BRILL

LANGUAGE DYNAMICS AND CHANGE 4 (2014) 189–221

brill.com/ldc

Behind Family Trees

Secondary Connections in Uralic Language Networks

Jyri Lehtinen
University of Helsinki
jyri.lehtinen@helsinki.fi

Terhi Honkola
University of Turku
terhi.honkola@utu.fi

Kalle Korhonen
University of Helsinki
kalle.korhonen@helsinki.fi

Kaj Syrjänen
University of Tampere
kaj.jaakko.syrjanen@uta.fi

Niklas Wahlberg
University of Turku
niklas.wahlberg@utu.fi

Outi Vesakoski
University of Turku
outi.vesakoski@utu.fi

Abstract

Although it has long been recognized that the family tree model is too simplistic to account for historical connections between languages, most computational studies of language history have concentrated on tree-building methods. Here, we employ computational network methods to assess the utility of network models in comparison with tree models in studying the subgrouping of Uralic languages. We also compare basic vocabulary data with words that are more easily borrowed and replaced cross-linguistically (less basic vocabulary) in order to find out how secondary connections

 | DOI: 10.1163/22105832-00402007

affect computational analyses of this language family. In general, the networks support a treelike pattern of diversification, but also provide information about conflicting connections underlying some of the ambiguous divergences in the trees. These are seen as reflections of unclear divergence patterns (either in ancestral protolanguages or between languages closely related at present), which pose problems for a tree model. The networks also show that the relationships of closely related present-day languages are more complex than what the tree models suggest. When comparing less basic with basic vocabulary, we can detect the effect of borrowing between different branches (horizontal transfer) mostly between and within the Finnic and Saami subgroups. We argue that the trees obtained with basic vocabulary provide the primary pattern of the divergence of a language family, whereas networks, especially those constructed with less basic vocabulary, add reality to the picture by showing the effect of more complicated developments affecting the connections between the languages.

Keywords

computational phylogenetics – phylogenetic networks – meaning lists – language evolution – family tree model – secondary connections

1 Introduction

The family tree model has been the most widely used way of illustrating historical connections between related languages both in traditional historical linguistics and in quantitative analyses (e.g., Swadesh, 1950, 1952, 1955; Ringe et al., 2002, and Gray and Atkinson, 2003 for Indo-European; Gray and Jordan, 2000 for Austronesian; Holden, 2002 for Bantu; Walker and Ribeiro, 2011 for Arawak; and recently Syrjänen et al., 2013 and Honkola et al., 2013 for the Uralic languages). This convention has been challenged on the grounds that the tree model only tells a limited story of the historical relationships of languages (e.g., Bloomfield, 1933: 311–318; Aikhenvald and Dixon, 2001: 4–9; McMahon and McMahon, 2005). The family tree model expresses the divergence history of languages by showing successive separation of subgroups based on shared innovations. However, languages are also connected by secondary connections that can be caused by language contact, such as borrowing, or innovations that have not spread through the entire language community, producing wavelike development.

In the tree models, the conflicting connections cannot be illustrated properly, but their existence can be inferred if the model does not yield a clear

branching order (e.g. in Syrjänen et al., 2013). Wave and network models have been developed to illustrate the secondary connections. Johannes Schmidt (1872) formulates the wave model as an alternative to the family tree model (Schleicher, 1861), which cannot take into account the innovations that have spread through language communities of a language family only partially or have spread across the boundaries of different language communities. Schmidt proposes that the relationships of languages should rather be expressed through overlapping waves, which would indicate characteristics that do not define clear-cut subgroups of the family.

These kinds of conflicting characteristics can be illustrated computationally with phylogenetic networks (Fitch, 1997). So far, in linguistics, phylogenetic networks have been used both in assessing historical connections of languages using structural features (e.g., Dunn et al., 2008 for languages of Melanesia) and for illustrating secondary connections between languages of well-known families (cf. Nelson-Sathi et al., 2011 on the Indo-European languages). The network models allow us to assess how "treelike" the historical connections are and where conflicting connections are located (Bryant and Moulton, 2003; Heggarty et al., 2010). Thus, networks can also visualize the effects of wavelike development and the effect of contact after divergence, and the use of both networks and trees provides a more comprehensive picture of the evolution of language families than trees alone.

The history of the Uralic language family has been thoroughly researched with traditional methods (for in-depth overviews of the linguistic and historical aspects of the different languages, see Sinor, 1988 and Abondolo, 1998b). While many of its subgroups have been firmly established, the mutual relationships of these subgroups have remained somewhat unclear. The ambiguities have led to the disbandment of the strictly binary branching family tree of the Uralic languages (e.g. Korhonen, 1981) in favor of highly polytomous (or "bushlike"; Häkkinen, 1984; Salminen, 1999, 2002) or somewhat polytomous illustrations of the relationships of the Uralic languages (Kulonen, 2002; Michalove, 2002; also in recent quantitative studies by Syrjänen et al., 2013 or Honkola et al., 2013).

More specifically, although Syrjänen et al. (2013) and Honkola et al. (2013) found clear support for the Finno-Ugric, Ob-Ugric and Finno-Saami branches among the contested subgroups (for language groupings, see Section 2.3), the further divergence of Finno-Ugric remained unclear across the board. In particular, in Syrjänen et al. (2013), the positions of Hungarian and Mari were uncertain, as their placement varied according to the data used. In this study we use networks in order to investigate the alternative connections between languages, which may explain their hitherto unresolved relationships in the Uralic family.

In quantitative approaches to language subgrouping, the focus has been on finding innovations in inherited word forms. This is done most often by using basic vocabulary: the most stable part of lexicon that is resistant to borrowings and, therefore, might be assumed to represent the history of languages without the confounding effect of secondary contacts. Basic vocabulary collected in meaning lists such as the Swadesh lists (see Section 2.1) actually includes borrowings and other items that do not optimally correspond to the criteria of basic vocabulary; however, the actual impact of the borrowings on the interpretation of the results has proven not to be very significant. McMahon and McMahon (2005) and Greenhill et al. (2009) found that data can bear 10–20% borrowings before the shape of the phylogeny is affected. This was seen also with Uralic languages, as the borrowings within the standard basic vocabulary lists did not confound the tree models (Syrjänen et al., 2013).

As the connections resulting from borrowings form an inseparable part of the histories of language families, they should not be neglected, but studied as a part of the evolutionary history of a language family. However, there are no studies experimentally testing the differences between using basic vocabulary and data carrying more information about borrowings in studying the historical connections within a language family. We suggest, therefore, that by taking into account the information provided by "less basic vocabulary" (i.e., words that are more prone to borrowing and replacement), and by exploring the alternative connections with network models, we should obtain a more diverse and realistic picture of language relationships.

To test this approach, we studied the evolutionary pattern of the Uralic language family, focusing especially on the vague position of Hungarian and Mari. We collected word lists for the languages, using meaning lists representing both basic and less basic vocabulary, and analyzed them with both phylogenetic tree and network methods. With the aid of less basic vocabulary and networks, we expect to expand our understanding of the history of the Uralic languages.

In the following section we provide justification for the use of meaning lists as a basis for language classification based on lexical data, and move on to the data and methods used for their analysis. Then, we present the results and discuss issues arising from their interpretation, and end with conclusions on Uralic language history and the general applicability of the methodology combining trees and networks, as used here.

2 Materials and Methods

2.1 *Justifying the Use of Meaning Lists*

Basic vocabulary, in the sense used here, should fulfill four criteria. It needs to be 1) resistant to borrowing, as the pattern of diversification is by definition dependent on lexical innovations and the transmission of those innovations within language communities, 2) historically stable (i.e., unlikely to be replaced) to ensure that the data represent ancient divergences, 3) universal as concepts, so data can be collected from any language, and 4) morphologically simple, because compounds and derivatives could make the data analysis equivocal (Tadmor, 2009).

Basic vocabulary data are obtained with meaning lists: words corresponding to each meaning are collected from all the languages studied, and the researcher then determines whether the words originate from the same historical protoforms. This can be done either by following etymological criteria, connecting words that originate in common ancestral forms (i.e. cognates), or by connecting words that share any kind of common origin, even borrowings from a common source (these are sometimes called "correlates" as a distinction from strict cognates, see McMahon et al., 2005; Heggarty, 2010). It should be noted that only those historical connections are considered that are identified among the meanings of the basic vocabulary list, so cognate words retained in a changed meaning are not considered. Essentially, the resulting basic vocabulary data identify lexical innovations on the basis of both vocabulary replacement and semantic change.

It has sometimes been argued that this meaning-based approach to data collection, excluding cognates that now have a different meaning, is not valid, as not all etymological cognates of words in the data are included. However, alternative approaches that connect language groups based on shared items of ancestral vocabulary are in conflict with a fundamental requirement in subgrouping. One of the basic rules of both biological phylogenetics (Wiley, 1981; Hennig et al., 1996) and linguistic subgrouping (e.g., Campbell, 1998) is that the branches of a tree are to be defined on the basis of shared innovations. This requirement is not met when considering only shared retentions of ancestral traits, i.e. reconstructed vocabulary.

Pusztay (1995), for instance, studies the connections of certain Uralic branches based on the number of shared items retained from the reconstructed vocabulary of the common ancestral protolanguage, using the etymological dictionary UEW (Rédei, 1988). This perspective produces results with striking differences to both traditional approaches (cf. both the treelike and bushlike views mentioned in the introduction) as well as computational analyses of basic vocabulary.

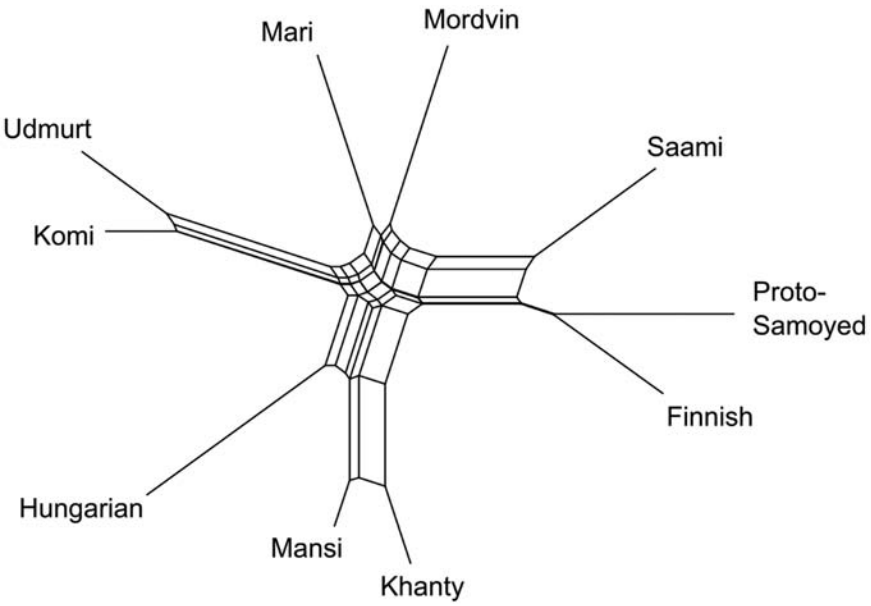

FIGURE 1 *Distances of the Uralic subgroups calculated from shared retentions of reconstructed Proto-Uralic vocabulary, based on data from Sammallahti (1988)*

We evaluate the approach of using shared ancestral protoforms by presenting the distances of the Uralic languages based on retention of reconstructed Proto-Uralic vocabulary in the etymological word list of Sammallahti (1988: 536–541) in a NeighborNet network (Fig. 1; see Section 2.5 for the method). The results are, again, notably different from both traditional and phylogenetic analyses. Strikingly, there was a close connection between Samoyed and Saami and especially Finnish. Instead of arguing that this constitutes evidence for Finnic and Saami diverging close to Samoyed, the most probable reason could be the conservative nature of Finnic and Saami (cf. Janhunen, 1981). This conservatism causes them to be positioned alongside Samoyed, the first subgroup to have diverged from Proto-Uralic (according to the traditional paradigm, see e.g. Korhonen, 1981; Michalove, 2002; Janhunen, 2009).

To sum up, using shared retention of ancestral traits does not provide a reliable alternative to the approach based on meaning lists. In the phylogenetic methodology, the subgroups have to be defined in terms of shared innovations, and this is what the basic vocabulary meaning lists aim at: when only those words are included in the data that are used for the required meanings, connections are formed between languages based on innovations in the combinations of form and function. A further important strategy to avoid the

effect of shared retentions is to use an outgroup for rooting that includes the features determined to be retentions from the common protolanguage (this was done in Syrjänen et al., 2013; see Section 2.4). By doing this, splits that the languages undergo after divergence from the protolanguage are determined by innovations that remove the languages from the ancestral traits included in the outgroup. We thus find it justified to utilize relationships between words based on meaning lists in studying the historical connections between languages, especially when the retentions in the data are determined by rooting.

2.2 *Meaning Lists Used in this Study*

The basic vocabulary lists used here are the same as in the previous phylogenetic study of Uralic diversification by Syrjänen et al. (2013): the 200-item Swadesh list (Swadesh, 1952), the later 100-item version (Swadesh, 1955; 7 meanings not on the 200-item list), and the 100-item Leipzig-Jakarta list[1] (Tadmor, 2009; 19 meanings not found in either of the Swadesh lists). Together, the Swadesh lists and the Leipzig-Jakarta list include 226 basic vocabulary meanings (Full Basic Vocabulary). From these, Syrjänen et al. (2013) separated the meanings that contained no attested borrowings in any of the Uralic languages and ended up with exactly 100 such meanings, resulting in a list called "Ura100" (for a more detailed description, see ibid.).

To study the effect of borrowings on the branching pattern of the Uralic languages, we compiled a new meaning list with less stable vocabulary. This list was put together from the data in the World Loanword Database (WOLD; Haspelmath and Tadmor, 2009b), the same used for creating the Leipzig-Jakarta list. We collected the meanings that were ranked from 401st to 500th based on their composite score (see Tadmor, 2009) and named this list the "WOLD401–500" list.

The meaning lists used for the separation of datasets in this study are summarized below.

– Full Basic Vocabulary (Swadesh 100 + Swadesh 200 + Leipzig-Jakarta combined; 226 meaning items)

1 One meaning in the original 100-item list, 'foot/leg,' was divided into two, 'foot' and 'leg', resulting in 101 items in the list used here, as in Syrjänen et al. (2013). The Leipzig-Jakarta list (Tadmor, 2009) was compiled from the 100 highest-ranked meanings in terms of criteria for basic vocabulary in the Loanword Typology Project (Haspelmath and Tadmor, 2009a).

– Ura100 (100 items)
– Leipzig-Jakarta (WOLD1–100; 101 items due to the splitting of original 'foot/leg' into two items)
– WOLD401–500 (100 items)

The datasets collected on the basis of these meaning lists contain gaps and multiple words for single meaning items for all languages in the sample. Therefore, the item counts for these word lists do not equal the number of items on the corresponding meaning lists. This is true especially for the WOLD401–500 dataset because the WOLD 401–500 meaning list contains many non-universal meanings, for which many Uralic languages lack words.

2.3 *The Collection of Lexical Data*

The word lists were collected for 18 Uralic languages, 17 of which were also included in Syrjänen et al. (2013) and in Honkola et al. (2013). For the present investigation, the dataset was expanded to cover Kildin Saami. Figure 2 shows a map of the languages studied, listed below in their respective lower-level subgroups.[2]

– *Saami languages*: Ume Saami, North Saami, Skolt Saami, Kildin Saami
– *Finnic languages*: Finnish, Karelian (Karelian Proper), Veps, Estonian, Livonian
– *Mordvin languages*: Erzya
– Mari (Meadow Mari)
– *Permian languages*: Komi (Komi-Zyrian), Udmurt
– Hungarian
– *Ob-Ugric languages*: Northern Mansi, Eastern Khanty
– *Samoyed languages*: Tundra Nenets, Selkup

Support for several higher-level groupings of these comes from both traditional historical study and computational subgrouping analysis (c.f. Syrjänen et al., 2013):

2 Subgroups are given according to the traditional view followed e.g. in the classification by Otto Donner (cf. Hovdhaugen et al., 2000: 178–179), but disregarding the "Volgaic" group that is considered to include Mari and Mordvin (Itkonen, 1997; Michalove, 2002). Our "Karelian" pertains to words common in both the Northern and Southern dialectal areas of Karelian Proper, as separate from Livvi (Olonetsian) and Ludian.

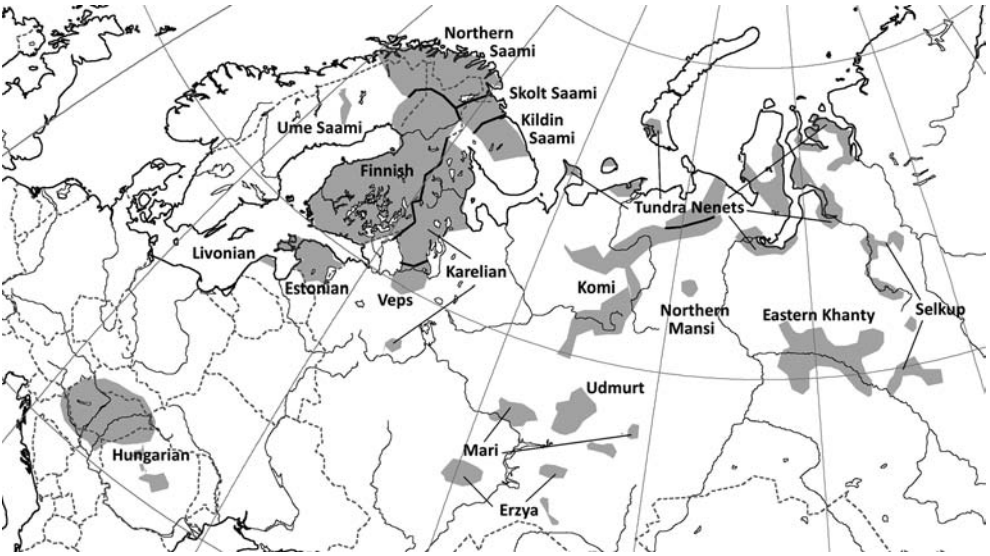

FIGURE 2 *Map of the Uralic languages included in the study*

- *Finno-Ugric*: Finnic + Saami + Mordvin + Mari + Permian + Hungarian + Ob-Ugric, i.e. all of the languages remaining after an initial separation of Samoyed
- *Ugric*: Hungarian + Ob-Ugric
- *Finno-Permian*: Finnic + Saami + Mordvin + Mari + Permian
- *Finno-Volgaic*: Finnic + Saami + Mordvin + Mari
- *Finno-Mordvin*: Finnic + Saami + Mordvin
- *Finno-Saami*: Finnic + Saami

The procedure for data collection of the basic vocabulary included here is described in Syrjänen et al. (2013). The additional words required for the WOLD401–500 list were collected from the same sources, whereas the entire data for Kildin Saami were taken from the World Loanword Database (Haspelmath and Tadmor, 2009b), because Kildin Saami was one of the source languages of the Loanword Typology Project (Haspelmath and Tadmor, 2009a).

Likewise, the historical connections between words were determined from the same sources as in Syrjänen et al. (2013). All words used in the same meaning, originating either from a common protolanguage and inherited through language transmission (according to etymological sources) or from borrowing from the same source, were assigned to the same correlate set. The use of correlates, allowing for looser connections than cognates (cf. McMahon et al., 2005),

was motivated by the possibility that the etymology of each word cannot be determined with certainty in all cases, and by the aim to show the connections caused by borrowing.

The final datasets were coded as binary matrices for both the phylogenetic tree analyses and the computation of networks. The data were converted into binary characters that correspond to etymological relationships, with the meanings belonging to a given correlate set marked as 1, meanings not belonging to a given correlate set marked as 0, and missing words (i.e., the meanings whose presence or absence in a language could not be ascertained from the references) marked with a question mark.

2.4 *Language Trees with Phylogenetic Methods*

The trees were calculated with a model-based Bayesian phylogenetic algorithm, implemented in the MrBayes software (version 3.2.1; Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003), which uses a Markov Chain Monte Carlo (MCMC) method for producing a distribution of trees that aims to approximate the most likely model of divergence reflected in the input data. A more detailed discussion of the technical background can be found in Syrjänen et al. (2013).

The analyses were run for 1 million generations with every 1,000th tree sampled, and the first 100 trees of the resulting sample discarded as "burn-in" (because the algorithm starts out with a random tree shape, it is to be expected that the early part of the sample is not representative of a likely tree). A Markov K substitution model was used for the computations, with an assumption of equal base frequencies and an equal probability of a change of character state in either direction. From the analysis of each dataset performed in this way, a consensus tree was produced from the final sample—a tree diagram that shows the branching pattern best represented by the sample. If a branching point was found in more than 50 % of the sampled trees, it was included in the consensus tree.

Syrjänen et al. (2013) and Honkola et al. (2013) infer the location of the first divergence with different approaches. Syrjänen et al. (2013) use a reconstructed Proto-Uralic language[3] as an outgroup for rooting, whereas in Honkola

3 In contrast to the traditional criterion in the study of Uralic languages, which accepts a reconstructed form as Proto-Uralic when its etymological descendants are found in both the Samoyed and Finno-Ugric branches, in Syrjänen et al. (2013) a Proto-Uralic etymology was included only when it was represented in both the Ugric and Finno-Permian subgroups of

et al. (2013), the initial divergence is inferred directly from the data by the BEAST software (Drummond and Rambaut, 2007). Both approaches consistently result in an initial split between Samoyed and Finno-Ugric, which is in agreement with traditional views (reviewed in Syrjänen et al., 2013). We employed the same datasets for both tree analyses and networks, and because no rooting point could be used for the networks, no Proto-Uralic reconstruction was included for rooting in the present study. Instead, we rooted the trees manually by setting the first divergence between the Samoyed languages (Tundra Nenets and Selkup) and the others, using the FigTree software (version 1.3.1).

In the consensus trees produced here, the branching points are marked with posterior probability values indicating the proportion of trees in the final sample that include the branching point in question. The values range from 0.5 (found in 50 % of the trees) to 1.0 (100 %), and, as a general guideline (following e.g. Huelsenbeck et al., 2001), values above 0.95 are considered to have very good support. As values slightly below this threshold are often noteworthy as well, we take posterior probability values over 0.90 into consideration as being tentatively supported by the data. Branches with lower values are regarded as collapsed, i.e., the branch is considered absent and its constituent branches are attached to the well-supported node below it, forming a polytomous branch (a branch splitting into three or more sub-branches). Moreover, branch lengths help in interpreting the support the trees provide for a given branching. Branch lengths essentially reflect the relative amount of change that has taken place along each branch, and thus long branches suggest higher rates of evolution than short ones, which have less data to support them.

2.5 *Using Network Methods in Establishing Secondary Connections between Languages*

Phylogenetic trees show the modeled divergence history as a sequence of discrete splits between different branches: they are *compatible* collections of splits, showing no connections in their structure that would conflict with other connections (with conflicting signals being implied only by low posterior probabilities). In contrast, network models can take into account connections between languages that do not accord with a strictly treelike model, and represent them directly in their graphical output. They can thus also represent *incompatible* collections of splits, where languages can be grouped by

Finno-Ugric, in addition to Samoyed. This was done so that the prior assumption of the initial binary split between Samoyed and Finno-Ugric, sometimes contested in the literature, would not cause unwarranted bias in the rooting.

crosscutting connections. Conflicting connections between languages and subgroups can be the result of borrowing, where material is transferred from one branch to another one that has split off earlier (see, e.g., Nelson-Sathi et al., 2011). Also, wavelike diffusion of innovations during the divergence of protolanguages, causing crosscutting patterns between offspring languages, can make the tree model alone unable to accurately represent the process of divergence (Heggarty et al., 2010).

Several computational network methods have been developed for evaluating conflicting connections in evolutionary histories. Some of them are character-state-based, in the same way as the Bayesian MCMC tree algorithm that we used for our tree models (cf. Section 2.4), and some are distance-based, like the NeighborNet algorithm we used in the network analyses. Character-based methods model connections by taking into account each "locus" (in our case, correlate set) at a time in the data. A prominent example, also used in linguistic studies, is the Network algorithm (Bandelt et al., 1995, 1999; Forster et al., 1998, 2006; Forster and Toth, 2003), which depicts alternative evolutionary paths detected in the data as reticulations. In contrast to the Bayesian MCMC tree algorithm employed for our tree models, Network is sensitive to undetected borrowings, whose coding may greatly influence the position of individual languages (McMahon and McMahon, 2005: 149–154).

In distance-based methods, the models are calculated based on distance matrices that record the amount of shared items in the data between pairs of languages, instead of going through differences between languages one correlate set at a time. These methods include, e.g., Split Decomposition (Bandelt and Dress, 1992) and NeighborNet (Bryant and Moulton, 2003). Split Decomposition has a tendency to produce more treelike graphs with low reticulation when dealing with large amounts of data (McMahon and McMahon, 2005: 158); consequently, it is not used here as a complementary model to the character-based tree models, because Split Decomposition does not provide much additional information with regard to the trees. NeighborNet, the algorithm used here, is a computationally simple way of illustrating the distances between all languages by means of a network of crosscutting connections.

The NeighborNet method goes through the data producing a set of splits, each of which separates the languages (or other taxa) of the sample into two sets.[4] The average distance of the sets separated by each split is referred to as

4 For languages A, B and C, the possible splits would be A|BC (separating A from both B and C), B|AC and C|AB. The amount of possible splits grows exponentially with more languages included in the sample.

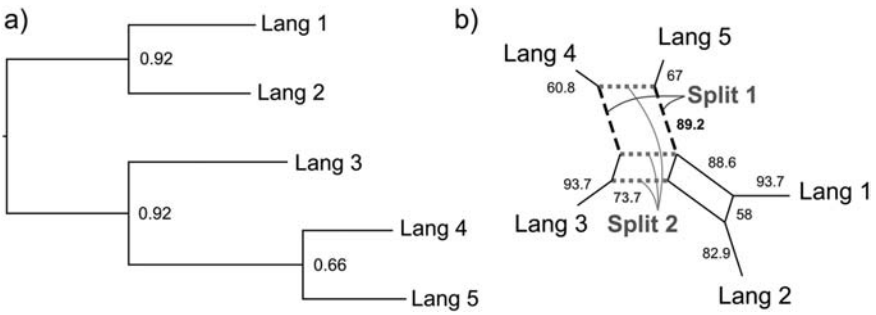

FIGURE 3 *a) A simulated tree; b) corresponding NeighborNet network with bootstrap values; two splits indicated with dashed lines*

the "weight" of that split. The resulting collection of splits can be viewed as a *split graph*, a network of connections that displays the weights of splits as the lengths of the corresponding lines that indicate distances between taxa. We generated our split graphs using the SplitsTree software (Huson, 1998; Huson and Bryant, 2006, version 4.13.1). For instance, in Fig. 3b, the parallel lines labeled "Split 1" separate languages 4 and 5 from the set formed by languages 1, 2 and 3 (see below for the indicated bootstrap values). Also, the relative distances can be seen from the sum of lengths of the lines leading from one language to another (through the shortest path): the distance of language 4 from 5 is proportional to the added weight of Split 2 and the weights of splits that separate each language from the network.

Even though the NeighborNet graphs are distance-based and do not aim to weed out any phylogenetic evolutionary signal, a network representation is more diverse than a discrete tree and helpful in assessing alternative connections, and NeighborNet can provide good estimates of historical connections (Bryant and Moulton, 2003). Figure 3 clearly shows where their greatest utility lies: whereas the phylogenetic tree in Fig. 3a, connecting languages 4 and 5 as a single branch, only shows that the closest language to that whole branch is language 3, the corresponding network (Fig. 3b) indicates that the branch containing languages 4 and 5 (shown as Split 1) is conflicted by the connection between languages 3 and 4 (Split 2), which weakens the former connection (languages 4 and 5) in the tree.

Another difference to the tree figures is that the networks are *unrooted*: splits indicate the distance between groups of languages, but the temporal order of the formation of the corresponding connections remains undetermined. The splits graph includes two measures for each split: the weight of a split, indicated by the length of the lines comprising it, and its robustness, represented by the *bootstrap values* placed near the corresponding splits. In the example provided

in Fig. 3b, Split 1 is longer than Split 2: the distance between languages on each side of the former split is greater than the distance for the latter, i.e., Split 1 has more weight. The amount of shared data based on the distance matrix is directly reflected in the length of the splits, and the shortest route from one language to another represents their relative distance.

Similarly, Split 1 is more robust than Split 2 based on its bootstrap value, 89.2 vs. 73.7 respectively (the value, 89.2 for instance, applies to both of the dark dashed lines making up Split 1). This robustness measure indicates how consistently the data support a given split; a short (weakly weighted) split can be considered an artefact of the data if it has a low bootstrap value, but the more robust splits are, the more significant they can be considered to be, even if weakly weighted. Bootstrapping is a statistical operation that divides the data randomly into a large number of smaller samples to see if a given conflicting signal is supported by the entire data or only a restricted set of characters (Felsenstein, 1985). This reduces the influence of any outliers that may be present in the data. Bootstrap values, however, are not as significant measures of the statistical reliability of splits as the posterior probability values are in the phylogenetic trees. Even splits with low bootstrap values are unambiguously present in the data and can be used to assess the connections between languages; the values only provide an additional statistical measure in determining the relative strengths of different connections. In the example in Fig. 3b, all bootstrap values have been included. In the NeighborNet splits graphs displayed in the results (Section 3, Figs 5 and 7), only values over 75 are indicated.

The bootstrap operation was performed in SplitsTree with 1,000 iterations. Bootstrap values were marked in our language networks on splits for values greater than 75, i.e., those splits that remained in 75 % or more of the bootstrap iterations. We also calculated the delta measure (δ) of phylogenetic reticulation (Holland et al., 2002) for our networks. Mean delta (mean of all δ values calculated for the whole data) is indicative of the treelikeness of the data based on the distance network; values close to 0 indicate that the data fit very well onto a tree topology, whereas values approaching 1 point to extensive reticulation causing conflicting or unresolved connections. Both delta scores and Q-residuals have been utilized to study reticulation in the history of different cultural patterns (Gray et al., 2010; Wichmann et al., 2011), but δ has been shown to be a more adequate measure of reticulation than Q-residuals. The delta value for each network is indicated in the captions for the respective figures.

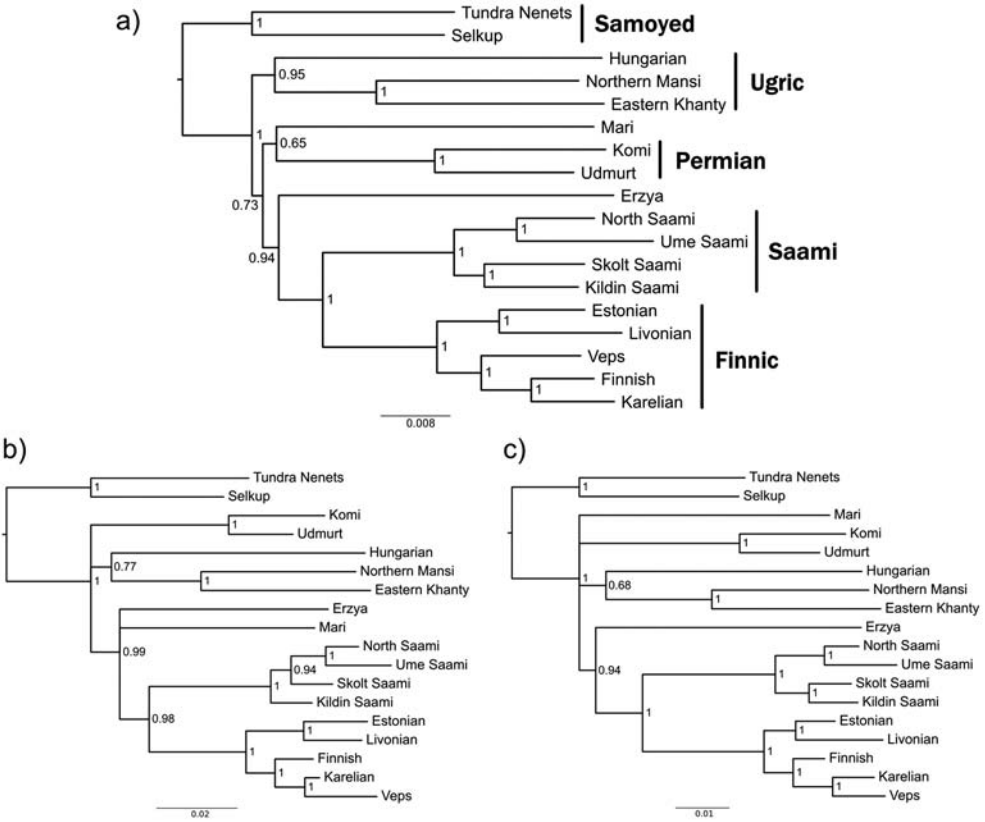

FIGURE 4 *Phylogenetic trees from the basic vocabulary datasets. Posterior probability values marked at nodes. a) Full Basic Vocabulary dataset (Swadesh lists + Leipzig-Jakarta; conventional low-level subgroups indicated); b) Ura100; c) Leipzig-Jakarta*

3 Results

3.1 *Basic Vocabulary Trees and Networks*

The three basic vocabulary trees (Fig. 4) all suggest a rather similar structure: after the first divergence between Samoyed and Finno-Ugric (determined manually on the basis of Syrjänen et al., 2013 and Honkola et al., 2013), Finno-Ugric diverges polytomously into several smaller groups (with low posterior probabilities; branches with support less than 0.9 are not considered). With the Full Basic Vocabulary dataset (Fig. 4a), these groups are Ugric (including Hungarian with posterior probability 0.95), Mari, Permian, and a tentatively supported Finno-Mordvin (0.94). With Ura100 (Fig. 4b), the groups diverging from Finno-Ugric are Permian, Hungarian, Ob-Ugric and Finno-Volgaic, and in the

Leipzig-Jakarta tree (Fig. 4c), these are Mari, Permian, Hungarian, Ob-Ugric and a tentatively supported Finno-Mordvin (0.94). Thus, the locations of Mari and Hungarian are the ones with the most variation.

There is further variability at the nodes separating Mordvin, Saami and Finnic. Full Basic Vocabulary and Leipzig-Jakarta trees suggest a Finno-Mordvin subgroup (0.94 in both trees) with an internal branching into Mordvin (represented here by Erzya) and Finno-Saami. Ura100 instead has a Finno-Volgaic branch (0.99) that diverges polytomously into Mari, Mordvin, and Finno-Saami. Also, although all three trees display fully supported, well-defined branches for the successive divergence of each individual Saami and Finnic language, the patterns vary between trees. Kildin Saami is placed either in an Eastern Saami branch together with Skolt Saami (Full Basic Vocabulary and Leipzig-Jakarta trees) or as the earliest of the Saami varieties studied to branch off (Ura100). A similar alternation pertains to the Northern Finnic languages Finnish, Karelian and Veps between different trees.

Having identified points of interest in the model of divergence, we turn to analyzing the data with NeighborNet graphs. The networks in Figs 5a, 5b and 5c (letters according to dataset, as in Fig. 4 above) show clearly defined splits with few conflicting connections, corresponding to many of the more stable subgroups seen in the trees. Groupings such as Finnic, Saami, Permian, Ob-Ugric and Samoyed consistently show up separated from the main body of the network by long (strongly weighted) splits that have few conflicting connections. The networks generally seem to support a treelike divergence pattern and have low mean delta values, indicating a low amount of reticulation ($\delta = 0.172$ in the case of Full Basic Vocabulary, $\delta = 0.1933$ for Ura100 and $\delta = 0.1787$ for the LJ network).

Other subgroups supported by the trees are weaker in the networks, for instance the Finno-Saami group, whose defining split is relatively short especially in the Ura100 network (Fig. 5b), where its bootstrap value also drops to 77.1. The corresponding branching point in the tree (Fig. 4b) is still strongly supported (0.98), which reflects the greater statistical power of the MCMC algorithm in finding the temporal pattern of diversification.

The networks also show a clear contrast with regard to the discrete but varying branching order of individual Finnic and Saami languages, as seen in the trees. Whereas different basic vocabulary trees show branches having full support (posterior probability values of 1.0, see Fig. 4), these are replaced in the networks by differently weighted sets of crosscutting splits between different languages. This is true of both the whole Saami area and the northern group of the Finnic languages (Finnish, Karelian and Veps). In all networks, crosscutting

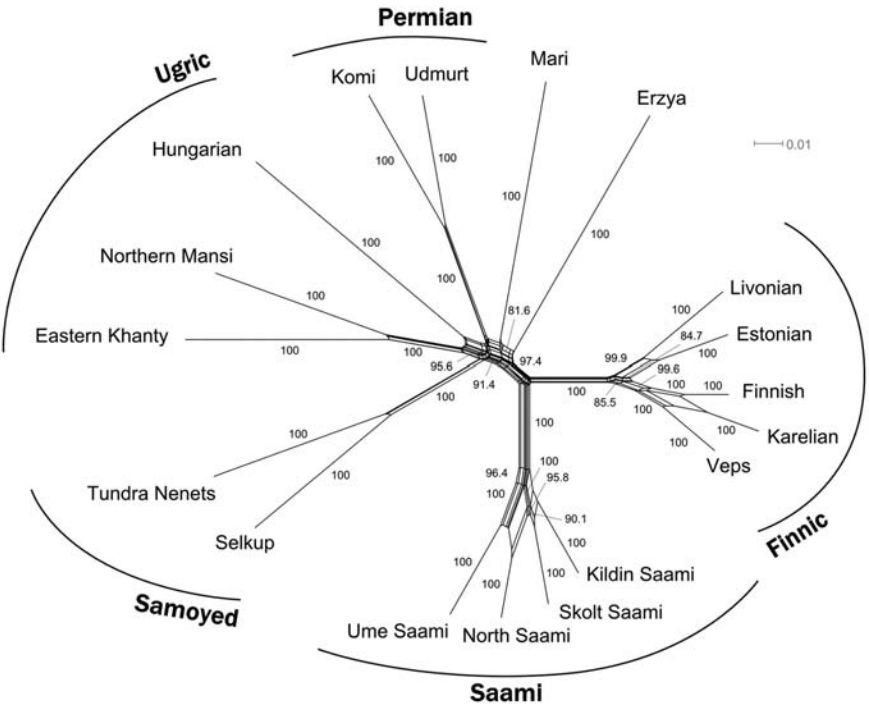

FIGURE 5A *Full Basic Vocabulary NeighborNet with low-level subgroups marked and bootstrap values greater than 75 indicated at the corresponding splits* ($\delta = 0.172$)

splits set Karelian more or less halfway between Finnish and Veps, contrasting with the neat bifurcations in the trees.

In the Full Basic Vocabulary network (Fig. 5a), the strong support value for the Ugric branch in the corresponding tree (0.95) is reflected as a fairly robust split (bootstrap value 95.6). This split conflicts with weak splits connecting Hungarian with Permian, Mari and Erzya, which it does not share with Mansi and Khanty. Even though only a Finno-Mordvin branch is seen in the tree (Fig. 4a), the network also shows a robust split that seems to separate a Finno-Volgaic group, including Mari in addition to Finnic, Saami and Mordvin (91.4). In the Saami subgroup the split separating the western languages Ume Saami and Northern Saami from the others is dominant, as reflected in the trees. The Skolt Saami/Kildin Saami grouping seen in the tree (Fig. 4a) is reflected as a weakly weighted split, which, however, is statistically robust, with a bootstrap value of 95.8.

The Ura100 network (Fig. 5b) shows no split whatsoever separating Finno-Mordvin from the rest; Mari and Erzya split off from the same point. Where the tree shows a three-way split of Finno-Volgaic (Fig. 4b), the corresponding

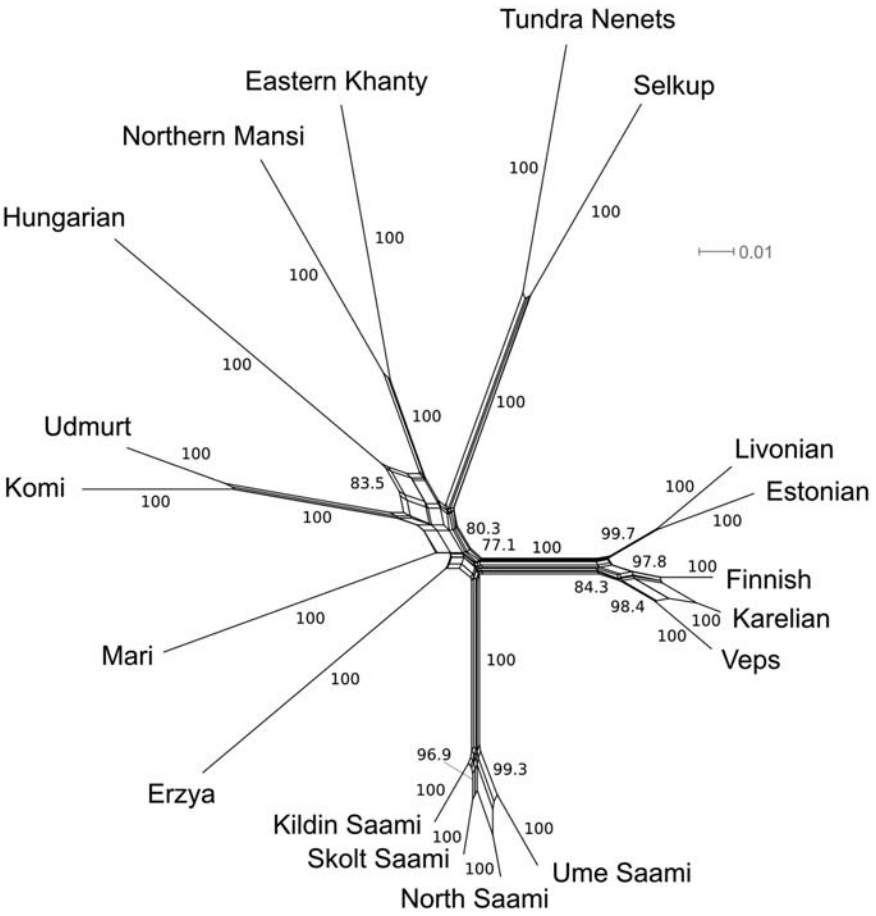

FIGURE 5B *Ura100 network* ($\delta = 0.1933$)

network also has a fairly robust split that separates Finno-Volgaic (80.3), but Mari, Erzya and Finno-Saami diverge approximately from the same point in the network. Again, a split corresponding to the Ugric branch is clear and robust (83.5), but this time Hungarian forms a connection with Permian (as well as a smaller one with Mari and Erzya) which, though less robust, has a weight comparable to the Ugric split.

In the Leipzig-Jakarta network (Fig. 5c), the secondary connections of Hungarian that conflict with a strong Ugric split (bootstrap value 89.2) are formed only with Permian. As reflected in the tree (Fig. 4c), there is no split separating Mari along with the other Finno-Volgaic languages from the remaining languages, but a split corresponding to Finno-Mordvin is clearly present and robust (90.8). Whereas in the Full Basic Vocabulary and Leipzig-Jakarta net-

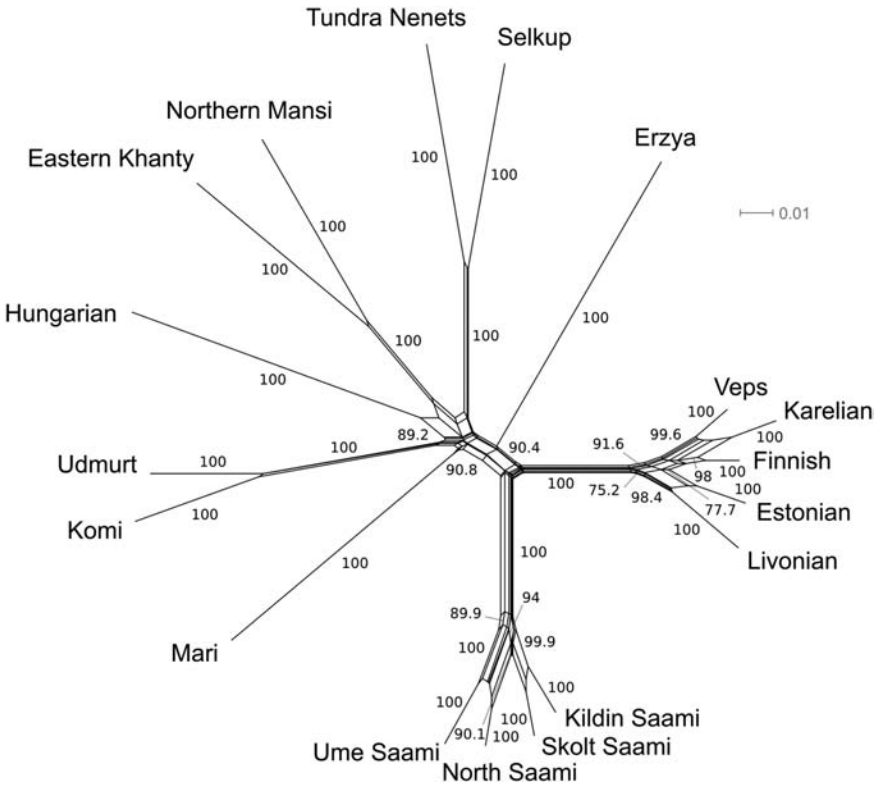

FIGURE 5C *Leipzig-Jakarta network* ($\delta = 0.1787$)

works the separation between the northern and southern groups of Finnic was fairly unequivocal, in the Leipzig-Jakarta network Estonian forms a discernible, robust (91.6) split that connects it with the northern group, which conflicts with its connection to Livonian.

3.2 *The WOLD401–500 Tree and Network*

The WOLD401–500 tree displays a stronger influence of secondary contacts than the basic vocabulary trees, as the data consist of less stable vocabulary (Fig. 6). After the rooted branching of Samoyed and Finno-Ugric, Ob-Ugric is suggested to be the next to diverge, with Finno-Permian + Hungarian as its sister group, which then diverges polytomously. Within Finnic, northern and southern groups are separated, and in the Saami languages Skolt and Kildin Saami form a sister group to North and Ume Saami. Thus, the well-established groups are also visible in the WOLD401–500 tree, whereas the intermediate divergences differ from those in the basic vocabulary trees.

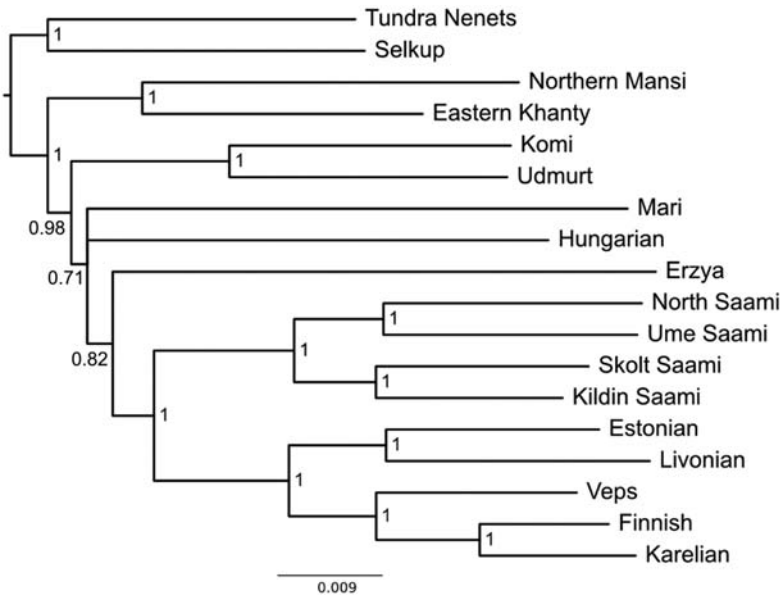

FIGURE 6 *Phylogenetic tree from the* WOLD401–500 *data*

The WOLD401–500 network (Fig. 7) is more reticulated than the other networks investigated here. However, its delta value is still quite low ($\delta = 0.244$). It retains the well-established groups (Saami, Finnic, Ob-Ugric, Samoyed and Permian). The split separating Finno-Saami from the rest is short but still robust (76.7), and clearly discernible splits connect some Finnic languages with Saami languages. The Saami languages are connected to all of Finnic except Livonian with a split that has a bootstrap value of 82.8. Skolt and Kildin Saami share a split with the whole of Finnic (76.8), and these Saami languages are also strongly connected to just Veps, Karelian and Finnish (95.7).

The relationships of Samoyed, Hungarian, Ob-Ugric, Permian and Mari remain unresolved because of the lack of splits clearly grouping any of these subgroups together. One of the more strongly weighted conflicting splits connects Selkup to the Permian languages, but even this split does not have a bootstrap value over 75.

With Erzya, Finnic and Saami there is no reduction of reticulation compared to the basic vocabulary networks, but there are more alternative connections. Splits conflicting with each other get more numerous and more robust compared with the basic vocabulary networks. There is a weak split separating the Finno-Mordvin languages from the rest, but Erzya is also connected just to Finnic with a more strongly weighted split.

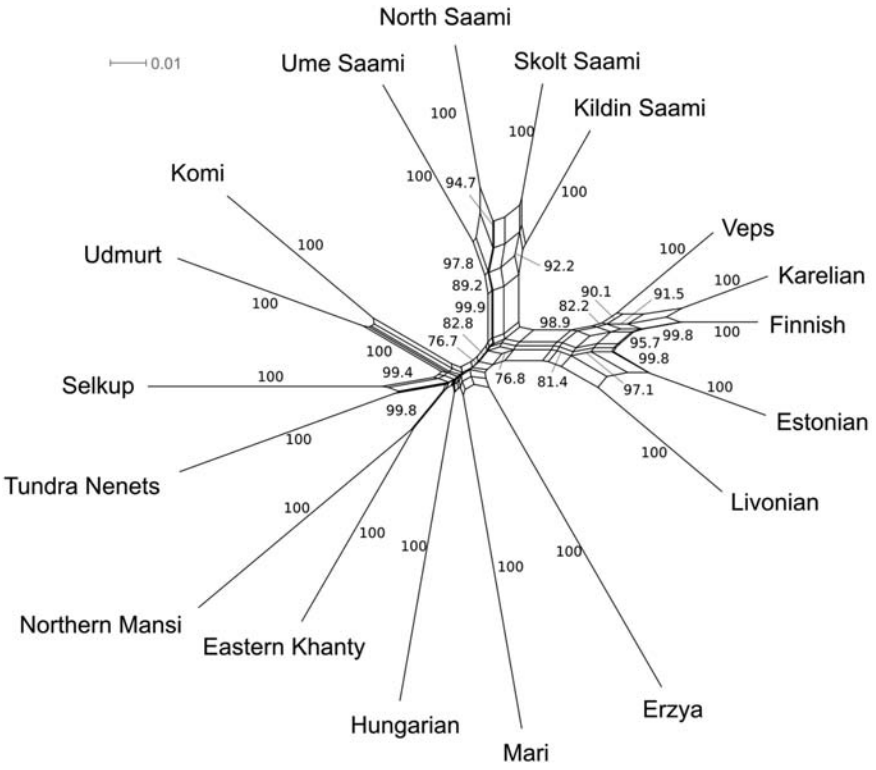

FIGURE 7 *NeighborNet from the* WOLD401–500 *dataset* ($\delta = 0.244$)

Unlike in the basic vocabulary trees and networks, Finnic does not clearly divide into northern and southern groups. Estonian shares one split with Veps, Karelian and Finnish (97.1) and another with just Karelian and Finnish (91.5), the splits now being more strongly weighted and more robust. Finnish and Karelian are now strongly connected to each other (99.8). There is also a fairly robust split (81.4) grouping Finnish, Estonian and Livonian together. The situation is more complex in the Saami subgroup as well. Ume and North Saami share a split (97.8) more robust than the conflicting ones (North + Skolt Saami, 94.7; North + Skolt + Kildin Saami, 92.2). Within the group, there is only a fairly weakly weighted further split connecting Skolt, North and Ume Saami, with a bootstrap value of 89.2.

4 Discussion

4.1 *Connections of the Uralic Languages in Light of Phylogenetic Networks*

The computational divergence analyses performed here and in earlier studies with tree-building algorithms have shown low support values and uncertain branching order for the intermediate divergences of the Uralic languages. This indicates that statistical models of treelike divergence cannot reliably establish the branching order during this period of diversification. We studied to what extent conflicting connections between the intermediate branches might have brought about the ambiguity in the branching pattern, and found only a modest degree of reticulation in the middle of both basic vocabulary and less basic vocabulary networks. Instead, the networks showed a largely "bushlike" structure where Samoyed, Permian, Ugric and Finno-Volgaic split more or less from the same point in the networks.

The modest degree of reticulation suggests that the inability of the tree models to find a clear branching order after the split of Samoyed is not caused by conflicting connections introduced by the data (e.g., borrowings). Instead, the pattern suggests that the early divergence of the Finno-Ugric branch was rapid and characterized by fast lexical development leading to the daughter branches. A divergence pattern involving a large rate of lexical development would make any kind of lexical dataset incapable of determining a discrete order of divergence.

The conflicting connections seen in intermediate branches may have been caused either by convergence in vocabulary, i.e., borrowing strong enough to pervade even basic vocabulary, or by a process of divergence that could not be properly modeled with a tree diagram. This latter scenario would leave us with the suggestion of wavelike divergence, a situation where lexical innovations spread only partially through ancestral dialect continua (Gray et al., 2010). We argue that the effect of the partial spread of lexical innovations between the early Finno-Ugric branches may be estimated by the degree of reticulation in networks.

Hungarian is one of the languages with several conflicting connections in the networks, making its position unclear (an issue which has also been noted earlier by Honti, 1998; Syrjänen et al., 2013; Honkola et al., 2013). In all the networks, Hungarian is connected to the Ob-Ugric languages, Khanty and Mansi, with a strongly weighted and robust split forming a Ugric branch. However, Hungarian shares splits also with Permian, which is most obvious in the Ura100 and LJ networks (Figs 5b and 5c), as well as Mari and Erzya (in all but the LJ network, cf. Figs 5a, 5b, and 7). These are in conflict with the

Ugric split, which, however, remains the most robust in the networks. The basic vocabulary networks suggest that the divergence of Hungarian did not happen from a discrete, uniform Ugric protolanguage. Instead, it may have originated in a Ugric language variety that shared some characteristics with Permian, Mari, and Erzya, caused by wavelike lexical diffusion in the diversification of Proto-Finno-Ugric. The observed patterns suggest that either ancient dialectal borders already existed in the Ugric protolanguage at the time of divergence of Hungarian, or that, after divergence from Proto-Ugric, Hungarian underwent strong areal influence with branches such as Permian.

Another language with varying positions in the trees is Mari. In the Full Basic Vocabulary and Ura100 networks (Figs 5a and 5b), Mari forms a strong split that separates it, along with Erzya, Finnic and Saami, from the other languages to form a Finno-Volgaic branch (this is seen in the trees only with the Ura100 data, cf. Fig. 4b). Conflicting with this, Mari forms additional splits both with Permian and Erzya and with Hungarian. These connections are likely the reason for its unclear position in the trees. As these connections are seen with the basic vocabulary data but not in the WOLD401–500 network (except for the connection with Erzya), it seems that the divergence of Mari also took place in a time of fast diversification of the Finno-Ugric protolanguage, characterized by much the same wavelike influence that was suggested to lie behind the ambiguities in the placement of Hungarian. However, for Mari, there is no indication of another branch that could have served as a source of strong alternative connections, unlike with the Hungarian-Permian connections, which were seen to conflict with the Ugric branch.

The lack of secondary connections between the Finno-Ugric subgroups indicates that recent contact between the branches has been scarce. This may be because Uralic languages have borrowed more from members of other language families spoken in the area than from other Uralic languages. For most languages spoken in the area of European Russia and Western Siberia, these contact languages have been the Iranian and Turkic languages and, more recently, Russian (Korenchy, 1988; Róna-Tas, 1988; Décsy, 1988); Hungarian has borrowed from Turkic, Slavic and Germanic languages (Imre, 1988). Our networks do not include the non-Uralic contact languages, so no conclusions on their impact on Uralic-internal connections can be drawn. Borrowing between Uralic branches, e.g. between Komi and Ob-Ugric and between Khanty and Nenets, is documented in Uralic studies (e.g. Abondolo, 1998a: 382–383). However, such contacts are not strong enough to be visible in the WOLD401–500 data.

The comparison of trees and networks provides more information about the evolutionary history of the closely related Finnic and Saami languages.

Interestingly, divergences within and between the Finnic and Saami subgroups gain almost full support in all the tree models, even though the location of languages varies between the datasets. The networks provide a much more diverse picture with a lot of conflicting connections. This shows that trees alone cannot be used to gain insight into the diversity of connections within recent dialect continua. While we can only conjecture about ancient dialect continua being the reason for the unclear divergences of the ancestral Finno-Ugric protolanguage, we can be more confident that this is a major reason for the network pattern seen in the Finno-Saami group. In general, this supports the use of network methods for obtaining information about recent periods of diffuse divergence of languages that are closely related today.

The Saami languages have until recently formed a continuum of dialects, in which the most important borders between neighboring language varieties are between the so-called Western Saami languages (including Ume and North Saami), Inari Saami, and the Eastern Saami languages (including Skolt and Kildin Saami; Korhonen, 1981: 17). This division is evident in the Full Basic Vocabulary, LJ and WOLD401–500 trees (Figs 4a, 4c, and 6), where Ume and North Saami are grouped together, as are Skolt and Kildin Saami from the eastern group. The corresponding networks (Figs 5a, 5c, and 7) show that, even though the splits defining this grouping are most robust and strongly weighted, there are also fairly robust splits that conflict with it. This might be a consequence of the late Saami dialect continuum and a relatively recent formation of language borders.

However, the Ura100 tree (Fig. 4b) supports the separation of Kildin Saami as the first of these four languages. It is possible that Ura100, being a quality-optimized meaning list for the Uralic languages, manages to find a deeper divergence preceding those that separate Eastern from Western Saami. This view agrees with Korhonen (1981: 19–20), who shows that the earliest and most fundamental phonological isoglosses separate the easternmost varieties, namely Kildin and Ter Saami, from the remaining Saami varieties.

The pattern of borrowing within the Uralic family can be seen most clearly in the increase of the network pattern between Finnic and Saami in the WOLD401–500 network (Fig. 7), compared with the basic vocabulary networks. According to, e.g., Korhonen (1981: 37–40), the direction of borrowing between these subgroups has mainly been from Finnic towards Saami. Several layers of borrowing are known in historical linguistics, starting with borrowing from Proto-Finnic to Proto-Saami, affecting all individual languages. More recently, loanwords in Saami have mainly been borrowed from the geographically closest Finnic languages, Finnish and Karelian. These most recent loans have not spread across the whole of the Saami area (Korhonen, 1988: 266–267). This can

also be seen in our results: Northern Finnic connects with just Skolt and Kildin Saami, which have bordered the Finnish and Karelian speech communities for a long time (Korhonen, 1981: 37–39).

In the WOLD401–500 network, within the Finnic branch, Estonian forms many connections with the Northern Finnic languages, especially Finnish and Karelian, reflecting strong and long-standing contacts around the Gulf of Finland (Miettinen, 1996). Livonian has largely been excluded from this contact group because of its geographical separation (Laanest, 1982: 26–29). Furthermore, the Northern Finnic group, for which the basic vocabulary networks support no clear internal division, now display a strong Finnish-Karelian connection. This indicates that the strongest contacts within the Northern Finnic group have existed between Finnish and Karelian. Borrowing has mainly occurred from Finnish into Karelian and not vice versa (Laakso, 2001: 202).

4.2 *Basic vs. Less Basic Vocabulary in Language Subgrouping*

We can expect the less basic vocabulary data to represent a) increased impact of borrowing between the languages studied, b) relatively weakened information about old connections because of a greater rate of replacements, c) patchier data because of the absence of words in the Uralic languages for some of the meanings, and d) ambiguity due to morphological complexity and the resulting difficulty of assigning the collected words to correlate sets. All these factors are apparent when comparing the Uralic basic vocabulary data to the WOLD401–500 data, which includes lexical material that does not accord to the criteria of basic vocabulary.

Heggarty (2010) notes that most approaches using meaning lists aim to restrict the lexical data to rule out the effect of borrowings and fast replacement, but argues that this loses valuable information about historical connections between languages. By comparing basic vocabulary with WOLD401–500 data, we can test the influence of different kinds of lexical data types on cross-linguistic connections. The tree models generated from basic vocabulary (Fig. 4) provide a more resolved divergence pattern than the less basic vocabulary does. The tree generated from less basic vocabulary (Fig. 6) is more polytomous (showing no clear bifurcation), as indicated by weaker support for the branching points. Moreover, the branch lengths of the individual languages in the WOLD401–500 tree are much longer than the lengths of the branches leading up to them, compared to the basic vocabulary trees. This indicates that much of the lexical change displayed by the WOLD401–500 data reflects recent lexical developments specific to individual languages, rather than innovations defining older divergences.

The NeighborNet graph produced with the WOLD401–500 data (Fig. 7) shows that the increased polytomy at the intermediate branches in the tree (uncertainty of the divergence order of the Finno-Ugric branches) is not caused by contacts between these branches at a stage after their initial divergence, which could be uncovered using these data. Instead, the branches with low support in the less basic vocabulary tree are reflected as lacking connections in the corresponding network. Compared to the basic vocabulary lists, the less basic vocabulary data thus makes the picture of the Finno-Ugric divergence more ambiguous. The decreased treelikeness and increased reticulation is also reflected in the delta value of the WOLD401–500 network, which is higher than for the basic vocabulary networks.

However, the remaining criteria of basic vocabulary (universality of concepts and morphological complexity) probably have less impact on the differences between models of basic and less basic vocabulary data than borrowing and internal lexical replacement. The fact that the WOLD401–500 data have more gaps in the word lists (because words for certain less universally lexicalized meanings cannot be found in various Uralic languages, see Section 2.2 above) may have caused the weakening of detectable connections. Morphological complexity of individual lexical items is not expected to adversely influence the structure of models, although it makes correlation judgments more complicated and is a reasonable indicator of the age of the forms in question. Thus, we expect that the greatest differences between the WOLD401–500 and basic vocabulary data and their interpretations are caused primarily by a) increased impact of borrowing and b) younger age of the lexical replacements and less information about old connections in the WOLD401–500 data.

Bearing in mind that, generally, human languages contain tens of thousands of words, the differences between the trees and networks produced by the WOLD401–500 list on the one hand and the basic vocabulary data on the other hand are significant, despite the fact that all of the concepts studied are, in principle, historically stable meanings. The meanings used for our less basic vocabulary list, although not of the kind most easily borrowed and replaced, still show the effect of recent contacts clearly, in comparison with the basic vocabulary. Certain hypothetical intermediate protolanguages of the Uralic subgroups, such as Proto-Finno-Saami and Proto-Finno-Volgaic, are more strongly supported by the basic vocabulary data than by the less stable meanings. Increased connections corresponding to these groupings in less basic vocabulary would support the role of contact in their formation. However, the fact that basic vocabulary shows these intermediate groupings more clearly than less stable meanings suggests that they are the result of lexical innovations in protolanguage stages, rather than borrowing after divergence.

We show in this study (see also Syrjänen et al., 2013) that correlational data on basic vocabulary can be used to obtain a treelike divergence pattern, and lexical data on less stable meanings may help in determining the role of relatively late contacts in the history of the languages studied. The diverse handling of lexical data with quantitative methodology is helpful in testing different hypotheses of language divergence in well-researched language families. However, this approach would likewise work for languages whose etymological connections are less thoroughly studied. Correlational relationships can be estimated for word lists collected from languages not researched well enough to reliably distinguish layers of borrowing from retained ancestral words. Comparing models based on basic vocabulary with models based on less stable meanings can provide a working hypothesis of historical connections as a starting point for further historical linguistic study.

4.3 *Phylogenetic Trees vs. Networks in Modeling the History of Languages*

Comparing trees and networks allows us to evaluate the sources of ambiguity in the phylogenetic signal of treelike divergence history. As argued by, e.g., Heggarty et al. (2010), compared to trees, the network models provide a more detailed picture of the evolutionary history of languages by combining the mechanisms of treelike and wavelike divergence. We can expect the evolutionary pattern of a language family to include both clear splitting events and diffuse diversification, which can be inferred by using different methods of analysis—preferably with varying datasets.

The comparisons allow for different kinds of diagnosis. Firstly, the trees display unclear divergence patterns for the intermediate branches, which correspond to a low amount of connections in the networks. Secondly, in other places, the networks show significant conflicting connections that render individual branches less firmly supported in the trees, pointing to wavelike diffusion of lexical material between the languages. Also, in some places the trees have strong support values, but the network illustrates clear secondary connections. This can be seen with closely related languages.

In applying the computational phylogenetic methodology to languages whose history is not well known, it would be essential to use a network representation of language connections to obtain implications of recent dialect continua. We can see recent wavelike diffusion between closely related languages bring about a branching order in the trees, with short branch lengths but still clearly defined. This indicates the statistical power of the trees in finding a bifurcating structure even in the presence of conflicting signal in the data. A single tree, with full support values for a bifurcating divergence of closely

related languages, would not imply the crosscutting connections shown in the networks. This shows the utility of using not only different types of datasets, but alternative methods to model different kinds of connections.

Similarly, we can see borrowing cause reticulation of the connections between languages. However, most branches in the trees are well supported, even with large amounts of borrowing affecting the connections seen in the networks. In other cases, unclear nodes in the trees do not correspond to secondary connections that would correlate with borrowing between Uralic branches. This indicates a situation where borrowing from other language families may have caused lexical replacement, resulting in ambiguous connections with other Uralic languages.

One of the advantages of networks is that they allow us to assess the usability of the tree models for data analysis (e.g. Bryant and Moulton, 2003). For instance, Gray et al. (2010) analyze lexical data from Polynesian and Indo-European languages and conclude (using the δ measure, among others) that the divergence pattern of Indo-European languages resembles a treelike pattern more than in the case of the Polynesian languages. In general, our networks support the estimates of Uralic divergence history made on the basis of tree models. Both methods uphold the lower-level subgroups with strong support and point to a largely treelike history of lexical replacement.

5 Conclusions

Comparing analyses of Uralic languages using both character-based tree methods and distance-based network models adds a layer of depth to the analysis of diversification of the Uralic language communities. Basic vocabulary data seem to provide support for several subgroups, and comparing these data to less stable vocabulary allows assessing the effect of borrowing between different Uralic branches. The ambiguous position of Hungarian is seen to be caused by secondary contacts with e.g. Permian languages, which conflict with the primary connections of Hungarian with the Ob-Ugric languages. The divergence of Mari can be placed within the same diffuse diversification of Finno-Ugric in which present-day Hungarian, Ob-Ugric and Permian also diverged as separate groups. In the case of Mari, the strongest connections seem to point towards a separate Finno-Volgaic stage after the Finno-Ugric diversification, followed closely by the separation of Mari from a residual Finno-Mordvin group.

We show here that comparing basic vocabulary to less basic vocabulary allows for separating the effects of more recent lexical change and borrowings from the primary divergence pattern of the language family. Patterns based on

vocabulary that is more prone to replacement and change (less basic vocabulary) do not contradict the signal obtained from basic vocabulary; on the contrary, the apparent conflicts provide more diversity in the study of evolutionary history of language families. We encourage the consideration of less stable meanings along with basic vocabulary, especially if more recent contacts are of interest.

NeighborNets on their own cannot provide support for different subgrouping hypotheses, but they can be used to refine models based on phylogenetic trees. Combination of these two models also allows for flexible modeling of language history, as binary and polytomous branching can take place alongside wavelike development in other periods of history.

Computational tree and network methods can be employed together as methods for obtaining quantitative, diverse models of well-researched language families. They can also be used to develop initial models of subgrouping for language families for which prior historical-comparative research is not available. Naturally, the use of more diverse data types provides a more complete picture, but no matter what kind of data is used, the comparison of models of treelike divergence with distance-based network models is more fruitful than using either method alone.

Acknowledgments

We thank Nikolett F. Gulyás, Paul Heggarty and Urho Määttä for their comments on earlier drafts of this article. We also thank Spiros Papakostas for providing assistance with the network methodology. This research has been funded by Kone Foundation for TH, KS and OV and by the University of Helsinki and Kone Foundation for JL and KK.

References

Abondolo, Daniel. 1998a. Khanty. In Abondolo (ed.), 358–386.

Abondolo, Daniel (ed.). 1998b. *The Uralic Languages*. London: Routledge.

Aikhenvald, Alexandra Y. and Robert M.W. Dixon (eds.). 2001. *Areal Diffusion and Genetic Inheritance: Problems in Comparative Linguistics*. Oxford: Oxford University Press.

Bandelt, Hans-Jürgen and Andreas W.M. Dress. 1992. Split decomposition: A new and useful approach to phylogenetic analysis of distance data. *Molecular Phylogenetics and Evolution* 1: 242–252.

Bandelt, Hans-Jürgen, Peter Forster, and Arne Röhl. 1999. Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution* 16: 37–48.

Bandelt, Hans-Jürgen, Peter Forster, Bryan C. Sykes, and Martin B. Richards. 1995. Mitochondrial portraits of human populations using median networks. *Genetics* 141: 743–753.

Bloomfield, Leonard. 1933. *Language*. New York: Henry Holt.

Bryant, David and Vincent Moulton. 2003. Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution* 21: 255–265.

Campbell, Lyle. 1998. *Historical Linguistics: An Introduction*. 2nd ed. Cambridge: MIT Press.

Décsy, Gyula. 1988. Slawischer Einfluss auf die uralischen Sprachen. In Sinor (ed.), 616–637.

Drummond, Alexei J. and Andrew Rambaut. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* 7: 214 (doi: 10.1186/1471-2148-7-214).

Dunn, Michael, Stephen C. Levinson, Eva Lindström, Ger Reesink, and Angela Terrill. 2008. Structural phylogeny in historical linguistics: Methodological explorations applied in Island Melanesia. *Language* 84: 710–759.

Felsenstein, Joseph. 1985. Confidence-limits on phylogenies: An approach using the bootstrap. *Evolution* 39: 783–791.

Fitch, Walter M. 1997. Networks and viral evolution. *Journal of Molecular Evolution* 44: s65–s75.

Forster, Peter, Tobias Polzin, and Arne Röhl. 2006. Evolution of English basic vocabulary within the network of Germanic languages. In Peter Forster and Colin Renfrew (eds.), *Phylogenetic Methods and the Prehistory of Languages*, 131–138. Cambridge: McDonald Institute for Archaeological Research.

Forster, Peter and Alfred Toth. 2003. Toward a phylogenetic chronology of ancient Gaulish, Celtic, and Indo-European. *Proceedings of the National Academy of Sciences* 100: 9079–9084.

Forster, Peter, Alfred Toth, and Hans-Jürgen Bandelt. 1998. Evolutionary network analysis of word lists: Visualising the relationships between Alpine Romance languages. *Journal of Quantitative Linguistics* 5: 174–187.

Gray, Russell D. and Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426: 435–439.

Gray, Russell D., David Bryant, and Simon J. Greenhill. 2010. On the shape and fabric of human history. *Philosophical Transactions of the Royal Society B* 365: 3923–3933.

Gray, Russell D. and Fiona M. Jordan. 2000. Language trees support the express-train sequence of Austronesian expansion. *Nature* 405: 1052–1055.

Greenhill, Simon, Thomas E. Currie, and Russell D. Gray. 2009. Does horizontal trans-

mission invalidate cultural phylogenies? *Proceedings of the Royal Society B* 276: 2299–2306.

Häkkinen, Kaisa. 1984. Wäre es schon Zeit, den Stammbaum zu fällen? Theorien über die gegenseitigen Verwandtschaftsbeziehungen der finnisch-ugrischen Sprachen. *Ural-Altaische Jahrbücher: Neue Folge* 4: 1–24.

Haspelmath, Martin and Uri Tadmor (eds.). 2009a. *Loanwords in the World's Languages: A Comparative Handbook*. Berlin: Walter de Gruyter.

Haspelmath, Martin and Uri Tadmor (eds.). 2009b. *World Loanword Database*. Munich: Max Planck Digital Library. Available at http://wold.livingsources.org/ (accessed May 18, 2011).

Heggarty, Paul. 2010. Beyond lexicostatistics: How to get more out of 'word list' comparisons. *Diachronica* 27: 301–324.

Heggarty, Paul, Warren McGuire, and April McMahon. 2010. Splits or waves? Trees or webs? How divergence measures and network analysis can unravel language histories. *Philosophical Transactions of the Royal Society B* 365: 3829–3843.

Hennig, Willi, D. Dwight Davis, and Rainer Zangerl. 1996. *Phylogenetic Systematics*. Urbana: University of Illinois Press.

Holden, Clare Janaki. 2002. Bantu language trees reflect the spread of farming across sub-Saharan Africa: A maximum-parsimony analysis. *Proceedings of the Royal Society of London B* 269: 793–799.

Holland, Barbara R., Katharina T. Huber, Andreas Dress, and Vincent Moulton. 2002. δ plots: A tool for analyzing phylogenetic distance data. *Molecular Biology and Evolution* 19: 2051–2059.

Honkola, Terhi, Outi Vesakoski, Kalle Korhonen, Jyri Lehtinen, Kaj Syrjänen, and 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).

Honti, László. 1998. *Ugrilainen kantakieli—erheellinen vai reaalinen hypoteesi?* Suomalais-Ugrilaisen Seuran toimituksia 228. Helsinki: Suomalais-Ugrilainen Seura.

Hovdhaugen, Even, Fred Karlsson, Carol Henriksen, and Bengt Sigurd (eds.). 2000. *The History of Linguistics in the Nordic Countries*. Helsinki: Societas Scientiarum Fennica.

Huelsenbeck, John P. and Fredrik Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17: 754–755.

Huelsenbeck, John P., Fredrik Ronquist, Rasmus Nielsen, and Jonathan P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294: 2310–2314.

Huson, Daniel H. 1998. Splitstree: A program for analyzing and visualizing evolutionary data. *Bioinformatics* 14: 68–73.

Huson, Daniel H. and David Bryant. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23: 254–267.

Imre, Samu. 1988. Die Geschichte der ungarischen Sprache. In Sinor (ed.), 413–447.

Itkonen, Terho. 1997. Reflections on Pre-Uralic and the 'Saami-Finnic protolanguage.' *Finnisch-Ugrische Forschungen* 54: 229–266.

Janhunen, Juha. 1981. Uralilaisen kantakielen sanastosta. *Journal de la société finno-ougrienne* 77: 219–274.

Janhunen, Juha. 2009. Proto-Uralic—what, where and when? In Jussi Ylikoski (ed.), *The Quasquicentennial of the Finno-Ugrian Society* [Suomalais-Ugrilaisen Seuran Toimituksia 258], 57–78. Helsinki: Suomalais-Ugrilainen Seura.

Korenchy, Eva. 1988. Iranischer Einfluss in den finnisch-ugrischen Sprachen. In Sinor (ed.), 665–681.

Korhonen, Mikko. 1981. *Johdatus lapin kielen historiaan*. Helsinki: Suomalaisen Kirjallisuuden Seura.

Korhonen, Mikko. 1988. The history of the Lapp language. In Sinor (ed.), 264–287.

Kulonen, Ulla-Maija. 2002. Kielitiede ja Suomen väestön juuret. In Riho Grünthal (ed.), *Ennen muinoin: Miten menneisyyttämme tutkitaan*, 102–116. Helsinki: Suomalaisen Kirjallisuuden Seura.

Laakso, Johanna. 2001. The Finnic languages. In Östen Dahl and Maria Koptjevskaja-Tamm (eds.), *Circum-Baltic Languages: Typology and Contact*. Volume 1: Past and Present, 179–212. Amsterdam: John Benjamins.

Laanest, Arvo. 1982. *Einführung in die ostseefinnischen Sprachen*. Hamburg: Buske.

McMahon, April and Robert McMahon. 2005. *Language Classification by Numbers*. Oxford: Oxford University Press.

McMahon, April, Paul Heggarty, Robert McMahon, and Natalia Slaska. 2005. Swadesh sublists and the benefits of borrowing: An Andean case study. *Transactions of the Philological Society* 103: 147–170.

Michalove, Peter A. 2002. The classification of Uralic languages: Lexical evidence from Finno-Ugric. *Finnisch-Ugrische Forschungen* 57: 58–67.

Miettinen, Timo. 1996. Suomenlahden ulkosaarten esihistoriaa. In Risto Hamari, Martti Korhonen, Timo Miettinen, and Ilmar Talve (eds.), *Suomenlahden ulkosaaret: Lavansaari, Seiskari, Suursaari, Tytärsaari*, 49–68. Helsinki: Suomalaisen Kirjallisuuden Seura.

Nelson-Sathi, Shijulal, Johann-Mattis List, Hans Geisler, Heiner Fangerau, Russell D. Gray, William Martin, and Tal Dagan. 2011. Networks uncover hidden lexical borrowing in Indo-European language evolution. *Proceedings of the Royal Society B* 278: 1794–1803.

Pusztay, János. 1995. *Diskussionsbeiträge zur Grundsprachenforschung* (*Beispiel: das Protouralische*). Veröffentlichungen der Societas Uralo-Altaica 43. Wiesbaden: Harrassowitz.

Rédei, Károly. 1988. *Uralisches Etymologisches Wörterbuch*. Band I. Wiesbaden: Harrassowitz.

Ringe, Don, Tandy Warnow, and Ann Taylor. 2002. Indo-European and computational cladistics. *Transactions of the Philological Society* 100: 59–129.

Róna-Tas, András. 1988. Turkic influence on the Uralic languages. In Sinor (ed.), 742–780.

Ronquist, Fredrik and John P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19: 1572–1574.

Salminen, Tapani. 1999. Euroopan kielet muinoin ja nykyisin. In Paul Fogelberg (ed.), *Pohjan poluilla: Suomalaisten juuret nykytutkimuksen mukaan*, 13–26. Helsinki: Societas Scientiarum Fennica.

Salminen, Tapani. 2002. Problems in the taxonomy of the Uralic languages in the light of modern comparative studies. In Александр Е. Кибрик (ed.), *Лингвистический беспредел: сборник статей к 70-летию А. И. Кузнецовой*, 44–55. Moscow: Издательство Московского университета.

Sammallahti, Pekka. 1988. Historical phonology of the Uralic languages with special reference to Samoyed, Ugric and Permic. In Sinor (ed.), 478–554.

Schleicher, August. 1861. *Compendium der vergleichenden Grammatik der indogermanischen Sprachen*. Weimar. [2nd ed. 1876.]

Schmidt, Johannes. 1872. *Die Verwandtschaftsverhältnisse der indogermanischen Sprachen*. Weimar: H. Böhlau.

Sinor, Denis (ed.). 1988. *The Uralic Languages: Description, History and Foreign Influences*. Handbuch der Orientalistik. Achte Abteilung, Vol. 1. Leiden: E.J. Brill.

Swadesh, Morris. 1950. Salish internal relationships. *International Journal of American Linguistics* 16: 157–167.

Swadesh, Morris. 1952. Lexico-statistic dating of prehistoric ethnic contacts: With special reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society* 96: 452–463.

Swadesh, Morris. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21: 121–137.

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

Tadmor, Uri. 2009. Loanwords in the world's languages: Findings and results. In Martin Haspelmath and Uri Tadmor (eds.), *Loanwords in the World's Languages: A Comparative Handbook*, 55–75. Berlin: Walter de Gruyter.

Walker, Robert S. and Lincoln A. Ribeiro. 2011. Bayesian phylogeography of the Arawak expansion in lowland South America. *Proceedings of the Royal Society B* 278: 2562–2567.

Wichmann, Søren, Eric W. Holman, Taraka Rama, and Robert S. Walker. 2011. Correlates of reticulation in linguistic phylogenies. *Language Dynamics and Change* 1: 205–240.

Wiley, Edward O. 1981. *Phylogenetics: The Theory and Practice of Phylogenetic Systematics*. New York: Wiley.