Clustering lexical variation of Finnic languages based on Atlas Linguarum Fennicarum

Honkola, T.[1], Santaharju, J.[2], Syrjänen, K.[3] & Pajusalu, K.[4]

Running title: Quantitative lexical clustering of Finnic languages

[1] Department of Biology, FI-20014 University of Turku, Turku, Finland, terhi.honkola@utu.fi, +358 407410453; Institute of Estonian and General Linguistics, University of Tartu, Jakobi 2, 51005 Tartu, Estonia.

[2] Organismal and Evolutionary Biology Research Programme, FI-00014 University of Helsinki, Helsinki, Finland, jenni.leppanen@helsinki.fi, +358 400700085

[3] Faculty of Information Technology and Communication Sciences, FI-33014 Tampere University, Tampere, Finland, kaj.syrjanen@tuni.fi

[4] Institute of Estonian and General Linguistics, University of Tartu, Jakobi 2, 51005 Tartu, Estonia, karl.pajusalu@ut.ee, +372 5267733

Abstract The article focuses on lexical relations of the Finnic languages. Here we studied whether lexical data is suitable for detecting the coarse-grained and fine-grained substructure within the Finnic group. We evaluated this by clustering old lexical variation from a dialectal dataset covering the whole Finnic speaker area (Atlas Linguarum Fennicarum; ALFE) using quantitative methods adopted from population genetics, and by comparing our results to groups suggested by earlier linguistic literature. We found the main lexical division between north-eastern and south-western Finnic. According to our lexical analysis, the Finnic languages are Finnish, North Estonian, South Estonian, Livonian, Karelian, Veps, and Votic-Ingrian. These groups matched well with the earlier suggested divisions, and we concluded that lexical data could be utilised more often in defining linguistic sub-structures, especially in linguistic situations that involve dialect continua.

Keywords: lexicon, dialect continuum, areal variation, quantitative analyses, Finnic language area

Introduction

The amount of linguistic variation in the world is immense. Linguistic variation can be classified in order to detect, for example, languages that belong to the same language family, languages that form a linguistic area (Sprachbund) and dialect borders between and within languages (Anttila 1989: 318; Chambers, Trudgill 1998: 89; Thomason 2000: 311). This is done in order to shed light on different aspects of the past of the languages in question, such as shared history, contacts or early divisions. In many linguistic studies, lexical evidence is often not the main evidence used, as lexical features are considered to be borrowed relatively easily when compared to grammatical features (Koponen 1991: 126; Chambers, Trudgill 1998: 97-99; Thomason 2000: 312). However, it has been noted that, at least with some languages, instability also exists in the non-lexical data; for example, in a study conducted with Austronesian languages, most grammatical features were noted to change faster than the items of basic vocabulary (Greenhill et al. 2017). Lexical innovations, once obtained, are rather stable, and differences in the innovations can provide a good basis for classification (Salminen 1998: 391).

Our study focused on investigating the suitability of lexical data in detecting divisions within the Finnic language group. The Finnic languages, which make up one of the westernmost sub-branches of the Uralic language family, form a dialect continuum around the Gulf of Finland (Figure 1). Estonian and Finnish have the status of national language and they are the language of the majority in their respective countries. Other Finnic languages are spoken as minority languages in Russia, Norway and Sweden, whereas in Latvia they have already become extinct. Historically, all Finnic languages are descendants of Proto-Finnic, and have split during the last ca. 2000 years due to different internal developments, which have subsequently been counteracted by the spread of innovations within the group (Laakso 2001: 204-207; Kallio 2007:

243-246; Kallio 2015a: 26; Lang 2018: 259-260). This group has intrigued researchers since the early 17th century (Ariste 1965: 80-87; Lang 2018: 227-229) and there are still different views concerning the number of Finnic languages (Tuomi et al. 2004: 11), what the main groups are (Salminen 1998: 392), and how the Finnic group has taken shape (Kallio 2015b: 93-94).

Both Glottolog 3.3 (Hammarström et al. 2018) and Ethnologue (Simons, Fennig 2018) list twelve Finnic languages: Finnish, Kven Finnish, Tornedalen Finnish, Karelian Proper, Livvi-Karelian, Ludian, Veps, Votic, Ingrian, North Estonian, South Estonian, and Livonian (Figure 1). Of these twelve, Kven Finnish, spoken in Norway, and Tornedalen Finnish, spoken in Sweden, have the official status of minority languages in their respective countries, while from a linguistic perspective they are traditionally considered to be dialects of Finnish (Kettunen 1930: 81, 95; Salminen 2007: 231) rather than separate languages. Of the Karelian varieties, Karelian Proper, Livvi-Karelian and Ludian could be granted a language status (Salminen 2007: 217; Pahomov 2017) but Livvi-Karelian and Ludian have also been regarded as dialects of Karelian (Laanest 1975: 26; Sammallahti 1977: 133; Viitso 1998: 99; Laakso 1991: 61-62). Ingrian has also been associated with Karelian (Kettunen 1960: 2; Turunen 1988: 59) while in particular Estonian scholars have considered it a separate language (Laanest 1975: 11; Viitso 1998: 96). It has also varied whether South Estonian is considered a separate language from North Estonian (Salminen 2007: 217; Hammarström et al. 2018; Simons, Fennig 2018), and the first one to diverge from the Finnic unity (Sammallahti 1977: 132-133; Kallio 2014: 163), or as a distinctive part of the Estonian language (Laakso 1991:97). Nowadays it is common to separate out at least seven (or eight) Finnic languages: Finnish, Karelian, Veps, Votic, Ingrian, Estonian, (South Estonian), and Livonian.

< Insert Figure 1 here >

The history of the Finnic group has been studied by determining the main groups of the Finnic languages and the subsequent divergences and convergences of the subgroups on a historical-comparative basis. Based on this deduction, two main lines of grouping have been suggested. In both of these, the Finnic languages are divided roughly into north-eastern (Finnish, Karelian, Ingrian, Veps) and south-western (Estonian, Livonian, Votic) groups (Lehtinen 2007: 155-167). The main difference between these two views is whether the Finnish language is considered to belong as a whole to the north-eastern group (Setälä 1916: 504; Ariste 1965:80; Raun 1971:49-98), or whether the western dialects of Finnish should be part of the south-western group (Ojansuu 1922: 139-145). In addition, the existence of an eastern group has been suggested (Itkonen 1983: 209-217), but its status appears to be less clear than those of the north-eastern and south-western groups (Itkonen 1983: 209-224). When specifically studying the order of divergences based on phonological and morphological features, there is evidence that South Estonian was the first to diverge, followed by either the north-south split (Sammallahti 1977: 132-133) or the divergence of Livonian (Kallio 2014: 156-163).

Differences in the classifications are partly due to differences in the linguistic material used in the classification, and partly due to differences in the criteria set for a language, which may also vary depending on the focus and scope of the study. For example, synchronic linguistics may be interested in more fine-grained sub-structuring than what is useful for historical linguistic research (Salminen 1998: 390-391). Most classifications of the Finnic languages have focused on phonological and morphological features, whereas lexicon has been utilised for this purpose to a lesser degree (Alvre 1973: 154-157; Kallio 2014: 155). However, with many languages, lexicon is essential for mutual intelligibility (cf. for example the lexical differences between Finnish '*pelata*' and Estonian *'mängida'* to the phonological differences between Finnish '*työ'* and Estonian '*töö*'). Therefore, the distribution of lexical variation could be assumed to coincide well with language borders. This may be especially the case with closely related languages, whose overall linguistic systems tend to be very similar to each other (e.g. Finnic languages) compared to languages belonging to different language families (e.g. Finnish vs. Japanese). However, it needs to be pointed out that for example in the case of Danish, Norwegian and Swedish, phonetic differences generally play a larger role in mutual intelligibility than differences in the lexicon (Gooskens 2007).

Here we classified the Finnic language group based on lexical data. We began by identifying the main lexical groups, followed by an exploration of the subgroupings that emerge when clustering lexical variation. Our aim was to study whether lexical variation is appropriate for detecting the main groups and subgroups of linguistic variation, i.e. whether the lexical groups we find are similar to the ones detected using other types of linguistic data. Notably, because we focused only on lexical variation, we cannot refer to the groups that we obtained as 'languages' or 'main groups' of the Finnic languages. Therefore, we refer to the groups we obtained from our analyses as 'lexical languages' and 'lexical main groups'.

We studied these questions using a dialect atlas of Finnic languages (ALFE; Tuomi et al. 2004; 2007; 2010), which describes linguistic variation in the Finnic languages at the beginning of the 20th century. At that time, Finnic speakers were still living in rural societies and the level of urbanisation was low (Tuomi et al. 2004: 17). To increase the time depth of our study, we limited our study to features that described old (pre-13th century) lexical variation in the Finnic area. Using this dataset, we aimed to resolve the deep lexical divisions within the Finnic group at both coarse-grained and fine-grained levels, and we hope to contribute to the discussion on the role of lexical data in linguistic classification.

We studied these questions using quantitative data clustering methods developed in population genetics. These methods have also been applied to cluster linguistic variation (Dunn et al. 2008, Reesink et al. 2009, Bowern 2012) as well as dialectal variation (Syrjänen et al. 2016, Honkola et al. 2018, Santaharju et al. ms.). We used two model-based clustering methods, Structure (Pritchard et al. 2000) and BAPS (Corander et al. 2003). Structure has been noted for its ability to identify the uppermost hierarchical division (Evanno et al. 2005: 2615, 2618), and we used it to study the main lexical division of the Finnic languages. BAPS, on the other hand, has been found to produce a more fine-grained division than Structure (Fjellheim et al. 2009: 24), and we used it to infer the number of lexical languages.

Materials

The atlas of Finnic languages, Atlas Linguarum Fennicarum (ALFE), has been compiled as a multinational effort of the Institute for the Languages of Finland, the Institute of the Estonian Language and Karelian Research Centre of Russian Academy of Sciences since the 1980s. The atlas was published in three parts (Tuomi et al. 2004; 2007; 2010), and is publicly available in a digital format in the AVAA service of the Institute for the Languages of Finland (https://avaa.tdata.fi/web/kotus/aineistot).

The atlas describes the spatial variation of Finnic languages and their dialects at the beginning of the 20th century, more specifically before 1939 (Tuomi et al. 2004: 17). It includes mainly lexical variation, but phonological and morphophonological variation of the collected lexical items have also been included. The decision to focus mainly on lexical variation in ALFE was due to the existing dialect atlases of Finnish (Kettunen 1940) and Estonian (Saareste 1938; 1941; 1955), which already covered phonological and morphological variation in these two languages. Preparations for the dialect atlas of Karelian also started in the 1930s, but it was published only in 1997 (Bubrih et al. 1997). It includes phonological and lexical variation.

Atlas Linguarum Europae (ALE; Weijnen 1975) served as a model for ALFE, and ALFE includes some ALE features which were considered to be appropriate for studying Finnic languages. In total, ALFE's questionnaire included over 300 questions, and it aimed to collect linguistic features describing different aspects of archaic human living, such as living conditions, means of livelihood, human anatomy, kinship terms, animals, plants, and handicrafts (Laanest, Jussila 1989).

The published atlas covers 298 different linguistic features. The linguistic variation of each feature is presented in map form, with the number of maps per feature varying from one to 12. In total, ALFE includes 687 map sheets. Most of the features that consist of individual maps document lexical variation across the whole Finnic area. Features consisting of more than one map use the additional maps to present different aspects or areal coverages of the studied phenomenon. For example, map number 24.1 describes the lexical variation of the term 'cowshed' in all Finnic languages, 24.2 presents the phonetic variation for the different terms for cowshed used in Finland and 24.3 presents what meaning a specific word for 'cowshed', *läävä*, has in different places where it is used. There is also variation in whether one map presents only one type of linguistic variation or whether there are two different types of linguistic variation within one map sheet. For example, map 2 shows both the lexical variation (*tuli, valkea, lekko, lämöi*) as well as the derivative variation of *tuli* and *valkea* (*tuluke*, *tuluk; valu, valkko, valkonen, valkos/kulta*, *valku*), whereas map 4.1 presents only the lexical variation between *haiku, savu* and *suits*. Because the atlas has been compiled as a joint effort of several participants, there is variation in how the linguistic features are structured on the maps.

The data has been collected from 259 main data collection localities scattered throughout the Finnic area (Figure 2; Tuomi 2004: 21). Depending on the language in question, these localities refer to municipalities (e.g. with Finnish and Estonian), villages (e.g. Veps) or main dialects (e.g. Livonian). The number of these localities varies greatly per language along with the size of the speaker areas. There are 185 main data collection localities from Finnish (within these, three from Kven Finnish in Norway and nine from Tornedalen Finnish in Sweden), 20 from North Estonian, 10 from South Estonian, 18 from Karelian Proper, seven from Livvi-Karelian, four from Ludian, seven from Veps, three from both Ingrian and Votic, and two from Livonian (Table 1). In addition to the main data collection localities, there are also small amounts of data from other localities within Finland and Estonia.

< Insert Figure 2 here >

The Finnish and Estonian data originates from the dialect archives of the Institute for the Languages of Finland and from the Institute of the Estonian Language. The data for Ingrian, Votic and Livonian is mainly from the same archives but also partly from the personal collections of Tiit-Rein Viitso and Arvo Laanest. Karelian and Veps have been collected mainly from informants after 1987. However, the data collected from Karelian and Veps can be considered to represent the linguistic situation prior to 1939 (Tuomi et al. 2004: 17), as the written standard languages created in the 1930s, were used only for a short period of time, and later written standards began to emerge as late as the 1990s (Anttikoski 1998: 119-126; Grünthal 2015: 23, 52).

To study the deep divisions of the Finnic languages based on lexical variation, we narrowed the atlas down to a set of maps which recorded old lexical variation with sufficient areal coverage. We classified different features as old if the main lexical division had taken place in pre-historical times (i.e. before the 13th century). The age of the division was estimated by the basis of current knowledge about the history of Finnic languages (SSA, EES). For example, map 65.1 with the feature 'evening' is classified as old, since both *ehto* and *ilta* are pre-historical stems with different distributions in the Finnic language area. The division between *elää* and *asua* ('be located') in map 14 is classified as young because the verb *asua* is a secondary derivative and its meaning varies considerably; furthermore, *elää* and *asua* often appear alongside each other in the Finnic languages.

We considered the feature to have sufficient areal coverage if there was data from ca. two thirds of the data collection localities in each map sheet (Table 1). With this, we wanted to ensure that our results were not skewed by large amounts of missing data from certain languages or dialects. We were more relaxed with the two-thirds requirement in Finland as the number of data collection localities was higher than in other areas; the other areas were checked carefully. As a result, when including only maps which recorded old lexical variation with sufficient areal coverage, our dataset included 184 linguistic features.

Table. 1. Number of main data collection localities per language or dialect

	# of localities in ALFE	# of localities required in each map sheet (“two-thirds requirement”)	# of localities in the full dataset	# of localities in the balanced dataset
Finnish	185		177	30
North Estonian	20	13	20	20
South Estonian	10	6	8	8
Karelian Proper	18	8*	18	18
Livvi-Karelian	7	4	7	7
Ludian	4	3	4	4
Veps	7	4	7	7
Ingrian	3	2	3	3
Votic	3	2	3	3
Livonian	2	1	2	2
In total	259		249	102

* Tver Karelian excluded from the calculations

All of the 259 data collection localities did not have information of all of the 184 linguistic features. We excluded localities with less than 78 collected features from our analyses. These included two localities from the South Estonian area (Leivu and Lutsi) as well as eight from the Finnish-speaking area (e.g. two from Kven Finnish and four from Tornedalen Finnish). As a result, our analyses ultimately had 249 localities (Table 1), each covering between 78 and 184 linguistic features.

As the number of data collection localities from the Finnish-speaking area was very large compared to the others (Table 1), we also made another, more balanced dataset with only 30 localities from the Finnish-speaking area. These localities were evenly distributed in southern and central Finland and covered both the eastern and the western main dialect. This area was chosen because of our interest in studying the deepest divisions, and here Finnish has been spoken for the longest time (e.g. Frog & Saarikivi 2015: 71). The number of studied localities in the balanced dataset was 102 in total.

The digital data found in the AVAA service was converted into a numerical format such that localities are presented as separate rows and different linguistic features and their variants are presented as separate columns (Table 2). This type of representation is commonly used by population genetic software. We used binary coding where the presence of a linguistic variant is marked with 1 and its absence with 0. Missing data is marked with -9. For example, in map 2 the variant used for the word 'fire' in Ala-Laukaa is *tuli* and in Alavus *valkea* (Table 2). The dialect data of Finnish (Kettunen 1940) has been analysed in a comparable format (Santaharju et al. ms). In this study, we were interested only in lexical variation. Therefore, we grouped the possible other types of variation (e.g. in derivatives) under one representative lexeme. For example, map 2 has three variants for the word *tuli* (*tuli, tuluke, tuluk*), but in our dataset all three of these variants are included in the column labelled *tuli*.

Table 2. An example of binary coding of the ALFE data

Map #	2				3.1	
	lekko	*lämöi*	*tuli*	*valkea*	*kynttelä*	*tuohus*
Ala-Laukaa	0	0	1	0	-9	-9
Alavus	0	0	0	1	1	0
Asikkala	-9	-9	-9	-9	-9	-9
Aunus	0	0	1	0	0	1

Some of the linguistic information which existed in the printed atlas was missing from the digital version. For example in the printed map 97.1, variants are recorded from all three Votic localities: Eastern Votic, Western Votic and Kukkuzi, but in the digital database only the variant of Kukkuzi is present. In cases when there was too little data from some languages in the digital database according to the "two-thirds requirement" (e.g. the data covers only 1/3 of the Votic localities, whereas 2/3 is required for the feature be included in the dataset (see Table 1)), we added the linguistic information from the printed atlas to the numerical database. We also removed duplicates from the digital dataset so that each locality was represented only once.

In sum, we analysed old lexical variation using two datasets, full and balanced, which differed from each other in the number of localities covering Finnish. With both datasets, we focused the analyses on the features that had sufficient areal coverage. In binary form both datasets consisted of 184 linguistic features represented by 1211 columns of data (Table 3).

Table 3. Summary of the used datasets

	Full dataset	Balanced dataset
# of localities studied	249	102
# of features studied	184	184

Methods

We analysed the data with two model-based clustering methods, Structure 2.3.4. (Pritchard et al. 2000) and BAPS 6.0 (Corander et al. 2003; Corander & Marttinen 2006; Corander et al. 2008a). Structure has been used earlier to cluster both variation within a language (Syrjänen et al. 2016; Honkola et al. 2018; Santaharju et al. ms.) and between languages (Dunn et al. 2008; Reesink et al. 2009; Bowern 2012) and was found applicable for these purposes. BAPS has been used to cluster the Finnish dialect data and estimate the exchange of linguistic variants between dialect groups (Santaharju et al. ms).

Structure and BAPS both use model-based methods based on Bayesian inference (Pritchard et al. 2000; Corander et al. 2003). Bayesian inference is a statistical approach to modelling complicated statistical problems whilst taking into account prior information (Puza 2015: 1). Population inference in both tools is based on two assumptions: first, they assume that the studied populations are in the Hardy-Weinberg equilibrium (HWE), meaning that the inferred populations remain stable through time. While the criteria of the HWE is rarely met in reality, its inexistence does not prevent the analysis being done but limits the inferences that can be made from the results. That is, if the populations have not remained stable over time – as is likely to be the case – one should see the clusters as representing the linguistic situation around the time of data collection, instead of making inferences of too far into the past (see Syrjänen et al. 2016). The second assumption is that the studied features are independent from each other, i.e. they are in linkage equilibrium. In the case of language data, it means that the studied linguistic features should be independent from each other. We studied the dependencies of the linguistic features using a comparison method originally proposed in Syrjänen et al. (2016: 263-264). We yielded a relative similarity estimate between each pair of features ranging from features with identical distribution of variants (coefficient of 1.0), to features with completely dissimilar distribution of variants (coefficient of 0.0). This was done for the full dataset and achieved by calculating Jaccard coefficients between features, where the features were represented as sets of locality pairs with identical feature variants. The results of this analysis were visualised as a heat map showing each feature compared with every other feature, with colours varying between yellow (coefficient 0.0) and red (coefficient 1.0).

Although Structure and BAPS are used to achieve the same goal – clustering variation into populations – they approach it in different ways. As a consequence, their results may differ from each other (Bohling et al. 2013: 82; Fjellheim et al. 2009: 24). One notable difference between Structure and BAPS is in the way in which they identify populations (mixture analysis) and the extent of admixture (admixture analysis). These are two separate analysis steps in BAPS (Corander & Marttinen 2006: 2834), whereas Structure performs them simultaneously (Pritchard et al. 2000). The two tools also use different algorithms in their analysis; Structure uses a standard Markov Chain Monte Carlo (MCMC) (Pritchard et al. 2000) while BAPS uses stochastic optimisation (Corander et al. 2006; Corander, Marttinen 2006). Structure and BAPS also differ in how the most likely number of clusters is estimated. In Structure, the calculation is repeated multiple times for different K values (inferred number of populations), and the number of clusters is inferred afterwards from the likelihood values of the analyses (Pritchard et al. 2000; Evanno et al. 2005). In BAPS, an upper bound for the inferred number of populations is specified before the analysis, and the partition with highest posterior probability is detected with repeated optimisation operations (Corander et al 2006). Within a single run, BAPS provides the list of best visited partitions, their logarithmic maximum likelihood values, and the probability for the optimal number of clusters (K) (Corander et al. 2006). In addition, BAPS can use geographical coordinates of the localities as prior information for the clustering analysis (Corander et al. 2008b). In contrast, geographical prior information in Structure has to be defined as integers, with each integer representing a specific sampling location (Pritchard et al 2010).

We ran a Structure analysis with both the full and the balanced datasets with K values 1-10, performing ten repetitions per each K value. In other words, we asked the software to divide the data for example into five clusters (K=5) and repeat this analysis 10 times. The repetitions are done to estimate the stability of the resulting clustering, as there may be several equally good solutions to cluster the data. The Structure analyses were run without inputting prior information of the geographical locations. The initial 10,000 generations were specified as burn-in, which refers to the discarded initial samplings from the MCMC chain that takes place before the analysis reaches the area of high likelihood. 100,000 steps following the burn-in were used for the actual analysis. We used the admixture model, which allows the individuals to originate from more than one ancestral population (Pritchard et al. 2000). This gives the model more realism, as it allows it to infer transitional areas between dialects and/or languages rather than sharp borders only. We summarised the repeated analysis for each K value with CLUMPP (Jakobsson, Rosenberg 2007), which summarises all the runs with the same K value under one combined result.

We also ran BAPS with both datasets (full and balanced), both with and without information of the geographical coordinates of the localities (four analyses in total). First, we detected the optimal number of clusters (i.e. the mixture analysis) from the range of K=2-20, with ten repetitions for each K value. Second, we detected the admixture proportions for each locality (i.e. admixture analysis). The admixture proportions were detected by using 500 iterations, 1000 reference localities (from each language/dialect) and 100 iterations per reference locality.

We determined the most likely number of lexical languages based on the BAPS analysis, using the list of best visited partitions with the highest marginal likelihood value. From the Structure results we obtained the most likely number of lexical main groups, using the ΔK metric (Evanno et al. 2005), which detects the K value with the largest difference in log likelihood values compared to the neighbouring ones.

We also wanted to study the capability of Structure to divide the data into a larger number of clusters. We did that in two different ways: first, we estimated the most likely number of clusters based on the mean log likelihood values (Pritchard et al. 2000), which are mean values of the individual likelihood values produced for each individual Structure run. Second, we studied the substructure of the main clusters by analysing them individually, as instructed in Evanno et al. (2005: 2616-2618). Calculations and plots related to the Structure analysis were done with the R package pophelper (Francis 2017).

Both BAPS and Structure produce a set of membership coefficients for each studied locality, which show the relative proportions of different ancestral populations. K=2 has two ancestral populations, K=3 has three etc. These membership coefficients sum up to 100 % for each locality. For example, when dividing the data into three groups, one locality could have 80 % membership for population A, 10 % membership for population B and 10 % for population C. In this case, the locality belongs most strongly to population A.

We plotted the results on a base map, which includes Finland, Estonia and parts of the neighbouring countries with municipal borders from ca. 1938. The base map, prepared by the Institute for the Languages of Finland, is available via the PaITuli web service (https://avaa.tdata.fi/web/paituli/latauspalvelu). The base map was slightly modified to accommodate the smallest languages. The polygons indicating the speaker areas of Livonian and Veps, spoken only in certain villages, were enlarged to help visually assess which populations these locations belong to. Therefore, the polygons are not in proportion with the size of the actual speaker area. New polygons were made for Kukkuzi and Leivu, and the names of Tihvinä and Tver were changed to Śeĺiśśa and Tolmačču to match those in the printed atlas and the digital database. Overlapping polygons were removed from the base map and the remaining ones were moved closer to reduce the amount of empty space in the language maps. The visualisations and modifications to the base map were done with ArcGIS.

We visualised the results on the maps in two ways. Firstly, we visualised the core and transitional areas for each language or dialect by colouring the localities. We considered localities with membership coefficient > 0.75 to belong to the core language or dialect area and coloured these with more saturated colours. Localities in which the coefficient was 0.50-0.75 were considered to belong to transitional areas, and these were coloured with less saturated colours. This type of visualisation was used for both the full and the balanced datasets. Secondly, we visualised the results of the balanced analyses with bar plots. This makes it possible to visualise all the membership coefficients of the location in question, including those below 0.50. We show the CLUMPP results for both the full and balanced analysis below and the clustering of the individual runs with the highest likelihood value in Supplementary Materials.

Results

The studied features had a low amount of linkage with each other (Supplementary Figure 1a) and we included them all in the analyses. The largest linkage percentages were between maps 153.1-155, 195.1-204.1 and 245-250.1 (Supplementary Figure 1b).

Clustering of the full data

Based on ΔK values, the Structure analysis suggested that the division into three clusters was the supported division, and thus the best uppermost hierarchical division (Supplementary Figure 2a). The mean likelihood values increased gradually until K=7. At this point, the likelihood values began to fluctuate more between the repetitions of different K values, and the mean likelihood decreased (Supplementary Figure 2b). The differences in the mean likelihood values were small between K=2 and K=7, indicating that the data is explained consistently with all these K values.

We summarised the ten individual Structure runs for each K value (outliers excluded) with CLUMPP. Finnish (blue in Figure 3) and Estonian-Livonian (orange in Figure 3) were the core areas with K=2, whereas the rest belonged to these clusters more vaguely. When the data was divided into three, northern (Finnish), southern (Estonian and Livonian) and eastern groups (Karelian, Veps, Ingrian; green in Figure 3) appeared. This was the best-supported coarse-grained division of the full data based on ΔK values. With K=4, Finnish was divided into Western (blue) and Eastern main dialect (brown) whereas the southern (Estonian and Livonian) and the eastern group (Karelian and Veps) remained the same as with K=3, with the exception of the easternmost part of Ingrian now being associated with eastern Finnish. With K=5, the largest changes again occurred within Finnish, with Eastern Finnish divided further into Savo (brown) and Southeast (gray) dialects. Some of the more northern Finnish speaking areas were now also part of the eastern dialects whereas with K=4 they were associated with the western dialects. With the K=6 summarisation the Savo area essentially disappeared, showing Finnish divided roughly into west (blue), east (grey) and north (pink). In addition, a combined group of Ingrian, Votic and Livonian (yellow) appeared. With K=7, the Savo cluster reappeared, but the seventh group had a strongly mixed ancestry, with no membership coefficients higher than 0.5; because of this, the seventh group does not appear on the map, and large areas of Finland are shown as white (Figure 3).

The highest log likelihood runs were largely similar to the CLUMPP summarisations of K=3-5 (Supplementary Figure 3). The K=2 division clustered the data into Finnish and the rest. When the data was divided into six groups, Veps was the new group to appear instead of Northern Finnish and with K=7, a group with Votic, Ingrian and Livonian appeared (Supplementary Figure 3).

< Insert Figure 3 here >

The BAPS analysis with the full dataset suggested that the division into five groups was the most supported one (with a probability of 0.997). However, a division into six groups was also present four times among the ten best visited partitions (the order of best visited partitions 5, 5, 5, 6, 5, 5, 5, 6, 6, 6). The five clusters found by BAPS were Western and Eastern Finnish, Estonian, the eastern group (Karelian and Veps) and a group of Ingrian-Votic-Livonian (Figure 4). In the full analysis, the result also remained identical when the analysis was done with geographical coordinates.

< Insert Figure 4 here >

Clustering of the balanced data

In the balanced Structure analysis with only 30 localities from Finland, the most supported uppermost hierarchical clustering based on ΔK split the data into two groups (Supplementary Figure 4a). The mean likelihood values increased gradually until K=5, after which the variation in the likelihood values between the repeated runs increased and the mean likelihood decreased (Supplementary Figure 4b). There were no large differences in the likelihood values between K=2-5.

In the CLUMPP summarisation of K=2, the data was divided into north-east and south-west, more specifically into Estonian and Livonian vs. the rest (Figure 5a). The admixture of these ancestral groups can be seen in Livonian, Ingrian, Votic, and the western Finnish localities (Figure 5b). With K=3, the eastern group with Karelian and Veps appeared (Figure 5c). Ingrian and Votic are a mixture of these three components (Figure 5d), as is Livonian, but with Estonian as the major component. Estonian influence can be seen in the south-west Finnish areas (orange tips in blue bars), and Finnish influence can be seen in the northernmost Karelian areas (blue tips in green bars). In addition, eastern influence is seen in the easternmost Finnish areas (green ends of blue bars). When clustering to four, a group of Ingrian-Votic-Livonian appeared (Figure 5e). Ingrian and Votic are the core area of this group, while Livonian is a mixture of this and Estonian (Figure 5f). In addition to the influences seen with K=3, there is Ingrian-Votic influence in Finnish in the Karelian Isthmus (yellow tips in blue bars). With K=5, membership coefficients do not exceed 0.5 in any of the studied localities, due to which no clear fifth cluster appears (Figure 5g). Instead, the "fifth cluster" appears as minor fractions in the peripheral areas of Veps, Western Finnish, Livonian and South Estonian (turquoise in Figure 5h). The highest log likelihood runs and the CLUMPP summarisations were very similar for K=2-4 (Supplementary Figure 5) in the balanced Structure analysis. With K=5, the highest log likelihood run differed from CLUMPP results and Veps was the new group to appear (Supplementary Figure 5).

< Insert Figure 5 here >

In the balanced analysis, the uppermost hierarchical clustering according to ΔK was K=2, a rough division into north-east and south-west (Figure 5a). These two groups were analysed separately, as suggested by Evanno et al. (2005), to detect the secondary division within these groups. We analysed 72 sampling locations from the north-east cluster and 30 from the south-west cluster with Structure with the same settings as the other analyses (see Methods). According to the ΔK value, the best supported subdivision of the north-eastern group splits the cluster into two, Finnish-Ingrian-Votic and Karelian-Veps (Figure 6a), with most of the admixture found in Ingrian and Votic (Figure 6b). When analysing the south-western group, the ΔK value peaked with K=3, suggesting a division into North Estonian, South Estonian and Livonian to be the most supported one (Figure 7). Consequently, the subdivisions of the main cluster suggested a total of five clusters.

< Insert Figure 6 here >

< Insert Figure 7 here >

BAPS found seven clusters in the balanced analyses (with a probability of 0.713) without the usage of geographical coordinates as a prior; however, the ten most visited partitions also included divisions into four, five, six and eight clusters (list of best visited partitions 7, 6, 6, 6, 5, 5, 4, 8, 8, 7). The seven clusters detected were Finnish, Karelian, Veps, Ingrian-Votic, North Estonian, South Estonian, and Livonian (Figure 8a). A small Vepsian influence was inferred in Ludian, and North Estonian influence could be seen in South Estonian. In addition, the Finnish spoken in the Karelian Isthmus had traceable influences of Karelian and Ingrian-Votic (Figure 8b). The BAPS analysis with geographical coordinates suggested the division of the balanced data into six clusters (with a probability of 0.973) while also four, five, seven and eight were among the ten most visited partitions (list of the best visited partitions 6, 7, 5, 4, 8, 8, 8, 8, 8, 8). The six clusters were otherwise identical to the seven clusters obtained without the coordinates, except Livonian and South Estonian was grouped together here (Supplementary Figure 6).

< Insert Figure 8 here >

Discussion

With both of the analysed datasets, Structure and BAPS detected groups which have been suggested earlier in the linguistics literature. However, there were notable differences in the nature of the clusters that appeared with the full dataset and those that appeared with the balanced dataset. With the full dataset, many of Structure's divisions took place within Finland, the area where most of the data points are, while Ingrian, Votic and Livonian, the areas with the smallest number of data points, were clustered together. In the balanced dataset, the produced clusters were more in line with the classical divisions of the Finnic languages, especially when analysed with BAPS.

Structure has been noted to be sensitive to unbalanced datasets, and as a result, most divisions take place in the area with the highest number of data points, while areas with a small number of data points are clustered together (Neophytou 2014; Puechmaille 2016). This also happened in our dataset, especially with the full data, where the level of imbalance was very high between the languages (e.g. Finnish: 177 data points vs. Livonian: two data points). However, as our main objective was to detect the lexical main groups with Structure and not the lexical languages, the level of imbalance was not as large between the main groups as between languages, especially with the balanced dataset.

BAPS has been found to tolerate data imbalance better than Structure (Neophytou 2014; 282). In our results, BAPS was able to detect groups with only a few localities and thus tolerate data imbalance with the balanced data, whereas with the full data this was not the case. Hence it seems that while BAPS tolerates more data imbalance than Structure, it tolerates only moderate amounts of it. To prevent possible skewing of the results due to data imbalance, we mainly focus on discussing both Structure and BAPS results produced with the balanced dataset.

According to the balanced Structure analysis, the main lexical division in the Finnic area is between two groups, north-east and south-west. Our result differ from the main groups suggested earlier (Setälä 1916: 504; Ariste 1965: 80; Raun 1971: 49-98) in that Votic is grouped together with the north-eastern group in our results instead of the south-western group, as it is in the traditional divisions. Of the languages that formed the north-eastern cluster in our analyses, Votic is, however, the language that is closest to the south-western group. This is in line with the mixed nature of Votic noted also for example in Kettunen (1960: 62-63, 215-216). Bar plot visualisations of the two clusters also show the connections of the western dialect of Finnish to the south-western group suggested by Ojansuu (1922: 139-145), as the western dialect of Finnish has south-western influence, which is missing from the eastern dialect of Finnish, Karelian and Veps. The early division of South Estonian from the Finnic unity (Sammallahti 1977: 132-133, Kallio 2014: 156-163) was not supported by our results. This is likely due to different kinds of data used in the analysis, as we used solely lexical data, while the suggested initial divergence of South Estonian is based on phonological and morphological data.

Regarding the number of Finnic lexical languages, seven clusters were found by BAPS with the balanced data. This division is more coarse-grained than the division into twelve languages listed in Ethnologue (Simons, Fennig 2018) and Glottolog 3.3 (Hammarström et al. 2018). Kven Finnish and Tornedalen Finnish, which both have official status as national languages in Norway and in Sweden (Sulkala 2010; 11, 13), did not appear in our results. This is likely due to their similarity with Finnish (Kettunen 1930: 81, 95) but also due to reasons of coverage; only a small number of localities represented these areas in the full data set, and in the balanced analyses these areas were excluded completely. Furthermore, as we focused on linguistic variation prior to 13th century, Kven Finnish and Tornedalen Finnish speaker communities had not yet been formed at that time. Otherwise, the clusters detected were largely in line with classifications suggested mostly in the linguistic literature (e.g. Itkonen 1983, Lehtinen 2007, Kallio 2014), with the exception that Votic and Ingrian were always clustered together.

Votic and Ingrian always clustered together in our analyses, even though it has been suggested that these two have different linguistic origins. Votic is closely related to North Estonian (Sammallahti 1977: 133, Kallio 2014: 162-163), while Ingrian is closely related to Finnish or Karelian (Viitso 2008: 64). Votic and Ingrian were spoken in the same area over several centuries, and Votic has obtained plenty of late loanwords from Ingrian (Itkonen 1983: 215), which explains their common cluster. It is also notable that in some cases Livonian clustered with Votic and Ingrian. This was unexpected, as there appear to be only two features in ALFE that are shared by these three languages and are different elsewhere: 'straight' (map number 95) and 'to lie' (map number 170). The Votic-Ingrian-Livonian group did not appear in the BAPS analysis of the balanced dataset. Therefore, its appearance is likely due to the difficulty of the analysis tools in handling unbalanced data.

Karelian Proper, Livvi-Karelian and Ludian clustered together with both Structure and BAPS, and with both the full and the balanced data. Thus, Karelian Proper, Livvi-Karelian and Ludian could be seen on lexical grounds as dialects of Karelian, rather than separate languages as suggested earlier (Laanest 1975: 26, Sammallahti 1977: 133; Viitso 2008: 64, Laakso 1991: 61-62). Karelian Proper, Livvi-Karelian and Ludian were also often clustered together with Veps. This is in line with the finding that Veps shares a notable amount of lexicon with Karelian and the eastern dialects of Finnish (Kettunen 1960: 25). The strong influence of Veps in Ludian (Kettunen 1960: 214) could also be seen in our results, especially in the Ludian localities of Kuujärvi and Haljärvi, which had the highest amount of admixture with Veps. Of the Livvi-Karelian localities, Vepsian influence could only be seen in the southeasternmost data collection locality – Kardaizet.

In the balanced analysis, when thirty data points were included from both Finland and Estonia, Estonia was divided into the North and South Estonian language areas, while the east-west division within Finnish did not appear. Notable differences between the main dialects have been detected in phonology, morphology and in lexicon in both Estonian (Kettunen 1960: 63) and Finnish (Rapola 1962: 32-105). However, it seems that the differences in this dataset were more pronounced within Estonian than Finnish.

Divisions within Finnish appeared only when the full dataset was analysed. The two-way division of Finnish split the language to east and west, reflecting the main dialect division of Finnish (e.g. Rapola 1962: 98-103). This division has been recently detected quantitatively by clustering the Dialect Atlas of Finnish (Kettunen 1940) using the Structure tool (Syrjänen et al. 2016). The three-way division within Finnish found using ALFE was either into west, Savo and Southeast dialects (e.g. CLUMPP K=5) or into east-west-north (CLUMPP K=6). The east-west-north division corresponds with an earlier three-way division of Finnish obtained with a lexical dataset of the Dictionary of Finnish Dialects (Leino et al. 2006). In the earlier quantitative analysis of the Dialect Atlas of Finnish, the third cluster to appear was that of south-western dialects (Syrjänen et al. 2016: 258). It should be noted, however, that the Dialect Atlas of Finnish is mainly morphophonological whereas our data is lexical, which may explain why this division is less prominent with ALFE.

The lexical features studied here were mostly independent from each other. However, a few features were found to be somewhat linked with each other. These features belonged to certain semantic categories, so that certain names of berries were linked with each other (bog whortleberry, cranberry, lingonberry, raspberry), as were also some colour terms (grey, red, yellow) and some fishing-related terms (oar, cone used in making fishing nets, burbot). When going back to ALFE, these features turned out to have very little variation in the Finnic area in general, meaning that they essentially repeated a similar spatial distribution, and thus had relatively small contribution to the produced clusters.

Atlas Linguarum Fennicarum is a remarkable collection of linguistic variation in the Finnic area, providing possibilities for studying the past of the Finnic area in several different ways. Studying the Finnic area has its challenges as a large amount of linguistic variation has already gone extinct without documentation, for example in the case of Livonian, and due to the large variety in the sizes of the speaker populations. Nevertheless, with the help of quantitative methods, large datasets such as ALFE can be analysed objectively from novel perspectives. While our aim here was to study the suitability of lexical data in detecting the linguistic subdivision of the Finnic dialect continuum, lexicon is only one part of language. Therefore, it is important to combine different types of linguistic evidence in future studies to obtain a more complete view of the linguistic past of Finnic.

Conclusions

We investigated here whether lexical data is suitable for detecting coarse-grained and fine-grained subdivisions of the Finnic group. We found that clusters produced here with quantitative tools and lexical data were largely in line with earlier divisions made with different types of linguistic data. We found support for the north-east – south-west main division, with minor differences compared to the traditional divisions. The estimated number of lexical languages was seven. The largest differences when compared to the traditional division were the substantial differences between North and South Estonian in our results, while the differences between Votic and Ingrian were smaller than assumed. In sum, the similarity between the lexical divisions produced here and the earlier divisions based on different types of data suggest that lexical variation is generally good material for linguistic clustering studies in a dialect continuum –like situations. With this work, we hope to encourage multidisciplinary work combining the vast amounts of linguistic knowledge of Finnic languages with modern quantitative methods in order to shed light on the past of the north-eastern Baltic Sea area.

Acknowledgements

We thank Tiit-Rein Viitso for valuable comments regarding ALFE and the dataset; Kone Foundation for the funding of this project (TH, KS, JS) and Outi Vesakoski and Unni-Päivä Leino, who obtained this funding for the BEDLAN research group; Timo Rantanen and Dmitry Kuznetsov for the language maps; Perttu Seppä for help with the data format.

References

Alvre, P. 1973, Läänemeresoome aluskeele varasest murdeliigendusest, eriti eesti ja soome keelt silmas pidades. – Keel ja Kirjandus 3, 151-162.

Anttikoski, E. 1998, Karjalan kirjakielen suunnittelu 1930-luvulla. – The Baltic-Finnic minorities of the Barents area and the literary language, Tromsø (Nordlyd: Tromsø University Working Papers on Language & Linguistics 26), 118-126.

Anttila, R. 1989, Historical and comparative linguistics. 2 rev. edn., Amsterdam, Philadelphia (Current issues in linguistic theory 6).

Ariste, P. 1965, Läänemere keelte kujunemine ja vanem arenemisjärk. – Sõna sõna kõrvale. Paul Ariste teaduslikust tegevusest, Tallinn (Emakeele seltsi toimetised 7), 80-105.

Bohling, J.H., Adams, J.R., Waits, L.P. 2013, Evaluating the ability of Bayesian clustering methods to detect hybridization and introgression using an empirical red wolf data set. – Molecular Ecology 22, 74-86.

Bowern, C. 2012, The riddle of Tasmanian languages. – Proceedings of the Royal Society B: Biological Sciences 279, 4590-4595.

Bubrih, D.V., Beljakov, A.A., Punzina, A.V., Sarvas, L. 1997, Dialektologičeskij atlas karel'skogo âzyka / Karjalan kielen murrekartasto, Helsinki (Kotimaisten kielten tutkimuskeskuksen julkaisuja 97).

Chambers, J.K., Trudgill, P. 1998, Dialectology. 2nd edn., Cambridge (Cambridge textbooks in linguistics).

Corander, J., Marttinen, P. 2006, Bayesian identification of admixture events using multilocus molecular markers. – Molecular Ecology 15, 2833-2843.

Corander, J., Marttinen, P., Mäntyniemi, S. 2006, A Bayesian method for identification of stock mixtures from molecular marker data. – Fishery Bulletin 104, 550-558.

Corander, J., Marttinen, P., Sirén, J., Tang, J. 2008a, Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. – BMC Bioinformatics 9, 539.

Corander, J., Sirén, J., Arjas, E. 2008b, Bayesian spatial modelling of genetic population structure. – Computational Statistics 23, 111-129.

Corander, J., Waldmann, P., Sillanpää, M.J. 2003, Bayesian analysis of genetic differentiation between populations. – Genetics 163, 367-374.

Dunn, M., Levinson, S.C., Lindström, E., Reesink, G., Terrill, A. 2008, Structural phylogeny in historical linguistics: Methodological explorations applied in island Melanesia. – Language 84, 710-759.

EES: Metsmägi, I., Sedrik, M., Soosaar, S. 2012, Eesti etümoloogiasõnaraamat, Tallinn, www.eki.ee/dict/ety/

Evanno, G., Regnaut, S., Goudet, J. 2005, Detecting the number of clusters of individuals using the software Structure: a simulation study. – Molecular Ecology 14, 2611-2620.

Fjellheim, S., Jørgensen, M.H., Kjos, M., Borgen, L. 2009, A molecular study of hybridization and homoploid hybrid speciation in *Argyranthemum* (Asteraceae) on Tenerife, the Canary Islands. – Botanical Journal of the Linnean Society 159, 19-31.

Francis, R.M. 2017, Pophelper: an R package and web app to analyse and visualize population structure. – Molecular Ecology Resources 17, 27-32.

Frog, Saarikivi, J. 2015, De situ linguarum fennicarum aetatis ferreae, Pars I. – RMN Newsletter 9, 64-115.

Gooskens, C. 2007, The contribution of linguistic factors to the intelligibility of closely related languages. – Journal of Multilingual and Multicultural Development 28, 445-467.

Greenhill, S.J., Wu, C., Hua, X., Dunn, M., Levinson, S.C., Gray, R.D. 2017, Evolutionary dynamics of language systems. – Proceedings of the National Academy of Sciences of the United States of America 114, E8822-E8829.

Grünthal, R. 2015, Vepsän kielioppi, Helsinki (Apuneuvoja suomalais-ugrilaisten kielten opintoja varten 17).

Grünthal, R., Sarhimaa, A. (toim.) 2004/2012, Itämerensuomalaiset kielet ja niiden päämurteet. Suomalais-Ugrilainen Seura.

Hammarström, H., Forkel, R., Haspelmath, M. 2018, Glottolog 3.3. Jena: Max Planck Institute for the Science of Human History, http://glottolog.org, Accessed 25.10.2018.

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

Itkonen, T. 1983, Välikatsaus suomen kielen juuriin. – Virittäjä 87, 190-229.

Jakobsson, M., Rosenberg, N.A. 2007, CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. – Bioinformatics 23, 1801-1806.

Kallio, P. 2007, Kantasuomen konsonanttihistoriaa. – Sámit, sánit, sátnehámit: riepmočála Pekka Sammallahti miessemánu 21. beaivve 2007, Helsinki (Suomalais-Ugrilaisen Seuran Toimituksia 253), 229-249.

Kallio, P. 2014, The Diversification of Proto-Finnic. – Fibula, Fabula, Fact: The Viking Age in Finland, Helsinki (Studia Fennica Historica 18), 155-168.

Kallio, P. 2015a, The stratigraphy of the Germanic loanwords in Finnic. – Early Germanic languages in contact, Amsterdam-Philadelphia (North-Western European Language Evolution Supplement Series 27), 23-38.

Kallio, P., 2015b, The language contact situation in prehistoric Northeastern Europe. – The linguistic roots of Europe: Origin and development of European languages, Copenhagen (Copenhagen studies in Indo-European 6), 77-102.

Kettunen, L. 1930, Suomen murteet: 2 Murrealueet, Helsinki (Suomalaisen Kirjallisuuden Seuran toimituksia 188).

Kettunen, L. 1940, Suomen murteet. 3 A, Murrekartasto, Helsinki (Suomalaisen Kirjallisuuden Seuran toimituksia 188).

Kettunen, L. 1960, Suomen lähisukukielten luonteenomaiset piirteet, Helsinki (Suomalais-Ugrilaisen Seuran toimituksia 119).

Koponen, E. 1991, Itämerensuomen marjannimistön kehityksen päälinjoja ja kantasuomen historiallista dialektologiaa. – Journal de la Société Finno-Ougrienne 83, 123-161.

Laakso, J. 1991, Itämerensuomalaiset sukukielemme ja niiden puhujat. – Uralilaiset kansat: tietoa suomen sukukielistä ja niiden puhujista, Porvoo-Helsinki-Juva, 49-122.

Laakso, J. 2001, The Finnic languages. – The Circum-Baltic languages: Volume 1, Past and present: typology and contact, Amsterdam–Philadelphia (Studies in language companion series 54), 179-215.

Laanest, A. 1975, Sissejuhatus läänemeresoome keeltesse, Tallinn, Keele ja kirjanduse instituut.

Laanest, A., Jussila, R. 1989, Itämerensuomalainen kielikartasto: kyselysarja, Helsinki, Kotimaisten Kielten Tutkimuskeskus.

Lang, V. 2018. Läänemeresoome tulemised, Tartu (Muinaisaja Teadus 28).

Lehtinen, T. 2007, Kielen vuosituhannet: suomen kielen kehitys kantauralista varhaissuomeen, Helsinki (Tietolipas 215).

Leino, A., Hyvönen, S., Salmenkivi, M. 2006, Mitä murteita suomessa onkaan? Murresanaston levikin kvantitatiivista analyysiä. – Virittäjä 1, 26-45.

Neophytou, C. 2014, Bayesian clustering analyses for genetic assignment and study of hybridization in oaks: effects of asymmetric phylogenies and asymmetric sampling schemes. – Tree Genetics & Genomes 10, 273-285.

Ojansuu, H. 1922, Itämerensuomalaisten kielten pronominioppia, Helsinki (Turun suomalaisen yliopiston julkaisuja, Sarja B, osa 1 nro 3).

Pahomov, M. 2017, Lyydiläiskysymys: kansa vai heimo, kieli vai murre? Helsinki (PhD thesis).

Pritchard, J.K., Stephens, M., Donnelly, P. 2000, Inference of population structure using multilocus genotype data. – Genetics 155, 945-959.

Pritchard, J.K., Wen, X., Falush, D. 2010. Documentation for Structure software. Version 2.3.

Puechmaille, S.J. 2016, The program Structure does not reliably recover the correct population structure when sampling is uneven: subsampling and new estimators alleviate the problem. – Molecular Ecology Resources 16, 608-627.

Puza, B. 2015, Bayesian Methods for Statistical Analysis, Australian National University eView (Free online access: OAPEN).

Rapola, M. 1962, Johdatus Suomen murteisiin, 2 rev. edn., Turku (Tietolipas 4).

Raun, A. 1971, Essays in Finno-Ugric and Finnic linguistics, The Hague (Indiana University publications, Uralic and Altaic Series 107).

Reesink, G., Singer, R., Dunn, M. 2009, Explaining the linguistic diversity of Sahul using population models. – PLoS Biology 7, e1000241.

Saareste, A. 1938, Eesti murdeatlas. Atlas des parlers estoniens. I vihik, Tartu.

Saareste, A. 1941, Eesti murdeatlas. Atlas des parlers estoniens. II vihik. Tartu.

Saareste, A. 1955, Petit Atlas des parlers estoniens = Väike eesti murdeatlas, Uppsala (Skrifter utgivna av Kungl. Gustav Adolfs akademien 28).

Salminen, T. 1998, Pohjoisten itämerensuomalaisten kielten luokittelun ongelmia. – Oekeeta asijoo. Commentationes Fenno-Ugricae in honorem Seppo Suhonen sexagenarii, Helsinki, (Suomalais-ugrilaisen seuran toimituksia 228), 390-406.

Salminen, T. 2007, Europe and North Asia. – Encyclopedia of the world's endangered languages, London, 211-282.

Sammallahti, P. 1977, Suomalaisten esihistorian kysymyksiä. – Virittäjä 81, 119-136.

Setälä, E.N. 1916, Suomensukuisten kansojen esihistoria. – Maailmanhistoria II, Helsinki, 476-516.

Simons, G.F., Fennig, C.D. 2018, Ethnologue: Languages of the World. Twenty-first edition. Dallas, Texas. www.ethnologue.com, Accessed 25.10.2018.

SSA: Itkonen, E., Kulonen, U-M. 1992-2000, Suomen Sanojen Alkuperä 1-3, Helsinki (Suomalaisen Kirjallisuuden Seuran toimituksia 556; Kotimaisten kielten tutkimuskeskuksen julkaisuja 62).

Sulkala, H. 2010, Introduction: Revitalisation of the Finnic minority languages. – Planning a new standard language: Finnic minority languages meet the new millennium. (Studia Fennica Linguistica 15), 8-26.

Syrjänen, K., Honkola, T., Lehtinen, J., Leino, A., Vesakoski, O. 2016, Applying population genetic approaches within languages: Finnish dialects as linguistic populations. – Language Dynamics and Change 6, 235-283.

Thomason, S. 2000, Linguistic areas and language history. – Languages in contact, Amsterdam (Studies in Slavic and general linguistics 28), 311-327.

Tuomi, T., Hänninen, A., Suhonen, S. 2004, Atlas Linguarum Fennicarum 1, Helsinki (Suomalaisen Kirjallisuuden Seuran toimituksia 800, Kotimaisten kielten tutkimuskeskuksen julkaisuja 118).

Tuomi, T., Hänninen, A., Viitso, T. 2007, Atlas Linguarum Fennicarum 2, Helsinki (Suomalaisen Kirjallisuuden Seuran toimituksia 800, Kotimaisten kielten tutkimuskeskuksen julkaisuja 118).

Tuomi, T., Hänninen, A., Rjagojev, V. 2010, Atlas Linguarum Fennicarum 3, Helsinki (Suomalaisen Kirjallisuuden Seuran toimituksia 1295, Kotimaisten kielten tutkimuskeskuksen julkaisuja 159).

Turunen, A. 1988, The Balto-Finnic languages. – The Uralic languages: description, history and foreign influences, Leiden, 58-83.

Viitso, T. 1998. Fennic, – The Uralic languages, London (Routledge language family descriptions), 96-114.

Viitso, T. 2008, Läänemeresoome murdeliigenduse põhijooned, Tartu–Tallinn – Liivi keel ja läänemeresoome keelemaastikud, 63-69.

Weijnen, A. 1975, Atlas Linguarum Europae (ALE): Introduction, Assen.

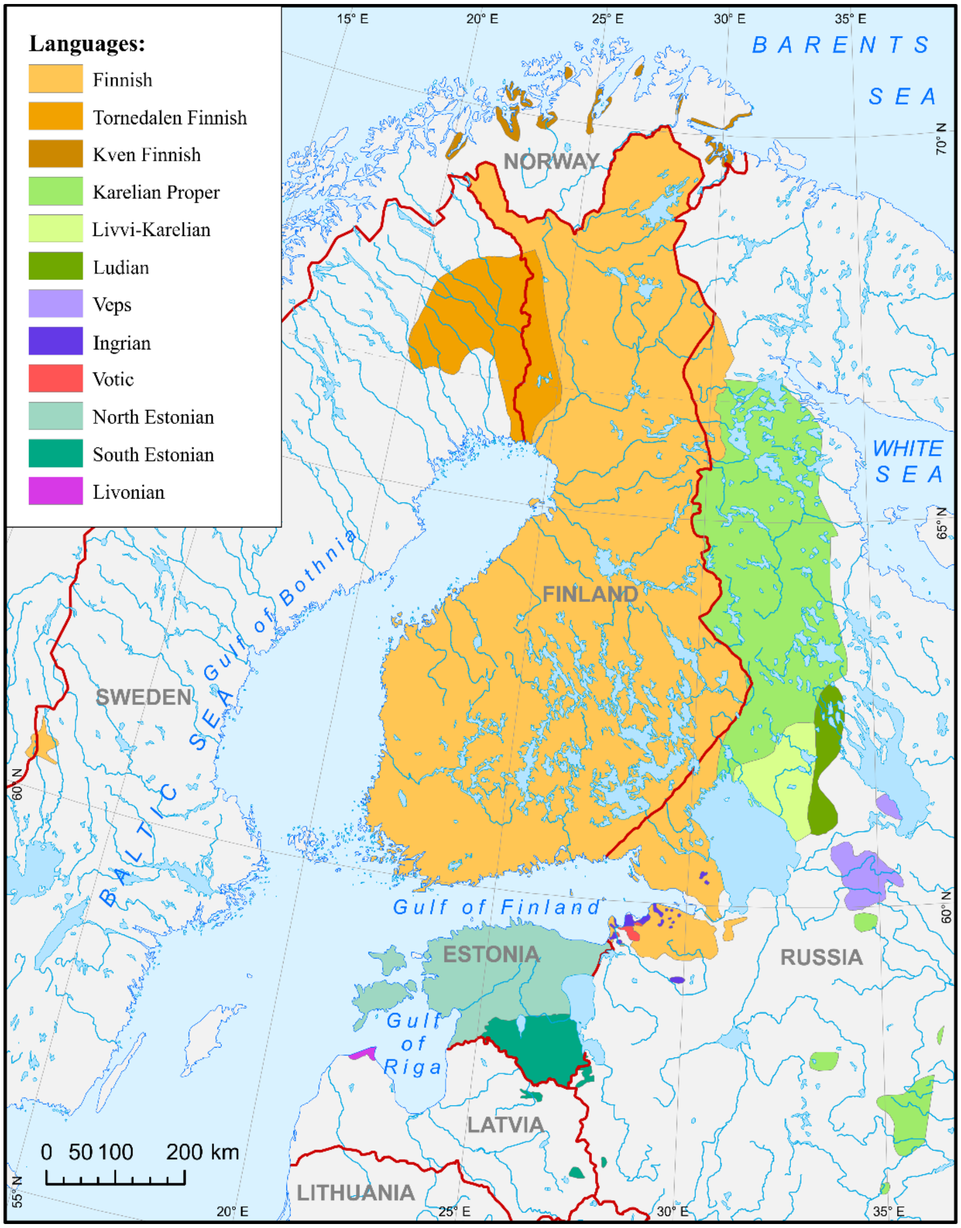
Languages:
Finnish
Tornedalen Finnish
Kven Finnish
Karelian Proper
Livvi-Karelian
Ludian
Veps
Ingrian
Votic
North Estonian
South Estonian
Livonian
BARENTS SEA
WHITE SEA
NORWAY
FINLAND
SWEDEN
RUSSIA
ESTONIA
LATVIA
LITHUANIA
Gulf of Bothnia
BALTIC SEA
Gulf of Finland
Gulf of Riga
0 50 100 200 km
15° E
20° E
25° E
30° E
35° E
70° N
65° N
60° N
55° N

Figure 1.

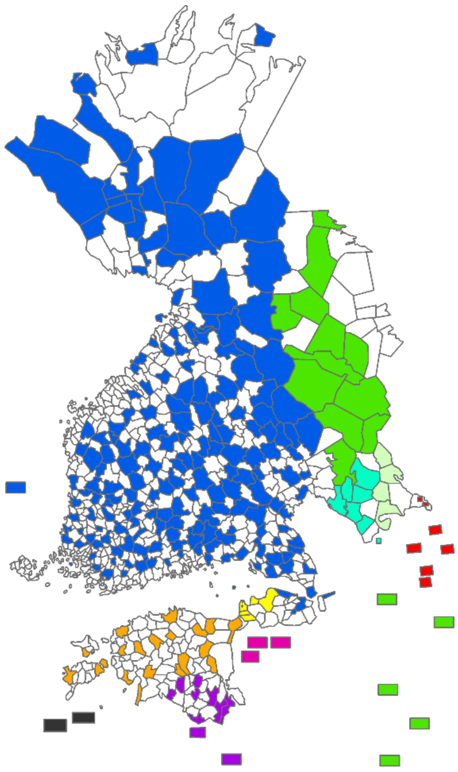

Figure 2.

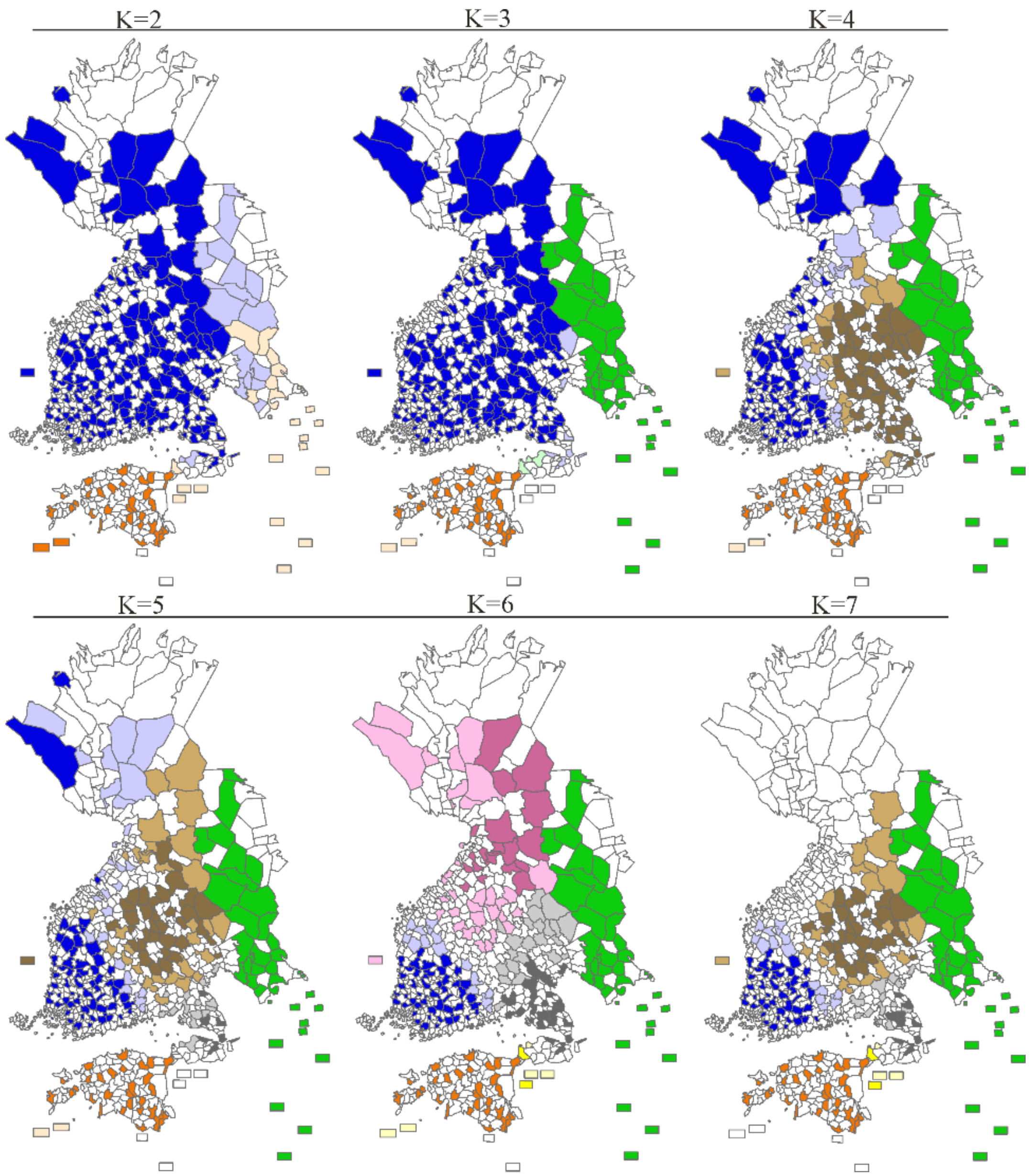

Figure 3.

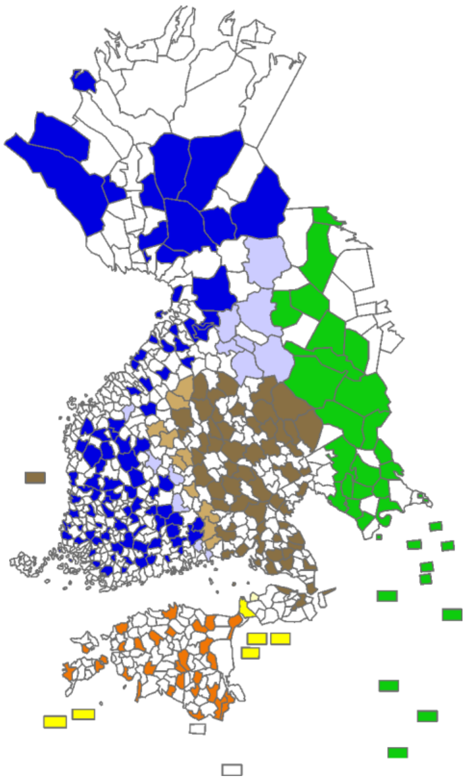

Figure 4.

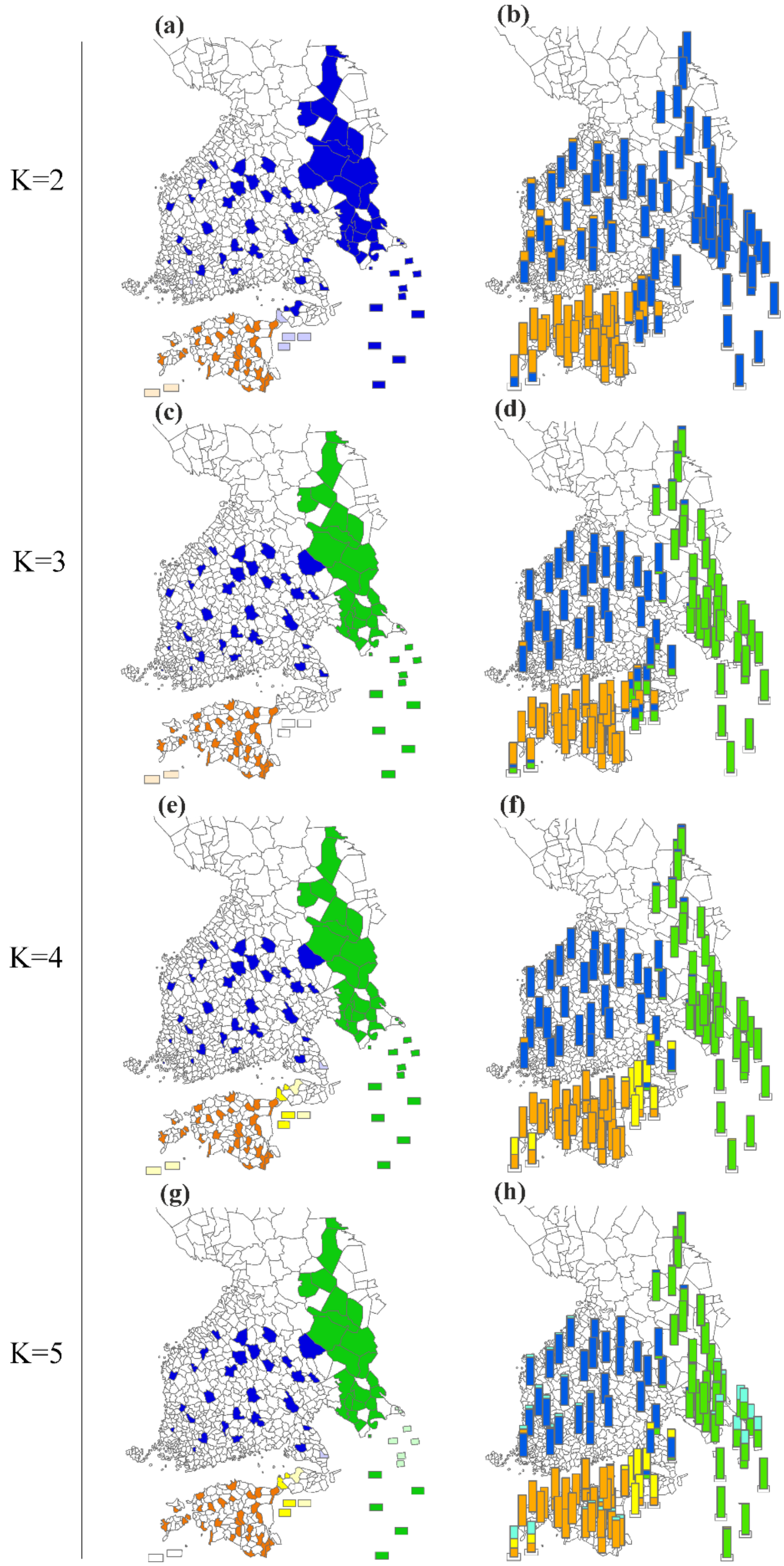

Figure 5.

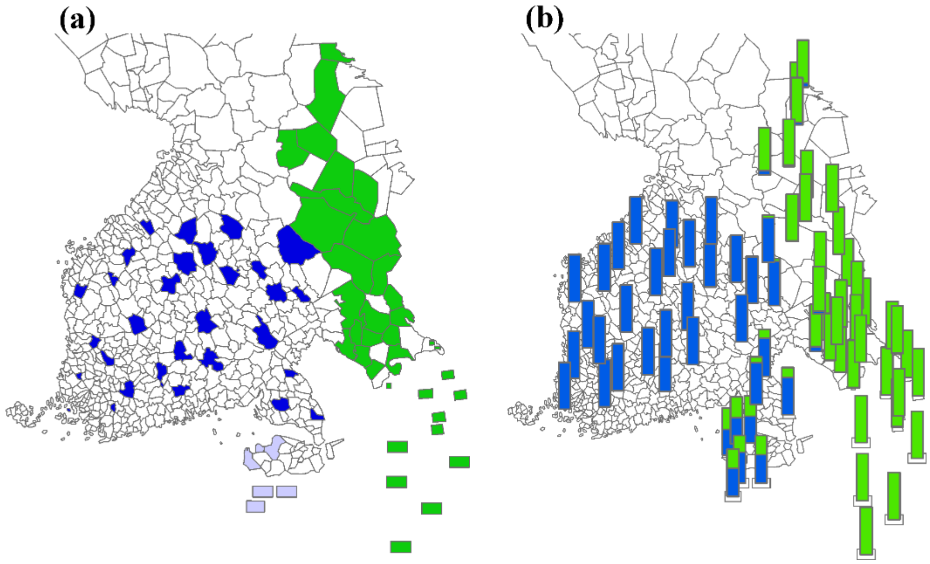

Figure 6.

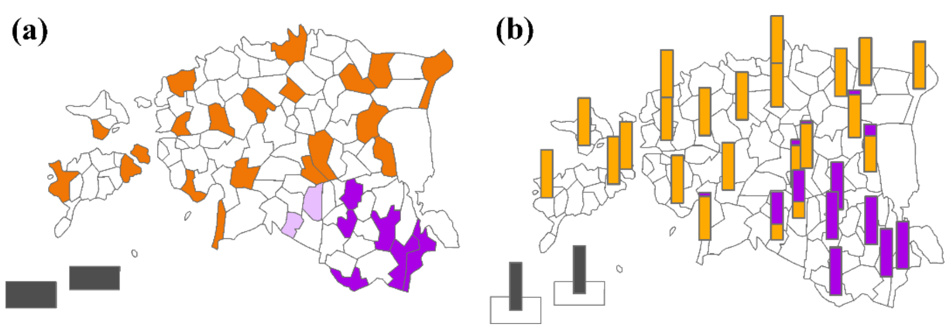

Figure 7.

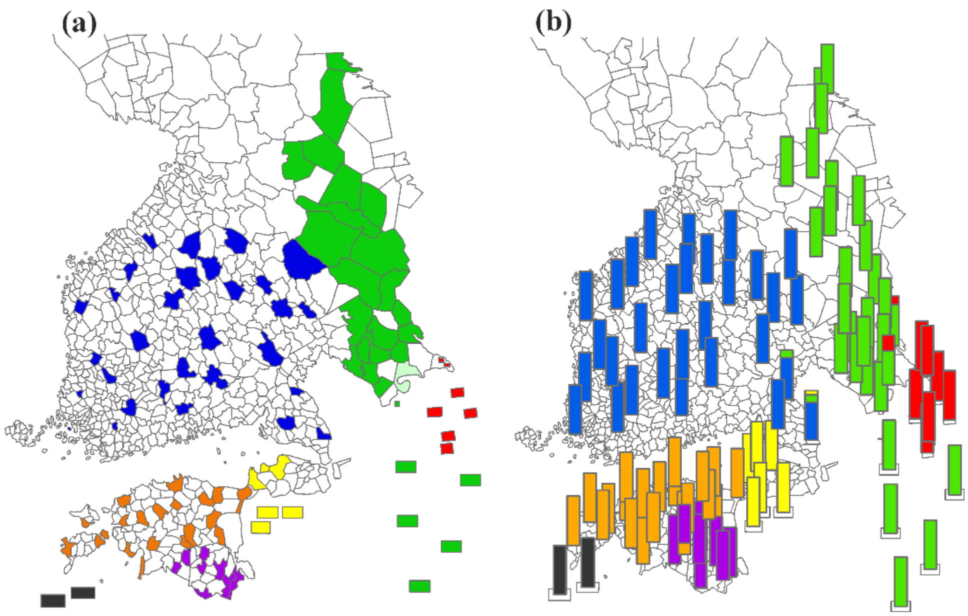

Figure 8.

Figure 1. Twelve Finnic languages according to the classification in Glottolog 3.3 (Hammarström et al. 2018) and Ethnologue (Simons, Fennig 2018). Speaker areas were drawn by BEDLAN / Timo Rantanen & Dmitry Kuznetsov, based on Grünthal and Sarhimaa (2004/2012) and Tuomi et al. (2004: 14).

Figure 2. Finnic languages according to ALFE and their main data collection localities. Blue = Finnish; dark green = Karelian Proper; turquoise = Livvi-Karelian; light green = Ludian; red = Veps; yellow = Ingrian; pink = Votic; orange = North Estonian; purple = South Estonian; black = Livonian. Rectangles represent data collection localities which are not in their exact geographical location either due to cartographic reasons, such as the blue rectangle outside of mainland Finland, representing the Forest Finns, who are situated in Sweden, or due to small or overlapping speaker areas, such as Votic and Ingrian.

Figure 3. Dialect divisions summarised by CLUMPP for K=2-7. Two shades of colour are used to separate the core (membership coefficient values of 0.75-1) and transitional (0.5-0.75) areas of the languages or dialects. The colours in K=6 represent the following languages or dialects: blue = Western Finnish; pink = Northern Finnish; grey = Eastern/Southeast Finnish; green = Karelian-Veps; orange = Estonian; yellow = Ingrian-Votic-Livonian. Localities coloured white in K=2 were not included in the analyses. Notably, localities that are coloured in K=2 but white in the other maps are data points with strongly mixed ancestry, with all the membership coefficients below 0.5.

Figure 4. Most supported division (K=5) based on the full dataset BAPS analysis. For colour-coding, see Figure 3 with the addition of brown = Eastern Finnish.

Figure 5. CLUMPP divisions of the balanced analyses for K=2-5. a) K=2 clusters b) visualised as bar plots, c) K=3 clusters d) visualised as bar plots, e) K=4 clusters f) visualised as bar plots, g) K=5 clusters h) visualised as bar plots. For the colour-coding, see Figure 3.

Figure 6. CLUMPP K=2 a) clusters and b) bar plots for the north-eastern part of the balanced data

Figure 7. CLUMPP K=3 a) clusters and b) bar plots for the south-western part of the balanced data.

Figure 8. BAPS division of the balanced data to seven clusters visualised as a) clusters and b) bar plots. For colour-coding, see Figures 2 and 3.