On Erzya and Moksha Corpora and Analyzer Development, ERME-PSLA 1950s

Rueter, Jack

2024-12

Rueter, J, Erina, O & Kabaeva, N 2024, On Erzya and Moksha Corpora and Analyzer Development, ERME-PSLA 1950s. in M Hämäläinen , F Pirinen, M Macias & M Crespo Avila (eds), Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages. The Association for Computational Linguistics, Kerrville, pp. 67–75, International Workshop on Computational Linguistics for Uralic Languages , Helsinki, Finland, 28/11/2024.

http://hdl.handle.net/10138/590946

On Erzya and Moksha Corpora and Analyzer Development, ERME-PSLA 1950s

Jack Rueter[1], Olga Erina[2] and Nadezhda Kabaeva[3]
[1]University of Helsinki
[2]Independent researcher
[3]Mordovian State University
[1]first.last@helsinki.fi
[2]first.jerina@gmail.com

Abstract

This paper describes materials and annotation facilitation pertinent to the «Erzya-Moksha Electronic Resources and Linguistic Diversity» (EMERALD) project. It addresses work following the construction of finite-state analyzers for the Mordvin languages, the gathering of test corpora, and the development of metadata strategies for descriptive research.

In this paper, we provide three descriptors for a set of new Erzya and Moksha research materials at the Language Bank of Finland. The descriptors illustrate (1) a low-annotation subcorpora set of the «Electronic Resources for Moksha and Erzya» (ERME); (2) the state of the open-source analyzers used in their automatic annotation, and (3) the development of metadata documentation for the «EMERALD» project, associated with this endeavor.

Outcomes of the article include an introduction to new research materials, an illustration of the state of the Mordvin annotation pipeline, and perspectives for the further enhancement of the annotation pipeline.

1 Introduction

The Mordvin languages, Erzya and Moksha, are spoken in settlements scattered throughout the Volga Basin (see Rueter 2013), but there have been settlements beyond that as well (see Sarv 2002).

Work with the description of these languages dates from the late 1600s in the form of word lists Witsen 1705, but it was not until the end of the 1830s that the first attempts at grammars were made for Moksha Ornatov 1838 and Erzya Gabelentz 1838–1839.

The Erzya grammar by Herr Conon von der Gabelentz is a striking study in that it illustrates the author's meticulous parallel-corpus-type knowledge of the Biblical texts. With this knowledge, Hhe was able to identify irregularities in the text and draws relatively accurate conclusions with regard to the meaning of morphological items[1]. The text of the Erzya Gospel is partially available now in the «Parallel Bible Verses for Uralic Studies» (PaBiVUS) corpus version 1 (see Helsingin yliopisto, FIN-CLARIN et al. 2020-06-07), but the next version of PaBiVUS will see the entire New Testament in Erzya from 1827, which can be aligned with translations into several other minority Uralic languages.

In this paper, we describe a new portion of the Electronic Resources Moksha-Erzya (ERME) and its annotation, which is soon to be introduced on the Language Bank of Finland Korp server.

The original intent of the ERME corpora was to provide the language community and researchers with citeable materials that distinguish writers and other language sources both geographically and in temporal space.

This meant the establishment of metadata features to served the purposes of the «EMERALD» project: provide parallel, consistent metadata for (1) Fieldwork and Early Literary Texts (FELT), with a focus on language materials collected in the Pre-Soviet Era, such as the Mordvin fieldwork collections by Heikki Paasonen and others (cf. Finno-Ugrian-Society, Suihkonen 2003); (2) non-central publications in the minority languages of the Soviet Union between the two World Wars[2], and (3) work

[1]The grammar in full can be accessed here: https://rueter.github.io/emerald/historical-mordvin-grammars/docs/gabelentz_hcvonder-versuch-einer-mordwinischen-grammatik-1838-39.html

[2]Here is a collection of openly licensed printed media including those from the era of minority-language popularization in the USSR. Outcomes of the Kindred Language Digitization pilot at the National Library of Finland with funding from the Kone Foundation «Language Programme» https://fennougrica.kansalliskirjasto.fi/

Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages, pages 67–75

with modern collection of the Mordvin language.

The metadata was seen as a means to align research and fieldwork documentation of the languages with texts from an era of language popularization, subsequent fieldwork conducted with these languages and their modern state. Thus, the metadata was organized to describe materials from the Heikki Paasonen collections of folklore, the Fenno-Ugrica collections at the National Library of Finland, the MORMULA corpora in Turku, the dialect archives at the Mordovian State University in Saransk (see Rueter 2020; Kabaeva 2021; Agafonova and Râbov 2021), and the ERME corpora at the Language Bank of Finland. This work was conducted in Turku, Helsinki and Saransk (2018–2022) in close association with corpora lemmatization, Constraint Grammar development and testing in Turku, dialect archive development in Saransk, and subsequent work with corpora in Helsinki.

In addition to metadata strategies for research citation, the annotation of the ERME corpora can be described as a combination of work with optical character recognition (OCR) and the development of finite-state descriptions for the morphology and syntax of the languages.

Over the years, attempts have been made to bring these two more tasks closer together. In the 1990s, OCR meant accurate recognition of individual characters. Since then attempts have been made to include finite-state work in OCR with hopes of word form recognition Silfverberg and Rueter 2015 for improved accuracy. Simple examples of finite-state approaches also include plain word lists or regular expressions used to represent numerous Uralic languages in OCR work in pilot projects conducted at the National Library of Finland[3].

Subsequent work with Mordvin lexica, morphology and syntax have played a major role in finding a purpose and collaboration beyond these languages. Work in lexica has meant collaboration in Erzya and Moksha with NorthEuraLex together with Mordovian State University staff [4], but it has also meant the development of dictionary editing (see Hämäläinen et al. 2021) and the enhancement of these dictionaries (see Alnajjar et al. 2022). Work with morpho-syntax, in turn, has opened connections to collaboration with specialist in the Komi languages in Syktyvkar (see Rueter et al. 2021) and the Universal Dependencies project with contributions to languages beyond Erzya and Moksha Zeman and et al. 2024. All of these together have contributed to utilizing the Giellatekno/Divvun[5] infrastructure GiellaLT (an infrastructure for Saami Language-Technology research and facilitation) and the open-source, shallow-transfer machine translation infrastructure Apertium[6]. The Apertium concept of shallow transfer makes it possible to draw parallels between lexicon, morphology, syntax and phraseology for the inspection of language diversity among closely related languages (see Rueter and Hämäläinen 2020, Rueter 2022). It also allows for a better comparison of closely related languages in linguistic research (cf (Rueter, 2023; Rueter and Kabaeva, 2024)).

The article proceeds by discussing the materials, metadata and methods, the state of the individual analyzers and prospects for future development.

2 Materials and methods

Despite a history of over 200 years of published texts in Erzya and Moksha, there is a dearth of searchable Mordvin text corpora consistently annotated for morphology, metadata and openly accessible. For this reason the Erzya-Moksha Electronic Resources And Language Diversity (EMERALD) project continues to augment metadata enriched materials available to the research communities (Rueter 2024).

The Electronic Resources for Moksha-Erzya (ERME) corpora versions one and two have been made available through Fin-CLARIN on the Language Bank of Finland Korp server since 2017. The first version of ERME focused merely on providing necessary metadata to facilitate a better alignment of text and authors, such that time-line and geographical plotting could

[3] https://www.doria.fi/bitstream/handle/10024/101915/Hakkarainen_Tallinn_19112014.pdf?sequence=2

[4] Erzya http://www.northeuralex.org/languages/myv and Moksha http://www.northeuralex.org/languages/mdf

[5] https://giellalt.github.io/

[6] https://wiki.apertium.org/wiki/Main_Page

be made consistent with analogous information available for Erzya-language fieldwork. In the second version of ERME, however, the materials were extended to include both Mordvin literary languages – Erzya and Moksha – with metadata for individual publications, and automatic annotation made possible with finite-state analyzers for the two languages.

Morpho-syntactic annotation has been accomplished using finite-state transducers describing the two languages (Lindén et al., 2013; Rueter et al., 2020). These, in turn, have been followed in the same pipeline by Constraint Grammar (CG) disambiguation, functions and dependencies tools[7]. The coding is open-source and facilitated in the Giellalt[8] infrastructure (cf. Moshagen et al. 2014) where the Erzya and Moksha languages share the progress in mutually applicable code that serves over 90 languages. At Giellalt, the code is reused whereever possible, i.e., analyzer code is flipped to make compatible generators, pedagogic descriptions are filtered to provide standard and descriptive models of the language for linguists, while normed filtering provides for spellers.

2.1 Materials

In the 2023, it was decided at the Language Bank of Finland that more Erzya and Moksha texts could be published on the Korp server if it involved less annotation. To this end, texts were selected from the Moksha-language journal «Mokša» and Erzya-language journals «Suran′ tolt» and «Sâtko». The texts represent original and translated writings from the late 1920s to the beginning of the 2000s. The right to use these texts in searchable corpora had been secured in Saransk as the beginning of the new millennium. The new portion of corpora was called ERME-Paragraph Segmentation Low Annotation (ERME-PSLA).

The majority of the texts was scanned in the 2010s with funding from the Finno-Ugrian Society and the University of Helsinki. Optical recognition was then conducted from 2017, and in 2024 the recognition changed from ABBYY Finereader to Transkribus[9], which would allow greater access to the OCR engine and models for recognition.

As the size of this material became apparent, it was decided that the corpora might further be divided according to decade and, of course, language. The first portion of the corpora came from the last four years of the 1950s (1956–1959), hence the name ERME-PSLA 1950s.

2.1.1 Figures

The size of the corpora can be measured as twenty-two issues in each language, i.e., four issues from 1956 and six issues from each of the subsequent years. The yield of those four years is 831 pieces written in Moksha, and 707 pieces in Erzya. This equates to approximately 91,017 sentences of Erzya and 92,432 sentences of Moksha, which is 803,406 and 902,518 words in Erzya and Moksha, respectively. The number of words might be compared with the analogous figures for ERME version 2 (Rueter and Erina 2023-03-23) – Moksha 855,435 and Erzya 2,041,196.

For the four years of publications in twenty-two issues for each language, there were 163 authors with pieces in Erzya, and 185 with pieces in Moksha. If we count the number of authors with four or more pieces, we arrive at thirty-five writers in Erzya, and fourty-three writers in Moksha.

The genres include poetry, story, short story, novel, essay, parody, critique, etc.

2.1.2 Errors

As is the case with most of the ERME materials, all of the ERME-PSLA corpora have been acquired through Optical Character Recognition (OCR). It goes without saying, some of the words in the corpora will be broken or unrecognized, which might lessen the value of the automatically annotated text. This shortcoming in the texts is one reason why the metadata includes page numbers and sentence enumeration; ORC errors might be located and corrected for future enhanced publications.

The probability of OCR errors exists. This can be seen in a simple comparison of ERME-PSLA 1950s figures against the number of word forms attested in the digitally transferred Erzya New Testament (NT) 2006 and Moksha New Testament 2016. In Table 1, the unique number of word forms is shown in parallel with word

[7]This site provides extensive information on Constraint Grammar https://edu.visl.dk/constraint_grammar.html

[8]https://github.com/giellalt/

[9]https://app.transkribus.org/

forms consisting of one to four letters in length followed by their ratio. The lower the ratio, the larger the number of short word forms in the language. Of the 7,648 unique Moksha word forms of four or less letters in length, 3,800 were not recognized by the Moksha analyzer. Likewise, of the 6,654 unique Erzya word forms of four or less letters in length, 3,082 were not recognized by the Erzya analyzer.

corpus	unique	unique 1–4	ratio
NT Erzya	18,439	874	21.10:1
1950s Erzya	99,287	6,654	14.92:1
NT Moksha	17,902	1,036	17.18:1
1950s Moksha	118,121	7,648	15.44:1

Table 1: OCR error statistics

2.2 Metadata

When ERME version two was published in the spring of 2023, it had twenty metadata features. Some of the attribute values were required by the Korp system, while others were introduced by the ERME documentation. The system required six attributes – a unique identifier, an ISO 639 language code, and timestamps – date from, date to, time from, time to. The remaining fourteen were optional.

The optional features were basically bibliographical. These consisted of <author>, <genre>, <number of pages>, <page range>, <publication place>, <publication name>, <publication year>, <publisher> and <bibliography> (an entire segment dedicated to bibliographic citation). After the information necessary for citation, came information which might help to explain linguistic characteristics of the text, namely, <corrector>, <electronic corrector> and <geographic origin of author> (this information is documented separately complete with coordinates). This was followed by two bit of statistics: <word count> and <character count>. All information was explicitly available in the corpora sources or it was readily derived. All texts were deemed original-language materials, and no extra information was needed for translation.

In the ERME-PSLA corpora, the twenty features shown above are elaborated upon. The <bibliography> segment is now written in Cyrillics, and it has a twin <bibliography iso9>, where all information in Cyrillics is converted according to the International Library Convension in ISO-9[10].

The challenge presented by journals is that individual pieces do not always explicitly indicate metadata important for locating a text in time and space. Thus, all important metadata, such as information on authors, titles, genres, and even correctors, electronic correctors are not readily available for all pieces. In fact, there are only three bits of information that can serve as consistent key identifiers: (1) the publication; (2) issue, and (3) page ranges. Information on authors, titles, genres and correctors are given whenever specifically stated in the publications or their sibling issues. Poetry and lyrics lacking titles are named using the first line of piece.

Journals are collections of pieces, such that a distinction must be made between the concepts of publication the container and the individual piece. In previous iterations of ERME corpora, this distinction has not been necessary, but the necessity for a feature <title> has already been encountered and facilitated in short stories by the Erzya author Pëtr Klûčagin 1997.

Journals are also collections of pieces representing both translations and original-language writing. This information is described in with the addition of features for <translator> and <translated>. While the former might readily be associated with a human actor, the latter requires explanation.

When approaching the concept of translation into Erzya and Moksha, we must all agree that anything written by Longfellow, Heine or Lenin has obviously be translated. Hence, the attribute <translated> in such pieces can easily be assigned the value <yes>.

A problem arises, however, when a piece by an Erzya or Moksha writer is indicated as having been translated from a different language by another native language writer, e.g., the play Ульнесь истямо тейтерь... 'There was such a girl...' by Ivan Antonov, 1957 was translated from the Russian by Aleksandr Ŝeglov. For this piece, the attributes translated and translator are given to attest certainty of translation, on the one hand, and knowledge of translator, on the other. Further investigation must be conducted before the value <no> can be assigned

[10] https://en.wikipedia.org/wiki/ISO_9

to the <translated> attribute.

Both journals present elements of folklore, such as riddles, lyrics and poetry. These require an additional human actor, the <collector>. Thus, materials collected by M. E. Evsev′ev might have subsequently been corrected by a <corrector>, while the language informant is provided with the <author> attribute.

Finally, it is important that versions and translations of a text be associated with each other for comparative studies, and that segments of larger texts be associated with each other. The concept of translation parallel is expressed in the feature <has parallel>. The <version> feature is initially used to indicate original version, but this may be altered in further development. Segment pointers are expressed with the attributes <continues from> and <continues to>. The attribute <comment> is reserved as a miscellaneous container, and the <content> attribute helps determine whether the piece will be contained in the corpora; the value <text> entails selection for use in the corpora.

2.3 Method

The low-annotation pipeline entails the same components found in the annotation pipelines for PaBiVUS version two and ERME version two. As these two sets of corpora that presuppose sentence-level annotation where the texts have been broken down to allow for sentence-level commenting, page numbers, indication of temporary change in genre and even language. ERME-PSLA makes the assumption that manual annotations are only made in the root element.

ERME-PSLA paragraph segmentation proceeds directly from the root element. Whereas previous work with ERME has presupposed sentence-level annotation, the PSLA pipeline utilizes features of the optical character recognition machines. These features include page and paragraph breaks. This has facilitated the numbering of pages and the recognition of paragraphs, which have simply been set off by double line breaks, and sentences have been recognized automatically with the help of sentence-final punctuation marks, such as full stop, question mark, exclamation mark and colon.

Although errors may have occurred in this part of the segmentation, the sentence-level texts are now ready for annotation.

Annotation at the sentence level involves minimal additional human input. The recognized sentences are automatically annotated with unique identifiers and page numbers for reference to the source texts.

3 FST models

The finite-state description of the two Mordvin literary languages started nearly fifteen years apart. Work on the Erzya analyzer was begun in the late 1990s, whereas work on Moksha was part of the «Creation of Morphological Parsers for Minority Finno-Ugrian Languages» project funded by the Kone Foundation «Language Programme», 2013–2014 (see Rueter 2014; Rueter et al. 2020). Despite attempts to make the descriptions as parallel as possible, there are still statistical differences in their coverage and accuracy afforded by the Erzya and Moksha analyzers.

Both analyzers are used in the analysis of corpora texts on different platforms and projects[11], system-wide spell checkers[12]. Their development is part of a collaboration with the «Experimental Treebanking for the Minority Moksha Language and Finite-State Descriptions» and «Experimental Treebanking for the Minority Erzya Language and Finite-State Descriptions» projects. The analyzers have been used in the annotation of treebanks on the Korp servers of both the Language Bank of Finland[13] in Helsinki, Finland, and Giellatekno/Divvun[14] in Tromsoe, Norway.

Evaluation and enhancement of these analyzers is important for improving community language facilitation. Brief statistical descriptions of the Erzya and Moksha analyzers have been given online through the Language Bank of Finland[15]. These provide annotational statistics based on materials of the upcoming version of PaBiVUS.

Below we offer an enhanced evaluation of the PaBiVUS annotation for comparison with figures for the outcome of the ERME-PSLA 1950s corpora. While notions of size are imme-

[11] https://universaldependencies.org/
[12] https://divvun.org/
[13] kielipankki.fi/korp/
[14] gtweb.uit.no/u_korp/
[15] https://www.kielipankki.fi/tools/giellalt_language_models/

diately obvious from the statistics, one must also bear in mind the cleanliness of the corpora. As noted above, in materials, ERME-PSLA 1950s materials are likely to have OCR errors, this might be observed in the a comparison of unique forms to unique misses in the modern New Testament texts and the ERME-PSLA materials.

3.1 The Erzya FST

The Erzya finite-state model has a relatively large lexical base of 176,832 lemma-stem pairs, and 1,370 continuation lexica. Together, these provide for a variety of inflectional patterns in verbs and words in the nominal categories.

Recently, a simple analysis of the Erzya model was published on the Language Bank of Finland website[16]. It provided statistics based on testing with the upcoming version two of Parallel Biblical Verses for Uralic Studies (PabiVUS). Modified statistics on the analyzer are given here, in which a distinction is drawn between results for the 1821–1827 version of the New Testament and the 2006 version. These results are aligned with results for the 1950s portion of ERME-PSLA (ERME-Paragraph Segmentation Low Annotation).

The three corpora can be distinguished in many ways. While notions of size are immediately obvious from the statistics, one must also bear in mind the cleanliness of the three. As noted above, in materials, ERME-PSLA 1950s materials are likely to have OCR errors, this might be observed in a comparison of unique forms to unique misses in the New Testament texts from 2006 and the ERME-PSLA materials. The New Testament (NT) figures for unique forms over unique misses renders a ratio of 48.53:1, whereas the correlating figures for ERME-PSLA are 5.13:1. The ratio for NT 1821–1827 is 2.54:1, which may be attributed to high variation in spelling in the older version of the Erzya New Testament.

The sizes of the three corpora illustrate a difference between two versions of the New Testament in Erzya and the text content of twenty-two issues of «Suran′ tolt», which is five and a half times the size of the New Testament. Their automated annotation is illustrated in Table 2.

corpus	NT 1821–1827	NT 2006	ERME-PSLA
words total	128,245	140,942	803,406
characters total	711,716	857,812	5,003,429
unique forms	22,569	18,439	99,287
unique misses	8,899	380	19,342
lines before hapax	1943	58	2,917
ambiguous PoS	8,943	449	30,352
unique amb. PoS	8,899	399	26,890
ambiguous dep.	29,120	9,999	91,625
unique amb. dep.	9,151	692	20,269

Table 2: Erzya annotation statistics

The Erzya analyzer and disambiguation do a better job than the dependency parser and subsequent conversion scripts from Constraint Grammar to Universal Dependencies presentation.

In the Erzya New Testament, 2006, the ratio for unique forms to unique forms with ambiguous parts of speech is 46.21:1, while the ratio for unique forms to unique forms with ambiguous or unrecognized dependencies is 26.65:1. The correlating figures for NT 1821–1827 and ERME-PSLA 1950s are: 2.54:1, 2.47:1 and 3.69:1, 4.90:1, respectively. Words total over ambiguous dependencies were 14.1:1 (NT 2006), 4.4:1 (NT 1821–1827), and 8.77:1 (ERME-PSLA 1950s).

3.2 The Moksha FST

The Moksha finite-state analyzer is younger than the Erzya model. It has been under construction since 2012, as part of the «Creation of Morphological Parsers for Minority Finno-Ugrian Languages» project, 2013–2014. Since then the analyzer has grown with approximately 189,476 lemma-stem pairs in the lexicon and 852 continuation lexica for facilitating complex morphology in the nominal and verbal categories.

A simple evaluation of the Moksha analyzer was recently published on the Language Bank of Finland website[17], in which test results for the model's performance on the New Testament texts in PaBiVUS version two were described.

Below, we provide a modified version of those statistics, where we draw only on the texts published in 2016, and compare them to analogous results for the ERME-PSLA 1950s corpus.

[16] urlhttps://www.kielipankki.fi/tools/giellalt_language_models/erzya/

[17] urlhttps://www.kielipankki.fi/tools/giellalt_language_models/moksha/

The two corpora can be distinguished by cleanlieness, size and accuracy. As noted above, in materials, ERME-PSLA 1950s materials are likely to have OCR errors, this might be observed in the a comparison of unique forms to unique misses in the New Testament texts from 2016 and the ERME-PSLA materials. The ERME-PSLA 1950s corpus for Moksha is over six and a half times the size of the New Testament materials from 2016. In fact, it is slightly larger than the ERME version two Moksha corpus from 2023. The NT figures for unique forms over unique misses renders a ratio of 21.99:1, whereas the correlating figures for ERME-PSLA are 3.56:1.

corpus	NT 2016	ERME-PSLA 1950
words total	136,718	902,518
characters total	793,393	5,559,553
unique forms	17,902	118,121
unique misses	814	33,195
lines before hapax	129	4,436
ambiguous PoS	1,075	46,361
unique amb. PoS	856	33,195
ambiguous dep.	10,853	117,888
unique amb. dep.	1,238	34,775

Table 3: Moksha annotation statistics

The Moksha analyzer and disambiguation perform better than subsequent function and dependency parsing followed by conversion scripts.

In the Moksha New Testament 2016, the ratio for unique forms to unique forms with ambiguous parts of speech is 20.91:1, while the ratio for unique forms to unique forms with ambiguous or unrecognized dependencies is 14.46:1. The correlating figures for ERME-PSLA 1950s are: 3.56:1 and 3.4:1. Total word forms divided by ambiguous dependencies give us the figures 12.6:1 (NT 2016), and 7.66:1 (ERME-PSLA 1950s).

3.3 FST retrospect

A Korp Vertical structure (VRT) validator under continuous development at the Language Bank of Finland is used to determine validity of the XML files, and additional scripts are run to assess the number of word forms lacking recognition, word forms with ambiguous part-of-speech readings and word forms with ambiguous dependencies. These figures have then been used in the evaluation of the individual analyzers, disambiguators, annotation for function and dependency, and conversion scripts.

It has been noted that over half of the unique missing word forms occur but once. The following assumptions have been made: High frequency of missing word forms would indicate need for lexical inspection and enhancement. High frequency of part-of-speech ambiguity may point to homonymy. And high frequency of dependency ambiguity may actually point to shortcomings in the CG-to-UD conversion scripts.

The annotation stops when the lexica or morphology are lacking or blocked. No non-described annotations are make.

4 Discussion and Conclusions

The statistics for the Erzya and Moksha analyzers were drafted before the present article was begun. During the course of writing the article, a number of problems were noted with regard to the CG-to-UD-format transfer and the quality of the OCR.

It was decided that enhancement and evaluation ought to be included in the release protocol of each new korpus for either of the languages.

In a brief inspection of the ERME-PSLA OCR quality, a lists of Erzya and Moksha words four letters or less in length were extracted from New Testament texts. These words had been human inspected and digitally transferred, which would guarantee their quality. In the future, however, is was decided that separate lists of this kind should be drafted and curated to be used in inspection of text validity.

Finally, more work should be allotted to the development of function and dependency parsing, as these along with conversion strategies would immensely improve the usability of ERME corpora.

Ethics statement

When dealing with an endangered language it is important to make sure that the research also contributes to the language community. This is the reason why we open-source our FST analyzers and components. We also work with data licensed to us by speakers of Moksha and Erzya with the intention of contributing to morpho-syntactic descriptions, tools and metadata practices for the languages. This means

that we are not conducting our research with no regard to the language community.

Acknowledgments

This research is supported by FIN-CLARIAH and Academy of Finland (grant 358720 Kielivarojen ja kieliteknologian tutkimusinfrastruktuuri).

References

Nina Agafonova and Ivan N. Râbov. 2021. Ulânovskoj oblasten novomalyklinskoj rajonon Èrzân velen kortavkstnèsè azorksčin nevtycâ suffikstnén baška ënksost. In Niko Partanen Mika Hämäläinen and Khalid Alnajjar, editors, Multilingual Facilitation, page 263–274. University of Helsinki Library. This book has been authored for Jack Rueter in honor of his 60th birthday.

Khalid Alnajjar, Mika Hämäläinen, Niko Tapio Partanen, and Jack Rueter. 2022. Using graph-based methods to augment online dictionaries of endangered languages. In Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 139–148, Dublin, Ireland. Association for Computational Linguistics.

Ivan Antonov. 1957. Ul'nes' istâmo tejter'... Suran' tolt, (№ 1):3–32. Genre = p'esa, translated from Russian = Ŝeglov, Aleksander.

Finno-Ugrian-Society. Suomalais-Ugrilaisen Seuran kenttätyökorpus.

Herr Conon von der Gabelentz. 1838–1839. Versuch einer mordwinischen grammatik. Zeitschrift für die Kunde des Morgenlandes., II(2–3):235–284, 383–419.

Mika Hämäläinen, Khalid Alnajjar, Jack Rueter, Miika Lehtinen, and Niko Partanen. 2021. An online tool developed for post-editing the new skolt sami dictionary. In Electronic lexicography in the 21st century (eLex 2021). Proceedings of the eLex 2021 conference, Electronic lexicography in the 21st century (eLex 2021). Proceedings of the eLex 2021 conference, pages 653–664, Czech Republic. Lexical Computing CZ s.r.o. Electronic lexicography in the 21st century (eLex 2021) ; Conference date: 05-07-2021 Through 07-07-2021.

Helsingin yliopisto, FIN-CLARIN, Jack Rueter, and Erik Axelson. 2020-06-07. Raamatun jakeita uralilaisille kielille, rinnakkaiskorpus, Korp.

Nadežda Kabaeva. 2021. Fonetičeskie osobennosti govora sela adaševo ûgo-vostočnogo dialekta mokšanskogo âzyka. In Ksenia Shagal with Heini Arjava, editor, Mordvin Languages in the Field, volume 10 of Uralica Helsingiensia, page 171–186. University of Helsinki Finno-Ugric Language Section and the Finno-Ugrian Society.

Pëtr Andreevič Klûčagin. 1997. Garec'. In Cëkan'ka, page 58–62. Mordovskoj knižnoj izdatel'stvas', Saransk.

Krister Lindén, Erik Axelson, Senka Drobac, Sam Hardwick, Juha Kuokkala, Jyrki Niemi, Tommi A Pirinen, and Miikka Silfverberg. 2013. Hfst—a system for creating nlp tools. In Systems and Frameworks for Computational Morphology: Third International Workshop, SFCM 2013, Berlin, Germany, September 6, 2013 Proceedings 3, pages 53–71. Springer.

Sjur Moshagen, Jack Rueter, Tommi Pirinen, Trond Trosterud, and Francis M Tyers. 2014. Open-source infrastructures for collaborative work on under-resourced languages. Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era, pages 71–77.

Pavel Ornatov. 1838. Mordovskaâ grammatika. Moskva: V" Sinodal'noj tipogrfii. Sostavlennaâ narečij mordvy mokši Tambovskoj seminarii professorom", magistrom", Pavlom" Ornatomvym".

Jack Rueter. 2013. The erzya language, where is it spoken? Études finno-ougriennes, 45.

Jack Rueter. 2014. The livonian-estonian-latvian dictionary as a threshold to the era of language technological applications. Journal of Estonian and Finno-Ugric Linguistics, 5(1):251–259. ESUKA – JEFUL 2013, 5–1: 253–261.

Jack Rueter. 2020. Корпус национальных мордовских языков: принципы разработки и перспективы функционирования/ действия. In Финно-угорские народы в контексте формирования общероссийской гражданской идентичности и меняющейся окружающей среды, pages 118–127, Russia. Издательский центр Историко-социологического института. Conference date: 08-10-2020 Through 09-10-2020.

Jack Rueter. 2022. Shallow-transfer problems in erzya-moksha conjunctive-preterite2 syncretism. The Journal of Brief Ideas.

Jack Rueter. 2024. On searchable mordvin corpora at the language bank of finland, emerald. Journal of Data Mining and Digital Humanities.

Jack Rueter and Olga Erina. 2023-03-23. ERME Ersän ja mokšan laajennettu korpus versio 2, Korp.

Jack Rueter and Mika Hämäläinen. 2020. Prerequisites For Shallow-Transfer Machine Translation Of Mordvin Languages: Language Documentation With A Purpose, pages 18–29. Ижевск:

Институт компьютерных исследований, Russian Federation.

Jack Rueter, Mika Hämäläinen, and Niko Partanen. 2020. Open-source morphology for endangered mordvinic languages. In Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS), pages 94–100, Online. Association for Computational Linguistics.

Jack Rueter and Nadezhda Kabaeva. 2024. On quantification and the ablative in erzya and moksha. In Hajner Réka, Kubínyi Kata, Dóra Pődör, and Tamm Anne, editors, European partitives in comparison. L'Harmattan Kiadó, Károli Gáspár Református Egyetem, Budapest, Magyarország. Budapest, Magyarország : L'Harmattan Kiadó, Károli Gáspár Református Egyetem (2024).

Jack Rueter, Niko Partanen, Mika Hämäläinen, and Trond Trosterud. 2021. Overview of open-source morphology development for the Komi-Zyrian language: Past and future. In Proceedings of the Seventh International Workshop on Computational Linguistics of Uralic Languages, pages 29–39, Syktyvkar, Russia (Online). Association for Computational Linguistics.

Jack Michael Rueter. 2023. Moksha Mordvin, 2nd edition edition, Routledge Language Family Series. Routledge, International.

Heno Sarv. 2002. Indigenous Europeans East of Moscow, Population and Migration Patterns of the Largest Finno - Ugrian Peoples in Russia from the 18th to the 19th Centuries., volume 17. Geographicae Universitatis Tartuensis. Dissertation.

Miikka Silfverberg and Jack Rueter. 2015. Can morphological analyzers improve the quality of optical character recognition? In First International Workshop on Computational Linguistics for Uralic Languages, volume 2 of Septentrio Conference Series, pages 45–56, Norway. Septentrio Academic Publishing. Proceeding volume: 2; International Workshop on Computational Linguistics for Uralic Languages ; Conference date: 16-01-2015 Through 16-01-2016.

P. M. Suihkonen. 2003. Metadata descriptions for combining information on multimodal data located at the university of helsinki language corpus server. In S. Darányi, editor, Proceedings of the Ḧigher Order Morphologies' Observer 2003C̈onference on Information Society: Cultural Heritage and Folklore Text Analysis Budapest University of Technology and Economics. L'Harmattan Kiadó, Károli Gáspár Református Egyetem, Budapest, Hungary.

Nikolaes Witsen. 1705. Noord en Oost Tartarye, Ofte Bondig Ontwerp Van eenig dier Landen en Volken Welke voormaels bekent zijn geweest. Beneffens verscheide tot noch toe onbekende, en meest nooit voorheen beschreve Tartersche en Nabuurige Gewesten, Landstreeken, Steden, Rivieren, en Plaetzen, in de Noorder en Oosterlykste Gedeelten Van Asia En Europa Verdeelt in twee Stukken, Met der zelviger Landkaerten: mitsgaders, onderscheide Afbeeldingen van Steden, Drachten, enz. Zedert naeuwkeurig onderzoek van veele Jaren, door eigen ondervondinge ontworpen, beschreven, geteekent, en in t licht gegeven, Door Nicolaes Witsen. Jan van der Deyster, Amsterdam By François Halma, Boekverkooper op de Nieuwendyk.

Daniel Zeman and et al. 2024. Universal dependencies 2.14. Universal Dependencies 2.14, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.