The Uralic Trove – The digital data infrastructure of speaker areas of Uralic languages

Outi Vesakoski[a], Michael Dunn[b], Meeli Roose[c] and Jenni Santaharju[a]

[a] *University of Turku, School of languages and translation studies*
[b] *Uppsala University, Department of General Linguistics*
[c] *University of Turku, Department of Geography and Geology*

Abstract

This paper presents the Uralic Trove, a collection of datasets related to the human past in the Uralic language speaker area with special focus on the area of Finland. All the datasets are made in large collaborations, and (apart from no 8) have initially been launched elsewhere – this paper aims at collecting a 'trove' where all these areal datasets are easily located. We briefly describe the contents of eight multidisciplinary open access datasets: The databases considering the whole Uralic family or its speaker area are 1) UraLex (a basic vocabulary database with cognate coding), 2) UraTyp (binary data of existence of given linguistic typological functions), 3) a Cartographical Database of Uralic Languages (GIS format) and another GIS database of Interdisciplinary Maps of North-West Eurasia. The Finnish datasets are more detailed and include 5) a digital, easy to use version of the Dialect Atlas of Finnish collected already 100 years ago, 6) AADA, Archaeological Artefact Database of Finland, 7) Historical Travel Environment model and 8) Historical Culture and Environment Database of Finland. Uralic Trove also includes two user interfaces, that are intended for easy visualization and access of the data.

Keywords [1]

Archaeological artefact, Environmental data, Finno-Ugric linguistics, Finnish dialects, Geographical Information System, Interdisciplinary human past

1. Introduction

Integrative approaches to building holistic human histories are little by little covering the globe, and more and more multidisciplinary data collections are published in open access. The work by the Finnish BEDLAN team (Biological Evolution and Diversification of Languages, www.bedlan.net) and Human Diversity consortium (www.humandiversity.fi), both at the University of Turku Finland, have integrated the North-West Eurasian area into this emerging network and data pool. This paper brings together the data collections produced within and around the BEDLAN project. This data collection, Uralic Trove, presents a new research infrastructure pooling our datasets regarding human diversity in the Uralic language speaker area (Fig. 1). With the Uralic Trove we aim to advance not only the study of the Uralic language family and North-Western Eurasia but we also hope in general to push forward the integrative studies of the human past as well as to support the further development of methodology for digital study of human diversity.

Currently, the Uralic Trove includes four datasets related to Uralic language speaker areas and four related especially to the area of Finland, as well as two user interfaces aiming at easy access and

[1] *Digital Humanities in the Nordic and Baltic Countries 2024: From Experimentation to Experience, May 27-31, 2024, Reykjavík, Iceland*
EMAIL: outi.vesakoski@utu.fi.fi.com (A. 1);michael.dunn@lingfil.uu.se (A. 2);meeli.roose@utu.fi (A. 3), jenni.santaharju@utu.fi (A. 4)
ORCID: 0000-0002-7220-3347 (A. 1); 0000-0001-5349-5252 (A. 2); 0000-0003-2776-9883 (A. 3); 0000-0002-7925-2715 (A. 4)

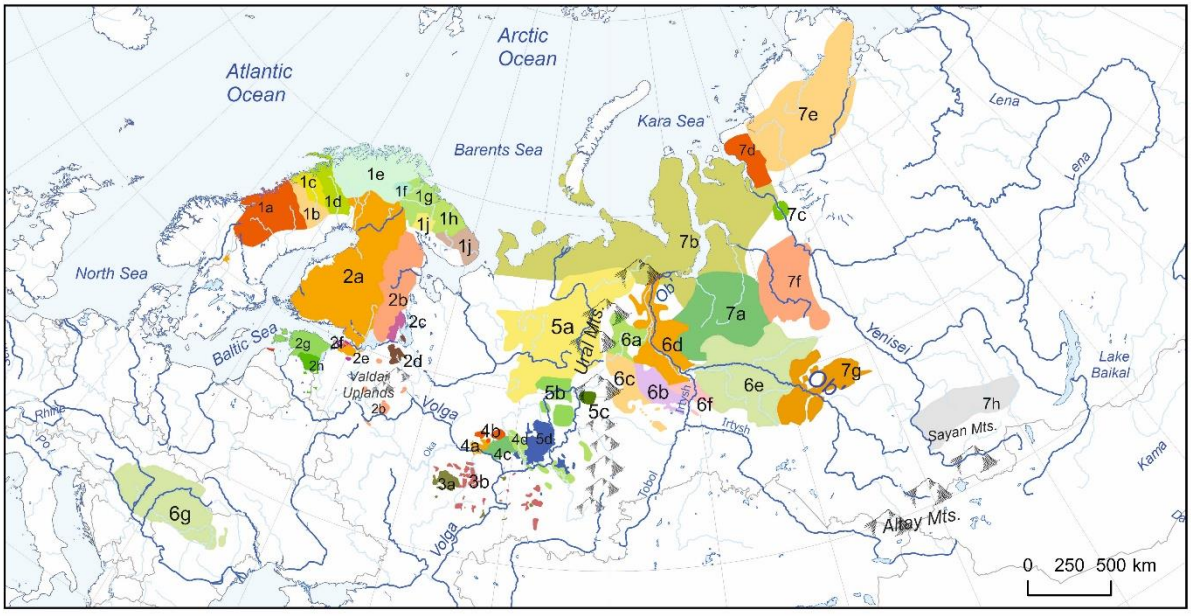
visualization of the datasets. All the data collections are or will be available in repositories and user interfaces. We have complied FAIR principles by making data…

…Findable: Most databases have been launched as part of an open access, peer reviewed research paper. The current paper is one step further: Here we promote the visibility of the areal data collections by reviewing them all in the same package. We also aim to expand the package with the databases currently in collection.

…Accessible: We aim to manage most of the databases in GitHub, which is a versatile repository where the team can update and curate each data according to needs. Some data sets can also be downloaded directly from github.com/BEDLAN, but all the data are or will be available in Zenodo or OSF or similar repositories. In Zenodo, we use the "BEDLAN" token to indicate the "community" the database in question is part of.

…Interoperable: The Uralic databases are interoperable with each other as well as with databases from other language families through standards identifications of the languages: ISO-639-3 identifiers and Glottocodes for languages, which - as well as updated coordinates - we have curated through the Glottolog initiative (Hammerström et al. 2024). The language data follows the emerging international standards such as usage of the CLDF format. The Human Diversity consortium is building a standard for multidisciplinary data describing human and environment variation in Finland. For now, the databases have been cross-analysed through municipality codes or coordinates (e.g. Lynch et al. 2022) or through building a grid over Finland and comparing the various attributes of each cell (as in Rantanen et al. 2021).

…**Reusable**: The data are stored in different formats (csv, excel, shapefiles).



Saami
South (1a) Ume (1b) Pite (1c) Lule (1d)
North (1e) Aanaar (1f) Skolt(1g) Kildin (1h)
Akkala (1i) Ter (1j)

Finnic
Finnish (2a) Karelian (2b) Ludic (2c) Veps (2d)
Ingrian (2e) Votic (2f) North Estonian (2g)
South Estonian (2h) Livonian (2i)

Mordvin
Moksha (3a) Erzya (3b)

Mari
Hill Mari (4a) North-Western Mari (4b)
Meadow Mari (4c) Eastern Mari (4d)

Permic
Komi (5a) Komi-Permyak (5b)
Komi Yazva (5c) Udmurt (5d)

Ugric
North Mansi (6a) East Mansi (6b)
West and South Mansi (6c)
North Khanty (6d) East Khanty (6e)
South Khanty (6f) Hungarian (6g)

Samoyedic
Forest Nenets (7a) Tundra Nenets (7b)
Forest Enets (7c) Tundra Enets (7d)
Nganasan (7e) Northern Selkup (7f)
Tomsk region Selkup (7g) Kamas & Mator (7h)

Figure 1: Uralic languages from the Geographical Database of Uralic Language Speaker Areas.

2. Data collections and web apps

The data collection processes started according to research needs. In the beginning of the project in 2009 the BEDLAN team created a basic vocabulary list for 17 Uralic languages with cognate (or correlate) coding in order to conduct phylolinguistic studies (Honkola et al. 2013, Syrjänen et al. 2013, Lehtinen et al. 2014). Today the data includes over 30 languages (de Heer et al. 2023). We also initiated a collection of Uralic linguistic typological data for the language family in 2013, but the initiative had limited success until we developed a collaboration with University of Tartu, Uppsala and with the Grambank initiative from the Max Planck Institute for Evolutionary Anthropology (Vesakoski 2023). For a spatial study of Uralic languages we needed a GIS-based data of language speaker areas. Such data did not exist, but through collaboration with prof. Jussi Ylikoski and the authorship of Oxford Guide for the Uralic Languages we were able to create the first ever polygon data of the language speaker areas. Finally, during the years of interdisciplinary work around the human past in the Uralic language speaker area we have created a large number of maps representing the linguistic, cultural, genetic and environmental landscapes of North-Western Eurasia. To make these maps available for other users we created another Zenodo release with PDFs of the maps, and will next add the thematic layers as shape files.

The Finnish language is member of the Uralic language family, and also a special focus for the studies within BEDLAN and the Human Diversity consortium. The BEDLAN project started from the digitized Dialect Atlas of Finnish (Kettunen 1940), which we publish as a corrected version in this volume (Santaharju et al. this volume). The Dialect Atlas is compiled using parishes of the time as study units. There is a variety of cultural and environmental information also collected per parish which BEDLAN has also digitalized. Both the Dialect Atlas and Cultural-Environment data are collected as attributes of parishes, and thus the dataset can be aligned with their parish (municipality) codes as in Honkola et al. (2018) or in Lynch et al. (2022). The cultural data only covers the centuries 1600-1800, whereas the third dataset,the Archaeological Artefact Database of Finland, AADA, provides spatial information and systematic typological classification of Finnish archaeological artefacts from the Stone Age to Early Medieval. AADA data compilation was an enormous task initiated already in 2013. Finally, as we wanted to understand how humans actually used the landscape and carried languages, cultures and genes over Finland, we created a GIS model of the (pre)historical human travelling environment. The travel environment model is versatile, allowing many different kinds of study of how humans potentially used the environment for travelling.

Besides offering the dataset in Zenodo, we have built an interactive user interface **URHIA - Uralic Historical Atlas** (https://urhia.fi) - to provide easy access to the static maps of Uralic language speaker areas (see below) and also to give lay audiences the possibility of creating their own maps. URHIA is presented in more detail in Roose et al. (this volume), but as an fundamental part of Uralic Trove, it will be presented also here. URHIA is built on the open-source spatial infrastructure GeoNode by GeoSolutions and integrated into UTU's spatial infrastructure (https://geospatial.utu.fi/resources/utu-geospatial-data-service/). Unlike conventional data repositories, URHIA is designed as an interactive spatial platform for researchers and lay audiences. The platform currently hosts two major datasets: the Cartographical Database of Uralic Language Speaker Areas (see below), which is visualized through the *Uralic Language Atlas* and the Archaeological Artefact Database of Finland (AADA) (see below), which is presented as the *Archaeological Artefact Atlas of Finland*. URHIA is an ongoing project aiming to transform spatial data into dynamic data showrooms, presenting thematic spatial datasets through interactive online maps. The inclusion of AADA represents a pioneering effort, being the first database of its kind in Finland—and possibly globally. This initiative marks a significant milestone in the digitization and accessibility of archaeological data, setting the stage for similar projects worldwide.

Another user interface included in Uralic Trove is the **Uralic Areal Typology Online** (uralic.clld.org) built to present the Uralic linguistic typological data (see below). We got a great opportunity to adopt the platform built by MPI-EVA (Forkel and Banks 2014) to promote the usage of

the Uralic typological data. In the next release, the web app will be expanded with UraTyp 2.0 and UraLex 3.0 (see below).

2.1. Uralic data collections

The Uralic language family (also known as Finno-Ugric family) consists of 30 to 40 languages. Uralic languages are spoken in Northwestern Eurasia adjacent to or amidst Indo-European languages, e.g., Scandinavian and Slavic languages, as well as near Turkic, Tungusic, and Yeniseian languages (Ket).

2.1.1. UraLex

UraLex is a basic vocabulary database with cognate assessments indicating etymological connections between words for a meaning in different languages. Basic vocabulary represents meanings that exist in most languages, and thus it does not include culture-specific meanings. Basic vocabulary words are expected to be relatively stable over time, and unlikely to be replaced through borrowing or semantic shift. These meanings include, for example, words like “mother”, “sun” and general concepts such as lower numerals, pronouns and body parts. In all, the basic vocabulary part of the lexicon is analogous to the most stable part of the genome; the part that has preserved best the trace of ancient relationships, whether of languages or populations. The meanings in the basic vocabulary lists are the same for datasets for different language families, and thus the basic vocabulary of the Uralic family can be cross-analysed with other families, especially if all the words are also provided in International Phonetic Alphabet transcription (IPA coding). IPA coding of the Uralic data will be published in UraLex 3.0. For the cross-comparison purposes, the data is published under the Lexibank umbrella (List et al. 2022), and is available from the lexibank repository https://github.com/lexibank/uralex .

The word lists were compiled from dictionaries and for the case of less studied languages, also from linguistic experts. The cognate information was compiled from etymological dictionaries and research papers. UraLex has had three major releases: The 1.0 release (Syrjänen et al. 2018) and corrected and updated 2.0 release (de Heer et al. 2021,) that cover 26 languages as well as reconstructed Proto-Uralic. The 3.0 release (data in de Heer et al. 2023, publication paper Vesakoski et al. ms) introduces more languages and coding for borrowing relationships as well as cognates. It will also be made available and visible through Uralic Areal Typology Online (see below).

2.1.2. UraTyp

UraTyp is a linguistic typological dataset (Norvik et al. 2022) consisting of 360 linguistic traits in the form of questions with binary answers (Fig. 2). The data was compiled from descriptive grammars and grammar sketches when available, or by interviewing language experts. The features actually represent two underlying typological questionnaires: Grambank questionnaire (Skirgård et al. 2023) and the Uralic-specific questionnaire.

Uralic specific traits (UT traits) are 165 traits that were developed to describe variation within Uralic languages (Norvik et al. 2022), since the level of granularity of the global Grambank traits is often insufficient to distinguish Uralic languages from each other. The UT questions were developed by researchers at the University of Tartu, collected with funding received from the University of Turku, and published with the expertise of University of Uppsala. At the moment we are expanding the UT traits to include also Uralic neighbors, and are also collecting examples of the usage of each linguistic trait in the language. The current version is in Zenodo 10.5281/zenodo.5236365 but is easily accessible in Uralic Areal Typology Online, uralic.clld.org.

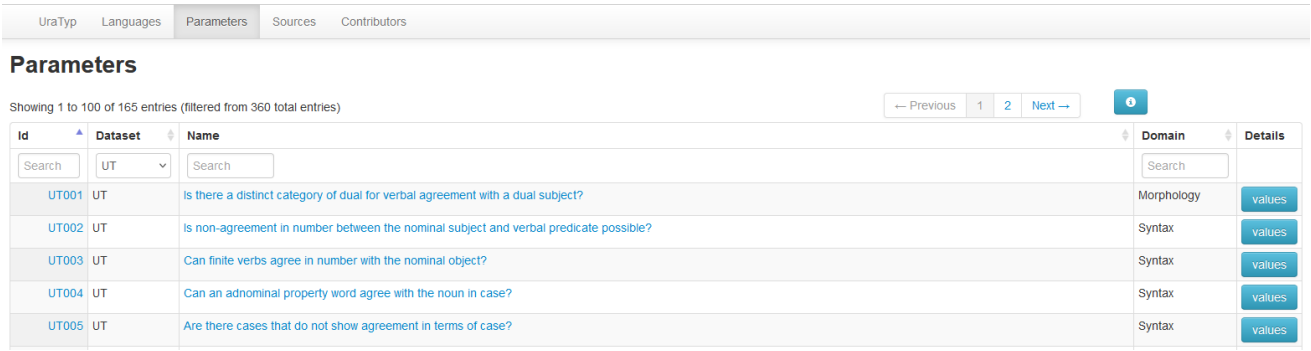


Figure 2: Example of UraTyp parameters from the window of the user interface Uralic Areal Typology Online.

Grambank data is a list of 195 grammatical features collected for cirka 2450 world languages with the aim of studying global typological diversity (Skirgård et al. 2023). BEDLAN contributed the Uralic languages to Grambank. The Grambank traits in UraTyp are found also in the Grambank database and, importantly, through the Grambank traits, Uralic language are fully interoperable with 2450 other word languages. Grambank data is easily accessible in grambank.clld.org

2.1.3. Geographic database of the Uralic languages

The Geographical Database of the Uralic languages contains spatial data with polygons of Uralic languages speaker areas (Rantanen et al 2022a; Fig 1). This dataset captures the geographic extent of Uralic language speaker areas by documenting both historical distributions, about 100 years old, and also contemporary distributions. Such a dataset did not exist earlier and we developed a standard for such work while creating the map collection. The process was (in brief) to find source maps from linguistic literature, which were first digitized, georeferred and then carefully reviewed by language specialists to ensure their accuracy (Rantanen et al., 2022b). The polygons can be bound to the Uralic linguistic dataset through standard language codes. The georeferred polygons allow the use of the speaker areas with other spatial data. For example, the GIS data is ready for mapmaking and has been used in many publications (e.g. Tambets et al. 2018 with genetic data) and in user interfaces (e.g. our own Uralic Areal Typology Online).

The dataset itself, along with static maps, is available in Rantanen et al. (2021), and the process is described in Rantanen et al. (2022a, 2022b). The Geographical database has been visualized in the Oxford Guide for Uralic Languages (Rantanen et al., 2022b), and the original electronic files of these maps in PDF format are accessible through URHIA and the whole data in Zenodo 10.5281/zenodo.4784188). Additionally, the database has been integrated into the URHIA interactive cartographical interface (https://urhia.fi; Roose et al. this volume).

2.1.4. Interdisciplinary Spatial Database of Human History in the North-West Eurasia

Interdisciplinary Uralic Spatial Database of Human History in North-West Eurasia is a collection of map visualizations and shapefiles created for various publications and grant applications. We now make these static maps and GIS files available for broader use in Roose et al. (2023). This collection, focusing on spatial information in linguistics, archaeology, geography, ecology and genetics within the Uralic language-speaking area, contains a vast array of multidisciplinary polygon data. The data is obtained

from various publications and from multidisciplinary collaboration. The metadata in Roose et al. (2023) will be expanded to include all the sources.

At the moment the Zenodo repository hosts only the static maps, but will be extended with the shapefiles and thorough metadata (10.5281/Zenodo.10081902). The shapefiles will allow for integration of each thematic layer into other georeferred datasets, which we hope will encourage further interdisciplinary studies of the human past and human-environment interactions. As data collections and mapmaking efforts continue to evolve alongside ongoing projects, the current repository will be updated dynamically.

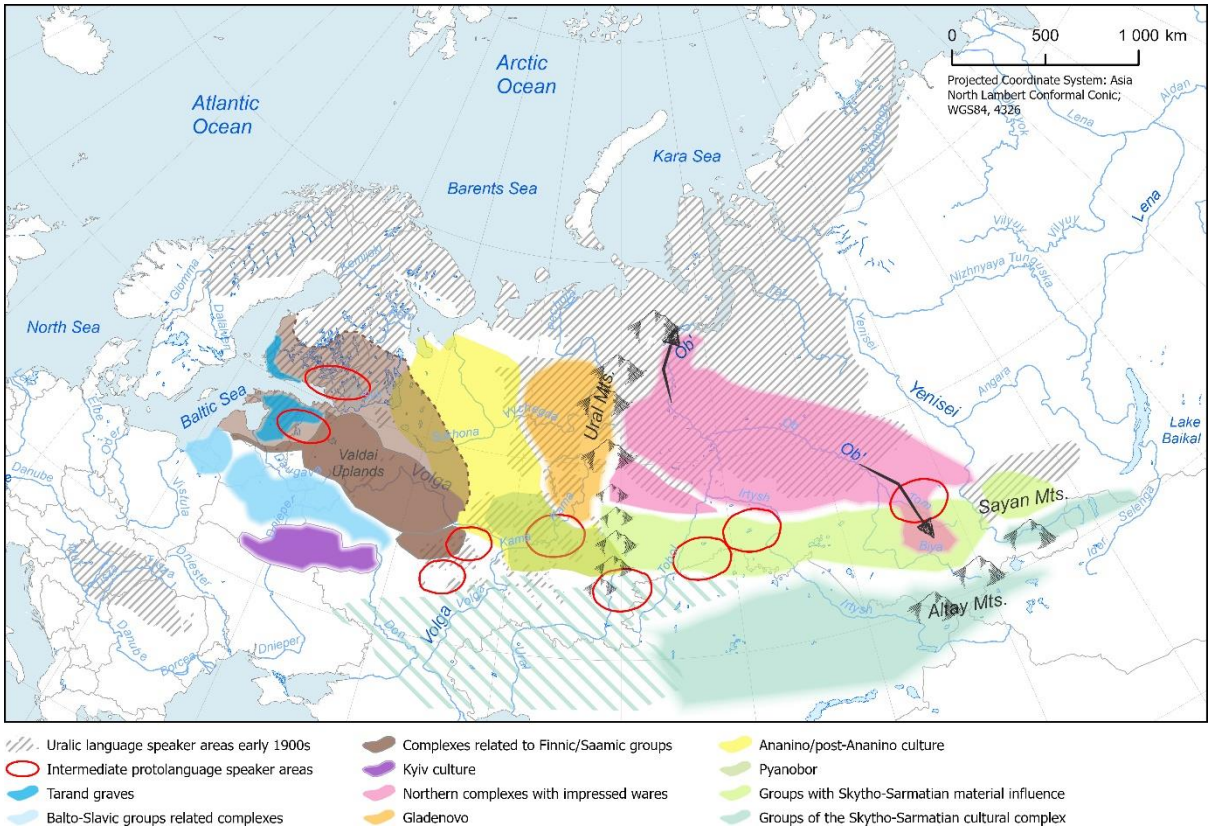


Figure 3: Example of the interdisciplinary map illustrating Uralic language speaker areas in 1900 alongside with some Early Bronze Age cultural complexes and suggested speaker areas of intermediate Uralic languages (such as Proto-Samojedic in the right). This is the original version of a map by Meeli Roose, and it is redrawn by the publisher in Vesakoski et al. (2025).

2.2. Finnish data collections

The Uralic Trove also includes data covering only the Finnish language area. A new profile area of University of Turku, the Human Diversity consortium (http:/www.humandiversity.fi) focuses especially on Finland for there are various unique multidisciplinary data sets available, e.g. on pre-industrial disease outbreaks (Nitch et al., accepted ms). The data sets can be cross-referenced e.g. through information based on parish codes, parish polygons and coordinates, or. by using a grid over Finland as in 2.2.4. At the moment the following datasets are published or close to being Open Access.

2.2.3. Preindustrial dialect landscape of Finland

Dialect Atlas of Finnish (Kettunen 1940) is a map collection of linguistic variation. It was compiled in the 1920s and 1930s to present the dialect landscape of Finnish at the turn of the century. Kettunen travelled through Finland interviewing people in Finnish-speaking municipalities in order to collect information of about 400 linguistic traits describing morpho-phonological and lexical variation of Finnish. Later, in 1940s, he published the Dialect Atlas with spatial variation of 213 of these traits in 525 municipalities (more about the data collection in Vesakoski et al. 2024). The maps are available at http://kettunen.free.nf. Embleton & Wheeler (1997, 2000) initiated the digitization process together with the Institute for Finnish Languages (KOTUS), and the outcome was further refined by the BEDLAN team. The corrected version was published without metadata by KOTUS (at http://urn.fi/urn:nbn:fi:csc-kata20151130145346403821). To meet the FAIR principles, BEDLAN now provides peer-reviewed documentation of the Atlas in Santaharju et al. (this volume) and the actual data with detailed data description in Santaharju et al. (2024a). The Atlas and will be added to URHIA (Uralic Historical Atlas, www.urhia.fi; see Roose et al. this volume)

Morphological trait described	Level1	Level2	Level3
Map 2: Gemination of consonants	Consonantism		
Map 3: Consonant palatalization	Consonantism		
Map 4: juopi/syöpi, juop/syöp, juo/syö etc.	Consonantism	History of plosives at the beginning of syllables	
Map 5: vasarata, vasaraa etc.	Consonantism	History of plosives at the beginning of syllables	
Map 7: vastakkaa/vastakka, vastaaka, vastatkaa etc.	Consonantism	History of plosives at the end of syllables	In front of voiceless consonants
Map 8: metsä, messä, mehtä, mehhä etc.	Consonantism	History of plosives at the end of syllables	In front of voiceless consonants

Figure 4: Examples of linguistic traits included in the Dialect Atlas of Finnish.

2.2.2 Historical environmental and cultural variation over Finland

The Uralic Trove also includes the Historical Culture and Environment Database of Finland that consists of environmental and cultural variation. It describes - parish by parish - how material and immaterial folk culture varies spatially: e.g. what kind of boat type or wedding customs were typical to the parish. The folk cultural data from 1600-1900 is further accompanied with information social and demographic data of for example wealth (taxation), child mortality and population numbers in 1880s. We were also able to reach data on environmental variation such as mean temperature and rainfall of a municipality in 1880s as well as geographical factors such as soil type and amount of watersheds as a percentage of the area of each parish. The cultural and environmental data from 1880s were used in Honkola et al. (2018) and Lynch et al. (2022). Part of the data is cultural-environmental, e.g. amount of farmed field or forests in a parish.

The original data comes from historical statistical yearbooks, historical atlases and geographical databases though some unchanging physical information such as soil types are based on modern geographical databases. The data was digitized by the BEDLAN team. We provide data per parish, and with parish coded, coordinates and polygon data. With these identifiers the data is interoperable with other spatial data over Finland. The data and metadata including approximate timing and typological classification is provided in Santaharju et al. (2024b) and thorough data description will be in Jordman et al. (in preparation). We aim also to add the data to URHIA (see above and Roose et al., this volume).

2.2.3 Archaeological Artefact Database of Finland

The Archaeological Artefact Database of Finland (AADA) is a spatial dataset offering typological categorisation, spatio-temporal context, location and photos of 48 000 artifacts stemming from the Early Mesolithic to the end of the Iron Age. The data was compiled over seven years (2013–2020) through meticulous fieldwork in Finnish museums, where each artifact was examined and documented individually. The data in AADA consists of detailed descriptions of artefacts (Pesonen et al. 2023) and related photographs (Moilanen et al. 2023). Each artifact is assigned with a collection number, which allows it to be linked to additional information about the archaeological site where it was discovered. The dataset also includes spatial coordinates, enabling integration with other spatial data for further analysis. The data is described in Pesonen at al. (2024) and is maintained in GitHub and published in Zenodo (Typological data in Pesonen et al. 2023 and the photos in Moilanen et al. 2023). To support advancements in digital archeology, the R code used to create the maps for the publication was also made available (Roose 2024). Additionally, AADA data will be integrated into the URHIA user interface (Roose et al., this volume).

2.2.4 Historical travel environment model over Finland

Historical travel environment model - a spatial model for historical travel effort - is a spatial dataset with terrain and landscape attributes (Fig. 5A) coupled with information of travel speed (Fig. 5B) in different environments. The outcome is e.g. prediction of which areas are directions were easy to travel (Fig. 5C). The ecological and geographical data is compiled from digital databases and digitized from literature. The travel speed estimates are made from historical sources that characterize the landscape in terms of travel effort given the environmental and human-related factors current up until the late 19th century. The data is described in Rantanen et al. (2021), and is published as data in Rantanen et al. (2025).

The data is organised into a 1 x 1 km grid over Finland. Each grid square includes information of the eco-geographical factors. The grid approach allows for versatile cross-analyses of data as any other spatial data could be added to the grid - including point data e.g. indicating archeological artifact types and location.

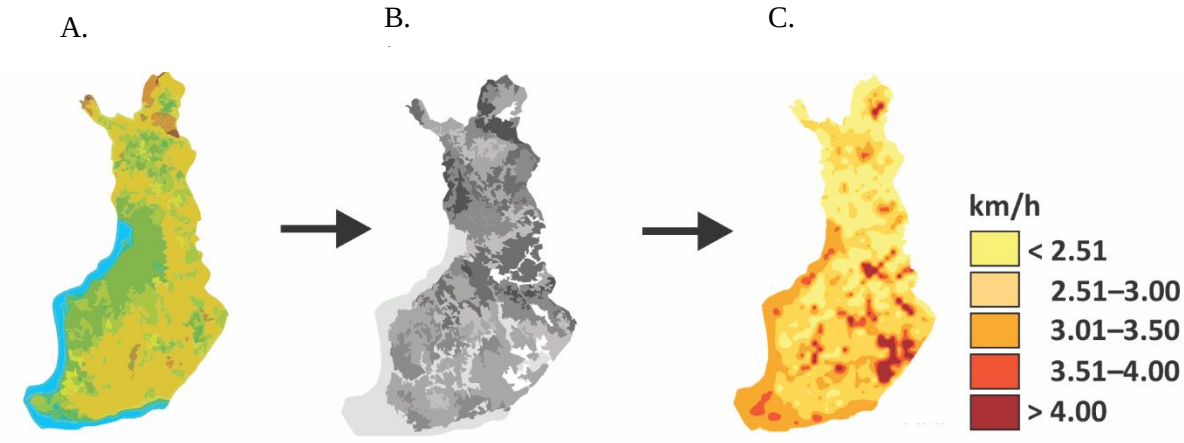


Figure 5: Example of the information within the travel environment model. The model is organized as 1 x 1 km grid. Each cell in the grid includes attributes indicating A) a value of seven thematic layers (an example shows the topographical variation) and B) travel cost, assuming walking, horse-back riding or rowing with a boat. Figure C) combines information of the thematic layer and travel cost layer to produce a travel environment landscape, showing the fastest travel speed for each cell. The outcome is a general overview of the variation in travel speeds in Finland – the darker the area, the faster the people were theoretically able to travel. Maps by Timo Rantanen, modified by Jenni Santaharju.

3 Concluding remarks

Building the Uralic Trove over the years since 2009 has been a long and complex process with new challenges with each data. The disciplines involved include:

1. Linguistics: Historical linguistics, typology, dialectology, etymology, contact linguistics; data compiled from dictionaries, grammars, through interviews etc.

2. Geography: GIS methods (Geographical Information Systems); data types are point, vector, polygon and grid; data collated from literature, other databases, Google Earth etc.

3. Archaeology: From the Mesolithic Stone Age to the Medieval era; point data and polygons; data compiled through inventory of museum items, literature searches, expert interviews

4. Cultural history: Data organised by municipalities; data acquired through digitizing atlases and literature sources.

5. Environmental sciences: Data organised by municipalities or in grid, point data or polygons; data compiled through various databases, digitizing historical atlases etc.

Each data set has its own qualities and demands, with its own standards we have had to adapt to or even develop ourselves. It also took time to learn the technical demands of Open Access publishing and interactive data handling in GitHub. During these years we have taught ourselves and also some of the user community to operate with GitHub and use data in GIS and R environments.

Today our position on Open Access data is totally contrary to 2009 when the first author started the BEDLAN project with the intention of keeping the data private, for use only by BEDLAN members. We now recognise that it was only through publication of open access datasets that our work was recognized internationally. This has stimulated interest in the study of human history in NW Eurasia, which has in turn made it easier to build collaborations and attract funding – thus contributing to the fulfilment of our scientific goal of building a picture of the holistic history of human past in NW Eurasia.

4 Acknowledgements

Much of the funding for data collection has been provided by the Kone Foundation (grant “Kielen murteet biologisen lajiutumisen näkökulmasta”, Urho Määttä as PI, “SumuraSyyni” and “AikaSyyni” Outi Vesakoski as PI, and “Kippo” with Unni Leino as PI). Outi Vesakoski was funded also by the Uralic Triangulation project (URKO) (2020–2022), supported by the Digital Humanities programme of the Research Council of Finland (grant no. 329259) and the Turku Institute for Advanced Studies (University of Turku). Meeli Roose was funded by University of Turku Graduate School BGG and Finnish Cultural Foundation (grant no. 0022088). Jenni Santaharju was funded with the Human Diversity consortium (HuDi) (Profi7 programme by Research Council of Finland, grant no. 352727). We are grateful to all co-authors and co-workers for participating in the development of this complex interdisciplinary data collection.

1. References

Embleton, Sheila and Eric Wheeler. 1997. “Finnish dialect atlas for quantitative studies.” Journal of *Quantitative Linguistics* 4.99-102.

Embleton, Sheila and Eric Wheeler. 2000. “Computerized dialect atlas of Finnish: Dealing with ambiguity.” *Journal of Quantitative Linguistics* 7.227-31.

Forkel, Robert and Sebastian Bank. 2014. “The CLLD toolkit. Language Comparison with Linguistic Databases: RefLex and Typological Databases, Nijmegen.” Zenodo. DOI: 10.5281/zenodo.10846846

Hammarström, Harald; Robert Forkel; Martin Haspelmath and Sebastian Bank. 2024. “Glottolog 5.1.” Leipzig: Max Planck Institute for Evolutionary Anthropology. DOI: 10.5281/zenodo.14006617

de Heer, Mervi; Mikko Heikkilä; Kaj Syrjänen; Jyri Lehtinen; Outi Vesakoski; Toni, Suutari; Michael Dunn; Urho Määttä and Unni-Päivä Leino. 2021. “UraLex 2.0 - Uralic basic vocabulary with cognate and loanword information. DOI: 10.5281/zenodo.4777568

de Heer, Mervi; Rogier Blokland; Michael Dunn and Outi Vesakoski. 2023. “Loanwords in basic vocabulary as an indicator of borrowing profiles.” *Journal of Language Contact*. 16: 54-103.

Honkola Terhi; Vesakoski, Outi; Korhonen, Kalle; Lehtinen, Jyri; Syrjänen, Kaj and., Wahlberg, Niclas 2013. “Cultural and climatic changes shape the evolutionary history of the Uralic languages.” *Journal of Evolutionary Biology* 26: 1244-1253.

Honkola, Terhi; Kalle Ruokolainen; Kaj Syrjänen; Unni-Päivä Leino; Ilpo Tammi; Niklas Wahlberg and Outi Vesakoski. 2018. “Evolution within a language: Environmental differences contribute to divergence of dialect groups.” *BMC Evolutionary Biology.* 18:132.

Kettunen, Lauri. 1940. “Suomen Murteet III A. Murrekartasto.” *Suomalaisen Kirjallisuuden Seura*. Helsinki.

Lehtinen, Jyri; Terhi Honkola; Kalle Korhonen; Kaj Syrjänen; Niklas Wahlberg and Outi Vesakoski. 2014. “Behind family trees: Secondary connections in Uralic language networks.” *Language Dynamics and Change*. 4: 189-221. DOI: 10.1163/22105832-0040200

List, Johann-Mattis; Forkel, Robert; Greenhill, Simon J.; Rzymski, Christoph; Englisch, Johannes and Gray, Russell D. 2022. “Lexibank, a public repository of standardized wordlists with computed phonological and lexical features”. Scientific Data 9, 316 (2022). DOI: 10.1038/s41597-022-01432-0

Lynch, Robert; John Loehr; Virpi Lummaa; Terhi Honkola; Jenni Pettay and Outi Vesakoski. 2022. “Socio-cultural similarity with host population rather than ecological similarity predicts success and failure of human migrations.” *Proceedings of the Royal Society B,* 289: 20212298.

Moilanen, Ulla; Petro Pesonen; Jarkko Saipio and Jasse Tiilikkala. 2023. “Archaeological artefact database of Finland (AADA)”. DOI: 10.5281/zenodo.10417383

Nitch, Aïda; Michael Briga; Tarmo Ketola; Terhi Honkola; Outi Vesakoski and Virpi Lummaa. “The spatial distribution of pertussis, but not measles or smallpox, in pre-health care Finland matches dialect groups." In revision.

Norvik, Miina; Yingqi Jing; Michael Dunn; Robert Forkel; Terhi Honkola; Gerson Klumpp; Richard Kowalik; Helle Metslang; Karl Pajusalu; Minerva Piha; Eva Saar; Sirkka Saarinen and Outi Vesakoski. 2021. UraTyp - Uralic typological data set DOI: 10.5281/zenodo.6392555

Norvik, Miina; Yinqi Jing; Michael Dunn; Robert Forkel; Terhi Honkola; Gerhard Klumpp; Richard Kowalik; Helle Metslang; Karl Pajusalu; Minerva Piha; Eva Saar; Sirkka Saarinen and Outi Vesakoski. 2022. “Uralic typology in the light of a new comprehensive data set.” *Journal of Uralic Linguistics* 1: 4-41.

Pesonen, Petro, Ulla Moilanen, Jarkko Saipio, Meeli Roose, Outi Vesakoski, Päivi (2023). Archaeological artefact database of Finland (AADA). DOI: 10.5281/zenodo.10437703

Pesonen, Petro, Ulla Moilanen, Meeli Roose, Jarkko Saipio, Jasse Tiilikkala, Usman Sanwal, Visa Immonen, Outi Vesakoski, and Päivi Onkamo. 2024. "Archaeological Artefact Database of Finland (AADA)." *Scientific Data* 11(1): 815.

Rantanen, Timo; Harri Tolvanen; Terhi Honkola and Outi Vesakoski. 2021. “A comprehensive spatial model for historical travel effort – a case study in Finland.” *Fennia* 199 (1) 61-88.

Rantanen, Timo, Outi Vesakoski, Jussi Ylikoski, Harri Tolvanen 2021. Geographical Database of the Uralic languages 2021. DOI: 10.5281/zenodo.4784188

Rantanen Timo; Harri Tolvanen; Meeli Roose; Jussi Ylikoski and Outi Vesakoski. 2022a. "Best practices for spatial language data harmonization, sharing and map creation—A case study of Uralic." *PLoS ONE* 17(6): e0269648.

Rantanen, Timo; Outi Vesakoski and Jussi Ylikoski. 2022b. "Mapping the distribution of the Uralic languages." in Marianne Bakró-Nagy, Johanna Laakso and Elena Skribnik (eds.), *The Oxford Guide to the Uralic Languages*. Oxford University Press.

Rantanen, Timo, Terhi Honkola, Harri Tolvanen and Outi Vesakoski. 2025. "Spatial model for historical travel effort (v1.0)" DOI: 10.5281/zenodo.10554593

Roose, Meeli; Tua Nylén; Petro Pesonen; Harri Tolvanen and Outi Vesakoski. 2024. "Uralic Historical Atlas (URHIA): Interactive Web App for Spatial Data." *Digital Humanities in the Nordic and Baltic Countries Publications*.

Roose, Meeli; Timo Rantanen; Henny Piezonka; Elina Salmela; Kerkko Nordqvist; Petro Pesonen, Ulla Moilanen, Terhi Honkola, Dmitri Kuznesov and Outi Vesakoski. 2023. "Collection of spatial information and maps of human past and environment in the Uralic languages speaker area". DOI: 10.5281/zenodo.10081902

Roose, Meeli. 2024. "AADA Archaeological Data Query and Map Visualisations in R." Zenodo. DOI: 10.5281/zenodo.11257952

Santaharju, Jenni; Kaj Syrjänen; Terhi Honkola; Jyri Lehtinen; Perttu Seppä; Outi Vesakoski and Unni Leino 2024a. "Digitized Dialect Atlas of Finnish by Lauri Kettunen. DOI: 10.5281/zenodo.10078078

Santaharju, Jenni, Paavo Jordman, Terhi Honkola, Timo Rantanen, Ilpo Tammi and Outi Vesakoski. 2024b. "Digitized historical cultural and environmental data of Finland." DOI: 10.5281/zenodo.10078077

Santaharju, Jenni; Kaj Syrjänen; Terhi Honkola; Seppä Perttu; Outi Vesakoski and Unni-Päivä Leino. This volume. "The digitized Dialect Atlas of Finnish by Lauri Kettunen." *Digital Humanities in the Nordic and Baltic Countries Publication*

Skirgård, Hedvig; Haynie, Hannah; Blasi, Damien (+ 102 authors). 2023. "Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss." *Science Advances*. 9, eadg6175.

Syrjänen, Kaj;, Honkola, Terhi; Korhonen, Kalle; Lehtinen, Jyri, Vesakoski, Outi and Wahlberg, Niklas. 2013. Shedding more light on language classification using basic vocabularies and phylogenetic methods: A case study of Uralic. *Diachronica* 30(3): 323-352.

Syrjänen, Kaj; Jyri Lehtinen; Outi Vesakoski; Mervi de Heer; Toni Suutari; Michael Dunn; Urho Määttä and Unni-Päivä Leino. 2018. "Lexibank/Uralex: UraLex Basic Vocabulary Dataset." Zenodo. DOI: 10.5281/zenodo.1459402.

Tambets, Kristiina; Bayazit Yunusbayev; Georgi Hudjashovet (+ 34 authors) 2018. "Genes reveal traces of common recent demographic history for most of the Uralic-speaking populations". *Genome Biology*. 19:139.

Vesakoski, Outi. 2023. "Karl Pajusalu ja uralilaisen typologisen aineiston synty." In *Pühendusteos Karl Pajusalule 60. sünnipäevaks.* Ed. by Eva Saar, Miina Norvik, Eva Velsker. Tartu: Tartu Ülikooli Kirjastus, 276–282.

Vesakoski, Outi, Jenni Santaharju and Lotta Aarikka. 2024. "Tieteen matkamiehen siivellä". In *Saarte keeled. Ellen Niidi juubeliraamat*, ed. by Mari Kendla. Emakeele Selts.

Vesakoski, Outi; Elina Salmela and Henny Piezonka. 2024. "Uralic archaeolinguistics". In *Oxford Handbook of Archaeology and Languages*, ed. by Martine Robbeets and Mark Hudson. In press.

Vesakoski, Outi; Tresholdi, Tiago; Soosaar, Sven-Erik; de Heer, Mervi; Maurits, Luke; Syrjänen, Kaj; Honkola, Terhi and Dunn, Michael. “Unravelling the disintegration patterns and chronology of the Uralic language family.” In preparation.