more popular with the older generation and apparently has varying popularity across regions. For example, it seems that there are more texts in Erzya there than in Udmurt. Nevertheless, relevant texts in *vkontakte* clearly outnumber those in *odnoklassniki*. Additionally, I download forums not associated with any social media service, if their primary language is one of those I am interested in. So far, I have only found forums of such kind for Erzya.

Although there are also blogs available in these languages, I did not include them in the social media corpora. Baldwin et al. (2013) show that the language of blogs could be placed somewhere between edited formal texts and social media by a number of parameters. This is true for most (although not all) blogs in minority Uralic languages, which on average contain less code-switching than social media and where the language variety seems closer to the literary standard. Nevertheless, blogs are undoubtedly a valuable source for linguistic research, which is why I downloaded them as well and included them in the "support corpora" (see below).

As a starting point, I take the URL lists of *vkontakte* pages collected by Orekhov et al. (2016).[4] I manually check all of them and remove those that were misattributed (which sometimes happens because the lists were compiled in an unsupervised fashion). An example of an erroneously marked page is a Russian group dedicated to Korean pop music where the users share the lyrics in Cyrillic transcription. Apparently, a transcribed Korean word coincided with one of frequent Udmurt tokens, which is why it ended up tagged as Udmurt.

As a second step, I perform manual search in Yandex search engine with an additional check in Google, using the same strategy as Orekhov et al. (2016). This allows me to enhance the original lists with URLs that were missed or did not exist in 2015, when the lists were compiled.

When the initial list of URLs is ready, I download the texts (posts and comments) and the metadata using the *vkontakte* API. The amount of data is small enough for it to be downloadable through a free public API with a limitation of 3 queries per second within several days. The texts with some of the metadata are stored in simple JSON files. User metadata is cached and stored in another JSON file to avoid the need of downloading it multiple times for the same user. Obviously, only texts and metadata open to the general public can be downloaded this way.

The final stage of the harvesting process involves limited crawling. The messages written by individual users are automatically language-tagged. For each user, I count the number of messages in the relevant language authored by them. All users that have at least 2 messages are added to the URL list and their "walls" (personal pages with texts and comments written by them or addressed to them) are downloaded as

[2] `https://vk.com/`

[3] `https://ok.ru/`

[4] The lists are available at `http://web-corpora.net/wsgi3/minorlangs/download`

well. The threshold of 2 messages was chosen to cut off instances of erroneous language tagging, which happen especially often with short messages. Besides, users with small message counts tend to have no texts in the relevant languages on their walls anyway.

2.2 Language tagging

The social media texts in minority Uralic languages are interspersed with Russian, so language tagging is of crucial importance to the project. There are standard techniques for language tagging, the most popular probably being the one based on character n-gram frequencies (Canvar and Trenkle, 1994). It is impossible, however, to achieve sufficient quality on minority Uralic social media data with these methods. The first problem is that the texts that have to be classified are too short. Mixing languages within one message is extremely common, which is why at least sentence-level tagging is needed in this case. In an overview of several n-gram-based methods, Vinosh Babu and Baskaran (2005) note that, although generally it is easy to achieve 95% or higher precision with such methods, "for most of the wrongly identified cases the size of the test data was less than 500 bytes, which is too small". This is always the case with the sentences, which most of the time contain less than 10 words. What's more, sentences in the relevant languages contain lots of Russian borrowings and place names, which would shift their n-gram-based counts closer to those of Russian. Classifying short segments with additional issues like that is still problematic with the methods commonly used at present (Jauhiainen et al., 2018, 60–61).

Instead of a character-based classification, I use a process which is mostly dictionary-based and deals with words rather than character n-grams as basic counting units. In a nutshell, it involves tokenization of the sentence, dictionary lookup for each word and tagging the sentence with the language most words can be attributed to. The classification is three-way: each sentence is tagged as either Uralic, or Russian, or "unknown". The last category is inevitable, although the corresponding bin is much smaller than the first two. It contains sentences written in another language (English, Tatar, Finnish and Hungarian are among the most common), sentences that comprise only emoji, links and/or hashtags, and those that are too difficult to classify due to intrasentential code switching. In the paragraphs below, I describe the algorithm in greater detail.

Before processing, certain frequent named entities, such as names of local newspapers and organizations, are cut out with a manually prepared regex. This is important because such names, despite being written in a Uralic language, often appear in Russian sentences unchanged. After that, the sentence is split into tokens by whitespaces and punctuation-detecting regular expressions. Only word tokens without any non-Cyrillic characters or digits were considered.

There are three counters: number of unambiguously Russian tokens (`cntR`), number of unambiguously Uralic tokens (`cntU`), and number of tokens that could belong to either language (`cntBoth`). Each word is compared to the Russian and Uralic frequency lists, which were compiled earlier. If it only appears on one of them without any remarks, the corresponding counter is incremented. If it appears only in the Uralic list, but is tagged as either a Russian borrowing or a place name without any inflectional morphology, `cntBoth` is incremented. The same happens if the word is on both lists, unless it is much more frequent, or its 6-character suffix is more common (in terms of type frequency), in one than in the other. (Exact thresholds here and in the paragraph below are adjusted manually and are slightly different for different

languages of the sample.) In the latter case, the corresponding counter, `cntR` or `cntU`, is incremented.

After all words have been processed, rule-based classification is performed. If one of the counters is greater than the others and most tokens in the sentence have been attributed to one of the languages, the sentence is tagged according to the winning counter. If there are many ambivalent words and either no Uralic words or some clearly Russian words in the sentence, it is classified as Russian. Finally, if counter-based rules fail, the sentence is checked against manually prepared regexes that look for certain specific character n-grams characteristic for one language and rare in the other. If this test also does not produce a definitive answer, the sentence is classified as "unknown".

There is a certain kind of texts in social media in minority languages that poses a serious challenge to this approach. In all languages I have worked with, there are groups designed for learning the language. They often contain lists of sentences or individual words with Russian translations. A simplistic approach to sentence segmentation places most of such translation pairs inside one sentence, which is then impossible to classify as belonging to one of the languages. To alleviate this problem, the language classifier tries splitting sentences by hyphens, slashes or other sequences commonly used to separate the original from the translation. If both parts can be classified with greater certainty than the entire fragment, and they have different language tags, the sentence remains split.

During the initial language tagging, "borderline" sentences, i.e. those whose `cntR` and `cntU` counters had close values, were written to a separate file. I manually checked some of them and corrected the classification if it was wrong. During second run of tagging, each sentence was first compared to this list of pre-tagged sentences. The tagging procedure described above was only applied to sentences that were not on that list. Finally, an extended context was taken into account. If a sentence classified as "unknown" was surrounded by at least 3 sentences with the same language tag (at least one before and at least one after it), its class was switched to that of the neighboring sentences.

The resulting accuracy is high enough for practical purposes and definitely higher than an n-gram-based approach would achieve. Tables 1 and 2 show the figures for Udmurt and Erzya. The evaluation is based on a random sample that contained 200 sentences for each of the languages. Actual cases of misclassification comprise only about 2% of sentences classified as Uralic. An additional 3% accounts for problematic cases, e.g. code switching with no clear main/matrix language. The share of sentences classified as "unknown" is 2.5% for Udmurt/Russian pair and 1.3% for Erzya/Russian; most of them are indeed not classifiable. Note that the figures below refer to sentences rather than tokens. Given that wrong classification overwhelmingly occurs in short sentences (1–4 words), precision measured in tokens would be much higher.

	correct sentences	**wrong language**	**mix / other**
Udmurt	95.5%	1.5%	3%
Russian	100%	0%	0%

Table 1: Accuracy of language tagging for Udmurt.

The described approach requires much more training data and annotation than the n-gram-based classification. Specifically, it relies on word lists for the respective lan-

	correct sentences	wrong language	mix / other
Erzya	94.5%	2.5%	3%
Russian	97%	1%	2%

Table 2: Accuracy of language tagging for Erzya.

guages that are long enough, contain some morphological annotation, annotation for Russian loanwords and place names, and frequency information. Such lists are readily available for Russian; I used a frequency list that is based on the Russian National Corpus and contains about 1 million types. However, it is much more problematic to obtain such lists for the Uralic languages. In order to do so, I had to collect a "support corpus" with clean texts and no Russian insertions for each of the languages first. Fortunately, this is achievable because there are enough non-social-media digital texts in them on the web. First and foremost, for each language there are one or several newspapers that publish articles in it. Apart from that, there are translations of the Bible, blogs (surprisingly, unlike social media, most of them do not contain chaotic code switching) and fiction. By contrast, Wikipedia, which is often a primary source of training data for major languages, is of little use for this purpose because Wikipedias in these languages mostly contain low-quality and/or automatically generated articles (Orekhov and Reshetnikov, 2014). The resulting lists contain around 230,000 types for Udmurt and around 100,000 types for Erzya, Moksha and Komi-Zyrian.

Although I am primarily interested in the Uralic data, all Russian and unclassified sentences are also included in the corpus. Omitting them in mixed posts would obviously be detrimental for research because it would be impossible to restore the context of Uralic sentences and therefore, in many cases, fully understand their meaning. However posts written entirely in Russian are also not removed if their authors or the groups where they appear have Uralic posts as well. This effectively makes my corpora bilingual, although not in a sense traditionally associated with this term (Barrière, 2016). One reason why this is done is facilitating sociolinguistic investigations of language choice in communication. Another is enabling research of contact-induced phenomena in Russian spoken by native speakers of the Uralic languages. A number of corpus-based papers has been published recently about regional contact- or substrate-influenced varieties of Russian, e.g. by Daniel et al. (2010) about Daghestan or Stoynova (2018) about Siberia and Russian Far East. The availability of corpora that contain Russian produced by Uralic speakers could lead to similar research being carried out on Uralic material.

2.3 Filtering and anonymization

After the language tagging, the texts undergo filtering, which includes spam removal, deduplication and anonymization.

Since the actual content is not that important for linguistic research, there is nothing inherently wrong with having spam sentences in the corpus, as long as they are written in a relevant language. However, the main problem with spam is that it is repetitive, which biases the statistics. In order to limit this effect, I manually checked sentences that appeared more than N times in the corpus (with N varying from 2 to 5, depending on the size of the corpus). Those that could be classified as being part of automatically generated messages or messages intended for large-scale mul-

tiple posting, were put to the list of spam sentences. If they contained variable parts, such as usernames, those were replaced with regex equivalents of a wildcard. Such variable parts make template sentences resistant to ordinary duplicate search, which justifies treating them separately. Most of such sentences come from online games, digital postcards or chain letters. The resulting list contains about 800 sentences and sentence templates. Sentences in texts that match one of the templates are replaced with a <SPAM> placeholder. Posts where more than half of sentences were marked as spam are removed.

Text duplication is a serious problem for social media texts, which are designed for easily sharing and propagating messages. Posts published through the "share" button are marked as copies in the JSON returned by *vkontakte* API. If multiple copies of the same post appear in different files, they are identified by their post ID. Only one copy is left in place, and all others are replaced by the <REPOST> placeholder. However, this procedure does not solve the problem entirely. Many posts are copies of texts that originate outside of *vkontakte*, and some copies of *vkontakte* posts are made by copy-pasting (and possible editing) rather than with the "share" function. As an additional measure, posts that are longer than 90 characters are compared to each other in lowercase and with whitespaces deleted. If several identical posts are found, all but one are replaced with the placeholder. However, there are still many duplicates or half-duplicates left, which becomes clear when working with the corpora. Some of the duplicates, despite obviously coming from the same source, have slight differences in punctuation, spelling or even grammar, which means they were edited. It is a non-trivial question whether such half-copies should be removed. In any case, this remains a serious problem for the corpora in question. By my informal estimate, as much as 15% of the tokens found in the corpora could actually belong to near-duplicates. Before applying more advanced approach in the future, e.g. shingle-based (Broder, 2000), the near-duplicates have to be carefully analyzed to determine what has to be removed and what has to stay.

Final step of the filtering is anonymization. The purpose of anonymization is to avoid the possibility of identifying the users by removing their personal data. Usernames and IDs of the users are replaced with identifiers such as F_312. The numbers in the labels are random, but consistent throughout each corpus. This way, the corpus users still can identify texts written by the same person (which could be important for dialectological or sociolinguistic research) without knowing their name. The names of the groups are not removed because there is no one-to-one correspondence between groups and users. Similarly, user mentions in texts are removed. Just like in other major social media platforms, user mentions in *vkontakte* are automatically enhanced with the links to the user pages and therefore are easily recognizable. All such mentions are replaced with a <USER> placeholder. All hyperlinks are replaced with a <LINK> placeholder. Finally, user metadata is aggregated (see Subsection 2.4). Only the anonymized corpus files are uploaded to the publicly accessible server.

2.4 Metadata and annotation

Each post together with its comments is conceptualized as a separate document in the corpus. There are post-level and sentence-level metadata. Both include information about the authors: the owner of the page (post-level) and the actual author of the post or comment (sentence-level), which may or may not coincide. Additionally, sentence-level metadata includes type of message (post/repost/comment), year of creation, and language of the sentence.

Author-related metadata primarily comes from the user profiles. It includes sex (which is an obligatory field) and, if the user indicated it, also their age, place of birth and current location. Simply copying the values for the latter three parameters would make it possible to identify the authors. However, these values are extremely important for any kind of sociolinguistic or dialectological research, so they have to be accessible in some way. As a compromise, these values are presented only in aggregated form. Exact year of birth is replaced with a 5-year span (1990–1995, 1995–2000, etc.) in all corpora. The solution for the geographical values has only been applied to the Udmurt corpus so far. The exact locations there are replaced with areas: districts (*район*) for Udmurtia and neighboring regions with significant Udmurt minorities; regions (*область/республика/край*) for other places in Russia; and countries otherwise. The correspondence between the exact values and areal values was established manually and stored in a CSV table, which at the moment has around 800 rows for Udmurt. Since there are a lot of ways to spell a place name (including using Udmurt names, which do not coincide with the official Russian ones), this is a time-consuming process[5], which is why I have not done that for the other corpora yet.

In order to make sure the birth places the users indicate are real at least most of the time, I read posts written by a sample of users. It is common for speakers in this region to live in cities and towns, but maintain ties with their original villages and describe them in their posts. In such descriptions, the speakers often explicitly indicate that they were born in that village. Additionally, place of origin is an important part of identity. This is why opening sections of most interviews in local press contain the information about the village the interviewee was born, along with their name and occupation. All this makes birth place information easily verifiable. In most cases, the place name indicated by the users was corroborated by the information I found in the texts. There were several cases, however, when instead of naming the exact place, the users wrote the district center closest to the real place of birth. This paradoxically makes the aggregated version of geographical data more accurate than the exact one.

The token-level annotation in the corpora includes lemmatization, part-of-speech and full morphological annotation, morpheme segmentation and glossing. This annotation is carried out automatically using rule-based analyzers, with the details (coverage, presence of disambiguation, etc.) varying from language to language. Additionally, the dictionaries used for morphological analysis were manually annotated for Russian borrowings, place names and other proper names, which is required for high-quality language tagging. Russian sentences were annotated with the *mystem 3* analyzer (Segalovich, 2003).

Social media texts in any language tend to be more "noisy" and difficult for straightforward NLP processing, having higher out-of-vocabulary rates (Baldwin et al., 2013). There are both standard and language-specific problems in this respect in the Uralic social media. The former include typos, deliberate distortions and lack of diacritics. An example of the latter is significant dialectal variation, which was to a certain extent accounted for in the morphological analyzers. The variation is explained by the facts that these languages were standardized only in the 1930s and that many people are not sufficiently well acquainted with the literary standards (or choose not to adhere to them).

[5]This process could be partially automatized, of course, e.g. using databases with geographical information such as DBpedia and distortion-detecting techniques such as Levenshtein distance. I would prefer this approach if I had to process tens of thousands or more place names. However, I believe that for my data, developing an automatic processing tool together with subsequent manual verification would take more time than completely manual processing.

The most frequent typos were included in the dictionaries. Some kinds of distortions, such as repeating a character multiple times, were removed before a token was morphologically analyzed (but not in the texts). Lack of diacritics is a common problem in Udmurt, Komi and Mari texts, as alphabets of these languages contain Cyrillic letters with diacritics that are absent from a standard Russian keyboard. They can be either omitted or represented in a roundabout way. Interestingly, the same letters are represented differently in different languages. In Udmurt, double dots above a letter are commonly represented by a colon or (less frequently) a double quote following it, e.g. *ӧ = o: / o"*. In Komi, the letter *о* in this context is most often capitalized or replaced with the zero digit. In all languages, similarly looking characters from Latin-based character sets can be inserted instead of Cyrillic ones. Alphabets of Erzya and Moksha coincide with that of Russian. Nevertheless, double dots above *ё* are often omitted, following the pattern used in Russian texts (where their use is optional). All these irregularities are taken care of during automatic processing.

3 Properties of the texts

3.1 Size and distribution of metadata values

After the language tagging, the corpus files were filtered to exclude users who wrote exclusively or almost exclusively in Russian. For each user wall, number of sentences classified as Russian, Uralic or Unknown was calculated. The file was excluded from the corpus either if it contained at most 3 Uralic sentences constituting less than 10% of all sentences, or if it contained at most 10 Uralic sentences constituting less than 1% of all sentences. If the number of sentences classified as "unknown" was suspiciously high, the file was checked manually.

The sizes of the corpora after filtering are listed in Table 3. The two columns on the right give sizes of Uralic and Russian parts of each corpus in tokens. It has to be borne in mind that some of the tokens belong to near-duplicates (see Subsection 2.2), so the actual sizes after proper deduplication may be lower. The figures for Komi-Zyrian and Moksha are preliminary, however it is clear that the total size of the Moksha *vkontakte* segment is tiny compared to the rest of the languages.

	#Groups	**#Users**	**Uralic part**	**Russian part**
Udmurt	335	979	2.66M	9.83M
Komi-Zyrian	87	408	2.14M	16.12M
Erzya	20 (+ forums)	111 (+ forums)	0.83M (*vk*: 0.4M)	5.23M
Moksha	17	17	0.014M	0.17M

Table 3: Corpus sizes.

In Table 4, year-by-year figures for Udmurt, Komi-Zyrian and Erzya are presented. The figures for 2018 are left out because the data for the entire year is not yet available. However, at least for Komi-Zyrian and Erzya they are projected to continue the trends observed in earlier data.

Vkontakte was launched in early 2007, which is why there are no texts in the corpora before this date. The only exception is one of the Erzya forums, `http://erzianj.borda.ru`, which was started in 2006. The dynamics look different for Erzya on the one hand and the Permic languages on the other. After an initial gradual

Year	Udmurt	Komi-Zyrian	Erzya (*vk*)	Erzya (forums)
2006	0	0	0	15.9
2007	1.0	0.7	0.01	70.7
2008	15.1	1.9	0.7	23.1
2009	14.3	6.0	2.6	64.3
2010	42.7	5.9	3.8	**105.6**
2011	101.7	14.3	11.3	79.0
2012	273.1	33.0	29.2	40.8
2013	424.1	55.4	28.3	15.8
2014	473.6	140.6	79.2	20.4
2015	429.8	251.4	**96.5**	11.3
2016	350.6	259.0	70.8	1.4
2017	**505.2**	**660.6**	44.5	0.01

Table 4: Size of Uralic parts of corpora by year, in thousands of tokens.

increase in the number of texts, which continued until 2014–2015, number of Erzya *vkontakte* texts started going down. Permic segments of *vkontakte*, by contrast, continued growing, although Udmurt had a two-year plunge. The number of groups also seems to grows continuously: Pischlöger (2017) reported 90 groups in 2013 and 162 groups in 2016 for Udmurt. Komi-Zyrian speakers were adopting social media at a lower pace, but at the moment, Komi-Zyrian segment outnumbers the Udmurt one in terms of token counts. The Erzya forums enjoyed peak popularity around 2010. The reason for that was most probably the discussions about development of an artificial unified Mordvin language out of the two existing literary standards, Erzya and Moksha. This idea was advocated by Zaics (1995) and Keresztes (1995) and supported by Mosin (2014). The initiative belonged to people in the position of power rather than e.g. writers or teachers (Rueter, 2010, 7) and was vehemently opposed by Erzya language activists. This possibility was actively discussed in 2009, which energized the activists and led to the spike in the number of forum posts. The controversy seems to have abated since then, and both forums are now defunct (although still accessible).

The gender composition is even more different in Udmurt and Erzya (counting only *vkontakte* texts), as can be seen from the Table 5. Three quarters of texts authored by users (rather than groups) in Erzya were written by males, while in Udmurt it is the females who contribute more. The Udmurt picture is actually close to the average: according to a 2017 study by Brand Analytics[6], 58.4% of all posts in *vkontakte* are written by females. I do not have any explanation for this disparity.

	F	M
Udmurt	59.5%	40.5%
Erzya	24.7%	75.3%

Table 5: Proportion of tokens by sex of the author in *vkontakte*.

[6] `https://www.slideshare.net/Taylli01/2017-77172443`

3.2 Linguistic properties

Literary standards were developed for minority Uralic languages only in the 1930s, although written literature in them existed earlier. During the Soviet times, the standard language was taught at schools, however, this is not obligatory anymore and even unavailable in many places. Dialectal variation is still significant within each language. While older speakers generally try to follow the literary standard when writing, the younger generation may not know it well enough. Their written speech is therefore influenced by their native dialects, as well as by Russian. This contrasts to official texts and press in these languages, where puristic attitudes prevail. In Udmurt, the official register with its neologisms is hardly comprehensible for many speakers (Edygarova, 2013). In Erzya, neologisms in press are often accompanied by their Russian translations in parentheses because otherwise nobody would understand them (Janurik, 2015). Texts in the social media are much closer to the spoken varieties, which makes them better suited for the research of the language as it is spoken today.

Dialectal variation is observable in vocabulary and morphology. Frequently occurring non-standard suffixes include, for example, the infinitive in *-n* and present tense in *-ko* in Udmurt, or dative in *-ńe* and 1pl possessive in *-mok* in Erzya. This makes dialectological research on the social media corpora possible in principle. The main obstacle to such research is corpus size. Only a minority of users indicate their place of origin. Divided by the number of districts these people were born, this leaves a really small number of geographically attributed tokens for all districts except the most populous ones (several thousand to several dozen thousand tokens in the case of Udmurt). In order to see a reliable areal distribution of a phenomenon, that phenomenon has to be really frequent in texts.

As a test case, I used three dialectological maps collected for Udmurt using traditional methods: the distribution of the affirmative particles *ben/bon* (Maksimov, 2007b); the word for 'forest' (Maksimov, 2007a); and the word for 'plantain (Plantago; a small plant common in Udmurtia)' (Maksimov, 2013). The distribution of the affirmative particles was clearly recoverable from the corpus data: having an average frequency of over 1000 ipm, they had enough occurrences in most districts. The distribution obtained from the corpus coincided with the one from the dialectological map, although it had lower resolution. Out of 7 different names for the forest available on the dialectological map (excluding phonetic variants), 5 were present among the geographically attributed tokens of the corpus (*ńules*, *tel'*, *śik*, *ćašša*, *surd*). The overwhelming majority of occurrences in all districts belonged to the literary variant, *ńules*, while each of the other variants had only a handful of examples. Nevertheless, all these occurrences were attested exactly in the districts where they were predicted to appear by the dialectological map. Finally, the map for the plantain had 27 variants. Given the number of available options and the low frequency of this word, it is not surprising that its distribution turned out to be completely unrecoverable from the corpus. To sum up, it is possible to obtain some information on areal distributions of high- or middle-frequency phenomena from the social media corpora. However, in most cases this information can only be used as a preliminary survey and has to be supplemented by fieldwork or other methods to make reliable conclusions.

4 Availability

All social media corpora (as well as the "support corpora", see Subsection 2.4) are or will be available for linguistic research through an online interface[7]. Udmurt and Erzya corpora are already online. Komi-Zyrian and Moksha are being processed and will be available in December 2018. Komi-Permyak and both Mari corpora are scheduled for release in the first half of 2019.

Unfortunately, due to copyright and privacy protection reasons it is hardly possible to simply redistribute the source files freely. Instead, I currently employ a solution whereby the texts are only available through a search interface where the users can make queries and get search hits. The search hits appear in shuffled order, and for each sentence found, only a limited number of context sentences can be seen for copyright protection. This is a solution that is commonly applied in the web-as-corpus approach.[8] All data is anonymized (see Subsection 2.3). *Tsakorpus*[9] is used as the corpus platform. Queries can include any layer of annotation or metadata and support regular expressions and Boolean functions. Additionally, all code used for data processing will be available under the MIT license in a public repository.

5 Conclusion

In this paper, I described the ongoing project with the goal of creating social media corpora for seven medium-sized minority Uralic languages. The processing pipeline for these corpora includes semi-supervised identification of the texts (mostly in the *vkontakte* social networking service), downloading them through the API, language-tagging, filtering and anonymization, and morphological annotation. The corpora and tools used to build them are or will be publicly available. Sizes of the corpora vary, but do not exceed 3 million tokens written in the Uralic languages. Apart from those, each corpus also contains Russian sentences written by native speakers of the Uralic languages or in groups where Uralic texts have been posted; Russian parts of the corpora are several times larger than the Uralic ones. The corpora are better suited for sociolinguistic research than more traditional resources and contain texts written in a less formal register than those of press and fiction. Greater dialectal variation in the texts make them a possible source for dialectological investigations, which, however, have to be supported by independent sources to make reliable conclusions. In any case, given the scarcity of texts available digitally for the languages in question, the social media corpora will be a valuable resource for any kind of corpus-based linguistic research on them.

Acknowledgments

I am grateful to Irina Khomchenkova, who helped me with filtering URL lists for Erzya, Moksha, Meadow and Hill Mari, to Boglárka Janurik, who pointed me to resources in

[7] `http://volgakama.web-corpora.net/index_en.html`

[8] Although it is really widespread in web-based corpora, this solution is often left implicit. For example, non-public-domain parts of the Leeds Corpora have a context limitation of 150 characters in the KWIC view and slightly larger limitation in the "expanded context" view. However, this limitation, which has existed right from the start, is not mentioned by Sharoff (2006). Similarly, the one-sentence limitation and shuffling in the Leipzig Corpora Collection is not reported by Biemann et al. (2007), although it is mentioned elsewhere (Schäfer and Bildhauer, 2013).

[9] `https://bitbucket.org/tsakorpus/tsakonian_corpus_platform`

Erzya, and to Svetlana Edygarova, who helped me with translating the abstract into Udmurt.

References

Aller Media Oy. 2014. Suomi 24 2001-2014 (näyte) -korpuksen Helsinki-Korp-versio. http://urn.fi/urn:nbn:fi:lb-2016050901.

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How Noisy Social Media Text, How Diffrnt Social Media Sources. In *International Joint Conference on Natural Language Processing*. Nagoya, Japan, pages 356–364.

Caroline Barrière. 2016. Bilingual Corpora. In *Natural Language Understanding in a Semantic Web Context*, Springer, Cham, pages 105–125.

Chris Biemann, Gerhard Heyer, Uwe Quasthoff, and Matthias Richter. 2007. The Leipzig Corpora Collection: monolingual corpora of standard size. In *Proceedings of Corpus Linguistics 2007*.

Rogier Blokland and Cornelius Hasselblatt. 2003. The Endangered Uralic Languages. In Mark Janse and Sijmen Tol, editors, *Language Death and Language Maintenance. Theoretical, practical and descriptive approaches*, John Benjamins, Amsterdam & Philadelphia, Current issues in linguistic theory, pages 107–142.

Andrei Z. Broder. 2000. Identifying and Filtering Near-Duplicate Documents. In Gerhard Goos, Juris Hartmanis, Jan van Leeuwen, Raffaele Giancarlo, and David Sankoff, editors, *Combinatorial Pattern Matching*, Springer Berlin Heidelberg, Berlin, Heidelberg, volume 1848, pages 1–10. https://doi.org/10.1007/3-540-45123-4_1.

W. B. Canvar and J. M. Trenkle. 1994. N-Gram-Based Text Categorization. In *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval*. pages 161–176.

Michael Daniel, Nina Dobrushina, and Sergey Knyazev. 2010. Highlander's Russian: Case Study in Bilingualism and Language Interference in Central Daghestan. In Arto Mustajoki, Ekaterina Protassova, and Nikolai Vakhtin, editors, *Instrumentarium of Linguistics: Sociolinguistic Approaches to Non-Standard Russian*, Helsinki, number 40 in Slavica Helsingiensia, pages 65–93.

Svetlana Edygarova. 2013. Ob osnovnyx raznovidnostjax sovremennogo udmurtskogo jazyka [On the fundamental varieties of the modern Udmurt language]. *Ezhegodnik finno-ugorskix issledovanij* 3:7–18.

Boglárka Janurik. 2015. The emergence of gender agreement in Erzya-Russian bilingual discourse. In *Language Empires in Comparative Perspective*, Walter de Gruyter, pages 199–218.

Heidi Jauhiainen, Tommi Jauhiainen, and Krister Lindén. 2015. The Finno-Ugric Languages and The Internet Project. *Septentrio Conference Series* (2):87–98. https://doi.org/10.7557/5.3471.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2018. Automatic Language Identification in Texts: A Survey. *arXiv:1804.08186 [cs]* ArXiv: 1804.08186. http://arxiv.org/abs/1804.08186.

László Keresztes. 1995. On the Question of the Mordvinian Literary Language. In Gábor Zaics, editor, *Zur Frage der uralischen Schriftsprechen [Questions of Uralic literary languages]*, Az MTA Nyelvtudományi Intézete, Budapest, Linguistica, Series A, Studia et Dissertationes, pages 47–55.

András Kornai. 2016. Computational linguistics of borderline vital languages in the Uralic family. Paper presented at International Workshop on Computational Linguistics for Uralic Languages 2. http://kornai.com/Drafts/iwclul.pdf.

Sergey Maksimov. 2007a. 'Les' v udmurtskix govorax: Dialektologicheskaja karta i kommentarij ['Forest' in Udmurt varieties: A commented dialectological map]. *Idnakar: Metody istoricheskoj rekonstrukcii* 2(2):56–69.

Sergey Maksimov. 2007b. Upotreblenie chastic ben-bon 'da' v udmurtskix dialektax [Use of the ben/bon 'yes' particles in Udmurt dialects]. *Nauka Udmurtii* 2(15):75–82.

Sergey Maksimov. 2013. Nazvanija podorozhnika v udmurtskix dialektax i ix proisxozhdenie [The Names of Plantain (Plantago) in the Udmurt Dialects and Their Origin]. *Ezhegodnik finno-ugorskix issledovanij* (4):7–17.

Mikhail Mosin. 2014. Sozdavat' li edinye literaturnye jazyki dlja ural'skix narodov? [Should unified literary languages be created for Uralic peoples?]. *Trudy Karel'skogo nauchnogo centra RAN* (3):76–82.

Boris Orekhov, I. Krylova, I. Popov, L. Stepanova, and L. Zaydelman. 2016. Russian Minority Languages on the Web: Descriptive Statistics. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016"*. RSUH, Moscow, Computational Linguistics and Intellectual Technologies, pages 498–508.

Boris Orekhov and Kirill Reshetnikov. 2014. K ocenke Vikipedii kak lingvisticheskogo istochnika: sravnitel'noe issledovanie [Evaluating Wikipedia as a linguistic source: A comparative study]. In Yana Akhapkina and Ekaterina Rakhilina, editors, *Sovremennyj russkij jazyk v internete [Contemporary Russian language on the Internet]*, Jazyki slavjanskoj kul'tury, Moscow, pages 309–321.

Saša Petrović, Miles Osborne, and Victor Lavrenko. 2010. The Edinburgh Twitter Corpus. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*. Association for Computational Linguistics, Stroudsburg, PA, USA, WSA '10, pages 25–26. http://dl.acm.org/citation.cfm?id=1860667.1860680.

Christian Pischlöger. 2017. Udmurtskij jazyk v social'noj seti "VKontakte": Kvantitativnye i (vozmozhnye) kvalitativnye issledovanija [Udmurt in the Vkontakte Social Network: Quantitative and (Possible) Qualitative Research. In *Elektronnaja pis'mennost' narodov Rossijskoj Federacii: Opyt, problemy i perspektivy [Digital literacy of the nations in Russia: Experience, challenges and perspectives]*. GOU VO KRASGSiU, Syktyvkar, pages 154–162.

Jack Rueter. 2010. *Adnominal Person in the Morphological System of Erzya.* Number 261 in Mémoires de la Société Finno-Ougrienne. Société Finno-Ougrienne, Helsinki.

Kevin P. Scannell. 2007. The Crúbadán Project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*. volume 4, pages 5–15.

Roland Schäfer and Felix Bildhauer. 2013. *Web Corpus Construction*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool.

Ilya Segalovich. 2003. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In *MLMTA-2003*. Las Vegas.

Serge Sharoff. 2006. Creating general-purpose corpora using automated search engine queries. In *WaCky! Working papers on the Web as Corpus*, pages 63–98.

Natalya Stoynova. 2018. Differential object marking in contact-influenced Russian Speech: the evidence from the Corpus of Contact-influenced Russian Speech of Russian Far East and Northern Siberia. In *Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue"*. RSUH, Moscow, pages 721–734.

Viljami Venekoski, Samir Puuska, and Jouko Vankka. 2016. Vector Space Representations of Documents in Classifying Finnish Social Media Texts. In Giedre Dregvaite and Robertas Damasevicius, editors, *Information and Software Technologies*, Springer International Publishing, Cham, volume 639, pages 525–535. https://doi.org/10.1007/978-3-319-46254-7$_4$2.

J. Vinosh Babu and S. Baskaran. 2005. Automatic Language Identification Using Multivariate Analysis. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing: Proceedings of CICLing 2005*, Springer, Berlin, Heidelberg & New York, pages 789–792.

Gábor Zaics. 1995. Skol'ko jazykov nuzhno erze i mokshe? [How many languages do the Erzya and the Moksha need?]. In *Zur Frage der uralischen Schriftsprechen [Questions of Uralic literary languages]*, Az MTA Nyelvtudományi Intézete, Budapest, number 17 in Linguistica, Series A, Studia et Dissertationes, pages 41–46.