Developing morphologically annotated corpora for minority languages of Russia

Timofey Arkhangelskiy

National Research University Higher School of Economics

tarkhangelskiy@hse.ru

Maria Medvedeva

University of Groningen

medvmr@gmail.com

Abstract

Despite recent progress in developing annotated corpora for minority languages of Russia, still only about a dozen out of about 100 have comprehensive corpora, and even less have computational tools such as machine translation systems or speech recognition modules. However, given that many of them have resources such as dictionaries and grammars, the situation can be improved at relatively low cost. In the paper we demonstrate the pipeline that can be used for developing such corpora, featuring the development of Udmurt and Adyghe corpora. The methods we describe are in principle applicable to any language for which certain kind of linguistic resources are available.

1. Introduction

Language corpora are one of the primary instruments of research in contemporary linguistics. Corpora allow researchers from all over the world to analyze raw language data rather than its interpretations by other linguists in grammars and articles. Compiling publicly available corpora is particularly important for more 'remote' languages most researchers have restricted physical access to. But, precisely because of their remoteness and poorer accessibility, there are no corpora for most such languages, while there are a multitude of quality corpora for most European languages.

In this paper we speak about developing corpora for minority languages of Russia. Despite their genetic diversity, these languages are similar in several respects, which makes certain approaches applicable to all or most of them. We will focus primarily on the cases of Udmurt and Adyghe and show that solutions we used in their development can be employed for creating corpora of other languages of Russia at a reasonably low cost. The Udmurt corpus was first released in 2014 and is available at http://web-corpora.net/UdmurtCorpus/search . Adyghe corpus is currently under development. The pilot version of the corpus currently has restricted access, but it is expected to be released later in 2016 at the same portal.

2. Languages of Russia and their corpora

There are 93 living indigenous minority languages spoken in Russia, according to Ethnologue (Lewis et al., 2016; the number should not be seen as precise because of the language vs. dialect uncertainty). All or almost all of them share several features important for corpus linguistics.

First, vast majority of them are written and have official orthography, which, with the exception of a

handful of Finnic languages, is based on Cyrillic alphabet. As virtually all of these orthographies were developed in the 1930s or later, they represent the phonology in a pretty straightforward fashion, unlike in English, Russian or other languages with long written tradition. Having been developed by professional linguists, these orthographies faithfully reflect all phonological distinctions. On the level of lexicon, these languages share numerous loanwords from Russian. On the level of grammar, all these languages are morphologically rich, having on average more morphologically expressed grammatical categories than Standard European languages. What this implies is that in order to be useful for a wide range of linguistic research, their corpora should have full morphological tagging including all morphological categories, rather than mere POS-tagging. Fine-grained morphological annotation is also essential for developing ulterior levels of annotation, such as syntactic parsing or anaphora resolution, in morphologically rich languages (see e.g. Goldberg and Elhadad, 2013 on syntactic parsing of Hebrew).

However, what is more important, is that quality linguistic resources have been created for these languages. Virtually all of them have grammars and many have extensive bilingual (usually X-to-Russian) dictionaries. These resources, as we will show, can be transformed into taggers relatively easily, and thus are crucial for low-cost corpus development.

Existing corpora of minority languages of Russia can be split into two groups: relatively small (almost always under 100,000 tokens) manually annotated collections, mainly containing spoken texts, and larger ones (at least several hundred thousand tokens, and usually more than one million) with automatic annotation. Numerous corpora of the former kind have been collected for various languages and dialects in linguistic expeditions since the 1960s. However, their size, which is naturally constrained by the amount of time and money required for their collection, is too small for many kinds of research, especially if the research involves statistics. In this paper, we focus on larger, automatically annotated (and mostly written) corpora, which are more suitable for low-cost development. To the best of our knowledge, such large corpora have been released for the following 13 minority languages of Russia belonging to five language families:

- East Caucasian: Avar[1], Dargwa[2], Lezgian[3];
- Indo-European: Ossetic[4], Romani[5];
- Mongolic: Buryat [6] (Badmaeva 2015), Kalmyk[7];
- Turkic: Tatar [8] (Suleymanov et al. 2011), Bashkir (Buskunbaeva, Sirazitdinov 2011), Khakas[9] (Sheymovich 2011);
- Uralic: Udmurt, Mari [10] (Bradley 2015), Komi[11].

Apart from those, there are several ongoing corpus development projects that we know of, including the Adyghe corpus project. There are also reports on developing corpora for Chuvash (Zheltov 2015), Tuva (Salchak, Bayirool 2013) and Yakut (Leontyev 2014), but the status of these projects is unclear.

3. The pipeline of corpus development

3.1. Collecting the texts

Books and other printed materials exist for most of the languages in question, but the cost of scanning, OCR and proofreading sufficient amount of texts is prohibitive for a low-budget corpus project. The only way to obtain a sufficiently large text collection at low cost is therefore the Internet. Unfortunately, this constraint makes it impossible to build corpora for the small and critically endangered languages that have very low digital vitality, in terms of Kornai (2013). However, it seems that more than one third of the languages in question are to some extent represented on the web. According to the estimates of Zaydelman et al. (2016), 30 to 40 languages of Russia have visible amount of texts on the Internet. The overall size of available texts

[1] http://web-corpora.net/AvarCorpus/search/
[2] http://dag-languages.org/DargwaCorpus/search/
[3] http://dag-languages.org/LezgianCorpus/search/
[4] http://corpus.ossetic-studies.org/search/ (Iron dialect), http://corpus-digor.ossetic-studies.org/search/ (Digor dialect)
[5] http://web-corpora.net/RomaniCorpus/search/
[6] http://web-corpora.net/BuryatCorpus/search/
[7] http://web-corpora.net/KalmykCorpus/search/
[8] http://web-corpora.net/TatarCorpus/search/
[9] http://khakas.altaica.ru/texts/
[10] http://corpus.mari-language.com/
[11] http://komicorpora.ru/

varies between a couple of thousand and several dozen million tokens. Our corpus of Udmurt, 13th largest minority language and probably the most digitally well-represented Uralic minority language, currently contains 7.3 million tokens, which covers the vast majority of all digitally available texts for this language. The volumes of the available data are slowly, but steadily growing: according to the year distribution of our texts, the growth rate is on average 0.7 million tokens per year in 2011-2015.

The texts available on the Internet fall mainly into one of the following groups: digital newspapers/mass media, blogs/social media and Wikipedia articles. For the genre composition of the Udmurt corpus see Table 1. It can be seen that the corpus is severely unbalanced, as the genre distribution is skewed in favor of press, followed by blogs with less than 6%. Our survey of texts in other minority languages available online suggests that the distribution is roughly the same for all these languages (again, with possible exceptions of Tatar and Bashkir). Lack of balance, which is inevitable in the proposed method of corpus development, is one of its largest downsides.

Genre	Tokens (millions)	%
press	6.64	90.56%
blogs	0.42	5.71%
New Testament	0.13	1.73%
Wikipedia articles	0.06	0.84%
non-fiction	0.03	0.40%
poetry	0.03	0.40%
fiction	0.02	0.36%
Total	7.33	

Table 1: Udmurt corpus genre composition

Now, the corpus does not include Udmurt posts from *vkontakte*, the most popular social network in Russia, which are estimated to contain more than 0.5 million tokens.

The resulting text collection resembles the corpora developed within 'Web as corpus' approach (Kilgarriff and Grefenstette, 2003). There is, however, an important difference between 'web as corpus' and the approach presented here. While the former aims at gathering vast amounts of data for NLP purposes, the objective of the latter is to collect *all* available texts in a given language, as the size of the collection is limited for minority languages. According to our estimate, there are less than 100 web domains containing texts in Udmurt. This order of magnitude allows for manual inspection of all relevant web domains (probably with the exception of Tatar and Bashkir, the most digitally viable of all these languages) and do not require extraordinary computational resources to process them.

Another potential pitfall in this process, besides poor balance, is low quality of texts on Wikipedia. While for larger languages Wikipedia is often used as a convenient and reliable source of linguistic data, Wikipedias in minority languages of Russia often contain a substantial number of automatically generated and thus linguistically useless content, which can be easily seen in their distorted frequency lists (Orekhov and Reshetnikov, 2014). If Wikipedia articles are to be used at all, they should be filtered (e.g. by length), after which normally only a small number of articles make it to the corpus. The corresponding figure in Table 1 shows the size of the Wikipedia subcorpus after filtering.

3.2. Morphological tagging

Given that the texts can be collected from the Internet and that tokenization is not much of a problem for minority languages of Russia, development of a morphological tagger is the most difficult step in corpus development. Both statistical and dictionary-based taggers require substantial amount of manual labor if built from scratch. The former have to be trained on sufficiently large manually annotated collections, while the latter require that a grammatical dictionary is compiled manually. However, the bilingual dictionaries available for the languages of Russia make compilation of a grammatical dictionary a much easier task. This fact, as well as the tradition of grammatical description of Russian that was started by Zaliznyak (1977), is the reason why all corpora listed in section 2 use the dictionary-based approach.

The idea is to manually write a formalized description of the morphology based on the grammars, and then transform a bilingual dictionary into a grammatical dictionary. In the Udmurt and

Adyghe projects we used the UniParser format and software for formalized description and tagging, which were also used for most other aforementioned corpora (Arkhangelskiy et al., 2012). There are also plenty of alternatives, including PC-KIMMO (Antworth, 1992), used in the Tatar corpus tagger, or giellatekno infrastructure (Moshagen et al., 2013).

The central problem in this step is the fact that bilingual dictionaries normally do not contain necessary grammatical information such as part of speech or declension / vowel harmony type; they have to be automatically restored. We combined three approaches to address this issue.

First, the form of the lemma in some cases clearly indicates its part of speech. In Udmurt, we tagged as verbs all lemmata ending in *-ɨnɨ* or *-anɨ* (markers of the infinitive). Manual check found that only one word, *ǯɨnɨ* 'half', was tagged incorrectly during this step.

All other parts of speech, however, did not have any markers that could be used as clues for part-of-speech tagging. In Adyghe, a polysynthetic language where bare stems are used as citation forms and parts of speech in general are not well differentiated, this was impossible altogether. The approach we used for these cases was using the tag given by a Russian tagger (specifically, mystem (Segalovich, 2003)) to the first non-abbreviated word of the translation. This worked surprisingly well: in Udmurt, around 85% of these tags proved to be correct. The wrong tags came primarily from two sources. First, some of the translation equivalents in both Udmurt and Adyghe dictionaries had several possible analyses, e.g. in adjectives which are commonly used as (substantivized) nouns. Second, Udmurt has an extensive (hundreds of items) inventory of ideophones, or imitative words that do not have Russian translation equivalents and are translated periphrastically, e.g. *č'iš-č'aš* 'about burning of wet wood'.

As the final approach we wrote some simple scripts. In Udmurt, the only additional field needed for tagging beyond part of speech is the conjugation type, which is determined by the last vowel of the stem (Winkler 2000: 45). In Adyghe, there is a regular *e/a* alternation in stems of a certain kind (Arkadyev and Testelets, 2009). Whenever the script sees an alternated stem in the lemma, it generates the base form and adds it to the list of stem allomorphs in the grammatical dictionary.

Apart from the challenge posed by part of speech tags, the excessiveness of the information in the dictionary can be an obstacle. One of its manifestations is abundance of synonymous translation equivalents, usage examples and phrases in dictionary entries, which are usually not needed in the corpus and thus have to be cut out. In Adyghe, for which several dictionaries were used as an input, this lead to especially long translations, since different dictionaries used different synonyms for translating the same word. This issue was addressed by passing the translation equivalent through a number of transformations. All secondary meanings and comments were removed by cutting out segments in parentheses, after semicolons and after colons if certain threshold length has been reached. In the case of Adyghe, the synonyms in translation equivalents were rearranged in a decreasing frequency order (according to the data from Russian National Corpus), so that the most frequent synonym appeared first and all the rest could be easily deleted during the manual proofreading.

Another manifestation of this problem lies in the list of words included in the dictionary. Apart from too many (potential) Russian loanwords that will probably never appear in a corpus, such as *aerosyomka* 'aerial photography', dictionaries for minority languages of Russia tend to include absolutely compositional and productive derivatives or word forms as separate entries. For both Udmurt and Adyghe, this involves, first and foremost, verbal derivation. In Udmurt, causative, detransitive and iterative forms of most verbs were included in the corpus, however only a handful of them have somewhat non-compositional meaning. If left as is, the tagger based on such a dictionary would give seemingly ambiguous results for words containing these affixes. For example, the word *vera-l'l'a-z* speak-ITER-PST.3SG will be ambiguously tagged as ITER.PST.3SG form of the verb *verani* 'speak' and as PST.3SG form of the verb *veral'l'anɨ* 'speak (repeatedly)'. These words were removed from the dictionary with a script that searched for a marker of one of these categories and checked if the remaining part was listed in the dictionary as a separate verbal stem. The situation is somewhat more difficult in Adyghe. Adyghe is a polysynthetic language, which means that the stem can attach numerous derivational affixes. While

most of these combinations are perfectly compositional, some are not, therefore manual check of all such complex stems is required. Words involving non-compositional combinations of stems and derivational affixes in Adyghe corpus get two levels of annotation, one for the original stem, the other for the combination of the stem and the affix (Arkhangelskiy and Lander, 2016). The interfaces enables users to search either for all occurrences of a given stem, or only those occurrences where it is not part of a non-compositional combination.

Finally, the dictionaries have to be extended manually by adding irregular words (mostly pronouns) and frequent regular words that were absent due to scarcity of the source dictionary or conversion errors. The Udmurt tagger, to which all pronouns and no more than a hundred other frequent lexemes were added manually, currently covers about 88% of the tokens in the corpus. Here is an example of an entry from the resulting Udmurt dictionary:

```
-lexeme
 lex: кизьыны
 stem: киз.
 gramm: V,I
 paradigm: connect_verbs-I-soft
 trans_ru: сеять, посеять, засеять
```

The entry contains fields indicating its lemma, stem, grammar tags, set of inflectional affixes and Russian translation.

4. Conclusion

The presented pipeline, which we used for developing the Udmurt corpus and which is currently used in the Adyghe corpus project, allows for relatively inexpensive construction of digital corpora. The proposed approach is applicable to digitally represented languages which have grammars and dictionaries. According to our estimates, there are still 15 to 20 minority languages of Russia that lack comprehensive written corpora but have enough resources so that this approach can be applied to them.

The resulting corpora will only have morphological annotation and will probably be severely unbalanced. However, development of such corpora constitutes a necessary step for introducing higher levels of annotation and for achieving better balance. Our ongoing experiments with OCRed Udmurt books suggest that adding a simple ngram-based postprocessor trained on corpus data may significantly improve its quality, reduce the cost of proofreading and thus eventually lead to adding books to the corpus. Finally, language models trained on such corpora enable other NLP tools for these other under-resourced languages (as an example, Yandex launched Udmurt-Russian machine translation service in 2016, which uses language model trained on the Udmurt corpus). This, in turn, can lead to preserving and revitalization of the minority languages.

References

Antworth, E. L. 1992. Glossing text with the PC-KIMMO morphological parser. Computers and the Humanities, 26(5-6):389-398.

Arkadyev, P. and Testelets, Ya. 2009. O trekh cheredovaniyakh v adygejskom yazyke [On three alternations in the Adyghe language]. Ya. Testelets (ed.), Aspekty polisintetizma [Aspects of polysynthesis]. 121-145.

Arkhangelskiy, T., Belyaev, O. and Vydrin, A. 2012. The creation of large-scaled annotated corpora of minority languages using UniParser and the EANC platform. Proceedings of COLING 2012: Posters, Ch. 9: 83-91.

Arkhangelskiy, T. and Lander, Yu. 2016. Developing a polysynthetic language corpus: problems and solutions. Computational Linguistics and Intellectual Technologies: papers from the Annual conference “Dialogue”.

Badmaeva, L. 2015. Natsionalnyj korpus buryatskogo yazyka: predposylki i perspektivnye puti razrabotki [Buryat National Corpus: prerequisites and future development trajectories]. Vestnik Buryatskogo gosudarstvennogo universiteta, 72.

Bradley, J. 2015. corpus.mari-language.com: A Rudimentary Corpus Searchable by Syntactic and Morphological Patterns. Septentrio Conference Series, 2:57-68.

Buskunbaeva, L. and Sirazitdinov, Z. 2011. Sistema razmetok v natsionalnom korpuse bashkirskogo yazyka [Annotation system in Bashkir National Corpus]. Proceedings of “Yazyki menshinstv v kompyuternykh tekhnologiyakh: opyt, zadachi i perspektivy”, Yoshkar-Ola: 46-51.

Goldberg, Y. and Elhadad, M. 2013. Word Segmentation, Unknown-word Resolution, and

Morphological Agreement in a Hebrew Parsing System. Computational Linguistics, 39/1:121-160.

Kilgarriff, A., and Grefenstette, G. 2003. Introduction to the special issue on the web as corpus. Computational linguistics, 29(3):333-347.

Kornai, A. 2013. Digital Language Death. PLoS ONE 8(10): e77056. doi:10.1371/journal.pone.0077056

Leontyev, N. 2014. Natsionalnyj korpus Internet-sajtov gazet na yakutskom yazyke [National corpus of newspaper web sites in Yakut]. Zhurnal nauchnyx i prikladnykh issledovaniy "Infinity", 4:35-36.

Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig (eds.). 2016. Ethnologue: Languages of the World, Nineteenth edition. SIL International, Dallas, Texas. Online version: http://www.ethnologue.com.

Moshagen, S. N., Pirinen, T. A., and Trosterud, T. 2013. Building an open-source development infrastructure for language technology projects. Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013), 343-352.

Orekhov, B. and Reshetnikov, K. 2014. K otsenke vikipedii kak lingvisticheskogo istochnika [Assessing Wikipedia as a linguistic source]. Sovremennyj russkiy yazyk v internete, Moscow: 310-321.

Salchak, A. and Bayirool, A. 2013. Elektronnyj korpus tuvinskogo yazyka: sostoyanie, problemy [Electronic corpus of Tuva: current state, challenges]. Mir nauki, kultury, obrazovaniya, 6 (43).

Segalovich, I. 2003 A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications. MLMTA'03. - Las Vegas: 273-280.

Sheymovich, A. 2011. Morfologicheskaya razmetka korpusa khakasskogo yazyka [Morphological annotation of the Khakas corpus]. Rossiyskaya tyurkologiya, 2(5):48-61.

Suleymanov, D., Khakimov, B., and Gilmullin, R. 2011. Korpus tatarskogo yazyka: kontseptualnye i lingvisticheskie aspekty [Tatar corpus: conceptual and linguistic aspects]. Filologiya i kultura, 26.

Winkler, E. 2001. Udmurt. Lincom Europa, München.

Zaliznyak, A. 1977. Grammaticheskiy slovar russkogo yazyka: slovoizmenenie [Grammatical dictionary of the Russian language: inflection]. Russkiy yazyk, Moscow.

Zaydelman, L., Krylova, I., Orekhov, B., Popov, I., and Stepanova, E. Russian minority languages: descriptive statistics. Computational Linguistics and Intellectual Technologies: papers from the Annual conference “Dialogue”.

Zheltov, P. 2015. Sozdanie natsionalnogo korpusa chuvashskogo yazyka: problemy i perspektivy [Development of Chuvash National Corpus: Challenges and perspectives]. Sovremennye problemy nauki i obrazovaniya, 1.