On the Role of New Technologies in the Documentation and Revitalization of Uralic Languages of Russia in Historical and Contemporary Contexts

Alexander Nazarenko
Independent researcher and enthusiast of Uralic languages,
amateur database and software developer
aleksanterinazarenko@gmail.com

Abstract

The Uralic languages spoken in Russia face significant challenges due to historical and sociopolitical factors, resulting in their endangered status. While only Finnish, Estonian, and Hungarian enjoy solid support as official languages, most Uralic languages suffer from limited resources and declining speaker populations. This paper examines the development of written Uralic languages, the impact of Russian language and its writing system to them, and the consequences of the lack of state interest in these languages for preservation efforts. Despite these challenges, technological advancements present valuable opportunities for revitalization. Existing projects, such as dictionaries and language corpora, highlight both the potential and shortcomings of current linguistic resources. Innovative approaches, including AI-based applications and user-driven platforms, can enhance engagement among people. By emphasizing the importance of high-quality linguistic data, this study advocates for a more proactive and collaborative effort in the preservation and promotion of Uralic languages.

1 Outline of the Problem

Only three Uralic language-speaking nations have succeeded in establishing their own states where the Uralic language has official status and is utilized in all aspects of life. The emergence of autonomous and independent political entities facilitated the development of sophisticated literary languages and supported the establishment of strong national identities and diverse cultures. Regrettably, the fate of other Uralic peoples has been less fortunate.

The earliest texts written in Finnish and Estonian are relatively late compared to many European languages; moreover, they are not significantly later than the early texts of Uralic languages spoken in Russia. It is known that the Komi people had their own writing system, *Važ Perym gižöm*, as early as the 15th century, which was used until the 18th century. The Old Permic script, like Cyrillic, is not entirely original; it incorporates features from Cyrillic, Greek, and Komi tamga signs, which are ultimately of Turkic descent. However, it emerged at a time when, according to current knowledge, no other Uralic peoples, except for the Hungarians, had any writing system. The first known sentence in the Baltic-Finnic languages was written in Cyrillic in the 13th century (*Birch bark letter no. 292*), while subsequent sentences date back to the second half of the 15th century and were written using the Latin script.

In most cases, the formation and development of written Uralic languages and their writing systems can be attributed to the expansion of various branches of Christianity, and their nature is closely tied to the church. Nevertheless, the influence of primary languages, such as Russian, and their characteristics, including writing and phonetic systems, was often more substantial in later stages. Essentially, most modern spellings of Uralic languages are based on the Russian variant of the Cyrillic script, which was designed and adapted to meet the needs of Russian and other Slavic languages but may not necessarily fulfill the requirements of the Uralic languages. Let's take a closer look at some of these impacts.

One example of this is the modern Moksha spelling, where the sound /ə/ can be represented by the letters а, о, е and the sound /æ/ can be represented by the letters я, е, э, depending on various factors. This variability can cause problems even for people who are familiar with the language. Another example is the Erzya sound /æ/, which has been entirely eliminated from the standard language, despite its presence in many dialects where it serves to differentiate meanings. For instance, /ˈkedʲ/ ('hand') (< Proto-Mordvinic **käd'*) and /ˈkædʲ/ ('skin') (< Proto-Mordvinic **ked'*). In the modern language created in 1922, the word кедь has been used for both meanings, likely due

Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages, pages 109–114

to the desire to simplify the language norms and maximize their alignment with Russian language standards. It is also known that, with the Bolshevik takeover, the idea of creating a common literary language for the Erzya and Moksha people was conceived, which ultimately ended in failure, as these two languages are not mutually intelligible and do not even form a clear continuum.

Texts composed in the 19th century and earlier display a relatively high level of linguistic quality, which is due to the fact that their authors were proficient in the languages, and the influence of Russian, particularly on grammar, was comparatively moderate during this period. The writing systems of that time vary, yet it is generally evident that an effort was made to create Cyrillic-based systems that prioritized phonetic accuracy. Compare the first three verses of Nikolai Barsov's Moksha translation of the Gospel of John from 1901:

> 1. Первай ульсь Вал, Валськæ Шкайсъль, Шкайськæ Валъль.
>
> 2. Сон первай кигæ Шкайсъль.
>
> 3. Сонь вельденза сембæ ушъдъзь улемаснън, а Соньфтемънза улемаснън ущъдыхьнень ёткста мезямътка исьть ушъдуфт.

with the same fragment, translated by Institute for Bible translation Helsinki, in 2003:

> 1. Ушеткссa ульсь Валсь, Валсь Шкайть мархтоль, и Валсь ульсь Шкай.
>
> 2. Сон ушеткссстокиге Шкайть мархтоль.
>
> 3. Сембось вельденза тиевсь, Соньфтомонза мезевок ашезь тиев.

The system used by Barsov has a distinct letter æ for the sound /æ/ and consistently employs the letter ъ to represent the sound /ə/ in all cases, ensuring that each sound corresponds to a specific letter. In contrast, in the modern language, the letter ъ no longer serves that function. The 1993 language reform was intended to reintroduce its use at the beginning and in the first syllable of words, for example, въръгаз /vərʲˈgaz/ ('wolf') (< Proto-Mordvinic **vərgas*) and търва /təˈrva/ (< Proto-Mordvinic **tərva*) instead of врьгаз, трва, but it was rejected.

There were also quite a few attempts to create a Latin script for Uralic languages, which are now written exclusively in Cyrillic. An Estonian linguist, Ferdinand Johann Wiedemann, possibly drawing on the earlier work of Hans Conon von der Gabelentz, *Versuch einer Mordwinischen Grammatik*, published in 1839, used a Latin writing system for the Erzya language in his work *Grammatik Der Ersa-Mordwinischen Sprache* and in his transliteration of the *Gospel of Matthew* (*Das Evangelium des Matthäus ersamordwinisch*) from 1865. Although this system is somewhat irregular, it represents a clear initiative to establish a Latin script for the language. Below is a short example from the 22nd chapter of the Gospel:

> *37. Jisus jovtaź tenze: vetškik es pazot vese sädeiset i vese oimset i vese prävset toń.*
>
> *38. Te uli ikel'tse i vesemede pokš zapoved.*
>
> *39. Ombotse že teń kond'amo: vetškik es malavikset koda es pŕat.*

Note that *Novum Testamentum Mordvinice litt. cyrill.* from 1821, which served as the basis for Wiedemann's adaptation, contains some characteristics of the ä-dialects.

In spite of these efforts, all literary Uralic languages of Russia and their respective writing systems were established only during the Soviet era. Many native words and structures, as well as widely occurring dialectal features, were discarded. Simultaneously, a vast number of words and structures were borrowed from Russian, completely overlooking the possibility of creating new words based on existing ones, as was done in Finnish in the 18th and 19th centuries. Compulsory education was also introduced, delivered almost entirely in Russian, apart from a brief period of "Korenizatsiia" in the 1920s and early 1930s. The introduction and implementation of new writing systems, particularly those based on the Latin alphabet, were abandoned in the 1930s and have been legally prohibited since the early 2000s.

Currently, all Uralic languages, except for Finnish, Estonian, and Hungarian, are considered endangered, and the speaker populations in Russia are experiencing a dramatic decline each year. The actual situation may be even more concerning than the statistics suggest. Proficiency in these languages among individual minority groups is low

and continues to decline. The ongoing advancement of technology and media has further diminished the role of smaller Uralic languages and contributed to the deeper assimilation of their speakers. Most users of these languages are older individuals who may not be well-versed in modern technology, which somewhat slows the process of assimilation, but conversely, means that the limited language resources available do not effectively meet their needs. Younger generations, despite having easier access to technology, often show little interest in these languages, and existing solutions fail to counter this trend.

However, technological evolution presents a valuable opportunity to address the challenges faced by these languages. The collection, organization, and visual presentation of high-quality linguistic data would not only allow enthusiastic individuals to learn and study independently but also bolster the efforts of scientists, researchers, teachers, and activists concerned about the future of endangered languages, thereby raising public awareness of their fate.

2 Main challenges in Light of Existing Solutions

The main issue in the collection and digitization of data related to smaller Uralic languages lies in its significant dispersion and inconsistency, along with the limited quantity and quality of available materials. Therefore, existing databases frequently suffer from inaccuracies, incompleteness, and a lack of refinement that makes them less suitable for interactive language projects. A constructive approach could involve enhancing and refining these databases through qualitative improvements, such as incorporating native vocabulary and grammatical structures. This would include integrating archaic and less commonly used terms, creating new words based on the existing linguistic material, and minimizing reliance on Russian loanwords whenever possible. Sometimes it leads to a dilemma between prioritizing the 'accuracy' of language data and the mostly subjective concept of language purity, which might not always align with the preferences of native speakers.

Another issue is that the materials essential for learning and researching Uralic languages are primarily in Russian, which makes them hard to analyze directly for non-Russian speakers. The quantity of materials available in English and the three major Uralic languages is decidedly insufficient, likely due to the fact that translation and direct data collection is time- and money-consuming.

Despite these challenges, several remarkable projects have emerged, including:

- dictionaries created by The Institute of the Estonian Language (*Eesti Keele Instituut*)[1]
- the Giellatekno dictionaries and Oahpa! tools managed at UiT The Arctic University of Norway (*UiT Norges arktiske universitet*)[2]
- Korp and other text corpora available in The Language Bank of Finland (*Kielipankki*)[3]
- dictionaries and materials created by the Institute for the Languages of Finland (*Kotimaisten kielten keskus*, *Kotus*)[4]
- Udmurt and Komi languages in Google Translate[5]

This list is not exhaustive; however, it effectively illustrates the landscape. Let's take a brief look at some of the projects mentioned above. The dictionaries presented by The Institute of the Estonian Language are notable for their relatively extensive vocabulary, numerous examples, and a clear effort to find suitable equivalents for terms missing in the target languages. This was achieved in part by assigning new meanings to words with closely related meanings, creating calques, and adding "descriptive equivalents", such as the Erzya word ардомапель ('vehicle'), derived from ардомс ('to go, travel') and the suffix -пель (used to form object names). While the dictionaries remain a highly reliable and innovative resource, their audience is understandably restricted to Estonian speakers.

The absence of corresponding dictionaries in the reverse direction, the lack of direct references to sources and literary examples, and the omission of transliteration are common issues found in many online dictionaries. The dictionaries available on the UiT The Arctic University of Norway website, while allowing bidirectional translations and containing an impressive amount of collected material and fairly extensive grammatical data, almost entirely lack usage examples for individual words

[1] https://eki.ee/keeleinfo/sonastikud/
[2] https://dicts.uit.no/
[3] https://www.kielipankki.fi/korp/
[4] https://www.kotus.fi/sanakirjat
[5] https://translate.google.com/

in the English section, and some translations may even appear rather unusual. On a positive note, some provide very interesting alternatives to the increasingly common Russicisms, even though they lack any confirmation in literature.

It is necessary to emphasize that citing sources is absolutely critical in the development of linguistic resources. Providing information about the authors of cited language data and innovations, such as neologisms, enhances the reliability of the documented information.

The role of language corpora in the documentation of endangered languages is indispensable, but they are also invaluable in the creation of language projects, such as online dictionaries and learning tools. According to the list on the FID FINNUG site[6], the Korp platform on the Language Bank of Finland website is the only tool that allows for the simultaneous display of the same texts in multiple Uralic languages, thus considerably facilitating their interpretation and comparative analysis. The size of the database and the number of available languages are distinctly unique within their field.

The support of Udmurt and Komi on Google Translate deserves special mention. Although the quality of the translation is not perfect (e.g. the phrase *Good night!* is translated into Udmurt as Бур уен! instead of the correct Ӟеч кӧл! or Ӟеч кӧлэ!), this is undoubtedly a significant step towards promoting these languages. Hopefully, in time, translation into other Uralic languages will be launched, with a particular emphasis on restoring their original forms.

For the purpose of comparison, the following is a list of several notable projects that have been developed in Russia:

- MarlaMuter Mari-Russian and Erzya-Russian dictionaries[7]
- FU-Lab dictionaries, primarily focused on Permic and Mari languages[8]
- Sámi dictionaries, Saamskije slovari (Саамские словари)[9]
- the Open corpus of Veps and Karelian languages VepKar[10]
- the Erzya corpus[11] and the Erzya social media corpus[12]
- the Moksha corpus[13]
- the National Corpus of the Udmurt Language (Национальный корпус удмуртского языка) with a dictionary[14]
- the LANGO.TO translator, which supports Erzya, Finnish and Estonian languages[15]

Comparing projects developed outside of Russia with those emerging within the country, it is regrettable to acknowledge that, in certain respects, the former demonstrate higher quality. This disparity is not surprising, as these projects often rely on existing works in Russian and are, in fact, digitized versions of books without any modifications. For example, MarlaMuter includes five digitized Mari dictionaries, an Erzya-Russian dictionary, and offers very useful features such as the ability to report typographical errors and buttons corresponding to letters with diacritical marks not present in the Russian alphabet. Additionally, it provides interfaces in both English and Russian.

The FU-Lab website contains 42 digitized dictionaries primarily focusing on Permic languages. Despite the extensive amount of gathered data, has a somewhat chaotic structure and lacks an interface in any language other than Russian, presenting an additional obstacle for individuals outside of Russia wish to study these languages. A similar issue is found with the Sámi dictionaries website.

VepKar, or the Open Corpus of Veps and Karelian Languages, is an example of a well-constructed website with many valuable materials, such as a speech corpus, an audio map with recordings, and a corpus-based dictionary that provides information about the specific region of Karelia from which each word originates, along with grammatical categories and relevant examples accompanied by Russian translations.

One of the advantages of the Erzya language corpora site is its capability for automatic transliteration of text according to the Uralic Phonetic Alphabet. It contains an extensive collection of

[6] https://fid.finnug.de/en/language-corpora/
[7] https://marlamuter.com/en/
[8] https://dict.fu-lab.ru/
[9] https://slovari.saami.su/
[10] http://dictorpus.krc.karelia.ru/en
[11] https://erzya.web-corpora.net/index_en.html
[12] https://erzya.web-corpora.net/erzya_social_media
[13] https://moksha.web-corpora.net/index_en.html
[14] https://udmcorpus.udman.ru/
[15] https://lango.to/

linguistic material, including, importantly, examples of colloquial language used in contexts such as online forums. It also includes translations for most terms in Russian.

The site of National Corpus of the Udmurt Language, in turn, includes an autonomous Russian-Udmurt and Udmurt-Russian dictionary, featuring usage examples and some audio recordings of pronunciations. This addition certainly enhances the usability of the corpus. A drawback is once again the lack of transcription and an English interface.

Finally, LANGO.TO offers an effective AI-based translator for the Erzya language and several other non-Uralic minority languages of Russia. It would not be overstatement to say that this represents one of the more intriguing initiatives of recent years, as AI has not been widely applied to the revitalization of endangered Uralic languages. The accuracy of translations between Russian and Erzya is quite impressive, especially considering the limited resources and the relatively undeveloped state of the language. In addition to Russian, it also supports Finnish and Estonian languages.

3 Summary and Example Solutions

Projects aimed at documenting and revitalizing endangered languages should, on one hand, include as much data as possible and reference specific sources, while, on the other hand, analyzing this data in terms of its quality and usefulness, and supplementing it with new information, such as grammatical categories, inflection, and usage examples. Websites and applications should feature a simple and accessible interface, offer multiple language versions, provide translations into English and major Uralic languages, and include transliterations or phonetic transcription for languages using the Cyrillic alphabet. Pronunciation recordings are invaluable for preserving the original pronunciation of the languages. Introducing new solutions, including experimental ones, with a particular emphasis on AI, is essential. At the same time, it is worth exploring how existing language corpora and "raw" lexical databases, such as those available on the Giellatekno Webdict[16], can be effectively utilized.

The implementation of the data does not have to be a difficult task, as demonstrated by my website Learn Erzya[17], where I utilized some databases from the aforementioned Giellatekno for testing purposes. I also incorporated an alternative Latin spelling of the Erzya language, as presented in the book by linguists Ksenija Djordjević and Jean-Léo Léonard, *Parlons mordve: erzya et mokša*, with minor modifications. The dictionaries feature a switch between the Latin spelling and the original Cyrillic spelling. Ultimately, I intend to replace these with databases containing verified data and additional elements such as phonetic transcription, transliteration, grammatical categories, automatic inflection, common phrases, usage examples, and examples from literature, possibly sourced directly from corpora.

The experimental transliteration of the databases from Giellatekno was carried out using a transliteration tool[18] that was partly based on the one the transliteration modules used in Wiktionary. The optimization of the code was facilitated largely by ChatGPT. It also has a phonetic variant[19].

One more initiative underway is creating language maps. Currently, there are three simple websites featuring maps that display specific words in various European[20], Uralic[21], and Mordvinic languages[22]. The latter pulls data directly from Wikisource. Moreover, it would be an exciting prospect to create a similar map using data from *The Dialect Dictionary of the Mordvin Languages Based on the Heikki Paasonen Materials*.

Another interesting option is creating open databases using MediaWiki.[23]

In summary, although the situation of the smaller Uralic languages is, to put it mildly, far from ideal, we are equipped with tools today that offer us almost limitless possibilities. The accumulation of accumulated knowledge and technological resources at our disposal is unprecedented, yet much of its potential remains untapped. Many sources are awaiting digitization and thorough analysis, without which the development of interactive tools is not possible. Of course, this is also a matter of

[16] https://gtweb.uit.no/webdict/

[17] https://aleksanterinazarenko.github.io/learnerzya/

[18] https://aleksanterinazarenko.github.io/learnerzya/transliteration-tool.html

[19] https://aleksanterinazarenko.github.io/transliterator/

[20] https://aleksanterinazarenko.github.io/interactivemap-europe/

[21] https://aleksanterinazarenko.github.io/interactivemap-uralic/

[22] https://aleksanterinazarenko.github.io/interactivemap-mordvinic-wiktionary/

[23] https://uralowiki.unaux.com/index.php/Main_Page?i=1

funding, which is allocated to these goals in a very limited capacity, shifting the entire burden onto enthusiasts and amateurs, along with their financial and time constraints. This has a direct impact on the results. Nevertheless, even this barrier can be overcome if social awareness and engagement are increased, and the only way to achieve this is by providing concrete, ready-made solutions. What has contributed to the decline of the smaller Uralic languages should be used as an instrument for their revitalization. The role of new technology in this process is not only underestimated but is also absolutely crucial.

Acknowledgments

Acknowledgements to Jack Rueter for his valuable comments on the article.

References

1821. *Novum Testamentum Mordvinice litt. cyrill.* Soc. bibl. Russica, Saint Petersburg.

2002. Law on the unified graphic basis. Federal Law of the Russian Federation No. 165-ФЗ, enacted on 11.12.2002.

2016. *The New Testament in the Mordvin-Moksha language*. Institute for Bible translation Helsinki, Helsinki.

N. P. Barsov. 1901. *Ot Ioanna svjatoe evangelie*. Helsinki.

A. P. Feoktistov. 2008. *Očerki po istorii formirovanija mordovskix pis'menno-literaturnyx jazykov*. Saransk.

Herr Conon Gabelentz. 1839. Versuch einer mordwinischen grammatik. *Zeitschrift für die Kunde des Morgenlandes*, 2:235–419.

J. S. Jelisejev. 1961. Vanhin itämerensuomalainen kielenmuistomerkki. *Virittäjä-lehti*, 65(1):134.

A. N. Kelina. 2003. *Mordovija. Ènciklopedija*, volume 1. Mordovskoe kn. izd-vo, Saransk.

A.N. Kelina and O.E. Poljakov. 2024. *Orfoèpičeskij slovar' mokšanskogo jazyka*. Saransk.

László Keresztes. 1986. *Geschichte Des Mordwinischen Konsonantismus*. Szeged.

E.A. Kosminskij. 1943. *Srednej pingetnen' istorijas'*. Mordovskoj gosudarstvennoj izdatel'stvas', Moscow.

V. I. Lytkin. 1952. *Drevnepermskij jazyk: čtenie tekstov, grammatika, slovar'*. Moscow.

Rein Taagepera. 2013. *The Finno-Ugric Republics and the Russian State*. Routledge, New York.

Ferdinand Johann Wiedemann. 1865a. *Gospel of Matthew (Das Evangelium des Matthäus ersamordwinisch)*. London.

Ferdinand Johann Wiedemann. 1865b. Grammatik der ersa-mordwinischen sprache: nebst einem kleinen mordwinisch-deutschen und deutsch-mordwinischen wörterbuch. *Impériale des Sciences de St.-Pétersbourg: VIIe Série*.

B. Š. Zaguljaeva. 1991. *Russko-udmurtskij razgovornik*. Udmurtija, Izhevsk.