Turun yliopisto
University of Turku

MACRO- AND MICROEVOLUTION OF LANGUAGES: EXPLORING LINGUISTIC DIVERGENCE WITH APPROACHES FROM EVOLUTIONARY BIOLOGY

Terhi Honkola

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
Sarja - ser. A II osa - tom. 311 | Biologica - Geographica - Geologica | Turku 2016

University of Turku

Faculty of Mathematics and Natural Sciences
Department of Biology
Section of Ecology

Supervised by

Dr. Outi Vesakoski
Department of Biology
University of Turku

Dr. Kalle Korhonen
Department of World Cultures
University of Helsinki

Prof. Niklas Wahlberg
Department of Biology
University of Turku

Reviewed by

Prof. Mark Pagel
School of Biological Sciences
University of Reading, UK

Docent Päivi Onkamo
Department of Biosciences
University of Helsinki, Finland

Opponent

Prof. Michael Dunn
Department of Linguistics and Philology
Uppsala University, Sweden

Cover image by Tiina Honkola

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin OriginalityCheck service.

ISBN 978-951-29-6374-4 (PRINT)
ISBN 978-951-29-6375-1 (PDF)
ISSN 0082-6979
Painosalama Oy - Turku, Finland 2016

“You either bet big or go home.
You gotta risk it to get the biscuit.”

-Shawn, Fired Up

“Toivottavasti sentään lukija, eikä vain lähellä tekijän ammattikuntaa oleva lukija, löytää näistä pahasti omakohtaisista ja sinänsä vähäpätöisistä muisteloista sekä matkakuvauksista myös sivistyshistoriallisesti mielenkiintoista, heimolaistemme elämää valottavaa tai muuta vetävää siksi paljon, että ankeasta alusta yli päästyään ei heitä kirjaa kesken.”

-Lauri Kettunen, Tieteen matkamiehen uusia elämyksiä

ABSTRACT

There are more than 7000 languages in the world, and many of these have emerged through linguistic divergence. While questions related to the drivers of linguistic diversity have been studied before, including studies with quantitative methods, there is no consensus as to which factors drive linguistic divergence, and how.

In the thesis, I have studied linguistic divergence with a multidisciplinary approach, applying the framework and quantitative methods of evolutionary biology to language data. With quantitative methods, large datasets may be analyzed objectively, while approaches from evolutionary biology make it possible to revisit old questions (related to, for example, the shape of the phylogeny) with new methods, and adopt novel perspectives to pose novel questions. My chief focus was on the effects exerted on the speakers of a language by environmental and cultural factors. My approach was thus an ecological one, in the sense that I was interested in how the local environment affects humans and whether this human-environment connection plays a possible role in the divergence process. I studied this question in relation to the Uralic language family and to the dialects of Finnish, thus covering two different levels of divergence. However, as the Uralic languages have not previously been studied using quantitative phylogenetic methods, nor have population genetic methods been previously applied to any dialect data, I first evaluated the applicability of these biological methods to language data.

I found the biological methodology to be applicable to language data, as my results were rather similar to traditional views as to both the shape of the Uralic phylogeny and the division of Finnish dialects. I also found environmental conditions, or changes in them, to be plausible inducers of linguistic divergence: whether in the first steps in the divergence process, i.e. dialect divergence, or on a large scale with the entire language family. My findings concerning Finnish dialects led me to conclude that the functional connection between linguistic divergence and environmental conditions may arise through human cultural adaptation to varying environmental conditions. This is also one possible explanation on the scale of the Uralic language family as a whole.

The results of the thesis bring insights on several different issues in both a local and a global context. First, they shed light on the emergence of the Finnish dialects. If the approach used in the thesis is applied to the dialects of other languages, broader generalizations may be drawn as to the inducers of linguistic divergence. This again brings us closer to understanding the global patterns of linguistic diversity. Secondly, the quantitative phylogeny of the Uralic languages, with estimated times of language divergences, yields another hypothesis as to the shape and age of the language family tree. In addition, the Uralic languages can now be added to the growing list of language families studied with quantitative methods. This will allow broader inferences as to global patterns of language evolution, and more language families can be included in constructing the tree of the world's languages. Studying history through language, however, is only one way to illuminate the human past. Therefore, thirdly, the findings of the thesis, when combined with studies of other language families, and those for example in genetics and archaeology, bring us again closer to an understanding of human history.

TIIVISTELMÄ

Monet maailman yli 7000 kielestä ovat syntyneet erkaantumisprosessin kautta. Tällöin yhdestä kielestä muotoutuu eri tekijöiden vaikutuksesta aikojen saatossa useampia kieliä. Kielten erkaantumiseen vaikuttavia tekijöitä on tutkittu aiemminkin ja myös laskennallisia menetelmiä käyttäen. Vielä on kuitenkin epäselvää mitkä kaikki tekijät voivat vaikuttaa kielten erkaantumiseen ja miten.

Tutkin väitöskirjassani kielten erkaantumiseen vaikuttavia tekijöitä. Lähestymistapani on monitieteinen, sillä sovellan laskennallisia evoluutiobiologian menetelmiä ja teorioita kieliaineistoon. Laskennalliset menetelmät mahdollistavat suurien aineistojen objektiivisen analysoinnin, kun taas evoluutiobiologisen lähestymistavan avulla voin muodostaa uudenlaisia tutkimuskysymyksiä ja käyttää uusia menetelmiä vastatakseni aiemmin esitettyihin kysymyksiin (esimerkiksi sukupuun muotoon liittyen). Tutkimuksessani keskityin selvittämään kielten erkaantumista ihmisen ekologian kannalta. Toisin sanoen olin kiinnostunut ympäristö- ja/tai kulttuuritekijöiden vaikutuksesta kielenpuhujiin ja siitä, voiko tämä kytkös olla osallisena kielten erkaantumisprosessissa. Tutkin kysymystä tämän prosessin kahdessa eri vaiheessa: sen alussa ennen kuin eriytyminen on kokonaan tapahtunut, ja sen jo tapahduttua. Murteiden eriytyminen vastaa prossessin alkuvaihetta, ja tutkin sitä suomen kielen murreaineistoa käyttäen. Tapahtuneita erkaantumisia tutkin sukupuista, joita tein uralilaisten kielten sanastoaineistosta. Koska uralilaisia kieliä ei ole aiemmin tutkittu vastaavanlaisin laskennallisin menetelmin eikä käyttämiäni populaatiogenetiikan menetelmiä ole käytetty aiemmin mihinkään murreaineistoon, testasin aluksi näiden menetelmien soveltuvuutta aineistojeni analysointiin.

Totesin biologisten menetelmien soveltuvan kieliaineiston analysointiin, sillä tulokseni vastasivat perinteisiä näkemyksiä sekä uralilaisen sukupuun muodosta että suomen murrejaosta. Lisäksi havaitsin, että erot ympäristöoloissa mahdollisesti vaikuttavat kielten erkaantumiseen. Tämä oli havaittavissa niin eriytymisprosessin varhaisissa vaiheissa murteiden välillä kuin myös koko kieliryhmän eriytymisiä tutkittaessa. Koska ihmisten tiedetään usein sopeutuvan vallitseviin ympäristöolosuhteisiin kulttuurisopeumien avulla, päättelin murretutkimusteni tuloksista, että juuri kieltenpuhujien kulttuurinen sopeutuminen paikallisiin ympäristöolosuhteisiin saattaisi toimia puhujapopulaatioita erottavana tekijänä ja täten kytköksenä ympäristöerojen ja kielellisen erkaantumisen välillä. Tämä voisi mahdollisesti selittää myös uralilaisten kielten erkaantumisia.

Väitöstutkimukseni tulokset tuovat uusia näkemyksiä kielten erkaantumiseen niin paikallisella kuin maailmanlaajuisellakin tasolla. Havaintoni ympäristöerojen mahdollisesta vaikutuksesta suomen murteiden muotoutumisessa herättää kysymyksen löytöni yleistettävyydestä myös muihin kieliin ja niiden murteisiin. Koska murteiden erkaantuminen on ensimmäinen vaihe kielen eriytymisprosessissa, on murteiden muotoutumista tutkimalla mahdollista myös selvittää, mitkä tekijät ovat aikaansaaneet maailmanlaajuisen kielten kirjon. Tästä syystä tarvitaan vastaavanlaisia tutkimuksia myös muiden kielten murteista. Esitän väitöskirjassani myös uralilaisten kielten laskennallisesti tehdyn sukupuun, jota voidaan verrata vastaavilla menetelmillä tehtyihin muiden kieliryhmien puihin. Tämän vertailun kautta on mahdollista selvittää onko kielisukupuiden muodossa jotain maailmanlaajuisia säännönmukaisuuksia, josta voi edelleen tehdä päätelmiä kieliin vaikuttavista lainalaisuuksista.

Ihmiskunnan historian ja esihistorian selvittäminen on haasteellinen palapeli, jossa eri tieteenalojen palasia yhteen sovittelemalla voidaan päästä lähemmäksi yleistä ymmärrystä menneisyydestä. Väitöstutkimukseni on pieni osa tätä kokonaisuutta, mutta yhdistelemällä havaintojani niin muista kieliryhmistä tehtyihin havaintoihin kuin myös esimerkiksi arkeologian ja genetiikan tuloksiin, olemme taas askeleen lähempänä tätä tavoitetta.

TABLE OF CONTENTS

LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following publications and manuscripts, referred to in the text by their Roman numerals.

I Syrjänen, K., **Honkola, T.**, Korhonen, K., Lehtinen, J., Vesakoski, O. & Wahlberg, N. 2013. Shedding more light on language classification using basic vocabularies and phylogenetic methods: A case study of Uralic. *Diachronica* 30: 323-352.

II Syrjänen, K.,*** Honkola, T.,*** Lehtinen, J., Leino, A. & Vesakoski, O. Applying population genetic approaches within languages: Finnish dialects as linguistic populations. Submitted manuscript.

III **Honkola, T.**, Vesakoski, O., Korhonen, K., Lehtinen, J., Syrjänen, K. & Wahlberg, N. 2013. Cultural and climatic changes shape the evolutionary history of the Uralic languages. *Journal of Evolutionary Biology* 26: 1244–1253.

IV **Honkola, T.**, Ruokolainen, K., Syrjänen, K.J.J., Leino, A.P.U., Tammi, I., Wahlberg, N. & Vesakoski, O. Evolution within a language: environmental differences contribute to dialect diversification. Manuscript.

*****Shared first authorship

Contributions to the original publications

	I	**II**	**III**	**IV**
Study question	OV, NW, KS, KK, TH	KS, TH, OV	TH, OV, NW, KK	TH, KR, OV, NW
Literature review				
-biology	TH, KS	KS, TH, OV	TH, NW	TH, OV
-linguistics	KS, KK	KS, TH	TH, KS	TH, KS
Material	JL, KS	JL, AL	TH, JL	IT, KS
Analyses	TH	TH, KS	TH, NW	TH, KR, OV, KS, IT
Interpretation of results	KS, TH, JL, OV, NW	TH, KS, OV, AL	TH, OV, NW, KK	TH, KR, OV, IT
Writing & commenting	KS, TH, JL, KK, OV, NW	KS, TH, AL, OV	TH, OV, KK, JL, KS, NW	TH, KR, IT, KS, AL, NW, OV

TH = Terhi Honkola, NW = Niklas Wahlberg, KS = Kaj Syrjänen, KK = Kalle Korhonen, OV = Outi Vesakoski, JL = Jyri Lehtinen, AL = Antti Leino, KR = Kalle Ruokolainen, IT = Ilpo Tammi

1. INTRODUCTION

Modern humans expanded from Africa ca. 70-50 000 years before present (YBP) (Soares et al., 2012, Wei et al., 2013) and it has been suggested that this expansion was induced by a shift in environmental conditions (Cohen et al., 2007, Scholz et al., 2007). After the original expansion out of Africa, humans have permanently inhabited all other continents except Antarctica. The Americas were the last ones to be populated ca. 20-15 000 YBP during and after the Last Glacial Maximum, when the sea level was low and the Beringian land bridge connected Eurasia and America (Goebel et al., 2008, Jobling et al. 2013). Climatic conditions also affected human populations in Europe, and climate has indeed been suggested to be the major driver of human population dynamics before, during and after the Last Glacial Maximum (Tallavaara et al., 2015), including the past 500 years (Zhang et al., 2011).

Humans have thus migrated considerably during the history of the species – often induced by environmental cues – and we have a wider global distribution than any other species. The success of humans as a species has often been connected to culture, which has enabled us to populate various types of environments ranging from rainforest to tundra (Mesoudi et al., 2004, Pagel & Mace, 2004). Language may be considered part of culture, and it has very likely played a role in the success story of the species. Language is nevertheless connected to human survival only indirectly; the impact of other cultural factors, such as techniques for hunting game or cultivating the land, or for protecting ourselves from the cold, has been more direct. Language can therefore be considered a neutral, or nearly neutral, marker of human cultural history (Mace & Jordan, 2011).

In this thesis, I have studied particular languages and their histories, from which I draw inferences as to the histories of the speaker populations. More specifically, my focus has been on the process of linguistic divergence, and on the effect on language speakers exerted in this process by environmental and cultural factors. Understanding the process of divergence is vital, as it has presumably played an important role in shaping the over 7000 languages existing in the world today.[1] In my approach, I underline that *Homo sapiens* is just another species existing in the biological realm and affected by its environment, which is why the role of environmental variation in the linguistic divergence process deserves to be studied.

I have explored these questions using a multidisciplinary approach and applying the quantitative methods and framework of evolutionary biology to linguistic data.[2]

---

[1] New languages may also appear through creolization, in which two languages (often a trading language such as English and a native language) become mixed, (e.g. Tok Pisin in Papua New Guinea) (Lewis et al., 2015).

[2] This thesis has been carried out as a part of the multidisciplinary BEDLAN (Biological Evolution and the Diversification of Languages) project. The project participants are mainly biologists and linguists in collaboration with geographers and historians.

Quantitative methods enable the analysis of large datasets objectively, and the framework of evolutionary biology provides theories and hypotheses with which to gain new insights into the process of linguistic divergence. Application of the evolutionary biology framework to study linguistic divergence is plausible as it may be seen as analogous with species divergences – a topic which has been studied extensively in evolutionary biology. Due to the novelty of this approach, I first evaluate the applicability of a biological methodology to linguistic data (**I**, **II**); I then apply a framework based on biological theory to the study of the linguistic divergence of speaker populations (**III**, **IV**). These two main themes, applying the methodology and applying the framework, are implemented at two levels: between languages, in the case of the Uralic language family (**I**, **III**), and within a single language, in the case of Finnish dialects (**II**, **IV**) (Table 1). I propose a distinction between these two levels corresponding to the two levels distinguished in biological evolution: macroevolution (between-species processes such as speciations and extinctions) and microevolution (within-species processes such as population divergence). I therefore refer to the processes examined in the Uralic languages as 'linguistic macroevolution' and those in Finnish dialects as 'linguistic microevolution'.

Table 1. Main themes (in columns) and levels of study (in rows) in different original publications (numbers in bold).

	Applying method	Applying framework
Between languages	**I**	**III**
Within a language	**II**	**IV**

In the following, I introduce certain central topics related to my thesis. Due to the multidisciplinary nature of this thesis, these sections cover a broad variety of disciplines. Reviewing several fields of research in detail is, however, beyond the scope of this thesis; I therefore focus strictly on the perspectives that are most important for my study. In particular, I introduce the recent wave of quantitative language studies, go through the proposed analogies on which these studies are based, and present the direction taken in the thesis compared to earlier studies (sections 1.1 and 1.1.1). I then provide a brief introduction to the particular nature of humans as a study species (section 1.2), explain the hypotheses related to species divergence (section 1.3), and summarize earlier discussion on the causes of linguistic divergence (section 1.4). Finally, I introduce the languages dealt with in the thesis – the Uralic language family (section 1.5) and Finnish dialects (section 1.6) – and present the aims of this thesis (section 1.7).

1.1. The new wave of quantitative language studies

My work is part of the recent wave of quantitative language studies, which was initiated a little over a decade ago (Pagel 2000a & b), and during which the number of applications of a modern biological methodology and framework to the analysis of language data has

increased considerably.[3] These studies fall roughly into three main groups: 1) studies focusing on untangling human prehistory through linguistic evidence, 2) studies focusing on the drivers of linguistic diversity, 3) studies focusing on language change. Below I briefly introduce these three groups.

In the first group, human prehistory, more specifically questions related to human dispersal, have been studied using phylogenetic and phylogeographic methods within various language families. In many of these studies, the timing of language divergences has also been quantitatively estimated to enable a better comparison with various types of data in drawing holistic inferences concerning past events (Gray et al., 2011). Together with the divergence times, the shape of the phylogeny and the branch lengths are also considered to indicate the speed of expansions (e.g. Gray et al., 2009). In individual studies, questions have related to the spread of languages in conjunction with agriculture (Diamond & Bellwood, 2003) in at least three language families, the Indo-European (Gray & Atkinson, 2003, Gray et al., 2011), the Bantu (Holden, 2002, Holden et al., 2005) and the Japonic (Lee & Hasegawa, 2011); the dispersal routes of the Bantu (Grollemund et al., 2015) and Arawak (Walker & Ribeiro, 2011) languages; the settlement of the Pacific (Gray & Jordan, 2000, Gray et al., 2009); the origins and dispersals of the Semitic languages (Kitchen et al., 2009); the origins of the Indo-European language family (Bouckaert et al., 2012); and the migrations of Na-Dene language speakers (Sicoli & Holton, 2014).

In the second group of studies, patterns of biological and linguistic diversity have been found to coincide; that is, both linguistic and biological diversities are at the highest level close to the equator and decrease towards the poles (e.g. Mace & Pagel, 1995, Moore et al., 2002).[4] This has prompted researchers to study why the patterns coincide – a question which is ultimately related to the drivers of linguistic diversification (Gavin et al., 2013). Proposed drivers include for example environmental features (e.g. the length of mean growing season (Nettle, 1999)), but also topographical (e.g. oceanic barriers (Lee & Hasegawa, 2014)) and sociocultural factors (e.g. political complexity (Currie & Mace, 2009)) (reviewed by Gavin et al. 2013). The findings of earlier studies, however, have not always been conclusive; there is some divergence for example between the findings related to the mean growing season arrived at by Nettle (1999) and by Gavin and Sibanda (2012). Further, in cases where correlations do appear, there are several alternative explanations for them; Michalopoulos (2012), for example, presents five possible mechanisms underlying the emergence of ethnolinguistic diversity. Studies concerned with the drivers of linguistic diversity thus still have a number of issues to resolve (Gavin et al., 2013).

The third group of language evolution studies focuses more on languages than on their speakers (although a language of course cannot exist without its speakers);

[3] For a review of the intertwined and multiphase history of linguistics and biology see Atkinson and Gray (2005).

[4] Just as there are various ways to specify biological diversity, linguistic diversity too may be referred to for example as *language richness*, defined as the "the number of languages within a given area" or as *phylogenetic language diversity* i.e. "the minimum total length of all branches needed to span a set of taxa on a phylogenetic tree" (Gavin et al., 2013).

their aim is to resolve whether regularities found in genetic material are also found in linguistic material. This has been confirmed to be the case at least for certain evolutionary features. Some examples: increased frequency of word use is connected to lower rates of lexical replacement (Pagel et al., 2007, Calude & Pagel, 2011) (cf. the stability of the functionally important gene sequences (Jobling et al., 2013));[5] global phonemic diversity follows a serial founder effect model and decreases with distance from Africa (Atkinson, 2011) (cf. the decrease of genetic diversity with distance from Africa (Prugnolle et al., 2005, Ramachandran et al., 2005)); population size affects the rates of gain and loss of new words (Bromham et al., 2015) (cf. the effect of the population size on rates of gain and loss of mutations (Hamilton, 2009)); and languages have been found to evolve in punctuational bursts (Atkinson et al., 2008) (cf. rapid bursts of change in species divergences (Pagel et al., 2006)).

Of the categories listed above, article **III** focuses on the first one, and article **IV** on the second. The common feature in these studies is that I infer aspects of human prehistory and past processes from modern linguistic material, which I use as a proxy for human populations similarly to the use of genetic data.[6] At this point, I should emphasize that I am *not* concerned in this thesis either with the general mechanisms of language change (i.e. innovation and acceptance) or with particular ones such as grammaticalization, borrowing or semantic change. Nor am I interested in the environmental factors affecting languages more directly: for instance whether populations living at a coastal site have a vocabulary filled with precise terms for various marine species (Sapir, 1912), or whether human sound systems are affected by environmental factors (Everett et al., 2015). Rather, I am concerned with the factors that affect speaker populations themselves, such as factors affecting their subsistence in a certain environment, whereby language is eventually affected as well. Environmental heterogeneity, for example, may lead to communicatively isolated speaker populations, in which processes of language change may gradually lead to language differentiation.

My studies, similarly to earlier ones, are based on the proposed analogies between biological and linguistic evolution, the range of which has been discussed for example by Croft (2000) and Pagel (2009) (see also **II**).[7] Here I introduce the analogies most relevant to the scope of this thesis.

Biological evolution may be defined as "change in the properties of groups of organisms over the course of generations" (Futuyma, 2009). This definition has three built-in parts, which form the core of evolution: heritability, variation and change-causing forces. Firstly, heritability forms the continuity between the past and the

[5] Lexis refers to the vocabulary of a language.

[6] Especially in the case of **IV**; in dealing with events of greater historical depth, as in **III**, there is a greater possibility of language shifts (i.e. a language is transmitted to a population which originally spoke another language).

[7] Certain aspects of the analogy between languages and species were touched on already by 19th-century scientists such as Lyell (1863), Darwin (1871) and Paul (1886), indicating a long-lasting interest on the topic.

present.[8] Heritable units sampled at present, such as words or phonemes in the case of language or genes and nucleotides in the case of species, can be used to infer past events, as the changes in these have accumulated in the course of time. Secondly, variation is the raw material for evolution. Without variation, transmission from one generation to the next would always occur similarly and change would not happen. Mutations create variation in genetic material, while innovations and errors in the transmission process produce variation in languages.[9] Thirdly, once variation exists, its frequency varies in time through the evolutionary forces of selection and drift. The analogy between linguistic and natural selection is perhaps seen as the most problematic (Itkonen, 2003): the difference is that speakers themselves select the preferred variants, while in nature the organism itself does not make a decision. In addition, linguistic selection by way of social acceptance can be considered to have a direction and a goal (i.e. it is teleological), while natural selection has neither (Itkonen, 2003). Drift (i.e. neutral processes) is more alike in these two fields, as in both of them variation accumulates over time and the frequencies of alternative variants vary randomly (Levinson & Gray, 2012). To reiterate: here I have studied existing linguistic variation, not the language-specific mechanisms of linguistic change (linguistic selection and drift). This should not be confused with processes affecting language speakers themselves (selective processes and drift affecting humans). Considering the scope of the thesis, the mismatch in the 'selection analogy' should thus not pose a problem.

1.1.1. Taking the wave to unexplored areas

The recent wave of quantitative language studies has until now largely focused on phenomena occurring above the language level, which I designate as 'linguistic macroevolution'. Biological processes taking place within a species are referred to as microevolution, which is why I designate processes taking place within a language as 'linguistic microevolution'. There has been some earlier discussion of linguistic microevolution, but these have been restricted to sociolinguistic aspects of language change (Levinson & Gray, 2012). In this thesis, I approach linguistic microevolution from an ecological perspective. In other words, I study dialect divergence by taking into account the physical surroundings of the speakers, as these factors also affect the spatial patterns of linguistic variation, and thereby dialect divergence as well.

I have studied linguistic microevolution by applying methods of population genetics to data on intra-lingual variation. This is not the first time population genetic methods have been applied to language data; Dunn et al. (2008), Reesink et al. (2009) and Bowern (2012)

[8] There are notable differences between biological and linguistic inheritance: in biological organisms, the transmission of genetic material occurs most commonly from parent to offspring (although, to note the variety existing in the biological realm, horizontal gene transfer also occurs), while with language it is the speaker community that plays the major role in the transmission process. This difference, however, does not erase the idea of heritability, as the most important thing – transmission from one generation to the next – still occurs (see also **II**).

[9] One difference between mutations and innovations is that mutations are always random, while innovations appear to serve a certain purpose such as the need for a certain lexical item.

used a population genetic clustering method to study ambiguities in language classification, and to determine the extent to which different source populations have contributed to the development of different languages. These studies, however, covered several languages; they were thus not 'microevolutionary', which entails focusing on a single language.

My interests lie in validating the applicability of population genetic tools to dialect data. I begin by applying a basic population genetic clustering method to dialects (**II**). This, to my knowledge, is the first time population genetic methodology has been applied to dialect data (in earlier studies clustering was used for languages (Reesink et al., 2009) or to a combination of languages and dialects (Bowern, 2012)). In addition, I investigate the drivers of linguistic divergence from the perspective of human ecology (**IV**). This means applying the hypotheses and methods of population biology to the study of the linguistic divergence of speaker populations. Compared to the approaches used in the studies mentioned in section 1.1 (second group), the advantage of my approach is that the underlying processes can be inferred from the emerging patterns (see section 1.3), leaving perhaps less room for speculation as to possible causal pathways. Another reason why studying the drivers of linguistic divergence within a language is important is that this is the level where the divergence process of speaker populations and languages is initiated.

In addition to the abovementioned issues related to linguistic microevolution, this thesis extends the scope of work on linguistic macroevolution to the Uralic language family, which has not previously been explored with quantitative phylogenetic methods. I therefore test the applicability of phylogenetic methods on the Uralic languages (**I**). By adding a new language family to the pool of language families studied with quantitative phylogenetic methods (see section 1.1), inferences as to global patterns of language evolution, or the tree of the world's languages, for example, are again one step closer. I also propose a more ecological perspective for the study of linguistic divergence taking into account changes in climate, and the effect of these changes on human populations and thereby also on their languages (**III**).

In sum, the central theme of the thesis is the study of linguistic divergence through human ecology, at the levels of linguistic macro- and microevolution. The application of biological hypotheses concerning the factors inducing the genetic divergence of species offers a new perspective on questions of linguistic divergence.

1.2. Humans as a species

There are currently more than 7 billion humans, speaking more than 7000 different languages (Lewis et al., 2015). These 7 billion people would be able to reproduce with each other,[10] but on average they would not be able to communicate with each other in an intelligible way.[11] This is an intriguing mismatch, which can be illuminated by looking

[10] The capability to reproduce is one way of determining which individuals belong to the same species. This is the 'biological species concept' (Mayr, 1942).

[11] Mutual intelligibility is one way of determining speakers of the same language, and the one applied here. However, languages may also be determined on political grounds: Norwegian, Swedish and Danish, for example, are at least to some extent mutually intelligible but politically different languages (Chambers & Trudgill, 1998).

at the origins of the process of linguistic divergence through the joint effort of several disciplines.

Histories of human populations can be studied on the basis of, for example, genetic, linguistic, anthropological, archaeological and historical data, depending on the time-depth in question; by combining independent lines of evidence, a multidisciplinary view of the human past can be obtained (Jobling et al., 2013). Combining the evidence from different fields, however, is challenging. First of all, different types of genetic material may provide conflicting results. Secondly, genetic material does not necessarily go hand in hand with linguistic material, as genes may be transmitted when linguistic material is not and vice versa. Thirdly, archaeological remnants are location-dependent signs of human existence, but are commonly void of evidence of the linguistic or genetic background of the people who left them. Thus, even though there are more data on humans than on any other species, and although the variety of data types is also greater for humans, drawing inferences as to the past is filled with challenges.

While the human species has its peculiarities, such as diverse languages, it is only one species among others. This means that, similarly to other species, humans too are dependent on their environment for their daily survival. Humans may adapt genetically to their local conditions, similarly to other biological organisms. However, due either to the challenges of convincingly detecting genetic adaptation or to its rarity (or possibly both), there are only a few examples where human populations can be said to have genetically adapted to their environment (Jobling et al., 2013). One remarkable example of this is that of the Tibetan highlanders, who have been found to be genetically adapted to living at high altitudes (Beall et al., 2010, Simonson et al., 2010). It has also been proposed that the emergence of human culture has been enabled through genetic changes in early humans (e.g. Somel et al., 2013); but as cultures and genes have been found to coevolve (Laland et al., 2010), it has been difficult to determine which came first – genetic changes enabling culture or cultural innovations directing genetic evolution (Fisher & Ridley, 2013).

Defining culture has been found to be challenging (Laland et al., 2010),[12] but one way of defining it is as "acquired information such as knowledge, beliefs and values, that is inherited through social learning, and expressed in behavior and artifacts" (Mesoudi et al., 2004). Cultural transmission has been found to take place for example in several primates (e.g. the macaque (Kawai, 1965)). It is, however, less common for culture to be cumulative in other species than it is in humans, where modern cultural variants are the product of historical build-up processes (Tomasello, 1999).[13] Human culture, therefore, has the three focal parts of evolution (see section 1.1): it is heritable, there occurs innovations and the frequencies of different cultural variants change in time. Hence it is not surprising that the analogies of biological evolution have been extended to cover cultural evolution as well (Mesoudi et al., 2004).

[12] The difficulty of defining culture has been compared to that of defining species (Ehrlich & Levin, 2005).

[13] Cumulative culture has been found for example in killer whales, whose dialect has been found to be cumulative (Filatova et al., 2013).

Related to the proposed evolutive character of culture, the question arises whether regularities found in biological evolution can be found in cultural traits as well. One topic of interest has been that of rates of change, based on findings showing that functionally important traits change more slowly than traits which are less selectively constrained; in other words, functionally important gene sequences change more slowly (Jobling et al., 2013), and the rate of lexical replacement is slower in frequently used items (Pagel et al., 2007). Recently, this has also been found to be the case with cultural traits. In a global comparison, cultural traits which were more directly connected to the local environment,[14] such as a subsistence economy or roofing materials, were found to evolve more slowly than traits related to social structures, such as class stratification and domestic organization (Currie & Mace, 2014).[15] Additionally, in the specific case of Polynesian canoes, the functional traits of the canoe were found to change more slowly than its stylistic features; the functioning of the boat was presumably more important for the success of the sailing voyage than were the ornaments (Rogers & Ehrlich, 2008). Such important cultural features, often related to subsistence, are commonly passed on vertically from generation to generation (Mace & Jordan, 2011). Limits on this process, however, are imposed by environmental conditions and their constraints; for example plant-based subsistence was found to be associated more with geographical proximity than with phylogenetic relationship (Mace & Jordan, 2011) – a pattern also noted in connection with the traditional use of medical plants (Saslis-Lagoudakis et al., 2014). The horizontal transmission of cultural features may thus also be adaptive, if traits are transmitted to serve a certain purpose (Mace & Jordan, 2011).

In sum, humans can react and adapt to changing environmental conditions by way of culture much more rapidly than by way of genes, which is most likely the reason both for the extensive cultural variety of humans and for our success as a species. Language is part of culture, but it is not part of the cultural core with which humans adapt to their environment. Rather, language can be seen as a neutral marker of cultural populations, which is why studying language and its divergence can also illuminate the cultural histories of speaker populations and their separation from each other. In my work, I consider culture very broadly, including variables which could be better described as demographic. This is because my primary interest is in the effect of environment on language speakers, and I consider it most important to distinguish between environmental variables and all others.

1.3. Hypotheses on population divergence, speciation and macroevolutionary events

Macroevolution involves processes which occur above the species level, and over great periods of evolutionary time (Futuyma, 2009). Indeed, speciations and extinctions, the

[14] Cultural traits closely connected to the environment were called "cultural cores" by Steward (1955), a pioneer in the field of cultural ecology.

[15] The authors, however, mention that these relative rates of change assume stable environmental conditions; if environmental conditions change, environment-related features presumably start to vary more.

two main macroevolutionary processes, generally take up numerous generations and long time-spans. These processes, however, are initiated at the population level, where changes occur locally and over a shorter time interval (referred to as microevolution). For this reason, drivers of speciation are often studied at the population level, where populations are not yet reproductively isolated but where the first step towards divergence has potentially been taken (Futuyma, 2009). The micro- and macroevolutionary levels are thus inseparably connected, even though macroevolution cannot be explained by microevolution alone (Reznick & Ricklefs, 2009).

At the microevolutionary level, population genetic structure in nature is considered to be mainly shaped by two factors: geographical distance and environmental differences (Sexton et al., 2013). Populations located geographically close to each other are more likely to exhibit gene flow, making them more alike than when populations are located at a distance from each other, when gene flow may be limited. Geographically isolated populations may then be exposed to genetic drift, the outcome of which depends on various factors, such as the size of the population and the pool of alleles it contains. This may result in differing allele frequencies in these populations and a pattern of isolation by distance (IBD) (Wright, 1943), where genetic distances between populations increase with geographical distance independent of environmental conditions. Where IBD occurs, it is considered to signal that the major role in structuring the spatial pattern of genetic variation is played by neutral processes rather than adaptation (Orsini et al., 2013).

Populations occupying differing selective environments may develop local adaptations despite on-going gene flow. However, if the environments are different enough, gene flow may be restricted due to the lower fitness of incoming individuals or alleles. In this case, the isolation of populations is caused by the environment, and a pattern of isolation emerges which has been called isolation by environment (IBE) (occasionally also referred to isolation by ecology or isolation by adaptation) (Shafer & Wolf, 2013). Due to the way the pattern of IBE takes shape, it is considered to indicate that adaptive processes have occurred and that ecological speciation may be underway (Shafer & Wolf, 2013).

Speciation through selection and adaptation to differing environmental conditions was one of Darwin's main ideas; now termed ecological speciation, it has become a popular research topic with an increasing amount of evidence for its existence (Shafer & Wolf, 2013). Ecological speciation can occur in either sympatry or allopatry, and thus is not dependent on the geographical context (Schluter, 2001), which is a central feature in traditional, geography-based modes of speciation (allopatry-parapatry-sympatry) (Schluter, 2001, reviewed in Futuyma 2009).[16] Ecological speciation is closely connected to the ability of individuals to obtain resources and reproduce in a certain type of environment: divergent selection in resource acquisition in different environments may lead to a critical reduction of gene flow between different environments, which in turn increases the likelihood of reproductive isolation (Rundle & Nosil, 2005). As the features which are important for resource acquisition may not be directly connected to

[16] In allopatry, populations are separated by a physical barrier. In parapatry, no physical barrier exists between adjacent populations. In sympatry, the diverging species occupy the same geographical area.

reproduction, reproductive isolation, which essentially completes the process of speciation according to the biological species concept, may arise as a by-product (Futuyma, 2009). Thus, while the pattern of IBE is connected to environmental speciation, IBD refers to parapatric speciation. In the latter case, geographical distance reduces the extent of gene flow and allows the emergence of novel mutations in distant populations, which, if separated for long enough, may become fixed and cause reproductive challenges (Rundle & Nosil, 2005, Shafer & Wolf, 2013)

At the macroevolutionary level, there are two contrasting views of the main drivers of evolutionary change; that is, whether this change is mainly driven by biotic interactions, such as predation or competition (the Red Queen model (Van Valen, 1973)), or by changes in the physical environment, such as the climate (the Court Jester model (Barnosky, 2001)) (for the reformulation of the Red Queen model in this thesis to suit human and language data, see section 1.4). These two models have been thought to play a role at different geographic and temporal scales, and are therefore not considered to be mutually exclusive: biological interactions are important locally and at short temporal scales, while climatic changes and tectonic events are thought to have more large-scale effects (Benton, 2009). Nevertheless, one possible reason for the conclusion that biotic factors have more local effects and abiotic factors more global ones is that it may be easier to connect mass extinctions or speciation booms detected from paleontological findings and from molecular systematic work to changes in climate (e.g. Peña & Wahlberg, 2008) than to assume them to be the result of biological interactions. Recently, this view of scale differences has been challenged, and the hypothesis of long-term evolution through biotic interactions has become more plausible (Voje et al., 2015).

These were the hypotheses used in my work. In **III** I apply the macroevolutionary Red Queen and Court Jester hypotheses in studying the drivers of the divergence of the Uralic languages. In **IV** I apply the hypotheses of population divergence (IBD, IBE, extended to cover cultural and administrative factors in the form of isolation by culture (IBC) and isolation by administration (IBA)) in studying the divergence of Finnish dialects.

1.4. Linguistic divergence

Linguistic divergence boils down to the question of what breaks up a single linguistic unity (to the extent that such is possible) into two or more geographically separate groups, i.e. dialects, which in time may become mutually unintelligible languages. Linguistic divergence thus originates at the dialect level, analogously to the origination of species divergences at the population level; as with populations and species, the level of linguistic microevolution is connected to that of linguistic macroevolution through linguistic divergence.

Language change is a prerequisite for linguistic divergence. It occurs for example through borrowing, lexicalization, grammaticalization, sound change and semantic change. The motivations behind these changes may include for example prestige, the principle of least effort, the principle of ‘one meaning – one form’ and/or various social

factors (Anttila, 1989, Trudgill, 2011).[17] These factors do induce language change, but they do not by themselves, as far as I know, induce linguistic divergence.

In addition to language change, linguistic divergence requires an additional isolating or group-enforcing factor;[18] to quote Robert Foley,"[l]anguage diversification is … fundamentally a process of inter-group boundary formation" (Foley, 2004). Since speakers are dependent on their surrounding environment, the ecological perspective in studying linguistic divergence becomes focal, and it is these isolating and/or group-enhancing factors that I focus on here. Once isolation or group formation has taken place, the processes of language change can occur differently in these separate groups, ultimately perhaps making the languages mutually incomprehensible.

Boundary formation through isolation or group-enforcing, then, is important for the divergence process. Isolating forces separate communicative groups by physical isolation through geographical barriers (e.g. mountains or seas), environmental barriers (e.g. dense forest or bog), man-made barriers (e.g. state or other administrative borders), or through plain geographical distance (Paul, 1886, Lee & Hasegawa, 2014). With the exception of geographical distance, these are parallel to the allopatric speciation mode, where the existence of barriers is needed to initiate the divergence process (Mayr, 1963). Geographical distance may induce linguistic divergence analogously to parapatric speciation (cf. section 1.3): new linguistic innovations appear in distant populations, and – if separated long enough without sufficient contact – the languages may become mutually unintelligible.

Group enforcement, on the other hand can be said to act through human will. There is no physical hindrance preventing mobility; people themselves decide where to move and where not, and with whom to communicate. Reasons for preferring to stay in one place and communicate with one's own group rather than an unfamiliar one may be for example cultural, social or political (Paul, 1886, Rapola, 1962, Britain, 2002). Recently, it has also been suggested that environmental heterogeneity induces linguistic diversification (Michalopoulos, 2012, Gavin et al., 2013). This has been suggested to take place through cultural specialization in resource acquisition strategies specific to certain environmental conditions. When specialization in different environments occurs, boundaries may also emerge between populations in different types of environment; this may then, as a by-product, lead to linguistic differentiation. Various similarities can thus be seen in the processes whereby group enforcement and ecological speciation lead to linguistic and species divergence (section 1.3).

On the whole, linguistic divergence is most likely the result of several different factors acting simultaneously within a language. Nevertheless, quantitative studies have thus far

[17] Prestige: for example, speakers imitate the language use of a culturally dominant group. The principle of least effort: for example, speakers simplify the pronunciation of a word. The principle of 'one meaning – one form': an 'ideal' situation where each meaning is expressed with one specific form and each form expresses one specific meaning. Social factors: for example, low social stability and large amount of contact with other groups may increase the rate of linguistic change in a speech community.

[18] Isolation and group boundary formation differ in that group boundary formation acts simultaneously in two ways, bringing individuals within the group closer and individuals outside the group further away. With isolation, no such distinction exists.

been conducted on a broad inter-language level (Gavin et al., 2013), where I believe the contributing factors acting on linguistic divergence on the intra-lingual level cannot be detected. Furthermore, in quantitative studies conducted within a single language, the focus has been on only one specific contributing factor, such as geographical distance (e.g. Heeringa & Nerbonne, 2001, Nerbonne, 2010). Studies are therefore needed in which linguistic divergence is investigated within a single language while taking into account several factors simultaneously.

I apply the biological micro- and macroevolutionary hypotheses presented in section 1.3 to the study of linguistic divergence. At the level of linguistic microevolution, I examine the divergence of Finnish dialects and quantify the extent to which geographical distance (IBD), and differences in cultural conditions (IBC), environmental conditions (IBE), and administrative history (IBA) explain linguistic differences within the Finnish language. I am thus dealing with several factors simultaneously, allowing me to estimate the relative contributions of these factors to the divergence process (**IV**). At the level of linguistic macroevolution, I examine the divergences of the Uralic languages. I transform the Red Queen hypothesis to better suit human-related data by considering biotic interactions as cultural interactions. As the abiotic variable related to the Court Jester hypothesis, I study temperature changes in relation to language divergence (**III**). To sum up: at both levels, I examine the roles played by the physical environment (IBD and IBE within a language; temperature between languages) and the 'biotic' environment (IBC and IBA within a language; cultural interactions between languages) in the process of linguistic divergence.

1.5. The Uralic languages and the cultural and climatic conditions of their speaker area

I apply quantitative phylogenetic methods and a macroevolutionary framework for the first time to the Uralic language family. The Uralic language family consists of more than forty languages, currently spoken by about 25 million speakers across northeastern Europe and Siberia (Abondolo, 1998, Salminen, 2007, Janhunen, 2009). Most of these languages are minority languages, spoken in Russia, Finland, Sweden, Norway and Estonia by some tens or hundreds of thousands of speakers or even fewer (Korhonen, 1991); some are already extinct (e.g. Livonian). Only Hungarian, Finnish and Estonian are spoken by more than a million speakers each and have the status of a national language (Korhonen, 1991).

The Uralic languages have been studied extensively in historical comparative linguistics for over a century (Hovdhaugen et al., 2000).[19] All these languages presumably evolved from Proto-Uralic (i.e. the common proto-language), supposedly spoken in the Volga-Kama area located within the borders of present-day Russia (Salminen, 1999, Häkkinen, 2009).[20]

[19] One of the pioneers in the field of Uralistics was M. A. Castrén, who collected data from Uralic language speakers during 1838-1848. For his inspiring trips to the north, see Castrén (1953) (Finnish translation of his travel journals).

[20] Other locations for the Proto-Uralic homeland have also been suggested, such as central Siberia (Janhunen, 2009).

The estimated divergence time of the Uralic proto-language into its immediate daughter branches varies from 7000-6000 YBP (Korhonen, 1981, Sammallahti, 1988, Janhunen, 2000) to 5000-4000 YBP (Kallio, 2006, Häkkinen, 2009, Janhunen, 2009), and the timings of the successive branchings thus vary as well. In addition to the debate over the divergence times of the Uralic languages, differing views have been proposed concerning the shape of the Uralic phylogeny. According to the textbook interpretation, the Uralic language phylogeny is fairly tree-like, with binary branchings (Korhonen, 1981). Others, however, have proposed classifications with polytomies (i.e. divergence into more than two branches simultaneously). These phylogenetic hypotheses vary from extreme polytomies, with all the main groups originating directly from Proto-Uralic (Häkkinen, 1983, Salminen, 1999, 2007, Saarikivi, 2011), to more conservative versions, where certain proto-languages show multiple simultaneous divergences (Kulonen, 2002, Michalove, 2002).

After the beginning of the Holocene, ca. 11 000 YBP, the boreal Northeastern Europe experienced an increase in temperature, peaking at the Holocene thermal maximum ca. 6500 YBP, when the temperature was 3.5 °C higher than at present. Subsequently, the temperature declined for a few thousand years, after which it increased again slightly (Kremenetski et al., 1997, Davis et al., 2003, Väliranta et al., 2003, Heikkilä & Seppä, 2010). Culturally the hypothesized speaker area of the early Uralic languages was inhabited by two main consecutive cultures: the Lyalovo culture (ca. 7000-5650 YBP) and the Volosovo culture (ca. 5650-3900 YBP) (Carpelan & Parpola, 2001). In addition to these, several other archaeological cultures have also been identified in Northeast Eurasia dating to the last 10 000 years (Carpelan, 1999, Carpelan & Parpola, 2001).

Studies on the Uralic languages have largely been non-quantitative, but some quantitative attempts to classify the Uralic languages have also been made (e.g. Taagepera, 1994). These quantitative studies, however, have not always been credited due to criticism directed at the field in general.[21] Here I apply a modern quantitative approach and a framework of linguistic macroevolution to the study of the Uralic languages. This language family constitutes an ideal case for testing the applicability of the method, as the family is relatively small compared for example to the quantitatively studied Indo-European and Bantu language families, with several hundred of languages each. In addition, the Uralic languages have been extensively studied with traditional linguistic methods, and it is therefore possible to compare the shapes of the phylogeny and divergence times obtained in quantitative and non-quantitative studies (**I**, **III**).

1.6. Finland in a linguistic, cultural and environmental context

Finnish belongs to the Finnic group of the Uralic languages, and currently has some five million speakers, the majority of them located in Finland (Lewis et al., 2015). This is the result of a more or less steady increase in population size; some 250 years ago Finland

[21] Lexicostatistics, especially in the form of glottochronology (developed by Morris Swadesh in the 1950s), was widely criticized, for example because the glottoclock assumed the rate of linguistic change to be constant, which has been found not to be the case. The development of the field largely came to a standstill for decades, due to this resistance.

still had only around 400 000 inhabitants (Haapala, 2007). There are differing views as to when and from where the speakers of Pre-Finnish arrived in Finland (e.g. Heikkilä, 2014) but in the Iron Age, around 1000 YBP, there were four main areas of settlement in Finland (the southwestern coast, the lake district in Häme, the Mikkeli area, and the coast of Lake Ladoga) (Lehtinen, 2007, Virrankoski, 2012). The settlers are considered to have been speakers of Finnish, and the modern dialects are thought to have emerged from these four groups (Lehtinen, 2007).

Systematic research on Finnish dialects is considered to have begun around the nineteenth century (Hovdhaugen et al., 2000). The Finnish language is traditionally divided into two main dialects, eastern and western (e.g. Rapola, 1962). Three-way divisions, however, have also been proposed (e.g. by Lenqvist 1777 in Rapola, 1962, Leino et al., 2006). Traditionally, the eastern and western dialect groups are subdivided into a total of eight dialects (e.g. Itkonen, 1964). Dialectal differences, which during the eighteenth century were still prominent, have leveled off since the nineteenth century, due for example to urbanization and industrialization.

Finnish dialectology was largely non-quantitative until the end of the millennium, since when various quantitative methods have been applied to Finnish language data both as a whole (Wiik, 2004, Leino et al., 2006, Hyvönen et al., 2007, Leino & Hyvönen, 2008) and with a focus on a specific dialect (Palander et al., 2003). Earlier quantitative studies have utilized various methods, none of which, however, have been based on models of population biology. When the results of these quantitative studies were compared to the traditional ones, the match was reasonably good (Leino et al., 2006, Hyvönen et al., 2007, Leino & Hyvönen, 2008). The most notable difference was that the lexical data supported the east-west-north trichotomy; this was not the case with morphophonological data.

Just as Finland is divided linguistically between east and west, the speakers of these dialects are genetically quite different from each other (Salmela et al., 2008, Neuvonen et al., 2015) – more different than for example the British from the Germans (Salmela et al., 2008). In addition, there are also cultural features which clearly have separate eastern and western variants. Examples include different forms of boat building, various food items (including cheese and sour milk, bread and pasties), vehicle types and agricultural traditions (Suomen Maantieteellinen Seura, 1929). Differing cultural features thus cover many areas of life. Compared to these genetic, cultural and linguistic divisions, differences in environmental features between the east and west are less obvious. Environmentally, most of Finland is located in the boreal forest zone; the southernmost parts belong to the hemiboreal zone and the northernmost to hemiarctic zone (Kersalo & Pirinen, 2009). In general, however, conditions become harsher in moving from southwest to northeast; for example snow-depth increases quite steadily and the mean temperature decreases (Kersalo & Pirinen, 2009).

In my work I take a new approach to quantitative dialect studies. I apply a microevolutionary framework and methodology to data on Finnish dialects, and in general adopt a more ecological perspective on the study of intra-lingual variation by taking into account environmental and cultural factors affecting speaker populations. Finnish makes

an excellent object of research as it has been widely studied by the methods of traditional dialectology. This makes it possible to estimate the applicability of population genetic methods, which are new in the field (**II**). In addition, the availability of large datasets of various cultural and environmental features makes it possible to assess the role of environmental and cultural features in the process of dialect divergences (**IV**).

1.7. Aims of the thesis

This thesis has three main aims which are somewhat intertwined with each other. Starting from the most practical one, I aim to bring new quantitative methods and biological frameworks to Uralistics and to quantitative dialectology. To do this, I test the applicability of the methods by comparing my results to those obtained in earlier studies (**I**, **II**). Secondly, I aim to identify the factors contributing to linguistic divergence, focusing specifically on the role of environmental and cultural factors in the process. I do this by applying a biological framework and hypotheses to language data at the levels of both macro- and microevolution (**III**, **IV**). The third and the most extensive aim is to unravel the histories of human populations with the aid of linguistic evidence. By identifying the factors which have contributed to linguistic divergence, it is possible to understand which factors affect the speaker populations in general and which therefore could have played a role in shaping not only patterns of linguistic diversity, but also those of genetic diversity.

2. MATERIALS AND METHODS

In this section I briefly outline the focal aspects of the materials and methods I have used. A more detailed description of the data and methods used will be found in the respective articles referenced here.

2.1. Data sets

Studies on linguistic macroevolution were conducted with lexical data for the Uralic languages (**I, III**). For **III**, information on temperature changes and historical cultural events was gathered from several sources. Linguistic microevolution was studied using dialect data for Finnish (**II, IV**). For **IV**, I also used data covering several extralinguistic variables.

2.1.1. Lexical data of the Uralic languages (I, III)

The analyses in **I** and **III** are based on data for seventeen Uralic languages in order to provide good coverage of all the traditional Uralic subgroupings:[22] Finnish, Karelian, Veps, Estonian, Livonian, North Saami, Ume Saami, Skolt Saami, Erzya, Meadow Mari, Komi, Udmurt, Hungarian, Northern Mansi, Eastern Khanty, Tundra Nenets and Selkup (for more information on the data collection and the subgroupings of these languages, see **I**). The geographical distribution of these languages is shown in Fig. 1.

Lexical data, more specifically core vocabularies, are commonly used to construct quantitative language phylogenies (e.g. Gray & Atkinson, 2003, Lee & Hasegawa, 2011, Grollemund et al., 2015). The core vocabulary consists of the most basic items of the lexicon, such as the lower numerals, pronouns, and nouns referring to parts of the body, i.e. words which presumably occur in every language. These items are also considered to be resistant to borrowing, making them useful for tracing historical relationships between languages. Morris Swadesh compiled two basic vocabulary lists in the 1950s: a 200-item list (Swadesh, 1952) which he revised later to a 100-item list (Swadesh, 1955). These lists have commonly formed the basis of lexicostatistical and later quantitative phylogenetic studies, but other vocabulary lists have also been compiled, for example to better fit certain language families, such as the CALMA list – Culturally and Linguistically Meaningful for the Andes – proposed by Heggarty (McMahon & McMahon, 2005) or as being more empirically based in general, such as the Leipzig-Jakarta list (Tadmor, 2009).

[22] The lexical data were collected as part of the BEDLAN project, and are currently being prepared for publication as part of the UraLex project.

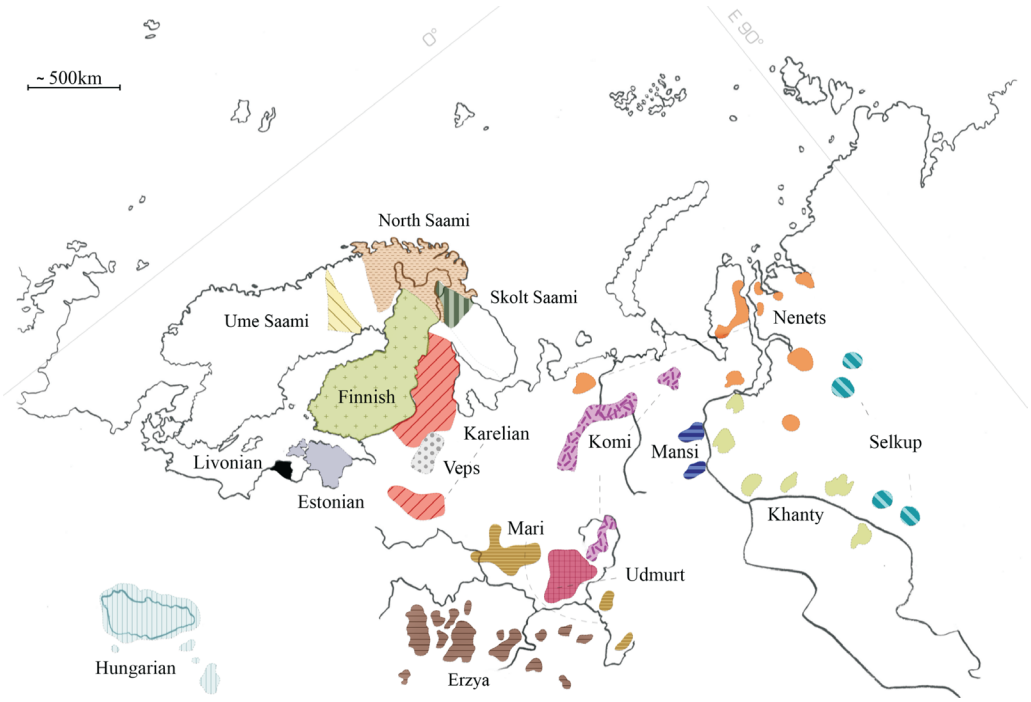

Figure 1. Map of the Uralic languages used in **I** and **III**. Map compiled from Abondolo (1998).

My studies were based on three core vocabulary lists: the 200-item Swadesh list (Swadesh, 1952), the 100-item Swadesh list (Swadesh, 1955), and the Leipzig-Jakarta list, taking the 100 most stable meanings from this list (Tadmor, 2009). These lists overlap to a considerable extent (i.e. the same items occur in different core vocabulary lists), and the total number of unique meanings in the full dataset is thus 226. Words representing these 226 meanings were collected for each of the seventeen languages studied (e.g. the items corresponding to 'fish' were *kala* in Finnish; *hal* in Hungarian; *kol* in Meadow Mari; *xalya* in Tundra Nenets) (Table 2). The chief sources for compiling the dataset were dictionaries (both common and etymological ones). In addition, the resulting word lists were checked by native speakers or experts on the language in question. Etymological dictionaries were used to determine whether the words carrying a certain meaning in different languages stem from a common ancestor (for a list of the dictionaries and informants used, see **I**). In choosing the word to represent a meaning, a strict semantic correspondence was required. In other words, the words needed to represent the meaning as precisely as possible. For example, 'sour milk' in Finnish is *piimä,* which shares a common ancestor with Estonian *piim.* The latter, however, refers to milk and not to sour milk, which is why Finnish *piimä* and Estonian *piim* do not correspond semantically. If the item 'sour milk' had been included in the basic vocabulary list (it is not), the Estonian word representing it could be *keefir,* which does not share a common ancestor with Finnish *piimä*.

Table 2. An example of the lexical data for the meaning 'fish' in some of the studied languages. Columns a and b contain the lexical items grouped according to cognacy relationships; items in column a belong to one cognate group while those in b belong to another. Columns A and B contain the same data coded in binary form.

	FISH			
Language	**a**	**b**	**A**	**B**
North Saami	guolli	–	1	0
Kildin Saami	kūll'	–	1	0
Finnish	kala	–	1	0
Ingrian	kala	–	1	0
Estonian	kala	–	1	0
Erzya	kal	–	1	0
Meadow Mari	kol	–	1	0
Komi-Zyrian	–	ćeri	0	1
Udmurt	–	ćorig	0	1
Hungarian	hal	–	1	0
Tundra Nenets	xalya	–	1	0

The Uralic languages have been studied extensively by historical linguists, who have determined which words have been inherited vertically from a common ancestral language within the family, and which have been introduced into the Uralic lexicon horizontally via borrowing. Lexical items inherited from a common protolanguage are referred to as cognates (e.g. Finnish *kala* and Hungarian *hal,* 'fish'). Words that have been borrowed for example at the proto-language stage from another language family, but which have been inherited vertically in the tree after the initial appearance, are referred to as correlates (e.g. the word for 'tooth' was borrowed in Proto-Finnic, and a modified form of the word is now found in all Finnic languages). Cognates and correlates are coded similarly in the data, as they both share the feature of being traceable to a common source and can thus be used to infer historical relationships between languages. These are analogous to homologies in biology which usually indicate a common ancestry within the tree, but which may also stem from an external source. For example viruses may be inserted into the host genome via horizontal gene transfer, and are then inherited vertically in the tree (Gasmi et al., 2015).

In collecting the lexical data, information was recorded as to whether the words stemmed from a common Uralic source or whether they were borrowings.[23] This information was then used to distinguish six qualitatively different sublists of the 226 items. The first contains the most stable part of the Uralic lexicon, i.e. items not represented by loanwords (i.e. only cognacy relationships included). This 'Ura100' list contains 100 items. As this is the optimized list for the Uralic languages, it was also used in **III**. The other five subsets were formed according to the number of attested borrowings in the languages. These borrowings may have come from another language family or from another language within the Uralic family. The first list was '1+ borrowings', referring

[23] The distinction between correlates and cognates, however, has not been made in other lexical databases (e.g. in IELex).

to items with one or more loans in the languages studied, the second '2+ borrowings', referring to items with two or more loans, and so on up to '5+ borrowings', with five or more loans.[24] As all the items in these lists were prone to borrowing according to the literature, their relationships represented correlates. In total ten lexical datasets were used in the analyses: Swadesh 200, Swadesh 100, Leipzig-Jakarta, full list 226, Ura100, 1+borrowings, 2+borrowings, 3+borrowings, 4+borrowings and 5+borrowings.

In addition to the lexical data collected for the seventeen Uralic languages, I used a reconstructed Proto-Uralic as an outgroup in the phylogenetic analyses in **I**. Reconstructed Proto-Uralic was not, however, used in **III**, where I used an analysis which produces a rooted tree automatically. The data were binary-coded according to cognacy/correlate judgements (words belonging to a given cognate/correlate set = 1, words not belonging to a given cognate/correlate set = 0) (Table 2). Missing characters (i.e. the small number of items whose presence or absence in a language could not be ascertained) were marked with a question mark.

2.1.2. Extralinguistic variables related to the Uralic speaker area (III)

To examine the Court Jester macroevolutionary hypothesis, i.e. whether linguistic divergences could be affected by abiotic changes, temperature data were compiled from studies by Kremenetski et al. (1997), Davis et al. (2003), Väliranta et al. (2003) and Heikkilä and Seppä (2010). In these studies, temperature estimates were based on lake sediment, pollen and peat data collected from several locations in the northeastern Europe. A generalization of changes in temperature on the western side of the Ural Mountains was created and illustrated as a color gradient (in Fig. 5). This generalization was considered feasible, as the Holocene Thermal Maximum seems to have had a remarkably similar pattern and timing throughout boreal northeastern Europe (Heikkilä & Seppä, 2010), and was in general followed by a gradual cooling (Kremenetski et al., 1997, Davis et al., 2003, Väliranta et al., 2003, Heikkilä & Seppä, 2010). To determine the possible role played by cultural interactions in the divergences of the Uralic languages, archaeological and historical information was collected from the scholarly literature as cited in section 3.4 (e.g. Carpelan & Parpola, 2001).

2.1.3. Finnish dialect data (II, IV)

The Dialect Atlas of Finnish (Kettunen, 1940) was used as the source of the language data in the studies on linguistic microevolution. Lauri Kettunen, compiler of the Atlas, traveled systematically around Finland during the 1920s and 30s to collect the data. It contains a total of 213 map pages (such as in Fig. 2), showing various linguistic (phonological, morphological and lexical) features and their variants collected from 525 Finnish-speaking municipalities. The data thus cover only geographical variation; the data are reported in terms of language variants per municipality rather than per individual, and thus no social variables are encoded in the data.[25] The number of linguistic variants

[24] Items in the 5+ borrowings list were therefore most susceptible to borrowings, while the 1+ borrowings list had a larger variety of borrowing-susceptible items.

[25] Due to this, sociolinguistic questions cannot be studied using Kettunen's dialect data alone.

per feature varies between 2 and 15, with 1-4 variants per municipality (several variants in one municipality are possible since Kettunen may have interviewed a couple of individuals per municipality). Swedish-speaking municipalities along the Baltic coast are not covered by the Dialect Atlas. The data were initially digitized at York University (Embleton & Wheeler, 1997, 2000) in collaboration with the Institute for the Languages of Finland (KOTUS). The digitization was checked for errors and further edited in collaboration with KOTUS and the BEDLAN project and the data is now available through the AVAA-portal (http://avoin-test.csc.fi/web/kotus/aineistot).

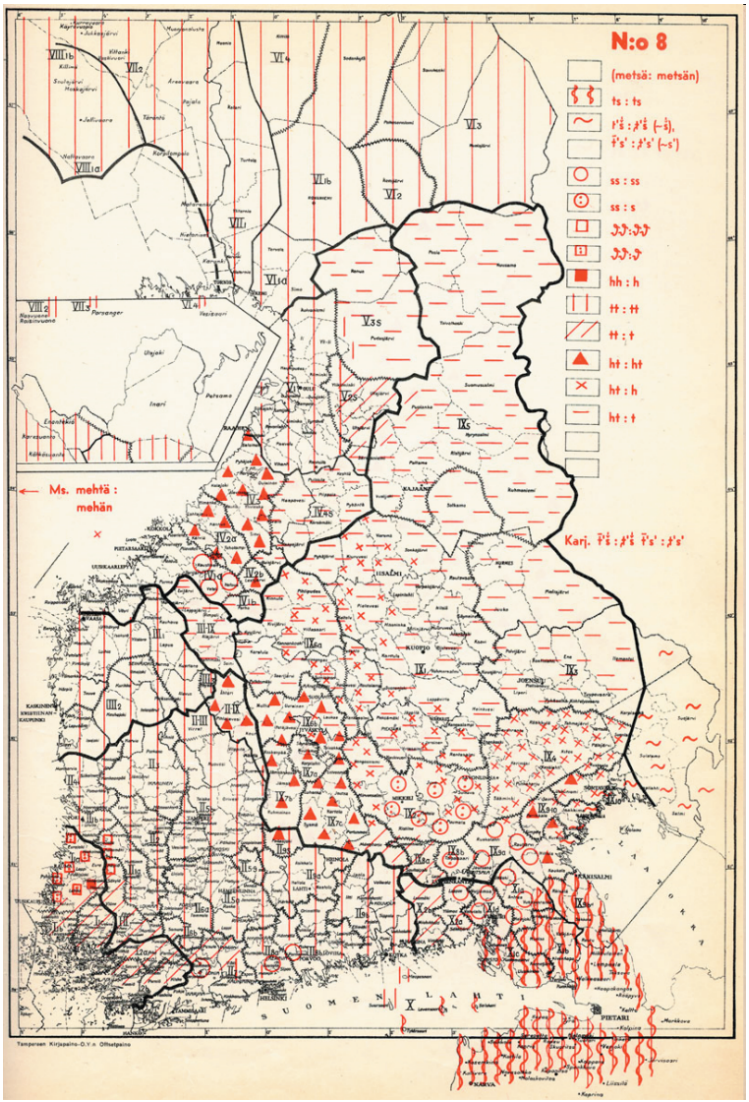

Figure 2. An example page of the Dialect Atlas of Finnish (Kettunen 1940), showing the geographical distribution of morphophonological variants for the word *metsä* 'forest'.

In **II**, data from all 525 municipalities were analyzed, but in **IV**, data were excluded for three kinds of localities, due to the inconsistent quality of the data on extralinguistic variables for these areas: Finnish-speaking areas outside Finland (where Kven,

Meänkieli and Ingrian are spoken), municipalities not part of Finnish territory at the beginning of the twentieth century (Karelian areas), and islands located in the Gulf of Finland. Three municipalities in northernmost Lapland, with fewer than twelve documented features, were also excluded. This reduced the number of municipalities studied in **IV** to 471.

As the language data obtained from the Dialect Atlas were analyzed with a population genetic clustering method, it was transformed to parallel the structure of genetic data: municipalities were likened to individuals, linguistic features to loci, and feature variants to alleles.[26] In the data collected, some 94 % of all study units had only one linguistic variant, and were thus analogous to biological haploids or diploids with a single type of allele (i.e. homozygous, that is, both alleles are similar);[27] 5.6 % had two variants per municipality, analogous to diploids with two alleles (i.e. heterozygous, that is, the two alleles are different). In 0.1 % of the municipalities studied, there were three or four variants per municipality, analogous to biological polyploids. Of these, the third and fourth variants were excluded from the analyses for the sake of simplicity.

For the analyses, the data were converted to both 'haploid' and 'diploid' format. Municipalities which originally had only one variant were already haploid; for those municipalities which had more than one variant, only one variant per municipality was included in the haploid data. To construct the diploid data, municipalities which originally had only one variant (haploids) were converted into diploid form by duplicating the existing variant; in other words, a single variant coded for example as [2] was duplicated as [2, 2]. Municipalities which originally had two variants were already in diploid format. In **II** the analyses were conducted with both diploid and haploid data, but the focus was on the diploid data since it also included variation within a municipality (i.e. it is more realistic). For the same reason, I used only diploid data in **IV**.

2.1.4. Extralinguistic variables related to Finland (IV)

In order to determine the possible role played by extralinguistic factors in the divergence of the Finnish dialects and in shaping the spatial pattern of linguistic variation within Finland, data were collected on four groups of extralinguistic variables: environmental and cultural variables, geographical distance and administrative borders. Geographic distance was used to study the isolating force of geographical distance (IBD), environmental and cultural variables to examine the isolating effect of differences in environmental and cultural conditions (IBE and IBC), and administrative borders to determine the isolating effect of belonging to different administrative areas (IBA) (see section 1.3). The data were primarily collected per municipality; for the dialect analysis, average values for core dialect areas (see section 2.2.3 and **IV**) were calculated for suitable variables from this municipality-based data.

[26] A locus is the location of a gene on a chromosome; alleles are variants of the gene located in the loci.
[27] Haploids have one, diploids have two and polyploids have more than two sets of chromosomes.

Table 3. Environmental and cultural variables used in the analyses of **IV**.

Environmental	Cultural
Mean annual temperature (ºC)	Forest land (% of total land area of a municipality)
Annual precipitation (mm)	Fields (% of total land area of a municipality)
Number of rainy days per year (> 1 mm)	Birth rate (annual mean for 1000 inhabitants)
Depth of snow cover (cm)	Death rate (annual mean for 1000 inhabitants)
Duration of snow cover (days per year)	Infant mortality: infants deceased during their first year of life (annual mean for 1000 born child)
The current annual increment per hectare of forest land (m^3)	Meadows (% of total land area of a municipality)
Growing stock on an average per hectare of forest (m^3)	Forests, wasteland etc. (% of total land area of a municipality)
The average quality of forest lands. The annual productive capacity when the forests are not thinned (m^3/ha)	Slash-and-burn area growing grain (per 100 ha of cultivated land)
Land area of the municipality (ha)	Average of inhabitants per residential building
Mean height of the municipality (meters above sea level)	Chimneyless peasant huts (% of all residential buildings)
Lakes (% of total land area of a municipality)	Immigration per 1000 inhabitants
Rivers (total river lengths per total land area of a municipality)	Emigration per 1000 inhabitants
Moraine (% of total land area of a municipality)	Number of Finnish speaking inhabitants
Clay (% of total land area of a municipality)	Number of Swedish speaking inhabitants
Gravel or sand (% of total land area of a municipality)	Total population number
Bedrock (% of total land area of a municipality)	Population density (total population/area)
Peat (% of total land area of a municipality)	Farmed area (% of total land area of a municipality)
Heat summation (ºC)	Population increase (immigration + birth rate)
	Population decrease (emigration + death rate)
	Population change (increase - decrease)
	Income per capita
	Taxes per capita

Data on 18 environmental and 22 cultural variables were collected from statistical yearbooks going back about a century and from historical atlases of Finland. The time when the dialect data were collected coincides quite closely with the time when the data on extralinguistic variables were documented.[28] These old data were supplemented by modern GIS data on relatively permanent physical variables, such as watersheds and soil types. Environmental variables were related to temperature, precipitation, snow,

[28] Despite the near-coincidence of the datasets, the extralinguistic factors naturally do not describe the area of Finland at the time when the dialects were taking shape. However, they cover the situation better than would be possible with modern data.

soil types, topography, watersheds and forest growth; the cultural variables were related to land use, house type, demography, subsistence strategy and level of income (Table 3). Some of these environmental and cultural variables were interconnected within the groups, for example the environmental variables 'temperature' and 'snow depth', and some were also linked between groups, for example 'field coverage' was classified as a cultural variable, since forests are cleared into fields by humans, but it also depends on soil type and thermal conditions, classified as environmental variables.

The data on administrative borders were a compilation of 16 administrative borders in the territory of Finland from the thirteenth to the nineteenth century. The data included the approximate eastern boundary after the second Swedish expedition to conquer Finland around 1250, the boundary drawn at the Treaty of Nöteborg between Sweden and Novgorod in 1323, nine different provincial boundary divisions, three different bishopric divisions, and two different divisions of judicial territories. The data were coded in binary form (a municipality either belonged to a particular administrative area or it did not). Geographical distances were measured both between municipalities and between core dialect areas (see section 2.2.3). For more details and the data source references see **IV**.

2.2. Analyses

The basic methodology adopted in this thesis to study language data is model-based and utilizes Bayesian inference and MCMC methods which, through millions of iterations, try different solutions and move toward the likelihood optima, where the data is best explained by the model (for more details on model-based methods and how they work, see **I** and **II**). Using these methods, I produce quantitative phylogenies of the Uralic languages (**I**, **III**) and cluster the Finnish dialects (**II**, **IV**). I also use more traditional distance-based clustering to verify the model-based approach in **II.** In addition to these methods, I perform various statistical analyses in **IV**.

2.2.1. Phylogenetic analyses of the Uralic languages (I, III)

To study the applicability of quantitative phylogenetic methods to the Uralic languages and the shape of the quantitative Uralic phylogeny, ten binary-coded lexical data sets (described in section 2.1.1) were analyzed with MrBayes v.3.2.1 (Huelsenbeck & Ronquist, 2001, Ronquist & Huelsenbeck, 2003), a Bayesian model-based method developed for phylogenetic analyses. The Bayesian inference with MCMC proceeds in phylogeny construction by repetitively calculating likelihood values for tree shapes and parameter values. It starts with a random tree shape and parameter values, and moves towards the optimum by accepting the tree shape and parameter values which get a higher likelihood than the previous one had (for more details see **I**).

In order to connect the Uralic phylogeny to archaeological periods and historical events on one hand, and to climatic changes on the other (to study the Red Queen and the Court Jester hypotheses, respectively), I estimated times for the language divergences along with phylogenetic relationships with the Ura100 data (**III**). This was done with

the BEAST v.1.5.4 software (Drummond & Rambaut, 2007). A relaxed linguistic clock was used in the analysis, allowing the rate of change in different branches to vary and making the 'clock' different from the traditional glottochronology where the rate of change is considered to be equal in all branches. I calibrated my tree topology with the three most reliable divergence time estimates of Uralic sub-branches based on the literature on the Uralic languages (Finno-Saamic 2000-3000 YBP, Permian 1300-1100 YBP and Samoyed 2200-2000 YBP) (for details see **III**).

As a general guideline when interpreting the phylograms produced by MrBayes, the posterior probability values above 0.95 are considered to have very good support (Huelsenbeck et al., 2001). Values slightly below this are less strong, but may also be noteworthy. In addition to posterior probability values, branch lengths are under observation, as they essentially reflect the amount of change that has taken place along each branch. Thus, shorter branch lengths indicate that only a small number of changes separates that clade from the rest and thus makes it a more uncertain group.

2.2.2. Quantitative clustering of the Finnish dialects (II)

In order to estimate whether population genetic methods are applicable to study languages I tested whether a focal method used in population genetic studies – clustering data with a model-based method – is applicable to language data. To study the applicability of the model-based clustering method to dialect data, I clustered the data both with a model-based clustering method (Structure software) (Pritchard et al., 2000), and with a more traditional, distance-based clustering method (K-medoids) (Kaufman & Rousseeuw, 1987) and compared the results. In addition, I compared these to the traditional dialect divisions of Finnish.

Structure uses Bayesian inference and MCMC method when detecting biological populations from the data (Pritchard et al. 2000). It infers the parameter set which gets the highest likelihood and thus explains the data best. Structure is one of the few methods of population biology that has previously been used with language data (Dunn et al., 2008, Reesink et al., 2009, Bowern, 2012), although not, as here, with pure dialect data. K-medoids has been previously applied to dialect data (Leino et al., 2006, Hyvönen et al., 2007). It works through a stepwise re-calculation process during which it minimizes the total distance between the medoid point (the data point selected as the center of the group) and the other points of each cluster (for further details see **II**).

Both Structure and K-medoids require the user to specify how many clusters (K) to infer. In my analyses K=2-20 clusters were inferred; in other words the data was split into two populations in the K=2 analysis, to three in K=3 analysis, and so on. The Structure analysis with each K value was repeated (20 times in **II**; 10 times in **IV**) to ensure the consistency of the results; K-medoids analyses were also repeated multiple times. Once the analyses were complete, the next step was to evaluate which K value (i.e. division into how many clusters) best explained the data.

Structure and K-medoids have different ways of evaluating the optimal number of clusters, as they differ with regard to the output they produce: Structure produces likelihood values and K-medoids produces distances. From the likelihood values obtained

in the repetitive runs for each K the average of the likelihood values may be calculated. From these it is possible to estimate which K value has the highest mean likelihood value and would thus best explain the data (mean log likelihood (Evanno et al., 2005)). Another way to estimate the most suitable number of clusters is to compare the mean likelihood values of successive K values and to find the K where the difference in likelihood values changes the most (Δ*K*-method (Evanno et al., 2005)). With K-medoids and its distance values the silhouette method was used (Rousseeuw, 1986). It compares the within-group and between-group dissimilarities of the data points describing how well a data point fits to its cluster compared to the neighboring one (for more details see **II**).

Structure analysis has two biological assumptions, Hardy-Weinberg equilibrium (HWE) and linkage equilibrium. It should, however, be noted that HWE is not an assumption in the same sense that linkage equilibrium is. Rather than requiring HWE for the analysis, it is used to create populations which would be as close to HWE as possible (for a more detailed discussion see **II**). Nevertheless, I investigated the possibility that this HWE "assumption" might affect the results and compared the Structure results to the traditional dialect divisions and to those of K-medoids, since as a distance-based method the last one is free of the biological assumptions. The assumption of linkage equilibrium in the case of the linguistic data is considered to refer to the correlatedness of different linguistic features. It was estimated by measuring the correlatedness of different map page pairs. More specifically it was done by calculating for each pair of map pages the number of cases where a pair of municipalities was linguistically identical on either map page and comparing that with the number of cases where a pair of municipalities was linguistically identical on both map pages. To obtain a correlation value for a pair of linguistic traits (map pages) the number of municipality pairs which were linguistically identical on both map pages was divided by the number of municipality pairs which were identical on at least one page (for further detail see **II**).

2.2.3. Calculations and statistical analyses used in dialect studies (IV)

I studied the variables shaping the spatial pattern of linguistic variation at two levels: between municipalities and between dialects. To obtain the dialect areas for the dialect analysis, I clustered the data into fourteen clusters with Structure (as explained in section 2.2.2 and **IV**) and used the core areas of these (i.e. where IC-values produced by the Structure analysis were >0.75) in the analyses. The details of the analyses with which the relationship of the spatial pattern of linguistic variation and extralinguistic variables were explored are summarized below, and covered in more detail in the Supplementary Information of **IV**.

Linguistic distances, i.e. the response variables, were calculated both for each pair of municipalities and each pair of core dialects. Between municipalities, these were calculated with a rough equivalent of Séguy's distance formula; the sum of linguistic variants differing between a pair of municipalities was calculated and divided with the total number of linguistic variants (Chambers & Trudgill, 1998). As a result, those pairs of municipalities which had a high number of differing linguistic variants were more different than those pairs which had smaller number of these. Between core dialect

areas, linguistic distances were calculated with F_{ST} adopted from population biology. F_{ST} calculation is based on differences in heterozygosities between subpopulations and the total population; if the heterozygosities in the total population and subpopulations are the same, population divergence has not happened, while if these differ, population structure may be inferred to exist. In addition, the more the heterozygosities differ, the larger the difference between the subpopulations (for further details see **II**).

Data on the explanatory variables (environmental, cultural, and administrative (see section 2.1.4)) were converted into distance matrices. Cultural and environmental variables were converted into distance matrices by subtracting; between dialects, the subtraction was performed with averages or sums calculated for the core dialect areas. The administrative data were converted into a Jaccard distance matrix to represent differences in administrative histories between each pair of municipalities. These were calculated by taking a pair of municipalities, summing the number of times these municipalities had belonged to different administrative areas (over all the studied administrative areas), and dividing it with the sum of the administrative areas to which either one or both of the municipalities had belonged. As a result, those municipalities which had often been part of the same administrative area were administratively closer to each other than those municipalities which had often belonged to different administrative areas (cf. Ross et al. 2013). Administrative data were not produced at the dialect level, as dialect areas typically embrace multiple administrative areas. Geographical distances were calculated both between each pair of municipal population centers and between the centroids of core dialect areas.

To exclude environmental and cultural variables that did not correlate with linguistic distance, or correlated with it only due to geographical distance, one-tailed partial Mantel tests (Smouse et al., 1986) were run for all explanatory variables. Variables showing a significant positive partial correlation ($p < 0.05$) with linguistic distance were included in the further analyses. I also determined the multicollinearity (i.e. the correlatedness) of all remaining cultural and environmental variables, and discarded variables with the highest correlations with several other variables; thus only one of the intercorrelated variables was left for further analysis.

I used multiple regression on distance matrices (MRM) analysis in model selection to find the environmental and cultural variables that best explained linguistic distances. MRM is a multiple regression analysis in which the variation of the response variable is explained with two or more explanatory variables. However, instead of using original data as in multiple regression, MRM uses distance values calculated from the original data. Model selection took place at both the dialect and the municipality level, separately for the environmental and cultural variables, with a backward elimination procedure. I removed the variable with the lowest coefficient in each round until the R^2 value dropped dramatically, resulting in environmental and cultural models with the largest explanatory power with the fewest variables. The variables remaining after model selection formed the 'final cultural' and 'final environmental' models.

To determine the extent to which the explanatory variables remaining in the final environmental and cultural models explained linguistic differences, I ran three sets of

MRM analyses: one for dialect-level and two for municipality-level linguistic distances. In the dialect analysis, I determined how much environmental (E) and cultural (C) differences and geographical distance (D) explained of dialect distances (E+C+D). In the first set of municipality analyses, these three (E+C+D) were also the explanatory variables. For the second set of municipality analyses, administrative distances (A) were included, and the environmental and cultural differences were grouped together (i.e. A+D+EC).

Finally, I performed a variation partitioning analysis to resolve the extent to which the explanatory variables explained linguistic distances, individually and jointly (e.g. Duivenvoorden et al., 2002, Macía et al., 2007). For the partitioning, I calculated R^2 values for each of the individual final models (e.g. for the dialect analysis these were E, C, D) and for all their combinations (EC, ED, CD, ECD). From these seven R^2 values, I calculated the relative proportions of the variation in linguistic differences explained by each of the eight subsections (three pure fractions, four joint fractions, and unexplained variation (shown in Fig. 6)). This was repeated for the municipality-level analyses with the same sets. In the analysis which included administrative differences and joint environment and culture, the individual effects were A, D and EC; the combined ones were AD, AEC, DEC, ADEC.

3. RESULTS AND DISCUSSION

3.1. Quantitative phylogenies of the Uralic languages (I, III)

Quantitative phylogenies produced with cognate-coded lexical data yield estimates of the shape (**I**) and timings (**III**) of the Uralic language phylogeny that are fairly similar to those proposed in traditional studies. This suggests the applicability of quantitative phylogenetic methods to the study of the Uralic languages. Next I relate my findings first to traditional Uralic phylogenies, then to quantitative phylogenies constructed for other language families. I also discuss what the shape of the phylogeny tells us about the history of the language family.

In my results, Proto-Uralic diverged into the Samoyed and Finno-Ugric branches in all phylogenies produced with core vocabulary lists of good and intermediate quality and size (i.e. full list, Sw207, Sw100, Leipzig-Jakarta, Ura100) (Fig. 3a), as well as in the Ura100 phylogeny constructed for the timing analyses (constructed with the same data but using a different model) (Fig. 5 in section 3.4; the figure also includes the main groupings of the languages). The same division was also obtained with a list of poorer quality (1+ borrowings), indicating its robustness. The division between the Samoyed and Finno-Ugric groups is also supported in the traditional linguistic literature based on lexical data (e.g. Janhunen, 2000); with other data types (phonological, grammatical) on the other hand, this division is less evident (e.g. Häkkinen, 2009). According to the results of my timing analyses, the divergence of Proto-Uralic occurred ca. 5300 YBP (Fig. 5). This timing is approximately intermediate in relation to those suggested earlier (in combination yielding dates of 7000-4000 YBP) (Korhonen, 1981, Sammallahti, 1988, Janhunen, 2000, Kallio, 2006, Häkkinen, 2009, Janhunen, 2009).

The Finno-Ugric branch shows no diversification during ca. 5300-3900 YBP; over the following millennium (ca. 3900-2900 YBP), the Finno-Ugric group then diverged into multiple branches (Figs. 3a and 5), suggesting a period of rapid divergences. This pattern is obtained with lists of both poorer quality (1+, 2+ and 3+ borrowings) and better quality (i.e. full list, Sw207, Sw100, Leipzig-Jakarta, Ura100) and with the timing analyses with Ura100. To my knowledge, the polytomous branching specifically of the Finno-Ugric group has not been proposed in earlier linguistic studies, although other forms of polytomous structures for the Uralic phylogeny have been suggested (Kulonen, 2002, Michalove, 2002, Häkkinen, 2007).

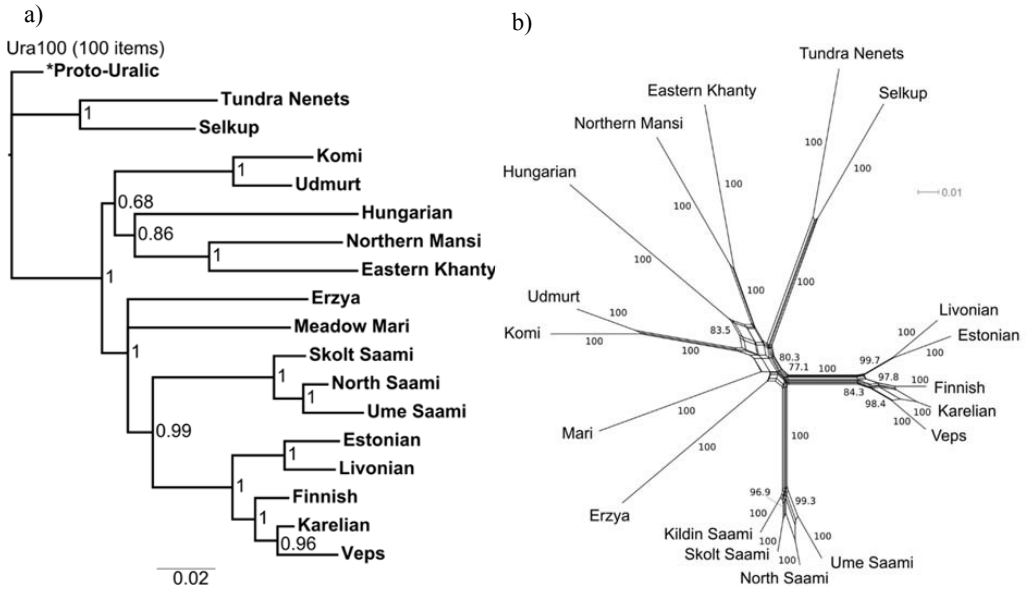

Figure 3. a) Phylogeny (**I**) and b) network (from Lehtinen et al., 2014) of the Uralic languages, both made with the Ura100 data.

Another phase of multiple rapid divergences in the Uralic phylogeny can be seen over the last 1500 years, during which the Finnic, Saami and Permian languages diverged within these groups from each other (Figs. 3a and 5) (**III**). Short branch lengths and high posterior probability values can be seen in the phylogenies constructed both with larger lists of good quality and with those of poorer quality, especially in the case of the Finnic and Saami groups (Figs. 3a and 5) (**I**). While nearly all divergences within Finnic and within Saami are well supported, they nevertheless show conflicting signals when visualized with the aid of networks (Fig. 3b) (Lehtinen et al., 2014). The conflict between the highly resolved tree and a complex network, seen especially when analyzing data which is more prone to borrowings than Ura100 (Fig. 7 in Lehtinen et al., 2014), most likely reflects recent divergences, coupled with recent contacts between closely related languages within these groups. This finding also highlights the importance of comparing quantitative phylogenies with networks, in order on the one hand to detect conflicting signals, on the other to quantify the tree-likeness of the network and thereby also to assess the reliability of the phylogenetic hypothesis, especially in the case of closely related languages.

The Uralic language family is not the only one to show a phase, or several phases, of rapid divergences. The Indo-European language family diverged into its main clades over a period of around one thousand years (between ca. 6000-7000 YBP) (Gray & Atkinson, 2003, Gray et al., 2011), the Semitic languages diverged into four main groups over ca. 1400 years (between ca. 5800-4400 YBP) (Kitchen et al., 2009), and the Austronesian phylogeny shows several pulses and pauses of divergences (Gray et al., 2009). The low posterior probabilities of the Arawak phylogeny (Walker & Ribeiro,

2011) also suggest rapid divergences in the deep branches, although timing estimates are not provided. In addition to these periods of rapid divergences located deep in the tree, at least the Indo-European and Semitic languages also show a period of rapid divergences during the last 2000 years (Gray & Atkinson, 2003, Kitchen et al., 2009, Gray et al., 2011), comparable to the late burst of divergences seen in the Uralic phylogeny (Figs. 3a and 5). It appears, then, that several quantitative phylogenies drawn up for language families spoken in different parts of the world follow a pattern whereby periods of rapid divergence alternate with times of fewer divergences. This raises the question of what it is that induces these divergences.

Unresolved phylogenies, with short branch lengths, are commonly associated with rapid population expansions, although languages may also spread geographically through language-shift, in which the language spreads to populations which originally spoke some other, now perhaps unknown, language. The early divergences of the Indo-European language family have been linked to the spread of farming from Anatolia, but the divergences of the major branches have also been connected to the population expansion of Kurgan horsemen (Gray & Atkinson, 2003, Gray et al., 2011). Recently, however, the steppe hypothesis, connected to the origin and spread of the Indo-European language family, has also received quantitative support (Chang et al., 2015). The settlement of the Pacific by Austronesian speakers has been linked to a set of technological innovations – a particular boat type and the ability to navigate by the stars – which enabled the spread of Austronesian language speakers to remote parts of the ocean (Gray et al., 2009).

Compared to the settling of Remote Oceania, which had not previously been settled by human populations, the Uralic speaker area most likely provides a different type of situation. Based on archaeological continuity in several locations and on substrate effects remaining in the Uralic languages, northeastern Europe was at least partly populated immediately after the retreat of the glaciers at the end of the last glaciation and before the Uralic languages spread to the area (Carpelan & Parpola, 2001, Janhunen, 2009). The question, then, is this: what was it that allowed the Uralic languages to spread, so that the Uralic languages survived down to the present while the other languages spoken in this area became extinct (e.g. Janhunen, 2009)?

Larger populations and military force are considered to have been unlikely explanations (Janhunen, 2009). It has been proposed that Uralic speakers had absorbed influences from their southern neighbors, through which they were perhaps more socially organized than populations living in the north (Mallory, 2001, Häkkinen, 2009). Such social structuring may have been connected to the transcultural Sejma-Turbino phenomenon: a network of warrior-traders distributing metal objects, weapons and other artefacts across large areas in northeastern Eurasia. It was thus not a cultural period in the traditional sense, as there are no Sejma-Turbino settlements (Carpelan & Parpola, 2001). Kallio (2006) has proposed and Häkkinen (2009) has agreed that this network is one plausible explanation for the spread of the Uralic languages and an inducer of possible language shift events. In section 3.4, I discuss other factors that may also have shaped the Uralic language family.

Phylogenies with polytomous branching may reflect the actual history of the language family, with certain periods of rapid divergence or language shift. However,

certain alternative reasons for the partly unresolved shape of the phylogeny need to be discussed. Firstly, more resolved phylogenies might be produced if more languages were sampled. It is, however, unlikely that the sampling of modern languages would change the picture, as they would contribute only to recent branches; the sampling of languages which have long been extinct is largely impossible, with the exception of those with written records – which at least in the case of the Uralic languages are very scarce. Secondly, language contacts may distort the pattern of binary branchings. The Uralic network, however, is very tree-like, exhibiting little reticulation in the deep divergences (Fig. 3b) (Lehtinen et al., 2014). There is thus no reason to suspect that the polytomous branching of Finno-Ugric would be explained by undetected borrowings. In addition, as in my studies (for more details see **I**), previous ones too (e.g. Greenhill et al., 2009) have found that quantitative language phylogenetics is fairly resistant to borrowings; thus these should not much affect the picture.

In summary, the phylogeny of the Uralic languages constructed using lexical data and its similarity to traditional Uralic phylogenies demonstrate the applicability of quantitative phylogenetic methods to the study of the Uralic language family. The lexicon is of course only one part of a language, and for example typological language features (e.g. the size and nature of the phoneme inventory; that is, how many phonemes there are and what kind of a system they form) are considered to be more stable and to represent old relationships better, as they are also resistant to borrowing. Furthermore, we can attempt to optimize the timings of the Uralic tree by testing the effect of including more languages and the effect of other datasets and models. Considering all the abovementioned issues is part of the ongoing work of the BEDLAN project.

3.2. Applicability of population genetic methods to Finnish dialects (II)

Population genetic clustering methods were found to be applicable to the Finnish dialect data: the clusters produced with the model-based method were largely similar to those obtained with a more traditional, distance-based clustering method, as well as to the traditional divisions of Finnish dialects. In what follows, I go through obtained clusterings in more detail, briefly discuss possible inferences related to the population history of Finnish speakers, and consider the biological assumptions of the Structure analysis.

Although my main purpose was to study the applicability of the population genetic clustering method called Structure to dialect data, I was also interested in what the optimal number of Finnish dialects is when determined quantitatively (i.e. into how many clusters the dialect data fits the best). Although the division into eastern and western dialects is considered to be the principal coarse-grained division, and the division into eight dialects to be the golden standard, other divisions have also been proposed in the traditional dialectological literature (division into three dialects (e.g. Leino et al., 2006), four dialects (Paunonen, 2006), seven dialects (e.g. Rapola, 1962); for a broader review see **II**).

The results of the Structure analysis offer basically two answers to this question: according to the likelihood values obtained from the analyses, divisions into K=2-14 (i.e.

two to fourteen dialect areas) are nearly equally good, while according to ΔK, a division into two dialects (east and west) is clearly the best. These results, however, do not exclude each other, as the ΔK results support the division on the uppermost hierarchical level. Silhouette values, with which the optimal number of K-medoids clusters was estimated, did not suggest any division to be notably better than the others, although there was a minor peak at K=16. Thus, the main division into east and west is clear from the Structure results as well, while none of the further divisions proposed earlier (into three, four, seven and eight clusters) was notably better than the others. To compare the clusterings more broadly, I focus here on divisions K=2-14 as suggested by the Structure results.

To estimate the applicability of Structure to intra-lingual variation data, I first compared the results to the clusters produced with K-medoids, which has previously been used to cluster dialect data (e.g. Leino et al., 2006). The results were fairly similar, especially when the data was divided into two to eight clusters; the only difference arose when the division was into three clusters. A comparison of the divisions into eight clusters is shown in Figs. 4a and b. When the data were divided into nine to fourteen clusters, there was more variability regarding which clusters appeared in which division (Fig. 8 in **II**); the majority of clusters were nevertheless the same in both analyses. This suggests that model-based and distance-based methods cluster Finnish dialects largely similarly.

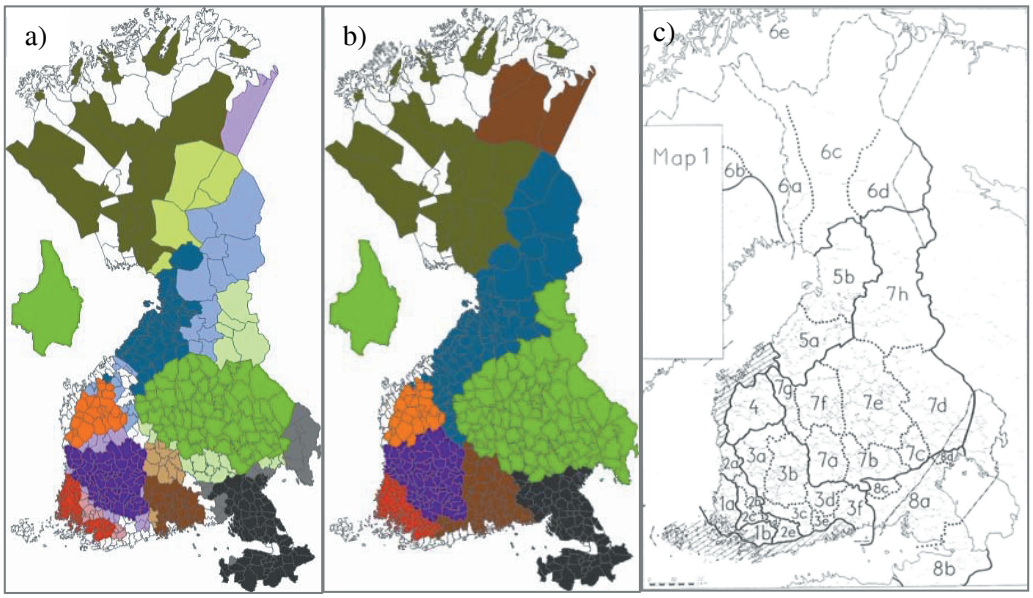

Figure 4. Three eight-way divisions of Finnish dialects (**II**). a) Division obtained with model-based Structure software; b) Division obtained with distance-based K-medoids; c) Traditional dialect division of Finnish (Itkonen, 1964). On map a, core dialect areas (including municipalities where the inferred cluster (IC) value obtained from the Structure analysis is > 0.75) are shown in darker shades than their corresponding transition areas (IC value 0.5-0.75). Municipalities colored white in the peripheral areas in maps a and b are those for which no data are available. Municipalities colored white in the central parts of map a represent areas of transitional dialects. Colors in maps a and b correspond with each other. The green isolated area next to Finland in maps a and b represents Värmland, an area in Sweden where people mainly from eastern Finland migrated to in the sixteenth century.

The order in which the clusters appear does not necessarily reflect the historical order of appearance of the dialects. However, the stability of clusters may have some connection to the settlement of Finland and the spread of dialects. When the data are divided into eight clusters, six of them appear on the western side of the east-west boundary: Southwest (red in Figs. 4a and b), Häme (purple), Southeast Häme + Päijät-Häme (brown), South Ostrobothnia (orange), Middle / North Ostrobothnia + North Kainuu + Kemijoki (blue), Far North (dark green). This leaves only two on the eastern side: Savo (green) and Southeast (gray). The western dialects remain relatively stable when the data are divided into nine to fourteen clusters: the majority of the new clusters appear on the eastern side of the boundary and split Savo in particular into several smaller groups. This may reflect differences in the histories of western and eastern Finns: in the east, the gradual expansion to the north and their slash-and-burn agriculture made them more mobile than people in the west, who had more stable settlements and land ownership over a longer time (Virrankoski, 2012).

When I continued validation of the method by comparing the obtained clusterings with the golden standard of the Finnish dialect division (Itkonen, 1964), I again found a relatively good match in all except one area (Fig. 4). The cluster that appeared in my results and is absent from the traditional divisions (e.g. in Fig. 4c) is Southeast Häme + Päijät-Häme (brown area in Figs. 4a and b); it is, however, mentioned in certain studies, where its difference from the surrounding dialects is recognized (e.g. Kettunen, 1930). The dialect area in Itkonen's division which did not appear in my eight-way divisions was the Southwest transitional (Fig. 4c, areas 2a-e). It did, however, appear in the quantitative clusterings with larger K values (K=10-14). As the name implies, it is a group of transitional dialects between Southwest and Häme; it is not always considered one of the main dialect areas, leading to the suggestion of a main division into seven dialects rather than eight (e.g. Rapola, 1962).

Concerning the biological assumptions of Structure, the similarity of the results acquired using both a method with biological assumptions (Structure) and one without them (K-medoids), as well as the traditional dialect division, supports the argument introduced in section 2.2.2 and in **II** that the Hardy-Weinberg "assumption" does not cause problems when clustering language data. Moreover, linkedness or correlatedness was not systematically found between linguistic features in noteworthy amounts. Even though uncorrelatedness was required by only one of the programs used, it can be pointed out that the uncorrelatedness of the observations should be considered in all cases, not merely when specifically stated, as in the case of Structure's assumptions.

3.3. Beyond phylogenies and population clustering

Quantitative phylogenies have become an important tool in resolving the historical relationships of both species and languages. Similarly, population genetic clustering methods are central in assessing the substructure of biological populations, and now may be also used to obtain linguistic clusters. Constructing phylogenies and inferring clusters is, however, often only the first step towards broader questions related to the

histories of the systems studied. The following steps consist of, for example, examining the drivers of macroevolutionary processes and the inducers of population divergence. These steps may jointly be seen as a biological tool chain applicable also to language studies. As linguistic phylogenies have been used in inferring the histories of the speaker populations in several language families (see sections 1.1 and 3.1), here I briefly discuss what the microevolutionary framework can bring to the field of dialectology.

Studies of population genetics are interested in such questions as whether population structures exist in a large mobile species such as the lynx (Rueness et al., 2003) or which factors have structured guppy populations (Crispo et al., 2006) – to name just two examples. Once we adopt a microevolutionary framework and a population genetic tool chain in studying dialectology, similar and other questions related to the histories of the speaker populations can be explored. An example of such a study is presented in section 3.5 and in **IV**.

A central step in the microevolutionary tool chain is the transformation of linguistic data into membership coefficients (IC values in Structure's results), where each municipality gets an IC value for each inferred population. IC values are frequencies which sum up to 1 and show how mixed the inferred populations are in each municipality. For example, when dividing into three populations (K = 3), a municipality may have a frequency of 0.4 for population A, a frequency of 0.4 for population B and a frequency of 0.2 for population C. These kind of fractions would indicate a large amount of admixture of the populations in the observed municipality; as a comparison, in a situation where the frequencies are 0.9 for population A, 0.1 for population B and zero for population C, population A is clearly dominant in the studied municipality. The output of the Structure analysis is in this frequency format when analyzing both genetic and linguistic data, and therefore in a data format applicable for further analyses. It is also important, however, to consider the differences between linguistic and genetic data as the object of study, both in assessing the applicability of a certain method and in drawing inferences from the results.[29]

In addition to enabling the usage of the population genetic tool chain, the frequency-type of output obtained in Structure's results enables a more flexible and thus also a more realistic illustration of dialect areas than when each municipality is assigned to only one dialect area, as with K-medoids (Fig. 4a vs. b). As another example, the frequencies may also be used to calculate linguistic diversity with, for example, the Shannon-Wiener index, which is commonly used in ecology to measure the diversity of ecological communities. Diversity values may, for example, be further compared to other spatial attributes to examine the occurrence of areas of high and low diversity.

In the following sections, I present the studies concerning the possible environmental and cultural inducers of linguistic divergence, at the levels of linguistic macro- and microevolution, and draw certain conclusions as to the histories of the speaker populations involved.

[29] Since there are often also historical data on humans, it is possible to test the applicability of the methods, as is done here in **I** and **II**. This is rarely possible with biological data.

3.4. Abiotic and 'biotic' changes shaping the history of the Uralic languages (III)

The deepest language divergences in the Uralic phylogeny took place soon after major cultural transitions during a cooling climate (Fig. 5). The possible causal pathways connected to these and to the spread of the Uralic languages are discussed below (and in **III**). Some of the divergences have taken place in historical times, and at least some of them were very likely induced by cultural drivers; the arrival of the East Slavic tribes in, for example, the Baltic area precedes the divergence of Finnish from Karelian and Veps (Fig. 5). Here, however, I focus the discussion on the deepest divergences and their possible environmental and cultural inducers, as these are both the least understood and those for which this new approach may have the most input.

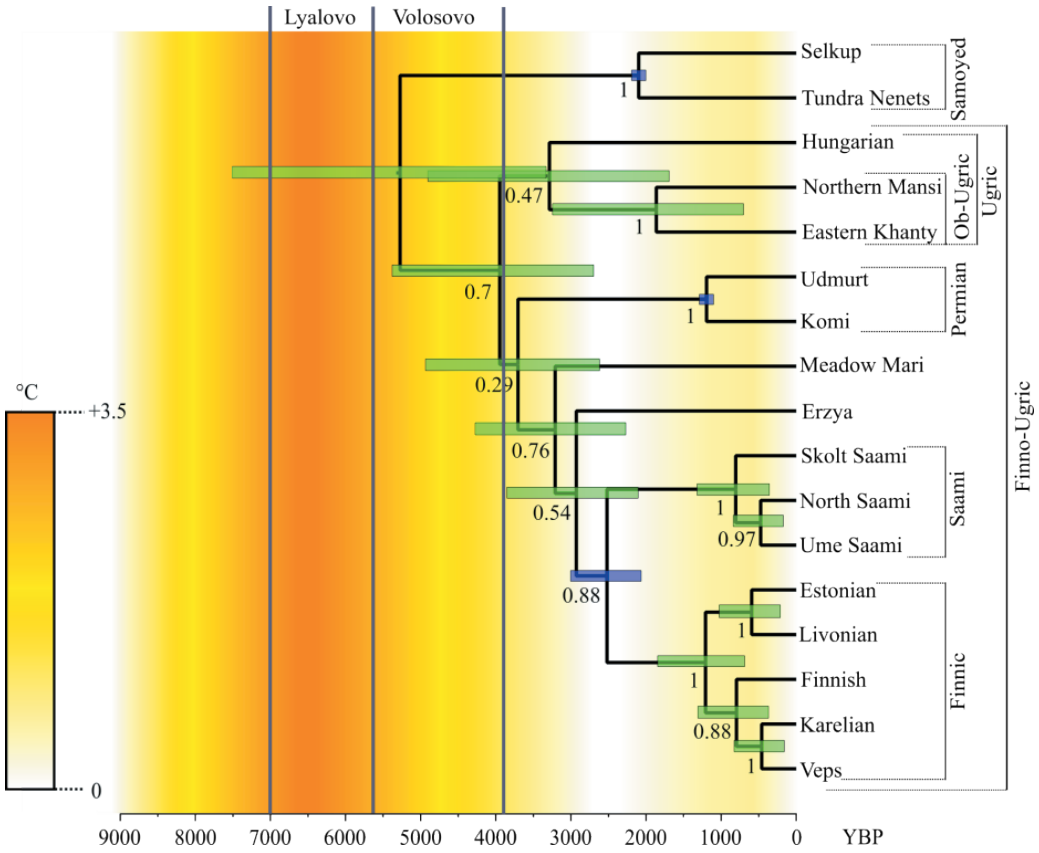

Figure 5. Uralic language phylogeny constructed with BEAST together with estimates of divergence times. Blue bars denote calibration time points and their ranges, which were used as uniform priors in the BEAST analysis; green bars represent 95 % HPD (highest probability density) for divergence times. Values below the nodes represent posterior probability values. Groupings on the right indicate subgroupings of the Uralic languages. Color gradient indicates variation in temperature in Uralic language speaker area. Gray vertical lines flank few most notable cultural periods.

According to my results, the Proto-Uralic divergence (ca. 5300 YBP) occurred after a period of warmth which lasted ca. 2000 years (the Holocene Thermal Maximum (HTM)) (Fig. 5). Changes in temperature induce changes in ecosystems, and during the HTM the biotic environment in the Volga-Kama area (the proposed area of Proto-Uralic speakers) was different from what it is today. In general during the HTM, vegetation zones moved northward and hunter-gatherer populations increased in size with rising temperatures (Tallavaara & Seppä, 2011). Local ecosystems, however, are affected by a number of climatic and geologic factors, temperature being only one among them, and regional variation in these tendencies is therefore likely to exist (Zhao et al., 2013). This makes it difficult to estimate the local effects of the temperature change.

If the rise in temperature did lead to a local increase in bio-productivity, and thereby also to an increase in the population size of Proto-Uralic speakers, possible pathways to the divergence of the Proto-Uralic could have been for example the following. First, the group of Proto-Uralic speakers grew so large that the internal integrity of the group was lost, dialectal differences became increasingly pronounced,[30] and in time these groups separated into different branches of the Uralic tree. Second, the increasing population size could have caused increased competition for resources, leading to migrations as an alternative to cope with the situation. The latter hypothesis is also probable in the case where the rise in temperature led to locally deteriorating conditions. However, it is important to bear in mind that hunter-gatherer populations have been found to be more flexible and prone to changing in relation to environmental conditions than has previously been assumed (Weber et al., 2013), and that they were presumably familiar with the more northern areas due to their mobility (Hertell & Tallavaara, 2011). Uralic hunter-gatherers could thus also have been very flexible in migrating to more suitable habitats and/or adapting to their local environments when conditions changed.

After the HTM, the climate cooled down. During this cooling period the Finno-Ugric branch diverged into four separate groups within about one thousand years (ca. 3900-2900 YBP) (Fig. 5). The cooling climate most likely induced changes in local environmental conditions, suggesting that some of the aforementioned processes could have recurred. I also found that the estimated time of the Sejma-Turbino network, ca. 4000 YBP (see section 3.1) coincided with the period of rapid divergences in the Finno-Ugric group (ca. 3900-2900). This network could thus have acted as a possible vector for the spread of the Finno-Ugric languages (cf. Kallio, 2006, Häkkinen, 2009).

Relating the divergence time estimates more broadly to shifts in cultural periods, we find that the Proto-Uralic divergence (ca. 5300 YBP) occurred after the transition from the Lyalovo to the Volosovo culture, ca. 5650 YBP (Fig. 5), and that the period of rapid divergences began right after the end of Volosovo period (ca. 3900 YBP), during the time of the Netted Ware culture (ca. 3900-2500 YBP). It is thus possible that these cultural changes had something to do with the language divergences. However, as the assumed culture of the early Uralic speakers depends on when and where Proto-Uralic is assumed to have been spoken, other possible language – culture connections have also

[30] Similarly to other natural languages, Proto-Uralic has also been suggested to have had dialects (Häkkinen, 2009, Janhunen, 2009).

been proposed (e.g. Häkkinen, 2009). Connecting archaeological periods to speakers of particular languages is thus extremely challenging, if not impossible, and a particular type of archaeological remains is most likely not evidence of speakers of just one language family. It may also be difficult to define or determine the effect of climatic change in cases of cultural changes other than collapse (with which climate has been proposed to be connected (e.g. deMenocal, 2001)).

In sum, there are several possible pathways through which climatic changes could have affected speaker populations. The different alternatives, however, are not mutually exclusive, but rather complement each other in shaping the Uralic phylogeny. Similarly, cultural phenomena, such as the Sejma-Turbino, could have acted together with more climate-induced demographic processes. Thus I agree that both 'biotic' and abiotic drivers are involved in language divergences, as stated in **III**. On the other hand, in view of the broad variety of cultural areas and changes in them in northeastern Europe alone (Carpelan, 1999, Carpelan & Parpola, 2001), I have become more critical towards the idea that the environmental impact on speaker populations necessarily induces shifts in archaeological periods by way of cultural change as I stated in **III**. At the very least this question needs closer scrutiny.

3.5. Extralinguistic variables shaping the spatial pattern of linguistic variation (IV)

In studying linguistic divergence within a language, I found that large cultural and environmental differences explained more of the linguistic differences within Finland than did geographical distance. This is an interesting result, considering that linguistic differentiation has more often been connected to geographical distance than to environmental differences. Below, I explain my findings in more detail and suggest how the connection between environmental and linguistic differences may have emerged.

I examined the question of factors contributing to linguistic diversification of the Finnish language at two different resolutions: coarse and fine-grained. The coarse-grained analysis took place between dialects, the fine-grained analysis between local linguistic variants, each spoken within a single municipality (henceforth 'between municipalities'). Environmental and cultural differences together with geographical distance explained 53.4 % of the linguistic differences between dialects, which is almost the same proportion explained by the same three categories of linguistic differences between municipalities (53.7 %). Adding administrative distances to the municipality analyses slightly increased the explanatory percentage (to 56.2 %). All in all, the percentage of linguistic differences explained by extralinguistic factors was relatively high.

In the dialect analysis, cultural differences (shown in yellow in Fig. 6a) explained the largest individual fraction of dialect differences (17.2 %), while between municipalities it explained the least (1.7 %; shown in yellow in Fig. 6b). This suggests a close interdependence between language and culture in Finland but only at the dialect level: minor cultural differences were less important. The cultural variables remaining in the

final model in the dialect analysis were related to house type (percentage of the poorest, chimneyless peasant huts out of all residential buildings) and land cover (percentage of farmed area of the total dialect area); in the municipality analysis variables were related to house type (percentage of the poorest, chimneyless peasant huts out of all residential buildings), land cover (percentage of forest land of the total area of the municipality) and subsistence type (percentage of slash-and-burn agriculture of the total area of the municipality).

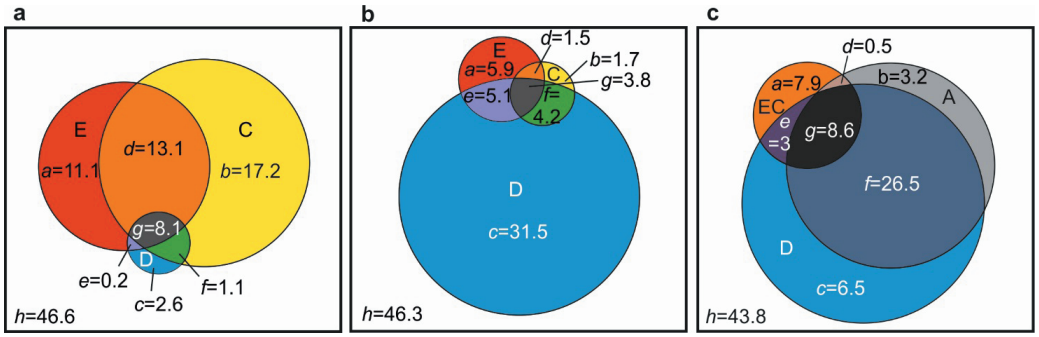

Figure 6. Partitioning of total variation in linguistic differences into components explainable by environmental (E), cultural (C), geographical (D) and administrative (A) distance. Individual (*a*-*c*) and joint (*d*-*g*) contributions of explanatory variables are specified; *h* shows the amount of unexplained variation. Circle sizes are approximations; values represent percentages of total variation. a) Relative proportions of E, C and D in explaining dialectal differences; b) Relative proportions of E, C and D in explaining linguistic differences between municipalities; c) Relative proportions of A, D and EC in explaining linguistic differences between municipalities.

The second largest individual fraction of dialect differences was explained by environmental differences (11.1%; shown in red in Fig. 6a). Environmental differences also explained the second largest fraction of the linguistic differences between municipalities (5.9 %; shown in red in Fig. 6b). Similarly to cultural differences, then, environmental differences matter when they are large enough. The environmental variables left in the final model in the dialect analysis were related to soil type (percentage of moraine and bedrock coverage of total dialect area) and waterways (river length in km per the dialect area). This was the case also in the municipality analysis as the variables left in the final model were percentages of lakes, clay soil and bedrock out of the total area of the municipality. It may be noted that in Finland even the largest differences in environmental conditions are subtle compared to the variety of conditions experienced in larger geographical areas. That is why it is remarkable that a pattern of linguistic isolation by environment emerged within Finland.

Geographical distance explained clearly the least of dialect differences (2.6 %; shown in blue in Fig. 6a), but on the other hand it explained the largest fraction of linguistic differences between municipalities (31.5 %; shown in blue in Fig. 6b). When administrative distances were included in the model, however, the geographical distance alone no longer explained the largest fraction of linguistic differences between municipalities; the joint contribution of geographical and administrative distances now

explained clearly the largest fraction (26.5 %; shown in dark blue in Fig 6c). This suggests that geographical distance does not act alone in shaping fine-grained differences in the spatial pattern of linguistic variation. The role of IBD in isolating language varieties in general (i.e. the idea that linguistic differences increase with geographical distance) (Heeringa & Nerbonne, 2001, Nerbonne, 2010) needs therefore to be reconsidered as well.

Environmental differences thus explained more of dialect differences than did geographical distance (Fig. 6a). According to the biological inference, the pattern of isolation by environment (IBE) (section 1.3) can be seen as indicating that adaptive processes have played a role in dialect differentiation within Finnish. Humans have been found to adapt to their environment by means of cultural features, and language can be viewed as a neutral marker of cultural populations (Mace & Jordan, 2011) (see section 1.2); I therefore find it plausible to conclude that the adaptive process which has produced the pattern of linguistic isolation by environment is cultural adaptation. More specifically, this adaptation is probably related to differences in subsistence strategies in different parts of Finland, as many of the environmental and cultural variables were, or could be, connected to subsistence (soil type, for example, is connected to plough type and to traditional subsistence-related skills). While I discarded the highest correlations from the variable data, many of these still correlated with each other (see Tables S2 and S3 in **IV**). This is why the results should be seen in terms of environmental and cultural effects as a whole, rather than in terms of individual variables.

It seems that, as hypothesized by Michalopoulos (2012) and Gavin et al. (2013), cultural specialization may indeed produce differing cultural spheres, which may in turn lead to group boundary formation, with linguistic diversification as a side product. This can be considered to be the case especially when the environmental differences are large, or when the range of behaviors suitable for the local environment is restricted (Currie & Mace, 2014). Different methods of slash-and-burn agriculture in eastern and western Finland can be considered an example of this kind of 'suitable behavior'. Since environmental conditions differ for example in terms of soil type and the length of the growing season, the slash-and-burn methods and plant varieties used in western Finland were not appropriate to be used in the east. Instead, the inhabitants of eastern Finland had a specific variety of rye and other special crops, which could tolerate the harsh conditions (Keto-Tokoi & Kuuluvainen, 2010).

After establishing that environmental differences contribute to dialect divergences, the big question now is whether these findings are generalizable to other languages and speaker areas, and whether it is possible that these processes underlie the coinciding patterns of linguistic and biological diversity (Harmon, 1996, Sutherland, 2003). The only way to resolve these questions is by investigating them in connection with dialects of other languages. Given the basic human needs that the pattern boils down to, however, I suggest that the Finnish language, and the environment in which it is spoken, are not an exception. Including more variables in the model might naturally change the percentage of linguistic differences the model explains. In addition, considering functional distances between dialect areas (i.e. where it is most cost-effective to move) instead of simple

geographical distances might change the relative contributions of different groups of explanatory variables. These questions will form the basis for my future work.

3.6. Unraveling linguistic divergence with approaches from evolutionary biology

I studied linguistic divergence at the levels of linguistic micro- and macroevolution. At both levels, I found it plausible to conclude that environmental features have played the ultimate role in linguistic divergences, acting through human ecology. This is remarkable, considering the great difference in the size between the speaker area of the Uralic languages and that of Finnish dialects, and differences in the time depth they cover.

I am not the only one to discuss the environmental conditions in which the speakers of Uralic lived (e.g. Salminen, 1999), or changes in those conditions (Häkkinen, 2009). Nor am I the first to suggest that the spread of a language family may be connected to environmental changes (climate-induced changes in vegetation, for example, provided a dispersal route for Bantu migration in Africa (Grollemund et al., 2015)). However, I am to my knowledge the first to discuss these two types of inducer of the spread of a language family: the demographic (population size increase) and the ecological (migration to cope with increased competition or to find more suitable habitats). The suggestion of subsistence-based cultural adaptation might also be extended from the context of dialect to that of the language family. For example the specialized reindeer-based economies of Saami and Samoyed speakers may suggest that similar kind of processes could have contributed also in shaping the Uralic language tree.

4. CONCLUSIONS

I found that quantitative biological methods are suitable for research concerning the Uralic languages and the dialects of Finnish. This suggests the general applicability of the methods used to linguistic data. I also found that environmental conditions and changes in them may play a role in isolating speaker populations from each other and dividing them into separate groups, in which the diversifying processes of language change can then take place. This underlines the importance of taking into account aspects related to human ecology in studying linguistic diversity and divergence.

Teams of scientists with various backgrounds have been shown to make more key discoveries than teams sharing a similar background and expertise (Dunbar, 1995, Dunbar, 1997). I hope that these findings and the findings reported in this thesis will encourage others as well to conduct more multidisciplinary research, as it is now finally possible. In order to reach this point, numerous essential steps have been taken in different fields over past centuries. In the case of my work these began in the nineteenth century, with the field trips of M. A. Castrén to the Uralic speaker areas and with the seminal work of Charles Darwin. They continued with the careful inspection of the historical relationships of these languages by historical linguists, and the further development of theories and quantitative methods in biology. The thesis now in the reader's hands is an attempt to integrate the knowledge and insights created in these, and also in other relevant fields.

However, rather than seeing the thesis as the end, I would like to see it as a beginning. A beginning of the field of quantitative Uralistics shaking off the ghosts of its glottochronological past, and the field of linguistic microevolution contributing not only methods from population genetics but also the framework itself – the questions asked and the ways these are investigated – to the study of languages and of the linguistic histories of speaker populations. To ensure a high standard, this should take place jointly between linguists and biologists in the future as well. By taking these aspects into account and deepening the multidisciplinary character of the work by entwining genetics and archaeology around the same question, both with the Uralic languages and with other language families, we will again be one large step closer to the ultimate goal: to resolve the holistic prehistory of humankind.

5. ACKNOWLEDGEMENTS

So how are the mysteries of the prehistory of the humankind investigated? PhD students, similarly to humans as a species, are affected by their environment. Therefore, the surrounding conditions in which the PhD studies were conducted deserve to be acknowledged.

First I would like to thank the Kone Foundation, which was brave enough to see the potential in this kind of a project already in 2008. At that time, I was still an undergraduate at the University of Jyväskylä, totally unaware of what would hit me in spring 2009, which is when I saw the announcement for this PhD student position. Along the way, I have also received funding from the Finnish Cultural Foundation Varsinais-Suomi Regional fund, the Ella & Georg Ehrnrooth Foundation, the University of Turku Foundation, and the University of Turku Graduate School, for which I am grateful.

In addition to the money, a project needs people. Being part of our BEDLAN project has been a great pleasure. Our initial team of biologists and linguists has later been complemented by research assistants and scholars from various fields. The project as a whole has been very valuable for me and my thesis, as without it my work could not have been completed with the breadth and depth it now has been. I would therefore like to thank all the BEDLAN people for their contribution and assistance during my PhD studies.

Certain individuals, however, deserve to be discussed and acknowledged in more detail, in particular my supervisors. Outi, I believe it is fair to say that you are the heart of the project – and at least occasionally also the brain. You get great ideas, and even though they are not all realizable (for example sending me to attend a linguistics field course in Siberia during the first year of my PhD studies), there is at least a lot of material to work with. You have taken good care of the project and the people in it, and I highly appreciate it along with everything you have taught me, both in and outside of science. Niklas, you have been very supportive throughout my PhD student years and encouraged me to ask questions when I did not know something. That is definitely something I have been doing a lot during these years, and I would say that multidisciplinary work is something where it is definitely needed. Kalle, the linguist of my supervisor team, it has been important that you have also been there for me. Thank you for all your help.

Kaj, in addition to sharing the experience of being PhD students in the BEDLAN project we also share half of the articles in our theses. I think it makes us dizygotic PhD twins. Thank you for your help throughout the years. In addition, I believe the whole project is thankful for your superhero skills with which you have tamed the python and speak the language of algorithms. Ilpo, you started as our research assistant to draw a map, but soon you became the lord of R with mysterious data collection and ArcGIS

skills. I am thankful of the many important skills I have learned from you. I would also like to thank Kalle Ruokolainen who bravely stepped into the world of language evolution and coauthored one paper. I stopped counting the times I came to ask you something because quite early on I ran out of fingers to do that. Thank you for your patience and for taking the time to discuss the mysteries of the language – environment relationship.

The Section of Ecology has provided me with a friendly (although temperature-wise a bit chilly) physical environment to work in. Thanks to cultural adaptation in the form of woolen shirts, mittens and a shawl, and the warm mental environment created by the staff, post docs and fellow PhD students, I have been able to survive. Thanks to Matti for assistance in computer-related issues and most importantly for the new computer I got last autumn. With the old computer I would probably still be waiting for the first version of the thesis Introduction to open. I would like to thank Niina and Tuija with the warmth of hundreds of liters of hot water for the Section coffees and for the hundreds of cups of tea I have drunk. I would also like to apologize to the people attending the Section coffees for neglecting those during last autumn, when I was busy writing my thesis. On the other hand, I did not have the time to keep up with celebrity gossip either, so I would not have been of very much use there anyway. I shall try to amend my manners in the future!

In addition to drinking tea, one also needs to feed herself with something more concrete than science: food. I would like to thank my regular and non-regular lunch company for joyful moments while enjoying e.g. Myssy's salads – including those who have already left the University of Turku, such as Liisa and Nanne. I think, however, that I have had lunch the most often with Jenni, so congratulations for this record! Thank you, Fiia, for reminding me that I should also leave the office at the end of the day. I may remind you of this same thing during the following months. I would also like to thank Chiara for broadening my mind food-culture–wise. I hope I was able to return the favor with makaroonivelli. Satu, Aino and Maria, thank you for sharing an office with me. You have made the office life very enjoyable. Thank you for your help with both scientific and non-scientific challenges, for giggles, candy and company (for further information, see the Acknowledgements in Dr. Kalske's thesis).

Outside the academic world I would like to thank KWP ry in particular. Behind all the bitterness, the point of our association is to say that friendship is important and something to cherish. I have enjoyed our harmonizing seminar weekends and the quality music we listen to at 90s gigs and at our own PMF midsummer festivals. Even though we have now lived several years in different locations around Finland, we have taken the time to gather together. This has meant a lot to me. Even though our lives may take us to different directions and to even longer geographical distances, I hope that these isolating factors will not come between our friendship (section 1.3 reference – LOVE IT!). The same goes for my Alkio-opisto friends. Even though we do not see each other quite that often, the things we do together are great fun, especially when saving Denmark! In addition to travelling and seeing friends, Irish dancing has turned out to be an excellent

way to balance my PhD studies, and I would like to thank my Irish dancing friends for the company and the sweaty moments. Right point and bow.

When I am not travelling to conferences or to see my friends, I travel to the place where any sensible person should go in their spare time, that is, to Kuortane, where I am from. It is located in Southern Ostrobothnia, where everything is bigger and better. The fact that my roots are there probably also gives an explanation to my charming personality. Maharottoman isoot kiitokset siis äiteelle, isälle, Tiinalle ja Terolle. Ootte mulle tärkeitä!

Finally, Aleksi, you have been like the sixth player on ice for my PhD project (without getting a two-minute penalty for that). It has been important for me to be able to reflect on all kinds of puzzling thoughts with you, and your vast amount of knowledge of linguistics and your critical way of thinking have also taught me a lot. Thank you.

6. REFERENCES

Abondolo, D., ed. 1998. *The Uralic Languages*. Routledge, London.

Anttila, R. 1989. *Historical and Comparative Linguistics*, 2 rev. edn. John Benjamins, Amsterdam, Philadelphia.

Atkinson, Q.D. 2011. Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science* **332:** 346-349.

Atkinson, Q.D. & Gray, R.D. 2005. Curious parallels and curious connections – phylogenetic thinking in biology and historical linguistics. *Syst. Biol.* **54:** 513-526.

Atkinson, Q., Meade, A., Venditti, C., Greenhill, S. & Pagel, M. 2008. Languages evolve in punctuational bursts. *Science* **319:** 588-588.

Barnosky, A.D. 2001. Distinguishing the effects of the Red Queen and Court Jester on Miocene mammal evolution in the Northern Rocky Mountains. *J. Vertebr. Paleontol.* **21:** 172-185.

Beall, C.M., Cavalleri, G.L., Deng, L., Elston, R.C., Gao, Y., Knight, J., Li, C., Li, J.C., Liang, Y., McCormack, M., Montgomery, H.E., Pan, H., Robbins, P.A., Shianna, K.V., Tam, S.C., Tsering, N., Veeramah, K.R., Wang, W., Wangdui, P., Weale, M.E., Xu, Y., Xu, Z., Yang, L., Zaman, M.J., Zeng, C., Zhang, L., Zhang, X., Zhaxi, P. & Zheng, Y.T. 2010. Natural selection on EPAS1 (HIF2a) associated with low hemoglobin concentration in Tibetan highlanders. *PNAS* **107:** 11459-11464.

Benton, M.J. 2009. The Red Queen and the Court Jester: Species diversity and the role of biotic and abiotic factors through time. *Science* **323:** 728-732.

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S.J., Alekseyenko, A.V., Drummond, A.J., Gray, R.D., Suchard, M.A. & Atkinson, Q.D. 2012. Mapping the origins and expansion of the Indo-European language family. *Science* **337:** 957-960.

Bowern, C. 2012. The riddle of Tasmanian languages. *Proc. R. Soc. B* **279:** 4590-4595.

Britain, D. 2002. Space and spatial diffusion. In: *The Handbook of Language Variation and Change* (J.K. Chambers, P. Trudgill and N. Schilling-Estes, eds), pp. 603-637. Blackwell Publishing, Malden.

Bromham, L., Hua, X., Fitzpatrick, T.G. & Greenhill, S.J. 2015. Rate of language evolution is affected by population size. *PNAS* **112:** 2097-2102.

Calude, A.S. & Pagel, M. 2011. How do we use language? Shared patterns in the frequency of word use across 17 world languages. *Phil. Trans. R. Soc. B* **366:** 1101-1107.

Carpelan, C. 1999. Käännekohtia Suomen esihistoriassa aikavälillä 5100...1000 eKr. In: *Pohjan poluilla: Suomalaisten juuret nykytutkimuksen mukaan* (P. Fogelberg, ed), pp. 249-280. Ekenäs Tryckeri Ab, Ekenäs.

Carpelan, C., Parpola, A. 2001. Emergence, contacts and dispersal of Proto-Indo-European, Proto-Uralic and Proto-Aryan in archaeological perspective. In: *Early Contacts between Uralic and Indo-European: Linguistic and Archaeological Considerations* (C. Carpelan, A. Parpola and P. Koskikallio, eds), pp. 55-150. Suomalais-Ugrilaisen Seuran Toimituksia 242, Helsinki.

Castrén, M.A. 1953. *Tutkimusmatkoilla pohjolassa*. Tammi, Helsinki.

Chambers, J.K., Trudgill, P. 1998. *Dialectology*, 2nd edn. Cambridge Univ. Press, Cambridge.

Chang, W., Cathcart, C., Hall, D. & Garrett, A. 2015. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* **91:** 194-244.

Cohen, A.S., Stone, J.R., Beuning, K.R.M., Park, L.E., Reinthal, P.N., Dettman, D., Scholz, C.A., Johnson, T.C., King, J.W., Talbot, M.R., Brown, E.T. & Ivory, S.J. 2007. Ecological consequences of early Late Pleistocene megadroughts in tropical Africa. *PNAS* **104:** 16422-16427.

Crispo, E., Bentzen, P., Reznick, D.N., Kinnison, M.T. & Hendry, A.P. 2006. The relative influence of natural selection and geography on gene flow in guppies. *Mol. Ecol.* **15:** 49-62.

Croft, W. 2000. *Explaining Language Change: An Evolutionary Approach*. Pearson Education, Harlow.

Currie, T.E. & Mace, R. 2009. Political complexity predicts the spread of ethnolinguistic groups. *PNAS* **106:** 7339-7344.

Currie, T.E. & Mace, R. 2014. Evolution of cultural traits occurs at similar relative rates in different world regions. *Proc. R .Soc. B* **281:** 20141622.

Darwin, C. 1871. *The Descent of Man*. Murray, London.

Davis, B.A.S., Brewer, S., Stevenson, A.C., Guiot, J. & Data Contributors. 2003. The temperature of Europe during the Holocene reconstructed from pollen data. *Quaternary Sci. Rev.* **22:** 1701-1716.

deMenocal, P.B. 2001. Cultural responses to climate change during the late Holocene. *Science* **292:** 667-673.

Diamond, J. & Bellwood, P. 2003. Farmers and their languages: The first expansions. *Science* **300:** 597-603.

Drummond, A.J. & Rambaut, A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7:** 214.

Duivenvoorden, J.F., Svenning, J.-C. & Wright, S.J. 2002. Beta diversity in tropical forests. *Science* **295:** 636-637.

Dunbar, K. 1995. How scientists really reason: Scientific reasoning in real-world laboratories. In: *Mechanisms of Insight.* (R.J. Sternberg and J. Davidson, eds), pp. 365-395. MIT Press, Cambridge, MA.

Dunbar, K. 1997. How scientists think: Online creativity and conceptual change in science. In: *Conceptual Structures and Processes: Emergence, Discovery, and Change* (T.B. Ward, S.M. Smith and S. Vaid, eds), pp. 461-493. American Psychological Assn, Washington, DC.

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.

Ehrlich, P.R. & Levin, S.A. 2005. The evolution of norms. *PLoS Biol* **3:** e194.

Embleton, S.M. & Wheeler, E.S. 2000. Computerized dialect atlas of Finnish: Dealing with ambiguity. *J. Quant. Linguist.* **7:** 227-231.

Embleton, S. & Wheeler, E.S. 1997. Finnish dialect atlas for quantitative studies. *J. Quant. Linguist.* **4:** 99-102.

Evanno, G., Regnaut, S. & Goudet, J. 2005. Detecting the number of clusters of individuals using the software Structure: A simulation study. *Mol. Ecol.* **14:** 2611-2620.

Everett, C., Blasi, D.E. & Roberts, S.G. 2015. Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots. *PNAS* **112:** 1322-1327.

Filatova, O.A., Burdin, A.M. & Hoyt, E. 2013. Is killer whale dialect evolution random? *Behav. Processes* **99:** 34-41.

Fisher, S.E. & Ridley, M. 2013. Culture, genes, and the human revolution. *Science* **340:** 929-930.

Foley, R.A. 2004. The evolutionary ecology of linguistic diversity in human populations. In: *Traces of Ancestry: Studies in Honour of Colin Renfrew.* (M. Jones, ed), pp. 61-71. McDonald Institute Monographs, Cambridge.

Futuyma, D.J. 2009. *Evolution*. Sinauer, Sunderland.

Gasmi, L., Boulain, H., Gauthier, J., Hua-Van, A., Musset, K., Jakubowska, A.K., Aury, J., Volkoff, A., Huguet, E., Herrero, S. & Drezen, J. 2015. Recurrent domestication by Lepidoptera of genes from their parasites mediated by bracoviruses. *PLoS Genet.* **11:** e1005470.

Gavin, M.C., Botero, C.A., Bowern, C., Colwell, R.K., Dunn, M., Dunn, R.R., Gray, R.D., Kirby, K.R., McCarter, J., Powell, A., Rangel, T.F., Stepp, J.R., Trautwein, M., Verdolin, J.L. & Yanega, G. 2013. Toward a mechanistic understanding of linguistic diversity. *Bioscience* **63:** 524-535.

Gavin, M.C. & Sibanda, N. 2012. The island biogeography of languages. *Global Ecol. Biogeogr.* **21:** 958-967.

Goebel, T., Waters, M.R. & O'Rourke, D.H. 2008. The late Pleistocene dispersal of modern humans in the Americas. *Science* **319:** 1497-1502.

Gray, R.D. & Atkinson, Q.D. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* **426:** 435-439.

Gray, R.D., Atkinson, Q.D. & Greenhill, S.J. 2011. Language evolution and human history: What a difference a date makes. *Phil. Trans. R. Soc. B* **366:** 1090-1100.

Gray, R.D., Drummond, A.J. & Greenhill, S.J. 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* **323:** 479-483.

Gray, R.D. & Jordan, F.M. 2000. Language trees support the express-train sequence of Austronesian expansion. *Nature* **405:** 1052-1055.

Greenhill, S.J., Currie, T.E. & Gray, R.D. 2009. Does horizontal transmission invalidate cultural phylogenies? *Proc. R. Soc. B* **276:** 2299-2306.

Grollemund, R., Branford, S., Bostoen, K., Meade, A., Venditti, C. & Pagel, M. 2015. Bantu expansion shows that habitat alters the route and pace of human dispersals. *PNAS* **112:** 13296–13301.

Haapala, P. 2007. *Suomen historian kartasto*. Karttakeskus, Helsinki.

Häkkinen, J. 2007. *Kantauralin murteutuminen vokaalivastaavuuksien valossa*. MA thesis, University of Helsinki.

Häkkinen, J. 2009. Kantauralin ajoitus ja paikannus: Perustelut puntarissa. *JSFOu* **92:** 9-56.

Häkkinen, K. 1983. *Suomen kielen vanhimmasta sanastosta ja sen tutkimisesta. Suomalais-*

ugrilaisten kielten etymologisen tutkimuksen perusteita ja metodiikkaa, PhD thesis. University of Turku, Turku.

Hamilton, M.B. 2009. *Population Genetics*. Wiley-Blackwell, Chichester.

Harmon, D. 1996. Losing species, losing languages: Connections between biological and linguistic diversity. *Southwest J. Ling.* **15:** 89-108.

Heeringa, W. & Nerbonne, J. 2001. Dialect areas and dialect continua. *Lang. Var. Change* **13:** 375-400.

Heikkilä, M. 2014. *Bidrag till Fennoskandiens språkliga förhistoria i tid och rum*, PhD thesis. University of Helsinki, Helsinki.

Heikkilä, M. & Seppä, H. 2010. Holocene climate dynamics in Latvia, eastern Baltic region: A pollen-based summer temperature reconstruction and regional comparison. *Boreas* **39:** 705-719.

Hertell, E., Tallavaara, M. 2011. High mobility or gift exchange – early Mesolithic exotic chipped lithics in Southern Finland. In: *Mesolithic Interfaces: Variability in Lithic Technologies in Eastern Fennoscandia* (T. Rankama, ed), pp. 11-41. The Archaeological Society of Finland, Saarijärvi.

Holden, C.J., Meade, A., Pagel, M. 2005. Comparison of maximum parsimony and Bayesian Bantu language trees. In: *The Evolution of Cultural Diversity: A Phylogenetic Approach* (R. Mace, C.J. Holden and S. Shennan, eds), pp. 53-65. Left Coast Press, Walnut Creek, CA.

Holden, C.J. 2002. Bantu language trees reflect the spread of farming across sub-Saharan Africa: A maximum-parsimony analysis. *Proc. R. Soc. B* **269:** 793-799.

Hovdhaugen, E., Karlsson, F., Henriksen, C., Sigurd, B. 2000. *The History of Linguistics in the Nordic Countries*. Societas Scientarium Fennica, Helsinki.

Huelsenbeck, J.P. & Ronquist, F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17:** 754-755.

Huelsenbeck, J., Ronquist, F., Nielsen, R. & Bollback, J. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* **294:** 2310-2314.

Hyvönen, S., Leino, A. & Salmenkivi, M. 2007. Multivariate analysis of Finnish dialect data – an overview of lexical variation. *Literary and Linguistic Computing* **22:** 271-290.

Itkonen, E. 2003. *What is Language? A Study in the Philosophy of Linguistics*. University of Turku, Turku.

Itkonen, T. 1964. *Proto-Finnic Final Consonants: Their History in the Finnic Languages with Particular Reference to the Finnish Dialects. 1:1, Introduction; the History of -K in Finnish*, PhD thesis. University of Helsinki, Helsinki.

Janhunen, J. 2000. Reconstructing Pre-Proto-Uralic typology spanning the millennia of linguistic evolution. In: *Congressus Nonus Internationalis Fenno-Ugristarum. Pars 2, Summaria Acroasium in Sectionibus Et Symposiis Factarum* (A. Nurk, T. Palo and T. Seilenthal, eds), pp. 59-76. University of Tartu, Estonian Finno-Ugrian Committee, Tartu.

Janhunen, J. 2009. Proto-Uralic – what, where, and when? *Mémoires de la Société Finno-Ougrienne* **258:** 57-78.

Jobling, M., Hollox, E., Hurles, M., Kivisild, T., Tyler-Smith, C. 2013. *Human Evolutionary Genetics*, 2nd edn. Garland Science, New York.

Kallio, P. 2006. Suomen kantakielen absoluuttista kronologiaa. *Virittäjä* **110:** 2-25.

Kaufman, L., Rousseeuw, P. 1987. Clustering by means of medoids. In: *Statistical Data Analysis Based on the L1-Norm and Related Methods* (Y. Dodge, ed), pp. 405-416. Amsterdam, North-Holland.

Kawai, M. 1965. Newly-acquired pre-cultural behavior of the natural troop of Japanese monkeys on Koshima islet. *Primates* **6:** 1-30.

Kersalo, J., Pirinen, P. 2009. *Suomen maakuntien ilmasto*. Ilmatieteen laitos, Helsinki.

Keto-Tokoi, P., Kuuluvainen, T. 2010. *Suomalainen aarniometsä*. Maahenki, Hämeenlinna.

Kettunen, L. 1930. *Suomen murteet: II Murrealueet*. Suomalaisen Kirjallisuuden Seuran Toimituksia, Helsinki.

Kettunen, L. 1940. *Suomen murteet: III A, Murrekartasto*. Suomalaisen Kirjallisuuden Seura, Helsinki.

Kitchen, A., Ehret, C., Assefa, S. & Mulligan, C.J. 2009. Bayesian phylogenetic analysis of Semitic languages identifies an early Bronze Age origin of Semitic in the Near East. *Proc. R. Soc. B* **276:** 2703-2710.

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

Korhonen, M. 1991. Uralin tällä ja tuolla puolen. In: *Uralilaiset kansat: Tietoa suomen sukukielistä ja niiden puhujista* (J. Laakso, ed), pp. 20-48. WSOY, Helsinki.

Kremenetski, C., Vaschalova, T., Goriachkin, S., Cherkinsky, A. & Sulerzhitsky, L. 1997. Holocene pollen stratigraphy and bog development in the western part of the Kola peninsula, Russia. *Boreas* **26:** 91-102.

Kulonen, U. 2002. Kielitiede ja Suomen väestön juuret. In: *Ennen, muinoin: miten menneisyyttämme*

tutkitaan (R. Grünthal, ed), pp. 102-116. Suomalaisen Kirjallisuuden Seura, Helsinki.

Laland, K.N., Odling-Smee, J. & Myles, S. 2010. How culture shaped the human genome: Bringing genetics and the human sciences together. *Nat. Rev. Genet.* **11:** 137-148.

Lee, S. & Hasegawa, T. 2014. Oceanic barriers promote language diversification in the Japanese islands. *J. Evol. Biol.* **27:** 1905-1912.

Lee, S. & Hasegawa, T. 2011. Bayesian phylogenetic analysis supports an agricultural origin of Japonic languages. *Proc. R. Soc. B* **278:** 3662-3669.

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

Lehtinen, T. 2007. *Kielen vuosituhannet: suomen kielen kehitys kantauralista varhaissuomeen*. Suomalaisen Kirjallisuuden Seura, Helsinki.

Leino, A. & Hyvönen, S. 2008. Comparison of component models in analysing the distribution of dialect features. *International Journal of Humanities and Arts Computing* **2**: 173-187.

Leino, A., Hyvönen, S. & Salmenkivi, M. 2006. Mitä murteita suomessa onkaan? Murresanaston levikin kvantitatiivista analyysiä. *Virittäjä* **110:** 26-45.

Levinson, S.C. & Gray, R.D. 2012. Tools from evolutionary biology shed new light on the diversification of languages. *Trends Cogn. Sci.* **16:** 167-173.

Lewis, P.M., Simons, G.F., Fennig, C.D. 2015. *Ethnologue: Languages of the World.* http://www.ethnologue.com.

Lyell, C. 1863. *Geological Evidences of the Antiquity of Man*. John Murray, London.

Mace, R. & Jordan, F.M. 2011. Macro-evolutionary studies of cultural diversity: A review of empirical studies of cultural transmission and cultural adaptation. *Phil. Trans. R. Soc. B* **366:** 402-411.

Mace, R. & Pagel, M. 1995. A latitudinal gradient in the density of human languages in North America. *Proc. R. Soc. B* **261:** 117-121.

Macía, M.J., Ruokolainen, K., Tuomisto, H., Quisbert, J. & Cala, V. 2007. Congruence between floristic patterns of trees and lianas in a southwest Amazonian rain forest. *Ecography* **30:** 561-577.

Mallory, J.P. 2001. Uralics and Indo-Europeans: Problems of time and space. In: *Early Contacts between Uralic and Indo-European: Linguistic and Archaeological Considerations* (C. Carpelan, A. Parpola and P. Koskikallio, eds), pp. 345-366. Suomalais-Ugrilaisen Seuran Toimituksia, Helsinki.

Mayr, E. 1942. *Systematics and the Origin of Species*. Columbia University Press, New York.

Mayr, E. 1963. *Animal Species and Evolution*. Harvard University Press, Cambridge, MA.

McMahon, A., McMahon, R. 2005. *Language Classification by Numbers*. Oxford University Press, New York.

Mesoudi, A., Whiten, A. & Laland, K.N. 2004. Is human cultural evolution Darwinian? Evidence reviewed from the perspective of *the Origin of Species*. *Evolution* **58:** 1-11.

Michalopoulos, S. 2012. The origins of ethnolinguistic diversity. *Am. Econ. Rev.* **102:** 1508-1539.

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

Moore, J.L., Manne, L., Brooks, T., Burgess, N.D., Davies, R., Rahbek, C., Williams, P. & Balmford, A. 2002. The distribution of cultural and biological diversity in Africa. *Proc. R. Soc. B* **269:** 1645-1653.

Nerbonne, J. 2010. Measuring the diffusion of linguistic change. *Philos. Trans. R. Soc. B* **365:** 3821-3828.

Nettle, D. 1999. *Linguistic Diversity*. Oxford University Press, New York.

Neuvonen, A.M., Putkonen, M., Översti, S., Sundell, T., Onkamo, P., Sajantila, A. & Palo, J.U. 2015. Vestiges of an ancient border in the contemporary genetic diversity of north-eastern Europe. *PLoS ONE* **10:** e0130331.

Orsini, L., Vanoverbeke, J., Swillen, I., Mergeay, J. & De Meester, L. 2013. Drivers of population genetic differentiation in the wild: Isolation by dispersal limitation, isolation by adaptation and isolation by colonization. *Mol. Ecol.* **22:** 5983-5999.

Pagel, M. 2000a. The history, rate, and pattern of world linguistic evolution. In: *The Evolutionary Emergence of Language* (C. Knight, M. Studdert-Kennedy and J. Hurford, eds), pp. 391-416. Cambridge University Press, Cambridge

Pagel, M. 2000b. Maximum-likelihood models for glottochronology and for reconstructing linguistic phylogenies. In: *Time Depth in Historical Linguistics* (C. Renfrew, A. McMahon and L. Trask, eds), pp. 189-207. The McDonald Institute for Archaeological Research, Cambridge.

Pagel, M. 2009. Human language as a culturally transmitted replicator. *Nature Reviews Genetics* **10:** 405-415.

Pagel, M., Atkinson, Q. & Meade, A. 2007. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* **449:** 717-720.

Pagel, M. & Mace, R. 2004. The cultural wealth of nations. *Nature* **428:** 275-278.

Pagel, M., Venditti, C. & Meade, A. 2006. Large punctuational contribution of speciation to evolutionary divergence at the molecular level. *Science* **314:** 119-121.

Palander, M., Opas-Hänninen, L. & Tweedie, F. 2003. Neighbours or enemies? Competing variants causing differences in transitional dialects. *Computers and the Humanities* **37:** 359-372.

Paul, H. 1886. *Principien der Sprachgeschichte*, 2nd edn. Max Niemeyer, Halle.

Paunonen, H. 2006. Lounaismurteiden asema suomen murteiden ryhmityksessä. In: *Kohtauspaikkana kieli – Näkökulmia persoonaan, muutoksiin ja valintoihin* (T. Nordlund, T. Onikki-Rantajääskö and T. Suutari, eds), pp. 249-268. Suomalaisen Kirjallisuuden Seura, Helsinki.

Peña, C. & Wahlberg, N. 2008. Prehistorical climate change increased diversification of a group of butterflies. *Biology Letters* **4:** 274-278.

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

Prugnolle, F., Manica, A. & Balloux, F. 2005. Geography predicts neutral genetic diversity of human populations. *Current Biology* **15:** R159-R160.

Ramachandran, S., Deshpande, O., Roseman, C.C., Rosenberg, N.A., Feldman, M.W. & Cavalli-Sforza, L. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *PNAS* **102:** 15942-15947.

Rapola, M. 1962. *Johdatus suomen murteisiin*, 2nd edn. Suomalaisen Kirjallisuuden Seura, Turku.

Reesink, G., Singer, R., Dunn, M. 2009. Explaining the linguistic diversity of Sahul using population models. *PLoS Biol*. **7**: 1000241.

Reznick, D.N. & Ricklefs, R.E. 2009. Darwin's bridge between microevolution and macroevolution. *Nature* **457:** 837-842.

Rogers, D.S. & Ehrlich, P.R. 2008. Natural selection and cultural rates of change. *PNAS* **105:** 3416-3420.

Ronquist, F. & Huelsenbeck, J.P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19:** 1572-1574.

Ross, R.M., Greenhill, S.J. & Atkinson, Q.D. 2013. Population structure and cultural geography of a folktale in Europe. *Proc. R. Soc. B* **280:** 20123065.

Rousseeuw, P.J. 1986. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20:** 53-65.

Rueness, E.K., Jorde, P.E., Hellborg, L., Stenseth, N.C., Ellegren, H. & Jakobsen, K.S. 2003. Cryptic population structure in a large, mobile mammalian predator: The Scandinavian lynx. *Mol. Ecol.* **12:** 2623-2633.

Rundle, H.D. & Nosil, P. 2005. Ecological speciation. *Ecol. Lett.* **8:** 336-352.

Saarikivi, J. 2011. Saamelaiskielet – nykypäivää ja historiaa. In: *Saamentutkimus tänään* (I. Seurujärvi-Kari, P. Halinen and R. Pulkkinen, eds), pp. 77-119. Suomalaisen Kirjallisuuden Seura, Helsinki.

Salmela, E., Lappalainen, T., Fransson, I., Andersen, P.M., Dahlman-Wright, K., Fiebig, A., Sistonen, P., Savontaus, M., Schreiber, S., Kere, J. & Lahermo, P. 2008. Genome-wide analysis of single nucleotide polymorphisms uncovers population structure in northern Europe. *PLoS ONE* **3:** e3519.

Salminen, T. 1999. Euroopan kielet muinoin ja nykyisin. In: *Pohjan poluilla: Suomalaisten juuret nykytutkimuksen mukaan* (P. Fogelberg, ed), pp. 13-26. Ekenäs Tryckeri Ab, Ekenäs.

Salminen, T. 2007. Europe and North Asia. In: *Encyclopedia of the World's Endangered Languages* (C. Moseley, ed), pp. 211-282. Routledge, New York.

Sammallahti, P. 1988. Historical phonology of the Uralic languages. In: *The Uralic Languages: Description, History and Foreign Influences* (D. Sinor, ed), pp. 478-554. E. J. Brill, Leiden.

Sapir, E. 1912. Language and environment. *American Anthropologist* **14:** 226-242.

Saslis-Lagoudakis, C., Hawkins, J.A., Greenhill, S.J., Pendry, C.A., Watson, M.F., Tuladhar-Douglas, W., Baral, S.R. & Savolainen, V. 2014. The evolution of traditional knowledge: Environment shapes medicinal plant use in Nepal. *Proc. R. Soc. B* **281:** 20132768.

Schluter, D. 2001. Ecology and the origin of species. *Trends in Ecology & Evolution* **16:** 372-380.

Scholz, C.A., Johnson, T.C., Cohen, A.S., King, J.W., Peck, J.A., Overpeck, J.T., Talbot, M.R., Brown, E.T., Kalindekafe, L., Amoako, P.Y.O., Lyons, R.P., Shanahan, T.M., Castañeda, I.S., Heil, C.W., Forman, S.L., McHargue, L.R., Beuning, K.R., Gomez, J. & Pierson, J. 2007. East African megadroughts between 135 and 75 thousand years ago and bearing on early-modern human origins. *PNAS* **104:** 16416-16421.

Sexton, J.P., Hangartner, S.B. & Hoffmann, A.A. 2013. Genetic isolation by environment or distance: Which pattern of gene flow is most common? *Evolution* **68:** 1-15.

Shafer, A.B.A. & Wolf, J.B.W. 2013. Widespread evidence for incipient ecological speciation: A meta-analysis of isolation-by-ecology. *Ecol. Lett.* **16:** 940-950.

Sicoli, M.A. & Holton, G. 2014. Linguistic phylogenies support back-migration from Beringia to Asia. *PLoS ONE* **9:** e91722.

Simonson, T.S., Yang, Y., Huff, C.D., Yun, H., Qin, G., Witherspoon, D.J., Bai, Z., Lorenzo, F.R., Xing, J., Jorde, L.B., Prchal, J.T. & Ge, R. 2010. Genetic evidence for high-altitude adaptation in Tibet. *Science* **329:** 72-75.

Smouse, P.E., Long, J.C. & Sokal, R.R. 1986. Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Syst. Zool.* **35:** 627-632.

Soares, P., Alshamali, F., Pereira, J.B., Fernandes, V., Silva, N.M., Afonso, C., Costa, M.D., Musilova, E., Macaulay, V., Richards, M.B., Cerny, V. & Pereira, L. 2012. The expansion of mtDNA haplogroup L3 within and out of Africa. *Mol. Biol. Evol.* **29:** 915-927.

Somel, M., Liu, X. & Khaitovich, P. 2013. Human brain evolution: Transcripts, metabolites and their regulators. *Nat. Rev. Neurosci.* **14:** 112-127.

Steward, J.H. 1955. *Theory of Culture Change: The Methodology of Multilinear Evolution*. Univ. of Illinois Press, Urbana.

Suomen Maantieteellinen Seura. 1929. *Suomen kartasto 1925 [Atlas of Finland 1925]*. Otava, Helsinki.

Sutherland, W.J. 2003. Parallel extinction risk and global distribution of languages and species. *Nature* **423:** 276-279.

Swadesh, M. 1952. Lexicostatistic dating of prehistoric ethnic contacts. *P. Am. Philos. Soc.* **96:** 452-463.

Swadesh, M. 1955. Towards greater accuracy in lexicostatistic dating. *Int. J. Am. Linguist.* **21:** 121-137.

Taagepera, R. 1994. The linguistic distances between Uralic languages. *Linguistica Uralica* **30:** 161-167.

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

Tallavaara, M., Luoto, M., Korhonen, N., Järvinen, H. & Seppä, H. 2015. Human population dynamics in Europe over the last glacial maximum. *PNAS* **112:** 8232-8237.

Tallavaara, M. & Seppä, H. 2011. Did the mid-Holocene environmental changes cause the boom and bust of hunter-gatherer population size in eastern Fennoscandia? *The Holocene* **22**: 215-225.

Tomasello, M. 1999. The human adaptation for culture. *Annu. Rev. Anthropol.* **28:** 509-529.

Trudgill, P. 2011. *Sociolinguistic Typology: Social Determinants of Linguistic Complexity*. Oxford University Press, Oxford.

Väliranta, M., Kaakinen, A. & Kuhry, P. 2003. Holocene climate and landscape evolution East of the Pechora Delta, East-European Russian Arctic. *Quaternary res.* **59:** 335-344.

Van Valen, L. 1973. A new evolutionary law. *Evol. Theor.* **1:** 1-30.

Virrankoski, P. 2012. *Suomen historia: Maa ja kansa kautta aikojen*. Suomalaisen Kirjallisuuden Seura, Helsnki.

Voje, K.L., Holen, ØH., Liow, L.H., Stenseth, N.C. 2015. The role of biotic forces in driving macroevolution: beyond the Red Queen. *Proc. R. Soc. B* **282**: 20150186.

Walker, R.S. & Ribeiro, L.A. 2011. Bayesian phylogeography of the Arawak expansion in lowland South America. *Proc. R. Soc. B* **278:** 2562-2567.

Weber, A.W., Jordan, P. & Kato, H. 2013. Environmental change and cultural dynamics of Holocene hunter–gatherers in Northeast Asia: Comparative analyses and research potentials in Cis-Baikal (Siberia, Russia) and Hokkaido (Japan). *Quaternary International* **290–291:** 3-20.

Wei, W., Ayub, Q., Chen, Y., McCarthy, S., Hou, Y., Carbone, I., Xue, Y. & Tyler-Smith, C. 2013. A calibrated human Y-chromosomal phylogeny based on resequencing. *Genome Research* **23:** 388-395.

Wiik, K. 2004. *Suomen murteet. Kvantitatiivinen tutkimus*. Suomalaisen Kirjallisuuden Seura, Helsinki.

Wright, S. 1943. Isolation by distance. *Genetics* **28:** 114-138.

Zhang, D.D., Lee, H.F., Wang, C., Li, B., Pei, Q., Zhang, J. & An, Y. 2011. The causality analysis of climate change and large-scale human crisis. *PNAS* **108:** 17296-17301.

Zhao, D., Wu, S. & Yin, Y. 2013. Responses of terrestrial ecosystems' net primary productivity to future regional climate change in China. *PLoS ONE* **8:** e60849.