The Uralic Trove (UraLaari) – The digital data infrastructure of speaker areas of Uralic languages and Finnish dialects

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 data sets related to the human past in the Uralic language speaker area and especially in Finland. We briefly describe the contents of four data sets related to the whole Uralic or Finno-Ugric language speaker area, along with how to find them. The Uralic databases are: UraLex (a lexical database of Uralic languages), UraTyp (a database of Uralic language typological data), a Cartographical database of Uralic languages speaker areas and another of interdisciplinary maps of North-West Eurasio. The Finnish databases are: Dialect Atlas of Finnish, AADA (Archaeological Artefact Database of Finland), Historical travel environment model and data collection of environmental and cultural variation.

Keywords [1]

Interdisciplinary studies, Human past, Finland, Spatial data, Language data

1. Introduction

Integrative approaches to build holistic human histories are little by little covering the globe, and the work by BEDLAN team (Biological Evolution and Diversification of Languages, www.bedlan.net) and Human Diversity consortium at the University of Turku, Finland, have integrated the North-West Eurasian area into this emerging network. This paper brings together the data collections produced within and around the BEDLAN project, presenting a new infrastructure combining our current and forthcoming datasets of human diversity in the Uralic language speaker area with a special focus on Finland (Fig. 1). With the Uralic Trove we aim to advance the integrative studies of the human past in the study area as well as to support the further development of methodology for digital study of the human past. In this paper we present the different datasets and interdisciplinary work behind them..

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

Findable: Most data sets 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 data collections by reviewing them all in the same package.

Accessible: We aim to manage each in GitHub, which is a versatile repository where the team can update and curate each data according to needs. The data sets are being collected under the BEDLAN organisation, github.com/BEDLAN. Some data sets can also be downloaded directly from GitHub, but

[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)

all the data are or will be available as snap shots in Zenodo or OSF or similar repositories. In Zenodo, we use a BEDLAN to indicate the "community" that the data is part of.

Interoperable: The Uralic dataset are interoperable within themselves and with similar data from other language families through usage of standards: ISO-639-3 abbreviations and Glottocodes for languages, which - as well as updated coordinates - we curated through the Glottolog initiative. The format of language data follows emerging international standards such as usage of CLDF format. Cross-analyses of the Finnish data in Uralic Trove and other owned by other projects could be done e.g. by comparing municipality-wise data (e.g. Lynch et al. 2021).

Reusable: The data are stored in different formats (csv, excel, shapefiles) that readily allow for combination to other datasets or that can be changed into reusable form. Part of the data are shapefiles/polygon data which necessitates expertise in geospatial analyses. However, we hope to ease the reusability of the data through the user interfaces that allow easy eyeballing of different datasets. We also intend to improve the data availability in the user interfaces for this reason. It should be noted that actually the most used resource so far has been the pdf-map collection showing language speaker areas that we offer in the URHIA web app. They are regularly used in the Finno-Ugric conference talks and when the web app had crashed, we got email from a newspaper reported asking when it will be fixed again. This indicates that actually the end user is often someone who only a map where the data is visualized.

In all, building the Uralic Trove over the years since 2009 has been a long and interdisciplinary process. The disciplines involved include

1. Linguistics: Historical linguistics, typology, dialectology, etymology, contact linguistics; data achieved 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 Mesolithic Stone Age to Mediaval 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, which we have had to learn to come to terms with during the journey. In some cases we have even built standards for data collection (Rantanen et al. 2022). Also, it 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. To promote the field of digital humanities, I strongly support master level courses of data handling so that the start of the PhD work would be smoother.

Today our take on OA data is totally contrary to 2009 when OV started the project and aimed at keeping the data for our team only. It was only through publication of OA datasets that our work was recognized internationally. Now we see that digital humanities using OA data could induce international interest to study human history in NW Eurasia, and that this would be a good thing. More focus on research will give more weight to the area, and it is easier to build collaboration and get funding – and build a picture of the holistic history of human past in NW Eurasia. We encourage Uralists and other researchers of the area to make their data OA, for more researcher will add their voices and cumulative insights, and eventually – hopefully - lead to holistic understanding of human history in Scandinavia and NW Eurasia.

2. Data collections included to Uralic Trove

The data collection processes started according to BEDLAN research needs. In the beginning of the BEDLAN project we 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), and today the list includes over 30 languages (de Heer et al. 2024). Ten year ago we started to collect typological data, UraTyp, for studying evolution of typological contacts within Uralic family and with their neighbours (Norvik et al. 2022). The initiative had limited success until we developed a collaboration with University of Tartu, Uppsala and with Grambank initiative from the Max Planck Institute for Evolutionary Anthropology (Vesakoski 2023). For spatial study of Uralic languages we also needed a GIS-based data of language speaker areas. Such data did not exist and neither were there methods to build such a cartographical database (Rantanen et al. 2022a). In collaboration with prof. Jussi Ylikoski (UTU) and the authorship of Uralic Guide for Uralic Languages we were able to create such data: We first digitized maps from literature, sent them for comment to the language specialists, and then redrew the polygons accordingly (Rantanen et al. 2022b). 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 landscape of North-Western Eurasia. To make these maps available for other users we created another Zenodo release with pdfs of the maps, and will soon add thematic layers too as shape files.

The Finnish language is part of the Uralic languages, and also a special focus for the studies within BEDLAN and the Human Diversity consortium. BEDLAN project started from digitized Dialect Atlas of Finnish, which we now finally will publish as corrected version. Archeological past of Finland is provided with Archaeological Artefact Database of Finland, AADA, which is a massive work initiated already 2013. To model (pre)historical human travelling in Finland we created a historical travel environment model allowing for versatile study of how human potentially used the environment for travelling. Finally, we offer our compilation of environmental and cultural data of historical Finland. .

Besides offering the dataset in Zenodo, we have targeted resources for building an interactive user interface URHIA, that is presented in more details elsewhere in this volume (Roose et al.). We also got a possibilit to adopt a user interface built by MPI-EVA to promote the usage of the areal linguistic data (uralic.clld.org).

2.1. Uralic data collections

The Uralic language family (also known asFinni-Ugric) 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).

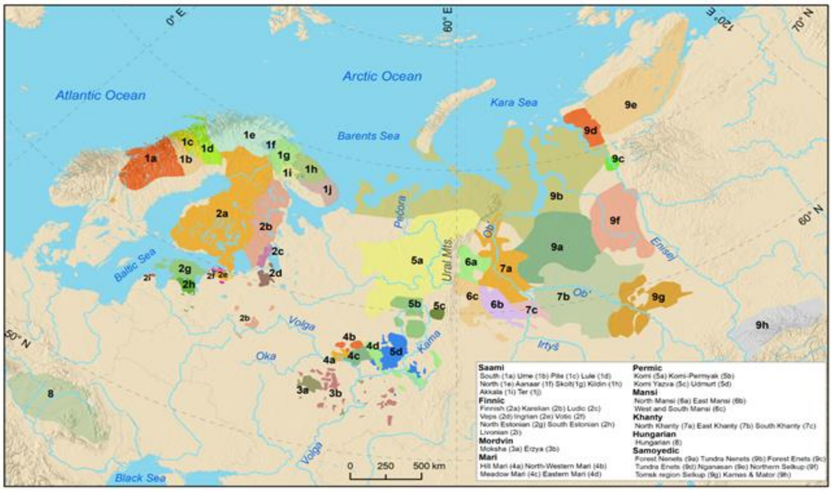

0° E
60° E
120° E
70° N
60° N
50° N
Atlantic Ocean
Arctic Ocean
Kara Sea
Barents Sea
Baltic Sea
Black Sea
Pečora
Ural Mts.
Ob'
Enisej
Volga
Oka
Kama
Irtyš
1a
1b
1c
1d
1e
1f
1g
1h
1i
1j
2a
2b
2c
2d
2e
2f
2g
2h
2i
3a
3b
4a
4b
4c
4d
5a
5b
5c
5d
6a
6b
6c
7a
7b
7c
8
9a
9b
9c
9d
9e
9f
9g
9h
0 250 500 km
Saami
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)
Mansi
North Mansi (6a) East Mansi (6b)
West and South Mansi (6c)
Khanty
North Khanty (7a) East Khanty (7b) South Khanty (7c)
Hungarian
Hungarian (8)
Samoyedic
Forest Nenets (9a) Tundra Nenets (9b) Forest Enets (9c)
Tundra Enets (9d) Nganasan (9e) Northern Selkup (9f)
Tomsk region Selkup (9g) Kamas & Mator (9h)

Figure 1: Uralic languages from Cartographical database of Uralic language speaker areas.

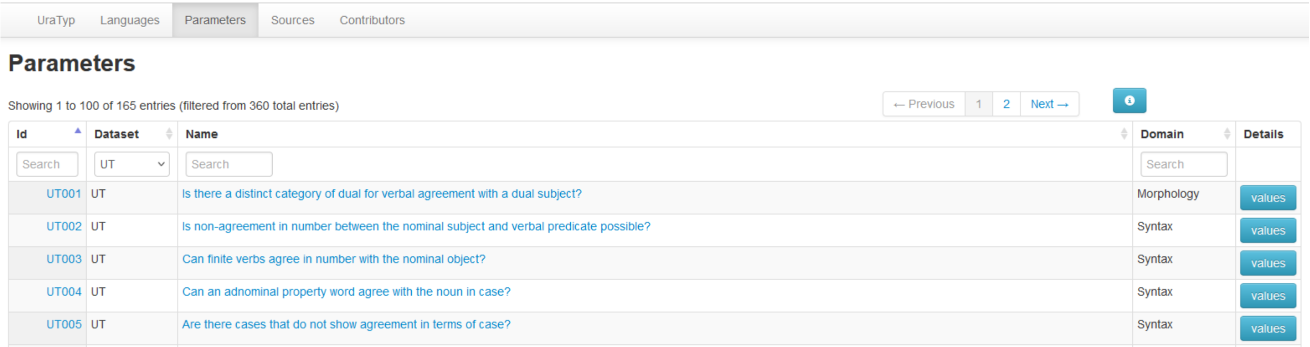
2.1.1. UraTyp

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 were compiled from descriptive grammars and grammar sketches when available, or by interviewing language experts. The data is being built in GitHub, published in Zenodo, and has an easily approachable method of access via the web app Uralic

Areal Typology Online built by MPI-EVA (Robert Forkel, uralic.clld.org). The features represent actually two underlying typological lists: Grambank data and UT data.

Grambank data (GB) is a list of 195 grammatical features collected for cirka 2450 world languages with the aim of studying global typological diversity. The list was produced by the Grambank initiative by Max Planck Institute (Skirgård et al. 2023). BEDLAN contributed the Uralic languages to GB. This is to say that the GB traits in UraTyp a) are found also from the Grambank database and that b) by the GB traits, Uralic language are fully interoperable with 2400 other word languages. Grambank data is easily accessible in grambank.clld.org

Uralic specific traits (UT) are 165 extra traits that were developed to describe variation within Uralic languages (Norvik et al. 2022), since the level of granularity of the Grambank traits is often insufficient to distinguish Uralic languages from each other. The UT list was 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 Uralic neighbors and collecting examples of the usage of each linguistic trait in the language. This data will be published by the next update of UraTyp. The current version in Zenodo DOI 10.5281/zenodo.5236365 but is easily accessible in Uralic Areal Typology Online, uralic.clld.org.



Parameters

Showing 1 to 100 of 165 entries (filtered from 360 total entries)

Id	Dataset	Name	Domain	Details
Search	UT	Search	Search	
UT001	UT	Is there a distinct category of dual for verbal agreement with a dual subject?	Morphology	values
UT002	UT	Is non-agreement in number between the nominal subject and verbal predicate possible?	Syntax	values
UT003	UT	Can finite verbs agree in number with the nominal object?	Syntax	values
UT004	UT	Can an adnominal property word agree with the noun in case?	Syntax	values
UT005	UT	Are there cases that do not show agreement in terms of case?	Syntax	values

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

2.1.2. UraLex

UraLex is a database of basic vocabulary data with cognate assessments indicating etymological connections between words for a meaning in different languages. This is a commonly used data type to construct quantitative language phylogenies (e.g. Gray & Atkinson 2003, Grollemund et al. 2015). The words listed for a particular language are the normal, everyday words corresponding to a standard list of meanings compiled for this project. The meaning lists include so-called Basic Vocabulary, parts of the lexicon which are likely to exist in most languages, and which are expected to be both semantically and morphologically simple. They are also expected to be relatively stable over time, and unlikely to be replaced through borrowing or semantic shift. These meanings include, for example, items of the lexicon which can be expected to exist in most languages, such as lower numerals, pronouns and body parts. The Uralix database contains basic vocabulary meanings from a union of several widely used sources, including the well established meaning lists by Swadesh 1952, 1955 (so called "Swadesh lists"), as well as the newer Leipzig-Jakarta list (Tadmor 2009). The full meaning list used in Uralex also includes a further set of meanings representing less-stable vocabulary that was first used in Lehtinen et al. (2014) to test the evolutionary behaviour of the less genealogically well-behaved lexicon.

The data is published under the Lexibank umbrella (List et al. 2022), and the data is available from the lexibank repository https://github.com/lexibank/uralex . Uralix has had three major released, the 1.0 release (Syrjänen et al. 2018) and corrected and updated 2.0 release (Syrjänen et al. 2021,) cover 26 languages as well as reconstructed Proto-Uralic. The 3.0 release (de Heer et al. 2023) introduces coding for borrowing relationships as well as cognates. It will also be made available and visible through Uralic Areal Typology Online.

2.1.3. Geographic database of Uralic languages

Geographical database of the Uralic languages instead is a spatial data with polygons of Uralic languages speaker areas (Rantanen et al 2022). It consist of multiple versions of each speaker area digitalized first from literature and corrected later by experts. Each language has at least two versions of the speaker areas: the current ones and the ones cirka 100 years ago. However, some of the languages have extinct since 20th century. Vizualized maps of the GIS data were part of the Oxford Guide for Uralic Languages (Rantanen et al. 2022b), and the original pdf's can be found also from URHIA. The data and static maps are available in Rantanen et al. (2021, data paper) in Zenodo DOI 10.5281/zenodo.4784188. As described in Roose et al. (2021) and Roose et al. (this volume), we also offer the raw data as an interactive cartographical interface in URHIA (https://urhia.fi).

The GIS data is readily usable for map making and has been used e.g. in many publications (e.g. Tambets et al. 2018) and user interfaces (e.g. our own Uralic Areal Typology Online). The speaker area data can be used in versatile ways: We have also made a 10 x 10 km grid over Eurasia with information if a given Uralic subgroup exists in the cells within the grid. This information is combined with spatial data of bio-geographical variation of the area to study if language speaker areas differ by their (a)biotic attributes (Roose, Nylen, Tolvanen & Vesakoski, ms).

2.1.4. Interdisciplinary spatial database of human history in the North-West Eurasia

Interdisciplinary spatial database of human history in the NW Euraasia is a collection of shapefiles and map visualizations that are done for various publications and grant applications. We now offer the static maps and GIS-files for further use in Zenodo (DOI:10.5281/zenodo.10376207). The collection of spatial information and maps of the past and environment in the Uralic languages speaker area consists of excessive amount of multidisciplinary data related to the vast region extending from Eastern Europe to Siberia, encompassing countries like Russia, Finland, and parts of Scandinavia. These datasets can be integrated for multidisciplinary purposes, allowing to explore human-environment interactions, migration patterns, and cultural evolution over time. As the data collections and mapmaking continue to evolve dynamically together with ongoing projects, the current repository will be updated accordingly. At the very moment the Zenodo only harbors the pdf maps.

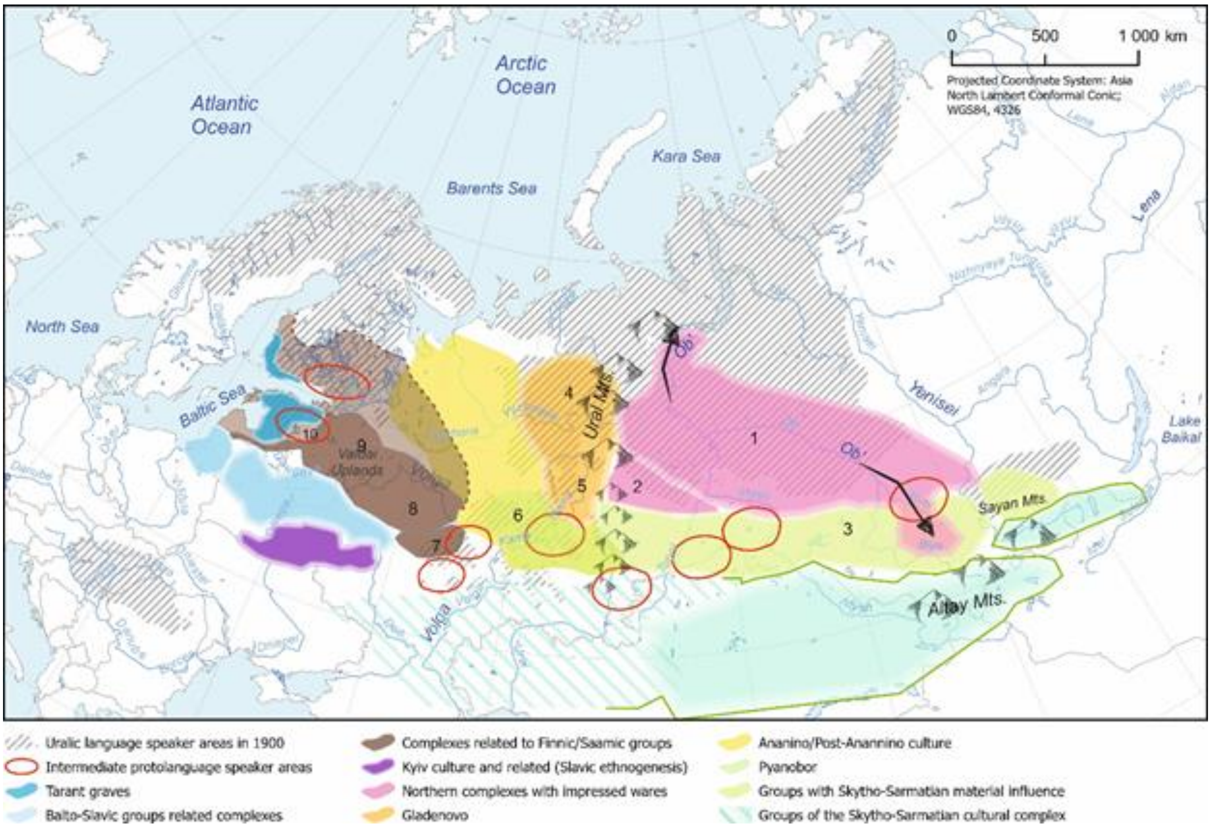

0 500 1 000 km
Projected Coordinate System: Asia North Lambert Conformal Conic; WGS84, 4326
Arctic Ocean
Atlantic Ocean
Kara Sea
Barents Sea
North Sea
Baltic Sea
Lena
Yenisei
Ob
Lake Baikal
Ural Mts.
Sayan Mts.
Altay Mts.
Valdai Uplands
Volga
1
2
3
4
5
6
7
8
9
10
Uralic language speaker areas in 1900
Intermediate protolanguage speaker areas
Tarant graves
Balto-Slavic groups related complexes
Complexes related to Finnic/Saamic groups
Kyiv culture and related (Slavic ethnogenesis)
Northern complexes with impressed wares
Gladenovo
Ananino/Post-Ananinno culture
Pyanobor
Groups with Skytho-Sarmatian material influence
Groups of the Skytho-Sarmatian cultural complex

Figure 3: Example of the interdisciplinary maps; original version of a map redrawn by the publisher to Vesakoski et al. (2024).

2.1.5. Uralic historical atlas, URHIA

We build an interactive web app Uralic Historical Atlas (URHIA, https://urhia.fi) for easy access to static maps and to provide a possibility for lay audience to create their own maps (Roose et al. 2021, Roose et al, this volume). URHIA is an open source spatial infrastructure GeoNode by GeoSolutions and integrated into UTU's spatial infrastructure (https://geospatial.utu.fi/resources/utu-geospatial-data-service/). It goes beyond a conventional data repositories for it is designed as an interactive spatial platform for researchers and lay audiences. Currently hosting the Uralic Language Atlas (based on above mentioned Cartographical Database of Uralic Language Speaker areas, Rantanen et al. 2022a) and the Archaeological Artefact Atlas of Finland (based on Archaeological Artefact Database of Finland, AADA, see below and Pesonen et al. 2024).

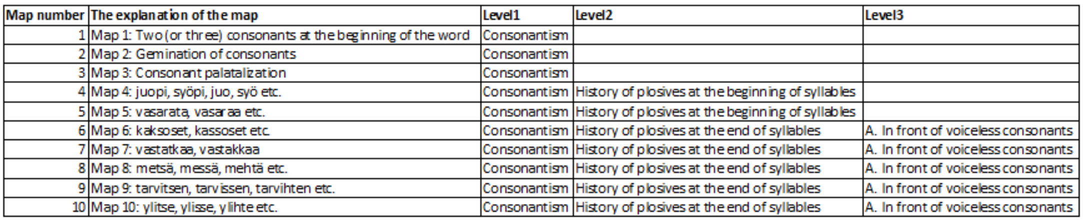
URHIA is still in process and aims to transform spatial data into a live data showroom, presenting thematic spatial datasets through interactive online maps. AADA represents a pioneering effort, being the first database of its kind in Finland and possibly globally. Its creation marks a milestone in the digitization and accessibility of archaeological data, setting the stage for similar initiatives worldwide.

2.2. Finnish data collections

The Uralic Trove also includes data covering only Finnish language area. A new profile area of University of Turku, the Human Diversity consortium (www.humandiversity.fi) focusses especially to Finland for there are multidisciplinary data sets available. At the moment we have published or are publishing the following:

2.2.1. Preindustrial dialect landscape of Finland

Dialect Atlas of Finnish (Kettunen 1930a; 1930b; 1940a; 1940b) is a map collection of linguistic variation, presenting dialectal landscape of Finnish in 1920s and 1930s. This is the time before mass migration to cities and major population increase from Karelian dialect are (400 000 refugees from Russian Karelia during the WW2). Kettunen interviewed people in 525 Finnish-speaking municipalities to collect information of 213 linguistic traits describing morpho-phonological and lexical variation of Finnish. The pdf maps are available at http://kettunen.free.nf. Embleton & Wheeler (1997, 2000) initiated the digitizing process of the data together with Insitute 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 the data with full documentation in Santaharju et al. (this volume) in Zenodo (https://zenodo.org/records/10078078).



Map number	The explanation of the map	Level1	Level2	Level3
1	Map 1: Two (or three) consonants at the beginning of the word	Consonantism		
2	Map 2: Gemination of consonants	Consonantism		
3	Map 3: Consonant palatalization	Consonantism		
4	Map 4: juopi, syöpi, juo, syö etc.	Consonantism	History of plosives at the beginning of syllables	
5	Map 5: vasarata, vasaraa etc.	Consonantism	History of plosives at the beginning of syllables	
6	Map 6: kaksoset, kassoset etc.	Consonantism	History of plosives at the end of syllables	A. In front of voiceless consonants
7	Map 7: vastatkaa, vastakkaa	Consonantism	History of plosives at the end of syllables	A. In front of voiceless consonants
8	Map 8: metsä, messä, mehtä etc.	Consonantism	History of plosives at the end of syllables	A. In front of voiceless consonants
9	Map 9: tarvitsen, tarvissen, tarvihten etc.	Consonantism	History of plosives at the end of syllables	A. In front of voiceless consonants
10	Map 10: ylitse, ylisse, ylihte etc.	Consonantism	History of plosives at the end of syllables	A. In front of voiceless consonants

Figure 4: Examples of linguistic traits included to the Dialect Atlas of Finnish. E.g. map number 7, i.e. trait number 7, tells about consonantism, and more precisely what is the variation of plosives in the end of syllables, and especially in front of voiceless consonants. As and example the map 7 uses the word *vastatkaa* “answer!”, that in some dialects is *vastakkaa.* The actual data tells which variant exist in the 525 studied municipalities (see also Santaharju et al, this volume).

2.2.2. Archaeological Artefact Database of Finland

Archaeological Artefact Database of Finland (AADA) is a spatial data offering typological categorisation, spatio-temporal context, location and photos of 49 000 artifacts stemming from the Early Mesolithic to the end of the Iron Age. The data was compiled by visiting Finnish museums and studying the artifacts one by one between 2013-2020. The data is described in Pesonen at al. (2024) and is maintained in GitHub and provided in Zenodo: the actual data in DOI10.5281/zenodo.10437703 (Pesonen et al. 2024 data paper) and the photos in DOI10.5281/zenodo.11256533 (Moilanen et al. 2024 data paper). To improve digital archeology, we also published the R code with which the maps were created for the publishing paper (Roose 2024). AADA data will be also part of URHIA user interface (Roose et al. this volume).

The data in AADA is description of artefacts. Each artefact has collection number, which with it can be combined to other information of the archeological site. The artefacts also have coordinates, which will allow for placing them alongside with other other spatial data.

2.2.3. Historical travel environment model over Finland

Historical travel environment model - spatial model for historical travel effort - is a spatial data with terrain and landscape attributes (Fig. 5) coupled with information of travel speed in different environments. The ecological and geographical data is compiled from digital databases and digitized from literature. The travel speed estimates are achieved 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 will be published in Zenodo as part of a manuscript in progress [Zenodo link will however appear here to the final version].

The data is organised into 1 x 1 km grid over Finland. Each grid includes information of the eco-geographical factors. The grid approach allows for versatile cross-analyses of data as any other attributes could be added to the grid - such as information of Mörby ceramic vessels exist in the given grid. However, applying such approach is only in progress now.

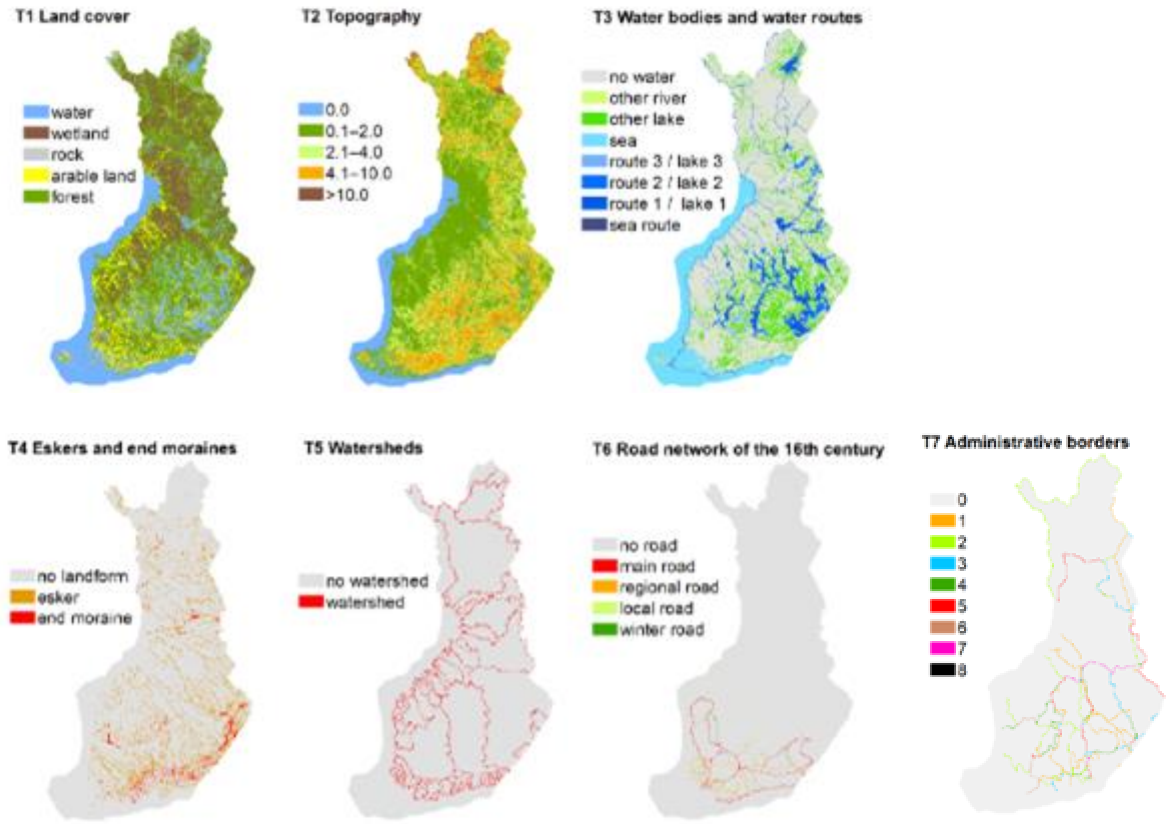


Figure 5: The thematic layers in the travel environment model. The area is dividied in 1 x 1 km grid and each grid has 7 different costs according to prevalence of different classes of these 7 landscape surfaces.

2.2.4. Historical environmental and cultural variation over Finland

The Uralic Trove will include also spatial data of environmental and cultural variation in Finland (used in Honkola et al. 2018 and Lynch et al. 2022) and folk-culture data from 1600-1800 used in Rantanen, Santaharju et al. (in preparation). Publishing these data as one launching paper is in process (Santaharju et al. in prep). The data will be offered in Zenodo (https://doi.org/10.5281/zenodo.13975170). Both these data sets are municipality-vice data and are readily comparable to e.g. municipality-vice data of dialect variation (Honkola et al. 2018) or to information of human movements within Finland (Lynch et al. 2022).

3. Acknowledgements

Funding for data collection has been achieved especially from Kone Foundation Kone Foundation (grant "Kielen murteet biologisen lajiutumisen näkökulmasta", Urho Määttä as PI, "SumuraSyyni" and "AikaSyyni" OV as PI, and "Kippo" with Unni Leino as PI). MR was funded mainly by UTU BGG graduate school. There has been vast amount people contributing to compile the database included now to UraLaari - they are the army authors of the different data papers.

1. References

Honkola, T., Ruokolainen, K., Syrjänen, K. J. J., Leino U, Tammi, I, Wahlberg, N. & Vesakoski, O. (2018). Evolution within a language: Environmental differences contribute to divergence of dialect groups. *BMC Evolutionary Biology*. 18:132.

Lehtinen, J., Honkola T., Korhonen K., Syrjänen K., Wahlberg N., Vesakoski O. (2014). Behind family trees: Secondary connections in Uralic language networks. *Language Dynamics and Change*. 4: 189-221. DOI: 10.1163/22105832-0040200

Lynch, R., Loehr, J., Lummaa, V., Honkola, T., Pettay, J. & Vesakoski, O. 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.

Norvik, M., Jing, Y., Dunn, M. Forkel, R., Honkola, T., Klumpp, G., Kowalik, R., Metslang, H., Pajusalu, K., Piha, Saar, E., Saarinen, S. & Vesakoski, O. Uralic typology in the light of new comprehensive data set. Journal of Uralic Linguistics 1: 4-41.

Rantanen, T., Tolvanen, H., Honkola, T., Vesakoski, O. 2021: A comprehensive spatial model for historical travel effort – a case study in Finland. *Fennia* 199 (1) 61-88.

Rantanen, Timo, Harri Tolvanen, Meeli Roose, *et al.* (2022). 'Best Practices for Spatial Language Data Harmonization, Sharing and Map Creation—A Case Study of Uralic'. *PLOS ONE* 17 (6): e0269648.

Rantanen, T., Vesakoski, O. & Ylikoski, J. 2022: Mapping the distribution of the Uralic languages. Marianne Bakró-Nagy, Johanna Laakso & Elena Skribnik (eds.), The Oxford Guide to the Uralic Languages. Oxford University Press.

Roose, M., Nylén, T., Tolvanen, H. & Vesakoski, O. 2021: User-centred design of multidisciplinary spatial data platforms for human-history research. *SPRS International Journal of Geo-Information* 10(7): 467.

Roose, Meeli, Timo Rantanen, Dmitri Kuznesov, *et al.* (2023). 'Collection of Spatial Information and Maps of Human Past and Environment in the Uralic Languages Speaker Area'. https://doi.org/10.5281/zenodo.10081902 [data set]

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*. submitted ms

Roose, Meeli, Timo Rantanen, Dmitri Kuznesov, et al. (2023). 'Collection of Spatial Information and Maps of Human Past and Environment in the Uralic Languages Speaker Area'. https://doi.org/10.5281/zenodo.10081902 [data set]

Santaharju, Jenni, Syrjänen, Kaj, Honkola, Terhi, Seppä Perttu, Vesakoski, Outi and Leino, Unni. *The digitized Dialect Atlas of Finnish by Lauri Kettunen. Digital Humanities in the Nordic and Baltic Countries Publications*. Submitted ms

Skirgård, Hedvig et al. (> 1 authors) 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. Science Advances. http://dx.doi.org/10.1126/sciadv.adg6175

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. https://doi.org/10.5281/zenodo.1459402.

Syrjänen, Kaj, Luke Maurits, Unni Leino, Terhi Honkola, Jadranka Rota, and Outi Vesakoski. 2021. “Crouching TIGER, Hidden Structure: Exploring the Nature of Linguistic Data Using TIGER Values.” *Journal of Language Evolution*, no. lzab004 (November). https://doi.org/10.1093/jole/lzab004.

Tambets, K., Yunusbayev, B., Hudjashov, G., Ilumäe, A-M., Rootsi, S., Honkola, T., Vesakoski, O., Atkinson, Q., Skoglund, P., Kushniarevich, A., Litvinov, S., Reidla, M., Metspalu, E., Saag, L., Rantanen, T., Karmin, M., Parik, J., Zhadanov, S.I., Gubina, M., Damba, L.L., Bermisheva, M., Reisberg, T., Dibirova, K., Evseeva, I., Nelis, M., Klovins, J., Metspalu, A., Esko, T., Balanovsky, O., Balanovska, E., Khusnutdinova, E., Osipova, L., Voevoda, M., Villems, R., Kivisild, T., Metspalu, M. 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. – Tartu Ülikooli Lõuna-Eesti keele- ja kultuuriuuringute keskuse aastaraamat XXI–XXII. Pühendusteos Karl Pajusalule 60. sünnipäevaks. Toim. Eva Saar, Miina Norvik, Eva Velsker. Tartu: Tartu Ülikooli Kirjastus, 276–282.

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

Vesakoski O, Salmela E ja Piezonka, H. (2024). Uralic archaeolinguistics. In *Oxford Handbook of Archaeology and Languages*, ed. by Martine Robbeets ja Mark Hudson. In press.

Datasets:

UraLex 1.0 2018. Kaj Syrjänen, Jyri Lehtinen, Outi Vesakoski, Mervi de Heer, Toni Suutari, Michael Dunn, Urho Määttä, Unni-Päivä Leino. lexibank/uralex: UraLex basic vocabulary dataset. DOI:10.5281/zenodo.1459402

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

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

UraTyp 2022. Miina Norvik, 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: Uralic typological data set DOI: 10.5281/zenodo.6392555

Collection of spatial information and maps of human past and environment in the Uralic languages speaker area. Meeli Roose, Timo Rantanen, Henny Piezonka, Eli Salmela, Kerkko Nordqvist, Petro Pesonen, Ulla Moilanen, Terhi Honkola, Dmitri Kuznesov, Outi Vesakoski. In making: DOI:10.5281/zenodo.10376207

Cost-distance model over Finland. Timo Rantanen, Harri Tolvanen, Terhi Honkola, Outi Vesakoski. In making.

Archaeological artefact database of Finland (AADA). Petro Pesonen, Ulla Moilanen, Jarkko Saipio, Meeli Roose, Outi Vesakoski, Päivi). Typological description of museum artefacts of Finland from Stone Age, Bronze Age and Iron Age with coordinates (https://zenodo.org/records/10437704)

Archaeological artefact database of Finland (AADA) Photographs. Ulla Moilanen, Petro Pesonen, Jarkko Saipio, Jasse Tiilikkala, Archaeological artefact database of Finland (AADA) Photographs Muhammad Usman Sanwal. (https://zenodo.org/records/11256533)

Digitized Dialect Atlas of Finnish by Lauri Kettunen. Jenni Santaharju, Kaj Syrjänen, Terhi Honkola, Jyri Lehtinen, Perttu Seppä, Outi Vesakoski, Unni Leino.