

Restructuring and visualising dialect dictionary data: Report on Erzya and Moksha materials

Niko Partanen

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

Jack Rueter

University of Helsinki
Department of Digital Humanities /
Language Bank of Finland
jack.rueter@helsinki.fi

Abstract

There are a number of Uralic dialect dictionaries based on fieldwork documentation of individual minority languages from the Pre-Soviet Era. In this article, we describe our methods, where we reuse dialect dictionary data in XML format, and visualize phonetic variants as linguistic isoglosses using a web application. The methods can be extended to other languages using a simple tabular structure. Our approach and application is suitable only for visualizing a small portion of the data present in large linguistic collections such as a dialect dictionary, and different tools must eventually be combined. However, simple and light applications appear to be a good solution as they are easily extended as needed.

1 Introduction

The dictionaries of endangered languages are very valuable in contemporary research. Many dictionaries, however, are not available digitally, and if they are, they may not have OCR accuracy that would make them fully searchable. The mere size and extent of dictionaries in large majority languages can make them challenging to process. Especially the work done with the Transkribus platform (Kahle et al., 2017) has made high quality text recognition available to an exceptionally large community. At the same time, the successful recognition of diacritical marks has opened many new avenues for further work on texts written using Finno-Ugric transcription, as reported by Partanen et al. (2022). The field is clearly moving toward the point where many dialect dictionaries will become digitally available.

Dialect dictionaries, however, present a relatively complicated data type, as the internal data structures are not always easily retrievable from the printed text, especially if we do not have all the formatting. Part of what contributes to this challenge is that the traditional dictionaries contain many different types of data: various derivations,

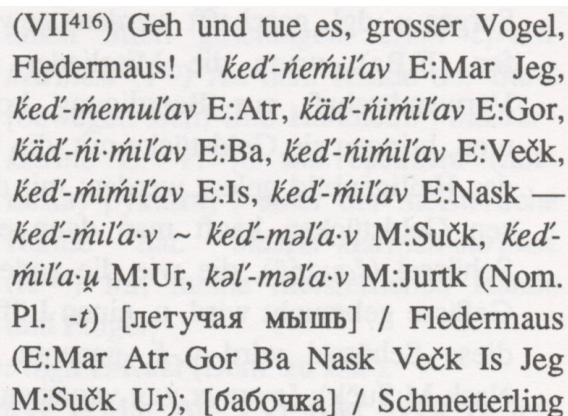


Figure 1: Example of an embedded word article in H. Paasonens Mordwinisches Wörterbuch. Band II (K-M) (Heikkilä et al., 1992, 678)

compounds, dialectal variants and example sentences, all appearing with various components of geographic data. The example in Figure 1 illustrates how the forms of the word *Fledermaus* ‘bat’ are presented inside a larger macro-article, and the geographic locations are presented with abbreviations. This data is very well structured and detailed, but it is organized for a printed dictionary.

When this data can be rendered in new ways, disconnected from the original layout of the printed pages, many new research questions and methods begin to appear. Data visualizations and interactive applications are often seen in the digital humanities, which, in many ways, are elementary for understanding the structures of more complex datasets.

In this study, we introduce methods and application we have developed to visualize and inspect geographically coded Erzya and Moksha dictionary data. The example application is built in the R language’s (R Core Team, 2021) Shiny framework (Chang et al., 2022), and is hosted on CSC – IT Center for Science’s Rahti service. Eventually we plan to host the application and store the data in the Language Bank of Finland.

Shiny is a framework for building and testing interactive web applications that can execute R code. In our opinion, Shiny is a very suitable tool for rapid prototyping, but we do acknowledge that different approaches should be investigated for long term deployment. Our application uses Leaflet JavaScript library through R's leaflet package (Cheng et al., 2024) and datatables JavaScript library through R's DT package (Xie et al., 2024), which all generate JavaScript, but the visualization is controlled through interactive R session within Shiny framework. There is some unnecessary overhead in this solution, as the same could be achieved with JavaScript alone. Yet, as the application in itself is fairly simple and very easy to maintain in the current form, this framework serves the current needs very well. All code is openly available in GitHub, with the documentation of our hosting solutions and most up-to-date URL.¹

The working model we have developed and present here connects especially to situations where we have the original dictionary as some kind of a digital file that contains the original formatting, or where the original formatting can be retrieved one way or another. This differs drastically from situations where the dictionary data is available as a database or within some software regularly used in dictionary compilation task. However, our situation is very realistic, as many printed dictionaries can be found in formats such as digital print files, text documents of some type, or we may have a version where text is retrieved through text recognition and, ideally, proofread carefully. If the data already exists in a database or other digital structure, importing it to our application would also be a trivial task.

2 The Erzya and Moksha dialect dictionary

The Erzya and Moksha dialect materials used in this dictionary represent fieldwork collections organized or performed by Heikki Paasonen at the end of the 1800s and beginning of the 1900s. Geographically, the fieldwork was extensive, representing over 200 collection points for the two languages combined. There is an inconsistency, however, in the representation of language materials from the various collection points, i.e., whereas there are 10,737 phonetically documented word forms found

for the Erzya village of *Marisevo*, on the one hand, there are only four word forms attributed to the Erzya village of *Kabaevo* (see Rueter, 2016, 134), on the other. Similar figures can be presented for Moksha, too. It may also be noted that the geographic granularity of Moksha-language collection represents a different level of polygons, i.e., Erzya materials appear to have more village-level representation, whereas their Moksha counterparts might be more readily associated with a *volost'* or *raion*-level representation.

Despite these shortcomings, the 'Mordvin Dialect Dictionary' is, in fact, the most extensive documentation of Erzya and Moksha vocabulary published. The materials come from the Mordwinisches Wörterbuch 'Mordvin Dictionary' 1990–1999 (Heikkilä et al., 1990, Heikkilä et al., 1992, Heikkilä et al., 1994, Heikkilä et al., 1996, Heikkilä and Kahla, 1998, Heikkilä and Kahla, 1999) based on the Heikki Paasonen works and collections (Paasonen, 1891, Paasonen, 1894, Paasonen, 1909, Paasonen, 1938, Paasonen, 1939, Paasonen, 1941, Paasonen, 1947, Paasonen, 1977a, Paasonen, 1977b, Paasonen, 1980, Paasonen, 1981).

The dictionary data were originally fed into a desktop in the 1980s and 1990s, and the resulting materials were converted into an XML UNICODE document based on style, size and font parameters. Even though there was a high consistency in the usage of fonts for distinguishing what nowadays could be handled with UNICODE ranges, some of the same problems that confound us today also occurred, namely, language abbreviations such as Erzya (E) and Moksha (M) and other look-alikes frequently required correction before the different languages and dictionary structure abbreviation data were clean.

The 2,703-page dictionary consists of 6,952 macro articles, each of which represents a distinct word root. The macro articles can be divided further into 21,754 stem word entries, which range in complexity from a single-stem article with Russian and German translations to a macro article containing multiple-stem articles with additional compound-word articles and etymologies.

In more complex articles, it becomes apparent that a stem word article can distinguish three separate sections where collection point data are mentioned. These sections are phonetic variants of a given cognate, the definitions, which may vary from place to place, and example contexts. Occa-

¹<https://github.com/rueter/Dictionary-Map-Viewer>

sionally, semantic cross-referencing is made where a word from one collection point may be associated with an entirely different word from another collection point. Etymological references indicate cognates in other languages, although a majority of the Uralic cognates or parallel forms were left out of the final version of the printed dictionary.

3 Related work

We contextualize our work within the cartography of the Uralic languages, and geographic visualization of the language documentation data and traditional fieldwork based data collected in early modern times. We do not position our work strongly toward dialect geography or research of geographic variation, primarily as we mainly have worked on visualization of the collected or already published data, but do not address how this data would be further used in research on these dialects and their variation. Naturally, the same datasets that are used in these applications can be also used in a wide array of different research purposes. The primary purpose of our application is to allow visual inspection of the data, helping to understand underlying geographic distributions and what kind of possible gaps or other structures there are. Naturally, the application may be extended in the future to allow more complicated tasks.

One problem with visualizing dialect data is that tools intended for semi-professional or professional cartographers, although powerful, are very specialized, have a high learning curve and are heavy to run. At the same time, the data we process when comparing dialectal variants is relatively simple. [Gawne and Ring \(2016\)](#) reviewed a number of light and practical programs that could be used in this task, and we believe their suggestions and observations are relevant also today. Another wide survey of visualization platforms was presented by [Roose et al. \(2021\)](#).

In the context of Uralic languages, the recent cartographic work by [Rantanen et al. \(2022\)](#) has been very important, as they have produced openly licensed maps about the distribution of the Uralic languages. As they primarily operate with polygon level, expressing the language areas, our work is very complementary to theirs. As a hypothesis, the collection points of the Paasonen's dictionary data should fall within the traditional Erzya and Moksha areas as shown in the dataset of [Rantanen et al. \(2021\)](#).

Indeed, it seems obvious that with rapidly changing technology we will be using new tools of visualization and analysis each decade. However, the dialectal data in itself remains valuable, even increasing in value as the data can be extended with other resources that deepen the geographic and temporal coverage. From this point of view the visualization is in all ways secondary, and the underlying data the key element.

4 Data Structure

We restructure the dialect dictionary entries so that each entry in the derived structure contains only individual word forms and their dialectal variants. We then introduce literary-language lemmas to plot the entries as individual items on a map. We separate the management of lexical data and coordinates, so that the 'location' connects the lexemes with their coordinates. This allows that the coordinate data can be stored in a separate table or in other format that is independent from the lexical data and does not need to be modified in several places at once. Similar structure was used also by [Gawne and Ring \(2016, 207\)](#).

We use columns 'base_form', 'variant', 'location' and 'language'. To manage the lexical data. Additional column that will need to be added when the materials are combined from various sources is 'source'. The Erzya and Moksha data could at later point be appended with contemporary dialect data, and the source for this information would then differ from Paasonen's. The data about the source will be stored in an additional table, as it contains information about the collection time, authors and correct citations. At the moment the application contains data only from one source, so the references can be stored at a higher level.

With the column 'base_form' we are currently rather free on what kind of content should be placed there. It is not possible to decide on one base lexeme that would match for both Erzya and Moksha, but especially when the visualization contains data from just one language, this seems like an easiest alternative. For the Erzya and Moksha application we also have added numbers for each lexeme, but we do not believe this is the best solution going forward. One possible approach is to use as the base form a descriptive translation that would then also be used to select the current lexeme for viewing.

In the original structure of the Paasonen's digitized dictionary, derivation articles are child articles

of a single macro article, and compound words are addressed in grandchild articles. Although various parts are connected to one another, they are related to different semantical lexical items, and compound word items may be mentioned as grandchild articles under each component macro article. The relations between root words (the first child article of the macro article), derivations (non-first child articles of the macro article), compound words and multiword contexts (grandchild articles of the macro article) are retained in the nested XML structure, so the original structure can, if needed, be retrieved.

We start by modifying the XML structure of the digitized dictionary where the layout has been converted to tagged elements.² This situation is very specific to the dictionary presently under inspection, and it does not necessarily serve as a model for further work. The data is read and rendered as a tabular structure where one row is one word form from one location. This structure is versatile for cases where there are different amounts of data from different locations for different words.

5 Application

We have currently set up two application prototypes. One visualizes the Erzya and Moksha data, and another serves as an editor for the data in our structure.³ The editor is very much at the preliminary testing stage, but it allows uploading and downloading the files that can be edited also locally. As shown in Figure 2, the application interface contains multiple elements. These are described below.

The application has a selection part on upper left corner where the user can browse the entries according to their German translations. Under that basic information about the word form is presented. All variants attested are displayed on an interactive map in the middle. The map is fully interactive, and when a dot is clicked, we see the name of the location, attested dialectal word form and the language (Erzya or Moksha). Under the map there is a table that displays all data rows. The table can be searched and filtered.

The example sentences connected to the entry are not currently displayed, but this could be added at a later stage. They are often attested for indi-

vidual locations, but we see it currently as an open question how to best display them. One possibility one could be to show all of them below other interface components.

Each entry with the same phonetic representation is coded with the same color on the map. This makes it easy to see different patterns and compare these word for word. The colors are currently selected automatically from a predefined palette.

Occasionally, the same collection point may have more than one dialectal variant. Here it was important that we apply a jitter function to both latitude and longitude readings that moves all points a bit randomly, so overlying data can be displayed. This does not seem to cause loss of information at the scale where this data operates. With locations very close to one another the impact of jitter should be monitored and checked, but the currently used values are effective for the data at hand. Another approach, tested by our collaborator Cinthia Ishida (Federal University of Pará), would be to overlay different shapes in these situations.

6 Conclusion

In the future the application will be extended for use with other dictionaries, especially for the Uralic languages, for which the dialect dictionaries have been created within the same research tradition. At the same time we are participating in a collaboration between researchers of the Uralic languages and the languages of the Amazonian region. This will allow for more extensive testing and will possibly necessitate adjustments for some additional information present in them. At the same time we aim to keep the structure simple enough so that different dictionary types can be readily used as data sources. It is not our goal, however, to visualize all the possible information in the original dictionary in one application. Instead, we envision that some of the same data might be transformed in various ways and displayed in applications that are more suitable for the aspects one is interested in. However, overlaying various different data types or displaying several maps side by side would be one feature that is so central for the usability of the application, that we may integrate this functionality very rapidly. At the same time our flexible data model makes it easy to reuse the same data in other novel environments as needed.

²Example of the original XML structure can be found in our GitHub repository: <https://github.com/rueter/Dictionary-Map-Viewer/tree/main/data>

³For the editor application, see: <https://github.com/nikopartanen/Dictionary-Map-Editor>

Mordvin Dialect Map

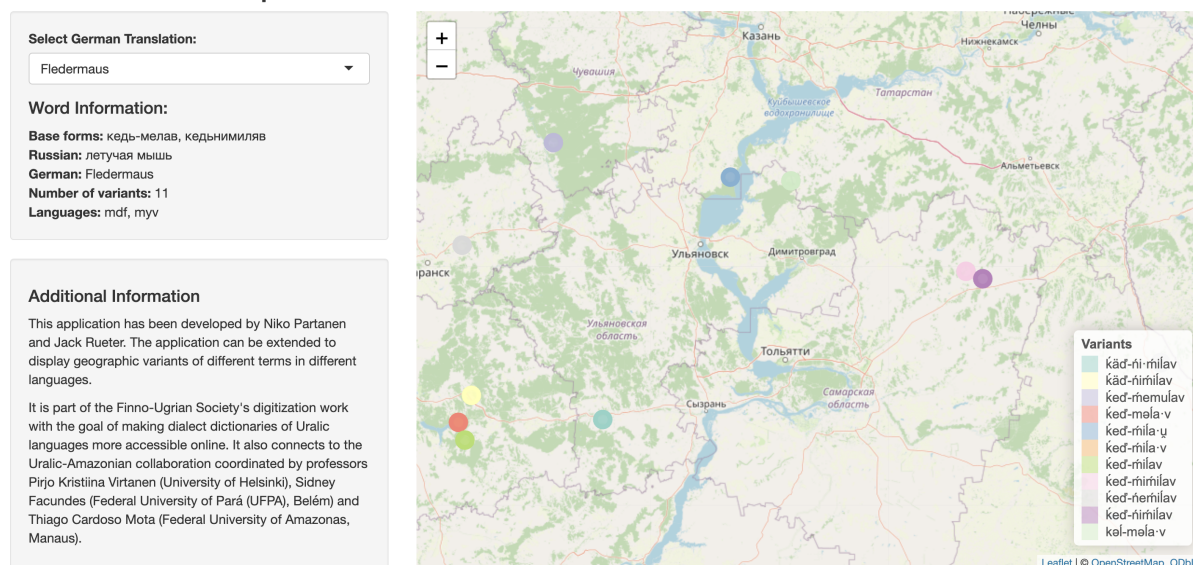


Figure 2: Screenshot of the application interface

Limitations

Our work with the visualization application has revealed continuity issues within our test dictionary. In fact, we have come to see that the ‘H. Paasonen Mordvin Dictionary’ material illustrates many of the shortcomings present in Uralic dialect dictionaries in general. First, while the dictionary distinguishes over 200 collection points, there are virtually no entries representing even 40 collection points. Second, while the collection points are distinguished with association to their literary language, none of the original articles make reference to literary-language word forms. Third, the individual word form articles make separate reference collection points in three divisions of a given article. They appear in the phonetic variant section; the definitions section, and the examples section, which, ideally, would have been aligned with the individual definitions.

These limitations actually point to the need for rendering the ‘H. Paasonen Mordvin Dictionary’ as a part of a maintained database for lexical research in the Mordvin languages. Such a database would make it possible to elaborate the representation of lesser represented collection points from within the H. Paasonen materials, on the one hand, and introduce collections from other times and geographic points, on the other. Such work would greatly help in the documentation of the two literary languages and even lesser documented Mordvin language forms. Furthermore, analogous work

could be envisioned for the development of work with other Uralic languages.

Currently the application is not ideal for displaying lexemes that have very extensive dialectal variation. This can be partially mitigated with a color palette, but tens of different colors are not visually easily distinguishable.

What we could currently recommend is to encode the data with coarser granularity. This would conceivably work well with data from sedentary communities in larger monolithic geographical settings, where no new settlements have been introduced. In the instances of settlements left of the Volga, however, we cannot assume large distributions of monolithic language variants, as this region has been subject to resettlement by different language groups and even different variants of the Mordvin languages. Thus, we are still looking for an ideal solution. Another way to approach coarser granularity, in this context, would be to break down distinct phonetic differences in a given word form and make several interlinked maps to illustrate the phenomena observed there. A good example might be seen in forms of the word for ‘butterfly’, where the separate maps could address first syllable vowel, stress placement, vocalization of the final /v/, palatalization of the central /m/ and so on. By addressing each phenomenon as a separate issue, we are able to reduce the number of variants, thus minimizing the color-coded distinctions required in an individual map.

Ethics Statement

We work with materials that have already been published and have undergone a rigorous editing process. We acknowledge that the material is part of the cultural heritage of the Erzya and Moksha people, and therefore steps have been taken to ensure the accessibility and availability of these materials to the language and research communities.

Acknowledgements

The work with this dictionary and application is part of the Finno-Ugrian Society's digitization work with the goal of making dialect dictionaries and other materials published by the Society more accessible online. We thank the two anonymous reviewers for their careful reading of our manuscript and their many valuable comments and suggestions. We also want to acknowledge and thank the Uralic-Amazonian collaboration coordinated by professors Pirjo Kristiina Virtanen (University of Helsinki), Sidney Facundes (Federal University of Pará (UFPA), Belém) and Thiago Cardoso Mota (Federal University of Amazonas, Manaus). The authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources. Very importantly, having been able to test the application and collect feedback as part of the authors' teaching in Belém at UFPA has been very important for us and improved the result in numerous ways.

References

- Winston Chang, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barbara Borges. 2022. *shiny: Web Application Framework for R*. R package version 1.7.2.
- Joe Cheng, Barret Schloerke, Bhaskar Karambelkar, and Yihui Xie. 2024. *leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library*. R package version 2.2.2.
- Lauren Gawne and Hiram Ring. 2016. Mapmaking for language documentation and description. *Language Documentation and Conservation*, 10:188–242.
- Kaino Heikkilä, Hans-Hermann Bartens, Aleksandr Feoktistow, Grigori Jermuschkin, and Martti Kahla, editors. 1990. *H. Paasonens Mordwinisches Wörterbuch. Band I (A-J)*, volume XXIII:1 of *Lexica Societatis Fenno-Ugricae XXIII & Kotimaisten kielten tutkimuskeskuksen julkaisuja 59*. Suomalais-Ugrilainen Seura & Kotimaisten kielten tutkimuskeskus. Zusammengestellt von Kaino Heikkilä. Unter Mitarbeit von Hans-Hermann Bartens, Aleksandr Feoktistow und Grigori Jermuschkin bearbeitet und herausgegeben von Martti Kahla.
- Kaino Heikkilä, Hans-Hermann Bartens, Aleksandr Feoktistow, Grigori Jermuschkin, and Martti Kahla, editors. 1992. *H. Paasonens Mordwinisches Wörterbuch. Band II (K-M)*, volume XXIII:2 of *Lexica Societatis Fenno-Ugricae XXIII & Kotimaisten kielten tutkimuskeskuksen julkaisuja 59*. Suomalais-Ugrilainen Seura & Kotimaisten kielten tutkimuskeskus.
- Kaino Heikkilä, Hans-Hermann Bartens, Aleksandr Feoktistow, Grigori Jermuschkin, and Martti Kahla, editors. 1994. *H. Paasonens Mordwinisches Wörterbuch. Band III (N-Ŕ)*, volume XXIII:3 of *Lexica Societatis Fenno-Ugricae XXIII & Kotimaisten kielten tutkimuskeskuksen julkaisuja 59*. Suomalais-Ugrilainen Seura & Kotimaisten kielten tutkimuskeskus.
- Kaino Heikkilä, Hans-Hermann Bartens, Aleksandr Feoktistow, Grigori Jermuschkin, and Martti Kahla, editors. 1996. *H. Paasonens Mordwinisches Wörterbuch. Band IV (S-Ž)*, volume XXIII:4 of *Lexica Societatis Fenno-Ugricae XXIII & Kotimaisten kielten tutkimuskeskuksen julkaisuja 59*. Suomalais-Ugrilainen Seura & Kotimaisten kielten tutkimuskeskus.
- Kaino Heikkilä and Martti Kahla, editors. 1998. *H. Paasonens Mordwinisches Wörterbuch. Band V: Russischer Index*, volume XXIII:5 of *Lexica Societatis Fenno-Ugricae*. Finno-Ugrian Society.
- Kaino Heikkilä and Martti Kahla, editors. 1999. *H. Paasonens Mordwinisches Wörterbuch. Band VI: Deutscher Index*, volume XXIII:6 of *Lexica Societatis Fenno-Ugricae*. Finno-Ugrian Society.
- Philip Kahle, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. 2017. Transkribus—a service platform for transcription, recognition and retrieval of historical documents. In *2017 14th iapr international conference on document analysis and recognition (icdar)*, volume 4, pages 19–24. IEEE.
- Heikki Paasonen. 1891. *Proben der mordwinischen Volkslitteratur. I. Band. H. 1*, volume 9 of *Journal de la Société Finno-Ougrienne*. Finno-Ugrian Society.
- Heikki Paasonen. 1894. *Proben der mordwinischen Volkslitteratur. I. Band. H. 2*, volume 12 of *Journal de la Société Finno-Ougrienne*. Finno-Ugrian Society.
- Heikki Paasonen. 1909. *Mordwinische Chrestomathie mit Glossar und grammatikalishcem Abriß*, volume 4 of *Apuneuvoja suomalais-ugrilaisten kielten opintoja varten — Hilfsmittel für das Studium der finnisch-ugrischen Sprachen*. Finno-Ugrian Society.
- Heikki Paasonen. 1938. *Mordwinische Volksdichtung: Band I*, volume LXXVII of *Mémoires de la Société Finno-Ougrienne*. Finno-Ugrian Society, Helsinki.

- Heikki Paasonen. 1939. *Mordwinische Volksdichtung: Band II*, volume LXXXI of *Mémoires de la Société Finno-Ougrienne*. Finno-Ugrian Society, Helsinki.
- Heikki Paasonen. 1941. *Mordwinische Volksdichtung: Band III*, volume LXXXIV of *Mémoires de la Société Finno-Ougrienne*. Finno-Ugrian Society, Helsinki.
- Heikki Paasonen. 1947. *Mordwinische Volksdichtung: Band IV*, volume XCI of *Mémoires de la Société Finno-Ougrienne*. Finno-Ugrian Society, Helsinki.
- Heikki Paasonen. 1977a. *Mordwinische Volksdichtung: Band V*, volume 161 of *Mémoires de la Société Finno-Ougrienne*. Finno-Ugrian Society, Helsinki.
- Heikki Paasonen. 1977b. *Mordwinische Volksdichtung: Band VI*, volume 162 of *Mémoires de la Société Finno-Ougrienne*. Finno-Ugrian Society, Helsinki.
- Heikki Paasonen. 1980. *Mordwinische Volksdichtung: Band VII*, volume 176 of *Mémoires de la Société Finno-Ougrienne*. Finno-Ugrian Society, Helsinki.
- Heikki Paasonen. 1981. *Mordwinische Volksdichtung: Band VIII*, volume 178 of *Mémoires de la Société Finno-Ougrienne*. Finno-Ugrian Society, Helsinki.
- Niko Partanen, Rogier Blokland, Michael Rießler, and Jack Rueter. 2022. Transforming archived resources with language technology: From manuscripts to language documentation. In *The 6th Digital Humanities in the Nordic and Baltic Countries 2022 Conference, Uppsala, Sweden, March 15-18, 2022*, pages 370–380. University of Oslo Library.
- R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Timo Rantanen, Harri Tolvanen, Meeli Roose, Jussi Ylikoski, and Outi Vesakoski. 2022. [Best practices for spatial language data harmonization, sharing and map creation—a case study of Uralic](#). *Plos one*, 17(6):e0269648.
- Timo Rantanen, Outi Vesakoski, Jussi Ylikoski, and Harri Tolvanen. 2021. [Geographical database of the Uralic languages \(v1.0\) \[data set\]](#).
- Meeli Roose, Tua Nylén, Harri Tolvanen, and Outi Vesakoski. 2021. User-centred design of multidisciplinary spatial data platforms for human-history research. *ISPRS International Journal of Geo-Information*, 10(7):467.
- Jack Rueter. 2016. Towards a systematic characterization of dialect variation in the Erzya-speaking world: Isoglosses and their reflexes attested in and around the Dubënki raion. In Ksenia Shagal and Heini Arjava, editors, *Mordvin Languages in the Field*, volume 10 of *Uralica Helsingiensia*, page 109–148. Finno-Ugrian Society.
- Yihui Xie, Joe Cheng, and Xianying Tan. 2024. *DT: A Wrapper of the JavaScript Library 'DataTables'*. R package version 0.33.