Open-Source Morphology for Endangered Mordvinic Languages

Abhinav S Menon
2020114001

Overview

Introduction

Designing for a Reusable API

Erzya and Moksha Language Pair

Current State and Ensuring Maintainability

Conclusions

Introduction

- Mordvinic languages are endangered
  - Erzya and Moksha

- In this paper
  - Open-source FST-based morphological analyser
  - “Highlight importance of design decisions”
  - Describe how to contribute

- Previous work on Erzya and Moksha

Designing for a Reusable API

- Open source is preferable
  - Different groups independently create same tools with different APIs

- GiellaLT framework
  - HFST
  - Supported by Apertium

Erzya and Moksha Language Pair

- Structure of Apertium
  - Morph analysis
  - Morpheme mapping
  - Morph generation

- Paradigms differ in some cases

Current State & Ensuring Maintainability

Current State

- Left out
  - Declension of nominals
  - Orthographic compounds
- Method of development
  - Agile :)
  - Unit testing

Future Contribution

- Anyone should be able to contribute
- XML-based dictionary
  - User-friendly UI
- Non-tech-savvy users can contribute

Conclusions

Core design principle: Open-source

Accessibility in usage & contribution

Views

- (Almost) Not a linguistics/CL/NLP paper
- Lots of fluff
  - “Moksha has a symmetric paradigm in the core cases, nominative, genitive and dative (see Table 2), with distinct word forms for each case-possessor combination slot and an additional distinction for number of possessa when the possessor is singular, whereas Erzya makes virtually no consistent distinction for case or number in the nominative and genitive – with one exception the specific nominative singular reading of the third person singular +N+SG+NOM+PXSG3 and option- ally in the first and second person singular of modern written literature (see Table 3).”
- Motivation, development process, API, interface
  - Sidesteps actual description of FSTs

References

- 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.

Open-Source Morphology for Endangered Mordvinic Languages

Jack Rueter
Dept. of Digital Humanities
University of Helsinki
jack.rueter@helsinki.fi

Mika Hämäläinen
Dept. of Digital Humanities
University of Helsinki
and Rootroo Ltd
mika@rootroo.com

Niko Partanen
Dept. of Finnish,
Finno-Ugrian
and Scandinavian Studies
University of Helsinki
niko.partanen@helsinki.fi

Abstract

This document describes shared development of finite-state description of two closely related but endangered minority languages, Erzya and Moksha. It touches upon morpholexical unity and diversity of the two languages and how this provides a motivation for shared open-source FST development. We describe how we have designed the transducers so that they can benefit from existing open-source infrastructures and are as reusable as possible.

1 Introduction

There are over 5000 languages spoken world wide, and a vast majority of them are endangered (see Moseley 2010). The Mordvinic languages Erzya and Moksha are no exception. One of the first NLP solutions that are typically developed along with lexical resources for any low-resourced language is a morphological analyzer (cf Zueva et al. 2020; Tyers et al. 2019; Lovick et al. 2018).

In this paper, we describe the development of an open-source FST (finite-state transducer) based morphological analyzer, lemmatizer and generator for Erzya and Moksha. We highlight the importance of certain design decisions to ensure the compatibility of our transducers with existing systems. In addition, we will describe how the transducers can be edited in a system that creates an abstraction layer behind a graphical user interface for FST development.

The unity and diversity of the Mordvin literary languages of today, Erzya and Moksha, has been a subject of research for over two hundred years. The first grammars were published in 1830s – Moksha in 1838 (Ornatov, 1838) and Erzya in 1838–1839 (Gabelentz, 1839). The subsequent 180 years brought scholars for fieldwork, grammars, dictionaries, and the popularization of the languages. 2002 saw the publication of the first monolingual dictionary of Erzya (Abramov, 2002), and the manuscript was proclaimed open by the author for future development. The Mordvin languages have continued to receive a fair share of linguistic research interest in the recent years (Luutonen, 2014; Hamari and Aasmäe, 2015; Kashkin and Nikiforova, 2015; Grünthal, 2016).

After the release first finite-state transducer for the closely related Komi-Zyrian (Rueter, 2000), it was only obvious that similar work should be done for Erzya Mordvin. Fortunately, over the past decade there has been an increasing number of publications on Erzya, relating to its morphology (Rueter, 2010), its OCR tools (Silfverberg and Rueter, 2015) and universal dependencies (Rueter and Tyers, 2018).

This document discusses open-source morphology development, which has greatly benefited from open-source projects, most notably achievements attributed to the GiellaLT infrastructure (Moshagen et al., 2014), i.e. Giellatekno & Divvun at the Norwegian Arctic University in Tromsø, Norway.

It is also very important that people new to language documentation be given opportunities to develop their understanding of languages through participation in projects. Here we must mention an important span of time 1988–1997, during which the first author did word processing for the 2073-page 'Dictionary of Mordvin Dialects' by Heikki Paasonen.

The transducers are available on GitHub for Erzya [1] and Moksha [2]. The nightly builds are available through a Python library called UralicNLP[3] (Hämäläinen, 2019).

[1] https://github.com/giellalt/lang-myv

[2] https://github.com/giellalt/lang-mdf

[3] https://github.com/mikahama/uralicNLP

Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS), pages 94–100
Virtual Conference, November 19, 2020. ©2020 Association for Computational Linguistics

2 Designing for a Reusable API

It is not uncommon that similar tools and methods are developed by different research groups in different parts of the world for language documentation purposes. To name a few, there are projects developing similar language documentation systems for African languages (Jones and Muftic, 2020), Indonesian languages (Nasution et al., 2018) and Yupik (Hunt et al., 2019), while at the same time, it has been shown that digital humanities projects that are seemingly different, still face the very same technical problems (Mäkelä et al., 2020). In arguably, any work conducted with endangered languages to document them and better serve the small community of speakers has a lot of value. However, the fact that the wheel gets reinvented over and over again leads to fragmentation in the resources available for smaller languages and makes their use more difficult as each individual tool exposes an API of their own.

Rule-based morphology can be implemented in many different systems. Popular tools include XFST (Karttunen et al., 1997), Foma (Hulden, 2009) and OpenFST (Allauzen et al., 2007). However, we use HFST (Helsinki-Finite State Technology) (Lindén et al., 2009) because it is the system used in GiellaLT (Moshagen et al., 2014).

GiellaLT provides our transducers with a list of quality attributes. Their infrastructure consists of work done for multiple endangered languages in such a way that the morphological resources can be used in a multitude of different contexts such as disambiguation (Trosterud, 2004; Ens et al., 2019), dependency parsing (Antonsen et al., 2010), online dictionaries (Rueter and Hämäläinen, 2019), spell checkers (Wiechetek et al., 2019), online creative writing tools (Hämäläinen, 2018), automated news generation (Alnajjar et al., 2019) and language learning tools (Antonsen and Argese, 2018).

In order to gain the added benefit from the GiellaLT infrastructure, we have to design our transducers so that they are compatible with HFST and that they follow a certain morphological tagset and that they take the input and output in a certain format. These requirements define the API our transducers need to implement. An example of this can be seen in Table 1.

Apertium (Forcada et al., 2011) is another open-source system that uses FST transducers for rule-based machine translation. They use their own transducer format, but fortunately also support HFST based transducers, and in fact Apertium type transducers can be compiled to HFST form as well (Pirinen and Tyers, 2012). However, their tagset is different from that of GiellaLT. This is solved by applying filters at the time the FST is compiled to produce a separate, Apertium-compatible, transducer automatically.

	input	output
analyzer	вирев	вирев+A+Sg+Nom+Indef вирь+N+SP+Lat+Indef
generator	вирев+A+Sg+Nom+Indef	вирев

Table 1: Examples of GiellaLT style input and output

3 Erzya and Moksha language pair

At the moment Erzya and Moksha infrastructure in GiellaLT can be considered to be in nearly equal standing, despite the fact that work on Erzya was started considerably earlier. This does not mean, however, that the resources for both languages are identical in all measures, although in basic numeric levels their sizes are comparable. There are many situations where additional consistency between the infrastructures for these two languages would be quite desirable and beneficial. We discuss next what these instances could be, and outline some of the major questions while working with this language pair.

The open-source machine translation infrastructure Apertium uses a shallow-transfer strategy. By definition, shallow transfer makes a morphosyntactic analysis of the source language, translates the lemmas from source to target language, generates morphology acquired from the source language in the target language and makes adjustments to the syntax of the target language. This concept of parallel lexica, morphology and syntax, would therefore, seem most effective in translation between closely related languages. In the case of the Mordvin pair, this means the use of mutual tags for describing mutual phenomena.

Mutual phenomena in Erzya and Moksha, however, can be distinguished as exact matches and fuzzy matches, as it were. While there is no doubt that the two languages share nearly the same categories of case, person, number and definiteness, it must also be noted that the paradigms do not share the same cellular structure (Keresztes 1999; Trosterud 2006; Rueter 2016). When the paradigm structure is diverse, it is suggested that a union of morphological tags be taken as a starting point. When one language distinguishes a category of

	+Nom	+Gen	+Dat
Sg+PxSg1 (my son)	цёразе	цёразень	цёразти
Sg+PxSg1 (my sons)	цёране	цёранень	цёраненди
Sg+PxSg2 (your son)	цёраце	цёрацень	цёрацти
Pl+PxSg2 (your sons)	цёратне	цёратнень	цёратненди
Sg+PxSg3 (his/her son)	цёрац	цёранц	цёранцты
Pl+PxSg3 (his/her son)	цёранза	цёранзон	цёранзонды
SP+PxPl1 (our son/sons)	цёранькe	цёранькoнь	цёранькoнди
SP+PxPl2 (your son/sons)	цёранте	цёрантень	цёрантенди
SP+PxPl3 (their son/sons)	цёрасна	цёраснон	цёраснонды

Table 2: Symmetric possessive declension of Moksha core cases

number, for instance, and the other does not, there comes a point where number must be determined. And, in order to facilitate a transition from the absence of the category of number to its presence and vice versa, the rudiments of tagging this category must be put in place, e.g. for the category of number we use SG ’singular’, PL ’plural’ and SP ’singular or plural’.

Diversity in the Erzya-Moksha language pair can be found in the core ranges of the categories definiteness, person and number. Both languages have an indefinite or basic declension, a definite or determiner declension and a possessive declension. For both languages, it can be stated that the indefinite declension distinguishes the category of number in the nominative alone, whereas number is always specified in the definite declension paradigms. The possessive declension, however, exhibits a salient rift in unity. Moksha has a symmetric paradigm in the core cases, nominative, genitive and dative (see Table 2), with distinct word forms for each case-possessor combination slot and an additional distinction for number of possessa when the possessor is singular, whereas Erzya makes virtually no consistent distinction for case or number in the nominative and genitive – with one exception the specific nominative singular reading of the third person singular +N+SG+NOM+PXSG3 and optionally in the first and second person singular of modern written literature (see Table 3).

Additional diversity is found in the subject-object paradigm, where portmanteau morphology enables the specification of first, second and third person subject and object provided both arguments are singular. When one or both of the arguments is not specifically singular, however, one form may serve to indicate more than one set of arguments, e.g. Erzya has a default non-past form -*samiź* which simply indicates a first person object with a second or third person subject when it is not true that both subject and object are specified, singular entities (see Table 4).

	+Nom	+Gen	+Dat
Sg+PxSg1 (my son)	цёрам	цёрам ∼ цёрань	цёрам туртов ∼ цёрань туртов ∼ цёранень
Pl+PxSg1 (my sons)	цёран ∼ цёрам	цёран ∼ цёрам	цёран туртов цёрам туртов
Sg+PxSg2 (your son)	цёрат	цёрать ∼ цёрат	цёрать туртов ∼ цёратень
PxSg2 Pl (your son)	цёрат	цёрат	цёрат туртов
PxSg3 Sg (his/her son)	цёразо	цёранзо	цёранстэнь ∼ цёранзо туртов
PxSg3 Pl (his/her sons)	цёранзо	цёранзо	цёранстэнь ∼ цёранзо туртов
SP+PxPl1 (our son/sons)	цёранок	цёранок	цёранк туртов
SP+PxPl2 (your son/sons)	цёранк	цёранк	цёранк туртов
SP+PxPl3 (their son/sons)	цёраст	цёраст	цёранстэнь ∼ цёраст туртов

Table 3: Asymmetric possessive declension of Erzya core cases

	Obj+Sg1	Obj+Pl1
Subj+Sg2	NA	*-samiź*
Subj+Pl2	*-samiź*	*-samiź*
Subj+Sg3	NA	*-samiź*
Subj+Pl3	*-samiź*	*-samiź*

Table 4: Erzya default first person object

Moksha, however, makes a further semantic split with regard to the category of person in the subject, namely, the form -*samaśt’* is used to indicate a second person subject, and -*samaź* is used to indcate a third or indefinite person subject (see Table 5).

The differences outlined here are not the same, but comparable, to those found in a recent study that evaluated the morphological differences between Komi-Zyrian and Komi-Permyak within the context of FST and treebank development (Rueter et al., 2020). With closely related languages any kind of resource reuse or transfer is rarely trivial, but through careful evaluation of linguistic features and differences we show that these goals are certainly possible and realistic.

4 Current State and Ensuring Maintainability

At the time of writing the transducers lemmas, stems and glosses were acquired through several channels. Statistics on the coverage of the Erzya FST can be seen on Table 6. The same statistics for

	Obj+Sg1	Obj+Pl1
Subj+Sg2	NA	*-samaśt’*
Subj+Pl2	*-samaśt’*	*-samaśt’*
Subj+Sg3	NA	*-samaź*
Subj+Pl3	*-samaź*	*-samaź*

Table 5: Moksha default first person object

word class	total	glossed	inflections	derivations
common nouns	24723	10754	450	8
proper nouns	50566	146	(450)	8
adjectives	12938	446	(450)	1
verbs	16133	4123	356	20
lemma:stem pairs	106293	15908		36

Table 6: Statistics for Erzya transducers

word class	total	glossed	inflections	derivations
common nouns	12851	9056	426	6
proper nouns	50070	267	(426)	6
adjectives	12043	4083	(426)	1
verbs	13449	11577	337	10
lemma:stem pairs	92716	26572		23

Table 7: Statistics for Moksha transducers

Moksha are shown in Table 7.

Although the largest bilingual dictionaries for Erzya and Moksha provide declension information, classification of declension for other nominals has not been dealt with. In the noun phrase, adjectives are declined only when they are promoted to NP head in ellipsis or in the predicative. Hence, adjectives, by nature, might be declined to virtually the same extent as common nouns, but they are subject to fewer derivations. Proper nouns, although most commonly encountered in the singular, may, in fact, be declined in the plural – place names when declined in the plural are down cased and their new semantics refer to inhabitants. Person names gain a sense of associative plural.

Derivations include morphologically new forms and orthographic compounds. Most salient are the ever present diminutive, which has a sense of endearment, diminutive and even partitive measure. Orthographic derivation can be observed in compound nouns, with collective sense, and verbs, expressing simultaneous activities. The challenge of orthographic compounds lies in the hyphened pair where both elements are inflected for the same grammatical categories, such that copula forms of compound nouns are included in the morphology of both stems, but the additive clitic attaches only to the second.

The transducer development is conducted in an agile fashion with nightly builds available for the end user via an open-source Python library called UralicNLP (Hämäläinen, 2019). The quality of the transducers is ensured by unit testing. We use test scripts written in YAML that contain a list of inputs and accepted outputs. Any changes in the transducers that break the tests will be immediately noticed and acted upon.

The source code of an FST is not the easiest to write for an average linguist or an NLP researcher. In the context of endangered languages, however, one would hope that people working on language documentation with a limited technical background or even community members could make an active contribution to the morphological tools. In order to address this need, we have embraced two levels of abstraction for FST development.

The first level of abstraction is the use of an XML-based dictionary. Lexical data is stored in XML form in the GiellaLT infrastructure, and to a great extent, they contain similar information to the lexicon of an FST: words (lemmas and stems) and their continuation lexicons that indicate how they are inflected. For this reason, we actually compile the FST lexicon from the XML-based dictionary. The XML dictionaries are, in fact, also used as source files for online dictionaries, and therefore they are an attempt to address the matter of reuse, i.e. transducer construction, online morphological dictionaries, source material for ICALL projects as well as rule-based machine translation.

With a recent effort of a TEI (Text-Encoding Initiative) compatibility layer in the GiellaLT XML dictionaries (Rueter and Hämäläinen, 2019), it was decided that editing the FSTs in the XML dictionaries should be open for a majority of people using modern lexicographic tools. This compatibility layer is important as TEI is an ISO-standard, which means that it will most likely outlive any individual project and remain usable in the future as well.

The second layer of abstraction is the open-source online user interfaces Ve'rdd (Alnajjar et al., 2020) and Akusanat (Hämäläinen and Rueter, 2018) that make it possible to edit XML based dictionaries for anyone without any technical background. This is important since direct edits to an XML might be daunting for a language community member who has no programming background. These two different levels of abstraction ultimately produce XMLs in the GiellaLT format that can be directly used to enhance the transducers.

To provide further comparison, a very similar mechanism to store lexicon externally from the FST has also been used successfully by Wilbur (2018) with Pite Saami. Also in this case same lexicon is used in FST, in a published dictionary (Wilbur, 2016), and in an online dictionary[4]. This shows that approach described here is practical,

[4] https://saami.uni-freiburg.de/psdp/ordbok/

and already used, in different endangered language contexts.

5 Conclusions

In this paper we have presented our work on open-source FSTs for Erzya and Moksha. Leveraging existing open source platforms has been the core design principle throughout the development. Without GiellaLT support, our transducer would be left out of a plethora of higher level services provided by the infrastructure such as spell-checking, constraint grammar based syntax, language learning tools etc. This would mean that we would need to build these resources from the ground up.

Another important design principle has been accessibility of the FSTs. Continuous integration makes it possible to commit to the nightly builds of the UralicNLP library, which makes our FSTs usable through a generic multilingual Python API. Accessibility has also been thought of from the point of view of the development. Using an XML based dictionary to generate the FST lexicon with a compatibility with the ISO-standardized TEI XML makes it easier for non-FST savvy people to contribute to the work. Furthermore, integration with a GUI such as Ve'rdd, makes it possible to crowdsource the development in the future by evoking community involvement of the native speakers.

We have had positive experiences from building on top of these other open-source solutions and would strongly recommend that other researchers working on the morphology of endangered languages investigate the existing open-source platforms before starting to build everything from scratch. The GiellaLT infrastructure could save you from two to three years of development. At the same time attention has to be paid to general APIs and interfaces that allow access to these tools at various levels of abstraction.

References

Kuzma Abramov. 2002. Валонь ёвтнема валкс. Mordovskoj knizhnoj izdateljstvasj. The manuscript of this dictionary was compiled by the Erzya national writer Kuz'ma Grigorievich Abramov, 1914-2008, whose activities as an Erzya writer spanned nearly 70 years.

Cyril Allauzen, Michael Riley, Johan Schalkwyk, Wojciech Skut, and Mehryar Mohri. 2007. Openfst: A general and efficient weighted finite-state transducer library. In *Proceedings of the Ninth International Conference on Implementation and Application of Automata, (CIAA 2007)*, volume 4783 of *Lecture Notes in Computer Science*, pages 11–23. Springer.

Khalid Alnajjar, Mika Hämäläinen, and Jack Rueter. 2020. On editing dictionaries for uralic languages in an online environment. In *Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages*, pages 26–30.

Khalid Alnajjar, Leo Leppänen, and Hannu Toivonen. 2019. No time like the present: methods for generating colourful and factual multilingual news headlines. In *Proceedings of the 10th International Conference on Computational Creativity*. Association for Computational Creativity.

Lene Antonsen and Chiara Argese. 2018. Using authentic texts for grammar exercises for a minority language. In *Proceedings of the 7th workshop on NLP for Computer Assisted Language Learning*, pages 1–9.

Lene Antonsen, Trond Trosterud, and Linda Wiechetek. 2010. Reusing grammatical resources for new languages. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*.

Jeff Ens, Mika Hämäläinen, Jack Rueter, and Philippe Pasquier. 2019. Morphosyntactic disambiguation in an endangered language setting. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 345–349.

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.

Herr Conon von der Gabelentz. 1839. Versuch einer mordwinischen grammatik. In *Zeitschrift für die Kunde des Morgenlandes.*, II. 2–3., pages 235–284, 383–419. Druck und Verlag der Dieterlichschen Buchhandlung.

Riho Grünthal. 2016. *Transitivity in Erzya: Second language speakers in a grammatical focus*, Uralica Helsingiensia, page 291–318. Finno-Ugrian Society, Finland.

Mika Hämäläinen. 2018. Poem machine-a co-creative nlg web application for poem writing. In *The 11th International Conference on Natural Language Generation Proceedings of the Conference*. The Association for Computational Linguistics.

Mika Hämäläinen and Jack Rueter. 2018. Advances in synchronized xml-media wiki dictionary development in the context of endangered uralic languages. In *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts 17-21 July 2018, Ljubljana*. Ljubljana University Press.

Arja Hamari and Niina Aasmäe. 2015. Negation in erzya. *Negation in Uralic languages*, 108:293.

Mans Hulden. 2009. Foma: a finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 29–32. Association for Computational Linguistics.

Benjamin Hunt, Emily Chen, Sylvia L.R. Schreiner, and Lane Schwartz. 2019. Community lexical access for an endangered polysynthetic language: An electronic dictionary for st. lawrence island yupik. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 122–126, Minneapolis, Minnesota. Association for Computational Linguistics.

Mika Hämäläinen. 2019. UralicNLP: An NLP library for Uralic languages. *Journal of Open Source Software*, 4(37):1345.

Kerry Jones and Sanjin Muftic. 2020. Endangered African languages featured in a digital collection: The case of the Ç,Khomani San, Hugh Brody Collection. In *Proceedings of the first workshop on Resources for African Indigenous Languages*, pages 1–8, Marseille, France. European Language Resources Association (ELRA).

Lauri Karttunen, Tamás Gaál, and André Kempe. 1997. Xerox finite-state tool. *Rapport technique, Centre de recherche Xerox de Grenoble.*

Egor Kashkin and Sofya Nikiforova. 2015. Verbs of sound in the moksha language: a typological account. *Nyelvtudományi Közlemények*, 111:341–362.

László Keresztes. 1999. *Development of Mordvin definite conjugation.* Suomalais-Ugrilaisen Seuran toimituksia, 233. Suomalais-Ugrilainen Seura, Helsinki.

Krister Lindén, Miikka Silfverberg, and Tommi Pirinen. 2009. Hfst tools for morphology–an efficient open-source package for construction of morphological analyzers. In *International Workshop on Systems and Frameworks for Computational Morphology*, pages 28–47. Springer.

Olga Lovick, Christopher Cox, Miikka Silfverberg, Antti Arppe, and Mans Hulden. 2018. A computational architecture for the morphology of upper tanana. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jorma Luutonen. 2014. Kahden sukupolven ersää – kielenhuoltoa ja muutoksen merkkejä. *Memoires de la Societe Finno-Ougrienne*, 270:187—201.

Eetu Mäkelä, Krista Lagus, Leo Lahti, Tanja Säily, Mikko Tolonen, Mika Hämäläinen, Samuli Kaislaniemi, and Terttu Nevalainen. 2020. Wrangling with non-standard data. In *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference*.

Christopher Moseley, editor. 2010. *Atlas of the World′s Languages in Danger*, 3rd edition. UNESCO Publishing. Online version: http://www.unesco.org/languages-atlas/.

Sjur Moshagen, Jack Rueter, Tommi Pirinen, Trond Trosterud, and Francis M. Tyers. 2014. Open-source infrastructures for collaborative work on under-resourced languages. The LREC 2014 Workshop “CCURL 2014 - Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era”.

Arbi Haza Nasution, Yohei Murakami, and Toru Ishida. 2018. Designing a collaborative process to create bilingual dictionaries of Indonesian ethnic languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Pavel Ornatov. 1838. *Mordovskaja grammatika / sostavlennaja na narechij mordvy mokshi Pavlom Ornatovym.* V Sinodalnoj tip., Moskva.

Tommi A Pirinen and Francis M Tyers. 2012. Compiling apertium morphological dictionaries with hfst and using them in hfst applications. *Language Technology for Normalisation of Less-Resourced Languages*, page 25.

Jack Rueter. 2010. *Adnominal person in the morphological system of Erzya*. Suomalais-ugrilaisen seuran toimituksia. Suomalais-Ugrilainen Seura, Finland.

Jack Rueter and Mika Hämäläinen. 2019. On xml-mediawiki resources, endangered languages and tei compatibility, multilingual dictionaries for endangered languages. *Rachel Edita O. ROXAS President National University (The Philippines)*, page 350.

Jack Rueter, Niko Partanen, and Larisa Ponomareva. 2020. On the questions in developing computational infrastructure for komi-permyak. In *Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages*, pages 15–25.

Jack M. Rueter. 2000. Xeljsinkisa universitetsa kyv tujalysj izhkaryn perymsa kyvjas simpozium vylyn lyddjomtor. In *V sbornike Permistika 6: Problemy sinxronii i diaxronii permskix jazykov i ix dialektov [Permistika 6: Problems in the synchrony and diachrony of the Permic languages and their dialects]*, volume 6 of *Permistika*, pages 154–158.

Jack Michael Rueter. 2016. *Towards a systematic characterization of dialect variation in the Erzya-speaking world: Isoglosses and their reflexes attested in and around the Dubyonki Raion*, number 10 in Uralica Helsingiensia, pages 109–148. University of Helsinki, Finland.

Jack Michael Rueter and Francis M Tyers. 2018. Towards an open-source universal-dependency treebank for erzya. In *International Workshop for Computational Linguistics of Uralic Languages*.

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. Volume: Proceeding volume: 2.

Trond Trosterud. 2004. Porting morphological analysis and disambiguation to new languages. In *SALTMIL Workshop at LREC 2004: First Steps in Language Documentation for Minority Languages*, pages 90–92. Citeseer.

Trond Trosterud. 2006. *Homonymy in the Uralic Two-Argument Agreement Paradigms*. Suomalais-Ugrilaisen Seuran Toimituksia 251. Suomalais-Ugrilainen Seura, Helsinki.

Francis M. Tyers, Jonathan Washington, Darya Kavitskaya, Memduh Gökırmak, Nick Howell, and Remziye Berberova. 2019. A biscriptual morphological transducer for crimean Tatar. In *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 74–80, Honolulu. Association for Computational Linguistics.

Linda Wiechetek, Sjur Moshagen, Børre Gaup, and Thomas Omma. 2019. Many shades of grammar checking – launching a constraint grammar tool for north sámi. In *Proceedings of the NoDaLiDa 2019 Workshop on Constraint Grammar - Methods, Tools and Applications*.

Joshua Wilbur. 2016. *Pitesamisk ordbok: samt stavingsregler*. Department of Scandinavian Studies, University of Freiburg.

Joshua Wilbur. 2018. Extracting inflectional class assignment in pite saami: Nouns, verbs and those pesky adjectives. In *Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages*, pages 154–168.

Anna Zueva, Anastasia Kuznetsova, and Francis Tyers. 2020. A finite-state morphological analyser for evenki. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2581–2589, Marseille, France. European Language Resources Association.