Tampere University

KAJ SYRJÄNEN

Quantitative Language Evolution

Case studies in Finnish dialects and Uralic languages

Tampere University Dissertations 433

Tampere University Dissertations 433

KAJ SYRJÄNEN

Quantitative Language Evolution

Case studies in Finnish dialects and Uralic languages

ACADEMIC DISSERTATION
To be presented, with the permission of
the Faculty of Information Technology and Communication Sciences
of Tampere University,
for public discussion remotely
on Friday 6th August 2021, at 12 o'clock.

ACADEMIC DISSERTATION
Tampere University, Faculty of Information Technology and Communication Sciences
Finland

Responsible supervisor and Custos	Dr. Unni Leino Tampere University Finland	
Supervisors	Dr. Outi Vesakoski University of Turku Finland	Dr. Urho Määttä Tampere University (retired) Finland
Pre-examiners	Dr. Simon Greenhill Max Planck Institute for the Science of Human History Germany	Dr. Rigina Ajanki University of Helsinki Finland
Opponent	Dr. Annemarie Verkerk Saarland University Germany	

The originality of this thesis has been checked using the Turnitin OriginalityCheck service.

Cover design: Roihu Inc.

ISBN 978-952-03-2003-4 (print)
ISBN 978-952-03-2004-1 (pdf)
ISSN 2489-9860 (print)
ISSN 2490-0028 (pdf)
http://urn.fi/URN:ISBN:978-952-03-2004-1

PunaMusta Oy – Yliopistopaino
Joensuu 2021

ACKNOWLEDGEMENTS

Over the years I have received a great deal of support and assistance. Without these people and organizations this thesis would still be unfinished, or very likely not exist at all.

The first and foremost of these is the BEDLAN project, along with the large number of projects that have continued where it left off (UraLex, AikaSyyni, SumuraSyyni, UraLex, URKO). These projects have not only made this dissertation financially possible, they have also spawned a network of people that share an interest both evolutionary techniques and uncovering the history of Uralic languages and speakers from different perspectives. Through these projects I have been introduced to aspects far beyond linguistics, including genetics, archaeology, history, geography and computational research. This constant exposure to things outside of my original area of expertise has had a significant impact on my workflow, as well as helped me see the advantages of true multidisciplinarity.

I would also like to thank all the funding organizations that have helped me make this dissertation by means of funding and providing a working space. These include the Kone Foundation, the Finnish Academy of Science and Letters, Tampere University and the University of Turku.

My PhD supervisors, both past and present, all have my gratitude not only for their help and support as well as their patience and understanding when things did not move forward as efficiently as one would hope, or life made things complicated. These include Urho Määttä, who headed the original BEDLAN project, and who not only spent countless hours talking about linguistic theory, talking about life in general and having coffee at the Pyynikki observatory. I also thank Carita Klippi for her help in getting me to attend my first scientific conference in Paris, as well as her help with the theoretical side of my work, especially at the beginning stage of my PhD journey. Unni Leino not only headed the UraLex project that provided my funding for several years, but also co-wrote two of the papers that make up this dissertation, and helped with the practical side of getting over the finishing line. Last but not least, I thank

Outi Vesakoski for spearheading most of the BEDLAN's research projects I have been a part of, vehemently seeking funding, actively cowriting all of the papers (occasionally to an annoying degree, which is not necessarily a bad thing), and above all, being both understanding and supportive with the entire project throughout the years.

I am also grateful for the coauthors of the articles, who have been valuable work colleagues over the years. I am grateful for Kalle Korhonen, who brought linguistic expertise to the first two papers of the dissertation; Jyri Lehtinen, who likewise brought his linguistic insight to the papers and was responsible for compiling a large part of the linguistic data that this dissertation uses; Niklas Wahlberg and Jadranka Rota, who introduced me to phylogenetic inference techniques such as Bayesian tree building and the TIGER algorithm; Luke Maurits, whose mathematical, programming and technical expertise was invaluable to getting the last article finished; and Terhi Honkola, who started this PhD journey at around the same time as I did, although got over the finishing line quite a bit before I did, and who was generally always ready to brainstorm and discuss any and all problems and challenges over the phone or Skype or emails.

In addition, numerous people from close colleagues to more distant acquaintances, have provided valuable input, comments and support for this dissertation and its articles over the years. These include, but are not limited to, Mervi De Heer, Michael Dunn, Mikko Heikkilä, Anne-Mai Ilumäe, Miina Norvik, Timo Rantanen, Jenni Santaharju, Miikka Silfverberg, Sven-Erik Soosaar and Kristiina Tambets.

Before the ongoing pandemic largely isolated everyone from the university community, I had the pleasure of sharing an office with various colleagues over the years, including Mikko Höglund, Tommi Kakko, Larisa Leisiö and Kirsi Sandberg. My thanks go to all of you, as well.

Finally, I also need to thank my family (Reetta, Elsa, Kaj, Sirpa, Riitta, Riikka, Marko, Terttu, Jussi and family, and all other relatives) and my congregation of friends, who helped keep me grounded when times were tough, and celebrated when things went well.

ABSTRACT

This dissertation focuses on quantitative analysis techniques that relate to language evolution. Over the last few decades an increasing number of linguists have acknowledged the benefits that quantitative analysis approaches can bring to the study of language data, such as increased objectivity, ease of replication as well as increased ability to handle large volumes of data. In addition to a more widespread acceptance of quantitative approaches in general, this has also brought analysis techniques from other research fields to linguistics. This includes methods from evolutionary biology, such as phylogenetic techniques, used to study historical relationships between different species, and population genetic techniques, which explore the relationships between the populations of a species. In the field of linguistics phylogenetic techniques can be used to quantitatively study related languages, while population genetic techniques are useful for studying more closely-related languages or varieties of the same language.

The present work focuses on phylogenetic and population genetic analysis techniques through four research articles that explore the applicability of these approaches in linguistic research, and also proposes new quantitative techniques to the toolkits of historical linguistics and dialectology. It also brings evolutionary research to both Uralic languages and Finnish dialects. The research is part of the work done by the multidisciplinary BEDLAN research initiative and its follow-up projects.

The language data used in the research comes from a dataset of Uralic basic vocabulary, compiled as part of the BEDLAN research initiative and its follow-up projects, and made publicly available as part of this dissertation's work. This language family has not previously been explored using phylogenetic techniques, but is well-researched from a traditional standpoint, making it a good test subject for these techniques. In addition to the basic vocabulary data, the dissertation also uses dialect data from the digitized version of Lauri Kettunen's Dialect Atlas of Finnish, error-checked and revised by the BEDLAN project and released by Kotus. Similarly to the Uralic language family, Finnish dialects also represent a good testing ground for new population genetic analyses due to the large amount of traditional research done on

them. In addition to real-life datasets, the dissertation's research also employs simulated language data as part of its methodological exploration.

The four research articles show that evolutionary techniques work well for modeling relationships both between languages and within dialects, based on how the overall results compare to the literature. The articles also bring new techniques from evolutionary biology to linguistics, including population genetic clustering as a way of inferring dialect areas and a phylogenetic metric called TIGER values as a way of estimating how treelike a linguistic dataset is. The research that makes up this work has also laid groundwork for multidisciplinary research on languages as part of the overarching study of human history.

TIIVISTELMÄ

Väitöskirjani keskittyy kielievoluutioon ja sen laskennallisiin analyysitekniikoihin. Viime vuosikymmeninä laskennalliset lähestymistavat ovat kasvattaneet suosiotansa kielitieteen piirissä tarjoten muun muassa korkeampaa objektiivisuutta, helpompaa replikoitavuutta ja tapoja tutkia aiempaa isompia aineistoja. Lisääntynyt suosio on tuonut mukanaan uusia analyysitekniikoita muilta tutkimusaloilta kielitieteeseen. Näihin kuuluu evoluutiobiologiasta lainattuja menetelmiä, kuten fylogeneettiset tekniikat, joilla tutkitaan biologisten lajien välisiä suhteita, sekä populaatiogeneettisiä tekniikoita, joilla tutkitaan lajinsisäisten populaatioiden suhteita. Kielitieteessä fylogenetiikkaa voidaan soveltaa samaan kielikuntaan läheistä sukua olevien kielten tutkimiseen, ja populaatiogeneettisiä menetelmiä puolestaan lähisukukielten tai saman kielen sisäisten varianttien tutkimukseen.

Fylogeneettisiin ja populaatiogeneettisiin analyysitekniikoihin perehdytään tässä väitöskirjassa neljän tutkimusartikkelin kautta. Artikkeleissa tutkitaan lähestymistapojen soveltuvuutta kielitieteelliseen tutkimukseen sekä ehdotetaan uusia laskennallisia lähestymistapoja historialliseen kielitieteeseen ja murteiden tutkimukseen. Tutkimus esittelee evoluutiopohjaista metodologiaa uralilaisten kielten ja suomen kielen murteiden tutkimukseen. Koko tutkimus on toteutettu osana monitieteistä BEDLAN-hanketta ja sen jatkohankkeita.

Tutkimuksessa käytetään kieliaineistona uralilaisista kielistä kerättyä perussanastoaineistoa, joka on kerätty osana BEDLAN-hanketta ja sen jatkohankkeita, ja julkaistu osana tätä väitöskirjatyötä. Uralilaisia kieliä ei ole aiemmin tutkittu fylogeneettisilla tekniikoilla, mutta kielikuntaa on kattavasti tutkittu perinteisemmillä lähestymistavoilla; tämän suhteen kielikunta on otollinen näiden uusien tekniikoiden koestamiseen. Perussanastoaineiston lisäksi väitöskirja käyttää BEDLAN-hankkeen korjaamaa ja Kotuksen julkaisemaa Lauri Kettusen murrekartaston digitaalista versiota. Uralilaisten kielten tavoin myös suomen murteet tarjoavat otollisen kokeilualustan uusille tekniikoille niiden kattavan tutkimushistorian ansiosta. Näiden kieliaineistojen lisäksi väitöskirja käyttää myös simuloitua kieliaineistoa metodologian testaamisessa ja evaluoinnissa.

Tutkimusartikkelien tulosten perusteella evoluutiopohjaiset analyysitekniikat toimivat hyvin sekä kielten että murteiden tutkimiseen ottaen huomioon miltä saadut tulokset näyttävät aiemmin julkaistun tutkimuksen valossa. Artikkelit tuovat myös uusia tekniikoita evoluutiobiologiasta kielitieteeseen mukaan lukien murrealueiden inferenssi populaatiogeneettisen klusteroinnin avulla ja fylogeneettinen TIGER-metriikka tapana arvioida kuinka puumainen rakenne kieliaineistolla on. Artikkelit tarjoavat myös perustaa jo käynnissä olevalle monitieteiselle ihmishistorian tutkimukselle.

CONTENTS

List of Figures

List of Tables

ORIGINAL PUBLICATIONS

Publication I Syrjänen, Kaj, Terhi Honkola, Kalle Korhonen, Jyri Lehtinen, Outi Vesakoski & Niklas Wahlberg. (2013). "Shedding more light on language classification using basic vocabularies and phylogenetic methods: A case study of Uralic." *Diachronica* 30(3), 323–352. DOI: 10.1075/dia.30.3.02syr.

Publication II Lehtinen, Jyri, Terhi Honkola, Kalle Korhonen, Kaj Syrjänen, Niklas Wahlberg & Outi Vesakoski. (2014). "Behind family trees: Secondary connections in Uralic language networks." *Language Dynamics and Change* 4: 189–221. DOI: 10.1163/22105832-00402007.

Publication III Syrjänen, Kaj, Terhi Honkola, Jyri Lehtinen, Antti Leino, & Outi Vesakoski. (2016). "Applying population genetic approaches within languages: Finnish dialects as linguistic populations." *Language Dynamics and Change* 6: 235–283. DOI: 10.1163/22105832-00602002.

Publication IV Syrjänen, Kaj, Luke Maurits, Unni Leino, Terhi Honkola, Jadranka Rota & Outi Vesakoski. (forthcoming). "Crouching TIGER, Hidden Structure: Exploring the nature of linguistic data using TIGER values." *Journal of Language Evolution.*

1 INTRODUCTION

1.1 Overview

The main focus of this dissertation is on exploring languages with evolutionary quantitative methods. A fundamental building block of this kind of approach is the idea that, on a general level, language change operates similarly to how evolutionary processes such as biological evolution do, making it possible to also apply similar quantitative techniques in both fields (see Leino *et al.* 2020). The idea of language change as a kind of cultural evolutionary process is by no means a new idea; as is discussed later in the background section, languages have been likened to biological organisms and language change to evolutionary processes already in the nineteenth century (e.g. Aronoff 2017, Atkinson & Gray 2005). What the most recent wave of evolutionary language study has brought to linguistics is a robust toolkit of quantitative techniques originating from the largely data-driven field of evolutionary biology. These include phylogenetic tools, such as Bayesian tree-building techniques, and population genetic tools, such as ancestral population inference.

The new tools adopted from evolutionary biology make it easier to study large and complex datasets, which are still more prominent in the biological realm, although linguistics also has a rich selection of digitally available material to be explored. Many of the tools are also designed to be more malleable than the earlier quantitative tools used in historical linguistics, such as lexicostatistical techniques (see e.g. Embleton 1986, McMahon & McMahon 2005), and are obviously more advanced in other ways. Some of these tools have also made their way to other research fields, and have facilitated multidisciplinary work between e.g. biology, linguistics and archaeology. Consequently, these techniques give hope to more ambitious research in the future where linguistic and extralinguistic data can be analysed together in a quantifiable way to shed light on language history as one of the many parts that make up the overall history of the human species.

While recent years have seen considerable interest in these new quantitative tools, their acceptance is not universal. Problems have been pointed out in several areas, including the theoretical basis of modeling language change as an evolutionary process, the challenge of coding linguistic data in a format suitable for these tools, and the suitability of the existing tools to explore linguistic data. This highlights the important fact that while evolutionary thinking has a long history in linguistics, the use of evolutionary techniques to study languages is a recent phenomenon and as such its tools are constantly being improved. As part of this dissertation I will also critically examine the challenges of using these techniques and discuss the problems.

Structurally this dissertation consists of four research articles, whose topics focus on testing and applying evolutionary approaches for language research, as well as showcasing new ones. Three of the articles focus on the exploration of related languages, analogous with species-level analyses in biology, while one article focuses on exploring variation within a language, analogous to population-level analyses in biology.

Two types of linguistic data are used in the articles: (1) lexical cognate data recording historical relationships between Uralic languages, collected specifically for the testing of these evolutionary analysis techniques and publicly released as part of this dissertation work, and (2) dialect atlas data recording regional linguistic variation in the Finnish language. As such, the articles also contribute to Uralic language research and Finnish dialect research, although their primary objective is the exploration of quantitative methodology and evolutionary approaches as a way of studying languages. In addition to the datasets mentioned above, one of the articles, which showcases TIGER, a new quantitative metric adopted from phylogenetics, uses simulated language data produced using generative models that mimic different linguistic scenarios.

Finally, it should be noted that the topic of this dissertation is fundamentally a multidisciplinary one, requiring expertise from the linguistic field as well as the biological field. As such, it would have been difficult if not impossible to do in isolation. This work has been done as part of the multidisciplinary BEDLAN (Biological Evolution and the Diversification of Languages) research initiative, involving both biologists and linguists, as well as its various follow-up projects, including UraLex, SumuraSyyni, AikaSyyni and URKO. This is also reflected in the articles that make up this dissertation, which are all co-authored.

1.2 Aims of the thesis

The four articles of this thesis revolve around both general-level questions that relate to current trends in quantitative language research, as well as questions specific to the study of the two main research objects, Uralic languages and Finnish dialects. The main emphasis of this dissertation is in methodology; thus the following two questions can be regarded as its primary research questions:

1. How applicable are phylogenetic and population genetic tools for studying language history and dialect variation? (Articles I-IV)

2. Can we introduce additional useful quantitative tools from the toolkit of evolutionary biology to the study of languages? (Articles III and IV)

However, as the work also involves the analysis of real-life linguistic data – basic vocabulary data from Uralic languages and dialect atlas data from Finnish – it also addresses the following two questions:

3. How are Uralic languages characterized by quantitative phylogenetic tools, including trees, networks and related metrics? (Articles I and II)

4. How are Finnish dialects characterized by population genetic clustering tools and related techniques? (Article III)

These make up the four main research questions of this dissertation.

2 THEORY AND BACKGROUND

As was already established, the main focus of this dissertation is on exploring quantitative evolutionary tools and their applicability to linguistic data. I will begin by unpacking the history of evolutionary thinking in linguistics. This will by no means be a comprehensive view of the topic of evolutionary concepts in linguistics, which in and of itself would be broad enough to fill a dissertation of its own (see e.g. Schleicher 1869, Sankoff 1973, Croft 2000; 2008, Ritt 2004, Atkinson & Gray 2005, Richerson & Boyd 2005, Livingstone 2003, Winter-Froemel 2008, Pappas & Mooers 2011, Pagel 2009, Aronoff 2017, Chomsky 2017, Creanza *et al.* 2017). In addition, it can be argued that, in many cases, the application of quantitative evolutionary techniques does not require an excessively detailed look at the relationship between linguistic change and evolution (see Leino *et al.* 2020 and further in this dissertation). However, it is valuable to provide a summary glance at the relationship between evolution and languages, pinpoint this study in the diverse field that involves evolution and languages, as well as introduce some of the core concepts that can be considered necessary for applying these quantitative techniques.

Following a look at the relationship between evolution and language, I will provide an overview of the quantitative side of evolutionary studies, focusing particularly on the type of tools applied in this dissertation. These include tools used for building phylogenetic trees and networks, such as MrBayes and NeighborNet, and metrics for evaluating treelikeness, such as δ (delta) scores, Q-residuals and TIGER values; these are central to Articles I, II and IV. It also takes a look at quantitative dialectology alongside tools used in population genetics for inferring population structure, such as the Structure algorithm, used extensively in Article III.

In the final section I will set the stage for the four studies by providing an overview of the Uralic language family and Finnish language from the perspective of its geographical dialects, which serve as this dissertation's macro-level and micro-level study objects, respectively.

2.1 Language change and evolution

2.1.1 Evolution in a linguistic context

One of the first problems we encounter is the question of what is meant by 'evolution' in a linguistic context. Historically, linguistics and the study of biological evolution crossed paths significantly in the latter parts of the nineteenth century; linguistic theories such as August Schleicher's *Stammbaumtheorie* was inspired by Darwin's theory of evolution, which at that point was still in its infancy. During this time 'evolutionary linguistics' essentially became more or less a synonym for what we now call 'diachronic linguistics' (Aronoff 2017), the latter replacing the former after connotations of evolution became less popular around the beginning of the twentieth century. Over the last few decades 'evolution' has resurfaced in linguistics, with more complexity in its meaning than it had earlier.

Da Silva (2010) identifies three senses for 'evolution' in a modern linguistic context:

1. evolution of the linguistic capacity
2. origin of contemporary languages
3. evolutionary models of language change

First of these senses focuses on the changes that led to the creation of the sociocognitive capacity for language as a communicative system. It is generally associated with sociobiology and evolutionary psychology. In a way this sense is most literal. Notably, this sense constrains 'evolution' to specifically refer to biological evolution rather than non-biological changes within a linguistic system, as seen in the following passage from Chomsky:

> [W]hat has evolved is not languages, which do not evolve, in the technical sense of the term, any more than states of the visual system evolve. Rather, what has evolved is the capacity for language (LC), analogous to the genetic basis for a mammalian, not insect, visual system. (Chomsky 2017)

Chomsky traces the origins of this line of research to the 1960s, specifically to the book *Biological Foundations of Language* (Lenneberg 1967). This line of research can be seen in e.g. studies of the FOXP2, a gene thought to be associated with the

human speech capacity (Marcus & Fisher 2003), as well as Chomskyan studies related to the so-called language organ.

The second sense, the origin of contemporary languages, relates to correlations between genetic, archaeological and linguistic diversification, as well as explorations of gene-culture coevolution (see e.g. Richerson & Boyd 2005). This approach is also sometimes termed the 'new synthesis', especially when talking about the interrelationships between genetic and linguistic connections (McMahon & McMahon 2005, Atkinson & Gray 2005). McMahon & McMahon (2005: 119) traces the kick-off point for this line of research to the work of Cavalli-Sforza *et al.* (1988), which included a far-from-perfect parallel linguistic-genetic tree. Cavalli-Sforza's work highlighted that in order for this type of research to be successful it needs to be truly multidisciplinary, i.e. there must be both genetic expertise and linguistic expertise, not just one or the other. Interestingly, this second sense is in many ways closest to the connotations of the historical term 'evolutionary linguistics', which, as mentioned earlier, was later replaced by the less loaded term 'diachronic linguistics' (Aronoff 2017).

Returning to Da Silva's three senses for 'evolution', the third and last sense involves the application of evolutionary concepts to the theory of language change. Most of the work on evolution as part of the theory of language change is fairly recent, such as e.g. Croft's Theory of Utterance Selection (Croft 2000, 2006, 2008). However, as mentioned earlier, the roots of this line of thinking, at least in a rough sense, trace back to the nineteenth century, as can be seen in the following passage from Darwin, where he suggests that language change and biological evolution operate under somewhat similar processes, albeit in different contexts:

> The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously the same. […] We find in distinct languages striking homologies due to community of descent, and analogies due to a similar process of formation. (Darwin 1871)

At the time, the line of thinking from the passage from Darwin was perhaps most prominently found in August Schleicher's *Stammbaumtheorie* (family tree theory), which characterized languages as diverging under a tree-like pattern, which bore resemblance to the Darwin's tree of life. Schleicher notably wrote an influential pamphlet titled *Die Darwinsche Theorie und Die Sprachwissenschaft* (Schleicher 1863), translated to English as *Darwinism tested by the science of language* (Schleicher 1869). The main argument of the pamphlet was that the results of

historical linguistics over the past half century essentially served as the first tangible evidence that Darwin's theory of evolution was indeed valid (Aronoff 2017, see also Muller 1870). Notably, Schleicher's *Stammbaumtheorie* was not unanimously accepted, which brought about alternative theories, such as the *Wellentheorie* (wave theory), proposed by Hugo Schuchardt and Johannes Schmidt in 1872; see François (2014) for examples of the criticism posed against the tree-model. Schleicher's positioning of linguistics as a natural science was also generally disapproved. Eventually Saussure's work shifted focus of linguistics from the study of historical linguistics akin to evolutionary processes to the synchronic study of grammar in the early twentieth century. The potential parallels between linguistic and biological systems would start to get more attention from the late 20th century onwards from various sides, including, for instance, the quantitative linguistic field, which included bioinformaticians and geneticists as well as linguists (see e.g. Sankoff 1973, Cavalli-Sforza & Wang 1986), evolutionary theories of language change (e.g. Croft 2000; 2006; 2008) and cultural evolutionary theories (e.g. Boyd & Richerson 1985, Richerson & Boyd 2005).

There is also room for interpretation as to what we mean when we talk about 'evolutionary models' of language change. Croft (2000) distinguishes three subtypes: *literal*, *analogical* and *generalized*. Essentially the same division is suggested in Winter-Froemel (2008), which however uses the terms *biologistic*, *metaphorical* and *generalized*, respectively; I will use Croft's terminology below. Of the three subtypes, the literal type assumes that language is a fundamentally genetic capacity and as such follows the principles of biology. Da Silva (2010) adds that the literal subtype is largely associated with Chomskyan linguistics, where specific universal properties of languages are argued to have a biological basis. The analogical subtype is focused on specific analogies identified between biological evolutionary processes and processes of language change, without committing to deeper connections between the theory of language change and evolution. Consequently, analogical similarities are generally quite superficial (for an example, see e.g. Table 1 further below). The third subtype, generalized, covers theories which are built on top of a generalized theory of selection and also attempt to commit to a deeper connection between the theory of evolution and the theory of language change. Croft's *Theory of Utterance Selection* (Croft 2000; 2006; 2008) is an example of the generalized type, and is based on David Hull's *General Analysis of Selection* (Hull 1988), on top of which he attempts to flesh out a full theory of language change based on evolutionary thinking.

Of the three senses of evolution discussed here, the third sense of the term – 'evolutionary models of language change' – is perhaps closest to the overall subject matter of this dissertation, which involves analyses of purely linguistic material using tools that are built on a kind of evolutionary model. Having said that, however, the analysis tools applied in this dissertation are also commonplace in studies that infer history by combining or correlating linguistic data with extralinguistic information, which corresponds with the second sense of evolution discussed above – 'origin of contemporary languages'. Examples of such studies within the Uralic language family include Honkola *et al.* (2013), Honkola *et al.* (2018); others include, for instance the phylogeographic analyses of Bouckaert *et al.* (2012), which apply combine geographical information with linguistic data.

With respect to Croft's subtypes, this dissertation falls mostly in the 'analogical' subtype, as it does not commit to a fully fleshed-out evolutionary theory. Rather, the work mainly addresses similarities between language change and biological evolution whenever they relate to the internal workings of the quantitative techniques themselves. At these points it is important to ensure that the principles underlying the techniques do not clash with our understanding of how language change works. With this in mind, the works that make up this dissertation do not need to be regarded as evolutionary approaches to linguistics; they can also be regarded simply as works exploring and showcasing new computational techniques for linguistics, all of which originate from the realm of quantitative biology.

2.1.2 Fundamental principles of evolutionary models

While the present work does not lean on any larger generalized theory of language change, the techniques that are used were originally devised within the field of evolutionary biology. As such, they require a certain degree of similarity to exist between how language change works and how biological evolution work (see Leino *et al.* 2020). In this section I go through the fundamental principles that evolutionary techniques generally lean on, without delving too deep in details. These can be thought of as the first stepping stone when going deeper into the details of each technique.

There are three fundamental principles that underlie evolution of any kind. Audersirk, Audersirk & Byers (2008) defines these in the context of biological evolution as *inheritance*, *variation* and *natural selection*. Formally these three processes

that define evolution trace back at least to Richard Lewontin's 1970 paper "The Units of Selection" (Lewontin 1970); notably, Lewontin's paper highlights the fact that these principles are, by intention, defined quite generally:

> It is important to note a certain generality in the principles. No particular mechanism of inheritance is specified, but only a correlation in fitness between parent and offspring. The population would evolve whether the correlation between parent and offspring arose from Mendelian, cytoplasmic, or cultural inheritance. (Lewontin 1970)

With this in mind, it should come as no surprise that these principles, in one form or another, are present in non-biological studies that revolve around evolution, including linguistic studies. Croft (2000), in his *Theory of Utterance Selection*, uses the terms *replication*, *variation* and *selection*, while Honkola (2016), following Futuyma's definition of evolution, uses the terms *heritability*, *variation* and *change-causing forces* for the same three principles. Article III of this dissertation, which focuses on analysing linguistic data at the population-level, likewise specifies three similar ingredients for applying biological methodology for language data: (1) the existence discrete heritable units (i.e. heritability), (2) the existence of physical carriers for the heritable units (i.e. variation), and (3) the existence of forces affecting the variant frequencies of the heritable units (i.e. selection or change-causing forces).

Heritability refers to the existence of heritable units, or 'replicators', that persist through time by being transmitted from one generation to the next. The heritable units may be defined at various levels of fine-grainedness[1], such as e.g. genes, alleles (variant of a gene) nucleotides and amino acids in the context of biology, and e.g. words, phrases, constructions or utterances in the context of languages. These units are carried over from one generation to the next quite differently in the two domains, with biological systems transferring their units by processes such as genetic inheritance and horizontal gene transfer (e.g. Gasmi *et al.* 2015, Campbell *et al.* 2008), and languages doing the same through communication and learning between the language users.

Variation in a biological context refers to the existence of genetic differences between the individuals of a biological population; in a linguistic context this can be interpreted as the existence of distinct language use patterns, or 'linguistic repertoires', as termed by Croft (2000), between the individuals of a linguistic

[1] Notably, different evolution-based linguistic frameworks, such as those by Ritt and Croft, also tend to have replicators that represent different levels of fine-grainedness (Pappas & Booers 2011).

population. What these specific differences are considered to be depends somewhat on the level one is investigating, which in biology may range from differences between the populations of one species (micro-level) to differences between species (macro-level); in both cases, different types of heritable units are explored. Similarly, variation and its nature is defined somewhat differently when looking at the relationships between within-language populations (lects) and relationships between languages. The data itself also plays a role in defining how linguistic variation is seen, as different data sets cover different types of linguistic units. What is common in both biological evolution and language change, regardless of the type of variational data point being observed, is the idea that variation gives rise to change.

Selection refers to the existence of internal or external forces that to some extent direct how variation changes from one generation to the next. In biological context perhaps the best known selective force is natural selection, defined as "the enchanced reproduction of organisms with variations that help them cope with their environment" (Audersirk *et al.* 2008), and in and of itself includes different subtypes of selection with different outcomes, such as directional selection (in which one extreme of continuous variation is favored), stabilizing selection (in which an intermediate state between continuous variation is favored) and diversifying selection (in which both extremes of continuous variation are favored). Similarly, we can model languages as a system that is affected by a combination of predictable and unpredictable changes (selective pressures), which gives rise to the observable variation and modifies it in various ways in the course of linguistic history.

In addition to the three aforementioned fundamental elements that make up an evolutionary system – heritability, variation and selection – more deep-level analogies or parallels are also occasionally suggested between biological and linguistic systems. For instance, Pagel (2009) provides such a list (reproduced in Table 1 without footnotes from the original), which lists many parallels suggested already in Darwin's *The Descent of Man* in 1871. It should be noted that detailed lists of parallels such as this are not necessarily useful from the practical perspective of computationally modeling language change as an evolutionary phenomenon. They may in fact be quite misleading if taken too literally, as they often imply mechanisms that are domain-specific, or include details that are only partially valid (e.g. Croft 2008: 220, List 2016). To give an example of a potentially misleading metaphorical parallel from Pagel's table: while biological mechanisms for

replication can be likened to teaching, learning and imitation in languages, language learning takes place more gradually (in small pieces) as opposed to biological replication. Linguistic information is also exchanged between already existing speakers, and the exchange does not produce a new speaker, unlike its biological counterpart. Another example of a misleading parallel in the table is 'drift', which in biological context refers to changes in the allele frequency in a population due to random selection, whereas in linguistics the term often refers to gradual change that is unconscious but directed rather than random. List (2016) also points out that the parallel terms 'homology' and 'cognates' are also misleading, as the biological notion of 'homology' has different subtypes ('orthology', i.e. genes related via speciation; 'paralogy', i.e. genes related via duplication; 'xenology', i.e. genes related with horizontal transfer of genetic material), whereas 'cognacy' refers specifically to shared descent from a common ancestor, and does not cover horizontal historical relationships, i.e. borrowing.

Table 1. Parallels between biological and language evolution, based on Pagel (2009).

Biological evolution	**Language evolution**
Discrete heritable units (for example, nucleotides, amino acids and genes)	Discrete heritable units (for example, words, phonemes and syntax)
Mechanisms of replication	Teaching, learning and imitation
Mutation (for example, many mechanisms yielding genetic alterations)	Innovation (for example, formant variation, mistakes, sound changes, and introduced sounds and words)
Homology	Cognates
Natural selection	Social selection and trends
Drift	Drift
Cladogenesis (for example, allopatric speciation (geographic separation) and sympatric speciation (ecological or reproductive separation))	Lineage splits (for example, geographical separation and social separation)
Anagenesis	Linguistic change without split
Horizontal gene transfer	Borrowing
Hybridization (for example, horse with zebra and wheat with strawberry)	Language Creoles (for example, Surinamese)
Correlated genotypes and phenotypes (for example, allometry and pleiotropy)	Correlated cultural terms (for example, 'hasta' and 'spear')
Geographic clines	Dialects and dialect chains
Fossils	Ancient texts
Extinction	Language death

In summation, the general principles of language change and biological evolution resemble each other on a general level, but this analogy does not necessarily extend into more fine-grained similarities. From the perspective of quantitative tools a fully analogous relationship is not a necessity, as each tool or analysis technique focuses on approximating a very specific evolutionary scenario as simply as it can, without modeling every piece from the underlying theory of biological evolution. Thus, as long as the specific quantitative tool that is being applied is not built on top of any fine-grained evolutionary assumptions that would be entirely nonsensical in the context of language change, it is likely that they can be successfully applied to study language change (Leino *et al.* 2020).

2.2 Quantitative tools of language evolution

As was already pointed out, quantitative evolutionary tools are generally such that they operate only on the basis of a small subset of evolutionary principles. In other words, they assume that specific evolutionary principles are valid, but do not require one to operate under a holistic evolutionary theory of language change. Notably, as the evolutionary principles vary between methods, one needs to be careful in choosing which tools are useful in a context such as linguistics.

2.2.1 Phylogenetic (macro-level) tools

The macro-level phylogenetic tools, used for inferring the internal classification of related languages, represent some of the most prominently used new tools adopted from evolutionary biology. They have indeed been applied to a wide variety of language families, such as Indo-European (e.g. Gray & Atkinson 2003, Bouckaert *et al.* 2012, Chang *et al.* 2015), Austronesian (Gray & Jordan 2000, Greenhill & Gray 2009), Japonic (Lee & Hasegawa 2011), Bantu (Holden 2002) and Uralic (Articles I and II, Honkola *et al.* 2013). The techniques are also discussed and explored in a broader context in e.g. McMahon & McMahon (2005), Nichols & Warnow (2008) and Greenhill *et al.* (2020). Phylogenetic tools have gradually developed in biology from 1950s onwards (Atkinson & Gray 2005, Felsenstein 2004), when aspects such as the rise of the 'modern synthesis', and the discovery of DNA called for more efficient quantitative tools to study the increasingly complex biological datasets

(Atkinson & Gray 2005, Felsenstein 2004). Among the first such approaches were the distance-based methods of Sokal & Sneath, developed in late 1950s.

Intriguingly, five years before Sokal & Sneath, Morris Swadesh popularized a similar distance-based calculation approach called *lexicostatistics* for historical linguistics. While being the most often-cited quantitative approach of this kind, Swadesh's lexicostatistics was not in fact the oldest quantitative means of exploring language classification (see e.g. Hymes 1983, Embleton 1986; 2000); the oldest precursor to lexicostatistics, according to Hymes's history of lexicostatistic techniques (1983: 68–69), was developed by Dumont d'Urville in the early 19th century, which did not become popular at the time. Hymes also notes (1983: 69–70) that Swadesh himself characterized lexicostatistics to be a reinvention of a technique introduced by Dixon and Kroeber in 1919. The lexicostatistic technique uses cognacy judgments of basic vocabulary meanings (discussed further in the Methods section) as a basis for building a distance matrix between languages (and occasionally dialects, as noted in Embleton 1986). The distance matrix could then be used to quantitatively assess how closely related the languages are to one another, and also to produce a tree classification. Swadesh also proposed a quantitative approach related to lexicostatistics called *glottochronology*[2], which aimed to infer divergence times for languages, based on the assumption that basic vocabulary undergoes a gradual process of change due to effects such as taboo-related change, substratum or superstratum effects, borrowing, semantic shift, semantic extension or narrowing. The glottochronological model assumed that the rate at which basic vocabulary items changed was constant, similarly to how radioactive decay works (Embleton 1986). Incidentally, a timing technique that is in essence identical to Swadesh's glottochronology was suggested by the anthropologist Paul Broca as early as 1862, based on the pre-lexicostatistical calculations of Tahitian-Hawaiian published by Dumont d'Urville (Hymes 1983: 93–108); while this technique did not draw inspiration from radioactive decay, it also had an assumption of a constant rate of change. As noted in Atkinson & Gray (2005) the assumption of a constant rate of change in glottochronology is quite similar to the idea of the molecular clock, which was introduced to biology in early 1960s, likely as a parallel development.

[2] Notably, the terms 'lexicostatistics' and 'glottochronology' are often confused with one another or used interchangeably, even though the scope and the objectives of these two approaches are distinct, with glottochronology being a specific extension and application of lexicostatistics (McMahon & McMahon 2005).

While in biology the quantitative tools continued to be developed quite actively, the same cannot be said of their linguistic counterparts. Swadesh's methods – particularly glottochronology – received mixed reactions after initial enthusiasm faded. Embleton (1986) devotes separate summaries for pro-glottochronological studies, anti-glottochronological studies and neutrally positioned studies, while Atkinson & Gray (2005) note that after initial enthusiasm Swadesh's methodology was heavily criticized and was eventually largely discredited in general historical linguistics (see also e.g. Crystal 1987). Embleton (1986; 2000) attributes the cold reception of the glottochronological methodology at least partially due to linguists treating the model as *deterministic* (i.e. one that produces the 'correct' solution) rather than *stochastic* (i.e. one that produces the 'most probable' solution given the data and the underlying model), which led to its virtual abandonment among linguists before the methodology had sufficient time to mature. The lexicostatistic model continued to be refined among those that did not abandon it, which largely included mathematicians and area-studies specialists, with improvements for taking into account various factors that the original techniques did not address, such as heterogeneous replacement rates, chance cognations and recurrent cognations, borrowing rates, and the presence of multiple synonyms (Embleton 1986; 2000, McMahon & McMahon 2005). Swadesh's techniques were also followed by other quantitative approaches (see e.g. Embleton 2000).

The biological phylogenetic toolkit is generally seen to have arrived to linguistics in the wake of the 'new synthesis' (McMahon & McMahon 2005), referring to crossdisciplinary studies that correlate linguistic and extralinguistic evidence quantitatively (see also section 2.1.1). Both McMahon & McMahon (2005) and Atkinson & Gray (2005) trace this trend back to a study by Cavalli-Sforza *et al.* (1988), which compared genetic and linguistic trees. This study generated a fair amount of controversy, as its results included e.g. genetic populations sampled on the basis of linguistic rather than genetic criteria, as well as controversial deep-level language groupings, such as Amerind, Nostratic and Eurasian, which are generally regarded as being impossible to prove due to linguistic entropy; most linguists consider that after around 8000 to 10000 years it becomes impossible to distinguish chance resemblances and vertical historical connections from one another (McMahon & McMahon 2005, Atkinson & Gray 2005). However, it also generated enthusiasm, and brought various studies that similarly correlated extralinguistic (often genetic) evidence with linguistic data, and also approached linguistic data using analysis tools from outside of linguistics. Such studies include Gray & Jordan (2000), which examined Austronesian expansion hypotheses on the basis of a

phylogenetic parsimony analysis of a Swadesh-list dataset done with biological tools. This was followed by other works that used parsimony analysis, such as Holden (2002) and Rexová *et al.* (2003), which explored Bantu and Indo-European data, respectively. Other phylogenetic and cladistic techniques were applied to language data at this time. McMahon & McMahon (2005) applied a wide variety of techniques, including Neigbour-Joining trees, maximum likelihood trees and NeighborNet networks. Another study which examined different biological tree-building techniques extensively was Nakhleh *et al.* (2005). Attention soon turned mostly to a newer and more flexible tree-building technique, Bayesian inference of phylogeny, which had also gained foothold in biology, with studies such as Gray & Atkinson (2003), Pagel *et al.* (2007), Dunn *et al.* (2008), Lee & Hasegawa (2011) and Bouckaert *et al.* (2012). These techniques also expanded to various directions where the phylogenetic analysis served only as part of an analysis pipeline; good examples of this are word rate analyses in Pagel *et al.* (2007) and the phylogeographic analysis of Bouckaert *et al.* (2012), which both integrate Bayesian phylogenetic techniques as part of a larger analysis chain. Some works also applied multiple techniques, such as Dunn *et al.* (2008), which used several types of analysis, including Bayesian inference of phylogeny, *NeighborNet* as well as the population genetic clustering tool *Structure*, to study a group of unrelated Papuan language families. The study is also notable in that it studies structural features coded as phylogenetic characters rather than basic vocabulary items, as most of the studies do. While the new quantitative methods have been better received than earlier quantitative attempts, they have also not been universally approved (see, for instance, the highly critical Pereltswaig & Lewis (2015) as well as a response to that criticism in Verkerk (2016).

Articles I and II involve the use of Bayesian inference of phylogeny using the software package *MrBayes* (Huelsenbeck & Ronquist 2001), with Article II and IV also applying non-Bayesian NeighborNet analyses (Bryant & Moulton 2004), using the tool *SplitsTree*. The focus of Articles I and II is in exploring the overall classification of Uralic using only linguistic material. On a larger scale they act as a validation process for related studies that analyze the Uralic language data alongside extralinguistic information using phylogenetic approaches (e.g. Honkola *et al.* 2013). This ensures our familiarity with the linguistic material that is being investigated. In addition to tools used for constructing phylogenetic trees and networks, Article IV, and to some extent Article II, explore supplementary metrics that estimate how tree-like a dataset is – a crucial step prior to a phylogenetic analysis especially when working with cultural data, which is often characterized by a considerable amount of reticulation, or nontreelike structure (see e. g. Wichmann

et al. 2011, Verkerk 2019, Nelson-Sathi *et al.* 2011, Gray *et al.* 2010, Jacques & List 2019). These include metrics used in existing studies - *δ scores* (Holland *et al.* 2002) and *Q-residuals* (Gray *et al.* 2010) - as well as a showcase of a new metric for this purpose - the TIGER value (Cummins & McInerney 2011).

2.2.2 Population genetic (micro-level) tools

The micro-level tools of evolutionary biology originate from the field of population genetics, and are used for quantitatively inferring within-species populations. In a linguistic context these can be likened to intralingual varieties, such as dialects. Thus, this research is essentially a type of *quantitative dialectology*, where dialect clustering is performed with the help of population genetic tools rather than more traditional dialectometrical tools. The use of quantitative approaches for studying and inferring intralingual varieties is not a new idea; while Jean Séguy's work on dialectometrical distances in 1973 is generally regarded as the beginning of quantitative dialect study, the dialectometrical research tradition also had predecessors (Nerbonne & Kretzschmar 2013), similarly to how there was quantitative language classification even before Swadesh's lexicostatistical approaches. Unlike lexicostatistics, however, quantitative dialect studies were generally better received than lexicostatistical approaches (Chambers & Trudgill 1998), perhaps due to changed attitudes or more understanding towards quantitative and statistical techniques at the time compared to the time when quantitative techniques were brought to historical linguistics.

The current toolkit of quantitative dialectology is quite large, covering a wide range of multivariate techniques (see e.g. Chambers & Trudgill 1998; Nerbonne & Wieling 2017). The main focus of this dissertation is in clustering techniques, which are used in both quantitative dialectology, where they aim to group individual observations together and thus identify dialect groups, and in population genetics, where they are used to infer a population structure based on genetic data. Unsurprisingly, both these approaches have relied on rather similar techniques in their research endeavours, including hierarchical and partitional clustering, principal component analysis and multidimensional scaling. A notable difference between these research areas is that from the 21st century onwards population genetics has increasingly adopted Bayesian model-based clustering approaches designed specifically for genetic clustering and the inference of genetic admixture (e.g. Novembre 2016 and references therein). Tools that work along these lines

include e.g. *Structure*, *BAPS*, *TESS* and *Geneland*, while quantitative dialectology has mainly relied on generic clustering techniques, especially hierarchical techniques such as e.g. UPGMA, WPGMA and Neighbor Joining (Nerbonne & Wieling 2017), although other techniques have also been applied to some extent (see e.g. Leino *et al.* 2006, Hyvönen *et al.* 2007). Perhaps the most attractive feature of population genetic techniques from a dialectological perspective is in their admixture modeling, essentially a type of 'fuzzy clustering' where each data point is assigned a degree of membership in multiple clusters. This contrasts with the more traditional 'hard clustering', where each data point is assigned to one and only one cluster. The concept of admixture fits well into many dialect data, which generally represent continuous linguistic variation across a geographical area.

Article III in this dissertation represents a large-scale pilot study on applying evolutionary analysis techniques from population genetics to Finnish dialect data. The main focus of this work involved inferring dialects and transitional dialects similarly to how populations and mixtures of populations within a biological species are inferred in population genetics. The main tool for this was the Bayesian population genetic tool *Structure* (Pritchard *et al.* 2000), used to infer populations from genotype data. Alongside this it also applies the more generic clustering technique (K-Medoids clustering), mainly to provide a non-evolutionary point of comparison for the Structure analyses. While the main focus of this work is on population clustering, the study also applies various population genetic metrics that are often used alongside these; these allow, for instance, the estimation of dialectal distances.

2.3 Overview of study objects

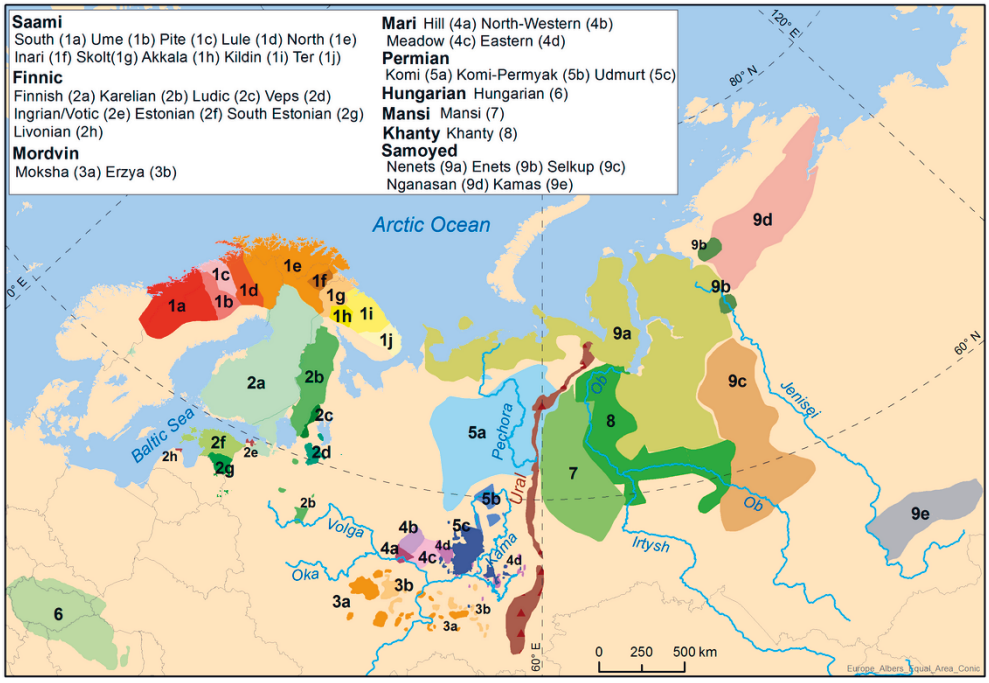
2.3.1 Uralic languages

Articles I and II, which focus on the macro-level of language evolution, explore the Uralic language family with the help of quantitative tools. Article IV also includes analyses of Uralic language data although for the most part it relies on different types of simulated language data. Here I provide a brief overview of this language family. A similar overview is also found in the articles themselves. Notably, this dissertation as a whole focuses more on quantitative methodology than Uralic

linguistics; consequently, this will by no means represent an exhaustive look at the Uralic language family and the research around it.

The Uralic language family includes around 40-60 languages (the exact number depends on factors such as what qualifies as a separate "language" and whether unattested languages are counted) with about 25 million speakers altogether. The Uralic speaker area (Figure 1) is geographically quite broad, ranging across Europe and Siberia (Salminen 2007, Janhunen 2009). The three largest languages of the family are Hungarian, Finnish and Estonian, the official languages of Hungary, Finland and Estonia, respectively; the other languages are generally minority languages. The language family has been studied since the beginning of modern historical comparative linguistics, and to some extent even before this (e.g. Korhonen 1986, Hovdhaugen *et al.* 2000).

Figure 1. Approximate geographical distribution of Uralic languages, based on Rantanen *et al.* (ms).



The Finno-Ugric *Stammbaum*, proposed by Otto Donner in the late nineteenth century, serves as the foundation for what is often called the 'standard paradigm' of Uralic phylogeny (Korhonen 1986, Hovdhaugen *et al.* 2000). In Donner's interpretation the ancestral Finno-Ugric language split into Ugric and Finno-

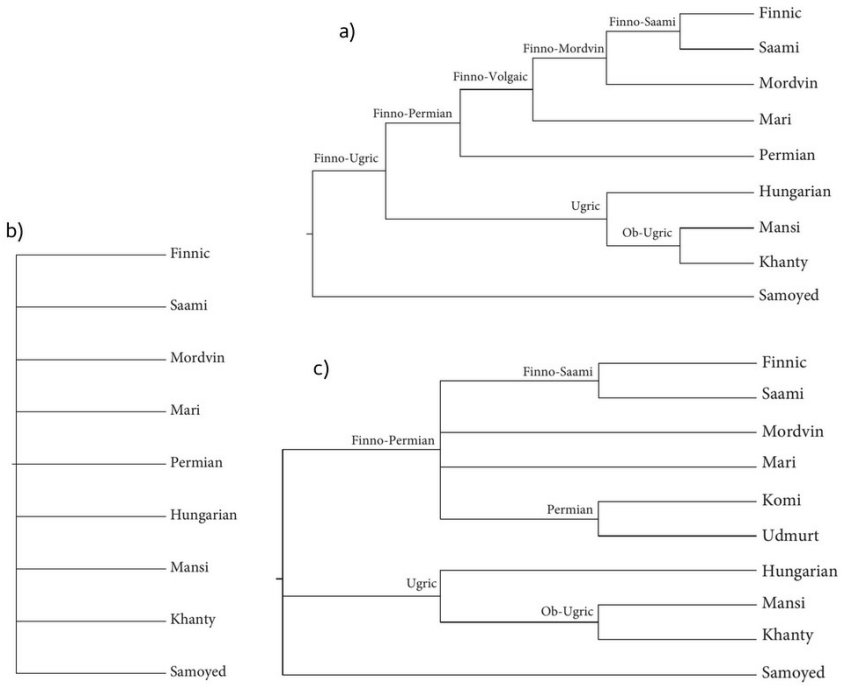
Permian branches; the latter then split into Permian and Finno-Volgaic, after which Mari and Mordvin (historically grouped together as the nowadays obsolete 'Volgaic' branch) diverged, followed by the splitting of the remaining languages (Finno-Saamic) into Finnic and Saami. By including the Samoyed branch alongside Finno-Ugric and connecting them through a common proto-stage (Proto-Uralic), we obtain the aforementioned 'standard paradigm' classification for the Uralic language family, a tree with strictly binary splits, with Samoyed as the first language to diverge. This kind of classification was commonplace until around the 1980s, and is reflected closely in, for instance, Korhonen (1981), shown in Figure 2 below; the only major difference between Korhonen's interpretation and the traditional 'standard paradigm' tree is is the lack of the Volgaic branch.

Although Uralic languages are generally accepted as forming a genealogical whole, the relationships within the Uralic language family have been debated since its inception; views of the Uralic language family are nowadays distinct from the 'standard paradigm' view. The nine 'shallow branches' of Uralic are generally considered valid and reconstructable as protolanguages (Aikio in press). To some extent, however, all of the 'deep branches' of the standard paradigm classification – the intermediate levels between the root and the tips of the tree – are nowadays considered either debatable or obsolete. The Volgaic branch was already absent from Korhonen's 1981 Uralic tree, Salminen (2002) presents an overview of the problems regarding the proposed innovations between the various deep branches. To briefly sum up some of his points: Finno-Ugric is predominantly supported by lexical evidence, which strongly distinguishes Samoyed from the remaining Uralic languages – however, there is little evidence for it on other linguistic levels; there are inconsistencies in the proposed Ugric and Ob-Ugric innovations, as they are either not consistently found amongst the Ugric members or are also found in Samoyed; Finno-Permian and Finno-Volgaic are both supported by only a few shared innovations, which may imply that they may not have been clearly distinguishable linguistic levels; Finno-Saami, quite a robust branch in its supposed number of shared innovations (see also Michalove 2002), also has potential problems due to the extensive borrowing history between Finnic and Saami. Other noteworthy sources that discuss the validity of the various traditional deep branches of Uralic include e.g. Michalove (2002), Helimski (2003) Tambovtsev (2004), Saarikivi & Grünthal (2005) and Ylikoski (2016).

The unresolved questions regarding the reconstructability of intermediate protolanguages in the Uralic language family have in turn introduced various

alternative branching hypotheses to replace the traditional, east-to-west spanning binary Uralic tree. These range from highly polytomous or "bushlike" classifications (e.g. Häkkinen 1984, Salminen 1999; 2002) to partially polytomous ones (e.g. Kulonen 2002, Michalove 2002). Figure 2 gives two examples of alternative Uralic classifications alongside a tree that closely reflects the 'standard paradigm'.

Figure 2. Branching hypotheses for the Uralic family, from Article 1: (a) Traditional tree based on Korhonen (1981); (b) highly polytomous classification, based on Salminen (2007); (c) partially polytomous classification, based on Kulonen (2002).



The analyses in Articles I and II, and to some extent also Article IV, shed quantitative light on the nature of the Uralic phylogeny, addressing the debated positions from the perspective of historical connections recorded for lexical data. They address the overall structure of the tree and the reliability of its various branches using different visualization techniques and metrics. The metrics on the

one hand provide estimations on how robust the different parts of a tree classification are, providing a rough baseline for how reconstructible the various branchings might be. In addition, the network analyses, while using essentially the same data as the tree models, visualize positions of ambiguity within the Uralic phylogeny, highlighting the effects of e.g. contact influence and ambiguous historical connections on the classification. Certain characteristics of the Uralic language family are also addressed as part of Article IV, such as how well the Uralic data fits into a tree, and what the largest version of the data looks like when viewed through the lens of a NeighborNet analysis (see section 3.1.1 for more detail on the Uralic data). Notably though, the main focus of Article IV is methodological exploration of TIGER values as a metric for treelikeness, and a bulk of the analyses are done using simulated data rather than Uralic data.

It is worthwhile to point out certain details from the terminology used in the articles when discussing the Uralic language family. Firstly, the articles make occasional references to both Volgaic and Finno-Volgaic branches when discussing the results; the use of these historical branch names should not be interpreted as any type of strong claim that they are historically attested, but rather as an attempt to make the text more economical. In retrospect, this point could have been made more clear in the articles, especially Article I. Secondly, the articles use 'Finno-Mordvin' when referring to the combination of Mordvin, Saami and Finnic branches; this grouping is also called 'Western Uralic' in the literature (see e.g. Häkkinen 2009, Ylikoski 2016).

2.3.2 Finnish dialects

Article III explores language evolution at the micro level, focusing on the dialects of Finnish. This chapter provides a dialectological overview of Finnish. This information can also be found in the article itself.

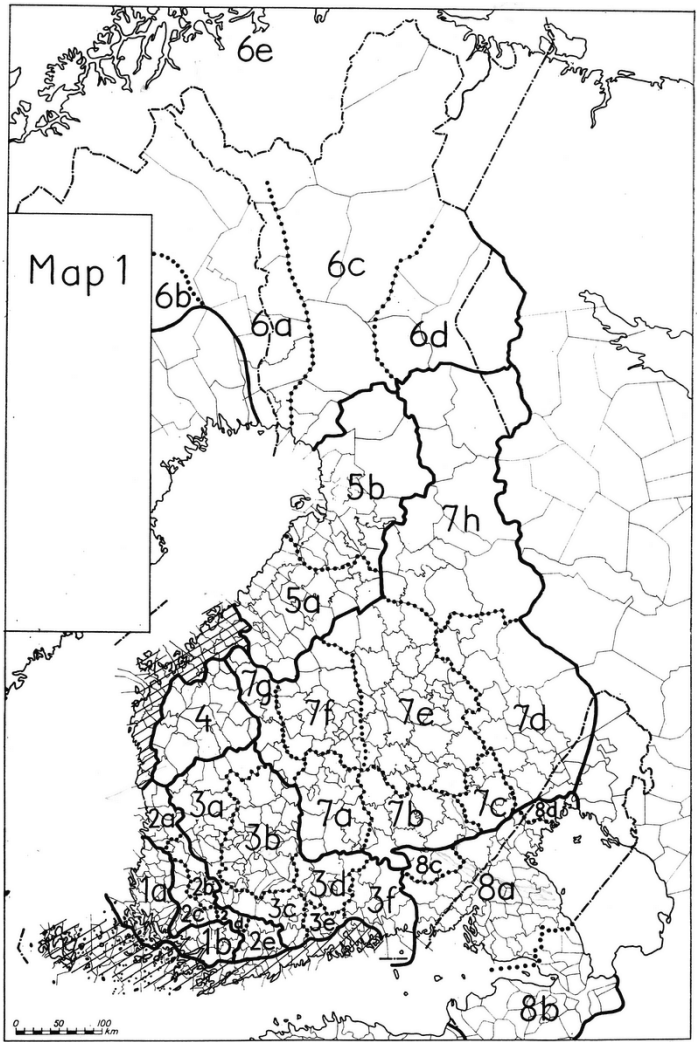
While subjective interpretations of Finnish dialects date all the way back to early written Finnish, as is apparent from Mikael Agricola's foreword for the New Testament (Agricola 1548), systematic dialect research of Finnish began around the nineteenth century. This research was motivated partially by growing interest in national history as well as fieldwork focusing on collecting oral tradition (Hovdhaugen *et al.* 2000). Dialectology remained one of the most actively researched subjects of Finnish linguistics until the mid-twentieth century, when

variationist studies turned their focus to sociolinguistics (Hurtta 1999). Finnish dialect research as a whole has mainly focused on fieldwork and detailed descriptions of specific dialects and dialect areas. Some of the more noteworthy works focusing on Finnish dialect variation as a whole include Kettunen (1930, 1940a, 1940b), Hakulinen (1950), Rapola (1969), and Hormia (1978).

A fairly good consensus exists on the division of Finnish dialects. Most commonly the language is split into two principal dialect areas, eastern and western. This dichotomy was described already in the 18th century by Vhaël (1733), becoming the default division in the early 19th century (Rapola, 1969; Wiik, 2004). It is considered to be the clearest general division of Finnish dialects, and serves as the foundation for various subdivisions, especially ones that emphasize morphological and phonological features.

The eastern and western dialects are generally split further into seven or eight generally clear main dialects (e.g., Itkonen, 1964; Savijärvi and Yli-Luukko, 1994). Slight variation on these areas can be found in e.g. Mielikäinen (1991) and Leskinen (1992). Itkonen (1964; 1989) is often regarded as the 'gold standard' of the eight-way division, splitting the western dialect area into Southwest, Southwest transitional, Häme (Tavastia), South Ostrobothnia, Middle/North Ostrobothnia, and Far North, and the eastern dialect area into Savo (Savonia) and Southeast (Figure 3).

Figure 3. Finnish eight-way dialect division from Itkonen (1964). This divides Finnish into Southwest (1a–b), Southwest transitional (2a–e), Häme (3a–f), South Ostrobothnia (4), Middle/North Ostrobothnia (5a–b), Far North (6a–e), Savo (7a–h), and Southeast (8a–d). 1–6 belong to the western dialect area, and areas 7–8 to the eastern dialect area.



While the east-west division remains the default for Finnish, three-way divisions have also been proposed. For instance, Erik Lencqvist's proposal from 1777 suggested a division into 1) the Turku dialect, covering parts of the Southwest and Southwest transitional dialects, 2) the Ostrobothnian dialect, which also included Häme, and 3) the Savo dialect (Rapola 1969). Mielikäinen (1991) and Paunonen

(1991; 2006) have suggested that synchronic typological features, among others, could be seen as support for the kind of three-way division proposed by Lencqvist, where Southwest becomes a main dialect area. Another proposed three-way division, originally proposed by Warelius (1848) and later discussed by e.g. Leino *et al.* (2006) and Hyvönen *et al.* (2007), splits Finnish into eastern, western and northern areas, with the northern area being essentially a mixture of eastern and western influence. The east-west-north trichotomy has been suggested to be more prominent at the lexical level, whereas the two-way division (east-west) is more prominent at the morphological and phonological levels.

There are also some grounds for suggesting four principal dialect areas in Finnish dialectological literature, as Paunonen (2006) suggests that, from a synchronic standpoint, Finnish should be divided into 1) Southwest dialects, 2) Western dialects (covering Southwest transitional dialects, Häme dialects, and South Ostrobothnian dialects), 3) Eastern dialects (covering Savo and Southeast dialects), and 4) Northern dialects, covering Middle/North Ostrobothnia and Far North.

Article III focused on testing population genetic clustering as a new quantitative solution for inferring dialect areas. In addition to exploring the aforementioned dialect divisions in the light of quantitative clustering, it also explored Finnish dialects using a more generic clustering technique called K-medoids clustering as well as tested various population genetic metrics with dialect atlas data.

3 MATERIALS AND METHODS

This section provides an general overview of the materials and the methods used in the four articles of this dissertation. The same information can also be found in the articles themselves, with the articles occasionally providing more detail than the descriptions here.

3.1 Materials

In this section I outline the two data sets that were analysed as part of this dissertation. The first of these is a basic vocabulary dataset of Uralic languages, which represents macro-level relationships between related languages. This dataset is the primary focus in articles I and II, whereas in article IV it serves as a real-life point of comparison for simulated datasets that have similar superficial properties. The second dataset is a digitized version of the Dialect Atlas of Finnish (Kettunen 1940a), and serves as the primary data in article III, where the focus was on micro-level relationships – that is, relationships that exist within a single language.

3.1.1 UraLex basic vocabulary dataset (I, II, IV)

Articles I, II and IV, which take a predominantly macro-evolutionary perspective and apply phylogenetic tools such as MrBayes (Huelsenbeck & Ronquist 2001), NeighborNet (Bryant & Moulton 2004) and the TIGER algorithm (Cummins & McInerney 2011), employ different versions of a dataset collected and expanded as part of the BEDLAN project and its follow-up projects over several years, and subsequently released under a Creative Commons license as the *UraLex Basic Vocabulary Dataset v. 1.0* (Syrjänen *et al.* 2018), or *UraLex* for short. The data records contemporary words and their historical relationships for meanings that mainly belong to so-called *basic vocabulary*. In this context basic vocabulary refers to a subset of the vocabulary usually denoted by semantically and morphologically

simple words that are relatively stable over time, culturally neutral (and consequently, likely to be found universally across different languages) and resistant to replacement by borrowing or semantic shift. The notion of basic vocabulary has a long history in linguistic research, tracing back to at least the seventeenth century (Hymes 1983: 65). Basic vocabulary is often associated especially with the lexicostatistical and glottochronological research tradition, started by Morris Swadesh in the 1950s. Basic vocabulary has been a staple of that field since its inception, and its use has continued to more contemporary methods, as well.

As Campbell notes, basic vocabulary is often intuitively understood to contain words for concepts such as "body parts, close kin, frequently encountered aspects of the natural world, and low numbers" (Campbell 2003: 263–264). A number of standardized basic vocabulary lists have been proposed over the years. Perhaps the most popular and well-known of these lists is the 200-meaning basic vocabulary list proposed by Morris Swadesh (Swadesh 1952), often noted to be the first formally specified basic vocabulary list, along with its revised 100-meaning version (Swadesh 1955), which eliminated various meanings deemed to be culture-specific, meanings without reliable matches in all languages, ambiguous object words, and non-independent meanings which may cause duplication. In addition to these two 'Swadesh lists', other basic vocabulary lists also exist, such as Tadmor's (2009) Leipzig-Jakarta list, which, unlike Swadesh's lists, is based on quantitatively evaluated criteria (susceptibility to borrowing, historical age in the language family, morphological simplicity and representativeness of the meaning in the vocabularies of different languages) assessed from languages around the world. Many variations of the standardized basic vocabulary lists exist; these generally optimize the contents for a specific set of languages. Examples of such modified basic vocabulary lists, given in McMahon & McMahon (2005: 43–44; 156–157), include the CALMSEA list (Matisoff 2000), the Australian language lists by O'Grady (1960) and Alpher & Nash (1999), and Heggarty's CALMA list. There also exists a modified Bantu basic vocabulary list (Bastin *et al.* 1999), and a proposed list of fifteen 'ultrastable' meanings, proposed originally by Dolgopolsky in 1964 (Trask 2000). Notably, Article I also used an optimized Uralic-specific basic vocabulary list ('Ura100') in addition to standardized basic vocabulary lists, as well as lists with more borrowing-susceptible items. The non-standard lists in Article 1 were based on supporting information on borrowing-susceptibility recorded in the dataset.

Perhaps a bit unintuitively, the primary data recorded in a basic vocabulary dataset such as UraLex does not consist of lexical data per se, but rather the historical relationships of lexical items. *Cognate relationships* specify whether a word denoting a given meaning contains a semantic root morpheme that has been vertically inherited from an ancestor language without considerable semantic shift[3]. Most basic vocabulary data sets in existence focus mainly on these type of relationships. There may also be words whose root morphemes have a shared historical origin through borrowing, or a combination of borrowing and vertical inheritance; when encoded as part of lexical datasets, these types of connection have been termed *historical connections* (Kessler 2001) or *correlates* (McMahon *et al.* 2005). UraLex includes both types of relationships, motivated by the desire to explore both horizontal and vertical relationships within the Uralic language family using phylogenetic tools. Both cognates and correlates, when appropriately represented, can serve as input for phylogenetic analysis tools.

The UraLex dataset was collected mainly by Jyri Lehtinen as part of the BEDLAN research initiative, while initial release version was edited as part of this dissertation project. An expanded and revised version of the dataset is already in the works. The data has undergone gradual expansion and correction process over the years, and this is reflected in the articles, which use different versions of the basic vocabulary dataset. Dictionaries (common and etymological) served as the primary sources for the dataset; a list of the sources can be found in the released dataset itself (Syrjänen *et al.* 2018). For each language investigated, contextually and semantically neutral words were collected for each form-meaning correspondence. If no single word could be found that alone represented the given meaning for a given language, a selection of near-synonymous words were collected to cover the desired meaning. Notably, this is against the strictest conventions used in collecting similar basic vocabulary lists – i.e. that each meaning is represented by one and only one lexeme per language. In practice, this does not pose a problem when this type of data is analysed in a binary representation; each cognate set can be represented as a separate column of binary values. With multistate analyses it may be necessary to choose one representative cognate set for each meaning; this was the case with the TIGER analyses of Article IV, where each meaning was represented by a single multistate character.

[3] As noted in List (2016), 'cognacy' and terms related to it are ambiguous, as they can also refer to shared ancestry between words that do not necessarily denote the same meaning. In the context of this work, these terms denote shared ancestry between words that also have the same meaning. This is also the standard meaning of the term with datasets similar to *UraLex*.

3.1.1.1 Article I version of the dataset

Article I uses effectively the first version of the basic vocabulary dataset. It covers 226 meanings from seventeen Uralic languages, chosen to provide a reasonably good overall coverage of the traditional Uralic sub-groupings. The languages included in this version are: Finnish, Karelian, Veps, Estonian, Livonian, North Saami, Ume Saami, Skolt Saami, Erzya, Meadow Mari, Komi-Zyrian, Udmurt, Hungarian, Northern (Sosva) Mansi, Eastern (Vahk-Vasyugan) Khanty, Tundra Nenets and Selkup. The 226 meanings in the data are essentially a combination of three standardized basic vocabulary lists – the 100-item Swadesh list (Swadesh 1952), the 200-item Swadesh list (Swadesh 1955) and the Leipzig-Jakarta list (Tadmor *et al.* 2009).

The historical relationships encoded within this version are essentially 'historical connections' (Kessler 2001) or 'correlates' (McMahon *et al.* 2005) rather than cognate relationships. That is to say they represent a mixture of vertical inheritance from a common language and transmission through borrowing within the language family.

When collecting the data, borrowing information was also recorded from the references (a list of which is provided in the supporting information of Article 1). This information was used to produce six more sublists of the meanings, based on the number of recorded borrowing events within each meaning. The first of these sublists, Ura100, includes all the meanings without recorded borrowing events – as such, it is essentially a Uralic-specific cognate list, based on the references used. The five remaining sublists consist of borrowing-susceptible meanings at different levels; these are detailed in Table 2, which outlines all the sublists specified as part of this version of the cognate corpus. All of these sublists were analysed separately as part of Article I. In addition, Article I included analyses of the Leipzig-Jakarta sublist, a Swadesh100 sublist, a Swadesh207 sublist and the full basic vocabulary dataset.

Table 2. Sublists of the basic vocabulary dataset used in Article I

Set	Meanings	Includes
Full	226	All meanings in the data set
Swadesh207	207	Meanings from the 200-word Swadesh list (Swadesh 1952) + 7 meanings from the revised Swadesh list (Swadesh 1955)
Swadesh100	100	Meanings from the revised Swadesh list (Swadesh 1955)
Leipzig-Jakarta	101	100 meanings with the highest rank from the Leipzig-Jakarta list of basic vocabulary (Tadmor 2009). The list includes 101 meanings as the item 'foot/leg' in the original list was split into 'foot' and 'leg', as in Swadesh207.
Ura100	100	Meanings with no attested borrowings according to the references employed
1+ borrowings	124	Meanings with 1 or more borrowings
2+ borrowings	69	Meanings with 2 or more borrowings
3+ borrowings	47	Meanings with 3 or more borrowings
4+ borrowings	32	Meanings with 4 or more borrowings
5+ borrowings	22	Meanings with 5 or more borrowings

Article I also employs a reconstructed Proto-Uralic included in the dataset, which is used experimentally as an *outgroup,* i.e. a reference group that allows phylogenetic tools to root the phylogeny and thus determine a starting point for the tree. This reconstruction includes only proto-forms whose original meanings most likely corresponded with the meanings of their counterparts in the daughter languages. For instance, if a Proto-Uralic reconstruction with the meaning "head" is given, and most of the daughter languages contain a counterpart for this word with the same meaning, the reconstruction is accepted into the data with this meaning. If the meaning given for a reconstruction is ambiguous (for instance, it denotes "fog", "smoke", and "steam") because of diverse meanings of its counterparts in the daughter languages, the word would not be accepted as a proto-form denoting one of these meanings, such as "fog". While it is common practice in Uralic linguistics to accept a Proto-Uralic reconstruction if its representatives can be found in Finno-Ugric and Samoyed, the two traditional main branches, the validity of this division has been questioned in recent work, as is mentioned in the text. In order to maintain the possibility that the initial branching did not necessarily occur between these language groups, only those forms were accepted as Proto-Uralic forms that had likely counterparts in three traditional Uralic subgroups: Ugric, Finno-Permian and Samoyed.

This version of the dataset is also documented in the supplementary information of Article I; parts of the description above are based on this documentation.

3.1.1.2 Article II version of the dataset

The material used in Article II is an edited and expanded version of the basic vocabulary dataset used in Article I. This covers 313 meanings instead of the previous 226, and eighteen Uralic languages instead of the previous seventeen. The additional meanings covered the so-called 'WOLD401-500' list of less basic vocabulary, also introduced in Article II. This list consists of meanings that rank between 401–500 according to the composite score that was used for creating the Leipzig-Jakarta list of basic vocabulary (Tadmor 2009). In practice this expands the dataset's representation of shallow and horizontal historical connections by adding borrowing-susceptible and less stable material. The expansion was motivated by the central research focus of Article II, which focused on exploring non-treelike connections within Uralic languages with the help of phylogenetic network analyses. In addition to the added meanings, the data used in Article II also covered an additional Saami language, Kildin Saami, to slightly improve the representation of Saami languages, which only covered three languages in the previous version.

3.1.1.3 Article IV version of the dataset

Alongside simulated language data, Article IV uses real-life data from the released version of the basic vocabulary data, *UraLex v. 1.0* (Syrjänen *et al.* 2018). This revised version, like the Article II version, includes 313 meanings, but adds eight languages, bringing the total number from 18 to 26. The eight additional languages included Nganasan, Võro (South Estonian), Inari Saami, Pite Saami, South Saami, Komi-Permyak, Ingrian and (Western) Votic. While *UraLex* includes both correlate relationships and cognate relationships, the analyses of Article IV only used the cognate relationships.

3.1.1.4 Data formatting used for basic vocabulary data

The format of the historical connections in the lexical data is such that each cognate or correlate set within each meaning is represented by a unique multistate character that reflects their historical relationships. For instance, the word denoting *moon* in Finnish in the dataset ('kuu') is historically connected to its counterparts in e.g. Estonian ('kuu') and Erzya ('kov'); the historical connection between these is marked with a shared character state (*a*). The corresponding words in Skolt Saami ('mään') and Kildin Saami ('mānn') are historically unrelated to the first set of words but are related to one another, so they are coded with a different shared character state (*b*). Their equivalent in Udmurt ('toleź') is not related to either of the aforementioned groups, which is marked with a third character state (*c*). The number of states varies from meaning to meaning, with certain meanings being such that they all share a root morpheme (e.g. *name*), while others being much more diverse (e.g. *to say*).

Phylogenetic tools are not directly designed to handle the type of multistate data found in basic vocabulary dataset. They are primarily intended for biological datasets, which record nucleotide or amino acid data, which have a prespecified number of character states for every character and are represented by specific set of character states. For instance, DNA nucleotides are generally represented by four characters – A, C, G and T – and tools that expect this type of data generally ignore any other character. An often-used workaround for this is to convert the multistate cognate data into a binary presence-absence pattern matrix, which these tools can readily analyse. In this format, each column corresponds to a single character state of a multistate character, and for each language this column gets one of three values: 1 (= present in the current language), 0 (= absent in the current language) or ? (= unknown). This binary representation approach is used in Articles I and II, where each historical connection is converted into a binary representation.

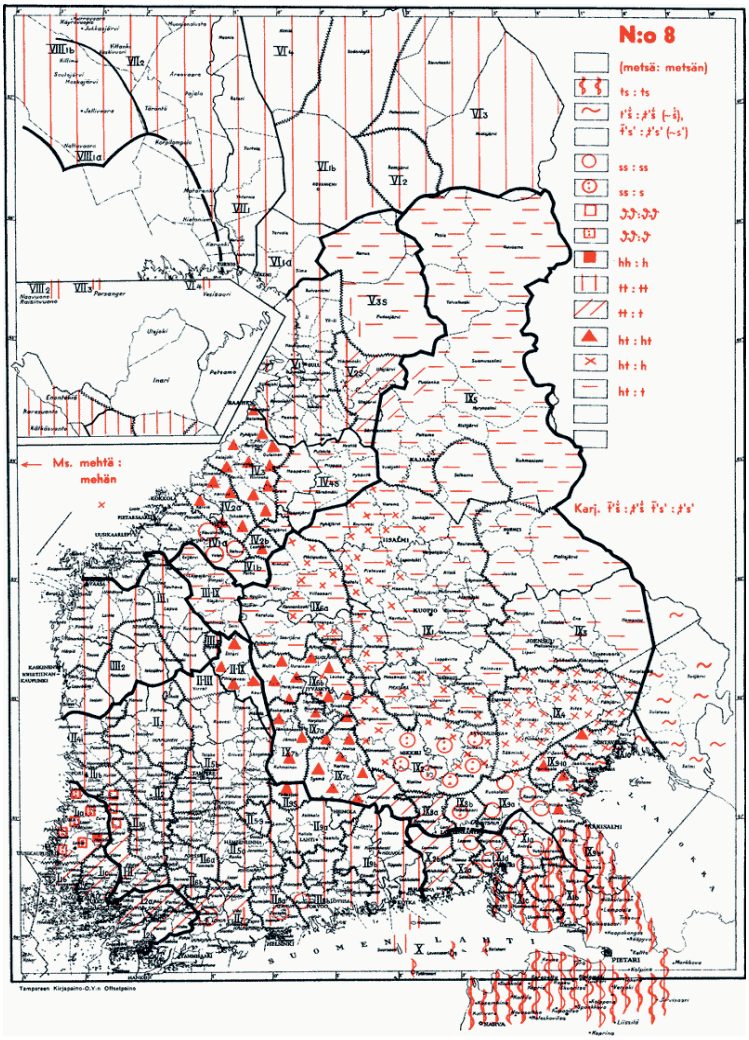
For Article IV the basic vocabulary data needed to be represented in a multistate form – one multistate character per meaning - in order to calculate TIGER values. As pointed out above, in many cases UraLex has multiple lexical reflexes for a meaning within a single language, which are all tied to a different cognate sets. These are generally found in cases where it has been difficult to define a single reflex due to e.g. use context limitations or lack of information related to frequency of usage. Article IV thus used a computationally implemented

"minimizing" strategy to choose a representative cognate set for each meaning from the dataset. The strategy essentially aimed to keep the total number of cognate sets for each meaning as low as possible. In practice this approach favours chronologically deep relationships (old historical connections between many languages) over shallow relationships (younger connections between few languages).

3.1.2 The dialect atlas of Finnish (III)

Article III differs from the rest of the dissertation in that it focuses on modeling dialects, rather than languages, from an evolutionary perspective. Its primary tool of analysis is Structure, a software designed for inferring within-species populations using genetic data (Pritchard *et al.* 2000). In Article III Structure is used to infer linguistic populations using dialect data that records intralingual variation. This data comes from a digitized version of the *Dialect Atlas of Finnish*, originally compiled by Lauri Kettunen between 1920 and 1930 (Kettunen, 1940a). It consists of 213 map pages (see Figure 4 for an example), each of which documents a specific morphological, phonological or lexical dialect feature, showing the variants that exist of the feature, and the geographical distribution of each variant across 525 Finnish-speaking municipalities, which serve as the data points. The number of variants for the dialect features range between 2 and 15, and a municipality may contain multiple overlapping variants; at most, the atlas contains 4 overlapping variants for the same municipality. The atlas covers the entire region of Finland, except for the exclusively Swedish-speaking areas on the western and southern coasts, as well as the northernmost part of Finland, which was predominantly Saami-speaking. It also includes Finnish-speaking areas in Ingria (Russia), Norway, and Sweden, as well as Karelian-speaking areas in pre-WWII Finland. The dialect atlas does not record responses from each interviewed informant individually; the data points represent the combined information from all the informants from each municipality.

Figure 4. Example page from the *Dialect Atlas of Finnish* (Kettunen 1940a), showing dialect differences in the pronunciation of the word *metsä* ('forest'), specifically how */ts/* is realized in the nominative and genitive forms. The legend on the top right lists the recorded variants, which correspond with specific symbols on the map itself.



Work on a computerized version of the dialect atlas (see Kettunen *et al.* 2021 for the most recent version) began already in 1997 (Embleton and Wheeler 1997; 2000), and was made available as part of the Finnish Dialect Atlas Project, funded by the Social Sciences and Humanities Research Council of Canada in co-operation with the Institute for the Languages of Finland (Kotus). An additional round of error-checking for this digital version was carried out by Jyri Lehtinen, as part of

the BEDLAN project; these corrections are incorporated in the released version of the data, available through the Etsin service (etsin.fairdata.fi).

3.1.2.1 Data formatting used for dialect atlas data

From a conceptual standpoint dialect data is quite similar to population genetic data; similarly to how populations within a species are defined through genetic variation – i.e. a specific set of *alleles* (variants of a gene), dialects can be regarded as 'linguistic populations' defined through linguistic variation – i.e. a specific set of linguistic variants. This conceptual similarity is discussed in more detail in Article III. With the aforementioned similarity in mind, the dialect data was converted to a format that reflects that of biological data from population genetics. Each data point (municipality) was likened to an individual, each dialect feature to a *locus* (position of a gene) and each variant of a dialect feature to an *allele* (variant of a gene). Using this approach, the dataset was converted to a representation that parallels that of *diploid* organisms (i.e. organisms with two alleles per each locus); this representation can code at most two overlapping linguistic variants of each feature per municipality, and accounts for around 99.9 per cent of the variation found in the dialect atlas. The data was also converted to a representation resembling that of *haploid* data, used to represent organisms with one allele per locus. This representation cannot account for cases where a municipality is characterized by more than one linguistic variant of a given dialect feature, and can thus represent less of the overall variation (around 94.3 per cent). Of these two the diploid data type became the main version for the Structure analyses due to its better coverage.

The data was also analysed using a more generic clustering approach (K-medoids clustering), with the difference that missing (empty) characters and absent data points were represented identically as zeroes, unlike with Structure, which retained the difference between missing and absent characters.

3.1.3 Simulated language data (IV)

The main focus of Article IV was to explore the values calculated by the TIGER algorithm (Cummins & McInerney 2011) as a metric for treelikeness. The majority of this exploration was done using simulated language data. The advantage of this

approach was that firstly, the extent to which the data was treelike was known and controllable. Secondly, with this approach it would be possible to not only to introduce a degree of nontreelike signal into a treelike dataset, but also construct datasets with a completely nontreelike structure. Superficially, the simulated datasets mimic the material found in the UraLex cognate corpus – cognate sets of meanings, represented as multistate characters. The models used to generate the simulated language data are also set up so that the overall distribution of cognate set counts across the simulated datasets resemble the cognate set distribution of the real-life reference dataset, UraLex.

Four types of data were generated for Article IV. The first datatype was purely treelike data, produced by generating a tree with the desired number of languages, the root node of the tree was given a cognate set; along each branch of the tree, cognate replacement events were drawn from a Poisson distribution, based on a change rate value drawn from a Gamma distribution, branch length and an overall cognate birth rate. This process is repeated for each simulated meaning. This model essentially produces a 'Dollo' style dataset, where each linguistic feature emerges only once and never resurfaces after being lost.

The second datatype mimicked treelike data with borrowings. The generation process goes as follows: first, a treelike dataset is generated as described above. After this, a borrowing susceptibility is drawn for each meaning from a Gamma distribution. Each node of the tree - representing the intermediary proto-stages of the simulated language family - is revisited, and at each of them a Bernoulli random trial is performed for each meaning to determine whether a borrowing event takes place; the probability of success for the trial is based on the randomly drawn borrowing susceptibility, an overall borrowing rate, and branch length. In the event of a success, a borrowing source node existing within the same time frame as the borrower is chosen, and the cognate class of that borrowing source node overwrites the cognate class of the borrower node. Following this, the subtree starting from the borrower node is 're-evolved' in the same way as when generating purely vertical data, essentially simulating vertical evolution following each borrowing event. Article IV included four simulated datasets generated with four different borrowing rate parameters: 0.05, 0.10, 0.15 and 0.20; these correspond to an expected 5%, 10%, 15% and 20% of datapoints being borrowed.

The third datatype, loosely resembling data that might have emerged from a dialect chain or a Sprachbund, involves simulating innovation and diffusion of

individual features within an ordered list, which represents the spatial ordering of the languages. A more detailed description of the process is provided in the supporting material of Article IV. Briefly described, this simulation generates the data using two kinds of operation. The first of these is *concatenation*, which conceptually mimics a situation where variants of a feature have evolved simultaneously and independently in neighbouring regions, each diffusing to their respective neighbouring languages. The second operation, *insertion*, mimics a situation where a geographically less widespread feature replaces a geographically more widespread feature in a small part of the overall geographical area, effectively causing a gap into the geographical continuum of a feature. These two processes combine to produces a dataset that is of a very different nature than treelike data, with a structure based on the geographical proximity of languages rather than descent with modification.

The fourth datatype, affectionately called 'swamp' data, is essentially unstructured data that superficially resembles aligned cognate data but includes no inherent structure, other than what is produced by pure chance. However, despite its generally unstructured nature, this data type is also set up so that the overall cognate class distribution resembles that of UraLex.

The code for the generative models was written in collaboration with Luke Maurits (one of the co-authors of Article IV), and is available as part of its supplementary material.

3.2 Analysis tools

In this section I will go over the main analysis tools employed in the articles and discuss the specific tools used in the four articles that are part of this dissertation. These, too, can be roughly divided into *macro-evolutionary* tools, which are adopted from phylogenetics and whose main purpose is to analyse relationships between related taxa (or in this case, languages), and *micro-evolutionary* tools, which originate from population genetics and are used for exploring within-species relationships (or in this case, intralingual relationships). Some general-purpose analysis tools are also part of the overall toolset, including R and Python, which are used for additional statistical analyses, data preparation and ad hoc data formatting.

3.2.1 Macro-evolutionary tools (I, II, IV)

3.2.1.1 MrBayes

The main analysis program employed in Articles I and II is MrBayes v.3.2 (Huelsenbeck & Ronquist 2001, Ronquist & Huelsenbeck 2003, Ronquist *et al.* 2012), a tool which uses MCMC methods for Bayesian inference of phylogenies. A general overview of its operation is given in Article I as well as other sources, such as Hall (2011) An overview of Bayesian phylogenetic inference techniques in a linguistic context can also be found from e.g. Dunn (2015) and Greenhill *et al.* (2020).

When using cognate character data, tools such as MrBayes essentially model the existence of words within a specific semantic slot that are historically connected across languages through their root morphemes. As time goes on, words used to denote a specific meaning are affected by processes such as borrowing and semantic shift and eventually get replaced by new words that are historically unrelated to the previous word. Cognate relationships, or known historical relationships between words that denote a specific meaning, can be used to construct a phylogenetic tree or network that aims to visualize and quantify the relationships between the different languages in the dataset. Bayesian inference does this by essentially modeling the replacement events of the cognate sets in the dataset, constructing a tree based on them. The analysis can use different types of substitution models, which establish rules for how the substitutions take place. The analysis can also incorporate prior knowledge to the analysis, such as set constraints on the topology of the phylogenetic tree to discard improbable solutions from the results.

In summary, Bayesian inference with MrBayes, after the researcher has set up the analysis by inputting the data (which consists of aligned sequence data in binary or multistate character format) and setting up the evolutionary model (see e.g. Dunn 2015) and potential priors, operates as follows:

1. the evolutionary model is initialized with a random tree and random values for any unknown parameters,

2. the starting tree's likelihood is calculated,

3. one of the unknown parameters of the model or the shape of the tree is slightly altered,

4. the likelihood of the tree is recalculated,

5. if the likelihood increases the new parameters are accepted as a new starting point. Otherwise the starting point remains unchanged.

Steps 2–5 make up one *generation* in MrBayes' terminology.

A single MrBayes analysis, or run, consists of several million generations, during which the model's parameters and the output tree undergo gradual improvement until they reach a stage where the tree and the parameters do not change significantly from one generation to the next. At this stage the run is said to have *converged* to the set of most likely trees and parameters given the data and the model; this is also called the *stationary distribution*. At specific intervals the results of the analysis (the tree and the unknown parameters) are sampled. However, as we are interested only in the most likely trees and parameters (i.e. those belonging to the stationary distribution), a certain number of initial samples are discarded from the results; this discarded set of samples is called the *burn-in*. The final result consists of the remaining sample, consisting of the distribution of most likely trees and unknown parameter values. A visualization of burn-in stage and the sampling stage are shown in e.g. Figure 7.2 in Dunn (2015) and Figure 11.2 in Greenhill *et al.* (2020).

Other programs using similar basic principle of operation include, for instance, BEAST (used in e.g. Honkola *et al.* 2013 to study Uralic languages) and PhyCAS.

The result of a phylogenetic analysis is a distribution of topologies and parameter values that give the best explanation for the data under the provided model. The resulting tree distribution can be visualized in various ways (see e.g. Dunn 2015; Greenhill *et al.* 2020); Articles I and II use *majority-rule consensus trees*, which are essentially trees that summarize the agreement between the entire set of trees from the results. These trees consist of all the clades present in at least 50 percent of the entire distribution of sampled trees. Each clade in the consensus tree is accompanied by a support value (posterior probability), which reflects the frequency of each clade in the sample. Figure 5a further below provides a visual example of a tree with support values. While there is no clear-cut rules on how the support values should be interpreted, the articles in this dissertation have generally

considered clades with a support value equal to or exceeding 0.95 as being very strongly supported (e.g. Huelsenbeck *et al.* 2001). Branch length, which in these results reflects the amount of change, is another useful metric alongside the posterior probabilities for estimating how well-supported a given clade is. Notably, this metric is only used informally in this study; for instance, in Article I we comment situations where a branch is considerably short compared to the other branches on the tree. Another point to make is that the meaning of the branch length depends on the type of tree.

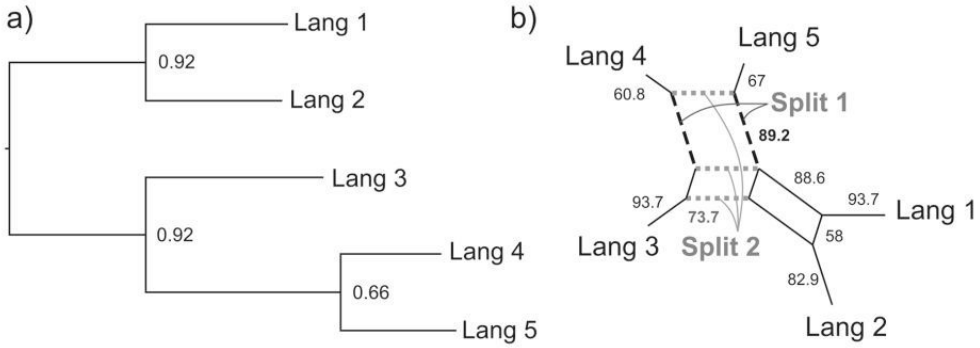
As their model, Articles I and II use a Markov K model, a simple model where the loss and gain of cognates is governed by a single rate parameter. The analyses are also done without setting explicit priors. The trees produced with MrBayes were visualized using a separate program called FigTree.

3.2.1.2 SplitsTree

SplitsTree (Huson 1998, Huson & Bryant 2006) is a phylogenetic software package which implements a number of published methods for constructing phylogenetic trees and networks. In this work it was used for creating phylogenetic networks for Articles II and IV. Phylogenetic networks, unlike phylogenetic trees, can explicitly visualize conflicting phylogenetic signals, which can be caused by complex divergence processes, such as horizontal gene transfer with biological species or borrowing with languages. SplitsTree includes several methods for constructing phylogenetic networks; the distance-based NeighborNet algorithm (Bryant & Moulton 2004) was used in Articles II and IV.

NeighborNet produces a graph consisting of *splits*, which separate the analysed taxa into various sets; two splits are shown in Figure 5 by dashed lines. In the figure split 1 separates languages 4 and 5 from languages 1, 2 and 3, while split 2 separates languages 3 and 4 from languages 1,2 and 5. The length, also called weight, of each split reflects how distant the sets are from one another; for instance, in the figure split 1 is longer than split 2, indicating that e.g. languages 4 and 5, which are separated from one another by the shorter split 2 – are closer to each other than e.g. languages 3 and 4 are to each other, as they are separated by the longer split 1. NeighborNets can also be accompanied by support values called *bootstrap value*, which range between 0 and 100, and like Bayesian posterior probabilities, specify how robustly supported each split in the split graph is.

Figure 5. Examples of a phylogenetic tree and a phylogenetic network, from Article II: (a) phylogenetic tree with support values; (b) phylogenetic network with bootstrap values.



NeighborNets can also serve as a visual aid for estimating how treelike a given data is (see Gray *et al.* 2010, Greenhill *et al.* 2010, Article IV). They can also be used in conjunction with numerical metrics of treelikeness, such as the δ score (Holland *et al.* 2002), and the Q-residual (Gray *et al.* 2010). These metrics provide values between 0 and 1, with values closer to 0 indicating more tree-like data, and values closer to 1 suggesting a less tree-like data.

3.2.1.3 δ score and Q-residual

The δ score (Holland *et al.* 2002), and the Q-residual (Gray *et al.* 2010) are currently among the most commonly used metrics for quantifying how treelike a phylogenetic dataset is. These metrics are also defined and discussed in e.g. Wichmann *et al.* (2011). They are especially useful in exploratory stages of phylogenetic work, as they provide information on how well a phylogenetic tree describes a dataset. Both techniques are based on *quartets* (groups of four taxa), and both estimate treelikeness based on deviations from the so-called *four-point condition*. Four taxa (or languages) (A, B, C and D), can be divided into two pairs in three different ways. Each of these divisions corresponds to a pair of distances:

1. (A,B), (C,D); (|AB| + |CD|)
2. (A,C), (B,D); (|AC| + |BD|)
3. (A,D), (B,C); (|AD| + |BC|)

The four-point condition is satisfied if the two largest of the three distances are identical. In other words, assuming *d1*, *d2* and *d3* are the three distances, ordered from longest to shortest, the four-point condition is satisfied if *d1* = *d2*. In this case the four taxa fit perfectly into a tree. Treelikeness of a specific taxon can be estimated by averaging deviations from the four-point condition across all the quartets that the taxon participates in, while the treelikeness of an entire dataset can be estimated by averaging deviations across all the quartets of that dataset.

The δ score estimates treelikeness using the formula *d1 - d2 / d1 - d3*; the score is 0 if *d1-d3* = 0. The Q-residual estimates treelikeness using the formula $(d1- d2)^2$, where *d1* and *d2* are distances normalized so that the average distance between the taxa is 1. Both methods provide a similar output, a value that falls between 0 and 1, measuring the amount of deviation from the four-point condition. Values closer to 0 suggest that the data fits a treelike structure, and the value increases with the existence of less treelike structure. Notably the metrics operate at different scales, with δ scores generally being much higher than Q-residuals.

Article II uses δ scores as part of its analyses, while both these metrics are used in Article IV, which explores TIGER values as an alternative technique for quantifying treelikeness. Article II uses SplitsTree to calculate δ scores, whereas δ scores and Q-residuals are calculated in Article IV using the Python package *phylogemetric* (Greenhill 2016).

3.2.1.4 TIGER

Tree-Independent Generation of Evolutionary Rates, or TIGER (Cummins & McInerney 2011), is a non-tree based algorithm designed for estimating similarities in the distribution of aligned phylogenetic data. The algorithm calculates stability estimates ('TIGER values'), originally intended for excluding rapidly changing (and thus uninformative) characters from phylogenetic data (Cummins & McInerney 2011). However, it was also adopted in biological studies as a computationally inexpensive technique for improving phylogenetic models by means of data

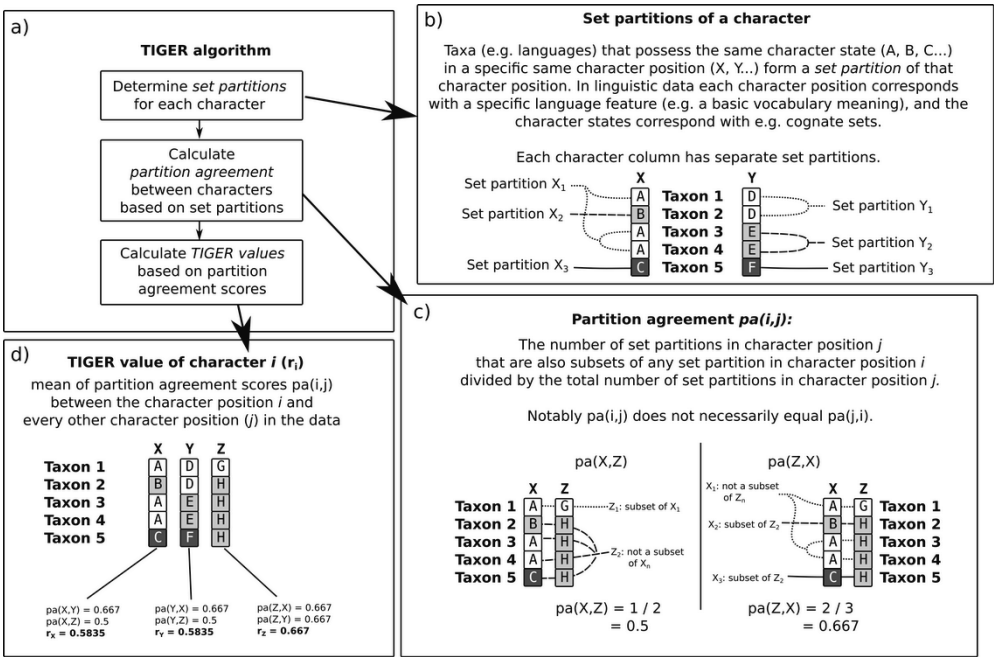
partitioning. Article IV of this dissertation explores TIGER values as a metric for treelikeness and heterogeneity for linguistic data, using both real-life data from UraLex, as well as simulated linguistic data.

The usefulness of estimating the treelikeness of a linguistic dataset stems from the fact that not all cultural data descends in a treelike pattern (e.g. Gray *et al.* 2010, Wichmann *et al.* 2011, Jacques & List 2019), which means that tree-based analyses do not necessarily make sense in all situations. In a worst-case scenario, a tree-based analysis of a non-treelike dataset can lead to serious misrepresentations of history. While solutions to addressing this problem exist e.g. in the form of similar metrics of treelikeness, such as the δ score and the Q-residual (see Holland *et al.* 2002, Gray *et al.* 2010, Wichmann *et al.* 2011), or the use of multiple tree-based analyses to identify conflicting patterns (Verkerk 2019 and to some extent Article I in this dissertation) and combining treelike and nontreelike representations (Nelson-Sathi *et al.* 2011), this toolkit remains relatively small and underexplored and consequently warrants more attention.

To infer TIGER rates from an aligned dataset, for each aligned character the set of all taxa is first partitioned into subsets by grouping together taxa with identical character states at that specific location of the alignment. After each position of the aligned data has been partitioned, 'partition agreements' are calculated for each character. Partition agreements for a specific character position are calculated by comparing its taxon set partition with the partitions of every other character position; each partition agreement score records how many of the sets in the compared character position's partition are subset of one of the sets in the partition of the character position whose TIGER value we are calculating. After each partition agreement has been recorded, a TIGER value of a character position is calculated as the arithmetic mean of the partition agreement scores between that character position and all the other character positions.

The resulting TIGER value is a number ranging between 0 and 1, with values closer to 1 indicative of more stable or consistent characters. Cummins & McInerney (2011) provides a detailed mathematical description of the algorithm; the general steps of the algorithm are summarized in Figure 6.

Figure 6. General steps of the TIGER algorithm, from Article IV.



For this work, a Python-based tool for calculating TIGER values was implemented on the basis of the mathematical description of the algorithm in Cummins & McInerney (2011). This version resolves calculation problems from the original TIGER implementation, is able to use arbitrary multistate data as its input, and also includes support for the linguistic CLDF format, used by the UraLex dataset, among other linguistic datasets. The tool is freely available in https://github.com/kasyrj/tiger-calculator.

3.2.2 Micro-evolutionary tools (III)

3.2.2.1 Structure and related tools

Structure (Pritchard *et al.* 2000) is a software for inferring population structure using genetic multilocus data, such as allele data, collected from individuals of the same species. Structure's models can produce both unmixed populations (where each individual belongs to one and only one population) and mixed populations,

although only the mixed population variant (*admixture model*) is used in Article III. In addition to selecting a proper model and setting up priors if necessary, the researcher must specify a K value for the analysis – that is, the number of populations for Structure to infer.

Like MrBayes, Structure is model-based and uses Bayesian inference and MCMC methods. The general operation of a Structure analysis can be summarized in similar steps as MrBayes:

1. The model is initialized using the chosen model settings and the K value; populations (or proportions of population memberships with the admixture model) are randomly assigned for each individual,
2. The overall likelihood of the solution is calculated,
3. One of the unknown parameters in the model (e.g. the assignment of individuals to different populations) is slightly altered,
4. Likelihood is recalculated,
5. If the likelihood increases, the new parameters are accepted as a new starting point. Otherwise the starting point remains unchanged.

Steps 2–5 make up one MCMC *repetition* in Structure's terminology (cf. MrBayes' *generation*). Generally each analysis is run several times in order to ensure that the results remain consistent. In Article III each of the Structure analyses was repeated 20 times.

When using the admixture model to produce mixed populations, Structure's output for every data point (individual) consists of a degree of membership in each inferred population (the *membership coefficient*). All of the membership coefficients of an individual sum up to (approximately) 1.0. The output also produces a log likelihood estimate, useful for determining how well different clustering solutions (runs with different K values) explain the data.

Two approaches were used for comparing runs with different K values – *mean log likelihood* and *deltaK* (Evanno *et al.* 2005). The mean log likelihood of a K value is calculated by first excluding potential outliers and then averaging the likelihoods of the remaining repetitions for that K value. This is done for all the tested K values.

The K with the lowest mean log likelihood is considered to be the best explanation for the data. The other metric, deltaK, is an extension of the mean log likelihood method, where we calculate how much difference there is between the mean log likelihoods of K and its neighbouring K values (K-1 and K+1). Thus, when the mean log likelihood of a K value differs significantly from the likelihoods of K-1 and K+1, deltaK is high. The place with the highest deltaK is likely to match the uppermost hierarchal level of a population structure.

It should be noted that Structure attempts to find populations that correspond to the 'Hardy-Weinberg equilibrium' (Pritchard *et al.* 2000), or HWE, as closely as possible. This is an idealized state where the allele frequencies of a population remain unchanged from one generation to the next, and is thus fairly unlikely for actual biological and linguistic populations, especially for extended periods of time. The implications of HWE are discussed in greater detail in Article III. Perhaps the most important point is that one should keep the HWE assumption in mind when making interpretations of Structure's results, and avoid interpretations that would not be unrealistic from the perspective of the HWE assumption, such as assuming an unrealistic time depth for the inferred populations.

In addition to HWE, Structure also assumes that the analysed loci are in 'linkage equilibrium' – meaning that the variables in the data are independent from each other. For Article III we analysed the dialect data in its entirety, but acknowledged that there are reasons to assume that some of the features may be more interlinked with each other, such as the various instances of consonant gradation. We experimented with an ad hoc method for detecting linkage, which operated as follows:

1. On map page A, we identify a set of *potentially linked* municipality pairs – that is, all municipality pairs which have identical dialect features (aside from being empty) on the page. We will call this set *Lp(A)*.

2. Similarly, we identify all potentially linked cases on another page B. We will call this set *Lp(B)*.

3. We discard those potentially linked municipality pairs from *Lp(A)* that contain no data on page B. We do the same for *Lp(B)*, removing those municipality pairs that contain no data on page A.

4. We identify actually linked cases between pages A and B as those municipality pairs that are potentially linked on both pages – in other words, $Lp(A) \cap Lp(B)$.

5. We calculate the total amount of linkage between two pages as actually linked cases divided by potentially linked cases on both pages, i.e. $Lp(A) \cap Lp(B) \ / \ Lp(A) \cup Lp(B)$; thus, we are essentially calculating the Jaccard index between $Lp(A)$ and $Lp(B)$. This may get values between 0 and 1, with 1 meaning that $Lp(A)$ is equal to $Lp(B)$ – in other words, every potentially linked data point is also actually linked.

The aforementioned steps were repeated for all unique map page pairs, and the results visualized as a heatmap (Fig. 13). In addition to Article III of this dissertation, the linkage test has also been applied in a quantitative study of Finnic languages using data from ALFE (Atlas Linguarum Fennicarum) (Honkola *et al.* 2019).

In addition to Structure, the analyses involved the use of additional tools designed specifically for Structure results. deltaK and mean log likelihoods were calculated using the R package *pophelper* (Francis 2014). In addition, a tool called *Structure Harvester* (Earl & vonHoldt 2001) was used to preprocess the results for another tool called CLUMPP (Jakobsson & Rosenberg 2007), which combines multiple analyses with the same K value into a single result. This can be useful for identifying *genuine multimodality* – i.e. possible situations where the different runs with the same K value might reveal qualitatively different clustering solutions. Finally, the geographic information system ArcGIS is used for producing map visualizations of the clusterings.

3.2.2.2 K-Medoids

K-Medoids (Kaufman & Rousseeuw 1987) is a distance-based general purpose clustering tool, used to assign data points into a prespecified number of unmixed non-hierarchal clusters. It is essentially an improved type of K-Means clustering, being less sensitive to outliers. It is used in Article III alongside Structure to provide a methodological point of comparison that is not purpose-built for population genetic data.

In general terms K-Medoids analysis proceeds as follows:

1. K data points are randomly selected as *medoids* (centers for the groups)

2. Each data point is assigned to the group whose medoid is closest to that point

3. Medoid points are re-evaluated: the total distance from each point in a group to all the other points in the same group is calculated, and compared to the total distance from the current medoid point to all the other points in the same group. The point with the lowest combined distance becomes the medoid point.

4. If the medoid point changed in step 3, steps 2 and 3 are repeated.

The result contains a cluster for each data point.

K-medoids clusterings using different K values can be compared using the *silhouette method* (Rousseeuw 1986). It examines the relationship between within-group dissimilarity (the average distance between data points within the same cluster) and between-similarities (a data points average distance to points in other clusters). The silhouette value compares within dissimilarity and the lowest between-dissimilarity, essentially describing how well a data point fits its current cluster compared to the neighbouring cluster. The silhouette value ranges between -1 (poorly classified point) and 1 (well-classified point). Across an entire dataset, we can examine the *average silhouette width*, which is the average of all the silhouette values for a clustering. These were calculated using the R package cluster (Maechler *et al.* 2014).

As with Structure, ArcGIS is used for producing map visualizations of the clusterings.

4 RESULTS AND DISCUSSION

This section goes through the main results of each article, beginning with the results of the phylogenetic analyses of Uralic languages in Articles I and II, followed by an overview of the population genetic clustering analyses of Finnish dialects as well as a discussion of the other of population genetic techniques applied in Article III, and finally, the usability of TIGER values as a linguistic metric for treelikeness and heterogeneity, investigated in Article IV. This section provides a general overview; further detail on the results can be found in the articles themselves.

4.1 Phylogenetic tools and Uralic languages (I, II)

Articles I and II focused on exploring the Uralic language family through phylogenetic methodology. Article I examined the Uralic phylogeny through multiple tree-based analyses with MrBeast using qualitatively and quantitatively distinct subsets of the lexical dataset, while retaining the model of sequence evolution the same in each analysis. This provided more insight into the details of the tree – that is, which of the branches were generally robust to changes in the analysed data and which ones were not. The analysed data consisted of 226 meanings and 17 languages.

Article II took a different viewpoint to the Uralic language family by adding phylogenetic network analyses (NeighborNet graphs) alongside phylogenetic tree (MrBayes) analyses, and by expanding the dataset to borrowing-prone lexical items that are outside of standardized basic vocabulary meanings; this increased the number of examined meanings to 313. Its dataset also included an additional language compared to Article I, making the total number of analysed languages 18. Article II also explored various subsets of the data, albeit to a lesser degree than Article I, but did so using both phylogenetic trees and phylogenetic networks. Through the network analyses it shed additional light on the conflicting patterns within the Uralic phylogeny.

I will begin by providing a general overview of the Uralic family based on the three articles. This will be followed by discussion on the usefulness of network methods as opposed to tree methods, based on Article II.

4.1.1 Overview of the Uralic language family based on phylogenetic analyses

Of the three articles that explored the Uralic phylogeny, Article I was the only one that made an attempt at determining the root position of the Uralic clade. The motivation to do so was to explore whether phylogenetic tools would give support to alternative hypotheses (see e.g. Häkkinen 1983; Salminen 1999, 2007; Saarikivi 2011), which do not follow the 'standard paradigm' approach of placing the root between Finno-Ugric and Samoyed. The root position test in Article I was based on an experimental technique of including a Proto-Uralic reconstruction as one of the analysed languages and using it as an outgroup. This reconstruction consisted of cognate sets for proposed proto-forms whose lexical reflexes were found from at least one Samoyed language, one Ugric language and one Finno-Permian language. This differs from the traditional criteria used for Proto-Uralic in Uralistics, where a match is required only from a Samoyed language and a Finno-Ugric language. The root position between Finno-Ugric and Samoyed showed up in six of the ten phylogenies (Swadesh207, FullBasic, Swadesh100, Leipzig-Jakarta, Ura100 and 1+ Borrowings) and was absent only with the four smaller and more borrowing-prone subsets (2+ Borrowings, 3+ Borrowings, 4+ Borrowings and 5+ Borrowings).

The rooting technique suggested a traditional root position, separating Samoyed languages from Finno-Ugric languages; this was the expected result considering the nature of the analyzed data (see e.g. Salminen 2002; Janhunen 2000). Having said that it must be emphasized that this technique, which relied on a reconstructed proto-language, was experimental and differed from how an outgroup would generally be selected in phylogenetic research. An ideal outgroup would be an actual (attested) taxon that is related to but not part of the so-called 'ingroup' – i.e. the taxa that we are classifying. With cognate data it is not possible to include a language from outside the language family to be used as an outgroup, since deep-level lexical historical connections (aside from some borrowings) are generally nonexistent, or speculative at best (for more on this, see e.g. passages on mass comparison in McMahon & McMahon 2005). The phylogenetic package BEAST

offers an alternative to using an outgroup, as it can quantitatively infer the most likely root position as part of the overall analysis; this approach was used in another article which falls outside the present dissertation (Honkola *et al.* 2013), and pointed towards the same conclusion regarding the root position as Article I's rooting technique, positioning it between Samoyed and the remaining languages. A recently proposed approach called 'rootstrap' (Naser-Khdour *et al.* 2020), a metric for quantitatively evaluating various root positions in phylogenetic data, could be useful for quantitatively determining root positions for phylogenetic data in future work. However, as of the time of writing, this method has not gone through peer-review, and its applicability to linguistic datasets remains undetermined. Unlike Article I, Article II used no outgroup for its analyses; instead the trees in that article were manually rooted between the Samoyed branch and the rest of the languages, based on the results of Article I and Honkola *et al.* (2013), as well as the fact that it is the traditional root position of Uralic. The network analyses likewise used no outgroup, as a rooting point cannot be specified for a NeighborNet.

Another point to consider in retrospect is that the models used in the articles could always be improved; for instance, the Markov K model used in both Articles I and II was largely chosen because of its simplicity, as it offered a consistent reference point to compare the phylogenies against. Many refinements could be done for the model to improve its fit with language data, such as incorporating rate variation via e.g. the gamma rate-heterogeneity model or phylogenetic partitioning and treating rates of gain (i.e. transitions from 0 to 1) distinctly from rates of loss (transitions from 1 to 0) (cf. e.g. Dunn 2009; 2015). However, the fact that the model could have been improved further is a truism for any study employing statistical methods.

The results of Articles I and II gave a reasonably consistent overall impression of the Uralic language family (Figure 7). From a superficial perspective the overall picture remained a fairly treelike one – close to a 'standard paradigm' classification of these languages (see e.g. Korhonen 1981). A closer inspection of the support values, branch lengths and also comparisons between trees made from different parts of the data allowed us to evaluate the result in greater detail. This further examination indicated uncertainty in specific parts of the tree, providing support to the criticism pointed towards the traditional binary tree model of the Uralic family (see e.g. Häkkinen 1983; 1984, Salminen 1999; 2002; 2007, Kulonen 2002, Saarikivi 2011, Ylikoski 2016, Aikio in press). The results suggested, for instance, that the branching events immediately following the Finno-Ugric stage took place fairly

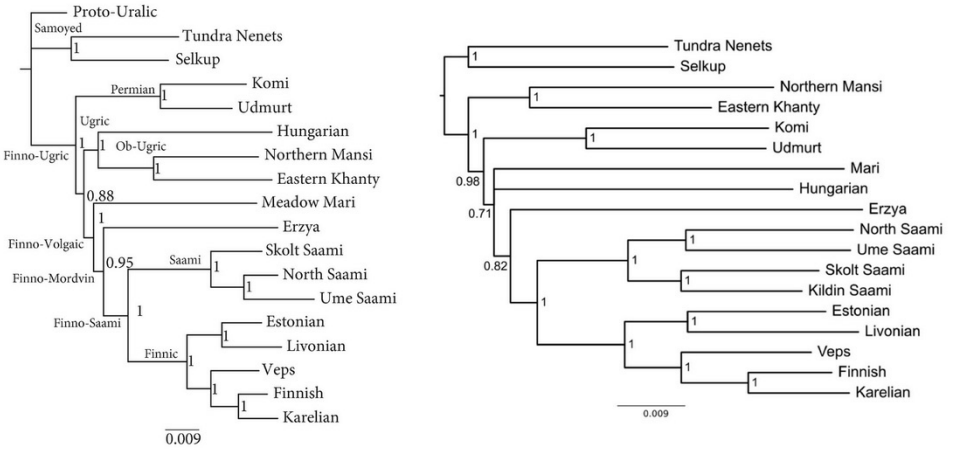
rapidly – possibly even polytomously – slowing down closer to the Finno-Saami level. Within our results the polytomous branching point was positioned within the Finno-Ugric clade, rather than the root of the Uralic tree, and did not strongly support an entirely unstructured, bush-like diversification of the shallow branches of Uralic. The rapid divergence pattern in the tree analyses is apparent from several factors in the articles, including subpar posterior probabilities of clades near Finno-Ugric (indicating that the entire distribution of trees did not unanimously support these branchings), short branch lengths (indicating that the number of inferred cognate set changes along that branch was small), and unorthodox branching orders and subgroupings, some of which only emerged with a specific subset of the data. Uncertain or ambiguous points around the Finno-Ugric root included Ugric, Permian, Finno-Volgaic (which, despite being generally regarded as obsolete, was present in several trees). Finno-Mordvin (Western Uralic) was also absent from the trees in Article 1 but was present and tentatively supported by the results of Article 2, which used a larger dataset. Languages such as Hungarian and Mari also showed instability regarding their positioning when comparing analyses done with different subsets of the lexical data, further indicating instability within the Uralic tree, especially regarding the deeper branches close to these languages. In addition to showing similar results in its tree analyses as Article I, Article II also showed weakly weighted splits (comparable to short branch lengths in phylogenetic trees) combined with a modest degree of reticulation around the initial branching point of Uralic. Article II's network analyses further indicated conflicting splits between e.g. Hungarian, Ob-Ugric, Permian, Mari and Erzya, likewise highlighting that there is considerable ambiguity around these languages within the Uralic classification.

The results firmly suggest that phylogenetic tools are able to reconstruct the big picture of Uralic language history quite successfully. Techniques such as trees and networks are especially efficient when used in combination, as it gives different perspectives to the same material. The tree-based analyses of the Uralic language family generally point towards a more treelike classification, but with some degree of ambiguity within that tree, especially closer to the root of Finno-Ugric, indicated by branch lengths, support values and differences between trees made from different parts of the data. The network analyses of Article II provides further insight into Uralic classification. In addition to showing a similar ambiguity around the Finno-Ugric stage, they also reveal more complex interrelationships than tree-based analyses could reveal between closely-related language groups, such as Finnic and Saami. The complex connections are especially prominent when analyzing

younger meanings that are more replacement-prone, such as those in the WOLD401-500 list of less basic vocabulary.

It is also worthy of note that Honkola *et al.* (2013), a study using BEAST and a different model of sequence evolution, came to similar general conclusions in terms of linguistic subgrouping as the articles presented here with respect to the general shape of the tree as well as the certainty and uncertainty within the branching order.

Figure 7. Phylogenetic tree analyses of Uralic data: (a) the full set of 226 meanings, from Article I; (b) WOLD401-500 meanings from, Article II.



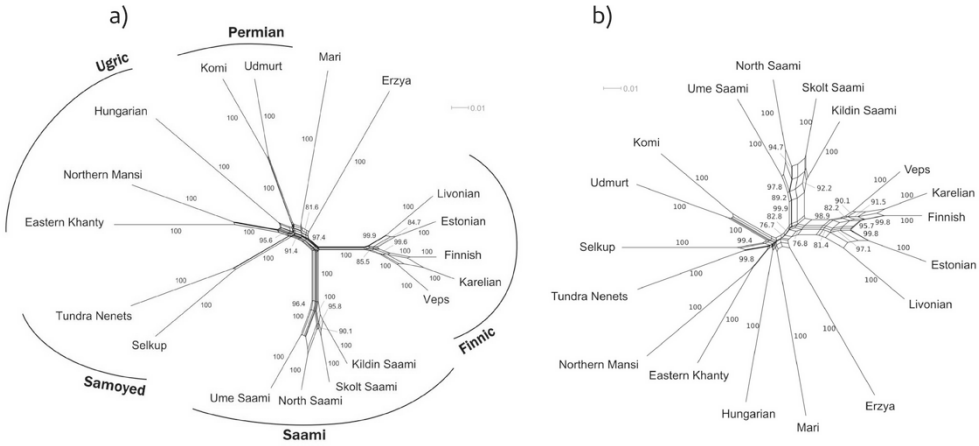
4.1.2 Tree analyses vs. network analyses of Uralic languages

For the most part the network analyses in Article II, which offer an alternative to the tree-based techniques for visualizing linguistic history, provided confirmation of the findings in Article I, with some additional advantages that provide further depth to the overall picture (Figure 8). By calculating the mean δ scores for the networks, used for estimating how tree-like a dataset is, we could confirm that the overall classification of the Uralic family, at least when observed using this dataset, was much more tree-like than network-like; the mean δ scores were quite low, ranging from 0.179 to 0.244; as a point of comparison, Table 1 in Gray *et al.* (2010) reports much higher δ scores for Polynesian languages than we observed for Uralic, ranging between 0.32 and 0.50. In many places the networks agreed with the general results of the tree-based analyses, with both of them identifying the

same robust groups under a similar hierarchy. Where the network analyses served especially useful were when examining closely-related groups with complex historical connections that are difficult to represent with a tree or include ambiguous connections. Among these are Saami and Finnic, for which the networks revealed complex reticulate connections not revealed well by the tree-based analyses, where there was little to no conflict between these clades in the form of posterior probabilities or branch lengths. Finnic languages, for instance, were systematically divided into a northern and a southern subgroup by the tree-based analyses, whereas the network analyses revealed connections between the northern and southern languages; in particular Estonian forms many connections with the northern Finnic languages. The networks also suggested a polytomous initial divergence pattern for the Uralic languages, showing Samoyed, Permian, Ugric and Finno-Volgaic splitting at around the same point in the network. The low δ scores (indicating the data to be quite tree-like) likewise supported the interpretation that the networks are more likely the result of polytomous diversification and less likely due to extensive borrowing confounding the analyses. As was the case with the tree-based analyses, these observations are in general agreement with the literature with certain interesting differences, such as the lack of clear evidence for a Finno-Mordvin (Western Uralic) proto-stage, which is reasonably well-supported by the literature (e.g. Ylikoski 2016). To sum up, it is clear that the network analyses are especially useful when dealing with sets of languages with more complex interrelationships, such as areal connections combined with genetic connections, as well as other properties that may not ideally fit a tree-based model.

There are also certain disadvantages in the network analyses compared to the tree-based analyses. The underlying model of NeighborNet is generally not malleable in the same way as Bayesian phylogenetic inference is, and networks, to my knowledge, cannot be used for e.g. the inference of divergence times or the representation of chronology. On the other hand, network-based analyses have been argued to provide a more natural representation for linguistic material than trees due to their ability to simultaneously represent horizontal connections such as language contact and and inheritance (McMahon & McMahon 2005), an argument which the results of Article II also supported. For that reason, network analyses continue to serve as a valuable tool in the quantitative linguistic toolkit.

Figure 8. NeighborNet analyses of Uralic data from Article II: (a) all 226 basic vocabulary meanings; (b) WOLD401-500 meanings.



4.2 Finnish dialects as populations (III)

In contrast with the macro-evolutionary perspective of Articles I and II, Article III approached the internal variation of one language – Finnish – using evolutionary tools from population genetics, which focuses on the study of within-species relationships. Consequently, this study served as a pilot study for increasing the scope of evolutionary language study from between-language relationships to within-language relationships.

The bulk of the study consisted of population clustering analyses, using the software package Structure, which were conducted to appropriately coded dialect data from the Atlas of Finnish Dialects (Kettunen 1940a). The coding approach has been explained in the Materials and Methods section. The population clustering analyses were also compared with results produced using K-medoids clustering, which produces similarly nonhierarchal groups but is not based on population-genetic principles, like Structure. The aim of these analyses was, on the one hand, evaluate how successful the population-genetic Structure package is in inferring dialects from this data, compared to K-medoids or existing knowledge on the dialect areas, and, on the other hand, produce a quantitative subdivision of Finnish dialects using these techniques. Article III also examined the overall nature

of Kettunen's dialect atlas using a technique designed for comparing how linked or correlated different map pages are to one another, as well as experimented briefly with other population genetic tools. It should be noted that the main purpose of Article III was not to propose a new dialect clustering of Finnish, but rather explore a new technique which might potentially be used for inferring divisions like these.

4.2.1 Dialect areas based on population genetic clustering

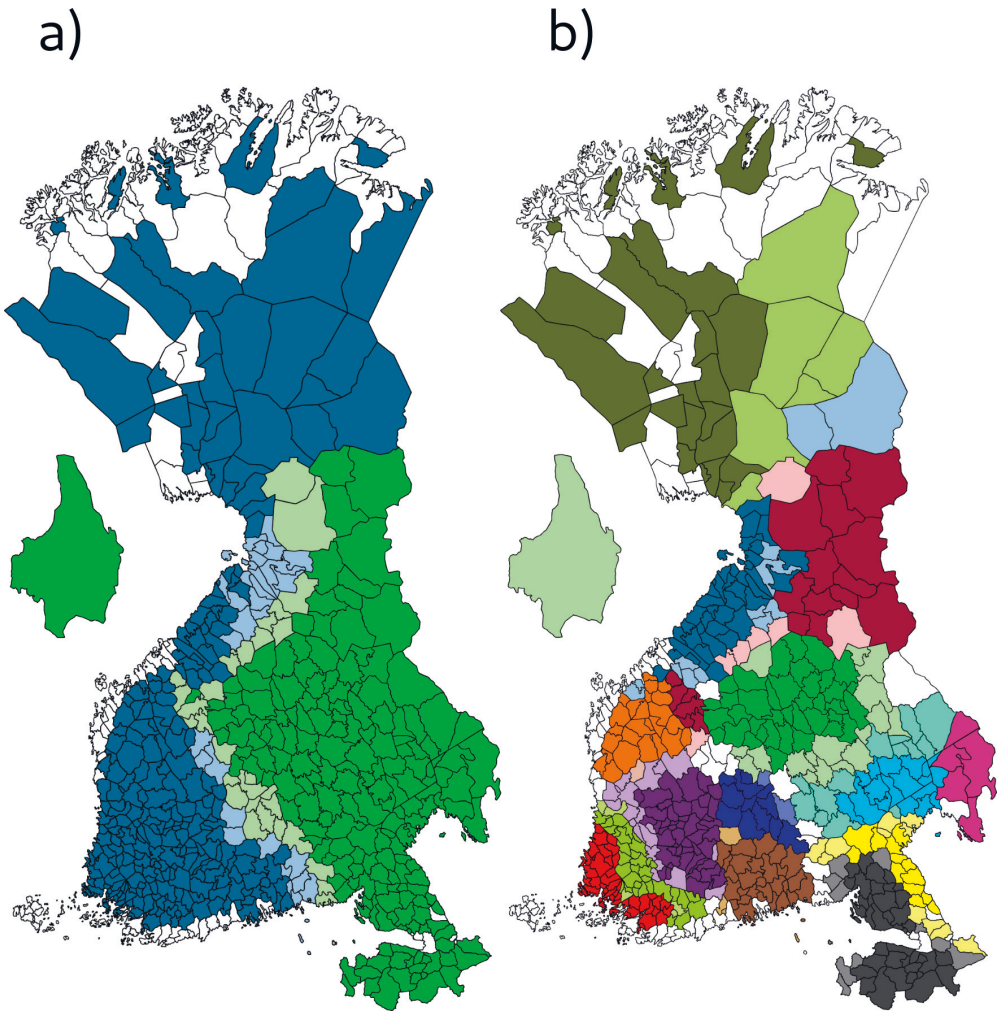
Structure's analyses were run for K-values 2-20, that is, we used Structure to divide the dialect data into into 2-20 dialect areas. Upon examining the likelihood values of these runs we found that the likelihood estimates plateaued at around K=14, after which they began to fluctuate significantly across different repetitions. As per Structure's guidelines, we focused our attention on the runs preceding this fluctuation, i.e. K=2-14. We also produced comparable analyses using K-medoids clustering.

As was already mentioned, likelihood estimates were highest with K=14 (Figure 9b), suggesting a 14-way division of Finnish as the best fine-grained dialect division. The likelihood estimates were generally fairly high with K values from 2 to 14, suggesting that all of them were much better subdivisions for the dialect data than those produced with higher K values. While K=15 and K=16 were also on average quite high, they were already much less consistent across repetitions. The *deltaK* metric, which is used to identify the uppermost level of a population structure based on identifying significant changes in the the likelihood value around a specific K value, suggested K=2 as the best coarse-grained hierarchy (Figure 9a). This agrees with Finnish dialectological literature, where the principal dialect division is between eastern and western dialects. The K=2 analysis agreed well with this division. For K-medoids, we examined the silhouette values of K-medoids analyses, which also aim to identify the optimal K-values of a cluster analysis. On a general level these agreed with Structure's likelihood estimates, showing fairly small differences in the silhouette values across the runs but rising as the K values rise. The silhouette values stabilized at K=6 and beyond, and ultimately gave the highest support for K=16, albeit by a very small margin compared to the next best values.

In addition to the K=2 division, which produced essentially the textbook dialect division into east and west, some of the divisions also bear resemblance to divisions from dialectological literature. With K-medoids, K=3 provides a fairly close match with the three-way division from Leino *et al.* (2006) and Hyvönen *et al.* (2007), while Structure's K=3 division appears to reflect an earlier three-way division originating from Lencqvist, which has also been discussed in Paunonen (1991; 2006) and Mielikäinen (1991). However, many of the quantitative divisions, while mostly resembling an existing division, have notable differences. For instance, the K=8 division produced by the analyses does not match up with the 'gold standard' eight-way division from Itkonen (1964), as it replaces its Southwest transitional dialect area with a Southeast Häme dialect area (see e.g. Rapola 1969; Wiik 2004). By and large, the analyses appear to agree quite well with what one would expect from the principally morphophonological dialect data that Kettunen's dialect atlas covers.

For lower K values (2-8) Structure and K-medoids clusterings generally agreed with each other, with the only exception being K=3, as discussed in the previous paragraph. With higher K values there were more differences between the two methods. In general the dialect areas made sense even though the clustering analysis did not include information on the geographical proximity of the different data points. However, K-medoids appeared to be more prone to data quality than Structure; with K=9-14 its results included a discontinuous cluster covering Border Karelia, Ingria and a small selection of border area municipalities. Upon closer inspection it became apparent that this cluster consisted of points that were poorly documented in the dialect atlas. Another matter where Structure is an improvement over K-medoids is its fuzzy clustering, which provides a more natural picture of the dialect areas by revealing transitional areas between dialect areas in addition to the dialect areas themselves.

Figure 9. Finnish dialect populations based on Structure clustering, from Article III: (a) K=2, the best coarse-grained division; (b) K=14, the best fine-grained division. Lighter colors indicate areas with more admixture between the inferred dialect clusters.



4.2.2 Linkage test

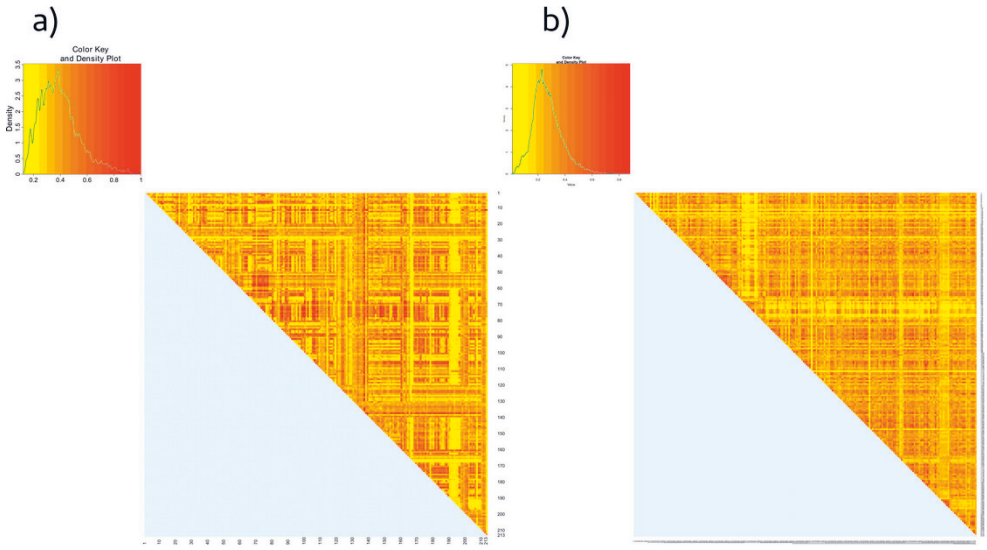
As part of exploring potential linkage in Kettunen's dialect atlas we devised a method for quantitatively comparing different map pages. The formula for this calculation is given in Materials and methods. Notably, while the linkage test's results published in Article III turned out to be erroneous due to a programming

error (incorrectly nested conditionals) in the original implementation of the linkage test. Here I provide the original description as well as its correction (Figure 10).

In the original version it was noted that the overall connectedness between linguistic features (map pages) in the dialect atlas is generally low, with certain features, such as those covering smaller geographical areas, showing up as having higher linkage. The original linkage test also produced biased results in situations where a geographically constrained feature, such as the one documented on map page 137, was consistently deemed as being strongly linked against most other pages.

A corrected version of the linkage analysis, similar to the original one, shows that the overall connectedness between the features is low, even lower than in the erroneous version. More significantly it shows that many of the aforementioned problems of the original test, such as high linkage across pages covering small geographical areas, disappeared with the corrected version. The corrected version is much clearer in showing sets of redundant features with potential linkage problems. Kettunen's dialect atlas is ordered so that similar phenomena are documented on consecutive pages, so linkage across redundant features mostly appear as triangles near the diagonal that goes from the top-left corner of the heatmap to the bottom-right corner. Following this diagonal, one can see, for instance, a reddish triangle around map pages 52-59, which all document the genitive forms of nouns. An even more pronounced triangle can be seen around map pages 191–195, which, according to the dialect atlas itself, all involve the development of vowel combinations ending in *a* or *ä*.

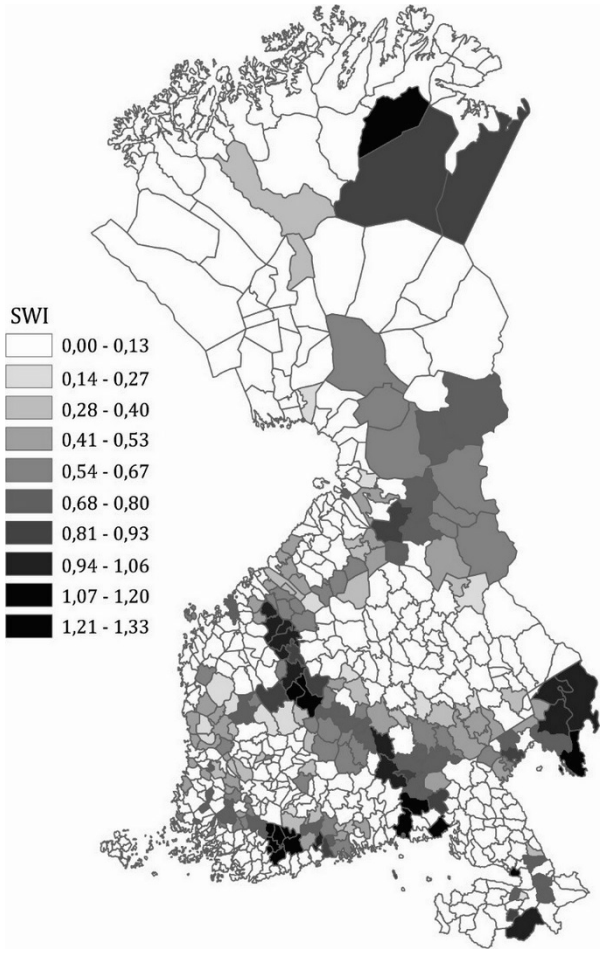
Figure 10. Linkage test heatmaps measuring connectedness between dialect features: (a) published heatmap from Article III; (b) corrected heatmap. Each cell horizontal and vertical cell correspond with a map page from the dialect atlas, organized in the same order as they are in the atlas itself. Red color indicates a value closer to 1.0 (more connectedness between features), whereas yellow indicates a value closer to 0.0 (less connectedness between features).



4.2.3 Linguistic diversity and the similarity of inferred populations

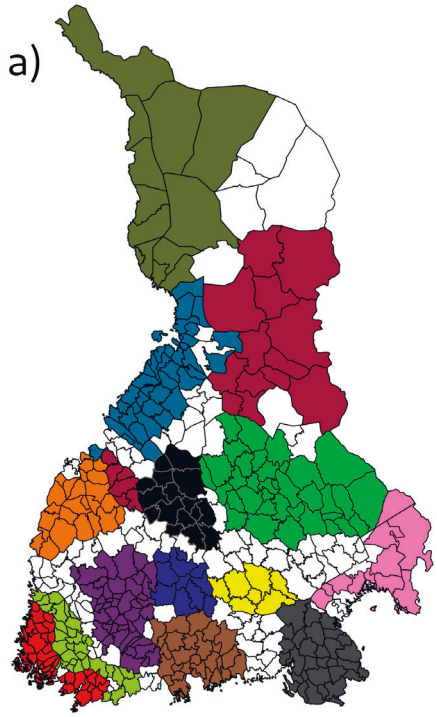
Article III also experimented with two further quantitative approaches from population genetics. The first of these were *Shannon-Wiener diversity index*, also known as *Shannon's entropy* (Legendre & Legendre 2012), which measures diversity in ecological communities, and can be calculated from the IC values of the Structure analyses. The index is low when the amount of linguistic diversity is low – that is, when the traits of a specific dialect are dominant in that municipality, and higher in areas with mixed dialects. Our tests with Finnish language data showed that this index was essentially maximized at the border areas between dialects, as one would expect (Figure 11).

Figure 11. Shannon-Wiener indices (swi) of each municipality in the dialect data, from Article III. swi were calculated after dividing the dialect data into seven populations. swi are divided into ten equal-sized classes: from the smallest swi, indicating the lowest amount of linguistic diversity (municipalities colored with white), to the class of the largest swi, indicating the largest amount of linguistic diversity (municipalities colored with black).



Another technique examined in Article III was *Fst* or *fixation index*, which can be used to measure how similar the inferred populations are in terms of linguistic variation. This in turn reveals whether there is an inferrable substructuring between the areas. In our experiments we calculated Fst values between 14 core areas inferred with Structure. The results lined up quite well in relation to what we know about the structure of Finnish dialects. For instance, Fst was maximized around the area of the main dialect division (east-west), and were lowest in areas with transitional dialects (Figure 12).

Figure 12. Fst analysis of dialect data, from Article III: (a) core dialect areas identified from a k=14 (14-cluster) Structure run, using an *ic* value threshold of 0.75; (b) pairwise *Fst* values, indicating the amount of linguistic difference of the populations, with color codes matching the clusters shown in (a).



b)

0													
0.70	0												
0.71	0.57	0											
0.50	0.55	0.62	0										
0.70	0.66	0.70	0.44	0									
0.76	0.62	0.53	0.65	0.70	0								
0.74	0.65	0.53	0.62	0.65	0.59	0							
0.81	0.76	0.55	0.75	0.78	0.58	0.69	0						
0.74	0.68	0.69	0.59	0.64	0.72	0.74	0.80	0					
0.78	0.72	0.44	0.69	0.77	0.65	0.60	0.61	0.80	0				
0.64	0.55	0.56	0.46	0.46	0.62	0.50	0.71	0.62	0.63	0			
0.81	0.78	0.60	0.73	0.78	0.57	0.66	0.54	0.81	0.67	0.69	0		
0.71	0.48	0.60	0.58	0.63	0.49	0.61	0.70	0.62	0.68	0.59	0.67	0	
0.74	0.65	0.42	0.66	0.73	0.61	0.62	0.65	0.74	0.52	0.59	0.67	0.65	0

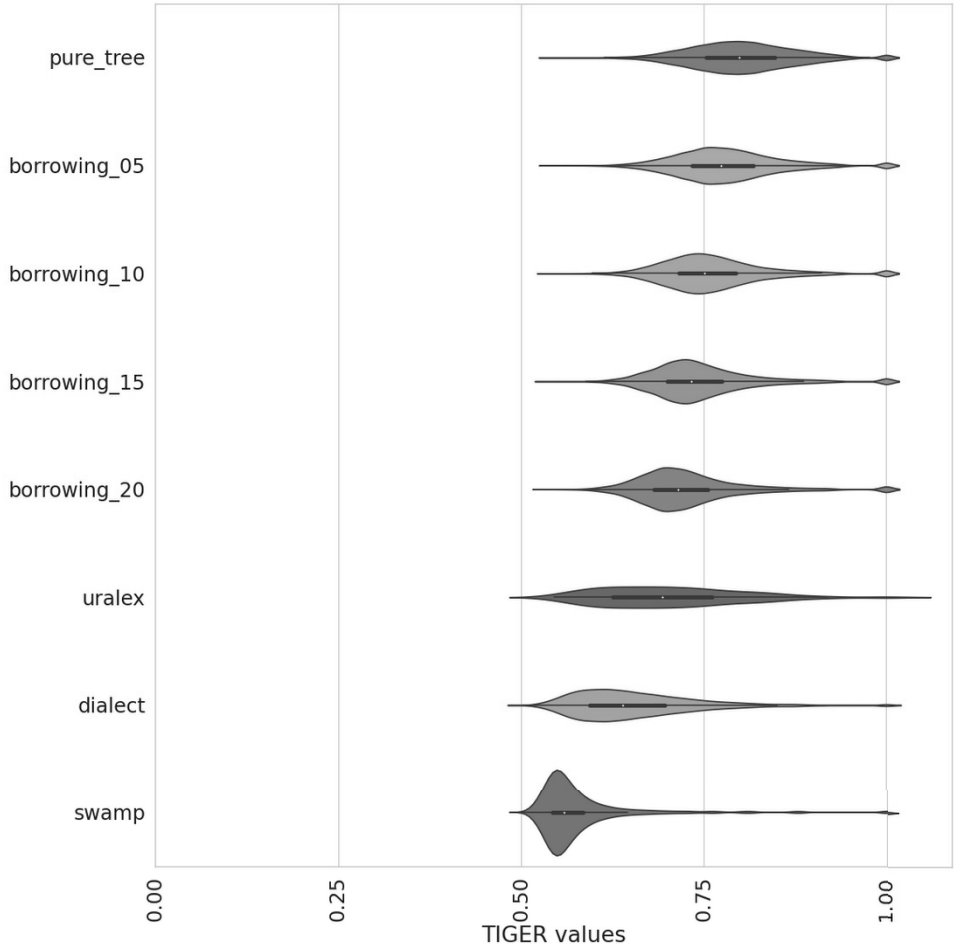
4.3 TIGER values as a metric of treelikeness (IV)

Article IV's primary aim was methodological exploration, as it focused on the application of TIGER values (Cummins & McInerney 2011) as a metric for treelikeness, similar to metrics such as the δ score (Holland 2002) and the Q-residual (Gray *et al.* 2010). This was explored using both real-life language data from the UraLex Basic Vocabulary Dataset (Syrjänen *et al.* 2018), as well as simulated language datasets that mimicked different linguistic situations: a purely treelike language dataset, a treelike dataset with different degrees of borrowing, a spatially structured dataset mimicking a dialect chain, and a completely unstructured 'swamp' dataset that superficially resembled cognate data. TIGER values, δ scores and Q-residuals were all calculated for all the datasets, and also visually examined using NeighborNets.

One shortcoming of the TIGER algorithm is that it cannot account for sets of multiple interrelated characters, such as the binary characters that represent the cognate sets of one meaning. For this reason it requires a multistate representation of the language data such that each meaning is represented by a one and only one multistate character. In contrast, UraLex can have multiple reflexes per language for a meaning. For the study in Article IV we algorithmically chose one cognate set for each language to represent each meaning, using a 'minimizing' strategy which selected the representative cognates so that the total number of cognate sets per each meaning would remain as low as possible. This type of strategy essentially favors deeper relationships over shallower ones.

The results showed that TIGER values performed well at detecting treelikeness of datasets. With simulated datasets the mean TIGER value was highest with purely tree-like data, gradually dropped as the degree of treelikeness reduced, and was minimized by the completely unstructured 'swamp' dataset (Figure 13). The TIGER values positioned the Uralic language dataset between the simulated dataset that had 20% borrowings, and the spatially structured dialect chain dataset. Considering what we know of the Uralic language family and the nature of the UraLex dataset, as well as the trees, networks and δ scores reported in Articles I and II, this seemed a reasonable estimate.

Figure 13. TIGER value distributions for simulated and real-life language data, from Article IV The simulated datasets, from most tree-like to least tree-like, are: pure_tree (purely tree-like data), borrowing_05 (tree-like data with 5% likelihood of borrowing), borrowing_10 (tree-like data with 10% likelihood of borrowing), borrowing_15 (tree-like data with 15% likelihood of borrowing), borrowing_20 (tree-like data with 20% likelihood of borrowing), dialect (simulated dialect chain data) and swamp (unstructured data). Higher TIGER values indicate a more treelike signal.



Compared to the existing metrics (Figure 14), TIGER values produced similar results as δ scores for the simulated datasets, with the obvious difference that δ scores are at their lowest with treelike data whereas the opposite is true with TIGER values. Q-residuals however, ranked the datasets differently, being

sensitive to reticulation, or webiness, of the data (maximized by the spatially structured dialect chain data) rather than the degree of treelikeness that the data has. The three metrics also characterized the real-life UraLex dataset differently, with the δ score characterizing it as being between a pure tree and a 5% borrowing dataset, whereas Q-residuals position it between a dialect chain dataset and the ustructured 'swamp' dataset. NeighborNets (Figure 15) revealed that the likely reason for this classification was the degree of reticulation or 'webiness', which was largest with the dialect chain data and UraLex, despite the two datasets otherwise being visually quite different from one another.

Figure 14. Comparison of TIGER values, δ scores and Q-residuals for different datasets, from Article IV.

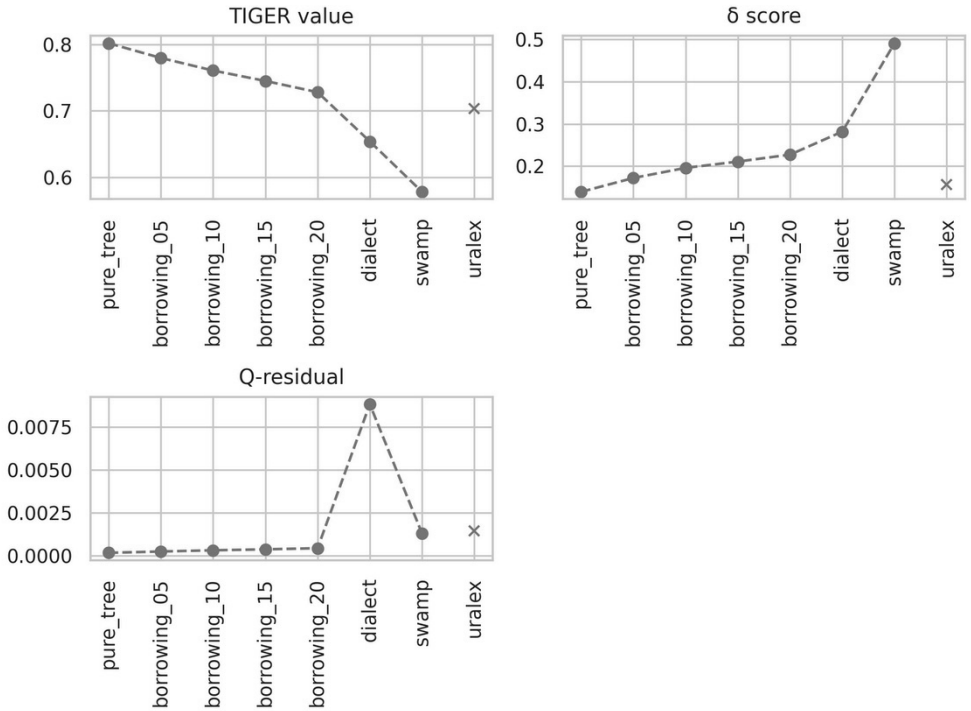
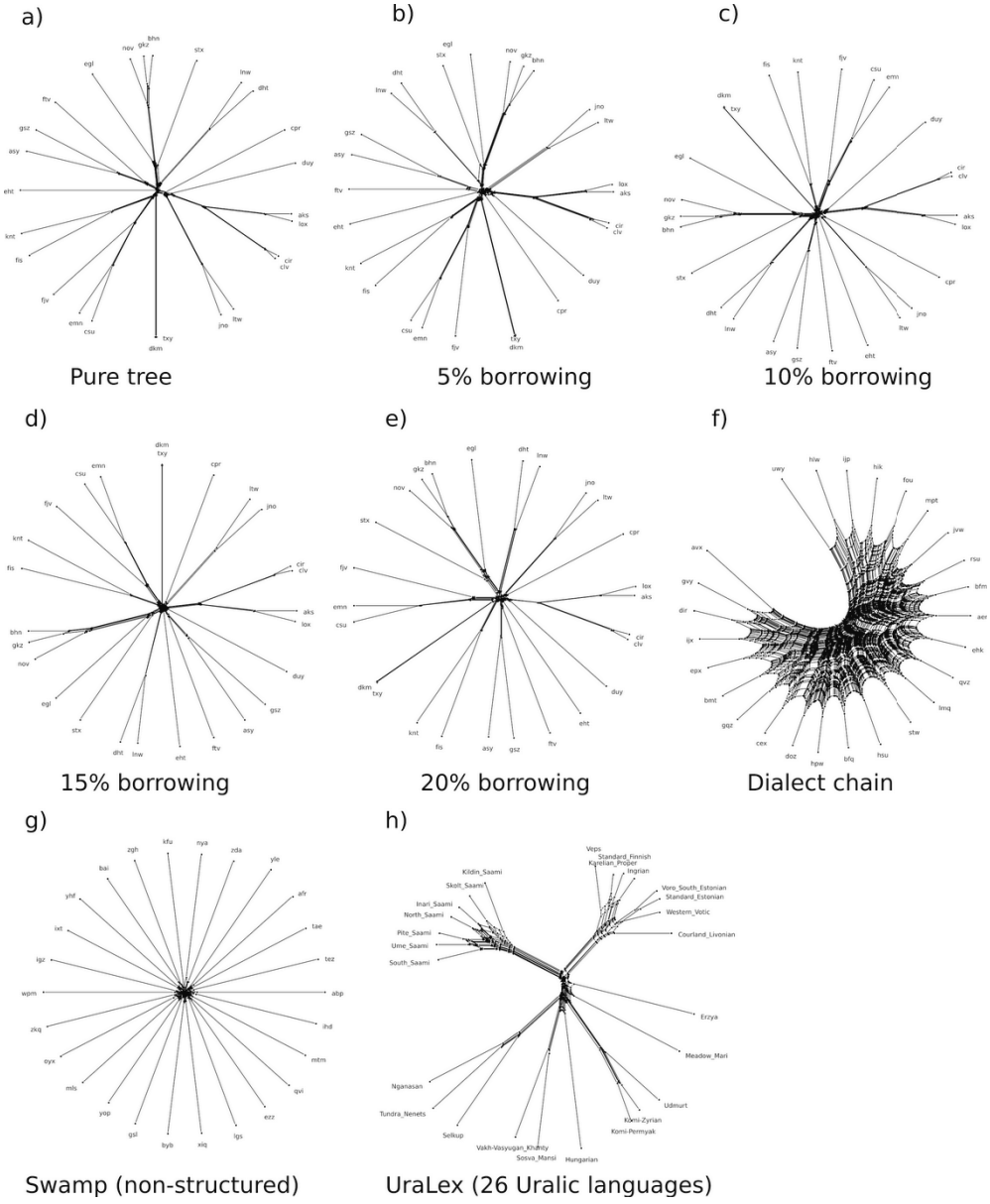
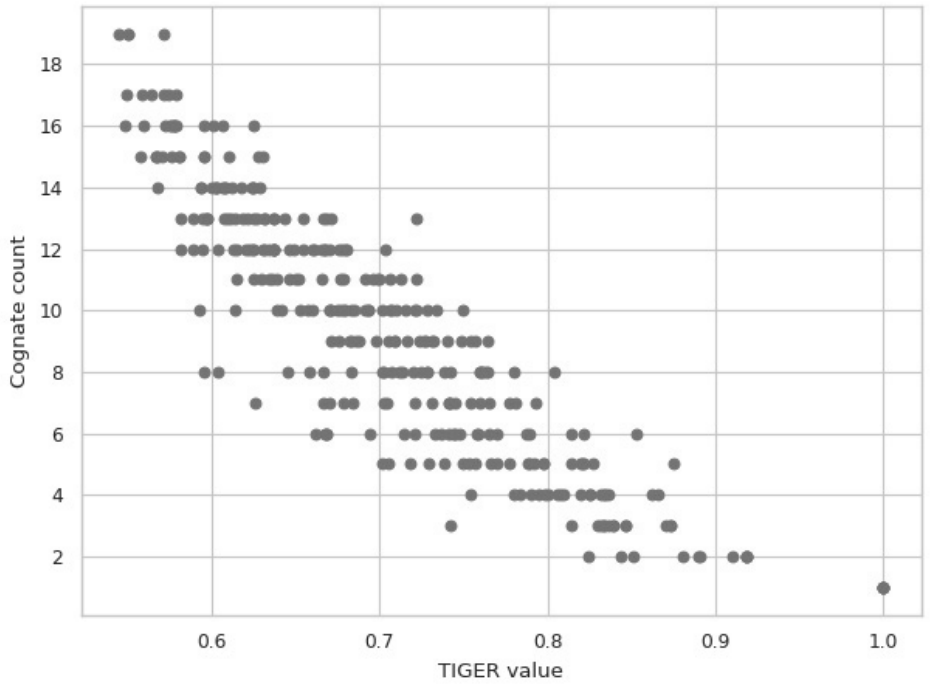


Figure 15. NeighborNets produced from simulated and real-life language data, from Article IV.



An examination of the TIGER values of individual meanings in UraLex showed that TIGER values were correlated with how many cognate classes each meaning had, but are not a simple proxy for cognate class count.

Figure 16. Scatterplot comparing TIGER values and cognate set counts of UraLex, from Article IV.



We also examined whether ad hoc linguistic differences such as basic vocabulary membership or semantic differences would be evident based on TIGER values. While basic vocabulary and non-basic vocabulary were distinguishable from one another based on TIGER values, the semantic classes in general were not distinguishable from one another by their TIGER value.

In conclusion, based on our results we found that TIGER values would serve as a good metric for estimating treelikeness, and even one that to an extent outperforms metrics such as δ scores and Q-residuals without adding too much computational complexity.

5 CONCLUSIONS AND FUTURE PERSPECTIVES

As the four studies presented here show, evolutionary tools are applicable to language data at both between-language and within-language levels. At both these levels they produce results that are compatible with the results of earlier research. An added advantage of these techniques comes from their ability to quantify ambiguity in the results through e.g. support values, branch lengths (with phylogenetic techniques) and population admixture (with population genetic techniques).

In the light of the phylogenetic analyses, the Uralic language phylogeny was mostly treelike, although not one that is necessarily fully resolved, especially between the ancestral stages between Finno-Ugric and Finnic. The network analyses also suggested that the Uralic family was more treelike than it is network-like, but with certain sections that are nontreelike, such as the relationships between Finnic languages and the Saamic languages. The uncertainty around the initial branching point of Uralic was also apparent from the network analyses.

It should obviously be kept in mind that the results of Articles I, II and IV represent the Uralic language history viewed through the lens provided by the UraLex Basic Vocabulary Database. This data consists solely of historical connections recorded for meanings from several basic vocabulary lists, alongside a smaller set of historical connections recorded for non-basic vocabulary meanings. As such, it excludes linguistic considerations not recorded by this type of data. With this in mind, it is likely that future revisions of the UraLex lexical cognate data – which are already in the works – as well as the inclusion of other kinds of data, such as the typological data currently collected by the UraTyp project, will surely shed further light on the overall picture of Uralic in future quantitative studies of this family.

The population genetic analyses of the Finnish dialect atlas similarly produced a very familiar picture of Finnish dialects, characterized by a strong east-west dichotomy. For the most part the dialect areas inferred by the analyses reflected the

traditional subdivisions (such as those in Figure 3), although with some differences such as a preference for a more fine-grained overall division (fourteen dialect populations rather than the traditional seven or eight). The fuzzy clustering technique from Structure's admixture model was able to produce a very natural picture of the Finnish dialects, visualizing both transitional dialect areas as well as clearer dialect areas quantitatively.

Also here it should be kept in mind that the dialect analyses in Article III obviously represent a predominantly morphophonological history of Finnish dialects, based on the data collected by Lauri Kettunen in the early 1900s, before large-scale urbanization took place. Other types of dialect data emphasize different kinds of results, as is apparent from e.g. the preference of a three-way dialect division of Finnish with lexical data[4] over the two-way division, which is prominent with Kettunen's data.

The new evolutionary quantitative tools are not just such that they repeat what we already know, but come with a number of added benefits compared to traditional work. The tools offer good repeatability of analyses, which is not only good for double-checking the results but also for updating old results as new type of data becomes available, or existing datasets are expanded. In addition to improved repeatability, the quantitative tools are accompanied by supplementary techniques that allow e.g. more intricate analyses or accurate objective visualization. In essence these are often not individual tools but rather parts of a larger tool chain. Quantifiable results also make the results more attractive for multidisciplinary work, as multiple fields can work with the shared language of numbers to shed light on historical topics that span beyond the history of languages. This will undoubtedly result in increased collaboration between different research disciplines that tackle with the history of humans and their culture from different perspectives.

The evolutionary toolkit is constantly being expanded to better tackle with new challenges that new subject matter has brought to it. In cultural history one such challenge involves the expansion of existing tools and techniques to better account for different types of historical connections, such as horizontal (non-treelike) and vertical (treelike), as many of the current tools remain primarily tree-based. Article IV of this dissertation also made an attempt to contribute to this challenge by

[4] See e.g. the analyses of a dictionary of Finnish dialects by Leino *et al.* (2006) and Hyvönen *et al.* (2007).

proposing TIGER values as a reasonably accurate but computationally inexpensive metric for quantifying treelikeness in a dataset. In a similar way, article III also focused on a largely unexplored (or at least underexplored) territory by applying various population genetic techniques to study intralingual variation. Another challenge is to further improve the compatibility of these methods with linguistic data, which differs considerably from its biological counterparts; tools such as BEAST are already seeing improvements in the form of models that account for linguistic data better. As these methods improve, it is also likely that this will feed into the nature of the datasets themselves, with more detailed information complementing the historical connections that make up most of the datasets today. One such dataset advancement, proposed by List (2016), would involve the inclusion of information about the degree of historical relatedness between historically connected items, such as the extent to which they retain morphophonological or semantic similarity with their source.

With respect to datasets, it is also likely that future work will see an increase in the use of datasets of different types, such as structural data alongside basic vocabulary data, which will likewise add another layer of complexity to the overall picture. Combining multiple types of data into one analysis will be a challenge, although some potential techniques for this exist, such as phylogenetic partitioning, which allows having separate models of character evolution for differently evolving portions of the data, such as basic vocabulary data and structural data. Finally, it is also likely that future research will see an increase in ambitious studies that attempt to combine multiple analysis techniques and information from multiple research fields to quantitatively explore cultural history from a more holistic perspective, in a similar way as e.g. the phylogeographic analyses of Bouckaert *et al.* (2012), or the dialect analyses of Honkola *et al.* (2018).

As the historical overview showed, the meaning of 'evolution' in the context of linguistics is a problem not only due to its multiple meanings but also because it is not necessarily clear whether the notion of 'evolution' is compatible with the notion of 'language change', which extends to the question of whether evolutionary techniques are fundamentally compatible with linguistic material. As has hopefully been shown here, this question is largely nonessential, as the techniques themselves do not model evolution in any holistic sense but rather only small parts of it, based on their intended use. The key challenge then is to ensure that the parts that the techniques do model are not incompatible with what we know about language change, and to ensure that we understand to what extent the results of these

techniques make sense with linguistic data. The aim of evolutionary analyses is not to replace the theory of language change with that of evolution, but rather provide new and robust tools to be used in a way that makes sense within the theoretical framework of language change.

6 REFERENCES

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

Agricola, Mikael. 1548. *Se Wsi Testamenti*. Stockholm: Amund Lauritzon. Facsimile Edition 1987. Helsinki: WSOY.

Aikio, Ante. (in press). "Proto-Uralic." In *The Oxford Guide to the Uralic Languages*, edited by Marianne Bakró-Nagy, Johanna Laakso & Elena Skribnik. Oxford: Oxford University Press.

Alpher, Barry, and David Nash. 1999. "Lexical replacement and cognate equilibrium in Australia." *Australian Journal of Linguistics* 19, 5–56.

Aronoff, Mark. 2017. "Darwinism Tested By The Science Of Language." In *On Looking into Words (and Beyond)*, edited by Claire Bowern, Laurence Horn, and Raffaella Zanuttini, 443–56. Berlin: Language Science Press.

Atkinson, Quentin D., and Russell D. Gray. 2005. "Curious Parallels and Curious Connections—Phylogenetic Thinking in Biology and Historical Linguistics." *Systematic Biology* 54 (4): 513–26. https://doi.org/10.1080/10635150590950317.

Audesirk, Teresa, Gerald Audesirk, and Bruce E. Byers. 2008. *Biology: Life on Earth with Physiology*. 8th ed. Upper Saddle River, NJ: Pearson Prentice Hall.

Bastin, Yvonne, Andre Coupez, and Michael Mann. 1999. *Continuity and divergence in the Banthu languages: Perspectives from a lexicostatistic study*. Tervuren, Belgium: Musée royal de l'Afrique centrale.

Bouckaert, Remco, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard, and Quentin D. Atkinson. 2012. "Mapping the Origins and Expansion of the Indo-European Language Family." *Science* 337 (6097): 957–60. https://doi.org/10.1126/science.1219669.

Boyd, Robert, and Peter J. Richerson. 1988. *Culture and the Evolutionary Process*. Chicago: University of Chicago Press.

Bryant, David, and Vincent Moulton. 2004. "Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks." *Molecular Biology and Evolution* 21 (2): 255–65. https://doi.org/10.1093/molbev/msh018.

Campbell, Lyle. 2003. "How to Show Languages are Related: Methods for Distant Genetic Relationship." In *The Handbook of Historical Linguistics*, edited by Brian D. Joseph, and Richard D. Janda, 262-82. Oxford: Blackwell.

Cavalli-Sforza, L. L., A. Piazza, P. Menozzi, and J. Mountain. 1988. "Reconstruction of Human Evolution: Bringing Together Genetic, Archaeological, and Linguistic Data." *Proceedings of the National Academy of Sciences* 85 (16): 6002–6. https://doi.org/10.1073/pnas.85.16.6002.

Cavalli-Sforza, L. L., and William S-Y. Wang. 1986. "Spatial Distance and Lexical Replacement." *Language* 62 (1): 38–55.

Chambers, J. K., and Peter Trudgill. 1998. *Dialectology*. Second edition. Cambridge: Cambridge University Press.

Chang, Will, Chundra Cathcart, David Hall, and Andrew Garrett. 2015. "Ancestry-Constrained Phylogenetic Analysis Supports the Indo-European Steppe Hypothesis." *Language* 91 (1): 194–244. https://doi.org/10.1353/lan.2015.0005.

Chomsky, Noam. 2017. "The Language Capacity: Architecture and Evolution." *Psychonomic Bulletin & Review* 24 (1): 200–203. https://doi.org/10.3758/s13423-016-1078-6.

Creanza, Nicole, Oren Kolodny, and Marcus W. Feldman. 2017. "Cultural Evolutionary Theory: How Culture Evolves and Why It Matters." *Proceedings of the National Academy of Sciences* 114 (30): 7782–89. https://doi.org/10.1073/pnas.1620732114.

Croft, William. 2008. "Evolutionary Linguistics." *Annual Review of Anthropology* 37 (1): 219–34. https://doi.org/10.1146/annurev.anthro.37.081407.085156.

Croft, William A. 2006. "The Relevance of an Evolutionary Model to Historical Linguistics." In *Current Issues in Linguistic Theory*, edited by Ole Nedergaard Thomsen, 279:91–132. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/cilt.279.08cro.

Croft, William. 2000. *Explaining Language Change: An Evolutionary Approach*. Longman Linguistics Library. Harlow, England; New York: Longman.

Crystal, David. 1987. *The Cambridge Encyclopedia of Language*. Cambridge: Cambridge University Press.

Cummins, Carla A., and James O. McInerney. 2011. "A Method for Inferring the Rate of Evolution of Homologous Characters That Can Potentially Improve Phylogenetic Inference, Resolve Deep Divergence and Correct Systematic Biases." *Systematic Biology* 60 (6): 833–44. https://doi.org/10.1093/sysbio/syr064.

Da Silva, Augusto Soares. 2010. "Replication, Selection and Language Change. Why an Evolutionary Approach to Language Variation and Change?" *Revista Portuguesa De Filosofia* 66 (4): 803–18.

Darwin, Charles. 1871. *The Descent of Man*. London: John Murray.

Dunn, Michael. 2015. "Language Phylogenies." In *The Routledge Handbook of Historical Linguistics*, edited by Claire Bowern and Bethwyn Evans, 190–211. London: Routledge.

Dunn, Michael. 2009. "Contact and Phylogeny in Island Melanesia." *Lingua* 119 (11): 1664–78. https://doi.org/10.1016/j.lingua.2007.10.026.

Dunn, Michael, Stephen C. Levinson, Eva Lindström, Ger Reesink, and Angela Terrill. 2008. "Structural Phylogeny in Historical Linguistics: Methodological Explorations Applied in Island Melanesia." *Language* 84 (4): 710–59. https://doi.org/10.1353/lan.0.0069.

Earl, Dent A. and vonHoldt, Bridgett M. 2012. "structure harvester: A website and program for visualizing structure output and implementing the Evanno method." *Conservation Genetics Resources* 4(2): 359–61.

Embleton, Sheila M. 2000. "Lexicostatistics/Glottochronology: From Swadesh to Sankoff to Starostin to Future Horizons." In *Time Depth in Historical Linguistics Vol. 1*, edited by Colin Renfrew, April M. S. McMahon, and Larry Trask, 143–65. Cambridge: The McDonald Institute for Archaeological Research.

Embleton, Sheila M. 1986. *Statistics in Historical Linguistics*. Quantitative Linguistics 30. Bochum: Brockmeyer.

Embleton, Sheila M., and Eric Wheeler. 1997. "Finnish dialect atlas for quantitative studies." *Journal of Quantitative Linguistics* 4(1-3), 99–102.

Embleton, Sheila M., and Eric Wheeler. 2000. "Computerized dialect atlas of Finnish: Dealing with ambiguity." *Journal of Quantitative Linguistics* 7(3), 227-31.

Evanno, Guillaume, Sebastien Regnaut, and Jérôme Goudet. 2005. "Detecting the number of clusters of individuals using the software structure: A simulation study." *Molecular Ecology* 14: 2611–20.

Francis, Roy M. 2014. pophelper: An r package for analysis of structure and tess files. r package version 1.0.0.

François, Alexandre. 2014. "Trees, Waves and Linkages: Models of Language Diversification." In *The Routledge Handbook of Historical Linguistics*, edited by Claire Bowern and Bethwyn Evans, 161–89. London: Routledge.

Frandsen, Paul B., Brett Calcott, Christoph Mayer, and Robert Lanfear. 2015. "Automatic Selection of Partitioning Schemes for Phylogenetic Analyses Using Iterative K-Means Clustering of Site Rates." *BMC Evolutionary Biology* 15 (1): 13. https://doi.org/10.1186/s12862-015-0283-7.

Gray, Russell D., David Bryant, and Simon J. Greenhill. 2010. "On the Shape and Fabric of Human History." Philosophical Transactions of the Royal Society B: Biological Sciences 365 (1559): 3923–33. https://doi.org/10.1098/rstb.2010.0162.

Gray, Russell D., and Quentin D. Atkinson. 2003. "Language-Tree Divergence Times Support the Anatolian Theory of Indo-European Origin." *Nature* 426 (6965): 435–39. https://doi.org/10.1038/nature02029.

Gray, Russell D., and Fiona M. Jordan. 2000. "Language Trees Support the Express-Train Sequence of Austronesian Expansion." *Nature* 405 (6790): 1052–55. https://doi.org/10.1038/35016575.

Greenhill, Simon J. 2016. "Phylogemetric: A Python Library for Calculating Phylogenetic Network Metrics." *The Journal of Open Source Software* 1 (2): 28. https://doi.org/10.21105/joss.00028.

Greenhill, Simon J., Quentin. D. Atkinson, Andrew Meade, and Russell D. Gray. "The Shape and Tempo of Language Evolution." *Proceedings of the Royal Society B: Biological Sciences* 277 (1693): 2443–50. https://doi.org/10.1098/rspb.2010.0051.

Greenhill, Simon J., and Russell D. Gray. 2009. "Austronesian Language Phylogenies: Myths and Misconceptions about Bayesian Computational Methods." In *Austronesian Historical Linguistics and Culture History: A Festschrift for Robert Blust*, edited by Alexander Adelaar and Andrew Pawley, 375–98. Canberra: Pacific Linguistics.

Greenhill, Simon J., Paul Heggarty, and Russell D. Gray. 2020. "Bayesian Phylolinguistics." In *The Handbook of Historical Linguistics, Volume II*, edited by Richard D. Janda, Brian D. Joseph, and Barbara S. Vance, 226–53. Hoboken: Wiley. https://doi.org/10.1002/9781118732168.ch11.

Hakulinen, Lauri. 1950. "Kansankielen sanakirjan koeartikkeleja." *Virittäjä* 54: 425-44.

Hall, Barry G. 2011. *Phylogenetic Trees Made Easy: A How-To Manual*. 4th ed. Sunderland: Sinauer Associates.

Helimski, Eugene. 2003. "Areal groupings (Sprachbünde) within and across the borders of the Uralic language family: A survey". *Nyelvtudományi Közlemények* 100: 156-167.

Holden, Clare Janaki. 2002. "Bantu Language Trees Reflect the Spread of Farming across Sub-Saharan Africa: A Maximum-Parsimony Analysis." *Proceedings of the Royal Society of London. Series B: Biological Sciences* 269 (1493): 793–99. https://doi.org/10.1098/rspb.2002.1955.

Holland, Barbara R., Katharina T. Huber, Andreas W. M. Dress, and Vincent Moulton. 2002. "δ Plots: A Tool for Analyzing Phylogenetic Distance Data." *Molecular Biology and Evolution* 19 (12): 2051–59. https://doi.org/10.1093/oxfordjournals.molbev.a004030.

Honkola, Terhi. 2016. *Macro- and Microevolution of Languages: Exploring Linguistic Divergence with Approaches from Evolutionary Biology*. Turku: University of Turku.

Honkola, Terhi, Kalle Ruokolainen, Kaj Syrjänen, Unni-Päivä Leino, Ilpo Tammi, Niklas Wahlberg, and Outi Vesakoski. 2018. "Evolution within a Language: Environmental

Differences Contribute to Divergence of Dialect Groups." *BMC Evolutionary Biology* 18 (1). https://doi.org/10.1186/s12862-018-1238-6.

Honkola, Terhi, Jenni Santaharju, Kaj Syrjänen, and Karl Pajusalu. 2019. "Clustering Lexical Variation of Finnic Languages Based on Atlas Linguarum Fennicarum." *Linguistica Uralica* 55 (3): 161. https://doi.org/10.3176/lu.2019.3.01.

Honkola, Terhi, Outi Vesakoski, Kalle Korhonen, Jyri Lehtinen, Kaj Syrjänen, and Niklas Wahlberg. 2013. "Cultural and Climatic Changes Shape the Evolutionary History of the Uralic Languages." *Journal of Evolutionary Biology* 26 (6): 1244–53. https://doi.org/10.1111/jeb.12107.

Honti, László. 1998. "ObUgrian". In *The Uralic Languages,* edited by Daniel Abondolo, 327-357. London: Routledge.

Hormia, Osmo. 1978. *Finska dialekter: En översikt.* Lund: Liberläromedel.

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

Huelsenbeck, John P., and Fredrik Ronquist. 2001. "MRBAYES: Bayesian inference of phylogeny." *Bioinformatics* 17, 754–55.

Huelsenbeck, John P., Fredrik Ronquist, Rasmus Nielsen, and Jonathan P. Bollback. 2001. "Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology." *Science* 294 (5550): 2310–14. https://doi.org/10.1126/science.1065889.

Hull, David L. 2010. *Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science*. Chicago: University of Chicago Press.

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

Huson, Daniel H., and David Bryant. 2006. "Application of phylogenetic networks in evolutionary studies." *Molecular Biology and Evolution* 23: 254–67.

Hurtta, Heikki. 1999. "Variaatiotutkimuksen myytit ja stereotypiat." In *Kirjoituksia sosiolingvistiikasta*, edited by Urho Määttä, Pekka Pälli, and Matti K. Suojanen, 53-101. Tampere: University of Tampere.

Hymes, Dell H. 1983. "Lexicostatistics and Glottochronology in the Nineteenth Century (with Notes toward a General History)." In *Essays in the History of Linguistic Anthropology*, 59–113. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/sihols.25.

Hyvönen, Saara, Antti Leino, and Marko Salmenkivi. 2007. "Multivariate Analysis of Finnish Dialect Data An Overview of Lexical Variation." Literary and Linguistic Computing 22 (3): 271–90. https://doi.org/10.1093/llc/fqm009.

Häkkinen, Jaakko. 2009. "Kantauralin ajoitus ja paikannus: perustelut puntarissa." *Journal de la Société Finno-Ougrienne* 92: 9-56.

Häkkinen, Kaisa. 1983. *Suomen kielen vanhimmasta sanastosta ja sen tutkimisesta*. Turku, Finland: University of Turku dissertation.

Häkkinen, Kaisa. 1984. "Wäre Es Schon Zeit, Den Stammbaum Zu Fällen? Theorien Über Die Gegenseitigen Verwandtschaftsbeziehungen Der Finnisch-Ugrischen Sprachen." *Ural-Altaische Jahrbücher, Neue Folge* 4: 1–24.

Itkonen, Terho. 1989. *Nurmijärven murrekirja.* Helsinki: Suomalaisen Kirjallisuuden Seura.

Itkonen, Terho. 1964. *Proto-Finnic Final Consonants. Their History in the Finnic Languages with Particular Reference to the Finnish Dialects. I:1 Introduction: The History of -k in Finnish*. Helsinki: Suomalaisen Kirjallisuuden Kirjapaino.

Jacques, Guillaume, and Johann-Mattis List. 2019. "Save the Trees: Why We Need Tree Models in Linguistic Reconstruction (and When We Should Apply Them)." *Journal of Historical Linguistics* 9 (1): 128–67. https://doi.org/10.1075/jhl.17008.mat.

Jakobsson, Mattias and Noah A. Rosenberg. 2007. "CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure." *Bioinformatics* 23(14): 1801–6.

Janhunen, Juha. 2000. "Reconstructing Pre-Proto-Uralic: Spanning the millennia of linguistic evo- lution." In *Congressus Nonus Internationalis Fenno-Ugristarum, vol. 1*, edited by Anu Nurk, Triinu Palo & Tõnu Seilenthal, 59–76. Tartu, Estonia: Eesti Fennougristide Komitee.

Janhunen, Juha. 2009. "Proto-Uralic - What, Where, and When?" *Mémoires de La Société Finno-Ougrienne* 258: 57–78.

Kaufman, Leonard and Peter Rousseeuw. 1987. "Clustering by means of Medoids." In *Statistical Data Analysis Based on the l1-Norm and Related Methods*, edited by Yadolah Dodge, 405–16. Amsterdam: North-Holland.

Kessler, Brett. 2001. *The significance of word lists*. Stanford: CSLI.

Kettunen, Lauri, Sheila Embleton and Eric S. Wheeler. 2021. "Murrekartasto, Lauri Kettunen". Kotimaisten kielten keskus. urn:nbn:fi:csc-kata20151130145346403821

Kettunen, Lauri. 1940a. *Suomen Murteet III A. Murrekartasto.* Helsinki: Suomalaisen Kirjallisuuden Seura.

Kettunen, Lauri. 1940b. *Suomen Murteet III B. Selityksiä Murrekartastoon.* Helsinki: Suomalaisen Kirjallisuuden Seura.

Kettunen, Lauri. 1930. *Suomen Murteet II. Murrealueet.* Helsinki: Suomalaisen Kirjallisuuden Seura.

Korhonen, Mikko. 1986. *Finno-Ugrian Language Studies in Finland 1828-1918*. Helsinki: Societas Scientiarum Fennica.

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

Kulonen, Ulla-Maija. 2002. "Kielitiede Ja Suomen Väestön Juuret." In *Ennen Muinoin: Miten Menneisyyttämme Tutkitaan*, edited by Riho Grünthal, 102–16. Helsinki: Suomalaisen Kirjallisuuden Seura.

Legendre, Pierre, and Louis Legendre. 2012. *Numerical Ecology*. Third English Edition. Amsterdam/Oxford: Elsevier.

Lee, Sean, and Toshikazu Hasegawa. 2011. "Bayesian Phylogenetic Analysis Supports an Agricultural Origin of Japonic Languages." *Proceedings of the Royal Society B: Biological Sciences* 278 (1725): 3662–69. https://doi.org/10.1098/rspb.2011.0518.

Leino, Antti, Saara Hyvönen, and Marko Salmenkivi. 2006. "Mitä Murteita Suomessa Onkaan? Murresanaston Levikin Kvantitatiivista Analyysiä." *Virittäjä* 110: 26–45.

Leino, Unni, Kaj Syrjänen, and Outi Vesakoski. 2020. "Linguistic change and biological evolution". In *Interdisciplinary Perspectives on the Philosophy and Science of Language,* edited by Ryan Nefdt, Carita Klippi, and Bart Karstens, 179–193. Cham: Palgrave Macmillan.

Lenneberg, Eric H. 1967. *Biological Foundations of Language*. New York: Wiley.

Leskinen, Heikki. 1992. *Karjalan kielisanasto 1. Idän ja lännen sanastoeroja.* Jyväskylä: Jyväskylän yliopisto.

Lewontin, Richard C. 1970. "The Units of Selection." *Annual Review of Ecology and Systematics* 1 (1): 1–18. https://doi.org/10.1146/annurev.es.01.110170.000245.

List, Johann-Mattis. 2016. "Beyond Cognacy: Historical Relations between Words and Their Implication for Phylogenetic Reconstruction." *Journal of Language Evolution* 1 (2): 119–36. https://doi.org/10.1093/jole/lzw006.

Livingstone, Daniel Jack. 2003. "Computer Models of The Evolution of Language and Languages." PhD dissertation, Paisley: University of Paisley.

Maechler, Martin, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. 2014. *cluster: Cluster Analysis Basics and Extensions.* r package version 1.15.3.

Marcus, Gary F., and Simon E. Fisher. 2003. "FOXP2 in Focus: What Can Genes Tell Us about Speech and Language?" *Trends in Cognitive Sciences* 7 (6): 257–62. https://doi.org/10.1016/S1364-6613(03)00104-9.

Matisoff, James. 2000. "On the uselessness of glottochronology for the subgrouping of Tibeto-Burman." In *Time depth in historical linguistics*, edited by Colin Renfrew, April

McMahon, and Larry Trask, 333-71. Cambridge: The McDonald Institute for Archaeological Research.

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

McMahon, April, Paul Heggarty, Robert McMahon, and Natalia Slaska. 2005. "Finding families: Quantitative methods in language classification." *Transactions of the Philological Society* 103(2), 147-70.

Michalove, Peter A. 2002. "The Classification of the Uralic Languages: Lexical Evidence from Finno-Ugric." *Finnisch-Ugrische Forschungen* 57: 58–67.

Mielikäinen, Aila. 1991. *Murteiden murros. Levikkikarttoja nykypuhekielen piirteistä.* Jyväskylä: Jyväskylän yliopisto.

Muller, Max. 1870. "Darwinism Tested by the Science of Language. Translated from the German of Professor August Schleicher." *Nature* 1 (10): 256–59. https://doi.org/10.1038/001256a0.

Nakhleh, Luay, Tandy Warnow, Don Ringe, and Steven N. Evans. 2005. "A Comparison of Phylogenetic Reconstruction Methods on an Indo-European Dataset." *Transactions of the Philological Society* 103 (2): 171–92. https://doi.org/10.1111/j.1467-968X.2005.00149.x.

Naser-Khdour, Suha, Bui Quang Minh, and Robert Lanfear. 2020. "Assessing Confidence in Root Placement on Phylogenies: An Empirical Study Using Non-Reversible Models." bioRxiv doi: https://doi.org/10.1101/2020.07.31.230144.

Nelson-Sathi, Shijulal, Johann-Mattis List, Hans Geisler, Heiner Fangerau, Russell D. Gray, William Martin, and Tal Dagan. 2011. "Networks Uncover Hidden Lexical Borrowing in Indo-European Language Evolution." Proceedings of the Royal Society B: Biological Sciences 278 (1713): 1794–1803. https://doi.org/10.1098/rspb.2010.1917.

Nerbonne, J., and W. A. Kretzschmar. 2013. "Dialectometry++." *Literary and Linguistic Computing* 28 (1): 2–12. https://doi.org/10.1093/llc/fqs062.

Nerbonne, John, and Martijn Wieling. 2017. "Statistics for Aggregate Variationist Analyses." In *The Handbook of Dialectology*, edited by Charles Boberg, John Nerbonne, and Dominic Watt, 1st ed., 400–414. Wiley. https://doi.org/10.1002/9781118827628.ch23.

Nichols, Johanna, and Tandy Warnow. 2008. "Tutorial on Computational Linguistic Phylogeny." *Language and Linguistics Compass* 2 (5): 760–820. https://doi.org/10.1111/j.1749-818X.2008.00082.x.

Novembre, John. 2016. "Pritchard, Stephens and Donnelly on Population Structure." *Genetics* 204: 391-393. https://doi.org/10.1534/genetics.116.195164.

O'Grady, Geoffrey N. 1960. "More on lexicostatistics." *Current Archaeology* 1, 338–339.

Pagel, Mark. 2009. "Human Language as a Culturally Transmitted Replicator." *Nature Reviews Genetics* 10 (6): 405–15. https://doi.org/10.1038/nrg2560.

Pagel, Mark, Quentin D. Atkinson, and Andrew Meade. 2007. "Frequency of Word-Use Predicts Rates of Lexical Evolution throughout Indo-European History." *Nature* 449 (7163): 717–20. https://doi.org/10.1038/nature06176.

Pappas, Panayiotis A., and Arne O. Mooers. 2011. "Phylogenetic Methods in Historical Linguistics: Greek as a Case Study." *Journal of Greek Linguistics* 11 (2): 198–220. https://doi.org/10.1163/156658411X600007.

Paunonen, Heikki. 2006. "Lounaismurteiden asema suomen murteiden ryhmityksessä." In *Kohtauspaikkana kieli – näkökulmia persoonaan, muutoksiin ja valintoihin*, edited by Taru Nordlund, Tiina Oinikki-Rantajääskö, and Toni Suutari, 249–68. Helsinki: Suomalaisen Kirjallisuuden Seura.

Paunonen, Heikki. 1991. "Till en ny indelning av de finska dialekterna." *Fenno-Ugrica Suecana* 10: 75–79.

Pereltsvaig, Asya, and Martin W. Lewis. 2015. *The Indo-European Controversy: Facts and Fallacies in Historical Linguistics*. Cambridge, United Kingdom: Cambridge University Press.

Pritchard, Jonathan K., Matthew Stephens, and Peter Donnelly. 2000. "Inference of population structure using multilocus genotype data." *Genetics* 155: 945–59.

Rantanen, Timo, Harri Tolvanen, Meeli Roose, Jussi Ylikoski, and Outi Vesakoski (ms). "Best practices for language distribution data harmonization, sharing and map creation – a case study of Uralic languages."

Rapola, Matti. 1969. *Johdatus suomen murteisiin.* 3rd. edition. Helsinki: Suomalaisen Kirjallisuuden Seura.

Rexová, Kateřina, Daniel Frynta, and Jan Zrzavý. 2003. "Cladistic Analysis of Languages: Indo-European Classification Based on Lexicostatistical Data." *Cladistics* 19 (2): 120–27. https://doi.org/10.1016/S0748-3007(02)00147-0.

Richerson, Peter J., and Robert Boyd. 2005. *Not by Genes Alone: How Culture Transformed Human Evolution*. Chicago: University of Chicago Press.

Ritt, Nikolaus. 2004. *Selfish Sounds and Linguistic Evolution: A Darwinian Approach to Language Change.* 1st ed. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511486449.

Ronquist, Fredrik, and John P. Huelsenbeck. 2003. "MRBAYES 3: Bayesian phylogenetic inference under mixed models." *Bioinformatics* 19, 1572–74.

Ronquist, Fredrik, Maxim Teslenko, Paul van der Mark, Daniel L. Ayres, Aaron Darling, Sebastian Höhna, Bret Larget, Liang Liu, Marc A. Suchard, and John P. Huelsenbeck. 2012. "MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space." *Systematic Biology* 61(3), 539–42. https://doi.org/10.1093/sysbio/sys029

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

Saarikivi, Janne. 2011. "Saamelaiskielet — nykypäivää ja historiaa." In *Saamelaistutkimus tänään*, edited by Irja Seutujärvi-Kari, Petri Halinen, and Risto Pulkkinen, 77–119. Helsinki: Suomalaisen Kirjallisuuden Seura.

Saarikivi, Janne, and Riho Grünthal. 2005. "Itämerensuomalaisten Kielten Uralilainen Tausta." In *Muuttuva Muoto: Kirjoituksia Tapani Lehtisen 60-Vuotisjuhlan Kunniaksi*, edited by Johanna Vattovaara, Toni Suutari, Hanna Lappalainen, and Riho Grünthal, 111–46. Helsinki: Helsingin yliopiston suomen kielen laitos.

Salminen, Tapani. 2007. "Europe and North Asia." In *Encyclopedia of the World's Endangered Languages*, edited by Christopher Moseley, 211–82. New York: Routledge.

Salminen, Tapani. 2002. "Problems in the Taxonomy of the Uralic Languages in the Light of Modern Comparative Studies." In *Лингвистический Беспредел: Сборник Статей к 70-Летию А. И. Кузнецовой*, edited by Александр Е. Кибрик, 44–55. Moscow: Издательство Московского университета.

Salminen, Tapani. 1999. "Euroopan Kielet Muinoin Ja Nykyisin." In *Pohjan Poluilla. Suomalaisten Juuret Nykytutkimuksen Mukaan*, edited by Paul Fogelberg, 13–26. Helsinki: Societas Scientiarum Fennica.

Sankoff, David. 1973. "Parallels between Genetics and Lexicostatistics." In Lexicostatistics in Genetic Linguistics, 64–74. The Hague: Mouton.

Savijärvi, Ilkka, and Eeva Yli-Luukko. 1994. *Jämsän äijän murrekirja*. Helsinki: Suomalaisen Kirjallisuuden Seura.

Schleicher, August. 1869. *Darwinism Tested by the Science of Language.* Translated by Alex V. W. Bikkers. London: John Camden Hotten.

Schleicher, August. 1863. *Die Darwinsche Theorie Und Die Sprachwissenschaft.* Weimar: Hermann Böhlau.

Sinor, Denis. 1988. "Introduction." In *The Uralic Languages: Description, History and Foreign Influences*, edited by Denis Sinor, xiii–xx. Leiden: Brill.

Swadesh, Morris. 1952. "Lexicostatistic dating of prehistoric ethnic contacts." *Proceedings of the American Philological Society* 96, 452-63.

Swadesh, Morris. 1955. "Towards greater accuracy in lexicostatistic dating." *International Journal of American Linguistics* 21, 121-37.

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

Tadmor, Uri. 2009. "Loanwords in the world's languages: Findings and results." In *Loanwords in the world's languages: A comparative handbook*, edited by Martin Haspelmath, and Uri Tadmor, 55–75. Berlin: Walter de Gruyter.

Tambovtsev, Yuri. 2004. "Uralic language taxon: Natural or artificial? (Typological compactness of Uralic languages taxons: Branches, subgroups, groups, families and superfamilies)." *Fenno-Ugristica* 26. 200-46.

Trask, Robert L. 2000. *The Dictionary of Historical and Comparative Linguistics*. Edinburgh: Edinburgh University Press.

Verkerk, Annemarie. 2019. "Detecting Non-Tree-like Signal Using Multiple Tree Topologies." *Journal of Historical Linguistics* 9 (1): 9–69. https://doi.org/10.1075/jhl.17009.ver.

Verkerk, Annemarie. 2016. "Phylogenies: Future, Not Fallacy." *Language Dynamics and Change* 7(1). https://doi.org/10.1163/22105832-00601013.

Vhaël, Bartholdus G. 1733. *Grammatica Fennica*. Åbo: Johan Kämpe.

Warelius, Anders. 1848. "Bidrag till Finlands kännedom i ethnographiskt hänseende." *Suomi* 7: 47–130.

Wichmann, Søren, Eric W. Holman, Taraka Rama, and Robert Walker. 2011. "Correlates of Reticulation in Linguistic Phylogenies." *Language Dynamics and Change* 1 (2): 205–40. https://doi.org/10.1163/221058212X648072.

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

Winter-Froemel, Esme. 2008. "Towards a Comprehensive View of Language Change: Three Recent Evolutionary Approaches." In *Paradox of Grammatical Change: Perspectives from Romance*, edited by Ulrich Detges and Richard Waltereit, 205–50. Amsterdam: John Benjamins Publishing Company.

Ylikoski, Jussi. 2016. "The Origins of the Western Uralic S-Cases Revisited: Historiographical, Functional-Typological and Samoyedic Perspectives." *Finnisch-Ugrische Forschungen* 63: 6–78.

PUBLICATIONS

PUBLICATION I

Shedding more light on language classification using basic vocabularies and phylogenetic methods: A case study of Uralic

Kaj Syrjänen, Terhi Honkola, Kalle Korhonen, Jyri Lehtinen, Outi Vesakoski & Niklas Wahlberg

Diachronica 30:3 (2013), 323-352
10.1075/dia.30.3.02syr

PUBLICATION II

Behind family trees: Secondary connections in Uralic language networks

Jyri Lehtinen, Terhi Honkola, Kalle Korhonen, Kaj Syrjänen, Niklas Wahlberg & Outi Vesakoski

Language Dynamics and Change 4:2 (2014), 189-221
10.1163/22105832-00402007

PUBLICATION III

Applying population genetic approaches within languages: Finnish dialects as linguistic populations

Kaj Syrjänen, Terhi Honkola, Jyri Lehtinen, Antti Leino & Outi Vesakoski

Language Dynamics and Change 6:2 (2016), 235-283
10.1163/22105832-00602002

PUBLICATION IV

Crouching TIGER, hidden structure: Exploring the nature of linguistic data using TIGER values

Kaj Syrjänen, Luke Maurits, Unni Leino, Terhi Honkola, Jadranka Rota & Outi Vesakoski

Journal of Language Evolution (forthcoming)