Journal of Data Mining and Digital Humanities

EPIsciences overlay journals

Dialect Cartography of Erzya and Moksha Languages: Digitized Historical Sources and Evaluation of the Contemporary Data

Niko Partanen[1] and Jack Rueter[1]

[1]University of Helsinki, Finland

Corresponding author: Niko Partanen , niko.partanen@helsinki.fi

Abstract

This study investigates the correspondences between a recent map of Uralic languages that also covers the Erzya and Moksha languages in detail. We discuss our point of view in linguistic cartography more generally, but especially within the context of Uralic languages, and address various difficulties that can be recognized in defining the speaker area boundaries and choosing settlements that should be included in the traditional or contemporary speech communities. We use the historical data of Heikki Paasonen, which, we believe, is a highly reliable indicator of at least some areas that should be included in the traditional distributions of these languages as points of comparison. This data is contrasted with the contemporary language maps.

Keywords

Uralic languages; Erzya; Moksha; language maps; dialectology

I INTRODUCTION

There is a long history of dialect cartography for Uralic languages, at least in connection with the rich fieldwork tradition of Uralic studies. When data has been collected from different localities, it implicitly creates an understanding of where these languages are spoken. These matters may also have real political consequences if it becomes disputed where each ethnic group has traditionally resided and what regions are their home regions. Very early publications have included language and dialect maps, and published text collections have always come with information about the settlements from which different texts have been collected and where the speakers represented there originated. This documentary data inevitably bears some cartographical evidence.

For many purposes, representing language areas as polygons is a good choice. Polygons are areas made up of line segments, and when displayed on a map they cover some geographic region. This works particularly well when we are working with languages whose speakers lead a nomadic lifestyle, where connecting the locations with specific dots may not give the correct picture. Polygons also allow some flexibility, i.e., if the exact areas where the language is spoken are not known, we can still use polygons for the rough representation of the correct region. Of course, one essential reason for using polygons in language maps is aesthetic: polygons scale very well to different zooming levels, and they can be perceived as visually more pleasant than individual dots. However, a lot of this depends on map design, and there are examples where the dots also have worked particularly well. At the same time, using dots we may avoid issues related to often necessary divisions and break-ups of terrain when polygons are used.

There are, however, Uralic languages such as Erzya and Moksha which are spoken in well established traditional agrarian settlements. These settlements are part of Western society and the Western scientific sphere, and they are often well established in different mapping services, such as OpenStreetMap and Google Maps. Even those settlements that no longer exist at the moment are often marked in OpenStreetMap as abandoned localities, as these places are still known in the collective memory of their language communities and may still have some elderly residents, at least who still know where these places were located.

More than five years ago, Rueter et al. [2020] published a Comparative Mordvin Database, which contained geographic locations of different Erzya and Moksha settlements. These derived maps are openly available and can be viewed online.[1] This project differs from the other ones in that no assumptions were made about geographical areas, but instead the focus was in locations where, according to some sources, Erzya and Moksha speakers live. Our sources include materials ranging from scientific publications to population censuses of the Russian Federation, e.g., we use Paasonen locale data in this study, which is part of the Mordvin-Varieties project, just for clarity. A new version of the database was published with the current study Rueter et al. [2025]

This collection of points in Russia is not in itself a language map, but more of a subset of Russian current and historical settlements and approximations of their locations, where we have appended information about Erzya and Moksha dialects, which, according to our understanding and knowledge, are spoken in those areas. Creating maps is one use of this dataset. In principle any Russian or European settlement may have individual Erzya and Moksha speakers, especially in contemporary times, our visualizations are moving very widely along the possible geographical region where Erzya and Moksha may be spoken, without tying it closely to specific regions or areas. One of the main goals of our settlement database work has been to allow visualizations of different corpora, i.e., when there is information about the birthplaces of individual authors, visualizing this could be useful for varying purposes. Of course, this kind of data has very extensive purposes, and our earlier work on dialect dictionary visualization is another good example [Rueter and Partanen, 2025]. Indeed, more comprehensive dialect isogloss visualizations would also be a logical next step when working with this data, and at this point we would be connecting dialectal features at a different level to these localities. In order to visualize occurrences in an individual corpora, instead of all Erzya and Moksha settlements, a more generic and larger database would probably be needed, and data sources such as Wikidata could also be used.

Since we mainly operate with settlements, there are some unique strengths in our approach. The settlements can be very accurately geocoded with contemporary computational tools and databases. We can thereby easily retrieve different names for them, and also acquire different related data: population, foundation year, date when first mentioned in the documents, higher level administrative entities, etc. With polygons, this is more complicated, even though we can always retrieve information about points that are included in a polygon, if needed. For this reason, the relationship of points and polygons is, in general, extremely important and relevant. Still, these different data representations can ideally strengthen one another, and both have their distinct advantages in different data visualisation environments.

[1] https://multilingualfacilitation.com/Mordvin-Varieties/

In 2021, Rantanen et al. [2021] published the Geographical database of the Uralic languages. This dataset is openly available in Zenodo, and the maps derived from this data have already had a very significant impact on the field of Uralic studies. The impact has primarily been illustrative. Many articles about these languages now have a visually pleasant and openly licensed map, which previously was often lacking. Rantanen et al. [2022a] and Rantanen et al. [2022b] describe, in detail, how the maps were originally created, especially in connection to the recently published Oxford Guide to the Uralic Languages, edited by Bakró-Nagy et al. [2022]. Another recent work that has been presented by Vesakoski et al. [2025] is connected to the larger infrastructure of speaker areas of Uralic languages.

Recently, Rueter and Partanen [2025] introduced a web application that allows displaying Erzya and Moksha lexical data as dialect maps. Each variant is associated with a location, and their distribution is coded automatically using different colors. The application is still under development, and the goal is to provide this feature for all entries in the Heikki Paasonen dialect dictionary: Heikkilä et al., 1990, Heikkilä et al., 1992, Heikkilä et al., 1994, Heikkilä et al., 1996. Rueter and Partanen [2025] mentioned in their study that adding the maps of Rantanen et al. [2021] to the application as a new layer would be a possible next step. This was tested in June 2025, resulting in novel observations that are discussed in this article. The goal of this work is to advance an open discussion on the geographical distribution of Uralic languages, and gradually improve the knowledge available about this topic.

II CURRENT STATE OF THE ART

The maps published in the URHIA project and discussed above indicate the currently highest state of the art. The maps were produced in a research project funded by the Academy of Finland in 2020-2022. Numerous specialists in Uralic languages were involved in the project, as well as GIS specialists. Rantanen et al. [2022a] describe how individual language experts expressed their opinions about the suitability of different data sources and areas displayed in them, and the experts themselves consulted numerous specialists and language speakers. Earlier maps and geographical sources were thoroughly studied and used as primary sources.

URHIA Erzya and Moksha map consists of three files: one for Erzya, one for Moksha, and one for areas where both Erzya and Moksha are spoken. For Erzya distribution, the sources are Ермушкин [1984], Feoktistow [1990], for Moksha Левина [2014], Feoktistow [1990], Feoktistov and Saarinen [2005], and for mixed area Feoktistow [1990] and Bartens [1999]. The maps in the original sources have been digitized, and they are also available in the dataset.

One of the main issues is that individual polygons are not named. They are marked for language variety, at times for the dialect, and source, among other attributes, but they do not have information about what locality they are intended to represent. Some sort of semantic convention would be useful in connecting polygons to other geographical entities. We understand that there are no predefined or widely used names for areas like these, but already something like ‘Erzya village cluster around settlement X’ would be enough. Or an explanation like ‘Moksha village cluster located in the North-Eastern part of raion X’. This would give the reader ways to tie the polygon into other sources, whereas at the moment there is just a region in the air under which we assume the language is spoken.

It must be mentioned that in some instances there are fairly specific dialect or other info comments, such as 'other info: Shoksha Erzya' or 'other info: Alatyr Erzya', the former referring to what elsewhere is known as Erzya mixed dialects, and the latter is given as another name for the North-Western dialect of Erzya. A problem here is that these descriptions and comments are mainly given for the Erzya and Moksha varieties spoken in the Republic of Mordovia and adjacent areas, which are fairly easily located and traditionally well described Mordvin varieties. This information would have been most useful in numerous diaspora regions, in which these comments are entirely absent. This shortcoming of minimal research outside of the Republic of Mordovia is seen in both local and foreign research, i.e., it has been difficult to coordinate regular meticulous research beyond the borders of the Republic.

Rantanen et al. [2022b] discuss how the maps, and especially the versions published in the recent handbook, were prepared. They emphasize that in their model the focus is on areas and not points that would represent the settlements. They also state that "The more detailed speaker areas are available for the more thoroughly mapped Western languages, which are also often languages with a larger number of speakers, such as Mordvin and many Finnic languages." Our current contribution will also be relevant in this light. It must also be noted that in the review of the handbook written by Kaheinen et al. [2024], the maps were highlighted as one of the more valuable contributions.

III DEFINING ERZYA AND MOKSHA SPEAKING AREAS

As discussed above, we base our conceptualization of where Erzya and Moksha are spoken on the locations of Erzya and Moksha settlements. Although larger settlements, towns and cities are inherently larger areas, these can still be represented fairly well by a point. Thus, if there is a known locality where these languages are spoken, we can represent it as a pair of coordinates. This is an approach starkly different from that used by Rantanen et al. [2022b], who state that their "language distributions are illustrated as speaker areas instead of being presented as points around the settlements", although leaving it unclear as to whether the underlying data of the language distribution is based on settlement data.

Another matter is how we define a locality where these languages are spoken. Especially in contemporary times, individual Erzya and Moksha speakers may be found in almost any settlement. It is clear that individual persons or families cannot be counted alone as speaking areas of these languages, and in some sense what we are most interested in here is what can be counted as traditional speaking areas, a factor which also operated as one option in Rantanen et al. [2021].

So, for the point of current exploration, we have chosen a subset of our data, placing special emphasis on the settlements where Erzya and Moksha have traditionally been spoken. We did this, first, by selecting all locations where Heikki Paasonen carried out fieldwork and collected data on these languages. It is possible that some of these localities are ones where Paasonen merely met Erzya and Moksha speakers who originally came from other places, although it should be possible to validate the locales in his materials citing what kind of settings these have been. In any case, it feels safe to assume that if it has been possible to collect large amounts of Erzya and Moksha folklore, folk music and dictionary data from some areas, then we can be fairly confident that these are indeed Erzya and Moksha speaking areas, at least by and large.

As a generalization, this does not always work. If, for example, we were to examine folklore materials gathered in the Nitra area of Slovakia, we would find that the richest materials for both Hungarian and Slovak come from virtually the same polygons. Some places, the soil is just richer.

Of course, the use of Heikki Paasonen's collections is limited to those localities where he visited. No arguement can be made about the localities that are outside the area where Paasonen worked, and similarly the presence of Paasonen's collection points does not exclude that Erzya and Moksha are also spoken somewhere else in the vicinity. Instead, it would probably be likely that the speaking areas are somewhat larger, and Paasonen did not exhaustively visit each settlement.

IV COMPARING ERZYA AND MOKSHA SETTLEMENT DATA AND URHIA MAP

We have analyzed each Erzya and Moksha polygon in the URHIA map and tried to connect them to Paasonen's settlement points presented in Rueter et al. [2020], while taking into account the limitations of the approach outlined above. There have been numerous instances that have required separate analysis. For example, the points and a polygon may coincide very clearly, in which case everything functions as expected.

In contrast, we also find instances where a polygon does not represent any Erzya or Moksha settlements known to us. This opens up the possibility that the polygon data refers to settlements that are not occupied any longer, or which have not been present in our data sources. For the most comprehensive understanding of the Erzya and Moksha spatiotemporal presence, analysing all these cases in detail is of extreme importance. What is problematic here is that, as explained above, the polygons do not have any identifiers or names that would have semantic significance. Therefore, it is not entirely clear how to decide which settlements they represent. We can see in the 2 how the point-based data collides with the polygons fairly well in some areas, but there are also wide mismatches. Especially in the whole Eastern part of the map displayed here where the points and polygons are by and large not in the same regions. When a more detailed view is taken, more nuances emerge. As those areas of the URHIA map that do not align at all with Paasonen's points may just represent different areas that Paasonen did not visit, we have chosen to focus in our analysis on specific details that can be analysed to some degree with the data at hand.

4.1 Mismatching polygons and points

These are exemplified in the Online Appendix Annotation 1. In some instances, it appears that a polygon of one map is close to a known settlement cluster of another that is not covered by any other polygon. In such cases, we must make an assumption that the polygon was intended to indicate these settlements. We can see this in Annotation 1 in the Figure 2. As the mismatch is only tens of kilometers, the issue is not very severe. However, if one were to try querying information about settlements under this polygon, the query would fail, since the settlements are outside the polygon. This illustrates how, even with a small mismatch between points and polygons, we lose one of the major benefits of using polygons.

From this point of view, the combination of point and polygon data would appear to be ideal in order to represent where these languages are spoken. However, as we can identify five settlement clusters in the central part of the map, south-east from Mordovia, and,

with all of these, the points and polygons are a little bit aside from one another, one has to ask whether the alignment and geo-referencing of the map has been perfectly accurate. However, we want to emphasize that these are still minor mismatches in these areas.

4.2 Settlements of Paasonen that have no polygons

These are exemplified in the Online Appendix Annotation 2. There are numerous instances of settlements where Heikki Paasonen collected large quantities of linguistic data, but which are not included in the URHIA maps. Even if the polygons had been somehow displaced at the point of georeferencing, the constellations are so different that there are clearly areas where Paasonen has collected data, but which are not included. This may partly be a matter of granularity, maybe some locations are so small that they were simply not associated with a polygon, but such decisions should have been made transparent somewhere. This situation is exemplified by the Annotation 2 in the Figure 2.

We would specifically want to draw attention to the area East from Mordovia, between the Samara Oblast and Tatarstan. This area contains a vast region of dense Erzya and Moksha population, which Paasonen has recorded. Only for the South-Eastern part of this area there is some alignment with the URHIA map. In our view the ideal path here would be to analyse all these settlements in detail, in order to understand which of them connect historically and linguistically together, and possibly assign them into some grouping polygons based on this information. If the whole area forms an unified whole, there could even be one larger polygon. Done this way, the polygons would already signal more complex and useful information beyond just indicating the possible existence of settlements under them.

4.3 Polygons, under which there are no settlements

These are exemplified in the Online Appendix Annotation 3. There are some polygons that do not match any points in Paasonen's data, or other points in Mordvinic Varieties in the geographic collection of Rueter et al. [2020]. We must assume that if there is a polygon on a map, there should be, or should have been, Erzya and Moksha localities. Even if the traditional speaking area is defined as the situation more than a hundred years ago, these settlements would have been expected to be present in the early 20th century, and if since abandoned, should be findable on some maps. As our data is not extensive, one cannot make any conclusions about this at this point, and refining the relationship of these polygons and known Erzya and Moksha localities is a matter of future work. The Annotation 3 in the Figure 2 illustrates one of these cases, although it must be highlighted that the most likely reason is that Paasonen did not visit this area. However, this emphasizes the complexity of connecting polygons with the intended localities.

V MOVING FORWARD

The approach that should be taken in the future, in our view, would be to re-examine the polygons digitized in the URHIA project and displayed in these digital maps in greater detail. We should analyze which polygon aims to point to which village or village cluster, and see to what detail this has been achieved in the current maps. If this leads to better matching of known settlements and the polygons – that would be a great development. The current infrastructure where the maps are hosted supports, in our understanding, very well such extending and collaborating approaches in this task, where further refinement is always possible.

The longer term goal would be to have a comprehensive database of localities where Erzya and Moksha have been spoken and are currently spoken. This would help to contextualize the existing fieldwork materials and also indicate gaps that could be filled, once new fieldwork becomes possible. Later, this work should be extended not only to Erzya and Moksha but for other Uralic languages, although the possible reanalysis of areas in URHIA maps, of course, needs separate consideration for each language.

VI DISCUSSION

In our view, one major challenge in dialect cartography stems from the situation where traditional printed maps were often produced with a relatively broad scale. If one centimeter is hundreds of kilometres, and few natural landmarks are present, the location of any given area is not necessarily precise. Indeed, small differences are easily caused by different projections, and in a small printed map the projection would usually not even have been indicated. The purpose of these maps was to provide a rough indication of where the languages or dialects were spoken, and what the general geographic relations of those areas were to one another as plotted against major landmarks. These maps were not intended for anyone to actually use for finding their way to these locations.

With digital maps, and especially digital datasets, however, we lose the comfort of coarse granularity. As digital maps can be zoomed indefinitely, the exact location of polygon boundaries becomes very important. The data sets can be used in combination with one another, and part of their core purpose is computational use. Thus, we need to aim for a situation where the polygons would be as solidly grounded on the actual settlements where the language is or has been spoken, or to real areas where the language speakers interact in cases where the language is not bound to permanent settlements.

References

Marianne Bakró-Nagy, Johanna Laakso, and Elena Skribnik. The Oxford guide to the Uralic languages. Oxford University Press, 2022.

R. Bartens. Mordvalaiskielten rakenne ja kehitys, volume 232 of Mémoires de la Société Finno-Ougrienne. Finno-Ugrian Society, Helsinki, 1999.

A. Feoktistov and S. Saarinen. Mokšamordvan murteet, volume 249 of Mémoires de la Société Finno-Ougrienne. Finno-Ugrian Society, Helsinki, 2005.

A. P. Feoktistow. Die dialekte der mordwinischen sprachen / Диалекты мордовских языков. In Kaino Heikkilä, editor, H. Paasonens mordwinisches Wörterbuch, volume XXIII of LSFU, pages XXXI–LXXXVI. Finno-Ugrian Society, Helsinki, 1990.

Kaino Heikkilä, Hans-Hermann Bartens, Aleksandr Feoktistow, Grigori Jermuschkin, and Martti Kahla, editors. H. Paasonens Mordwinisches Wörterbuch, volume Band I (A-J) of Lexica Societatis Fenno-Ugricae XXIII and 1 Kotimaisten kielten tutkimuskeskuksen julkaisuja 59. Suomalais-Ugrilainen Seura and Kotimaisten kielten tutkimuskeskus, Helsinki, 1990.

Kaino Heikkilä, Hans-Hermann Bartens, Aleksandr Feoktistow, Grigori Jermuschkin, and Martti Kahla, editors. H. Paasonens Mordwinisches Wörterbuch, volume Band II (K-M) of Lexica Societatis Fenno-Ugricae XXIII, 2 Kotimaisten kielten tutkimuskeskuksen julkaisuja 59. Suomalais-Ugrilainen Seura and Kotimaisten kielten tutkimuskeskus, Helsinki, 1992.

Kaino Heikkilä, Hans-Hermann Bartens, Aleksandr Feoktistow, Grigori Jermuschkin, and Martti Kahla, editors. H. Paasonens Mordwinisches Wörterbuch, volume Band III (N-Ŕ) of Lexica Societatis Fenno-Ugricae XXIII, 3 Kotimaisten kielten tutkimuskeskuksen julkaisuja 59. Suomalais-Ugrilainen Seura and Kotimaisten kielten tutkimuskeskus, Helsinki, 1994.

Kaino Heikkilä, Hans-Hermann Bartens, Aleksandr Feoktistow, Grigori Jermuschkin, and Martti Kahla, editors. H. Paasonens Mordwinisches Wörterbuch, volume Band IV (S-Ž) of Lexica Societatis Fenno-Ugricae XXIII, 4 Kotimaisten kielten tutkimuskeskuksen julkaisuja 59. Suomalais-Ugrilainen Seura and Kotimaisten kielten tutkimuskeskus, Helsinki, 1996.

Kaisla Kaheinen, Riku Erkkilä, and Toivo Qiu. The Oxford guide to the Uralic languages: A major albeit uneven handbook. Finnisch-Ugrische Forschungen, (69):227–234, 2024.
Timo Rantanen, Outi Vesakoski, Jussi Ylikoski, and Harri Tolvanen. Geographical database of the Uralic languages, May 2021. URL https://doi.org/10.5281/zenodo.4784188.
Timo Rantanen, Harri Tolvanen, Meeli Roose, Jussi Ylikoski, and Outi Vesakoski. Best practices for spatial language data harmonization, sharing and map creation—a case study of Uralic. PLOS One, 17(6):e0269648, 2022a. doi: https://doi.org/10.1371/journal.pone.0269648.
Timo Rantanen, Outi Vesakoski, and Jussi Ylikoski. Mapping the distribution of the Uralic languages. In The Oxford Guide to the Uralic Languages. Oxford University Press, 2022b.
Jack Rueter and Niko Partanen. Restructuring and visualising dialect dictionary data: Report on Erzya and Moksha materials. In Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities, pages 41–47, 2025.
Jack Rueter, Olga Erina, and Niko Partanen. rueter/mordvin-varieties: Comparative mordvin database, January 2020. URL https://doi.org/10.5281/zenodo.3627624.
Jack Rueter, Olga Erina, and Niko Partanen. rueter/mordvin-varieties: Comparative mordvin database, October 2025. URL https://doi.org/10.5281/zenodo.17464539.
Outi Vesakoski, Michael Dunn, Meeli Roose, and Jenni Santaharju. The Uralic Trove (UraLaari)–the digital data infrastructure of speaker areas of Uralic languages and Finnish dialects. Digital Humanities in the Nordic and Baltic Countries Publications, 7(3), 2025.
Г. И. Ермушкин. Ареальные исследования по восточным финно-угорским языкам (эрзя-мордовский язык). Наука, Москва, 1984.
М. З. Левина. Мокшень диалектологиясь. Диалектология мокшанского языка: учебное пособие. Издательство Мордовского университета, Саранск, 2014.

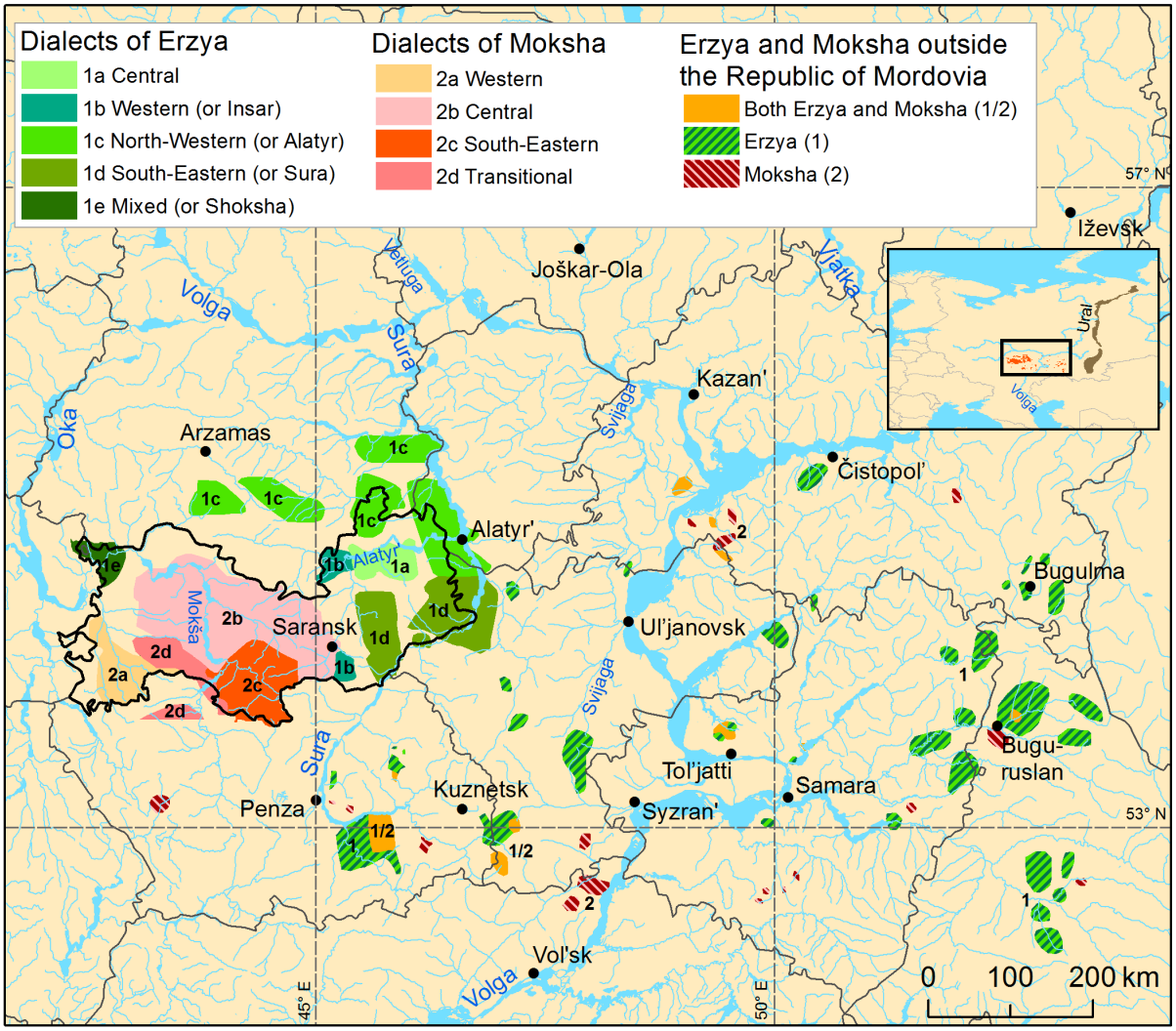
A ONLINE APPENDIX

We provide for the study an Online Appendix that is also published in Zenodo Rueter et al. [2025], and for which the source code is available in GitHub.

The Online Appendix is available at the following address:

https://multilingualfacilitation.com/Mordvin-Varieties/Dialect_Cartography_of_Erzya_and_Moksha_Languages

Figure 1: Map of the Erzya and Moksha dialects by Rantanen et al. [2021]



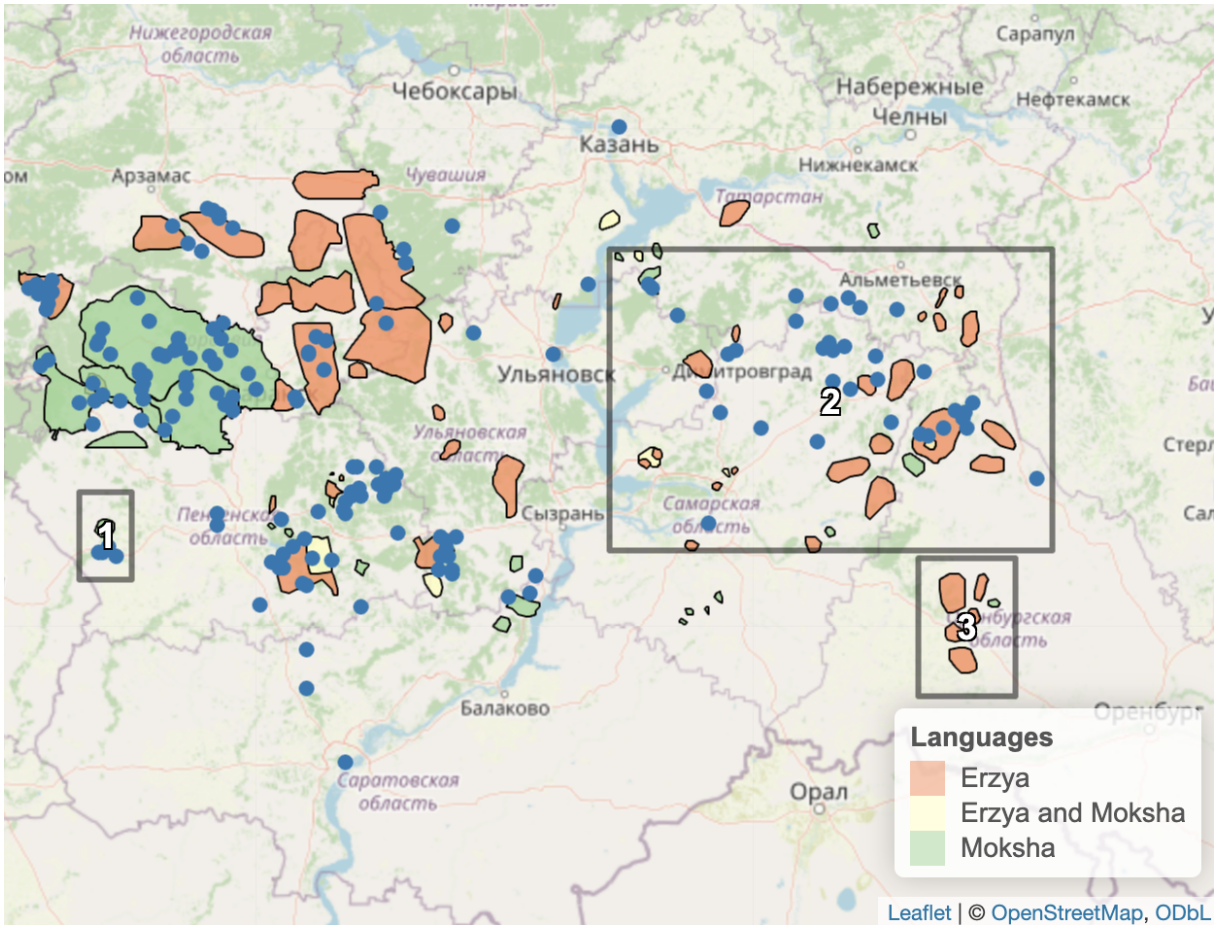


Figure 2: Collection points of Paasonen and URHIA polygons