

Web Corpora of Volga-Kama Uralic Languages¹

Timofey Arkhangelskiy

This paper presents corpora of five minority Uralic languages that belong or are adjacent to the Volga-Kama area, which has been characterized as a Sprachbund (Bereczki 1983, Helinski 2003). A total of 11 corpora contain written and, in one case, spoken texts in Udmurt, Komi, Meadow Mari, Erzya and Moksha languages. The described resources are “web corpora” both in terms of their accessibility (all of them are accessible through a web-based query interface) and, in most cases, in terms of the medium (almost all texts come from web resources, such as digital newspapers and social media). The paper describes the corpora from the user perspective. The main focus is on the search capabilities and on certain research questions that can be studied with the help of these corpora. All corpora are available at <http://volgakama.web-corpora.net/>.

1 Introduction

Linguistic corpora as research tools and corpus linguistics as a methodology have experienced exponential growth since the 1990s. Multiple general-use reference corpora, as well as thousands smaller research-specific corpora, have been developed for major languages of the world. The Uralic family is no exception. For example, already in early 2000s there existed a number of large annotated corpora for Hungarian, such as the Hungarian National Corpus (Váradi 2002); somewhat smaller, but syntactically annotated Szeged corpus (Csendes et al. 2004); vast Hungarian web corpus (Halácsy et al. 2004); historical corpus (Pajzs 2000), etc. However, the minority Uralic languages spoken in Russia, even the largest and most vital ones, had a different fate. Until mid-2010s, only digital text collections of a limited size were created for some of them, e.g. by Suihkonen (1998), or small spoken corpora recorded by researchers in the field. First reasonably large publicly available written corpora for these languages only started appearing in 2014-2015, when the first versions of the literary Komi corpora (by the Syktyvkar-based FU-Lab team headed by Marina Fedina), the Udmurt corpus (by Maria Medvedeva and Timofey Arkhangelskiy) and Mari corpora (Bradley 2015) were created.

The corpora described in this paper were mostly developed in 2017-2019 by Timofey Arkhangelskiy with the purpose of filling this gap. The two exceptions are the “main” Udmurt corpus, which was started earlier in collaboration with Maria Medvedeva, and the spoken Udmurt corpus, which contains the data collected by Ekaterina Georgieva (see below). All corpora are available at <http://volgakama.web-corpora.net/>.

Since the languages in question share many properties such as some grammatical features or Cyrillic-based orthography, and have comparable level of digital presence, or digital vitality (Kornai 2016), similar methods and tools were used for developing the corpora. The vast majority of texts in all written corpora come from the web; my goal was to collect all or most texts written on the internet in the relevant languages. For each language, a rule-based morphological analyzer was developed; all of them are open source and can be found through the links in the respective corpus pages. Each analyzer contains a grammatical dictionary and a formalized description of the inflectional (as well as some

¹ The work is supported by RFBR grant 20-512-14003 ASCF_a “Linguistic diversity in the Volga-Kama region. Typology and language documentation between Volga and Urals”.

productive derivational) morphology. Since the analyzers are dictionary-based, not all words in the corpora will have a morphological analysis. Words which are not covered in the dictionary or that contain spelling mistakes or non-standard/dialectal affixes do not receive analyses. The proportion of analyzed words is different for different corpora and varies between 80% and 96%. Also, most analyzers do not take word's context into account. This leads to ambiguity, whereby each word receives all potentially possible analyses, even though only one of them is correct in the given context. For instance, an Erzya token *valdo* can in principle be analyzed either as the base form of the adjective *valdo* 'bright', or as the ablative of the word *val* 'word' (*val-do* word-ABL).² Without disambiguation, both analyses will be assigned to each *valdo* token in the entire corpus.

More detailed technical information about the corpus development process can be found in (Arkhangelskiy 2019).

2 Sources

For each language, two written corpora were created: a "main" corpus and a social media corpus. The latter contains texts from social media (*vkontakte*, which is the most popular social media platform in Russia, and, in some cases, forums), while the former contains all other digital texts. Other social media, such as Facebook, Twitter or Odnoklassniki, presumably contain far fewer posts in minority Uralic languages than *vkontakte*, and were not included at this stage.

The reason for this dichotomy is that linguistic properties of these two types of texts are so different that different processing pipelines and different metadata are required for them. One significant difference is code switching, which is ubiquitous on social media, but rather limited or nonexistent in other texts (even in blogs). As a consequence, the social media corpora contain sentence-level language tagging and offer an option of searching in Russian sentences written on pages that also contain Uralic posts. The number of misspellings and dialectal material is also higher in social media, which is why a slightly different approach was taken for tagging them. The social media corpora are generally smaller than their "main" counterparts and contain between 0.014 and 3.59 million words in the target languages (as well as several times more words in Russian). Their sizes are summarized in Table 2.

The "main" corpora mainly consist of contemporary digital press but include other digital texts as well. Table 1 presents the genre distribution in the five "main" corpora and their total sizes. The "other" column subsumes fiction, scientific papers, Bible translations, Wikipedia articles (filtered by quality), official texts and some other genres. Most texts in the corpora were written between 2010 and 2019, but there are some earlier texts as well.

Metadata for both kinds of corpora include year of creation (exact date in the case of newspaper articles), title and author (when known). The main corpora also contain genre metadata. The social media corpora contain information about relevant distinctions, e.g. whether the text was taken from a post or a comment, or whether it appeared on a group page or a personal page. Additionally, it includes sociolinguistic data about the

² The following abbreviations are used in the paper: 1 = first person, 2 = second person, ABL = ablative case, FUT = future tense, ILL = illative case, M = million, NOM = nominative case, NP = noun phrase, P = possessive suffix, PL = plural, SG = singular.

authors (in aggregated, non-identifying form) whenever the authors indicated them in their profile.

Language	size in words	press (%)	blogs (%)	other
Udmurt	9.57M	91.3%	5.1%	3.6%
Komi-Zyrian	1.75M	100%	0%	0%
Meadow Mari	2.63M	84%	0%	16%
Erzya	2.3M	67.4%	6%	26.6%
Moksha	1.74M	86.4%	0.7%	12.9%

Table 1: *Size and composition of the “main” corpora*

Language	size in words (Uralic part)	size in words (Russian part)
Udmurt	2.66M	9.83M
Komi-Zyrian	2.14M	16.12M
Meadow Mari	3.59M	15.1M
Erzya	0.83M	5.23M
Moksha	0.014M	0.17M

Table 2. *Size of the social media corpora*

Although the sizes of these corpora are several orders of magnitude smaller than those of e.g. contemporary Hungarian corpora, it is likely that the majority of digital texts available in these languages on the web has been included. A significant expansion of these corpora would necessarily require adding digitized texts from traditional media (books and newspapers), which requires a much higher level of time and resources.

The only spoken corpus so far contains transcribed Udmurt recordings made by Ekaterina Georgieva in several Udmurt dialects (Arkhangelskiy and Georgieva 2018). Although very different in its size and composition from the rest, it was processed using approximately the same pipeline and published through the same search interface as the other corpora.

3 Search capabilities

For the linguistic data to be reusable, it is crucial that they come with a tool that allows for complex search queries. As an example, the literary Komi corpus by FU-Lab, which is amazing in terms of its contents (over 50 million words of texts in a variety of genres, spanning almost a century), only allows very basic search requests, and therefore is difficult to use in some kinds of research.

All corpora described in this paper are published through the *tsakorpus* search platform that I started developing in 2017 and maintain now.³ When developing it, I had several primary objectives:

- Provide an intuitive user interface that would allow complex linguistic queries without the need to learn a full-fledged query language such as CQP, used in Corpus Workbench (Evert and Hardie 2011), or AQL, used in ANNIS (Krause 2019).

³ <https://bitbucket.org/tsakorpus/tsakorpus>

- Treat various corpus types (written, sound-aligned, parallel etc.) in a uniform way.
- Make sure the platform is fast enough to enable even sophisticated queries on mid-sized corpora (1–100 million words) with heavy annotation.
- Make the platform ambiguity-friendly. When it comes to POS tagging, it is assumed in most corpora of major languages that each analyzed word can have exactly one analysis. It might indeed be possible to choose one analysis out of several theoretically correct ones based on the context with very high precision, e.g. using neural networks trained on large manually tagged datasets, for major languages. However, for under-resourced languages this is usually not the case. Since there are no such datasets for them, any kind of statistical analyzer that only leaves one analysis for each word will make too many mistakes. Even with a 5% error rate the linguist risks not being able to find many relevant, but incorrectly tagged examples. Keeping ambiguous analyses makes the linguist's work more time-consuming, but reduces the chances of missing something important in the data.

The tsakorpus platform is open-source and language-independent. Since its creation, it has been used in a number of projects other than the one described here, e.g. INEL Selkup corpus (Brykina et al. 2020; <https://inel.corpora.uni-hamburg.de/SelkupCorpus/search>), Spoken corpus of Khakas (Maltseva and Sokur 2020, https://linghub.ru/oral_khakas_corpus/), or Bashkir National Corpus (<http://bashcorpus.ru/>). The search interface is available in English and Russian.

There is a concise description of the search functionality in the Help window in each corpus (orange question mark at the top of the page). Instead of listing individual features, I will now describe a single research question that requires building a rather complex query, to demonstrate the capabilities of the platform. Udmurt Social media corpus will be taken as an example; the same search functionality is available in all other corpora (although the grammatical tags are language-specific).

Just as in other Volga-Kama languages, most spatial relations in Udmurt are expressed by inflected postpositions, or relational nouns, which have a nominal or pronominal dependent. In Standard Udmurt, the only available construction of this kind requires the dependent to be in the nominative and not cross-referenced on the head, as in Example 1. This is prescribed in most grammars and textbooks. However, there are other options available in the dialects. In one of them, 1st and 2nd person pronominal dependents are still in the nominative, but trigger appropriate possessive marking on the head, as in Example 2 (which is highly unusual for an Udmurt NP). This option has been mentioned in the grammar by Winkler (2011) without any remarks about its dialectal nature; other than that, it is unknown where exactly and why this construction exists.

(1) *mon dor-i*
 I.NOM at-ILL
 ‘towards me / to my place’

(2) *mon dor-a-m*
 I.NOM at-ILL-P.1SG
 ‘towards me / to my place’

Since the social media corpus contains geographical metadata (place of birth and current location) for some authors, it would make sense to search the second construction and see whether its approximate areal distribution can be established.

Here is how an appropriate search request can be built in the web interface:

- By default, tsakorpus shows one block of search fields that corresponds to one search term. Since the construction in question involves two words, a second block should be added by clicking the plus sign (“add word”) in the right-side pane of the first block.

- If your search includes multiple words, the default behavior is to find all sentences that include all of them regardless of their mutual order or distance. Since we want the first word to be located immediately to the left of the second, a distance requirement has to be added. This is done by clicking the “add distance” button (two arrows pointing in opposite directions) in the second block. The default values (distance of at least 1 word and at most 1 word from the word #1) describe exactly the scenario that we need.

- The first word, i.e. the dependent, has to be a personal pronoun of first or second person. The easiest way to specify this constraint is to list all four possible variants in the Lemma field or in the Word field.⁴ The expression that has to be put there is *мон|мон|му|мӱ*. The pipe symbol stands for logical OR in the Word, Lemma and Grammar fields; the words separated by it are the lemmata of the Udmurt 1SG, 2SG, 1PL and 2PL pronouns, respectively. Putting this string in the Word field means that the first word in the construction must coincide exactly with one of these four options. Since in the case of pronouns, the lemma coincides with the nominative form, this will be sufficient for our purposes. If, instead of that, this expression is pasted in the Lemma field, by default it means that all forms of these four pronouns must be found. In our case, we would have to additionally specify that only the nominative has to be found by typing *nom* in the Grammar field. The *nom* tag stands for the nominative (or, in the case of nouns, unmarked accusative); the entire tagset, i.e. the list of grammatical tags used in the corpus, can be found at the start page of each corpus. Instead of typing, the values can also be selected from a pop-up window that appears after clicking the button at the right end of the Grammar field. The two methods (putting the pronouns in the Word field or putting them in the Lemma field while specifying their case) may look the same; nevertheless, the latter yields more precise results. The reason for that is that some frequent misspellings, such as missing diacritics in *мӱ* you.PL.NOM, are handled correctly by the analyzer. Since the misspelled word *mu* will be found by the lemma+case query, but missed by the word query, the lemma+case query is preferable in the case of noisy texts.



- The second word can be any relational noun with a 1st or 2nd person possessive suffix. Additionally, we will limit the search to the three most frequent spatial cases that relational nouns combine with: locative (inessive), illative and elative. This constraint can be set by putting the following expression in the Grammar field of the second block: *rel_n,(1sg|2sg|1pl|2pl),(loc|el|ill)*. Again, the pipe symbol stands for the logical OR; comma stands for AND, and parentheses are used for grouping.

- Finally, a metadata constraint has to be added to narrow down the search. In tsakorpus, two kinds of metadata are distinguished. The first kind is text-level metadata, such as title, author, or creation year of the text. Their values can be used for limiting the search to a subset of corpus texts, e.g. all texts written by a certain author, by clicking the “Select subcorpus” button. The second kind is the sentence-level metadata, which pertain to individual sentences. In the case of social media corpora, sentence-level metadata contain the information about the author of each particular sentence or post, while text-level metadata refer to the owner of the page where that post was written. Since we are

⁴ I am omitting the 1pl inclusive pronoun (Maksimov and Panina 2018), which coincides with a possessive form of the reflexive pronoun, because it behaves differently in this respect.

interested in the areal distribution of the phenomenon in question, only those sentences are relevant for which the author’s place of birth (which is an approximation of their dialect) is known. Since only a minority of users indicate their birth place in their profile, the “non-empty birth place” requirement will cut off many irrelevant search hits and thus save the researcher’s time.


Sentence-level metadata requirements can be set by clicking a downwards arrow in any of the two blocks. In our case, the “Account type (post-level)” field should be set to *user*, so that posts authored by groups are excluded. The “Birth place (post-level)” has to be set to $\sim(\textit{unknown}|\textit{other})$, where \sim stands for negation. This expression will cut off sentences written by users whose birth place is either not indicated (which is expressed by the value of *unknown* in the corpora), or indicated, but not recognized by the geographical classifier at annotation time (the value of *other*).


Udmurt social media corpus RU | EN |  

Word #1

Word:

Lemma:

Grammar: 

Gloss: 

Translation (ru):

2nd lemma:

2nd transl. (ru):

Metafield_author_id_post:

Year (post-level):

Sex (post-level):


Current place (post-level):

Birth place (post-level):


Birth year (post-level):

Account type (post-level):

Post type:

Analyses: 


Position in sentence:


Language/tier: 


Word #2

Word:

Lemma:

Grammar: 

Gloss: 

Language/tier: 

Distance to word #

from

to

Full-text search: Precise match

Search sentences
Search words / lemmata
Select subcorpus

Figure 1: Search query in Tsakorpus interface of the Udmurt social media corpus

Clicking “Search sentences” will yield a number of search hits (21 as of May 2020), where the construction in question is highlighted. The examples are sorted randomly. First, this prevents the user from reconstructing the entire text, which would be a copyright violation. Second, in the case of a large number of results, the user can easily see how the construction in question behaves on average by looking at the first 100 or 200 sentences, for which it is crucial to have an unbiased sample.

The final step is going through the sentences found and assessing them manually. As it almost always happens, only a part of the search hits contain the construction that is

being looked for. For instance, the sentence in (3) technically conforms to the query. However, the pronoun there is the subject rather than the dependent of the relational noun, which has no overt dependent:

- (3) *Berjtsk-o-d* *ton* *dor-a-m.*
 return-FUT-2SG you.SG.NOM at-ILL-P.1SG
 ‘You will return to me.’

After sifting through the hits, we find that only 5 sentences make it to the final list of genuine examples. Sentence-level metadata for each of them can be seen in the upper right corner when hovering the mouse pointer over the sentence.

4 Social media corpora and dialectology

The corpora presented here can be used for researching a number of topics in the areas of lexicography, morphology and syntax. However, the metadata in the social media corpora make it possible to conduct research on sociolinguistics and dialectology. This prospect seems especially important to me, since these disciplines have not benefited from corpora as much as other areas of linguistics. Besides, dialectological research with its fieldwork in multiple locations is a very expensive and time-consuming undertaking. Therefore, it is important to know to which extent social media data can be used to learn about areal distributions of words and grammatical phenomena.

As I have demonstrated elsewhere (Arkhangelskiy 2019), the social media corpora can be used in studies of dialectal vocabulary. By comparing the data extracted from social media corpora with the results of traditional dialectological surveys, I showed that although corpus data does not provide enough information on some varieties, the information it does provide does not contradict the facts established by traditional dialectology. Therefore, social media corpora can be used as incomplete, but relatively reliable sources of dialectological data. As such, they can be used in preliminary studies, e.g. when planning dialectological fieldwork.

Since Uralic dialectology has paid much more attention to phonology and vocabulary than to morphosyntax, relatively little is known about dialectal distribution of syntactic constructions such as the one described in Section 4. Social media corpora could prove a great help here. The examples of the non-standard construction found in the corpus belong to the authors born in Igra and Sharkan districts, which allows us to very roughly outline the area where this phenomenon exists. My preliminary fieldwork shows that it indeed exists there, while being either infrequent or altogether nonexistent elsewhere.

5 Future work

The corpora described in this paper were last updated in 2018–2019. In order to keep them up to date, I am working on a semi-automatic pipeline that would make it easy to add new texts from social media, blogs and newspapers each 6 months. Geographical metadata has to be added to the social media corpora to enable the dialectological research described above; right now, it is only available in Udmurt and Meadow Mari (to a certain extent) corpora. Another direction of improvement is the functionality of the search platform; I

expect the next major release to be ready in late 2020. Finally, I am collaborating with other teams who have spoken corpora of Volga-Kama Uralic languages in order to make them available through tsakorpus and provide the functionality necessary for searching them. At the moment, this includes a spoken Meadow Mari corpus (Anna Volkova, Aigul Zakirova, Linguistic Convergence Laboratory at Higher School of Economics); I will be happy to collaborate with other researchers and teams as well.

6 Conclusion

I have presented 11 corpora of five Uralic languages of the Volga-Kama area. All of them have morphological annotation and are publicly available through a web interface. These corpora can be used in various kinds of linguistic research, such as lexicography, morphology and syntax. Additionally, the social media corpora may be used in studies of sociolinguistics and dialectology. I hope that these corpora will help linguists who specialize in these under-resourced Uralic languages and boost the research on them.

References

- Arkhangelskiy, Timofey. 2019. Corpora of social media in minority Uralic languages. In Tommi A. Pirinen, Heiki-Jaan Kaalep & Francis M. Tyers (eds.), *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, 125–140. Tartu: Association for Computational Linguistics. <https://doi.org/10.18653/v1/w19-0311>
- Arkhangelskiy, Timofey & Georgieva, Ekaterina. 2018. Sound-aligned corpus of Udmurt dialectal texts. In Tommi A. Pirinen, Michael Riebler, Jack Rueter, Trond Trosterud & Francis M. Tyers (eds.), *Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages*, 26–38. Helsinki: Association for Computational Linguistics. <https://doi.org/10.18653/v1/w18-0203>
- Berezki, Gábor. 1983. A Volga-Káma vidék nyelveinek areális kapcsolatai. In: Balázs János (ed.), *Areális nyelvészeti tanulmányok*, 207–236. Budapest: Tankönyvkiadó.
- Bradley, Jeremy. 2015. Corpus.mari-language.com: A Rudimentary Corpus Searchable by Syntactic and Morphological Patterns. In *Proceedings of the First international workshop on computational linguistics for Uralic languages*. Septentrio Conference Series. Tromsø: Septentrio Academic Publishing. <https://doi.org/10.7557/5.3468>
- Brykina, Maria, Orlova, Svetlana & Wagner-Nagy, Beáta. 2020. INEL Selkup Corpus. Version 1.0. Publication date 2020-06-16. Archived in Hamburger Zentrum für Sprachkorpora. <http://hdl.handle.net/11022/0000-0007-CAE5-3>. In Wagner-Nagy Beáta; Alexandre Arkhipov, Anne Ferger; Daniel Jettka & Timm Lehmberg (eds.), *The INEL corpora of indigenous Northern Eurasian languages*.
- Csendes, Dóra, Csirik, János & Gyimóthy, Tibor. 2004. The Szeged Corpus: A POS Tagged and Syntactically Annotated Hungarian Natural Language Corpus. In P. Sojka, I. Kopeček & K. Pala (eds.), *Text, Speech and Dialogue*, 41–47. Berlin/Heidelberg: Springer. https://doi.org/10.1007/978-3-540-30120-2_6
- Evert, Stephan & Hardie, Andrew. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 Conference*. Birmingham, UK.

- Halácsy, Péter, Kornai, András, Németh, László, Rung, András, Szakadát, István & Trón, Viktor. 2004. Creating open language resources for Hungarian. In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa & Raquel Silva (eds.), *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, 203–210. European Language Resources Association. Lisbon, Portugal.
- Helmski, Eugene. 2003. Areal groupings (Sprachbünde) within and across the borders of the Uralic language family: A survey. *Nyelvtudományi Közlemények* 100. 156–167.
- Kornai, András. 2016. Computational linguistics of borderline vital languages in the Uralic family. In Tommi A. Pirinen, Simon Eszter, Francis M. Tyers & Vincze Veronika (eds.), *Proceedings of the Second International Workshop on Computational Linguistics for Uralic Languages*. Szeged: Szegedi Tudományegyetem. (Available online at <http://kornai.com/Drafts/iwclul.pdf>, accessed on 05.11.2018.)
- Krause, Thomas, 2019. ANNIS: A graph-based query system for deeply annotated text corpora. Humboldt-Universität zu Berlin, PhD thesis.
- Maksimov, Sergey & Panina, Tatjana. 2018. On the category of clusivity in the Udmurt language. *Linguistica Uralica* 54(3), 213–224. <https://doi.org/10.3176/lu.2018.3.05>
- Maltseva, Vera & Sokur, Elena. *Spoken corpus of the dialects of Khakas*. Moscow: Institute of Linguistics; Moscow: Linguistic Convergence Laboratory, NRU HSE. (Available online at https://linghub.ru/oral_khakas_corpus/, accessed on 04.08.2020.)
- Pajzs, Júlia. 2000. Making Historical Dictionaries with the Computer. In Ulrich Heid, Stefan Evert, Egbert Lehmann & Christian Rohrer (eds.), *Proceedings of EURALEX 2000*, 249–259. Stuttgart: Universität Stuttgart.
- Suihkonen, Pirkko Marjatta. 1998. *Documentation of the Computer Corpora of Uralic Languages at the University of Helsinki*. Helsinki: Department of General Linguistics, University of Helsinki. Technical paper.
- Váradi, Tamás. 2002. The Hungarian National Corpus. In Manuel González Rodríguez, Carmen Paz Suarez Araujo (eds.), *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, 385–389. Las Palmas, Canary Islands, Spain.
- Winkler, Eberhard. 2011. *Udmurtische Grammatik* (Veröffentlichungen der Societas Uralo-Altaica 81). Wiesbaden: Harrassowitz Verlag.

Timofey Arkhangelskiy
Universität Hamburg
timarkh@gmail.com