# Languages of Russia: Using Social Networks to Collect Texts

Irina Krylova, Boris Orekhov, Ekaterina Stepanova, and Lyudmila Zaydelman(✉)

National Research University Higher School of Economics, Moscow, Russia
krylova93@gmail.com, nevmenandr@gmail.com,
stepanovayekaterina@gmail.com, luda.zaidelman@yandex.ru

**Abstract.** In this paper we outline a method of finding texts in minor languages of Russia in social networks by the example of VKontakte. We find language-specific markers – special tokens that contain letter combinations unique to a certain language and highly frequent in texts in this language. We use Yandex.XML to generate lists of web-pages that contain texts in these languages. We then download data from web-pages in the https://vk.com domain through Vkontakte API.

**Keywords:** Minor languages · Lexical markers · Social networks

## 1 Introduction

There are over a hundred national languages in Russia, excluding Russian and languages that are official in other countries. In this paper we use a term "minor languages" despite the fact that some of the languages count more than a million native speakers. However, linguistic tools for all of these languages are equally few. The lack of tools – first and foremost, the corpora – results from the lack of digitized texts in these languages. The Wikipedia seems to be the most obvious source of such texts and it does, in fact, contain sections in many of Russian national languages. But as Orekhov and Reshetnikov [1] showed in their paper, the Wikipedia is rarely (and in case of minor languages – very rarely) a relevant linguistic source of texts. The goal of this project is to collect texts in Russian national languages which will then be used to create text datasets (like marked-up corpora, sets of n-grams and so on).

## 2 Why We Use Social Networks

Web-pages of regional newspapers and local municipal bodies are quite common on the Internet, and either all or most of the texts found in such web-pages are written in the national language of the respective region. Nevertheless, we chose against using these pages as the main source of texts in favor of the social networks, primarily Russia's most popular one – VKontakte (https://vk.com). We do not discard common web-pages completely, though, as shown later in the paper.

P. Braslavski et al. (Eds.): RuSSIR 2015, CCIS 573, pp. 179–185, 2016.
DOI: 10.1007/978-3-319-41718-9_11

There is a number of reasons behind this decision. Firstly, even though a social network contains pages in all kinds of different languages, these pages are still identical as far as their structure is concerned. For us as developers this means that we are able to create a universal page-processing tool once we understand what the structure of the page is. With usual web-pages, on the contrary, the structure would differ from one web-page to another.

Secondly, social networks often provide an API and this makes page-processing even easier. An API offers a number of methods that allow third party software to access the social network's data. This means that at this point the structure of the pages becomes irrelevant and the problem comes down to using the necessary method.

Finally, it is the social aspect of social networks. They encourage natural, live communication between actual people, and the texts produced in the process are natural as well, as opposed to automatically generated ones often found on the Wikipedia. Pischlöger [2] mentions additional advantages of social networks for minor languages from the point of view of users in that social networks are cheap, easy to use, and provide communication over long distances; he also states that informal use of language in social networks lowers borders that exist in written language, which is important for people without formal education.

## 3 How It Works

In this section we describe the technical aspects of collecting texts written in the minor languages of Russia.

There are different techniques for gathering web-corpora. For example, Boleda et al. [3] in order to collect Catalan corpus used as initial (seed) list of domains from a Spanish search engine Buscopio and then crawled other web-pages that either had the.es suffix or were assigned by IP to a network located in Spain. After that they used language filtering to separate Catalan from other languages, applied duplicate detection and successfully gathered 166 million word corpora.

The most popular method for large corpora gathering is to use search engine queries to gather seed URLs from this search engine result page and then crawl these URLs. This method called WaC (Web as Corpus) was first proposed by Baroni et al. [4] and is now used by various researchers [5, 6].

In [5] Guevara used this method to collect a Norwegian corpora. He made up a list of frequent words based on a dump of the Norwegian Wikipedia, then took the top 2000 of them and used different pairs of these words to find pages in Norwegian. Finally, by limiting the results to the pages in the no domain, he gathered a list of seed web-pages in Norwegian.

In our work we mostly follow the steps described in [5], limiting the results to the vk.com domain. In general, in our project we search for different sites in national languages of Russia, however this paper describes only the part of work that deals with the most popular Russian social network. Instead of the top frequent words from the Russian Wikipedia, we use the so-called lexical markers. Unacceptability of most frequent words from Wikipedia for collecting corpora of national languages of Russia was proved by Orekhov and Reshetnikov in [1] at the example of Bashkir, Tatar and some other languages. Most frequent words in these languages, according to the respective versions of the Wikipedia, are water-related terms “river” and “basin”, which contrasts with the more common idea of function fords being the most frequent in a language. The concept of lexical markers and the process of selecting them is discussed in the next section.

### 3.1 Lexical Markers

A lexical marker of a language is a word that is unique to the language and therefore uniquely defines it. We use such words to find web-pages (including pages on social networks) that contain texts in Russian national languages. We collect lexical markers manually from grammars, vocabularies and phrasebooks for the languages in question. Obviously, automatic marker search would be preferable but it is currently impossible as explained later.

Our goal is to find as many web-pages as possible so lexical markers need to be frequent in the language. As a result all of the collected markers are function words.

On the other hand, we wanted to avoid finding pages that contain texts in languages other than the one we look for. That is why the markers are required to be graphically unique and not to occur in other languages. We understand, that a marker is unique by posting it to the Yandex search engine and analyzing search result pages. Mostly, it is not that difficult to identify the page language, as most of them contain nation or country name in the title.

Apart from these compulsory restrictions on markers there is an additional one: markers should only contain Cyrillic symbols. In texts found on the Internet symbols containing diacritics are often replaced with their Cyrillic analogues that are either graphically or phonetically similar to them, i.e. Bashkir “ң” replaced by “н” or “ө” replaced by “о”. This phenomenon is called “everyday written language” [7]. Some of the manually collected markers contain diacritics so it is necessary to provide a way to replace them consistently, though the symbol pairs may vary depending on the language. It is also necessary to check if a marker remains unique after the replacement procedure. This additional restriction can be easily explained: people who speak Russian national languages usually have a Russian keyboard layout and they often forget or just do not want to switch the layout to Udmurt, thus they write without using diacritics.

The combination of these restrictions is the reason why making the marker search automatic is impossible. Obviously, a rather small set of texts would be quite sufficient for determining the most frequent words in any of the languages we are interested in. However, we would also need to make sure that the markers do not match any word in any other language. This task requires far larger collections of texts or at least a method to create such collections. This brings us back to the problem outlined in the introduction: such collections currently do not exist and developing a method to create them is the goal of our project.

The number of markers we were able to find varied for different languages: while for Tabasaran we have six markers that meet all of the restrictions (Tabasaran words for “how many”, “if”, “someone”, “bigger/greater”), for Tatar we only have three (Tatar words for “whole”, “then”, “again”), two of which contain diacritics. We provide the translations for these words but not the words themselves to minimize the number of documents where the markers would occur in an “artificial” linguistic environment.

### 3.2 Collecting URLs

When we have a set of lexical markers for a language, the next step is to find the web-pages that contain texts written in this language. Our tool of choice for the task is Yandex.XML – a Yandex service that enables automatic search queries to Yandex search engine [8]. The number of queries we can make is limited, in our case the limit is 1000 queries per day. For every language we send each of its marker in a separate query and by combining the resulting domain lists for every marker we get a list of domains that contain texts in a given language.

The next step is to search web-pages inside a domain. We use Yandex.XML for this as well by sending queries that contain the name of a domain and a marker both corresponding to a given language. Currently we use only web-page lists that we get from queries with domain name set to https://vk.com, but we store all web-page lists and plan to extract texts from them.

We made a decision to work with community pages rather than users’ personal pages. The decision is based on our assumption that an average user would use languages other than the national language on his page because he would have friends who do not speak it. On the contrary, a community unites people who share common interests or, more importantly, a common language. For this reason first, manually composed lists of web-pages in the https://vk.com domain only contained URLs to communities. Unfortunately, the majority of the pages in the lists we get from Yandex.XML are URLs of users’ personal pages, e.g. 286 personal pages out of 450 total for Tatar.

### 3.3 Processing a VKontakte Page

When we have a list of VKontakte web-pages, we proceed to extract all necessary information by sending queries to VKontakte API. The end result of page processing is a JSON file that contains the following information:

- file metadata: name of the language, creation date.
- list of posts on the community wall.
- lists of comments for each post.
- information about the author of a post or a comment: user id, first name, last name, gender, date of birth, city; user ids and names can be used for corpora cue mark-up, which is useful for linguistic and sociolinguistic research.

### 3.4 VKontakte API Limitations

While the use of VKontakte API undoubtedly simplifies and speeds up VKontakte page processing, there are also a number of restrictions. Firstly, there is a 3 requests per second limitation for all of the methods. Secondly, there are method-specific limitations, e.g.

comment and wall post retrieving methods can only return up to 100 comments or posts respectively [9]. These limitations combined with a large number of languages and communities do not allow us to proceed as fast as we would prefer, but we are constantly improving our processing toolset to increase the processing speed.

## 4 Preliminary Results

We have been working on this project for a relatively short time and it is still far from complete, but we already have some results. We have downloaded a number of communities devoted to or using minor languages of Russia, the table underneath provides detailed statistics:

**Table 1.** Intermediate results: the number of downloaded communities for each language

| Language | Total (found manually) | Downloaded | Total (found automatically) | Downloaded | Overlap |
|---|---|---|---|---|---|
| Adyghe | | | 26 | 5 | |
| Avar | | | 17 | 7 | |
| Bashkir | 135 | 107 | 787 | In progress | 11 |
| Buryat | | | 181 | 59 | |
| Chuvash | | | 315 | 90 | |
| Erzya | | | 76 | 26 | |
| Ingush | | | 270 | 87 | |
| Kalmyk | | | 182 | 54 | |
| Karachay-Balkar | | | 42 | 16 | |
| Khakas | 4 | 3 | 4 | 2 | 1 |
| Komi-Zyrian | | | 89 | 39 | |
| Lak | | | 2 | 0 | |
| Mari | | | 161 | 54 | |
| Udmurt | 72 | 53 | 769 | In progress | 33 |
| Tabasaran | | | 1 | 1 | |
| Tatar | | | 450 | 138 | |
| Tuvan | | | 444 | In progress | |

We were originally provided with community lists for three of the languages (Bashkir, Khakas, Udmurt), which were manually composed during previous research. Since then we were able to automatically generate new web-page lists for all languages for which there had markers using our wrapper for Yandex.XML. Table 1 shows that the automatically generated lists for Bashkir, Khakas and Udmurt are larger than manual ones, as one would expect. However, a large part of the generated lists are actually URLs that we would not normally consider. Firstly, there are a lot of URLs for users' personal pages, with which we decided not to work. Secondly, in some cases several URLs correspond to a single community, e.g. URLs for communities and URLs for individual posts in these communities.

This does not explain why the overlap is so small, even though we could expect the manual lists to be subsets in automatically generated ones. The main reason is the method we used to determine the overlap: for each of the three languages we simply calculated the number of URLs found in both lists. Unfortunately this method overlooks cases when different URLs refer to the same web-page, e.g. https://vk.com/public85682520 and https://vk.com/novostibarum. Once we download Bashkir and Udmurt communities we should be able to take this into account and provide more accurate and hopefully higher figures.

There is one other problem that we came across when we studied the data for the downloaded communities and that is identifying the language. Because we use lexical markers to find VKontakte communities, we are certain that a given minor language is used in these communities. We cannot, however, guarantee that other languages like Russian are not used as well. Indeed, we find examples of Russian and a minor language used not just in separate replies in a dialogue but in a single reply: *То самое чувство когда понимаешь, что 1 шырпы менан утты яндырдып ебарден*. In this line found in a Baskir-speaking community we see both code-switching and use of the aforementioned "everyday written language" (word *менән* written as *менан*). The question is: what is this language? Obviously, we cannot say if it is Russian or Bashkir, the answer in this case, as well as in many other cases, lies somewhere in between. Let's now look at another line found in an Udmurt community: *Стив Джобс – почетной удмурт*. What language is this? Any automatic language identifier would recognize this as written in Russian, even though there is a minor mistake in that an adjective почетной and a noun удмурт do not syntactically agree, which is actually rather common for texts found in social networks. However, from the point of view of Udmurt grammar this line is absolutely correct being an actual Udmurt translation for "Steve Jobs is an honorary Udmurt".

We are definitely not the only ones aware of the code-switching problem. C. Pischlöger provides several examples [10, 11] of Udmurt and Russian switching in VKontakte. He calls this phenomenon "suro-pojo" ("суро-пожо" – "mix" in Udmurt) and states that this mix characterizes the contemporary situation of a living minor language. From the words of his informant, "only foreigners speak clear Udmurt".

## 5 Conclusion

The approach proposed in the paper has proved to be quite effective. We currently have lexical markers for 97 languages of Russia. These markers were used to generate web-pages lists via Yandex.XML and we have so far collected lists of web-pages in the https://vk.com domain for 32 languages and lists of web-pages in other domains for 18 languages. We have also downloaded (completely or partially) communities related to 17 languages using VKontakte API. These are very early numbers and we expect them to increase as we continue our work and collect larger sets of texts in languages of Russia. We plan to share our corpora with the community as soon as we have got more structured and marked-up data.

**Acknowledgements.** We thank Timofey Arkhangelskiy for pointing out difficulties of language identification by the example of Udmurt.

## References

1. Orekhov, B.V., Reshetnikov K.Yu.: To the assessment of Wikipedia as a linguistic source (К оценке Википедии как лингвистического источника), Contemporary Russian on the Internet (Современный русский язык в интернете), Moscow, Jazyki slavjanskoy kul'tury, pp. 310–321 (2014)
2. Pischlöger, C.: Besermyan in the internet: social networks as a chance for language maintaining? (Бесермяне в интернете: социальные сети как шанс для сохранения родного языка?), Problems of ethno-cultural interaction in the Ural-Volga region: history and the present (Проблемы этнокультурного взаимодействия в Урало-Поволжье: история и современность), Samara, pp. 216–219 (2013)
3. Boleda, G., Bott, S., Meza, R., et al.: CUCWeb: a Catalan corpus built from the web. In: Proceedings of Second Workshop on the Web as a Corpus at EACL 2006 (2006)
4. Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E.: The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. Lang. Resour. Eval. **43**(3), 209–226 (2009)
5. Guevara, E.: NoWaC: a large web-based corpus for Norwegian. In: NAACL HLT 2010 6th Web as Corpus Workshop, pp. 1–7 (2010)
6. Ljubešić, N., Erjavec, T.: hrWaC and slWac: compiling web corpora for Croatian and Slovene. In: Proceedings of 14th International Conference, Pilsen, Czech Republic, pp. 395–402 (2011)
7. Zaliznyak, A.A.: Old Novgorod dialect (Древненовгородский диалект), Moscow, Jazyki slavjanskoy kul'tury (2004)
8. Yandex.XML – https://tech.yandex.ru/xml/
9. VK API – https://vk.com/dev/api_requests
10. Pischlöger, C.: Udmurt and Besermyan languages in social networks (Удмуртский и бесермянский языки в социальных сетях). In: Proceedings of International Science-Practical Conference, Dedicated to 260-Anniversary of V.G. Korolenko Материалы Международной научно-практической конференции, посвященной 260-летнему юбилею В.Г. Короленко.), Glazov, pp. 187–190 (2013)
11. Pischlöger, C. Notes from Murjol underground: super Udmurts in cyberspace (Запис(к)и из Мурҗол Underground: Super удмурты в Cyberspace). In: Proceedings of IV International Science-Practical Conference "Florov's Readings" (Материалы IV Международной научно-практической конференции "Флоровские чтения"), pp. 56–59. Glazov pedagogical institute, Glazov (2014)