

Uralic typology in the light of a new comprehensive dataset

Miina Norvik^{1,2,3} (0000-0001-5781-3916), Yingqi Jing^{1,2} (0000-0003-2831-5701), Michael Dunn¹ (0000-0001-5349-5252), Robert Forkel⁴ (0000-0003-1081-086X), Terhi Honkola^{2,3} (0000-0002-3330-0857), Gerson Klumpp³, Richard Kowalik⁶ (0000-0003-4903-997X), Helle Metslang³ (0000-0002-6397-6532), Karl Pajusalu³ (0000-0001-5554-5049), Minerva Piha^{1,5} (0000-0003-1307-8491), Eva Saar³ (0000-0002-1849-6425), Sirkka Saarinen⁵ (0000-0002-0747-6517), Outi Vesakoski^{2,5} (0000-0002-7220-3347)

¹ Department of Modern Languages, Uppsala University, Sweden

² Department of Biology, University of Turku, Finland

³ Department of Estonian and General Linguistics, University of Tartu, Estonia

⁴ Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology

⁵ Department of Finnish and Finno-Ugric Languages, University of Turku, Finland

⁶ Department of Linguistics, University of Stockholm, Sweden

Abstract

This paper presents the *Uralic Typological Database* (UraTyp 1.0), a typological dataset of 35 Uralic languages and a total of 360 features, mainly covering the levels of morphology, syntax, and phonology. The features belong to two different datasets: 195 features' definitions originate from the Grambank (GB) database, developed for comparison of world language typology, whereas 165 features (UT) have been designed specifically to describe the typological variation within the Uralic language family. We present a series of analyses of the dataset demonstrating its scope and possibilities. The complete data set correctly identifies the main Uralic subgroups in a Principal Components Analysis, whereas GB data alone is insufficiently granular to detect this family-internal structure. Similar analyses limited to various typological subdomains also give variable results. A model-based admixture analysis identifies four distinct areas of historical interaction: Saami, Finnic, the Volga area and Ob-Ugric.

Keywords: Uralic languages, typology, areal linguistics, syntax, morphology, phonology, quantitative linguistics

1. Introduction

Uralic languages have been mostly neglected in global comparisons of language typology. Where Uralic languages are included, most often it is only Finnish, Hungarian, and to some extent Estonian – all large state languages from Western Europe – which are treated as representative (see, e.g., Greenhill et al. 2010). Whereas typological information on these languages is readily available, access to documentation of some other members of the Uralic family additionally requires familiarity with scientific literature in various languages (e.g., Russian, Finnish, German, Hungarian, Swedish); the data is often scattered across different types of sources originating from different time periods. Despite the fragmentation of sources, the family is relatively well studied, and in recent years the documentation and description of these languages has improved in both breadth and depth. Still, due to the lack of systematically organized comparative typological data on Uralic languages, outstanding questions about the structural diversity of the family are still to be answered.

The lack of coherent, family-wide typological data was the main driving force for creating the Uralic Typological Database (UraTyp), the first large-scale typological database where Uralic languages from all the main branches are equally represented. The data includes 360 binary features belonging to two different datasets: 195 features defined in the Grambank database (“GB data”) developed for comparative typological investigation of the languages of the world, and 165 designed by the authorship (MN, KP, HM, GK, ES) to specifically capture the Uralic typology (“UT data”). The current paper has two main aims. First, the database is offered for public use with this paper and thus we introduce the database: how the feature lists were created, how the data was collected and how to access the data. The second aim is to visualize the linguistic patterns and typological diversity within the Uralic languages. With quantitative analyses we clustered the data to see if it accurately identifies the conventional branches of the Uralic languages; a subsidiary question was to see whether the GB traits alone could represent Uralic typological variation in global cross-comparisons. In addition, as the features represented three main levels of language structure (phonology, morpho-lexicon, and syntax), we explored whether typological diversity at each of these three levels would similarly identify the Uralic language groups. Finally, we wanted to demonstrate that typological data can be used also for diachronic studies. We present model-based clustering analyses which allow for historical inferences based on the evolution of Uralic

typological diversity. In all, we demonstrate the possibilities for combining extensive data and quantitative approaches for studies of the Uralic language family.

The early studies of the Uralic languages were descriptive and influenced by linguistic understanding of their dominating contact languages, such as German or Russian. Since the 19th century comparisons between different Uralic languages and subgroups have been carried out mainly in a historical-comparative framework. The functional-typological study of these languages has developed only during the last decades focusing mainly on certain morphosyntactic features, such as negation (Miestamo et al. 2015) or the essive (De Groot 2017). The book on negation contains 17 chapters on different languages and a typological overview, whereas the book on the essive includes 21 languages in separate chapters and a typological overview. Although there exists in-depth research on particular typological features, research on a multitude of features belonging to various linguistic levels is lacking. Need for systematic data sets where the values of features are defined in a similar way to enable family-wide typological studies has been expressed in several studies (see Klumpp et al. 2018; Miestamo 2018; Veenker 1985).

The traditions of Uralic historical linguistics have had an impact on defining research questions of typological studies, too. For instance, the location and spread of the Proto-Uralic homeland, the typological profile of Proto-Uralic, and historical contacts with other language families are still topical issues in the newest studies related to Uralic languages (e.g., Nichols 2021; Grünthal et al., in press). Besides that, the areal-typological approach has been important both in studies of Uralic languages (Helimski 2003) and in broader studies of specific language areas, such as the Circum-Baltic area (Koptjevskaja-Tamm & Wälchli 2001). The position of Hungarian within the Uralic family and its position with regard to Standard Average European has been a question of special interest (see Haspelmath 2001; Laakso 2020). The relationship between various language levels in the typological shift of Uralic languages has recently also found attention (cf. Klumpp et al. 2018; Nichols 2021). These studies, however, are mainly qualitative, or are based on a limited set of features.

There are also large-scale quantitative studies that make use of the family-wide data of the Uralic languages but they have mainly been based on lexical data (Honkola et al. 2013; Lehtinen et al. 2014; Syrjänen et al. 2013). For instance, Syrjänen et al. (2013) studied the robustness of the shape of the Uralic language family by analyzing different sets of basic vocabulary with Bayesian phylogenetic methods. Including lexical data in the research means considering cognates and

sound changes, thus, working with a similar linguistic material as comparative linguists do. As regards structural data, there are studies that take a closer look at a particular subfield. For example, Pajusalu et al. (2018) used 33 phonological traits from 28 Uralic languages, and found that phonological traits divide the family primarily into western and central-eastern clusters. Although typological data can act as an independent source of information, together with lexical data it can provide us with a more complete picture of the Uralic languages and their past.

As the current typological diversity is the product of diachronic processes, it is important to look at synchronic diversity in terms of the diachronic processes which produced them. This cannot be studied from individual features only but family-level typological surveys are needed. Past decades have witnessed the emergence of various online databases that contain structural information, the World Atlas of Language Structures (WALS; see Dryer & Haspelmath 2013) being the most widely known. Currently, WALS contains information on 2,662 languages and 192 features. There is information on 27 Uralic languages but their representation varies: in the entire database, Veps has 1 entry (i.e., feature studied), Livonian – 5, Erzya – 39, Udmurt – 46, Estonian – 60, Mansi – 63, Hungarian and Finnish – 155. Due to unequal coverage in WALS and the fact that Finnish, Hungarian, and Estonian are the most widely studied Uralic languages, other languages have often been neglected or are included sporadically in large-scale global comparisons. For example, Greenhill et al. (2010) only include Hungarian and Finnish in their study and Dediu & Levinson (2012) consider 3 to 12 Uralic languages depending on the feature. This western geographic bias has been a persistent problem for a balanced understanding of the typological diversity of the Uralic family.

The most recent attempt at global data collection is the Grambank initiative developed at the Department of Cultural and Linguistic Evolution at the Max Planck Institute for the Science of Human History and at the Max Planck Institute for Evolutionary Anthropology (see Skirgård et al., submitted). While WALS aims at global coverage of interesting typological features, the Grambank data was meant to thoroughly cover typological diversity across language families. The developers of Grambank have set a goal to provide structural information about half of the world's languages (<https://glottobank.org/#grambank>). We joined the project to help to collect the Uralic languages for the Grambank database. With our experience in many different scientific traditions (notably Finnish and Russian), and by involving language experts, the Grambank part of the data now includes maximally complete data from as many as 29 Uralic language varieties. Grambank

data aims at analyzing the global linguistic diversity (Skirgård et al., submitted ms). Accordingly, the linguistic features collected in Grambank are designed to differentiate language families from each other and the data only partially fulfils the need to study typological variation within the Uralic family. We wanted to supplement the Grambank list of features with Uralic-specific data. Alongside this paper we will publish the Uralic typological data including both the GB and UT features (see Section 2.3 for data availability). Further we offer the data for easy use in the visual user interface Uralic Typological Database (UraTyp) in (<https://uralic.clld.org/>). The visual user interface is built in co-operation with MPI-EVA.

The paper proceeds as follows. In Section 2, we introduce the UraTyp database. We first give an overview of various attempts to create such a database and then proceed to explain how the UraTyp database was created; finally, we provide information on how to access the data. Section 3 presents the methods and results of analyses carried out using the UraTyp database. Section 4 provides a discussion and Section 5 draws main conclusions.

2. UraTyp & Uralic languages in Grambank

2.1. Previous systematic documentation of Uralic typological diversity

Whereas there are typological databases containing information on the languages of some branches of Uralic, e.g. the Typological Database of the Ugric Languages (Havas et al. 2015), attempts to build a database covering the whole Uralic family have turned out to be unsuccessful. The first known attempt to systematize the very extensive but inconsistent information available on the structures of the Uralic languages was the “Dialectologia Uralica” project initiated by Wolfgang Veenker (Hamburg) in the 1980s. He presented his idea at the Congress for Finno-Ugric Studies in 1980, and organized the kick-off symposium in Hamburg in 1984. Veenker introduced a unified model for classifying and systematizing morphological information of different languages and dialects. His method consisted in coding morphological forms on a uniform basis: in the paradigms every morphological form becomes a code; every category has its own position in the code and its values are marked by numbers. Veenker also planned to develop models for phonology and morphonology. The first attempts to apply the morphology model for different languages showed

that it worked well. Nevertheless, the project requiring broad international cooperation did not take off; above all, the Iron Curtain was an obstacle. (Veenker 1985; Hausenberg & Kokla 1988.)

In the 2000s, at the initiative of Ferenc Havas (Budapest), a new attempt was made to establish international cooperation in compiling a typological database of Uralic languages. Havas presented his idea at the Congress for Finno-Ugric Studies in 2005 and 2010; in 2010, he also organized a workshop on this topic. In 2008, a kick-off conference of the Uralic Typology Database Project was held in Vienna, focusing on morphosyntactic and syntactic features. Havas' idea was based on a functional-typological approach: to collect and present typological specifications of different syntactic, morphosyntactic and morphological functions in Uralic languages and present them in the form of tables (Havas 2010). As there were also no possibilities to implement this project in its entirety, the Typological Database of Ugric Languages was launched at Loránd Eötvös University in Budapest. It has been publicly available since 2015 and it contains 200 morphophonological, morphological, morphosyntactic and syntactic features from four Ugric varieties: Synja Khanty, Surgut Khanty, Northern Mansi, and Hungarian (Havas et al. 2015).

In 2016, the workshop “Typology of Uralic Languages: Towards Better Comparability” at the 49th Annual Meeting of the Societas Linguistica Europaea in Naples was organized by Gerson Klumpp (Tartu), Lidia Federica Mazzitelli (Bremen), and Fedor Rozhanskiy (Tartu). Outcomes of the workshop are presented in a publication that contains systematic typological overviews. E.g., Klumpp et al. (2018) analyze 25 phonetic and grammatical features in 30 Uralic varieties, and Pajusalu et al. (2018) study 33 word-prosodic and segmental features of 28 Uralic varieties.

The UraTyp database presented in this paper grew out from seed money by Kone Foundation in 2013 but it became fully-fledged only thanks to funding from the University of Turku (*Kipot ja kielet* ‘Pots and languages’, 2018–2020). Collecting the GB features for Uralic languages and creating the UT list of features and collecting them was a joint effort between the University of Turku, University of Tartu and Uppsala University. The team cooperated with the Grambank initiative to ensure a similar data collection procedure within all the GB languages.

Here, it is also worth noting that databases like Grambank and UraTyp do not replace or downgrade previous databases but rather supplement them by introducing new angles to study typological diversity and by opening up new possibilities for various types of quantitative and qualitative analyses.

2.2. Creating the UraTyp database

2.2.1. *Uralic languages in Grambank*

The Grambank database was designed to give a broad typological overview of languages, coded in terms of abstract features – usually in the form of presence or absence of a particular grammatical function/feature. The features included (all in all 195) were chosen as the result of several generations of development of questionnaires, workshopped by a broad consortium of linguists. Within the Grambank project languages were coded by trained coders, for the most part using published sources. Each feature had a “patron”, who was responsible for documenting coding decisions and communicating standardized coding principles to the coders. (Full description of Grambank features will appear in Skirgård et al. submitted *ms.*)

We employed a coder trained within the Grambank project (Richard Kowalik) to collect the Grambank features from those Uralic languages not yet included in the Grambank database. Before our contribution, Grambank coders had already initiated the collection of some Uralic languages, but coverage of this language family was rather sporadic due to lack of typologically oriented literature. We now succeeded in covering the Uralic languages better for we used language experts as data source besides grammar descriptions and grammar sketches. We chose this strategy since for several Uralic languages, comprehensive literature is not yet available but several of the experts are working on grammatical descriptions of the languages. Therefore, answers for features were also extracted from the experts’ unpublished grammar sketches or their private databases. Consulting language experts yielded far more complete coverage of data than is typical for typological databases of this kind. The list of language experts can be found in the Acknowledgments and in full details from the data release (Norvik et al. 2021; see also Appendix 2). From the authors of this paper GK, RK, HM, MN, KP, MP, ES also acted as language experts.

2.2.2. *Defining the UT features*

In order to make the UraTyp database compatible with the Grambank database, we adopted the general principles used for designing the GB questionnaire and collecting the data. Thus, the questions about typological features were created in a way they could be answered in a binary form: Yes/Present or No/Absent. To compare, in WALS, the number of answer options varies

from question to question. The following example illustrates how the information is collected/presented in these three databases:

- (1) Grambank: Is there overt morphological marking on the verb dedicated to past tense? – Yes/No
- (2) UraTyp: Can tense be expressed overtly on the negative marker? – Yes/No
- (3) WALS: The Past Tense –
 - a. Past/non-past distinction marked; no remoteness distinction,
 - b. Past/non-past distinction marked; 2–3 degrees of remoteness distinguished,
 - c. Past/non-past distinction marked; at least 4 degrees of remoteness distinguished,
 - d. No grammatical marking of past/non-past distinction (Dahl & Velupillai 2013)

The UT features were designed not to repeat but to elaborate on the GB features, keeping in mind the specifics of the Uralic languages. To achieve this, we familiarized ourselves with the GB features. Whereas Grambank does not contain any features on phonology, the UT list was designed so that about 1/3 of the features would ask about phonology. Differently from GB features, the UT features were provided with glossed linguistic examples to illustrate the feature. Glossed examples can also be found in WALS and the Typological Database of the Ugric Languages. The final version of the UT questionnaire included 165 questions, each one providing information on a typological feature: 51 on phonology, 53 on syntax, 55 on morphology and 6 on the lexicon.

When creating the UT questions, our goal was to have them as broad as possible, while also capturing the variation within the Uralic family. The questions were also designed to be specific enough to avoid ambiguous answers. Following the procedure used to develop Grambank, every question was provided with information on what to consider when answering a question (e.g., in the case of UT questions on phonology it was asked not to consider recent loanwords). The descriptions are included in the user interface but they are also available in Norvik et al. (2021).

Regardless of the attempt to have the UT questionnaire finalized by the time the coding process started, almost every language brought up a handful of issues that had to be tackled. Some of the issues required us to make some refinements in the questions or in the accompanying descriptions (the GB feature list had undergone this iterative process earlier). For instance, the question *UT045 Is copula needed for predicate nominals in the 3rd person form of the present*

tense? specifically asks about the 3rd person form of the present tense. Although for some languages (e.g., Estonian) no specification was needed, in Surgut Khanty, there is no copula in the 3rd person form of the present tense but its use is optional in the 1st and 2nd persons, while in the past tense copula is obligatory in all persons (Csepregi & Gugán, to appear). Such revisions of questions and descriptions throughout the process enabled us to diminish ambiguity and keep the amount of missing information to a minimum. Occasionally, this meant re-coding the respective features in the case of languages that had already been coded. This could be easily done as we were working with around 30 languages and the answers were provided with examples. The UraTyp 1.0, which is released together with the present article, represents the finalized version of the database; the respective datasets were also used to carry out the analyses presented in this article.

2.2.3. Coding the Uralic languages

Most of the UT questions were answered during face-to-face meetings between the language expert and the coder (see the full list of language experts in Norvik et al. 2021). When filling out the UT questionnaire, the language experts were encouraged to provide examples or check examples provided by the coder whenever the answer was 1 ‘yes’ and add comments if necessary (answers for GB questions are occasionally followed by a comment). Depending on what was available, the language experts consulted grammar books/sketches, text collections, fieldwork data or used their knowledge as a native speaker. Although consulting experts and giving examples (and sometimes also commenting on them) made the process time-consuming, this adds a valuable qualitative dimension to the database while also reducing the number of “no information” entries to the minimum. Furthermore, comments and examples turned out to be useful in the process of revising and correcting the data (see above) as they enabled us to understand the reasoning behind an answer and discover obvious mistakes whenever a question was misinterpreted. Working with experts was also preferred when filling out the Grambank questionnaires for the Uralic languages. The outcome in the case of both datasets is a combination of literature search by coder and expert and interviews with the experts.

Throughout the process, it was necessary to agree upon several recurring coding-related questions brought up by our team members or language experts.

- ***What language / state of the language are we coding?*** The general principle was to code the modern-day standard language whenever possible. Still, there are languages (such as Finnish) that have high prestige, official status and a literary standard, and where in addition to dialects there is a spoken variety that may be quite different from the literary standard. In this case, we coded the literary language but, if something was especially dominant in the spoken language, we included such information as well. As regards Uralic languages which are included in our database but are not in active use anymore (e.g. Ingrian), have gone extinct (Kamas) or have no literary standard but exist in the form of several dialects (e.g. Ludic), we chose one particular language variety and considered what is/was characteristic or more widely spread in it. In some instances, this meant coding the language of the mid-20th century.
- Another issue, related to the previous one, is the **doculect** a coding is based on. A doculect is a ‘documented lect’, i.e., a linguistic variety as described in a specific source (Good & Cysouw 2013). Any grammar or data source will inherently reflect a certain linguistic variety, also with respect to time/a chronolect. For instance, the (written) material that answers for Eastern Mansi are based on is more than 110 years old, whereas other languages (e.g. South Saami) are coded based on contemporary, spoken language. Such differences in doculects are inherently part of any large-scale typological data set.
- ***How to ensure consistency in providing answers?*** To achieve this, following the GB principles, the questions were designed so that it would be possible to ask for the presence/absence of a feature via its function and not by its name (e.g., case names) as these might be put to different uses depending on a tradition of language description. Still, as not all the necessary information can be included in questions, the information on what to take into account when providing answers was included in the accompanying descriptions (as also done in GB). For example, in the case of questions on differential object marking (UT110–112) we only asked to think about finite clauses. Such decisions were made to avoid instances where an answer 1 ‘yes’ can mean several things. We also had joint training for the coders of UT features to ensure consistency in understanding and explaining the questions. The GB features were coded according to the training procedure of the Grambank team.

- ***How to deal with foreign influence?*** No language lives in a vacuum, thus, ignoring all the foreign influence would distort the picture and give as a result a language that never existed. However, we decided not to consider the very recent foreign influence or loss of features that go hand in hand with language death. For example, in the case of the question *UT158 Is there a constraint on word-initial r?* we specified in the description that when answering the question recent Russian loanwords are not considered. Still, it was not always easy to decide. The comments section was used to provide extra information about possible foreign influence.
- ***How common does a feature have to be in order to count as 1 ‘yes’?*** In general, this issue was avoided by careful formulation of the questions. Occasionally it was specified in the question whether something should be common or possible, e.g., *UT019 Is it common to use a verb in the present tense for future time reference?*, *UT100 Is it possible to use singular with paired body parts/clothing/accessories that accompany them?*. Again, general principles on when to code 1 ‘yes’ or 0 ‘no’ were specified in the descriptions. Whenever it was necessary to add some extra information it was done in the comments section of the language in question.

2.2.4. Combining GB and UT data into UraTyp

The 165 UT questions with additional 195 questions in the GB questionnaire make up the Uralic Typological Database (UraTyp) with 360 features. The features belong to different typological domains, including e.g. 12 questions of agreement (5 in GB and 7 in UT data; see the exact distribution in Figure 1). To illustrate the variation of themes we further listed the questions by keywords (Figure 2). The distribution by domain (Figure 1) is meant to give a broader picture of the features included in UraTyp, while the keywords (Figure 2) provide a more fine-grained distinction of features. It is important to note that both Figures 1 and 2 include 147 out of 195 GB questions as the remaining 48 questions were not considered relevant for the Uralic languages (e.g., questions on gender distinctions are irrelevant as none of the Uralic languages expresses this). The UraTyp database, however, includes all 360 questions in order to allow for comparison of Uralic languages to global Grambank data (see more in Section 5).

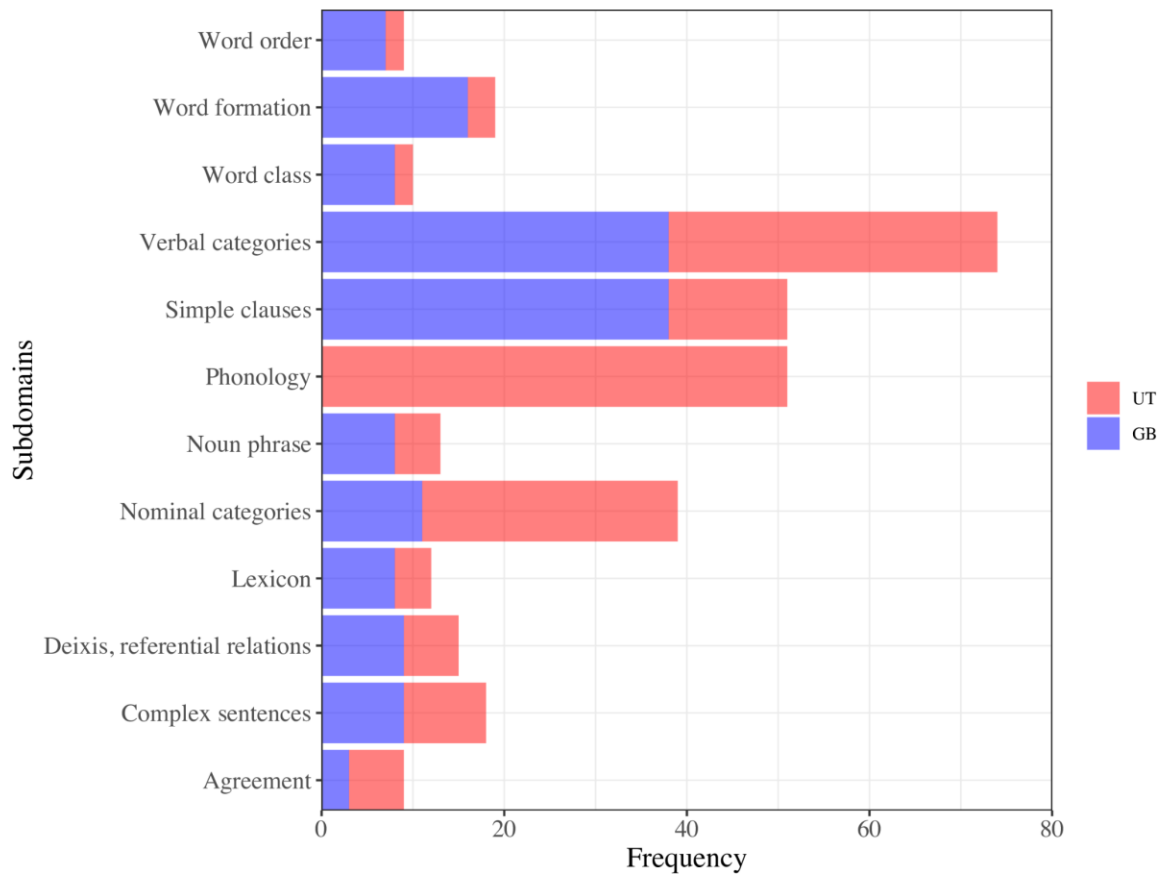


Figure 1. Distribution of features in the UraTyp data divided into different typological subdomains and colored by the division into GB and UT features.

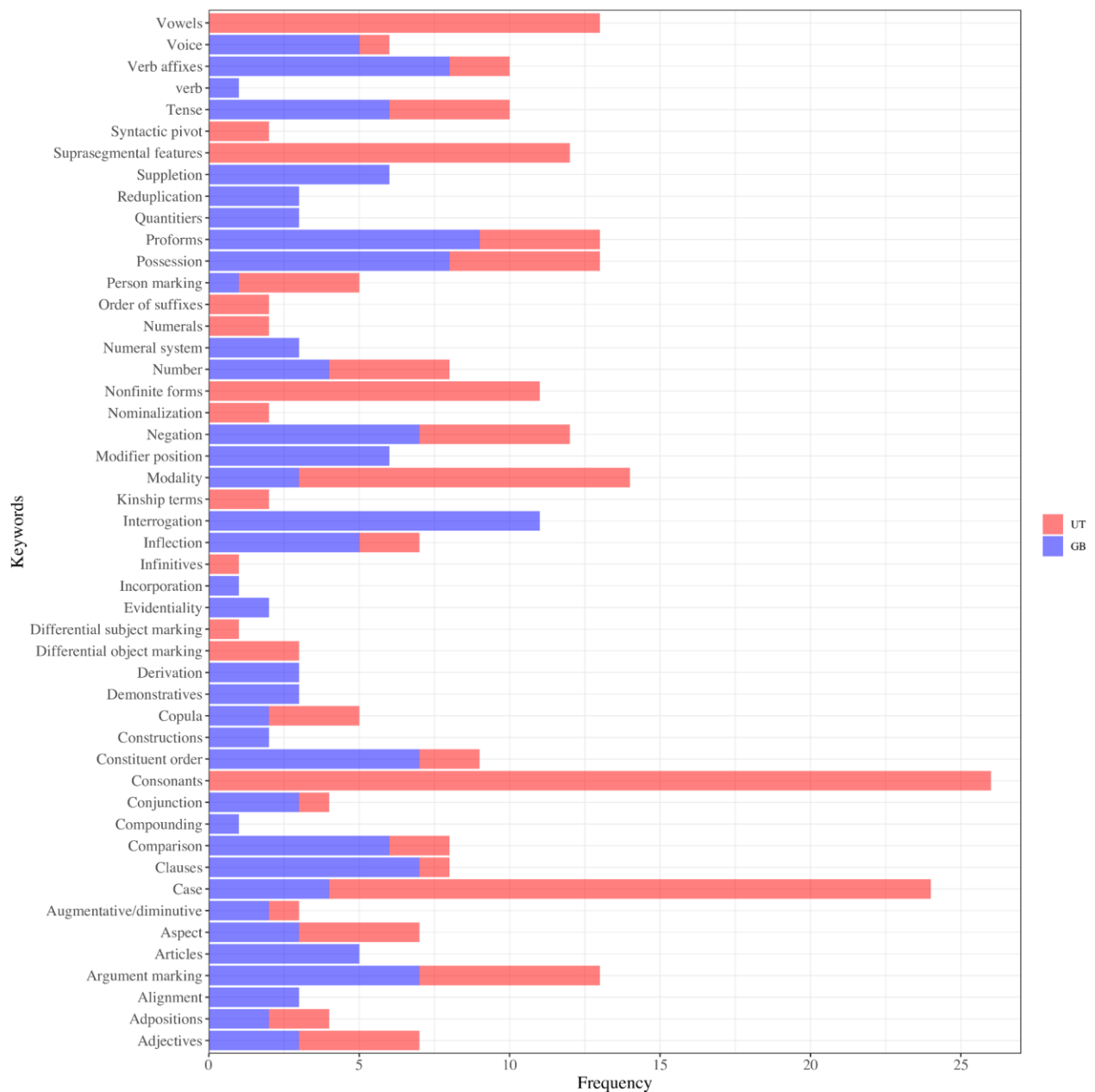


Figure 2. Distribution of features in the UraTyp data divided into different keywords and colored by the division into GB and UT features.

As Figures 1 and 2 reveal, some domains are covered only or mostly by the UT questionnaire, some only or mostly by GB questionnaire. For instance, only the UT questionnaire includes questions on phonology, non-finite forms, differential object marking, and order of suffixes, but only the GB questionnaire contains questions on interrogation, alignment, and various types of word formation. As a result, the two parts complement each other although they can be used on their own. This is licensed by the fact that on several occasions the questions covered by the GB

part are of a more general type, whereas the UT questions are more specific. To exemplify, in the GB questionnaire there are questions that ask whether there are morphological cases for core arguments and oblique NPs, whereas the UT list allows for a more detailed approach, e.g., *UT086 Is there a locative case that marks goal?*, *UT094 Is there a separate case for marking accompaniment, which is different from the instrumental?*, etc.

2.3. Data availability

At the time of submitting this paper, there are 29 Uralic languages / language varieties coded with the GB questionnaire and 33 with the UT questionnaire. The total number of languages is 35, covering all the branches of the Uralic language family. For a majority of languages, we have both GB and UT parts; one or the other part is missing only for a few languages. Komi-Permyak and Ume Saami are only found in GB, whereas UT includes 6 languages (Inari Saami, North Saami, Kazym Khanty, North Mansi, Tundra Nenets, Udmurt) that are not currently covered in GB data (see also Figure 3 in Section 3).

In conformance with the FAIR principles for scientific data management (Wilkinson et al. 2016), the Uralic Typological Database is published under a Creative Commons Attribution license as a CLDF dataset (Forkel et al. 2018) in the Zenodo repository. The Zenodo platform ensures that the data will be findable and accessible in the long term, while the CLDF specification provides a framework for interoperability. Choosing a CC-BY license makes sure that data reuse is possible. Anticipating future changes of the data – e.g., to adapt to new releases of the Glottolog language catalog (Hammarström et al. 2021) – the data is curated in a version-controlled repository on GitHub at <https://github.com/cldf-datasets/uratyp>, using the CLDFBench toolkit (Forkel & List 2020).

The Uralic typological database UraTyp includes the data collected using the UT and GB questionnaires. While the dataset is easily accessible for automated reuse from Zenodo (Norvik et al. 2021), we also implemented a web application based on the clld toolkit (Forkel et al. 2020) at <https://uralic.clld.org>, which allows for browser based, interactive exploration of the dataset. A detailed description of the data, including the list of features and their descriptions, the list of language experts and their contributions, and the list of sources can be found in Norvik et al. (2021).

3. Statistical analyses of the UraTyp data

In order to explore the data and visualize its typological diversity, we analyzed the data in two different ways. First, we made a principal component analysis (PCA) to study (dis)similarities between Uralic subgroups and further explored which typological features make the distinction between particular subfamilies or languages. Second, we clustered the UT data using a model-based admixture analysis to demonstrate how this typological data can be used to make historical inferences. With PCA we studied the complete UraTyp dataset, and also ran separate analyses on GB and UT datasets. The separate analyses were carried out for two reasons: First, the joint UraTyp database does not currently contain the GB and UT data for all the languages (see Figure 3 below), and by separate analyses we were able to visualize all the language relationships. Second, the global GB data to be published in Skirgård et al. (submitted ms) will allow for family cross-comparisons. However, it is unclear how well the GB traits cover the diversity within a given language family. Our comparisons between Uralic GB and UT data as well as GB and the whole UraTyp will serve as background information for the future comparisons between typological diversity within Uralic and that of other language families. We also ran separate PCAs within the diverse typological domains of the UT (phonological, morpho-lexical and syntactic) and GB (morphological and syntactic) datasets to see how these different typological subdomains perform in describing the internal variation within the Uralic family. With these analyses we aim at answering the following research questions: (1) how the different datasets perform in dispersing the conventional branches of the Uralic language, (2) to what extent the groupings differ if we consider the main levels of language structure (phonology, morphology/lexicon, syntax) separately. Below, we explain the methods and main results of the analyses. All the statistical code for the analyses is provided in Appendix 1.

3.1. Overview of the variation in UraTyp data

The UT data covers a total of 165 typological features (questions) across different domains: phonology (51 features), morphology/lexicon (61 features) and syntax (53 features) collected from 33 languages. The 195 GB features that are characterized as morphosyntactic could also be divided between morphology/lexicon (98 features) or syntax (97 features); the data comes from 29 languages. The coverage of the data is high with only few unanswered or unclear entries in the data (Figure 3). The high quality was due to exploiting the knowledge of language experts to supplement the data available in the published literature. Figure 3 shows the variation of the binary answers for the languages.

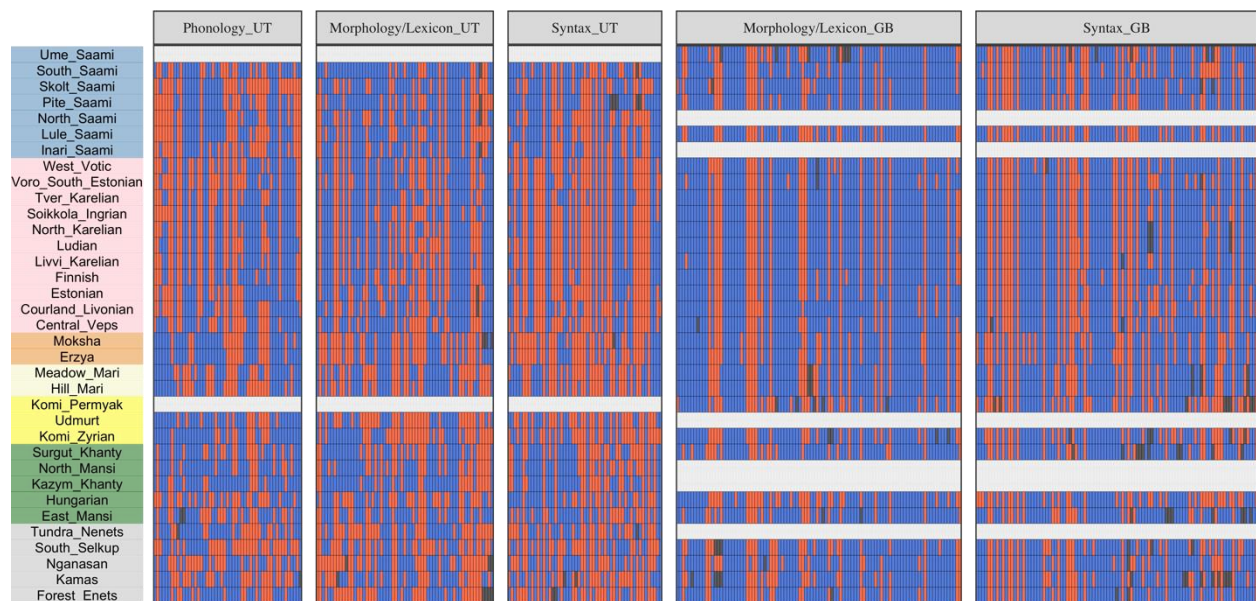


Figure 3. Overview of typological features in the UraTyp database (UT and GB). Each column represents one question, i.e. a linguistic feature. The colored cells indicate the presence (red) or absence (blue) of certain features (answered as 1 or 0 in the data, respectively), and gray cells indicate that these features are uncertain or unknown (indicated as “?” in the data). White rows in the plot represent missing data, i.e. unmatched languages between UT and GB datasets. Vertical cells of the same color indicate invariant features (features which are present or absent in all languages – common in the GB features, but by design unusual in the UT features). The y axis labels are colored by language subfamilies.

Among the 165 questions that form the UT data, there are only 3 invariant features: all Uralic languages need a copula for predicate nominals in the past and/or future tense (UT046), they distinguish between animate and inanimate objects in interrogative pronouns (UT106), and S and A can be morphologically conflated across clause boundaries (UT014). In comparison, the number

of invariant features among the answers collected with the GB questionnaire for Uralic languages is higher, since the GB questionnaire was designed to capture world typological variation rather than variation within a particular family. Out of 195 features in GB 64 were invariant (10 all 1s; 54 all 0s in Figure 3). Some features that get 1 ‘yes’ for all the Uralic languages included in GB are the following: presence of morphological cases for oblique non-pronominal NPs (GB02) and oblique independent pronouns (GB073), occurrence of verbal affixes or clitics that turn intransitive verbs into transitive ones (GB113), the use of postpositions (GB075), a decimal numeral system (GB333). As regards instances where the values 1 and 0 stand for a particular ordering, all Uralic languages behave in the same way. For instance, in ordering the numeral and the noun in the NP: the numeral precedes the noun (GB024) or the adnominal possessor noun and possessed noun in a pragmatically unmarked clause: the adnominal possessor precedes the possessed noun (GB065). As shown in Figure 3 (gray cells), there are only very few occasions where the information is uncertain or non-existing in the UT (0.6%) and GB (2%) datasets.

3.2. Clustering UraTyp and its subsets with PCA

For an initial exploration of the typological relationships of these languages we used **Principal Component Analysis** (PCA). PCA is used for reducing the dimensionality of high-dimensional data in order to summarize its major parameters of variation. It is used for example in human genetic studies to squeeze information from thousands of genetic components (e.g., SNPs) into a smaller number of dimensions which summarize the most significant variation in the data, and which can be plotted against each other to indicate how (dis)similar human populations are. In the UraTyp data, each feature can be considered as an independent axis of variation. PCA rotates the frames of reference within this multidimensional space to create new axes – mathematically equivalent in that they preserve the relative positions of all the data points – but which are selected so that the points vary the most of the first axis, the next most of the second axis, and so forth. The amount of variance “explained” by each axis is measurable, and axes which explain too little variance (e.g. under a conventional threshold of 10% or 5% of the total variance in the dataset) can be ignored as irrelevant or uninformative. In our analyses reported below the first two dimensions explain 30% or more of the variance, and the third dimension explains less than 10%, so only the

first two dimensions (PC1 and PC2) are reported in the paper. The amount of variance (around 9%) captured by the third dimensions (PC3) can be found in Appendix 1.

We have run PCA analyses of the joint UraTyp data and separately for all the UT features and GB features. The results of these analyses are presented in the first row of Figure 4. As the UT and GB languages did not completely match, we had 33 languages in UT analyses, 29 languages in GB analyses and 27 languages in UraTyp analyses. We have excluded those features that are completely invariant or contain some missing values in certain languages in our PCA analysis. Specifically, 23 out of 165 features in UT data and 120 out of 195 features in GB data are removed in the PCA analysis.

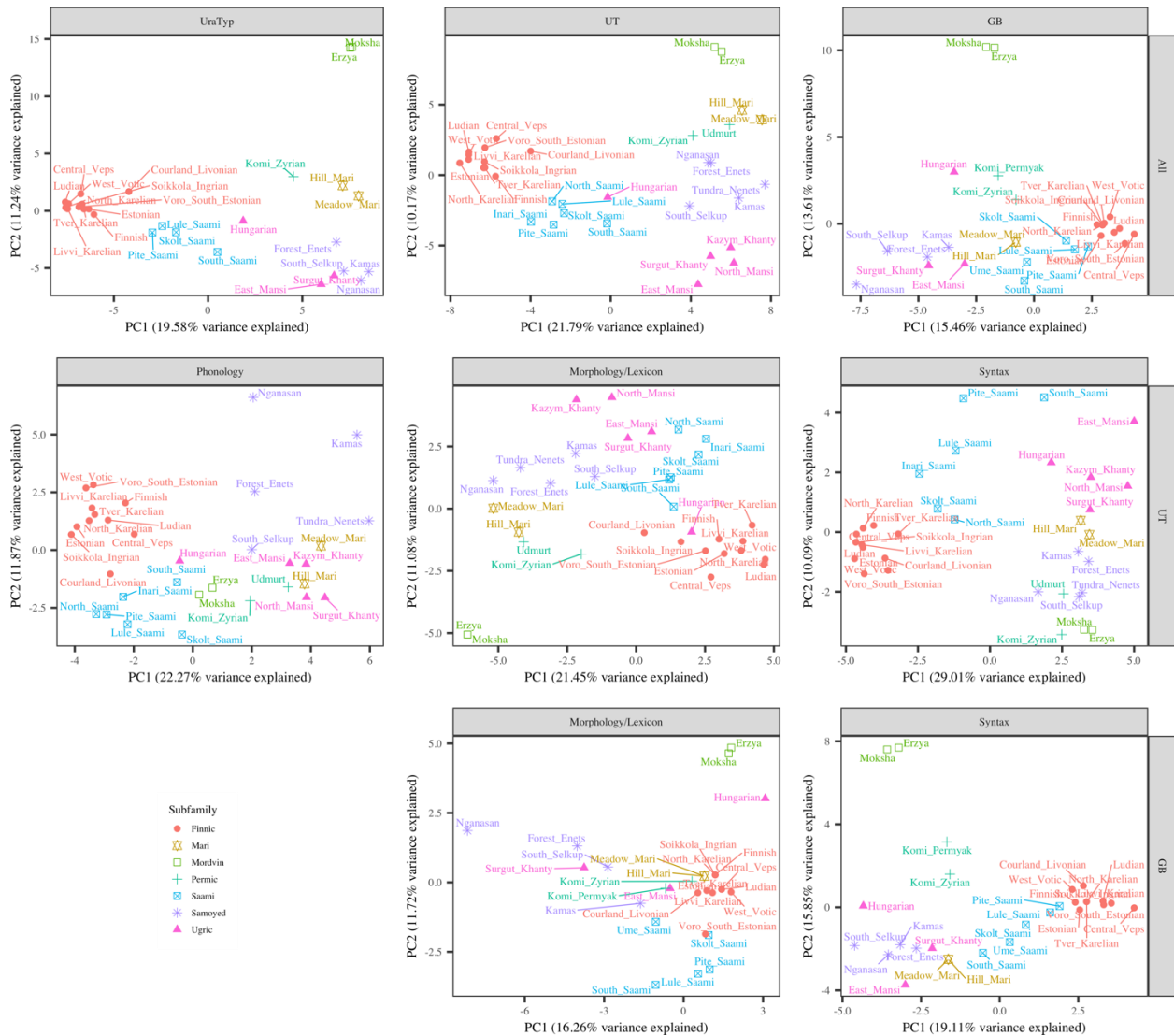


Figure 4. Principal Component Analysis of all Uralic typological data (UraTyp): The panels show (left to right) all UraTyp data, all UT data, all GB data in the top row, and then subsets of the UT data filtered by

typological domains: phonology, morphology/lexicon, and syntax. The GB data partitioned by morpho-lexical and syntactic domains are shown in the bottom row. Each language is projected onto the two-dimensional space (PC1 vs. PC2) and colored by its language subfamily. The number of features in each analysis (N): UraTyp (N=214), UT (N=140), GB (N=74), Phonology in UT (N=47), Morphology/Lexicon in UT (N=52), Syntax in UT (N=42), Morphology/Lexicon in GB (N=32) and Syntax in GB (N=42).

At first glance, our results show a good distinction between the Uralic subgroups, especially in the subplot of the UT dataset. In all cases except for GB features alone, the first principal component (PC1) accounted for 20–22% of the variance, the second 10–13% and the third 9–11%. The first panel shows the pooled UraTyp data, and could be considered as the best indication of Uralic typological variation. However, the downside is that it only includes the languages that overlap in the two datasets; thus only 1 Permic and 4 Saami languages are currently available. In general, UT performed the best: It disperses the subfamilies sensibly and clearly away from each other, unlike GB, where most languages are closely located in PC2. UT also locates Ob-Ugric languages separately from Samoyedic, whereas neither GB nor UraTyp finds differences between them. Saami languages cluster near each other, but only UT and UraTyp place them further away from the Mari and Permic languages. In UT, GB, and UraTyp, Mordvin languages form a distinct group, as Mordvin shows obvious differences from other subfamilies in PC2. In UT, the Morphology/Lexicon domain distinguishes the Mordvin languages, while in GB, the Mordvin languages appear special due to Morphology/Lexicon as well as Syntax. In all three cases (UT, GB, UraTyp), the Finnic languages formed a distinct cluster and Hungarian was positioned separately from the other Ugric languages with which Hungarian is traditionally grouped.

We divided the UT data into domains, and studied if the phonological, morpho-lexical or syntactic features would cluster the languages differently. The middle panel in Figure 4 reveals an intriguing difference in subgroup diversity across domains. All the domains locate Finnic languages together, but morpho-lexical features place Hungarian within the cloud of Finnic languages. The Saami languages constitute a coherent group when considering only phonological or morpho-lexical features (Morphology/Lexicon_UT), while syntactic features move the Pite and South Saami to the margin. All domains locate Samoyedic languages relatively close to each other, but the languages seem to vary a lot in their phonological features creating a large cloud. Mordvin languages are always near to each other, as are Mari languages, too. Interestingly, morpho-lexical features make Mordvin languages to deviate from the other Uralic languages. Phonologically, Mari languages appear within the cloud of Ob-Ugric languages, syntactically they are neighbors, but

morpho-lexical features take them apart. Finally, only syntactic features place Hungarian to the Ugric cluster.

As the GB features do not cover phonology, they were divided between Morphology/Lexicon and Syntax to make them comparable to the UT domains. In broad terms, the syntactic features' subset of the GB data shows an eastern (Samoyedic, Ob-Ugric, Permic, Mari) and a western (Finnic, Saami) cluster, whereas Mordvin appears the most different from the other languages. The morpho-lexical features (Morphology/Lexicon_GB) suggest a single large cluster, with Hungarian and Mordvin as the outliers.

3.3. What distinguishes Uralic subfamilies?

We further studied which typological features are responsible for the clustering of the languages seen above. We did this by calculating mutual differences between two subfamilies. More precisely, for each typological feature, we calculated the degree of dissimilarity for all pairs of languages between two subfamilies. For example, when looking at the contrast between short and long vowels in unstressed syllables between Finnic (11 languages) and Mari (2 languages), we can have 22 mutual comparisons of the feature values in a language and aggregate them into a single score of differences. In this case, Mari languages do not contrast between long and short unstressed vowels, whereas 6 out of 11 Finnic languages have this feature. This means 12 mutual comparisons are different, yielding a difference score of 0.55 (12/22). We then plotted these scores in Figure 5 to inspect the overall performance of each typological feature in discriminating between two subfamilies in each domain.

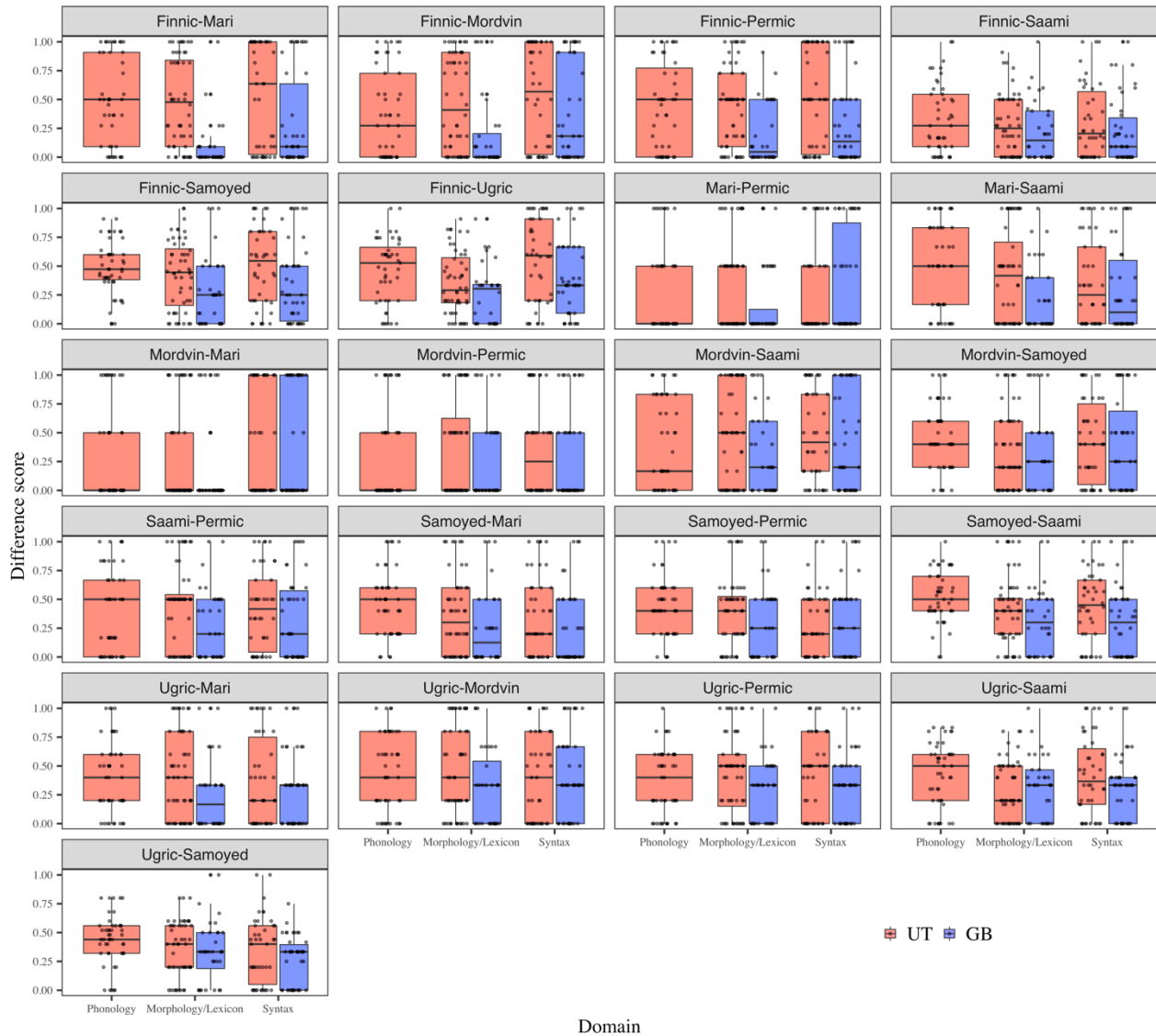


Figure 5. Box plots of difference scores for each typological feature between language subfamilies. Each data point represents the aggregated score of differences for each typological feature between two subfamilies. A higher score (close to 1) indicates that this feature is more distinctive between two subfamilies, while a lower score (close to 0) indicates that this feature is more similar between two subfamilies. The colored bars show the first and third quartiles of the data (25th and 75th percentiles), and the horizontal line within them shows the median. Points are jittered horizontally for legibility. In each figure the leftmost box-whiskers plot represents the phonological traits (UT data only), 2 middle plots represent the morpho-lexical features and the rightmost plots represent the syntactic traits.

We can make some broader generalizations about these results: when we compare Finnic and most other subfamilies (Mari, Mordvin, Permic, Samoyed, and Ugric), syntactic features in UT tend to be more distinguishable than phonological and morpho-lexical domains. We further studied which features exactly were the ones making the differences (Table 1 and Tables A1–A5 in Appendix 1). For example, Finnic and Ugric differ considerably in syntax, such as differential

subject marking and adnominal word agreement, whereas only two phonological features (rising diphthongs and palatal nasal [ɲ]) show contrastive values (Table 1). Phonological features in UT can also achieve good performance in differentiating between Saami and other branches like Finnic, Mari, and Permic, or between Samoyed and other groups. For the Samoyed-Saami comparison of the same Phonology_UT features the scores are mostly higher than 0.5 (although less than 1.0), meaning that there is a tendency for pairs of Samoyed and Saami languages to have opposite scores for each feature. Morpho-lexical features in UT only show slightly larger differences between Mordvin and Saami languages than syntactic features do.

Table 1. Most distinctive features between Finnic and Ugric in UraTyp. The last two columns show the presence (1s) or absence (0s) of certain features. The value 1/0 shows presence/absence in the entire group. More comparisons between Finnic and other subfamilies can be found in Appendix 1, Tables A1–A5.

ID	Domain	Definition	Finnic	Ugric
UT004	Syntax	Can an adnominal property word agree with the noun in case?	1	0
UT011	Syntax	Does object marking correlate with aspectuality?	1	0
UT012	Syntax	Does the marking of the object depend on the verb conjugation form?	1	0
UT013	Syntax	Is there differential subject marking?	1	0
UT049	Syntax	Does existential negation have a separate operator?	0	1
UT050	Syntax	Can standard negation be asymmetric in the present tense?	1	0
UT061	Syntax	Can subject in at least some participial subordinate clauses be encoded in the same way as in independent clauses?	0	1
UT079	Syntax	Does the numeral modification of a noun have an effect on its case selection?	1	0
UT102	Syntax	Can the possessor in the attributive construction be marked with the nominative?	0	1
UT131	Phonology	Are there rising diphthongs?	1	0

UT149	Phonology	Is there a palatal nasal [ɲ]?	0	1
GB074	Syntax	Are there prepositions?	1	0
GB132	Syntax	Is a pragmatically unmarked constituent order verb-medial for transitive clauses?	1	0
GB133	Syntax	Is a pragmatically unmarked constituent order verb-final for transitive clauses?	0	1
GB184	Syntax	Can an adnominal property word agree with the noun in number?	1	0

As already illustrated in Figure 4, GB features in general are less informative in distinguishing between subgroups, since they are more similar across all Uralic languages. This is especially evident in Figure 5 where GB features tend to have lower difference scores across-the-board than UT features. In particular, in the case of Finnic-Mari comparison, where the points in GB data are mostly at or around the zero line, meaning that the languages from both these subfamilies have the same values (i.e., a difference of 0). The comparison between Finnic and Mari languages, in turn, tend to be different in UT data (median above 0.5), meaning that for more than half of the features the language pairs mostly have opposite scores. See also Tables A1-A5 in Appendix 1 for the most distinctive features across different domains between subfamilies.

3.4. Diachronic patterns of typological admixture

The PCA clusters the data by current similarity. For a more historical perspective we further clustered the UT and UraTyp data with **model-based admixture analyses**. Admixture analysis is commonly used in population genetics to infer the population structures and detect the recent admixture histories between populations (Pritchard et al. 2000; Alexander et al. 2009). Admixture models work by inferring the clusters which existed in the past, and showing how these ancestry components are reflected in the present variation; which ancestral entities are the building blocks of the contemporary typological variation. Admixture analyses are an alternative for achieving historical inference from typological data: As typological features typically reflect a restricted design space, the similarities between languages may be not only inherited or may be just due to a limited number of alternatives e.g. in word order. However, with carefully selected typological

features also genealogical analysis has been conducted (Ceolin et al. 2020). Phylolinguistic tree-building algorithms better capture the genealogical relationships of cognate data such as basic vocabulary with cognate or correlate coding, where the probability of chance similarity is low (Greenhill et al. 2020). Tree-building algorithms are currently built to measure vertical evolution, whereas admixture models are aimed at studying horizontal transmission of material. Thus, they fit especially well to linguistic typological data.

Admixture models have been applied in linguistics to study language families (Reesink et al. 2009) and dialects (Syrjänen et al. 2016; Honkola et al. 2018; 2019). The application of model-based clustering algorithms to linguistic data is discussed in depth elsewhere (Syrjänen et al. 2016; Syrjänen 2021). Here we carried out an admixture analysis using the typological variation of each language as input. These analyses aim at clustering the languages into subgroups without prior information of their relatedness. Besides clustering the languages, we also studied which number of ancestral clusters (K) best fits the data, and further studied the ancestry components of each of the attested languages.

A general assumption in model-based clustering analyses is “linkage disequilibrium”, i.e., the characters (e.g., SNPs in genetic data or typological features here) must show independent variation. While there are multiple methods to address this with biological data, for language data there is no established method to measure the independence of the characters (but see Syrjänen et al. 2016; Syrjänen 2021 for dialect data). In both biological and linguistic data some dependencies always exist. An important difference between biological and linguistic data is that with genetic data the researcher does not know a priori how the genetic components are linked – the features are given by nature. The linguistic features, on the other hand, are designed by the experimenters and it is thus possible to influence the amount of dependencies existing in the data. For both the GB and UT data lots of effort was spent on selecting features that are logically independent, and we believe that this makes them suitable for model-based cluster analysis. However, further development of a formal methodology for measuring the independence of linguistic data would be urgently needed.

The second assumption in model-based clustering analyses is the “Hardy-Weinberg equilibrium”, which states that there is a random mating pattern between the sampled individuals (discussed for languages in Syrjänen et al. 2016 and Syrjänen 2021). In our case this would mean that each language should be equally likely to be in contact with each of the other languages. This

probably does not hold true, since geographically neighboring languages are more in contact than expected by random chance (something which is very much the case for genetic populations, too). Here we do not study the Hardy-Weinberg assumption in more detail, but consider the areal effect in the interpretation of results.

As a model-based clustering algorithm we used the sNMF algorithm implemented in the LEA package in R (Frichot & François 2015). The inference algorithm used in sNMF (sparse non-negative matrix factorization) is based on a fast version of STRUCTURE, which is widely used in population genetic clustering analyses. Comparison between STRUCTURE and K-medoids is found in Syrjänen et al. (2016), and comparison to BAPS is found in Honkola et al. (2019). The sNMF produces results which give estimates of ancestry coefficients that are similar to STRUCTURE (François 2016); it also includes a convenient method for estimating the best fitting number of ancestral clusters (K) based on cross-validation and cross-entropy.

We have run 2000 iterations for the sNMF algorithm for K-values 1–7. To determine the best fitting K value, we then assessed the predictive value of each K through 10-fold cross-validation. It turned out that K=4 gave the best predictive accuracy in terms of minimizing cross-entropy (Figure 6B). We estimate uncertainty of the K clusters and ancestry proportions with 50 repetitions of the admixture analysis (see Figure A9 for the consistency across 50 replicates in Appendix 1).

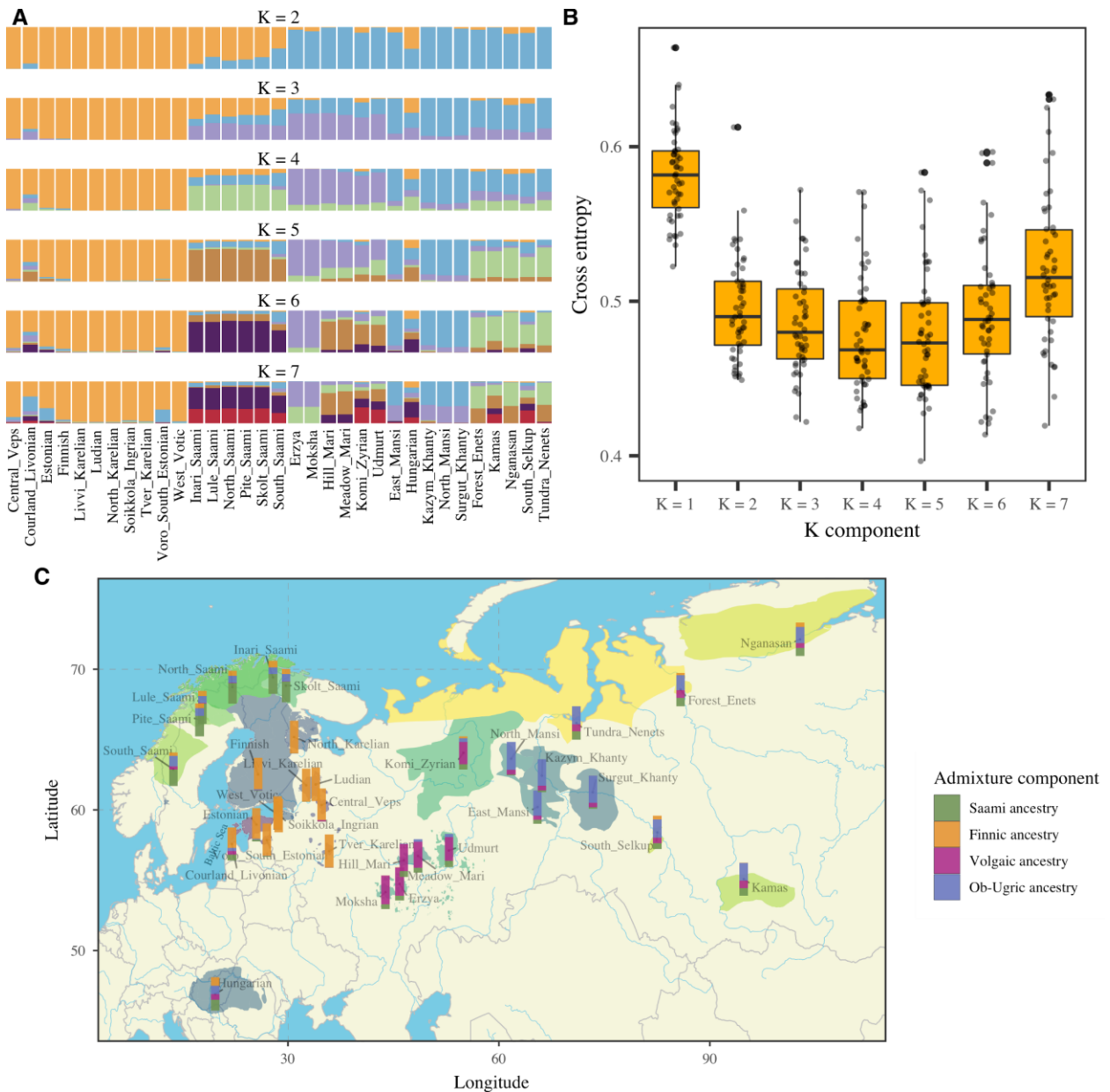


Figure 6. Inferred number of components (K) and ancestral coefficients in the admixture analysis of UT data. The bar charts (subplot A) provide the estimated ancestral proportions for each K component, which are averaged across 50 repetitions of the analyses. The corresponding predictive accuracy via cross-entropy for each K value is shown in subplot B. A lower cross-entropy value means better predictive performance in cross-validation. To examine the geographical distribution of each language and its ancestry, we draw the ancestral coefficients on the map (subplot C) by using the admixture results of $K=4$. The colored areas on the map represent the language speaker areas from Rantanen et al. (2021).

Under the preferred number of ancestral populations, $K=4$, the contemporary languages can be shown to be made up of ancestral populations with clear geographical separation (Figure 6 A & C). These inferred ancestry components could be called Saami ancestry, Finnic ancestry, Volga area ancestry and Ob-Ugric ancestry, and in most of the languages in the sample only one of these

components predominates. Hungarian is an exception, with no clearly dominant ancestry component. The Samoyedic languages also have a mixture of all components, but with a larger contribution from the Ob-Ugric component than the others. The clusters capture well the Uralic subfamilies (Saami, Finnic, Volgaic [incl. Mordvin, Mari and Permic] and Ob-Ugric [and Samoyed]) but at the same time, the clusters are likely to reflect areas of historical interactions. Samoyed did not have an ancestral population of its own when dividing the data into four. However, when dividing it to five, the Samoyed ancestral component appeared (see Figure 6 A).

4. Discussion

We present here a new comprehensive database of linguistic typological features from the Uralic languages. The database design is derived from two sources. The UT data (Uralic-specific typology) include typological features specifically selected to differentiate between the Uralic languages. We show here that the selection of features was successful: As a sanity check we clustered the data with two different methods, and indeed the known subfamilies (as identified also by historical linguistics) can be located only by using this part of the UraTyp dataset. Furthermore, we show that when the features are divided into different functional domains, then the domains carry different signals of language relationships. However, UT does not include all the relevant structural features, and it is built to be used alongside the GB features. The GB data alone clusters the subfamilies in mostly sensible ways, but does not identify so many differences between them as the UT data does. Note that the analyses here were conducted on a subset of GB features, since GB includes many features that are invariant for Uralic languages. It remains, however, important to record all the 195 characters of GB, as they serve an important function in allowing for comparisons of the Uralic family to the rest of the languages of the world. The location of Uralic languages in the typological space of the world's languages will be the focus of future studies.

In general, the analyses show that a combination of both datasets is needed to cover the variation in Uralic typology, and that the different domains reveal different aspects of linguistic relationships between the languages. The PCA analyses carried out using only the GB data showed that the GB data does not perform as well in distinguishing the subfamilies as the UT data does. The complete set of GB data grouped together all the languages except Mordvin. The position of

Hungarian was also somewhat outlying. To compare, all the subdomains within the UT data place Hungarian within the cluster of other languages, and all subdomains except for phonology place both the Mordvin languages, Moksha and Erzya, as a cluster on the periphery. This pattern is preserved when all the data (UT + GB) is combined into the All_UraTyp dataset: the features that place Hungarian on the periphery in the GB data are not enough to overwhelm the evidence of the Uralic-specific features (although Hungarian is placed centrally on the chart, separately from the other Ugric languages).

While principal component analyses visualize the synchronic diversity of Uralic languages, admixture models are able to give insights into the diachronic processes of language family evolution reflecting a joint effect of phylogenetic and areal relatedness. The admixture model suggested that the diversity in the UT data is best described with four ancestral groups. The clusters identified by admixture models are considered as ancestry components in genetic data, and we follow the same logic in interpreting the typological results. One of the ancestry components is maximized in West Votic and is found in all Finnic languages. This “Finnic” ancestry is very stable and exists through all the K-values (Figure 6A). In K=4 Saami languages form another ancestry component. Contrary to tree-models of the family (e.g., Syrjänen et al. 2016), admixture analyses suggest that structurally, Mordvin, Mari and Permic languages share a single component that could be called “Volgaic ancestry”, probably reflecting linguistic interaction in the area. The last group has similarities maximizing in Mansi and Khanty, thus this could be called “Ob-Ugric ancestry”. Samoyedic languages best fit to this group suggesting another contact zone to the east of the Ural Mountains. Unsurprisingly, given its complex contact history and geographical separation from the rest of the family, Hungarian shows a mix of components, indicating that it cannot really be fit to any of these groups. The admixture analyses for the large typological data thus complement the earlier tree-models that model the vertical evolution of the language family. The admixture analyses instead model the interactions between languages. The results suggest a communication network within the languages in the north of the Volga area and in the east of the Ural Mountains.

Although the best fitting number of clusters was four, the admixture analysis indicates the first division of the Uralic family into western (Finnic, Saami) and central-eastern (incl. Mordvin, Mari, Permic, Ugric, Samoyed) groups (see Figure 6a, K=2). Most of the PCA plots also cluster Ob-Ugric and Samoyed languages to the same space and keep Finnic and Saami languages near to each other. Thus, both the admixture analyses and principal component analyses hint at typological

similarities between Ob-Ugric and Samoyedic languages – this could also be seen as some common linguistic structure occurring with the easternmost Uralic languages. The division into western and eastern languages was also found in phonological study by Pajusalu et al. (2018). In this data, the clearest east-west division is seen in the All_UraTyp plot, where it is likely to stem from All_UT data and further from Phonology_UT data, which is ultimately based on the data in Pajusalu et al. (2018). Nevertheless, while the GB data does not find clear subgroups, it is more in line with the broader understanding of western Uralic (Grünthal 2019; Ylikoski 2016).

As regards PCA analyses performed on the UT and UraTyp data, the Saami and Finnic clusters turn out to be quite uniform in terms of containing languages of one subfamily. However, there is still one language that stands out in both groups – South Saami and Livonian, respectively. Both of these are the southernmost languages in their respective groups. South Saami appears in a cluster with other Saami languages in the overall analysis, but stands somewhat more out in phonology and syntax; as regards syntax, it appears at the end of the continuum (see Figure 4). With respect to phonology, the language was coded based on the phonological analysis by Kowalik (forthcoming) which focuses on phonological contrasts and which differs from other/previous descriptions. As suggested by the admixture analysis, South Saami also shows a component of “Volgaic ancestry”. This does not come as a surprise as the language has several (well-known) features that are shared with languages in the Volgaic area: a genitive predicative possession construction, the basic word order SOV, or morphologically distinct case endings for the accusative and genitive. The deviation of South Saami from other Saami languages could also appear through different contact languages. Based on Germanic loanwords Southern Proto Saami – the predecessor of South Saami – was the first one to drift off from Common Proto Saami at an early point of time (Piha & Häkkinen 2020). Piha and Häkkinen (2020: 119) support the hypothesis that Southern Proto Saami would have arrived to the southern parts of Norrland, Sweden slightly earlier (100–200 years at the most) than other current Saami languages expanded to their current locations around 300–400 CE (about other Saami languages reaching northern Fennoscandia, see Heikkilä 2011: 76; Aikio 2012: 79; Magga 2014: 43). The more southern and earlier spread to the Swedish side of the Baltic Sea could have even allowed for contacts with unknown languages other than the ones the other Saami languages met in the current Finnish area and in Lappland (Aikio 2012; see also Piha 2018: 172–175).

Courland Livonian, in its turn, besides common Finnic typology, shares several similarities with the Saami and Mordvin languages. Some of these features are typical also in other Southern Finnic languages such as the use of a demonstrative stem in the function of the 3rd person pronoun or the lack of vowel harmony. Livonian has a number of fusional traits which are similar to Saami languages. These similarities, however, do not indicate direct contacts between these languages but are more likely to be a result of a stronger influence from neighboring Indo-European languages than what northern and eastern Finnic languages have had (cf. Grünthal 2015). On the other hand, like Saami and Mordvin languages, Livonian lacks several innovations that have spread in most Finnic languages, e.g., a grammatical device for immediate future (UT020) or the loss of affricates (UT152, UT153).

Central Uralic, i.e., Mordvin, Mari and Permic languages cluster together into a Volgaic group in our admixture analysis. These historically divergent languages share a series of innovations, such as a dative case, case compounding, variable order of possessive suffix and case suffix across the paradigm, a special negator that combines with non-finite verb forms, etc. The majority of these innovations are promoted by the long-lasting influence of Turkic languages (Bereczki 1984: 307–314; Csúcs 1990; Isanbaev 1994; Róna-Tas 1988: 760–774; Saarinen 1997: 388–396). Also, there has been a strong impact of Proto-Permic on Mari (cf. Bereczki 1977: 57–76). At the same time, they have preserved some Proto-Uralic phonological features, for example, presence of unrounded vowels (UT137 or UT138) and lack of contrastive length of vowels and consonants. Still there are significant differences between these languages caused by various (socio)linguistic factors. Our finding reflects a hypothesis of a Volga–Kama Sprachbund (i.e., an area of linguistic convergence) that has been suggested because of the many similarities found in the Uralic (Mari, Mordvin, Udmurt) and Turkic (Chuvash, Tatar, Bashkir) languages spoken in the region. These languages are known to originally resemble each other typologically and as they have a long history of mutual influence, the similarities between them are easily explained by borrowing and copying (see Johanson 2000). Close contacts between the Uralic and Turkic speaker populations are also reflected by their overall genetic similarity.

Hungarian is an interesting case. Although historically Hungarian is most closely related to the Ob-Ugric subfamily, i.e., Mansi and Khanty languages (Abondolo 1998; Hajdú 1952; Honti 1979; 1997), it separated from them in the early phase of evolution of the Uralic family. Some authors however suggest that the apparent Hungarian–Ob-Ugric unity is an artefact caused by areal

contacts between early phases of Hungarian and Ob-Ugric languages (Aikio 2018; Gulya 1977; Helimski 2003). The PCA plots show that while Hungarian is an outlier from the other Ugric languages in terms of phonology and morphology/lexicon, in terms of syntax Hungarian in fact clusters with the Ob-Ugric languages Khanty and Mansi (UT data), or forms a wider cloud that also includes the Samoyed languages (GB data). This probably indicates that Hungarian really does cluster with Ob-Ugric when considering Uralic-specific typological traits, but that the GB features do not have the resolution to capture these relationships. The admixture analyses (Figure 6, but also Figure A2 in Appendix 1) suggest that Hungarian has properties similar to Ob-Ugric languages, but it fits as well to any other cluster. At higher K-values (K=4) its largest component is also the major component of Saami languages; this component is not shared with Ob-Ugric languages. The common features of Hungarian and Saami languages are listed in Table A6 (see Appendix 1). Their common features in phonology are mostly innovative from the Uralic point of view, such as extralong syllables (UT121), word-initial consonant clusters (UT141), and consonant clusters in syllable coda (UT143). In morphology and syntax, their shared features are more conservative. In any case, all other features except the syntactic features in UT data are consistent with the view that Hungarian would be synchronically a one-language branch. This idea is also diachronically plausible: Hungarian speakers departed from their closest linguistic relatives very early, moved through vast territories inhabited by peoples speaking many other languages, and settled in Pannonia together with several of their allies, who also spoke a number of different languages. As a consequence, they have preserved both eastern and western features and got all kinds of innovations from their neighboring peoples.

In several of our analyses, Samoyed languages grouped together with Ob-Ugric languages, forming an eastern cluster of the Uralic family (see All_UraTyp and All_GB in Figure 4, Figure 6). They form a distinctive group not only in the case of K=5 clustering (see Figure 6A) but also when considering UT features separately (see All_UT in Figure 4). One reason for the Samoyed languages not being reflected separately in K=4 clustering is a likely typological dispersion of southern and northern Samoyed languages in the different contact areas of Siberia. Still, regardless of the very early split of the Samoyed languages from the rest of the Uralic family, they preserve on average about 57% of ancestral characters with the Ob-Ugric subfamily. Another reason may be that some Samoyed features are shared with languages of the western periphery, e.g., radical consonant gradation is found in Saami, Finnic and Samoyed languages. In phonology, Samoyed

languages have retained some old “eastern” features that occurred in Proto-Uralic, such as a lack of consonant clusters in syllable coda and a constraint on word-initial *r*.

The first results of the study suggest that there could be differences caused by various areal effects that act upon pronunciation and grammar in different ways. For future study of structural diversity of the Uralic languages it will also be important to discuss the main contact languages in the area. As Nichols (2021) shows, Uralic languages, especially eastern Uralic languages, include features that can be regarded as characteristic of the Inner Asian type (e.g., head-final word and morpheme order, strongly configurational syntax), while western features in several cases include loss of object indexation, attrition of possessive person inflection. Nichols also claims that Proto-Uralic had a number of typological traits that are typical of the North Pacific Rim and not of western Eurasia. Inclusion of the main contact languages would be important also for gaining a better understanding of the position of Hungarian and other languages that stand out in the group (e.g. Courland Livonian, South Saami).

It is possible that the large-scale quantitative studies would benefit from further refinement of the database as regards the features and answers to them. For instance, currently syntactic features in general proved to be more informative for clustering Uralic languages than morpho-lexical or phonological ones. It is possible to add features without affecting the need to make changes in the already existing ones. The addition of morphological features might have an effect on better differentiating the Samoyed languages. Whereas Uralic languages have been described as less complex than northern Eurasian languages in general, the morphology of Samoyed languages is regarded to be among the most complex ones. This is seen as a result of post-Proto-Uralic and postbranch developments (see Nichols 2021: 364; Grünthal et al., in press). Morphology was less informative than syntax also in the case of GB data. Further refinement of the data also involves checking the features with missing values (see Section 2.3). The current set-up of the data allows for flexible corrections and additions to the data e.g. through crowd-sourcing at the source repository.

An issue inherent in all large-scale typological databases is that data points for different languages may be based on doculects that are of different character, e.g. differ in age or to what extent they reflect a specific dialect or a potential standard variety of a language. In the current project, doculects range from written material of extinct languages (e.g., Kamas) that was documented more than a century ago to doculects that reflect contemporary spoken language (e.g.,

Võro). A historical perspective on particular features, or data for different doculects of the same languoid, would be a welcome complement.

5. Conclusions and future perspectives

In typological studies, Uralic languages are often represented by Hungarian, Finnish, and sometimes also by Estonian. However, as both clustering analyses reveal, these languages do not perfectly represent the diversity of Uralic typology. This means that there is a need for the UraTyp database allowing for a more complete picture on Uralic typology. UraTyp is the first online platform where a large number of Uralic languages receive equal attention and where the different domains of language structures are well-covered. This also brings into the picture such Uralic languages that have hardly ever made it to typological databases.

The statistical analyses that were carried out to explore the data and demonstrate the possibilities of use of the database confirmed that it was worthwhile adding the Uralic-specific traits. Regarding our first research question on the performance of the data in dispersing the conventional branches of the Uralic languages, we found that, as regards the PCA analyses, the addition of UT features enabled us to disperse the subfamilies more sensibly and clearly away from each other than using only the GB data. The admixture analysis on the UT data was able to find four main clusters: the Finnic, Saami, Volga area (incl. Mordvin, Mari and Permic languages), and Ob-Ugric (and Samoyed), which reflect the main branches of the traditional Uralic tree but offer also indication of (secondary) areal contacts. The admixture analysis was not run on the joint UraTyp dataset as currently the languages collected with the GB and UT questionnaires do not overlap entirely.

As regards our second research question about which typological domain (phonology, lexicon/morphology, syntax) separates the groups from each other, the analyses revealed that different domains can be responsible for different groupings. For instance, whereas phonological and morpho-lexical features placed Hungarian within the cloud of Saami or Finnic languages, the syntactic features kept it in the Ugric cluster.

Even though the UraTyp database contains a vast amount of data on the Uralic languages, there are several ways to improve and expand it. First, it is possible to add new questions as well

as new languages / language varieties without a need to make changes in the general structure. Second, the GB questions could also be provided with examples as done with the UT part. Although such work is time-consuming, it would add a valuable qualitative dimension to the data. Third, a further possibility to expand the database is to add cognacy information. For instance, in the case of an answer 1 ‘yes’ (e.g. *UT081 Is there a genitive case that has an affix?*) it is possible to add information on its source. This would open up new perspectives for diachronic analyses and would allow, for example, a joint analysis of the typological and the lexical data. Fourth, coding the UT questions also for the contact languages would allow one to investigate language contacts more thoroughly. The data framework flexibly allows for updating the data. All these kinds of additions would bring us closer to a comprehensive understanding of the past of the Uralic family.

Acknowledgments

We are very thankful to the language experts for their time and expertise:

Niina Aasmäe, Marianne Bakró-Nagy, Mariann Bernhardt, Rogier Blokland, Jeremy Bradley, Josefina Budzisch, Márta Csepregi, Svetlana Edygarova, Ulla-Maija Forsberg, Nikolett F. Gulyás, Arja Hamari, Mervi de Heer, Heinike Heinsoo, Katri Hiovain-Asikainen, Csilla Horváth, Sulev Iva, Markus Juutinen, Olle Kejonen, Olesya Khanina, Nikolay Kuznetsov, Miika Lehtinen, Olga Melentjeva, Petter Morottaja, Ilja Moshnikov, Irina Nikolaeva, Irina Novak, Péter Pomozi, Alexandra Rodionova, Mária Sipos, Sven-Erik Soosaar, Elena Skribnik, Denys Teptiuk, Helen Türk, Jussi Ylikoski, Beáta Wagner Nagy, Joshua Wilbur, Nina Zaiceva.

Funding information

Kipot ja Kielet research project (2018–2020) led by Päivi Onkamo and Outi Vesakoski; several Kone Foundation projects: UraLEX (2014–2016) led by Unni-Päivä Leino; SumuraSyyni (2013–2016) and AikaSyyni (2017–2021) led by Outi Vesakoski: OV, TH and YJ were funded by AikaSyyni. Research was also supported by the Turku University Foundation (Terhi Honkola) as well as Collegium for Transdisciplinary Studies in Archeology, Genetics and Linguistics, University of Tartu (2018–), and partly by the Estonian Research Council (grants No. PUTJD926, PRG341, PRG1290).

References

- Abondolo, Daniel Mario (ed.). 1998. *The Uralic languages* (Routledge Language Family Descriptions). London, New York: Routledge.
- Aikio, Ante (Luobbal Sámmol Sámmol Ánte). 2012. An essay on Saami ethnolinguistic prehistory. In Riho Grünthal & Petri Kallio (eds.), *A linguistic map of prehistoric northern Europe* (MSFOu 266), 63–118. Helsinki: Finno-Ugrian Society.
- Aikio, Ante. 2018. Notes on the development of some consonant clusters in Hungarian. In Sampsa Holopainen & Janne Saarikivi (eds.), *Περὶ ὀρθότητος ἐτύμων – Uusiutuva uralilainen etymologia* (Studia Uralica Helsingiensia 11), 77–90. Helsinki: Finno-Ugrian Society.
- Alexander, David H. & John Novembre & Kenneth Lange. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* 19. 1655–1664. (<https://doi.org/10.1101/gr.094052.109>)
- Bereczki, Gábor. 1977. Permi-cseremisiz lexikális kölcsönzések [Permic–Mari lexical borrowings]. *Nyelvtudományi Közlemények* 79. 57–76.
- Bereczki, Gábor. 1984. Die Beziehungen zwischen den finnougriischen und türkischen Sprachen im Wolga–Kama-Gebiet. *Nyelvtudományi Közlemények* 86. 307–314.
- Ceolin, Andrea, Cristina Guardiano, Monica Alexandrina Irimia & Giuseppe Longobardi. 2020. Formal syntax and deep history. *Frontiers in Psychology* 11. (<https://doi.org/10.3389/fpsyg.2020.488871>)
- Csepregi, Márta & Katalin Gugán. The syntax of Khanty. (http://www.nytud.hu/oszt/elmnyelv/urali/publ/Khanty_Syntax_first_draft.pdf) Manuscript (Accessed 2021-12-21.)
- Csúcs, Sándor. 1990. *Die tatarischen Lehnwörter des Wotjakischen*. Budapest: Akadémiai Kiadó.
- Dahl, Östen & Viveka Velupillai. 2013. The past tense. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (<http://wals.info/chapter/66>) (Accessed 2021-04-04.)

- De Groot, Casper. 2017. *Uralic essive and the expression of impermanent state*. Amsterdam, Philadelphia: John Benjamins.
- Dediu, Dan & Stephen C. Levinson. 2012. Abstract profiles of structural stability point to universal tendencies, family-specific factors, and ancient connections between languages. In Alex Mesoudi (ed.), *PLoS ONE* 7(9). e45198. <https://doi.org/10.1371/journal.pone.0045198>.
- Dryer, Matthew S. & Martin Haspelmath (eds.). 2013. *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (<https://wals.info/>) (Accessed 2018-11-03)
- Forkel, Robert, Sebastian Bank, Christoph Rzymiski & Hans-Jörg Bibiko. 2020. cld/cld: cld – a toolkit for cross-linguistic databases (v7.2.0). Zenodo. (<https://doi.org/10.5281/zenodo.3968247>)
- Forkel, Robert & Johann-Mattis List. 2020. CLDFBench: Give your cross-linguistic data a lift. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck et al. (eds.), *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 6995–7002. Paris: European Language Resources Association (ELRA). (doi:10.17613/8t0e-w639)
- Forkel, Robert, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymiski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping & Russell D. Gray. 2018. Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data* 5(1). 180205. (<https://doi.org/10.1038/sdata.2018.205>)
- François, Olivier. 2016. Running structure-like population genetic analyses with R. R tutorials in population genetics, University of Grenoble-Alpes, 1–9.
- Frichot, Eric & Olivier François. 2015. LEA: An R package for landscape and ecological association studies. *Methods in Ecology and Evolution* 6(8). 925–929. (<https://doi.org/10.1111/2041-210X.12382>)
- Good, Jeff & Michael Cysouw. 2013. Languoid, doculect, and glossonym: Formalizing the notion “language”. *Language Documentation & Conservation* 7. (<http://scholarspace.manoa.hawaii.edu/handle/10125/4606>) (Accessed 2021-08-31.)
- Greenhill, Simon J., Q. D. Atkinson, A. Meade & Russell D. Gray. 2010. The shape and tempo of language evolution. *Proceedings of the Royal Society B: Biological Sciences* 277(1693). 2443–2450. (<https://doi.org/10.1098/rspb.2010.0051>)

- Greenhill, Simon J., Heggarty Paul & Russell D. Gray. 2020. Bayesian phylolinguistics. In R. D. Janda, B. D. Joseph & B. S. Vance (eds.), *The handbook of historical linguistics*, vol. 2, 226–253. New Jersey: Wiley-Blackwell.
- Grünthal, Riho. 2015. Livonian at the crossroads of language contacts. In Santeri Junttila (ed.), *Baltic languages and white nights* (Uralica Helsingiensia 7), 12–67. Helsinki: Suomalais-Ugrilainen Seura.
- Grünthal, Riho. 2019. Canonical and non-canonical patterns in the adpositional phrase in Western Uralic: Constraints on borrowing. *SUSA/JSFOu* 97. 9–34.
- Grünthal, Riho, Volker Heyd, Sampsa Holopainen, Juha Janhunen, Olesya Khanina, Matti Miestamo, Johanna Nichols, Janne Saarikivi & Kaius Sinnemäki. In press. Drastic demographic events triggered the Uralic spread. *Diachronica*. John Benjamins.
(<https://researchportal.helsinki.fi/en/publications/drastic-demographic-events-triggered-the-uralic-spread>) (Accessed 2021-08-31).
- Gulya, János. 1977. Megjegyzések az ugor őshaza és az ugor nyelvek szétválása kérdéséről [Comments on the issue of the separation of the Ugric homeland and the Ugric languages]. In Bartha, Antal et al. (eds.), *Magyar őstörténeti tanulmányok*. 115–121. Budapest: Akadémiai Kiadó.
- Hajdú, Péter. 1952. Az ugor kor helyének és idejének kérdéséhez [On the question of the place and time of the Ugric age]. *Nyelvtudományi Közlemények* 54. 264–269.
- Hammarström, Harald, Robert Forkel, Martin Haspelmath & Sebastian Bank. 2021. Glottolog 4.4. Leipzig: Max Planck Institute for Evolutionary Anthropology.
(<https://doi.org/10.5281/zenodo.4761960>, available online at <http://glottolog.org>) (Accessed 2021-08-31.)
- Haspelmath, Martin. 2001. The European linguistic area: Standard Average European. In Martin Haspelmath, Ekkehard König, Wulf Oesterreicher & Wolfgang Raible (eds.), *Language typology and language universals* (Handbücher zur Sprach- und Kommunikationswissenschaft, 20.2), 1492–1510. Berlin: Mouton de Gruyter.
- Hausenberg, Anu-Reet & Paul Kokla. 1988. Unificirovannaja sistema opisanija dialektov v primenenii k komi i marijskim glagol'nym formam. *Soviet Finno-Ugric Studies* 24. 19–26.
- Havas, Ferenc 2010. The Uralic typology database project. Paper presented at the Eleventh International Congress of Finno-Ugric Studies, Piliscsaba, 9–14.8.2010.

- (<https://slidetodoc.com/the-uralic-typology-database-project-ferenc-havas-budapest/>)
(Accessed 2021-11-28.)
- Havas, Ferenc, Márta Csepregi, Nikolett F. Gulyás & Szilvia Németh. 2015. Typological Database of the Ugric Languages. Budapest: ELTE Finnugor Tanszék. (utdb.elte.hu)
(Accessed 2021-06-09.)
- Heikkilä, Mikko. 2011. Huomioita kantasaamen ajoittamisesta ja paikantamisesta sekä germaanisista etymologioista suomalais-saamelaisille sanoille. *Virittäjä* 1. 68–82.
- Helimski, Eugene. 2003. Areal groupings (Sprachbünde) within and across the borders of the Uralic language family: a survey. *Nyelvtudományi Közlemények* 100. 156–167.
- Honkola, Terhi, Outi Vesakoski, Kalle Korhonen, Jüri Lehtinen, Kaj Syrjänen & Niklas Wahlberg. 2013. Cultural and climatic changes shape the evolutionary history of the Uralic languages. *Journal of Evolutionary Biology* 26. 1244–1253. (doi:10.1111/jeb.12107)
- Honkola, Terhi, Kalle Ruokolainen, Kaj Syrjänen, Unni-Päivä Leino, Ilpo Tammi, Niklas Wahlberg & Outi Vesakoski. 2018. Evolution within a language: environmental differences contribute to divergence of dialect groups. *BMC Evolutionary Biology* 18(1). (132. doi:10.1186/s12862-018-1238-6)
- Honkola, Terhi, Jenni Santaharju, Kaj Syrjänen & Karl Pajusalu. 2019. Clustering lexical variation of Finnic languages, based on Atlas Linguarum Fennicarum. *Linguistica Uralica* 55(3). 161–184. (doi:10.3176/lu.2019.3.01)
- Honti, László. 1979. Features of Ugric languages (Observations on the question of Ugric unity). *Acta Linguistica Academia Scientiarum Hungaricae* 29. 1–25.
- Honti, László. 1997. *Az ugor alapnyelv kérdéséhez* [On the question of the Ugric protolanguage]. (Budapesti Finnugor Füzetek 7). Budapest: ELTE BTK Finnugor Tanszék.
- Isanbaev 1994 = Исанбаев, Н. И.: *Марийско-тюркские языковые контакты. Часть вторая*. Йошкар-Ола: Марийский научно-исследовательский институт языка, литературы и истории им. В. М. Васильева.
- Johanson, Lars. 2000. Linguistic convergence in the Volga area. In Dicky Gilberts, John A. Nerbonne & Jos Schaecken (eds.), *Languages in contact* (Studies in Slavic and General Linguistics 28), 165–178. Leiden: Brill.
- Klumpp, Gerson, Lidia Federica Mazzitelli & Fedor Rozhanskiy. 2018. Typology of Uralic languages: Current views and new perspectives. Introduction to the special issue of ESUKA

- JEFUL. *Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics* 9 (1). 9– 30. (<https://doi.org/10.12697/jeful.2018.9.1.01>)
- Koptjevskaja-Tamm, Maria & Bernhard Wälchli. 2001. The Circum-Baltic languages: An areal-typological approach. In Östen Dahl & Maria Koptjevskaja-Tamm (eds.), *The Circum-Baltic languages: Typology and contact. Volume 1: Grammar and typology* (Studies in Language Companion Series 55), 615–750. Amsterdam, Philadelphia: John Benjamins.
- Kowalik, Richard (forthcoming). *A grammar of spoken South Saami*. Stockholm University doctoral dissertation.
- Laakso, Johanna. 2020. Contact and the Finno-Ugric languages. In Raymond Hickey (ed.), *The handbook of language contact*, 2nd edition, 519–535. Wiley-Blackwell.
- Lehtinen, Jyri, Terhi Honkola, Kalle Korhonen, Kaj Syrjänen, Niklas Wahlberg & Outi Vesakoski. 2014. Behind family trees. *Language Dynamics and Change* 4(2). 189–221. (<https://doi.org/10.1163/22105832-00402007>)
- Magga, Ole Henrik. 2014. Lullisámegiela muohtasánit. *Sámi dieđalaš áigečála* 1. 27–49.
- Miestamo, Matti. 2018. On the relationship between typology and the description of Uralic languages. *Journal of Estonian and Finno-Ugric Linguistics* 9(1). 31–53. (<https://doi.org/10.12697/jeful.2018.9.1.02>)
- Miestamo, Matti, Anne Tamm & Beáta Wagner-Nagy (eds.). 2015. *Negation in Uralic languages* (Typological Studies in Language 108). Amsterdam: Benjamins.
- Nichols, Johanna. 2021. The origin and dispersal of Uralic: Distributional typological view. *Annual Review of Linguistics* 7(1). 351–369. (<https://doi.org/10.1146/annurev-linguistics-011619-030405>)
- Norvik, Miina, Yingqi Jing, Michael Dunn, Robert Forkel, Terhi Honkola, Gerson Klumpp, Richard Kowalik, Helle Metslang, Karl Pajusalu, Minerva Piha, Eva Saar, Sirkka Saarinen & Outi Vesakoski. 2021. Uralic Typological database - UraTyp. Zenodo. (<https://doi.org/10.5281/zenodo.5236365>)
- Pajusalu, Karl, Kristel Uiboed, Péter Pomozi, Endre Németh & Tibor Fehér. 2018. Towards a phonological typology of Uralic languages. *Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics* 9(1). 187–207.

- Piha, Minerva. 2018. Combining Proto-Scandinavian loanword strata in South Saami with the Early Iron Age archaeological material of Jämtland and Dalarna, Sweden. *Finnisch-Ugrische Forschungen* 64. 118–233. (<https://doi.org/10.33339/fuf.66694>)
- Piha, Minerva & Jaakko Häkkinen. 2020. Kantasaamesta eteläsaameen. Lainatodisteita eteläsaamen varhaisesta eriytymisestä. *Sananjalka* 62. 102–124. (<https://doi.org/10.30673/sja.95727>)
- Pritchard, Jonathan K., Matthew Stephens & Peter Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155. 945–59.
- Rantanen, Timo, Outi Vesakoski, Jussi Ylikoski & Harri Tolvanen. 2021. Geographical database of the Uralic languages. Zenodo. (<https://doi.org/10.5281/zenodo.4784188>)
- Reesink, Ger, Ruth Singer & Michael Dunn. 2009. Explaining the linguistic diversity of Sahul using population models. *PLoS biology* 7(11). e1000241. (<https://doi.org/10.1371/journal.pbio.1000241>)
- Róna-Tas, András. 1988. Turkic influence on the Uralic languages. In Denis Sinor (ed.), *The Uralic languages. Description, history and foreign influences*, 742–780. Leiden, New York, København, Köln: E. J. Brill.
- Saarinen, Sirkka. 1997. Language contacts in the Volga region: Loan suffixes and calques in Mari and Udmurt. In Heinrich Ramisch & Kenneth Wynne (eds.), *Language in time and space. Studies in honour of Wolfgang Viereck on the occasion of his 60th birthday*, 388–396. Stuttgart: Franz Steiner Verlag.
- Skirgård, Hedvig, H. J. Haynie, Harald Hammarström, D. E. Blasi et al. Grambank data reveal global patterns in the structural diversity of the world’s languages. Submitted manuscript.
- Syrjänen, Kaj. 2021. Quantitative language evolution: Case studies in Finnish dialects and Uralic languages. Tampere University doctoral dissertation. (<https://trepo.tuni.fi/handle/10024/132864>) (Accessed 2021-08-12)
- Syrjänen, Kaj, Terhi Honkola, Kalle Korhonen, Jyri Lehtinen, Outi Vesakoski & Niklas Wahlberg. 2013. Shedding more light on language classification using basic vocabularies and phylogenetic methods: A case study of Uralic. *Diachronica* 30(3). 323–352. (doi:10.1075/dia.30.3.02syr)

- Syrjänen, Kaj, Terhi Honkola, Jyri Lehtinen, Antti Leino & Outi Vesakoski. 2016. Applying population genetic approaches within languages: Finnish dialects as linguistic populations. *Language Dynamics and Change* 6 (2), 235–283.
- Veenker, Wolfgang (ed.). 1985. *Dialectologia Uralica. Materialien der ersten internationalen Symposium zur Dialektologie der uralischen Sprachen 4.–7. September 1984 in Hamburg* (Veröffentlichungen der Societas Uralo-Altaica 20). Wiesbaden: Harrassowitz.
- Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3 (1). 160018. (doi:10.1038/sdata.2016.18)
- Ylikoski, Jussi. 2016. The origins of the western Uralic *s*-cases revisited: Historiographical, functional-typological and Samoyedic perspectives. *Finnisch-Ugrische Forschungen* 63. 6–78.

Appendices

Appendix 1. Statistical analyses of Uralic typological database

The executable R script and input data for data analysis can be found in an Open Science Foundation repository (DOI: 10.17605/OSF.IO/2ZG7H). Please note that you can find a pdf of the whole of Appendix 1 but also separate files for executable R code and for each figure separately.

Appendix 2. Uralic Typological dataset (UraTyp)

The full data is available as a Zenodo repository:

Norvik, Miina, Yingqi Jing, Michael Dunn, Robert Forkel, Terhi Honkola, Gerson Klumpp, Richard Kowalik, Helle Metslang, Karl Pajusalu, Minerva Piha, Eva Saar, Sirkka Saarinen & Outi Vesakoski. 2021. Uralic Typological database - UraTyp. Zenodo. (<https://doi.org/10.5281/zenodo.5236365>)

Appendix 3. Author contributions

Development of the project

The UraTyp project was developed by OV, TH, MD, KP, and MN.

Design and management of the questionnaire

The UraTyp questions were developed by GK, HM, MN, KP, and ES. Grambank questions were developed at the Department of Cultural and Linguistic Evolution at the Max Planck Institute for the Science of Human History and at the Max Planck Institute for Evolutionary Anthropology by a team of people lead by Russell Gray: <https://glottobank.org/people.html>

Introducing the principles and training the coders

To ensure that the UraTyp questions were created and coded following the principles used in Grambank, we were instructed by Harald Hammarström and MD.

Feedback to the questionnaire

The final draft of the UraTyp questionnaire was commented by Jeremy Bradley and Ksenia Shagal. Ksenia Shagal also gave advice on composing the questions on nonfinites.

Coders

MN (coordination of the coders), MP, and ES coded the UraTyp questions. RK and MN coded the Grambank questions.

Members of the project acting also as language experts:

GK, RK, HM, MN, KP, MP, ES

Statistical analyses

YJ, MD, OV

Database and user-interface

YJ, RF, Luke Maurits

Writing

First versions of sections were written as follows: Section 1 by MN and KP; Section 2.1 by HM, 2.2 by MN, 2.3 by RF and YJ, 3 by OV, MD, and YJ, and 4–5 by MN, OV, MD, and KP. All authors commented and improved on the text.

Address for correspondence

Corresponding author

Miina Norvik

Ülikooli 18,

50090 Tartu, Estonia

miina.norvik@ut.ee