Language reconstruction methods put to the test

a comparison between traditional and computational phylogenetic reconstruction methods with data from Scandinavian language varieties

Stephan Eekman
University of Amsterdam
RMA Linguistics
14 October 2015
Supervisors: Prof. dr. A.P. Versloot
dr. R.A. Cloutier
Prof. dr. M. Dunn

Introduction

The genesis of historical linguistics is traditionally attributed to the Calcutta based English orientalist Sir William Jones, who in his lecture to the Asiatic Society on 2 February 1786 said the following about Sanskrit and the languages of Europe (Allan 2013:143):

"*The Sanskrit language, whatever its antiquity, is of a wonderful structure; more perfect than the Greek, more copious than the Latin, and more exquisitely refined than either; yet bearing to both of them a strong affinity, both in the roots of verbs and forms of grammar, than could possibly have been produced by accident; so strong, indeed, that no philologer could examine the Sanskrit, Greek and Latin, without believing them to have sprung from some common source, which, perhaps, no longer exists.*"

Although similar ideas had been put forward before Jones' lecture, this was the first time that the idea of a common ancestor of the languages of Europe and India was embraced by other linguists who further worked out the idea proposed by Sir William Jones. A revolutionary contribution to the development of historical linguistics was made by the linguist August Schleicher who was first to summarise the linguistic history of a language family in a tree model in 1853 (Allan 2013:154). The tree model is still used extensively in historical linguistics and has even been adopted into evolutionary biology.

Not much later, in 1872, the German linguist Johannes Schmidt presented a new model, the wave model, as a better alternative to the tree model (Allan 2013:163). Whereas the tree model suggests that after splitting events innovations can no longer spread between the split-off branches, the wave model projects linguistic innovations as circles that spread through the linguistic area. Innovations can occur in different places and have their own reach. In some cases these circles will overlap, which denotes that the variety in the overlapping area has introduced innovations from several different origins. This idea of language evolution seems to be more in line with what happens in the real world, but it has never replaced or become as popular as the tree model as has two major disadvantages. First, wave models are much harder to interpret and second they do not show time scales as clearly as tree models.

Around the same time, a more systematic approach in historical linguistics was introduced by the Neogrammarians (Allan 2013:160). They are responsible for the idea of phonological rules, which state that phonological innovations occur in all words that contain the particular phonological environment that causes the innovation. They considered these phonological rules to be without exception, a view which has since been abandoned. The language family histories known at present are still by and large inferred from sound changes and the same applies to much of our knowledge about proto-languages.

Despite the fact that sound changes and phonology were most commonly used for the reconstruction of language phylogenies, other parts of language were heeded as well albeit to a lesser extent. However, whatever part of language was used in the reconstruction, the methods applied were always qualitative and the final outcome based on judgments by the historical linguist. In an attempt to introduce a quantitative method into historical linguistics in order to avoid subjective judgements affecting the outcome, Morris Swadesh (1952) invented what has generally become known as lexicostatistics. In this method, the linguist compiles a list of 100 or 200 culturally independent meanings for the languages to be investigated and subsequently calculates the lexical distance between the languages. Later, Swadesh even proposed a fixed rate of loss of lexemes that would enable the linguist to extend the lexical distance to chronological distance between languages. This method bears the name glottochronology. No such rate seems to exist, however, as Bergsland and Vogt (1962) for example pointed out for Icelandic and Norwegian. They found that Icelandic had a substitution rate of 4% per

millennium, whereas Norwegian had one of 20%. The validity of glottochronology was also questioned and criticised by other linguists (Fodor 1961, Chretien 1962 & Guy 1980). The obvious flaws and shortcomings of the method eventually led to its abandonment by the majority of the linguistic community.

In recent times, a renewed interest for quantitative methods in historical linguistics has emerged. A number of new methods have been designed, one of which is computational phylogenetics. Computational phylogenetics is a group of statistical methods that have been adopted from evolutionary biology (Nichols & Warnow 2008:760). In essence these methods calculate the similarity between languages and convert this similarity into some type of model, usually a tree model. The statistics used are methodologically reliable, but the question remains how the splits in the model relate to events in the real world. In contemporary evolutionary biology, DNA is most often used as the data on which the phylogeny is calculated. This has shown to be much more reliable than comparing the physical characteristics of species, as shared physical features can easily have been produced by chance. In order to extend these methods to the study of language, one needs to find out how to capture the DNA of language, if that even exists.

Ancestors pass on their DNA to the next generations and the same usually applies to their language. In this process of transfer only small changes occur. Even though, the process of genetic and linguistic evolution are very similar, some differences do exist. In linguistic evolution, linguistic features from an L2 are commonly adopted into the L1. Such type of transfer is very unusual in biological evolution. A herd of cows living among a herd of zebras will never develop stripes because of their contact with the zebra community, but the speakers of a Celtic language in contact with speakers of Romance languages could easily adopt Romance features into their Celtic mother tongue. So, linguistic features are not only transferred vertically, from generation to generation, but also horizontally, between distinct linguistic entities. In order to uncover the genetic history of a language phylum, one needs to find a method to filter out the horizontally transferred material from the vertically transferred material, as horizontally transferred material blurs the genetic signal. Not doing this properly might lead to incorrect conclusions (e.g. English is a Romance language, because of the high number of Romance loan words in the language).

In this study databases with different types of linguistic data (lexical, phonological, morphosyntactic and combined) will be compiled for Scandinavian language varieties, standard languages and dialects, and trees will be deduced from each of these databases. The outcome of each analysis will be compared to the tree that has been reconstructed by historical linguists through qualitative research. The main objective of the comparison of the outcome of these two reconstructive methods is to enhance the validity and reliability of both. Questions to be asked then are the following: What type of data gives the best results in computational phylogenetic analysis and what structures do the other types of data reveal? Does the phylogenetic analysis uncover events in the history of the language family that traditional methods could not reveal?

The Scandinavian languages are used in this study as a test case because their history is well documented as well as the history of their peoples. The oldest written sources in Scandinavian stem from the Proto-Nordic period so the entire history of this branch of Germanic is attested in written material (Barnes 2012). Research investigating the validity of computational phylogenetic analysis of language has been conducted before, but mostly on Indo-European as a whole (Gray & Atkinson 2003, Nakhleh et al. 2005, Nichols & Warnow 2008, Bouckaert et al. 2012, Chang et al. 2015). These studies show that these methods are capable of recognising the distinct branches of Indo-European, such as Slavic, Romance and Germanic, but if that is the highest level of accuracy attainable by computational phylogenetics one could wonder if it really is an asset to the traditional reconstructive methods. Dissecting the distinct branches of Indo-European from the pool of Indo-European languages is relatively easy

to accomplish considering that the linguistic distances between languages within a clade are much smaller than those between clades. The focus on the Scandinavian languages will challenge the phylogenetic methods to detect more recent relations showing less variation in the modern languages as there has been less time for the languages to diversify. Also constantly changing internal social relations pose a challenge for uncovering the factual history of the languages of Scandinavia.

Moreover, some scepticism towards the validity of computational phylogenetics in general seems to exist within the linguistic community. Pereltsvaig and Lewis (2015) discussed some fallacies and controversies of this method. One of their main concerns is the effect that shared retentions could have on the outcome of an analysis. Their observation is indeed correct, but shared retentions are not expected to pose a problem in most analyses, as disturbing shared retentions are usually randomly distributed over all the languages in the sample. The role of shared retentions is discussed more elaborately in the discussion section of this study.

Another shortcoming of computational phylogenetics is the exclusive dependence on lexical data, according to Pereltsvaig and Lewis. This is not an issue in this study as phonological and morphosyntactic data are tested as well. Pereltsvaig and Lewis are also reluctant about the way the lexical data are collected and coded. They mention for example that some lexemes have acquired new meanings (such as hound in English) or that some lexemes were lost as independent words, but did survive in compounds (such as were in werewolf). They are not sure how these should be coded. In order to treat the data for every language equally, strict rules were taken into account during the collection and coding procedures. These rules are explained in detail in the methodology section.

If the outcome of this study is such that the computational phylogenetic analysis shows a pattern which is compatible with reconstructions from historical linguistics, it might not only enhance the validity and reliability of both methods, but it will also show how phylogenetic analysis of language can be applied to language families of which the history is poorly documented or not documented at all. Moreover, it might even reveal social or linguistic events in the history of Scandinavian that hitherto have been unknown. Lastly, representational tools and models will also be compared. Both tree and network models (NeighborNet split graphs) will be produced and compared.

The paper is arranged as follows. First, the history of the Germanic languages as it has been reconstructed by historical linguists will be presented with a particular focus on the North Germanic languages. At the end of this section, the history of this branch of Indo-European is summarised in a tree that will serve as a tool for comparison. The next section deals with the methodology of the phylogenetic analysis in this study. The first part of the section presents the language sample that is used. Then, the different databases will be presented and the coding procedures will be explained. The next part of the section explains how the data are analysed and converted into a tree. The third section consists of an exploration of the data by means of wavelike models called NeighborNet split graphs. In the next section the results of the analysis, i.e. the tree models, are presented and briefly discussed. Finally, the main questions of this study are discussed in detail in the discussion section at the end.

METHODOLOGY

The first section of this chapter gives a brief overview of the history of the Germanic languages, as it is reconstructed by historical linguists. Dialects, except for Gutnish, are usually not included in descriptions of the history of the Germanic languages, so they are not included in this overview either. In most cases it is safe to consider dialects as later divergences within the branch of the standard language. The tree presented in this section will serve as a reference tree for the computational phylogenetic analyses in this study.

In the next section the language sample of the computational phylogenetic analyses in this study is presented and the selection procedure is explained. Subsequently, the databases employed are discussed as well as how the data were coded. Then, the procedure of how the data are converted into a tree structure is explained. In the last part of this chapter, the data are explored by means of NeighborNet split graphs.

Historical linguistics: The history of the Germanic languages

The languages in this study all belong to the Germanic branch of the Indo-European language family. How the development from Proto-Indo-European into Proto-Germanic occurred is still a topic of discussion. Consensus has been reached on the location of the homeland of Germanic, as well as when it first was spoken. The homeland of Germanic is traditionally believed to be the southern parts of Scandinavia and northern Germany (Kausen 2012:109; Ramat 1998:381; Speyer 2007:16). Here, the Germanic people settled around 1000 CE. During the last half of the first millennium CE, the Germanic tribes started to expand, initially most south- and eastwards. At this point the Germanic language area was still a dialect continuum.

The first group of languages to leave the Germanic dialect continuum were the East Germanic varieties. About 200 CE the speakers of these dialects started to move to the southeast to the lands of present-day Ukraine, Romania and Bulgaria. The best known member of the East Germanic branch is Gothic, which we know mostly from Wulfila's bible translation from the 4th century AD. Only little is known of the other members of the East Germanic sub-family. The East Germanic branch probably went extinct around 1700 AD with the death of Crimean Gothic.

The other branches of Germanic, traditionally known as North and West Germanic remained within the dialect continuum and consequently formed a separate Northwest Germanic branch in opposition to the East Germanic branch. The Northwest Germanic dialect continuum probably ceased to exist when the Angles colonised Britain from the south of Jutland in the 5th century AD. The massive migration of the Angles caused a gap in the Northwest Germanic dialect continuum (Kuhn 1955). This inhibited linguistic innovations from spreading through the entire continuum, which eventually led to the split between the West and North Germanic branches around 500 AD.

The development of the West Germanic languages is quite complex as splits were mostly gradual instead of abrupt and later migrations blurred the original partitions even more. As the West Germanic languages are not the focus of this study, their development will not be discussed in more detail, and it suffices to know that they belong to a different branch of Germanic than the North Germanic languages.

The common ancestor of all North Germanic languages, Proto-Nordic, was originally spoken in southern Sweden, southern Norway and Denmark. The first split within the North Germanic branch occurred in the 7th century between the West and East Scandinavian languages. West Scandinavian was spoken in the western parts of Norway and East Scandinavian in Sweden and Denmark. Also here, the split was not abrupt but gradual, so

varieties in the border region of the two branches might exhibit features of both. During the Viking pillages in the subsequent centuries, West Scandinavian spread to parts of Britain, Orkney and Shetland, the Faroe Islands, Iceland, Greenland and possibly even North America. The Vikings from the eastern lands spread their language to the shores of the Baltic Sea, the inlands of Russia and also parts of Britain.

Besides West and East Scandinavian, some scholars have also considered Gutnish, the language of the island of Gotland in the Baltic Sea, to be a separate branch of North Germanic, whereas others treat it as a sub-branch of East Scandinavian. The Old Gutnish language does in fact lack a number of typical East Scandinavian innovations and has undergone some typically Gutnish changes as well. The special position of Gutnish is probably a result of the fact that it is spoken on an island, so the contact between Gutnish and the mainland varieties might have been limited, which makes it conceivable that it split off from Proto-Nordic as a separate branch.

West Scandinavian split off into western Norwegian varieties, Faroese, Icelandic and Norn, the language of Orkney and Shetland. As stated before, West Scandinavian spread from western Norway to the islands of the Atlantic Ocean. This linguistic expansion occurred in the 9th century. Norn is very sparsely documented, so very little is known about its development. The same also applies to a lesser degree to Faroese, older stages of which have hardly been written down. When it comes to Icelandic, the situation is completely different. Icelandic has had a lively literary tradition since about the beginning of the second millennium. Most of the documents written in Old Norse come from Iceland. Although the Atlantic islands were already colonised in the 9th century, the distance from the homeland did not lead to an immediate split. According to Kuhn (1955) a number of innovations occurred in the entire West Scandinavian language area after the colonisation of the islands in the Atlantic Ocean. From about the 11th century, the languages of Norway started to develop more into the direction of the East Scandinavian languages. At present, Norwegian is part of a dialect continuum running from the Danish-German border through Sweden to the west coast of Norway. The insular varieties, Faroese and Icelandic, are not part of this continuum as they are not mutually intelligible with their neighbouring varieties.

East Scandinavian split up into two languages, Danish and Swedish. In the beginning, the split was gradual, but nowadays a clear border can be drawn through The Sound, the strait that separates Denmark from Sweden. It is hard to say when Danish and Swedish started to develop in separate ways, but most estimates lie in the same period: 11th century (Kausen 2002:111), 1200 (Speyer 2007:22) and 1300 (Harbert 2007:19). Bokmål Norwegian is often considered to be East Scandinavian as well, but that is only partly correct. Norway was part of the Danish kingdom until 1814 and the standard written language was standard Danish until then. The spoken language of the Norwegian elite around that time can be described as written Danish with East Norwegian pronunciation and a number of Norwegian features. After Norwegian independence from Denmark, several efforts were made to construct a new Norwegian written standard. One of these standards, Bokmål, was based on the speech of the Norwegian elite who spoke a Danish-Norwegian koiné-like language. Bokmål Norwegian can therefore be considered a descendant of standard Danish, with some Norwegian material and might rather be classified as a hybrid or patchwork of West and East Scandinavian. Nynorsk, on the other hand, was not based on the Danish standard but based on the dialects of West Norway, which were less affected by the Danish standard. Therefore, Nynorsk is classified as a "genuine" West Scandinavian variety.

As the above description of the history of the Germanic languages makes clear, most splits started gradually instead of being abrupt, and some older splits were even blurred by later migrations or developments, for instance. This means that the Germanic linguistic history is not well illustrated in a tree that assumes abrupt splits, which is also pointed out by several scholars (Ramat 1998:385; Kausen 2002:109). Nevertheless, trees are still commonly used to give an

overview of the history of the Germanic languages. In Figure 1, the linguistic history of Germanic is summarised in a tree. As the East and West Germanic branches are of little interest for this study they are not fully given.

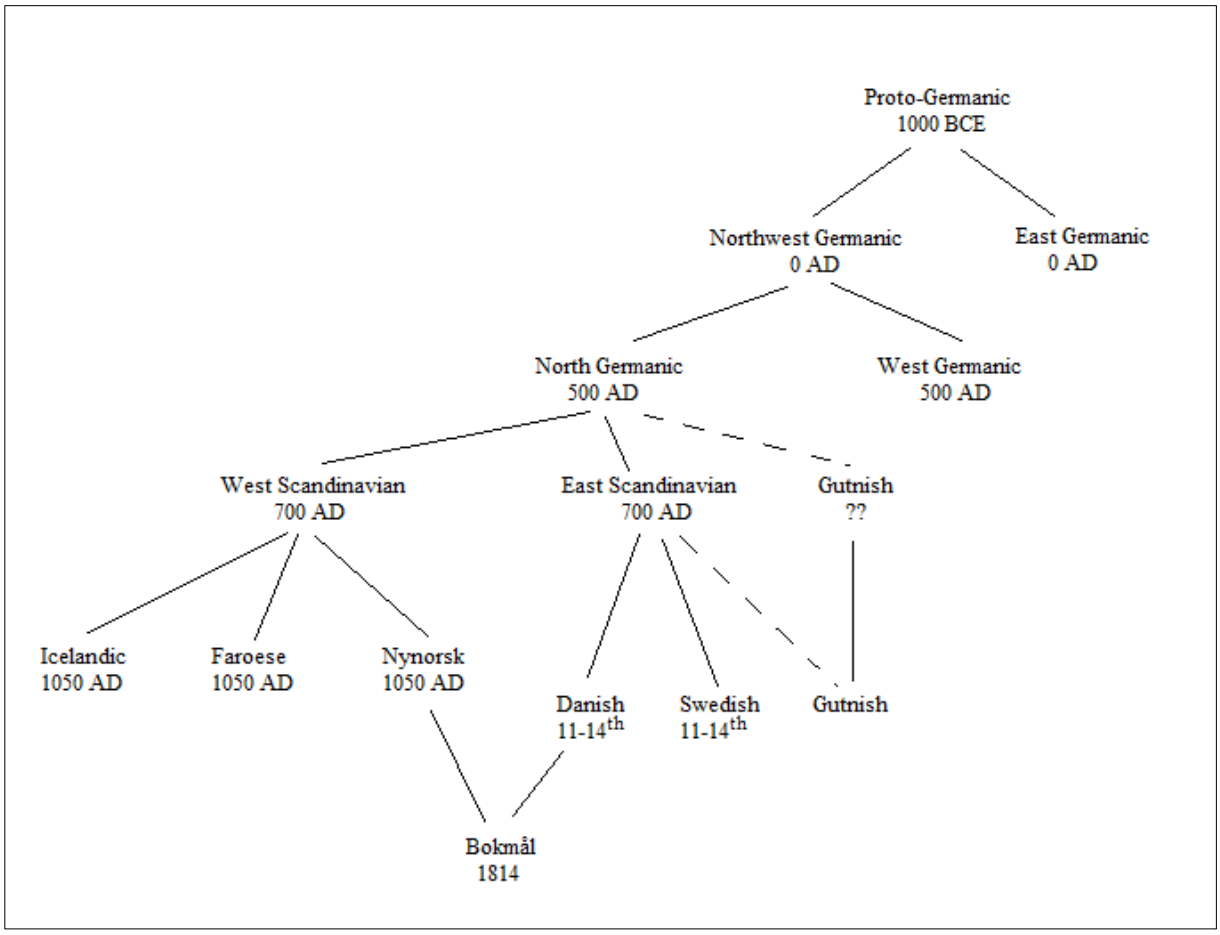

FIGURE 1: The family tree of the Germanic languages. Disputed splits are indicated by dashed lines.

The tree in Figure 1 will serve as the tree that represents the history of the Scandinavian languages as it has been reconstructed by historical linguists. Hence, it is the tree to which the phylogenetic trees will be compared. It should be noted that only standard languages and Gutnish are mentioned in the tree in Figure 1. As stated before, any other dialects can be considered to be later divergences in the branch of the standard language.

Computational Phylogenetic Analysis: Language sample

The dataset consists of data for sixteen Germanic languages. Fourteen of the languages belong to the North Germanic branch. The other two are West Germanic. Three of the North Germanic languages represent older stages of North Germanic. The contemporary languages are listed in Table 1, according to their traditional classification.[1] The ancient languages are listed in Table 2. The map in Figure two shows where the modern North Germanic varieties are spoken.

[1] The Danish dialect of Fjolde is no longer spoken, but as it only went extinct a few decades ago, it is still listed as a contemporary language.

TABLE 1. The twelve contemporary languages in the sample. A thick horizontal line denotes a split on the higher level, whereas a thin line indicates a split on the lower level. A dotted line represents a disputed split.

TRADITIONAL CLASSIFICATION		LANGUAGE	ABBREVIATION
NORTH GERMANIC	WEST SCANDINAVIAN	Icelandic	IC
		Faroese	FO
		Norwegian Stavanger	NO_S
		Norwegian Nynorsk	NO_N
	WEST & EAST	Norwegian Bokmål	NO_B
	EAST SCANDINAVIAN	Övdalian/Elfdalian	OEV
		Swedish	SE
		Danish	DK
		Danish Bornholm	DK_B
		Danish Fjolde	DK_F
	GUTNISH	Gutnish Lau	GU
WEST GERMANIC	ISTVAEONIC	Dutch	NL
	INGVAEONIC	West Frisian	FR

TABLE 2. The three ancient North Germanic languages in the sample, accompanied by an approximation of when they were spoken and a list of their modern descendants in the language sample. A language in brackets indicates that it might have already split off in a stage preceding the ancient language.

LANGUAGE	ABBREVIATION	ERA	MODERN DESCENDANTS
Old (West) Norse	ON	900-1200 A.D.	West Scandinavian
Old Swedish	OS	1225-1375 A.D.	Swedish, (Övdalian/Elfdalian)
Old Gutnish	OG	????-1600 A.D.	Gutnish Lau

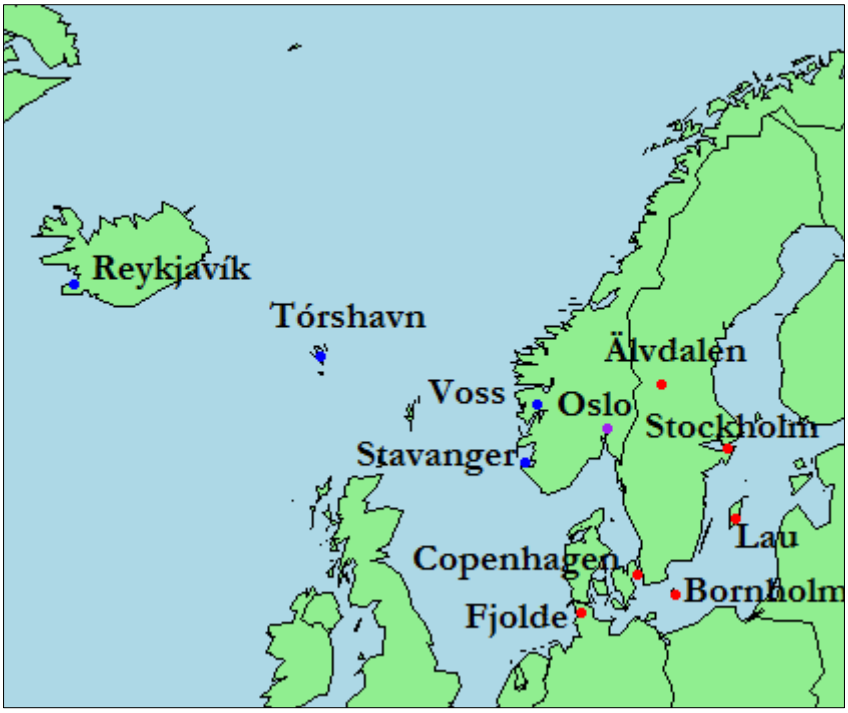

FIGURE 2. A map of the locations of the North Germanic languages in the sample. Standard languages are put in the capital of the country. For Nynorsk, Voss is chosen as reference city. Red dots stand for East Scandinavian languages, blue ones for West Scandinavian and purple means a combination of West and East Scandinavian.

The languages in the sample are selected in such a way that each sub-branch of North Germanic - West Scandinavian, East Scandinavian and Gutnish - is represented by an ancient language and at least one modern language. On the level of individual languages, the standard language was selected as well as one or two of the linguistically most distant dialects with respect to the standard variety. The standard languages are chosen, because their history has been extensively described and the data for these varieties are readily available. The most aberrant dialects are chosen in order to get a sample as linguistically diverse as possible, but also because their classification might be problematic in some way. The ancient languages are included to calibrate the time scale of the tree. Moreover, it will also be interesting to see if the ancient languages are assigned to the same clade as their modern descendant(s). The West Germanic languages are in the sample to define the root of the tree. In the subsequent paragraphs the selection process will be discussed in more detail.

The branch of West Scandinavian languages is represented by the insular Scandinavian languages, Icelandic and Faroese, as well as western Norwegian varieties and their common ancestor Old Norse. The insular Scandinavian languages are spoken by relatively small populations, compared to the other North Germanic languages, so the dialectal variation is limited and mostly confined to the area of phonology. Therefore, only the standard languages are included. The case of Norwegian is rather exceptional as the southwestern dialects are traditionally classified as belonging to the West Scandinavian branch, whereas the other dialects are considered East Scandinavian or a mixture of the two. The West Scandinavian branch of Norwegian is represented by the dialect of Stavanger and the regiolect known as Nynorsk. The classification of Stavanger and Nynorsk Norwegian might be challenging for the method, as these vernaculars have been under constant influence of the East Scandinavian languages. As a result, these present-day varieties are much closer to East Scandinavian than to the other West Scandinavian languages. Therefore, Norwegian is an interesting case for testing the method.

East Scandinavian consists of two standard languages - Swedish and Danish - and possibly one historical language: Gutnish. The Swedish clade is represented by Swedish, Övdalian (also known as Elfdalian) and Old Swedish. Övdalian is the dialect spoken in the Älvdalen municipality in the remote northwest of the Dalecarlia region. Övdalian is a very peculiar dialect as has been shown by previous research. Dahl (2005) showed that the lexical distance between standard Swedish and Övdalian is even larger than between Swedish and the other East Scandinavian standard languages and comparable to the lexical distance between Swedish and modern Icelandic. Övdalian exhibits many archaisms in both lexicon and morphosyntax and is consequently much closer to the more archaic West Scandinavian languages in many respects compared to the other contemporary East Scandinavian languages. Therefore, it will be interesting to see in which sub-family it will be grouped.

Old Swedish is the ancient language in this clade and the ancestor of modern Swedish. It might also have been the ancestor of Övdalian, but there is no consensus on when Övdalian started to develop as an independent language and split off from Swedish. The off-branching of Övdalian is believed by some scholars to have occurred even before the Old Swedish period, so in that case it would not have been a descendant of Old Swedish (Levander 1925; Ringmar 2005).

The twigs on the Danish branch of the tree are standard Danish and the dialects of Fjolde and Bornholm. The dialect of Fjolde was until recently spoken in the parish of Viöl in Schleswig Holstein, Germany (Bjerrum & Bjerrum 1974). It was the southernmost Danish dialect and completely surrounded by a German speaking population. Therefore, the dialect developed in a rather different way than other Danish dialects and is full of German loanwords. The dialect of Bornholm is spoken on the island of Bornholm which is located in the Baltic Sea approximately 150 kilometres east from the rest of Denmark. Its proximity to the Swedish mainland has led to a number of shared features with Swedish which are absent in other Danish dialects. I included two dialects of Danish, as the grammars of neither of them contained enough data for all the databases. Usually I included the dialect for which I had the most data, but when I had enough data for both dialects I included both of them. This did not affect the outcome of the analysis in any serious way.

Norwegian Bokmål is the written standard of Norwegian that is based on Danish. Although several adaptions have been made to Bokmål over the years to make it closer to the spoken language, it is still by and large a continuation of Danish. Therefore, Bokmål contains both West and East Scandinavian traits, which will make it interesting to see where it will show up on the tree.

The Gutnish branch comprises Old Gutnish and the Gutnish spoken in the village of Lau. As already mentioned, no consensus has been reached so far on the position of Gutnish in the family tree. Some scholars consider it a sub-branch of East Scandinavian, whereas others treat it as a branch on its own situated alongside Old West and Old East Norse. So, it will be interesting to see which of these views is supported by the analysis. The present-day dialect of Lau might be just as interesting as its ancestor in the analysis. As Gotland has been a part of Sweden since 1645, the vernaculars on the island have been heavily influenced by the Swedish standard language. Most of the Gutnish vernaculars are even considered to be dialects of Swedish nowadays. Therefore, it might be challenging for the analysis to recognise the Gutnish origin of the vernacular of Lau.

Dutch and West Frisian do not belong to the North Germanic languages, but to the closely related West Germanic languages. As they are in another sub-family of Germanic, the first split is always between them and all the North Germanic languages. In this way, the tree can be rooted neatly.

Computational Phylogenetic Analysis: Databases and character coding

In this study, four types of databases are used. The first three databases contain lexical, phonological and morphosyntactic features and the fourth one combines items from all the aforementioned databases in a combined database. In this chapter, each of the databases is discussed. It will provide the reader with information on how the items in the individual databases were selected, what problems were encountered in the process of data collection and what sources were used. It will also show how the data in the databases are converted into the right format for the analysis.

Lexicon

The lexical database consists of two sub-databases. The template for the first lexical sub-database is the extended Swadesh list which contains 207 meanings. Swadesh (1952) constructed this list to be able to map linguistic diversification by means of quantitative data. The list is thought to consist of basic concepts that are culturally independent and hence applicable to all of the world's languages. Although the reasoning behind the Swadesh list seems sound, it has not sufficiently been tested whether a list of this type gives better results than lists which consist of culturally dependent concepts. So, the template for the second lexical database is a list of vocabulary concerning domestic animals, which is thought to be culturally dependent. The list is constructed especially for this study.

The lexical data were retrieved from English-Scandinavian dictionaries. These translations were double checked in monolingual Scandinavian dictionaries. Occasionally, native speakers were consulted for doubtful cases. For the dialects, standard language-dialect or dialect-standard language dictionaries were consulted. In the cases of Old Swedish and Old Gutnish, ancient language-Swedish dictionaries were used. Unfortunately, there are no Swedish-Old Swedish/Gutnish dictionaries, so a number of uniquely Old Swedish or Old Gutnish lexemes might be missing in the database. However, this hardly affects the analysis as the only information in these lexemes would be that these languages are different from all the others. For Old Norse, Old Norse-English and Old Norse-Scandinavian dictionaries were used. Also, a digital Old Norse-English dictionary was consulted, which made it possible to search through the dictionary using the English entry, so that etyma unique to Old Norse could also be added to the database.

When all the needed lexical data were collected, the items were coded for cognacy.[2] Lexemes are cognates if they derive from the same historical stem. This does not necessarily mean that the words have to be identical, but only that they need to have developed from the same word in the proto-language. Each set of cognates belongs to a separate cognate class and is assigned a certain value. For one meaning, several cognate classes can be available and languages are allowed to have more than one lexeme for each meaning. A certain cognate class might occur more than once in the database, when it is used for more than one meaning on the list. In such cases the cognate classes are treated as different items in the analysis. Table 3 illustrates the cognate classes for the meaning 'cloud'.

[2] Cognacy data were retrieved from De Vries (1961) for Scandinavian and Philippa (ed.) (2003-2009) for Dutch. The cognacy judgments for the dialects are my own.

TABLE 3. The cognate classes for the meaning ‘cloud’.

	*skiwjan	*wulkan, *wulkōn	moln
ON	ský		
OS	sky		
OG			
IC	ský		
FO	skýggj		
NO_S	sjy		
NO_B	sky		
OEV			muoln
SE	sky		moln
GU	sköi		(måuln)
DK	sky		
DK_F	skøj		
NL		wolk	

There are three cognate classes available for the meaning ‘cloud’. Two stem from at least the Proto-Germanic period. Their Proto-Germanic reconstructions are listed at the top of the table. The third class goes back to an innovation that has occurred in the Swedish language area. There, a new lexeme for cloud was formed from the adjective *mulen* ‘cloudy’.

Then the data are translated into binary character states. Each cognate class is represented by one character. A character can have one of two states, either 1 or 0. The value 1 is assigned if a language has a lexeme belonging to the cognate class for the meaning in question, whereas the value 0 is assigned if it does not. This seems pretty straightforward but there are a few problems to take into account. As is shown in table 3, not a single lexeme for cloud has been attested in Old Gutnish. As it lacks cognates for every cognate class, it should get a 0 for each character. This might, however, not be the best way to deal with this. It is actually very conceivable that Old Gutnish had a lexeme belonging to one of these cognate classes. Assigning 0 to all these classes would set Old Gutnish apart from all the other languages, as it does not share a cognate with any of the other languages, although this might not have been the case. In order to avoid this problem, missing data are assigned the value ‘?’. When a character is assigned the value ‘?’, it is treated as either present or absent in the analysis. So, no distance measures will be made between Old Gutnish and the other languages based on the meaning ‘cloud’.

A second problem is illustrated by Gutnish. One of the Gutnish words for cloud is *måuln*. This word is in round brackets in the table as it is a loan from Swedish, as the proper Gutnish form would have been *muln*. Loan words most often do not give any information about the genetic history of a language, but rather about the contact situations it has been in. As the main purpose of this analysis is to reconstruct the genetic history of North Germanic, loan words are assigned 0 as well. Nevertheless, a separate analysis is run as well in which loan words are also taken into account. Table 4 illustrates how the cognate classes are translated into binary states.

TABLE 4. Binary coding of the cognate data for the meaning cloud.

	COGNATE CLASS		
	A	B	C
ON	1	0	0
OS	1	0	0
OG	?	?	?
IC	1	0	0
FO	1	0	0
NO_S	1	0	0
NO_B	1	0	0

	COGNATE CLASS		
	A	B	C
OEV	0	0	1
SE	1	0	1
GU	1	0	0
DK	1	0	0
DK_F	1	0	0
NL	0	1	0

A few additional remarks about the selection of words should be made. In order to optimise the lists, a few conditions were taken into account in the selection procedure. Each meaning on the list is clarified by a short description. The complete word list with descriptions can be found in the appendix. Such a description could state that the desired word describes the concept as neutrally as possible. In those cases, a word was not allowed to have any connotations and should have the same semantic reach as the English word, or as close as possible. A good example of this is the concept 'bone', which is accompanied by the description '*general term for human and animal bones (anatomical)*'. Words for joint or specific bones are thus ruled out.

Sometimes semantic shifts have occurred and two cognates might no longer have the same meaning. In Swedish, for example, the word *snok* is cognate to English *snake* and Old Norse *snákr*. Originally this word could be used to describe any type of reptile belonging to the suborder *serpents*, just like English *snake*. The Swedish word *snok*, however, has acquired a more specific meaning and is now used to describe types of snakes belonging to the family *colubridae*. The description for the concept snake reads '*generic term for snake'*. This is exactly what the Old Norse *snákr* meant as well as its English cognate. Its Swedish cognate, however, has a narrower semantic scope and is therefore excluded from the list, even though it is a cognate of Old Norse *snákr* in the traditional sense.

In other cases, the description stated that the concept had to be applicable to as many cases as possible. The concept leaf is described as '*applicable to as many plants as possible*', for example. North Germanic languages usually have two words for English 'leaf', such as ON *lauf* and ON *blað*. They are not synonyms, however. The word *lauf* can only be used for leaves from deciduous trees whereas *blað* is also applicable to petals and cabbage leaves for example. Consequently, only *blað* was included in the list as it is applicable to more types of plants.

In some cases, the opposite strategy was used and the description narrowed down the scope of a concept to a very specific meaning. For 'rotten', the description reads '*of trees*', which indicates that only words meaning 'rotten of trees' are included in the list. Words related to ON *rutinn* 'rotten in general' are therefore excluded from the list in favour of more specialised terminology, such as ON *morkinn*, ON *meyrr*, OEV *fuoskun*, SE *multen*, DK_F *mosk* and NL *vermolmd* all meaning specifically 'rotten of wood'.

A special remark should be made about the list for NO_S. The source for this dialect is a short grammar (Berntsen & Larsen 1925). The word list at the end of the grammar contains all the words that occurred in the grammar. So, it is by no means a complete dictionary of this dialect. In some cases, the word list in the grammar gives one word for a certain meaning, while Nynorsk, the written language that is closest to the Norwegian spoken in Stavanger, has other words for it as well. In such cases a '?' is assigned to the cognate classes that are found in Nynorsk but are not mentioned in the dictionary, as it is very conceivable that the word might be used in the dialect as well. The reason to do this is that the distance between NO_S and the other varieties otherwise might turn out larger than it in fact is. Following this procedure, fifteen

‘?’s are added to the list. When no cognate class is attested for a meaning all, ‘?’ is assigned to all cognate classes, as is the usual procedure. For the list of domestic animals, only Nynorsk was included as only very few terms could be retrieved from the grammar.

In two cognate classes, a morphological development had occurred in some of the languages. This has happened in one of the cognate classes for the meaning ‘nose’. An overview of all cognates in this cognate class is shown in Table 5.

TABLE 5. The lexemes for nose stemming from Proto-Indo-European **nās-*, **néh$_2$s-* with morphological glossing.

COGNATE CLASS STEMMING FROM PIE **nās-*, **néh$_2$s-*					
ON	nǫs-ø	nose-NOM.SG	OEV	-	-
OS	näs-a	nose-NOM.SG	SE	näsa	nose
OG	nas-ar	nose-NOM.PL	GU	nas-ar	nose-PL
IC	-	-	DK	næse	nose
FO	nøs-ø	nose-NOM.SG	DK_F	nees	nose
NO_S	nase	nose	NL	neus	nose
NO_B	nase	nose			

There are two deviant groups in this cognate class. The first group are Icelandic and Övdalian which both have lost this cognate. The second one, and the more interesting one, are Old Gutnish and Gutnish. Contrary to all the other languages, they have a *plurale tantum* for this meaning. According to the methodology, this cognate class is translated into only one character. This means that the information that is revealed by the shared plural ending in Old Gutnish and Gutnish would be ignored. Therefore, this cognate class is translated into two characters. One character for the presence or absence of a cognate stemming from **nās-*, **néh$_2$s-*, and one for the presence or absence of the plural ending. The binary coding for this cognate class is illustrated in Table 6.

TABLE 6. Binary coding for the cognate class **nās-*, **néh$_2$s-*.

	ON	OS	OG	IC	FO	NO_S	NO_B	OEV	SE	GU	DK	DK_F	NL
**nās-*, **néh$_2$s-*	1	1	1	0	1	1	1	0	1	1	1	1	1
plural ending	0	0	1	0	0	0	0	0	0	1	0	0	0

A similar type of innovation has occurred in one of the cognate classes for ‘to breathe’. The Old Norse cognate in this class is *anda*. This cognate is found in all of the North Germanic languages, with the exception of Old Gutnish for which no lexeme for ‘to breathe’ is attested. In the Swedish language area, however, something peculiar has happened with this lexeme. In Swedish, *anda* has been replaced by *anda-s* breathe-MID, with an ending denoting middle voice.[3] The same form is found in Övdalian and in Gutnish we find *andäs*, which might be a loan from Swedish. Ironically, the same form with the middle voice ending, is also found in the West Scandinavian languages (ON *anda-sk*; IC/FO/NO_N *anda-st*) meaning quite the opposite – ‘to die’. To account for this innovation, this cognate class is also represented by two characters: One for the presence or absence of *anda* and one for the presence or absence of the middle voice ending.

[3] The original *anda* has disappeared completely in Swedish.

A special note should be made about the meaning ‘at’. The clarification of this meaning reads ‘*generic locative, no spatial specification*’. It is, however, hard to tell which locative is the most generic. In order to make this meaning more applicable, I decided to only include the translation of ‘at’ in the sequences ‘at a party’ or ‘at a wedding’. In all languages but Old Gutnish a translation for at least one of the two sequences could be retrieved. Languages for which translations for both sequences were found, usually use the same preposition in either sequence. The only exceptions in the sample are NO_N and NO_B which use *på* in ‘at a party’, but *i* in ‘at a wedding’. Neither of the sequences is attested in NO_S.

Phonology

The phonological database is a separate database which consists of 104 phonological developments that have occurred since the Proto-Germanic period. As in the lexical databases, the presence or absence of an innovation was converted into a binary code with 1 for present and 0 for absent. The presence or absence of these developments are coded for Old Norse, Old Swedish, Old Gutnish, Icelandic, Faroese, Nynorsk Norwegian, Stavanger Norwegian, Bokmål Norwegian, Övdalian, Swedish, Gutnish, Danish, Bornholm Danish, Fjolde Danish, Dutch and West Frisian. As I am not an expert on West Germanic, only a few uniquely West Germanic developments are included in the database. This should suffice for rooting the tree, but might not be enough for dating the first split between West and North Germanic in the tree.

The list of phonological developments was compiled from Bandle (1973) and Groenke (1998). Both are handbooks on the history of the Scandinavian languages with a special focus on phonological developments. Unfortunately, these works take mainly developments that occurred in the standard languages into account, so for the non-standard varieties I deduced some phonological developments from their grammars and dictionaries.

As phonological innovations are not always without exceptions, one cognate was selected for each development. The reason for this is that it is more convenient to code for the presence or absence of a phonological development in one lexeme than in the language as a whole. The presence or absence of the phonological development in the selected cognate is coded by 1 or 0. Some phonological developments had different reaches depending on the phonological environment of the affected sound and sometimes the implementation of a certain change is even word specific. In Swedish, for example, the sequence /ld/ is maintained in the word *eld* ‘fire’, but has assimilated in *kväll* ‘evening’. In such cases, two to five words are selected to map the spread of the development as accurately as possible. This is also a way to make the differences between the languages in the sample as large as possible.

A few further remarks should be made about the coding of the features. In a small number of cases, an older phonological change is made undone by a new mutation. An example of this is found in Danish. The Old Nordic word for ‘to sing’ was *syngva*. Before the sequence /ngw/, breaking occurred which meant that /y/ was broken into /iu/, hence in Old Danish we find *siungæ*. A later change in the phonology of Danish affected /iu/ and changed it back to /y/. As a result, the word for ‘to sing’ is *synge* in Modern Danish. This form does not reveal that breaking affected this word in an earlier stage of the language. As one of the aims of this study is to test whether computational phylogenetics could also be used with languages of which the history is unknown, breaking in ‘to sing’ is coded absent for Danish.

Another process that reversed an earlier development is morphological analogy. In West Scandinavian, nasal assimilation led to the assimilation of /nt/ to /t:/ among other things. So, the original tense conjugation of ‘to bind’, must have been something like *bind-bant-bundinn*. Because of the nasal assimilation in West Scandinavian, Old Norse has *bind-batt-bundinn*, with *batt* instead of *bant*. The past tense without /n/ is still found in most of the West Scandinavian

varieties, but in Faroese and Stavanger Norwegian the original /n/ has been reintroduced in this position. This /n/ is not the product of a phonological development as it only occurs in this particular word, but probably the result of morphological analogy. As back mutations like these are not caused by phonological changes, they are assigned the value '?' instead of coding them as being absent.

In some cases, the selected word has been replaced by a loan word from another Scandinavian language. As loan words do not tend to exhibit the phonological changes that occurred in the variety in question, the original word was retrieved from compounds that were not affected by the borrowing. In the Stavanger dialect for example, the word for 'winter', *vinter*, is probably borrowed from Bokmål Norwegian or Danish. The original *vetr* is still found in compounds such as *vetrong* 'cattle of one winter old'. In such cases, only the original form is included in the database.

Borrowing can also affect the pronunciation of certain sounds in a word. In Faroese the sequence /hj/ is usually pronounced /tʃ/. However, in the selected word *hjálpa* 'to help', the sequence <hj> is pronounced /j/ as in Danish, as opposed to the same sequence in the genuine Faroese word *hjá* 'by', which is pronounced /tʃɔa:/. When a selected word turns out to have an irregular pronunciation, it is replaced by another word with the regular pronunciation.

Morphosyntax

The morphosyntactic database consists of 61 items which are coded as being either present or absent for Old Norse, Old Swedish, Old Gutnish, Icelandic, Faroese, Nynorsk Norwegian, Stavanger Norwegian, Bokmål Norwegian, Övdalian, Swedish, Gutnish, Danish, Bornholm Danish, Fjolde Danish and Dutch. The items in the list are developments of new morphology or syntactic constructions and features whose parameters change easily over time. Loss of morphology or syntactic variation is not taken into account. The full list of features can be found in the appendix.

An example of new morphology is the development of a new noun declension in Swedish. In Old Swedish, the plural of neuter words was identical to the singular. Hence, the word *hiärta* could mean both 'heart' and 'hearts'. In modern Swedish, a new declension has been developed which consists of neuter words ending in a vowel (Petterson 2005). The *–n* of the definite plural ending (*hiärta-na* 'heart-DEF.PL'), was reanalysed as a plural ending, giving rise to a new declension in which the singular and plural are no longer identical. The plural of *hjärta* 'heart' is therefore *hjärtan* in modern Swedish.

A feature of which the parameters can change over time is for example the way numbers over 20 are expressed. In Old Norse '21' was *tuttugu ok einn*, literally 'twenty and one'. In Old Swedish the order is different, though, and the form *en ok tiughu* 'one and twenty' is attested. These are the only two possible orders for Scandinavian, so all languages either have the pattern of Old Norse, that of Old Swedish or both.

Loss of morphology and syntactic variation is not included in the list. The reason for this decision is that loss of categories is to be expected in Germanic-speaking communities. Especially the mainland varieties of Germanic have a substantial number of speakers as well as L2 speakers, two factors which correlate with the loss of inflexions and variation (Trudgill 2011). So, the loss of the case system and most of the verb endings in large parts of the mainland Scandinavian varieties is probably not the result of common ancestry, but that of the similarity of the communities in which the languages are spoken. In other words, the loss of features does not necessarily need to reveal common ancestry and might even cause noise in the analysis. Therefore, the loss of features is not taken into account in this study.

A common feature of lexicon, phonology and morphosyntax is that parts of these categories can be borrowed into another language. Although borrowing of morphosyntactic features is less likely to occur than borrowing of lexical items, it does occur sometimes. In the majority of cases it is hard to tell whether a certain construction is the result of borrowing. Hence borrowing was not controlled for in this database.

A difficulty with syntax is that the order in which different parts appear might not be wholly unrelated. Greenberg (1963) was the first to describe a number of linguistic tendencies which state that the positions of certain elements are interrelated, but none of his universals or tendencies seem to affect the features in this database. However, an observation that is relevant for the data in this study was made by Koeneman & Zeijlstra (2010). They argue that the richness of verb morphology affects the placement of the verb in sub-clauses. The sample of Scandinavian languages does in fact show this distinction. The languages which have rich verb morphology do not exhibit verb movement in the sub-clause, whereas the ones with poor verb morphology do. Therefore, this feature is excluded from the database.

The data for the morphosyntactic database are retrieved from the grammars of the respective languages. Also, the Nordic Atlas of Language Structures (NALS) was consulted, but only little was taken from there, as most of the maps only focused on a subpart of Scandinavia. As for the lexical and phonological data, the presence or absence of a feature is translated into a binary matrix with 1 for presence and 0 for absence. In some cases a certain feature was not applicable to one or several of the languages. Consider, for example, features 49 and 50, which concern the placement of the article in the construction "such a". Some of the languages, such as Icelandic, do not have an indefinite article so there is no way to know if it would precede or follow "such". In those cases, the value "?" is assigned.

Combined database

The combined database consists of data from the three earlier mentioned parts of language, viz. lexicon (domestic animal terminology)[4], phonology and morphosyntax. The three databases contain different numbers of categories, however. The morphosyntactic database contains only 61 items, whereas the phonological and lexical ones have 104 and 222 items respectively. So, plainly concatenating rows of character state values of the three databases would give an unequal amount of power to each of them. This would not be a problem, if the data in each database would give the same result, but as we do not know that the databases are represented in equal measure. 61 items are randomly selected from the lexical and phonological databases by a random number generator.

Although the number of items from each database is the same now, it does not necessarily mean that all databases have the same amount of impact on the outcome of the analysis. The specifications of the samples are summarised in Table 7.

TABLE 7. The specifications of the three database samples in the combined database.

	# items	# 1's (Dutch included)	# 1's (Dutch excluded)
LEX	61	201	186
PHON	61	338	329
MOR-SYN	61	348	327

[4] The domesitc animal terminology database was chosen in order to already make a preselection.

Table 7 shows the number of 1's in each of the samples with and without Dutch. The table shows that the number of 1's is much lower for the lexical database. This means that the average number of languages sharing 1's in that database is much lower than in the two other ones. To avoid having too many small or one-language clusters, I decided to gather a similar number of 1's for every database.

A reason to exclude Dutch from the calculations is that fact that Dutch is represented to varying degrees in the three samples. A comparison of the third and fourth columns shows that the number of 1's in Dutch represent 7% of all 1's in the lexical sample, 3% of those in the phonological sample and 5% in the morphosyntactic dataset. As this skewed distribution might influence the results and considering the fact that the main focus of this study is on the Scandinavian languages, Dutch was excluded as a precaution.

An additional number of categories are therefore randomly selected from the lexical database until the number of 1's similar to the number of 1's in the phonological and morphosyntactic samples is reached. At 89 categories, the number of 1's in the sample is 328 and the specifications of the combined dataset are shown in Table 8.

TABLE 8. The specifications of the three database samples in the combined database.

	# items	# 1's (Dutch included)	# 1's (Dutch excluded)
LEX	89	355	328
PHON	61	338	329
MOR-SYN	61	348	327

So, the combined dataset consists of present/absent data for 211 items and of a total of 984 1's. The advantage of such a combined database is thought to be that it levels out the imperfections and problems of the indivual databases and gives a better summary of the history of the phylum under analysis.

Computational Phylogenetic Analysis: Conversion into a tree

Likelihood methods

The next step in the process is to convert the observed data into tree structures. There are several likelihood methods that can be applied to build a tree. A likelihood method quantifies the probability that the observed data is produced by a particular mechanism (Dunn 2014). The mechanism at play here is an evolutionary process. Hence, the mechanism consists of a set of mathematical descriptions of evolutionary processes. The outcome of the analysis is summarised in a set of tree models containing a tree topology and branch lengths that can be assigned several meanings. It should be noted that the tree model is presupposed and does not affect the likelihood calculations.

The challenge is to maximise the likelihood of the parameters of the tree and to find the best tree topology in the space of possible trees. This is more challenging than one might think, because there is at present no algorithm that can find the best tree within a reasonable amount of time. This problem cannot be overcome by random sampling in the tree space as only a very small number of trees in the tree space have a high likelihood score. So, most random tree samples will contain no high likelihood tree at all. To cope with this problem, Bayesian Monte Carlo Markov chain (MCMC) sampling is used. This algorithm starts from a random point in the tree space and searches for the region of highest likelihood from there. It compares the

likelihood of its current position to those of other points in its vicinity. If the likelihood of another position turns out to be higher, it moves to the new location and the process of finding a higher likelihood is continued from there. To avoid getting stuck at a local maximum, the algorithm sometimes also allows slightly lower likelihoods.

The search is continued until the average standard deviation of split frequencies is lower than 0.01. Every *n* steps the tree parameters and likelihood scores are saved to the posterior sample. Thus, after the analysis the posterior sample contains of a large number of trees. Not all trees in the posterior sample have similar likelihoods, however. Figure 3 illustrates the typical course of a search.

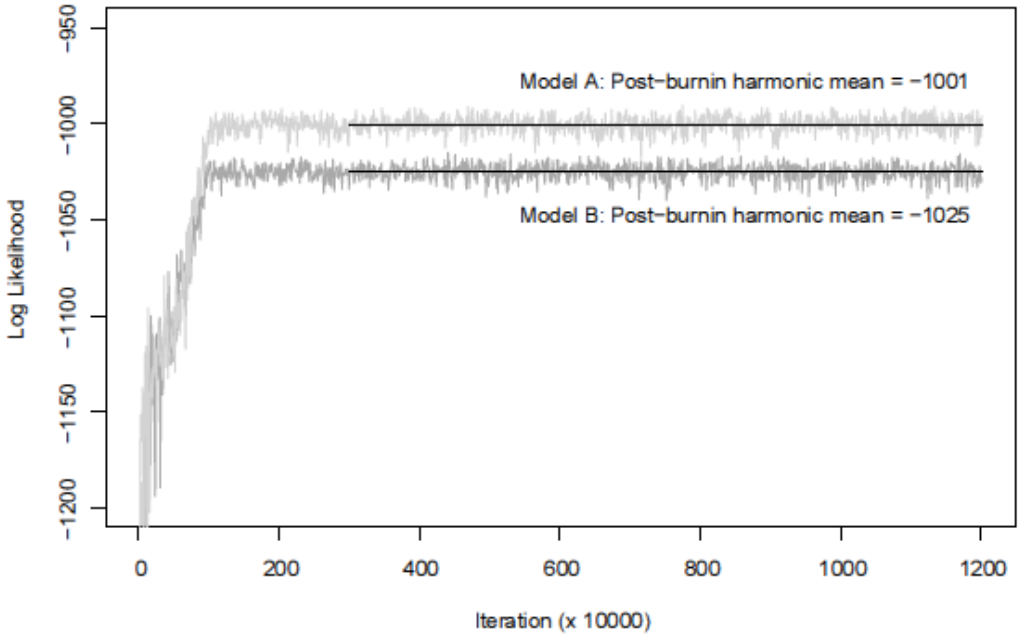

FIGURE 3. Likelihood trace MCMC sampling under two models. (Dunn 2014)

The course of the likelihood in Figure 3 shows that it oscillates wildly at the beginning of the search, but that it after a certain number of iterations reaches some kind of equilibrium. The samples from the initial period are therefore discarded, whereas the rest is saved to the posterior sample as they all represent more or less equally likely trees. The final phylogenetic tree is a sort of summary of all the trees in the posterior sample.

Evolutionary model

Before the data can be analysed and a phylogeny can be inferred, a number of settings need to be chosen. One of these settings is the evolutionary model. In order to be able to carry out the analysis, suitable substitution and rate models need to be set. There are three possible substitution models to choose from: F81, HKY and GTR. The only models that can work with binary data are F81 and HKY. F81 assumes that the rates of innovation and loss of cognates/features are equal, whereas HKY allows them to vary. As it is more realistic to assume that the rates of innovation and loss are different from each other, the HKY model is used in this study.

The rate model that is used is the gamma rates model. This model works with the assumption that the rates of characters evolve according to a gamma distribution. It assumes

that characters belong to a certain number of rate classes. A larger number of classes would improve the fit of the tree to the data, but as the amount of data in this study is quite small, any new classes will probably be very similar to those that already exist. Moreover, increasing the number of classes would have a highly negative effect on the computation time. Therefore, the default setting of four classes is maintained.

Dating the tree

Swadesh (1952) already put forward a method to calculate the age of the splits in a phylogenetic tree. He thought he had discovered a constant rate of change of neutral vocabulary, i.e. the meanings on the Swadesh list. Languages that split 1000 years ago would share 81% (±2%) of the meanings on the Swadesh list. Unfortunately, the constant rate of change does not occur in language evolution. Not all the branches of the tree change at a constant rate and even within the history of one branch, the rate of change can vary over time. Research has pointed out that the rate of change tends to be accelerated in the period directly after a splitting event or in times of social insecurity, for example (Trudgill 2011). Also contact situations and population size play a role in the rate at which a language changes (Trudgil 2011).

The existence of differences in the rates of change between branches can be inferred quite well from the tree of North Germanic languages. All members of the tree are descendents from Proto-Nordic. Yet, Icelandic is much closer to the ancestor language than the mainland varieties (Bergsland & Vogt 1962). In other words, Icelandic has changed at a much slower rate than Norwegian, Swedish and Danish. This applies to both vocabulary and morphosyntax. This means that glottochronology is not the right tool to use for calculating the age of the splits in the tree. Moreover, glottochronology only maintains a constant rate of change for the vocabulary on the Swadesh list and does not give any constant rate of change for morphosyntactic or phonological data.

Luckily, some other methods have been developed to estimate the ages of the splits in a tree. The likelihood method also includes rate of change in its analysis as clock models. The most basic type of clock model is a strict clock, which just as glottochronology assumes a constant rate of change over the whole tree. The rate of change does not need to be 81% per 1000 years, but could theoretically take every value between 0% and 100% per 1000 years. As already mentioned before, the tree of Scandinavian languages offers clear evidence to reject a constant rate of change along all its branches.

Beside strict clocks, also a number of so-called relaxed clocks have been created. These clock models fit better to the evolutionary process of languages, as they allow rates to vary across the tree. The MrBayes software package contains several such relaxed clocks. For every analysis three types of relaxed clocks (cpp, tk02 and igr) as well as a strict clock are tested. The clock model resulting in the highest harmonic mean was used for the final analysis of the data. In order to be able to date the tree, an estimate has to be given of the age of the root of the tree. In all analyses this date is set to 1400 to 2400 years ago. Also, the ages for the older language stages are included in the analysis and set to 800 to 1200 years for Old Norse, 800 to 650 years for Old Swedish and 800 years for Old Gutnish.

Such data are usually not available for languages of which no historical records have been found. However, some tricks can be applied to acquire some information about the history of those languages. Archaeological findings might shed light on when an area was first inhabited by a certain tribe or people, which gives an indication of the point at which the language of that area might have split off from the other languages. Hence, these methods can also be employed to languages for which we have no older written records. As we do have

written records of older stages of Scandinavian, dating based on archaeological findings do not need to be taken into account.

Tree representation

The posterior tree sample can be summarised in different types of trees. In this study the types of trees are restricted to consensus and maximum clade credibility (MCC) trees.

Consensus trees summarise the data in such a way that conflicting splits are resolved by putting them one level higher on the tree. In order to exemplify this, consider a posterior sample in which 35% of the trees show the clade (A,B,C|D,E) and 65% show the clade (A,B|C,D,E). The conflict in this example is that in some samples C belongs to the same clade as A and B, but in the majority of the samples it is clustered together with D and E. Instead of projecting C together with D and E as the majority of the samples suggest, a consensus tree gives the clade (A,B|C|D,E). So, instead of depicting Proto-ABCDE splitting off into two branches, the tree gives a trifurcating split to avoid the conflict. In all the trees in the sample A and B and D and E belong to the same clades, but as C might belong to either of them the split of C is situated at the same point as that of (A,B) and (D,E).

Behind the construction of an MCC tree lies another procedure. To build an MCC tree one needs the probability of every branch in the tree. The probability of a branch is the proportion of the posterior sample in which it is found. Then, the probabilities of every branch in the trees are multiplied. The tree with the highest probability of all its branches combined is the MCC tree. So, the MCC tree is not strictly a summary of all the trees in the posterior tree sample, but rather the most representative tree in that sample.

Both types of tree have their advantages and disadvantages. One of the major disadvantages of consensus trees is that the avoidance of conflicting branches might lead to the choice of a clade that has a lower likelihood than either of the conflicting clades. To go back to the aforementioned example, (A,B|C|D,E) might have a lower likelihood than either (A,B,C|D,E) and (A,B|C,D,E). Besides, the resulting tree might not even be one of the trees in the posterior sample. The main drawback of MCC trees is that they only allow for bifurcating splits. This is in particularly problematic with dialect continua that show gradual instead of abrupt diversification. Considering the above example of languages A, B, C, D and E again, C might be in the middle of the dialect continuum and share innovations with both (A,B) and (D,E). Clustering C with either of those branches would suggest either a Proto-ABC or a Proto-CDE language. Such an ancestor language might in reality have never existed, though. To explore the pros and cons of both tree types, both are procured and shown in the results section.

Exploring the data: NeighborNet split graphs

Before deducing a phylogenetic tree from the data, the data are explored first. This means checking how tree-like the data are. This is done by means of a NeighborNet split graph. As opposed to trees, split graphs do not force the data into a tree-like structure but render a network representation of it. A NeighborNet split graph is a practical tool for visualising how tree-like the data are. When the data are completely tree-like, the split graph looks like an ordinary unrooted tree. The less tree-like the data are, the more boxes emerge in the split graph denoting conflicting distance relationships in the data.

A NeighborNet split graph is constructed according the following lines (Bryant & Moulton 2004). At the outset, all taxa in the distance matrix are represented by a single node in space. Then the two closest taxa are paired up as neighbours. This process is reiterated until a

third neighbour is found. At that point the three nodes are collapsed into two nodes, which replace the three nodes in the distance matrix. This procedure is illustrated in Figure 4.

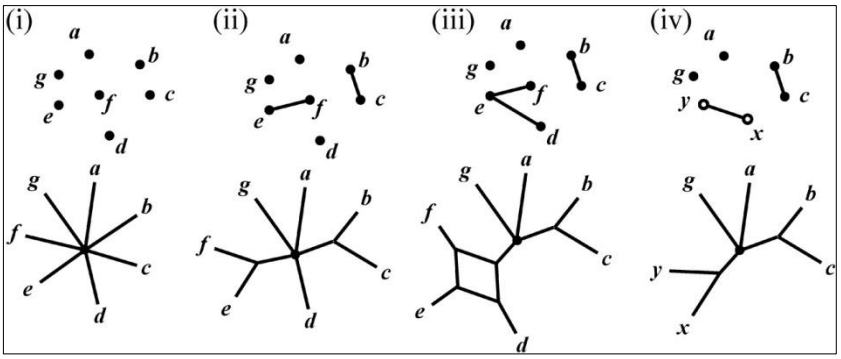

FIGURE 4. (i) each taxon is represented by a node. (ii) the two closest nodes are paired up, in this case *e* and *f* and *b* and *c*. (iii) d is also identified as a neighbour of e. (iv) the paired up nodes *d,e,f* are collapsed into two nodes *x,y*. (Bryant & Moulton 2004)

Then the search for a third neighbour starts all over and this process is repeated until only three nodes are left. Then the process is reversed and the nodes are fully expanded. The expanding gives the splits in the split graph. This process is illustrated in Figure 5.

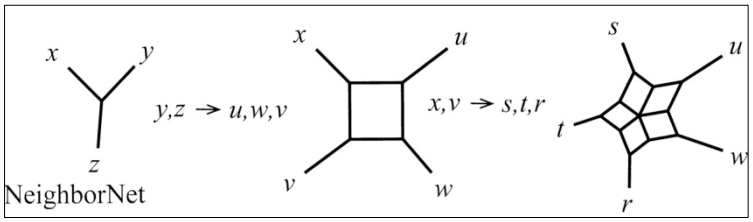

FIGURE 5. The nodes *y,z* are expanded to the original nodes *u,w,v*. Then *x,v* are expanded to *s,t,r*. (Bryant & Moulton 2004)

The result is a network that shows the amount of conflicting signal in the data by the degree of reticulation. Conflicting splits can be explained in several ways. The most evident explanation is borrowing. In the event of the split of two separate groups, the individual languages within those groups can still borrow linguistic material from the languages from the other group. However, this should not occur in this study, the lexical part at least, as loan words are excluded.

A second reason for conflicting signal is the loss and retention of ancestral material. Imagine the languages *a, b, c, d, e* and *f* which relate to each other as (*a,b,c|d,e,f*). Language *a, b, c, d, e* and *f* all developed from Proto-*a,b,c,d,e,f*. Proto-*a,b,c,d,e,f* contained a lexeme in cognate class A. This cognate was still present in both Proto-*a,b,c* and Proto-*d,e,f*. In the modern descendents, however, cognate class A has been lost in languages *c* and *e*. This will produce the conflicting signal (*a,b,d,f|c,e*) as the presence of cognate class A is considered a shared innovation of languages *a, b, d* and *f*, even though that is not the actual case. So, shared retentions also cause some of the reticulation in the split graph.

Conflicting signals may also be explained by the nature of language change in a family. In a dialect continuum for example, absolute splits are unlikely to occur. Linguistic changes under such circumstances are expected to spread in a wave-like fashion. In the language family (*a,b,c|d,e,f*) the fundamental split is between languages *a, b* and *c* and languages *d, e* and *f*. If there is transboundary communication in the adjacent languages *c* and *d*, however, changes in

language *c* might also spread to language *d*. A split between two languages at some point in time does not mean they can no longer share innovations.

Conflicting splits caused by borrowing events or the loss of ancient material usually do not contain information on the genetic history of a phylum, so their invisibility in tree representations might rather be seen as an advantage of this type of model. The fact, however, that tree representations presuppose absolute splits means that developments that spread wave-like are not shown in the final tree model. Therefore, it is useful to visualise the data in a network before opting for the construction of a phylogenetic tree.

Perfect tree-like histories are not to be expected as borrowing, loss of ancestral material and wave-like changes can overrule the initial split in different branches, but some phyla will have developed in a more tree-like way than others. Unfortunately, there are no formal procedures to decide whether the data are tree-like yet. Holland et al. (2002) introduced the delta score as a tool for data exploration. Gray et al. (2010) use both the delta score and Q-residual score to analyse the shape of the evolution, but mention that these two measures have not been proven sufficiently powerful and robust yet. When the data are completely tree-like the delta score will be zero, and one if the data are not tree-like at all. As there is some debate about the actual significance of delta and Q-residual scores they will not be further discussed in this paper. They will, however, be given for every split graph.

Neighbor-Net split graphs

In this section the data are explored and analysed in NeighborNet split graphs. The split graph of the Swadish list lexical data is shown in Figure 6.

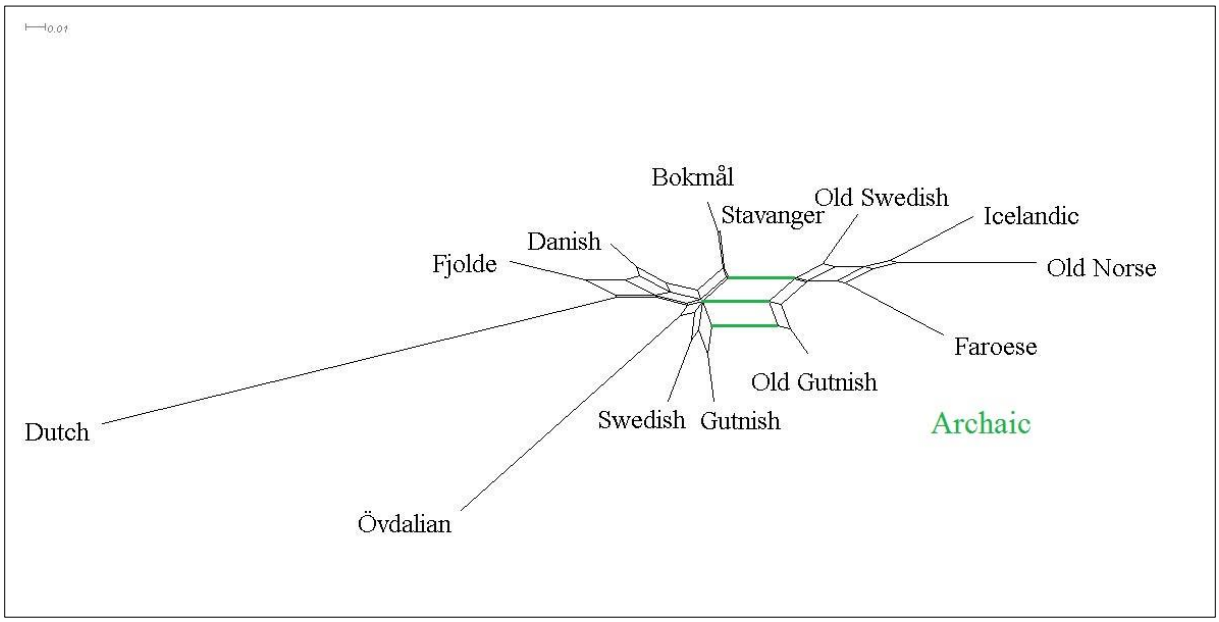

FIGURE 6. The Neighbor-Net split graph for the Swadish list data. The fundamental split seems to be between the retentive, archaic varieties and the more innovative ones. The lines signalling the the grouping of the archaic varieties are highlighted in green.
Fit = 96.426, δ-score = 0.3013 and Q-residual score = 0.01755.

The well-known split between West and East Scandinavian cannot be discerned from the split graph. The fundamental split here seems to be (NL, DK_FJ, DK, OEV, SE, GUT, NO_B, NO_S | OG, OS, IC, FA, ON). The driving factor behind this partition is probably the retention of cognate classes in insular Scandinavian which have been lost in the mainland varieties. So, the first

impression is that this dataset does not give an accurate picture of the history of the Scandinavian languages.

Nevertheless, the data seem to be quite tree-like and some subgroupings can be discerned. In the left half of the network, a clear Danish (DK, DK_FJ) and Norwegian (NO_B, NO_S) branch can be detected. Besides, there is also a weaker Swedish branch (SE, OEV, GUT). In the right half, there is a clear insular Scandinavian branch (IC, FA, ON). Old Swedish and Gutnish do not seem to have very close neighbours beside each other. However, there seems to be some evidence that connects Old Gutnish to the Swedish branch.

The split graph for the domestic animal terminology data is shown in Figure 7. In this network, the split between West and East Scandinavian is clearly visible. The edges indicating the clustering of the West Scandinavian branch are highlighted in blue. There is also some evidence for an East Scandinavian branch. The edges signalling this clustering are in red. As the red edges are shorter than the blue ones, there is stronger evidence for the West Scandinavian branch than for the East Scandinavian one.

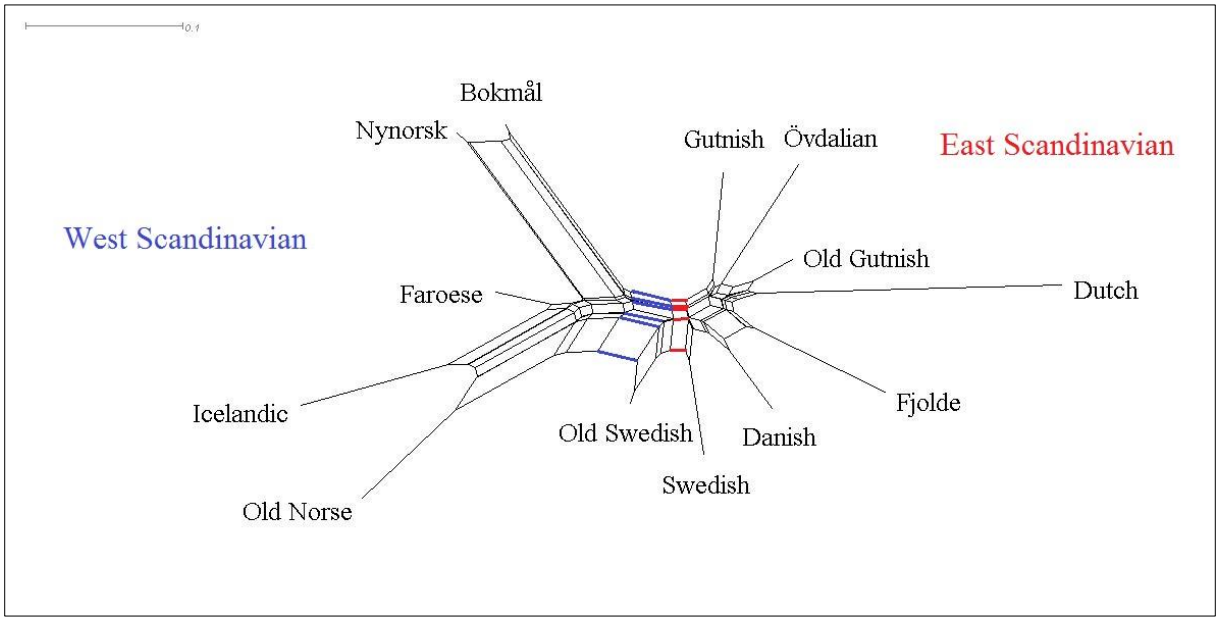

FIGURE 7. The Neighbor-Net split graph for the domestic animal terminology. The fundamental split seems to be between West and East Scandinavian. The lines setting apart the West Scandinavian varieties are highlighted in blue, the ones setting apart the East Scandinavian varieties in red.
Fit = 97.319, δ-score = 0.3189 and Q-residual score = 0.01538

Another conspicuous difference between the West and East branch are the average edge lengths. The edges in the western half are significantly longer than the ones in the eastern half. This might suggest there have been more innovations in the West Scandinavian languages, but it might also mean that the western varieties have preserved archaic cognate classes that have disappeared altogether from the languages in the east or a combination of both. Unfortunately, one cannot be sure on the basis of a split graph representation.

The picture in figure 7 looks slightly more reticulate than the network in figure 6. This indicates that the domestic animal terminology data are less tree-like than the data from the Swadesh list. Nevertheless, some subgroupings can be discerned in the network. Beside the West/East split, there are also a number of secondary splits. In the West Scandinavian languages, there is a clear Norwegian as well as an insular Scandinavian branch. In East Scandinavian, three secondary partitions are discernable. The first one is the Swedish branch which comprises Swedish and Old Swedish, which according the network does not belong to

either the West or East Scandinavian branch. There is also evidence for a separate Danish branch, consisting of Danish and the Danish dialect of Fjolde. At the top of the eastern half, there seems to be some evidence for a cluster of Swedish *Randmundarten* in which Övdalian, Gutnish and Old Gutnish are included.

The split graph for the phonological data is shown in Figure 8. The split between West and East Scandinavian can also be discerned in this network. The blue lines denoting the evidence for a separate West Scandinavian group are clearly longer than the red lines indicating the data suggesting a separate East Scandinavian group. In other words, the West Scandinavian languages have more shared features than the East Scandinavian languages.

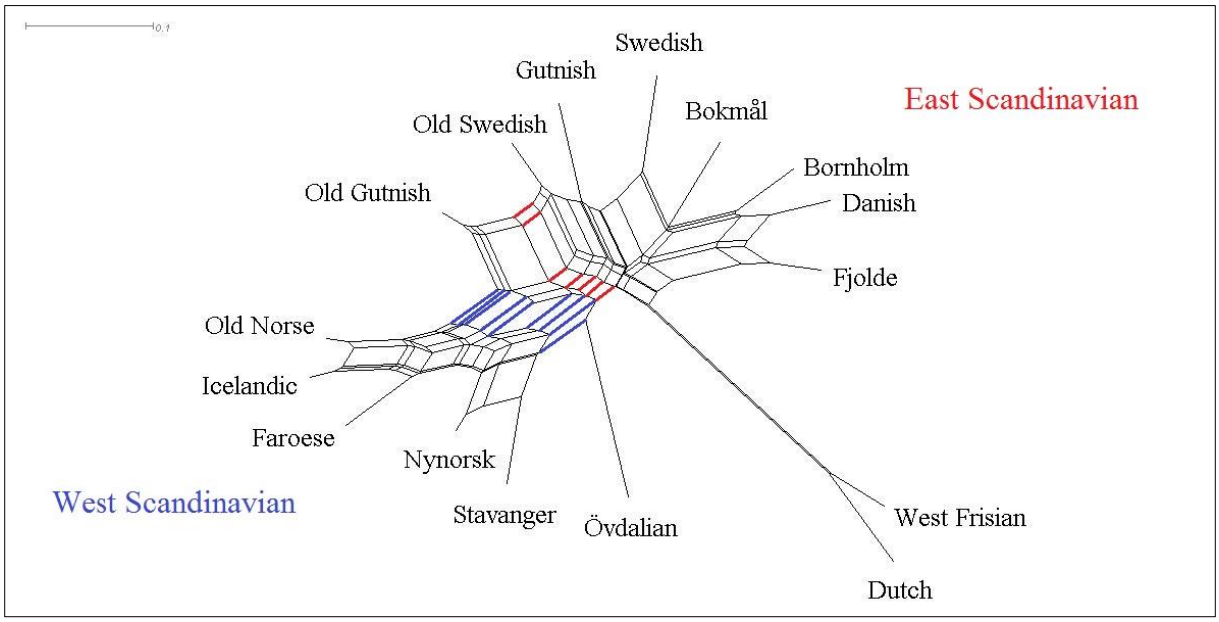

FIGURE 8: The Neighbor-Net split graph for the phonological data. The fundamental split seems to lie between the West and East Scandinavian varieties. The lines signalling the grouping of the West Scandinavian languages are highlighted in blue, the lines signalling the grouping of the East Scandiavian ones in red.
Fit = 93.392, δ-score = 0.3645 and Q-residual score = 0.05369

A special remark should be made about Övdalian and Old Gutnish. Both these languages exhibit phonological developments that are indicative of both West and East. Övdalian is spoken in the centre of Scandinavia, so in the frontier zone of the western and eastern subgroups. The presence of features from both branches implies there was no definite line between West and East Scandinavian. Varieties spoken in the border zone might therefore exhibit phonological features of both groups. This explanation does not apply to Old Gutnish, though, as that variety was spoken on one of the easternmost islands in the North Germanic language area. Its position in the tree might be the result of the fact that it did not take part in some Pan-East Scandinavian developments such as monophthongisation, compare Old Norse *steinn*, Old Swedish *sten* but Old Gutnish *stain*, and the fact that it probably by chance shared some innovations with the West Scandinavian languages such as nasal assimilation, compare Old Norse *vetr*, Old Swedish *vinter* and Old Gutnish *wittr*.

In general, the varieties are located in the expected positions. West Scandiavian is divided into an insular and Norwegian branch and in the East there are two clear Swedish/Gutnish and Danish branches. One very notable difference between this network and the one for the domestic animal terminology is the placement of Bokmål Norwegian. In Figure

7 there was no doubt it belonged to the West Scandinavian branch but in the phonological network it is located in the centre of the East Scandinavian subfamily. The explanations for this difference are several. Firstly, the heartland of Bokmål Norwegian is just as that of Övdalian located in the borderland of the West and East branches. So, it might combine western and eastern features in one variety. Another aspect that should be taken into account is that Bokmål Norwegian is based on the older written Danish standard, which accounts for a number of East Scandinavian traits in the pronunciation of Bokmål Norwegian. When Norway gained independence from Denmark, Bokmål was introduced as the new national language. In an endeavour to make it closer to the language actually spoken in Norway some typical Norwegian vocabulary was introduced into the new standard language. Therefore, the lexical data suggest Bokmål Norwegian to belong to the West Scandinavian branch, whereas the phonological data puts it between its geographical neighbour Swedish and its former ruler Danish.

In general, the phonological network looks a little more reticulate than either lexical one. All expected subgroups can be easily detected, though. The only real problematic subgroup here might be the Swedish/Gutnish one as it developed in quite conflicting ways. Its members clearly introduced shared Swedish/Gutnish innovations but at the same time some members also developed into the direction of languages outside the group. This might be caused by the older languages, as they were spoken before certain shared innovations had appeared. In Figure 9 the same data are shown, but now the older language stages are omitted.

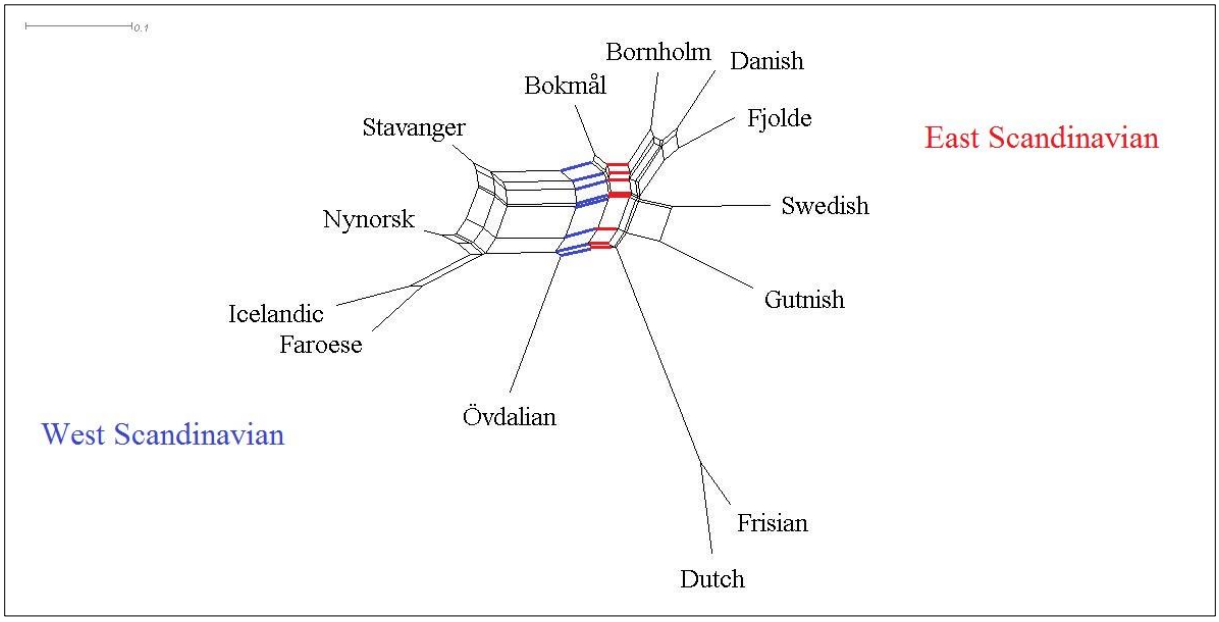

FIGURE 9. The Neighbor-Net split graph for the phonological data. The lines signalling the grouping of the West Scandinavian languages are highlighted in blue, the lines signalling the grouping of the East Scandiavian ones in red. Ancient languages have been omitted.
Fit = 94.858, δ-score = 0.3507 and Q-residual score = 0.03945

Indeed, the sizes of the boxes in the Swedish/Gutnish branch have decreased now, but now another problem is revealed in the network in Figure 9. There is not only a clear vertical split between West and East, but also a horizontal one in the extension of the Swedish edge, giving the split (NO_S, NO_B, DK_B, DK, DK_F | SE | NO_N, IC, FA, OEV, NL, WFR, GUT). This is probably the corollary of the Danish hegemony of Norway, which caused the varieties spoken in the larger Norwegian cities to assimilate to the pronunciation of Danish. The reasons for the

clustering of Övdalian and Gutnish with the West Scandinavian languages are the same as discussed above.

The phonological variation among the modern Scandinavian varieties is not unequivocally explained by common ancestry but are also the result of language contact, in particular contact with Danish. This is problematic for the construction of a phylogenetic tree, as it only can show one of the developmental paths. Consequently, these data are not particularly suitable for conversion into a tree-like structure as the result would not be in agreement with the actual history of the language family. It would spuriously suggest an unequivocal signal and might sketch an entirely wrong picture of the genetic relations of the languages of Scandinavia. This should be kept in mind when constructing the phonological tree.

In Figure 10 the network for the morphosyntactic data is shown. Also here, the split between the West and East Scandinavian branches is expressed by the split graph. The basis for a separate East Scandinavian branch is very weak though, as is indicated by the red lines that are so short that they rather appear to be dots. Furthermore, Old Swedish, Old Gutnish and Övdalian are not even included in the East Scandinavian cluster according to this figure. As there is so little evidence for an East Scandinavian branch, it is more sensible to speak of a Swedish/Gutnish and a Danish branch in this figure.

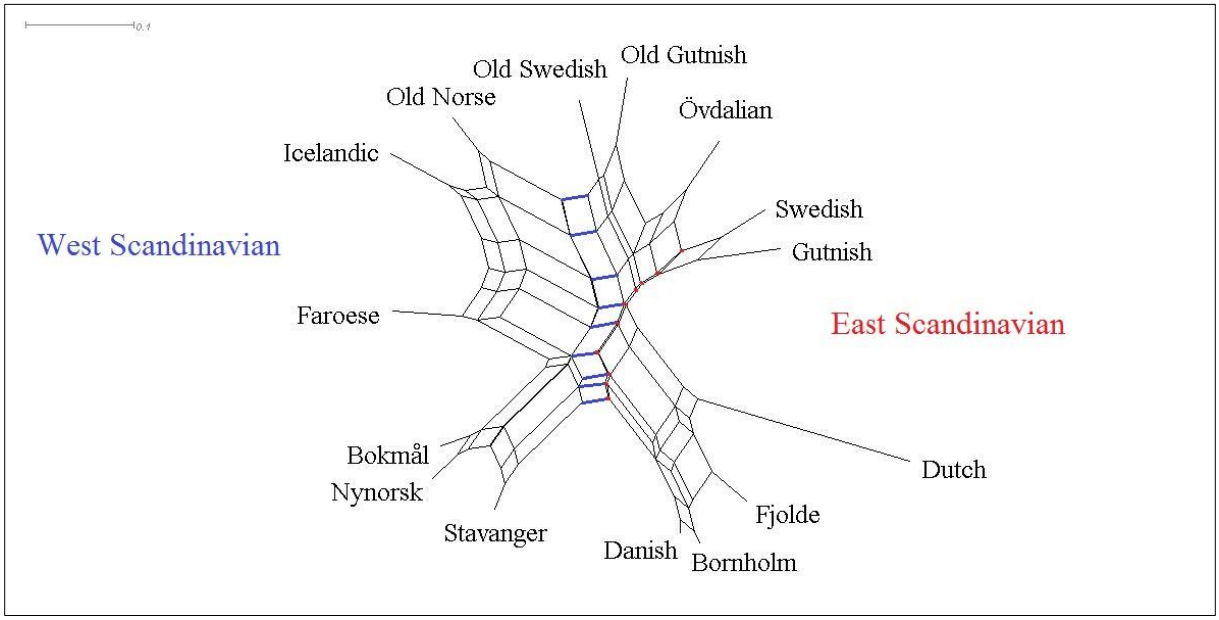

FIGURE 10. The Neighbor-Net split graph for the morphosyntactic data. The lines highlighted in blue signal the clustering of the West Scandinavian varieties, whereas the ones highlighted in red signal the clustering of East Scandinavian.
Fit = 91.822, δ-score = 0.3719 and Q-residual score = 0.08596

Compared to the network for the phonological data, this one looks even more reticulate. This indicates that conflicting signals are at play. Part of the reticulation might be explained by the fact that Old Swedish and Old Gutnish are in the transition between Proto-Norse and modern East Scandinavian varieties and have not yet implemented all the typical innovations. Yet, this does not explain all the conflicting signals. Again, language contact with Danish has caused noise in the ancestral signal. As figure 11 shows, all the languages spoken in the former kingdom of Denmark share a number of innovations that are absent in the other Scandinavian languages. The edges setting apart the Danish-influenced languages are highlighted in orange in figure 11.

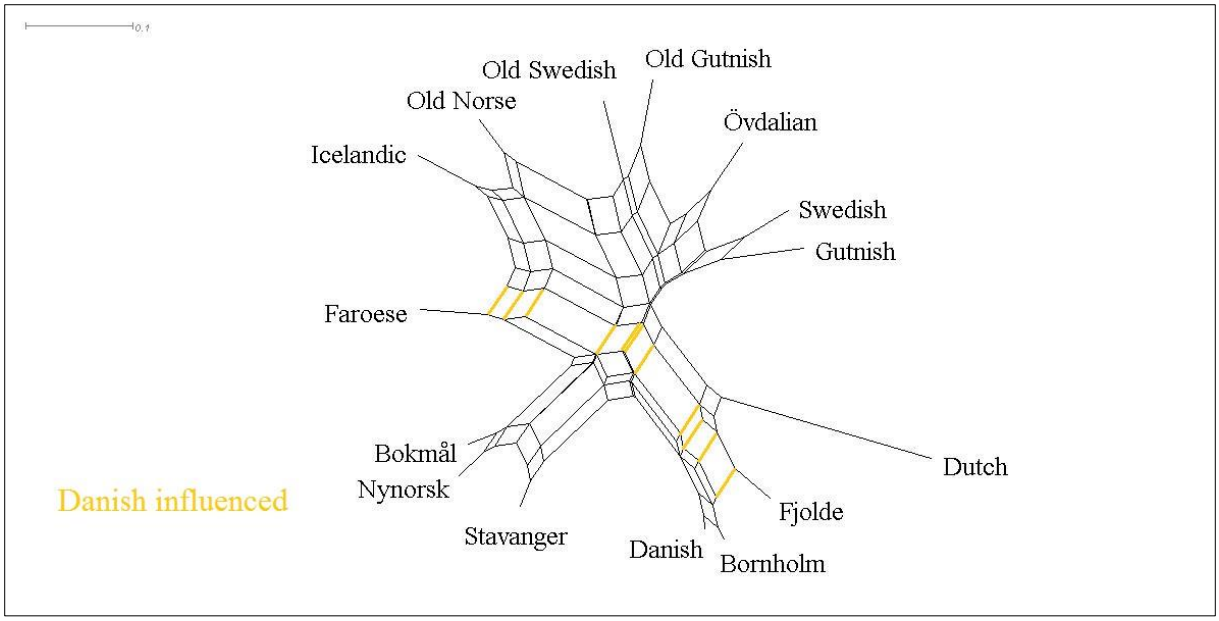

FIGURE 11: The Neighbor-Net split graph for the morphosyntactic data. Another fundamental split seems to exist between the languages spoken in the historical kingdom of Denmark. The lines signalling this group are highlighted in orange.
Fit = 91.822, δ-score = 0.3719 and Q-residual score = 0.08596

The original West-East split is blurred by the changes that spread from Danish to the other languages spoken in the kingdom of Denmark. A situation like this can easily be shown in a split graph, but causes problems for trees as it assumes unidirectional developments after the split of two groups. All this taken into account, as well as the high level of reticulation, makes these data very bad for a phylogenetic analysis.

Data from the lexical, phonological and morphosyntactic databases were gathered in a combined database which resulted in the network in Figure 12.

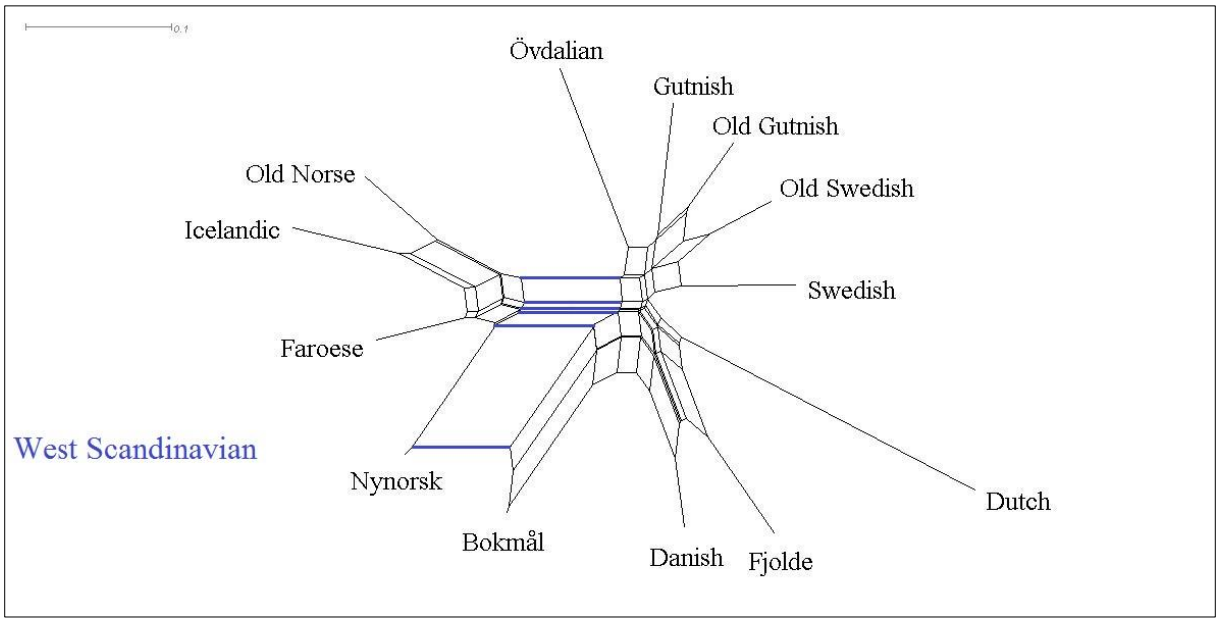

FIGURE 12. The Neighbor-Net split graph for the combined data. The partition of West Scandinavian is signalled by the lines highlighted in blue.
Fit = 94.333, δ-score = 0.3959 and Q-residual score = 0.04287

The signal for a separate West Scandinavian branch is quite clear again and indicated by the edges in blue. An East Scandinavian branch cannot be equally well discerned, but there are clear clusters of Swedish/Gutnish and Danish varieties. Influence from Danish on the languages on its former colony is hardly sensed in this network. Only Bokmål Norwegian seems to have been weighed down by the influence from Danish. This is not very surprising, though, as Bokmål Norwegian is the continuation of the older written Danish standard in Norway.

The reticulation in the network is indicative of a reasonable amount of conflicting signal. However, a great deal of the conflicting signal can be explained by Bokmål Norwegian. It has both the highest δ-score, 0.44975, and Q-residual score, 0.06376. Bokmål Norwegian is in some way a smorgasbord of typically West and East Scandinavian as well as Danish features and consequently shares innovations with every branch in the family. Also the Swedish/Gutnish cluster exhibits a high number of conflicting signals. The average δ-score is 0.42186 which is significantly above average. The conflicting signals are partly confined to the cluster itself, but some varieties also share features with West Scandinavian languages. The splits shown in the split graph are listed in (1).

(1) (WEST, ÖV | GUT, OG, OS, SE)
(ÖV, GUT, OG | OS, SE)
(ÖV | GUT, OG, OS, SE)
(IC, ON, ÖV, GUT, OG, OS | SE)

In addition to this, Old Gutnish and Old Swedish have both two closest neighbours, giving the pairs (OG, GUT), (OG, OS) and (OS, SE). So, in conclusion the varieties belonging to the Swedish/Gutnish cluster are all more closely related to each other than to the other Scandinavian languages, according to the network, but their internal ordering is quite obscure as a result of conflicting signals. This is, however, not an uncommon pattern for dialect continua, hence the Swedish/Gutnish language area might be considered a dialect continuum based on the combined database.

In conclusion, the lexical data seem to be most tree-like. The superiority of lexical data in this respect is probably a result of the exclusion of loanwords. The exclusion of loanwords has as a result that horizontal developments that occurred after splitting events are neglected in the analysis. So, the influence from Danish that is detected in the phonological and morphosyntactic data is not a factor in the lexical data. This makes the data less equivocal and thus more suitable for conversion into a tree. However, it should be noted that the tree in that case might be an oversimplification of the real history of the Scandinavian languages as secondary developments, most often emanating from Denmark or northern Germany, are not represented in the tree. On the other hand, it should do better at retrieving the common ancestor of two random languages in the family, which is the main goal of computational phylogenetics.

Results

In this section the phylogenetic trees are presented and briefly discussed. Also, some comparisons are made to the family tree as it is assumed to be by historical linguistics, i.e. the tree in figure 1 in section 2. More profound comparisons between the two types of tree as well as interpretations on the differences and resemblances are discussed in the discussion section.

The Swadesh tree

The consensus tree deduced from the Swadesh lists is shown in Figure 13. Edges representing a taxon or taxa traditionally considered to belong to the West Scandinavian languages are made purple. Red edges denote that the taxon or taxa connected to those are traditionally classified as East Scandinavian. If the tree were in agreement with the traditional views on the history of the Scandinavian languages, the Scandinavian branch would split off in one blue and one red clade.

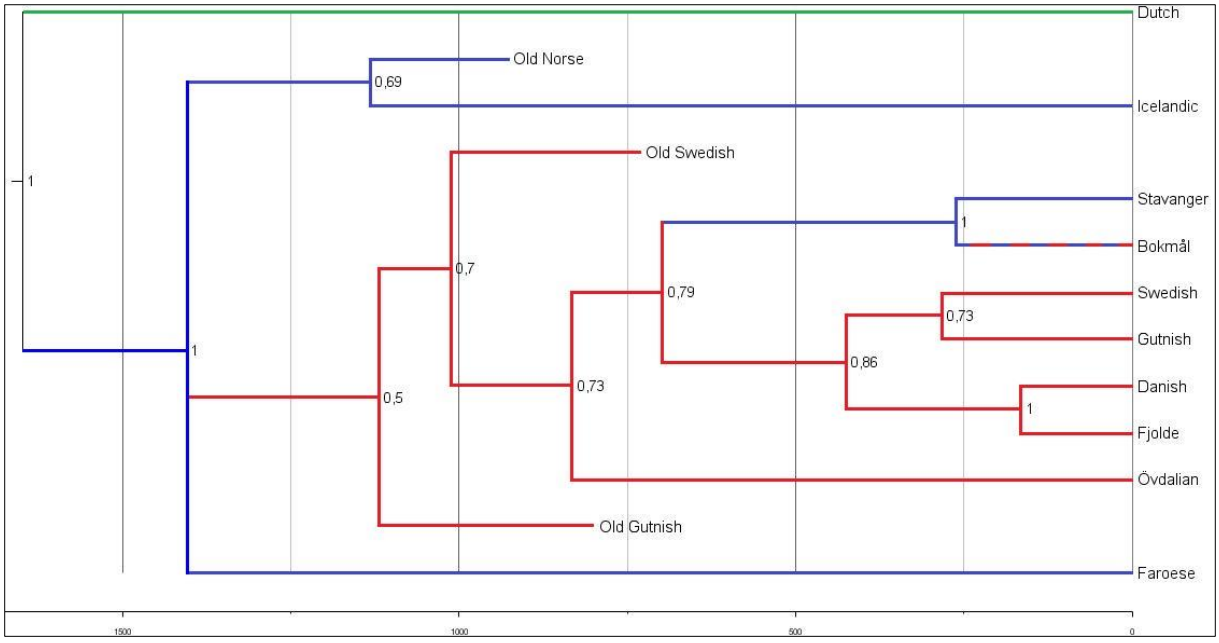

FIGURE 13. The consensus tree for the Swadesh data. Green edges denote West Germanic languages, blue West Scandinavian and red East Scandinavian. A time scale is added below the tree. The numbers at the nodes are the posterior probability values.

The tree in Figure 13 is not consistent with the traditional views, however. The East Scandinavian languages are clustered neatly together in one clade, but with Norwegian as an "intruder". The West Scandinavian branch has been torn apart and its members emerge in three different places. Two branches are considered to be direct splits from Proto-Nordic, one branch consisting of Old Norse and Icelandic and a separate branch for Faroese. The third branch, Norwegian, is found in a subgroup among the East Scandinavian varieties.

The older language stages are found in more or less expected places. Old Norse is clustered together with its descendant Icelandic. Old Gutnish and Old Swedish are believed to be the ancestors of all East Scandinavian languages. As the variation between the early East Scandinavian languages was rather small, their position in the tree seems quite sensible.

Even though the splits are not really in accordance with the conventional history of North Germanic, the dating of the splits does tally much better with traditional estimations. The

first split between West and North Germanic is dated around 300 AD, which lies perfectly within the range that has been proposed by historical linguists. The first splitting event within North Germanic is put around 600 AD, which also is in line with traditional reconstructions. The only real incongruities are found in the Swedish/Danish clade. The splits (SE | GUT) and (DK | DK_F) as well as (SE, GUT | DK, DK_F) are dated to have occurred much later than is traditionally assumed. This might be due to influence of the standard language on the dialects which might have caused the dialects to assimilate to the standard language.

The MCC tree for the same data is shown in Figure 14.

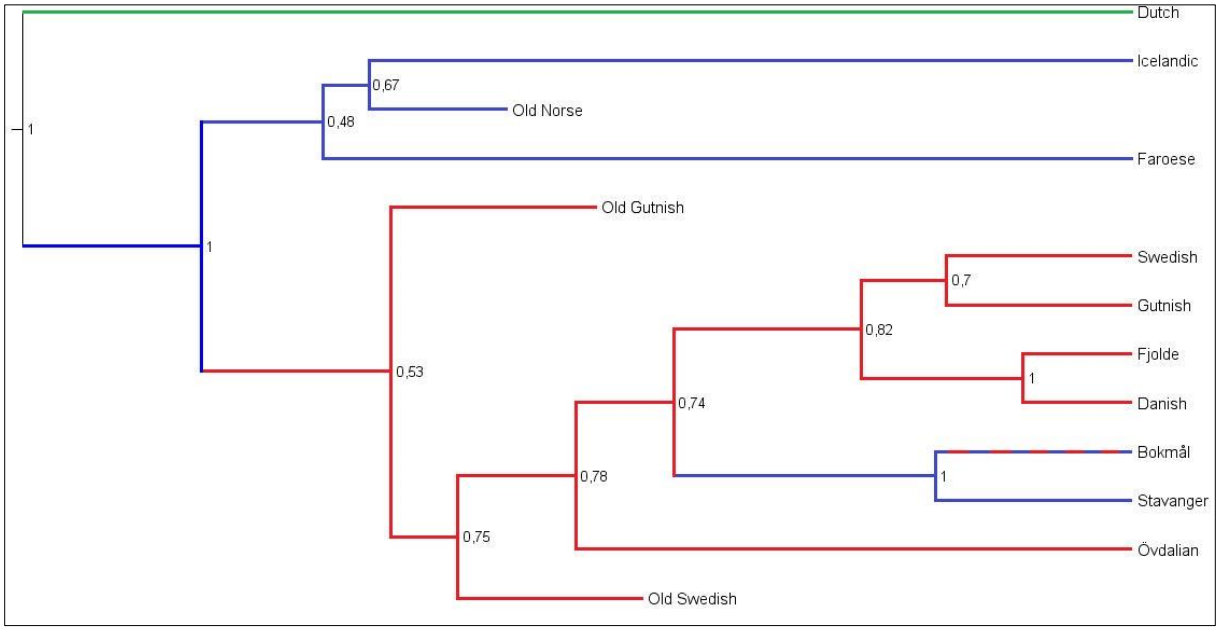

FIGURE 14. The MCC tree for the Swadesh data. Green edges denote West Germanic languages, blue West Scandinavian and red East Scandinavian. The numbers at the nodes are the posterior probability values.

The MCC tree shows by and large the same ordering as the consensus tree. The only difference is the position of Faroese, which is now clustered together with Icelandic and Old Norse, constituting a distinct West Scandinavian branch, in which Norwegian is still missing. Besides, Old Norse is only considered the ancestor language of Icelandic and not that of Faroese. The dating of the splits in the tree are more or less identical to those of the consensus tree.

The domestic animal tree

Figure 15 shows the consensus tree for the domestic animal lexical data. Judging from the colours of the edges, this dataset shows a phylogeny that is much more in accordance with conventional trees from historical linguistics.

A more thorough inspection of the tree confirms this first impression. The internal ordering of the two sub-branches, West and East Scandinavian, shows very much the same structure as conventional trees. Old Norse is considered to be the ancestor of all West Scandinavian languages, which is perfectly in line with what is known from historical linguistics. Then the West Scandinavian clade splits off in three sub-groups, Icelandic, Faroese and Norwegian, around 1150 which is quite close to general estimations for this split.

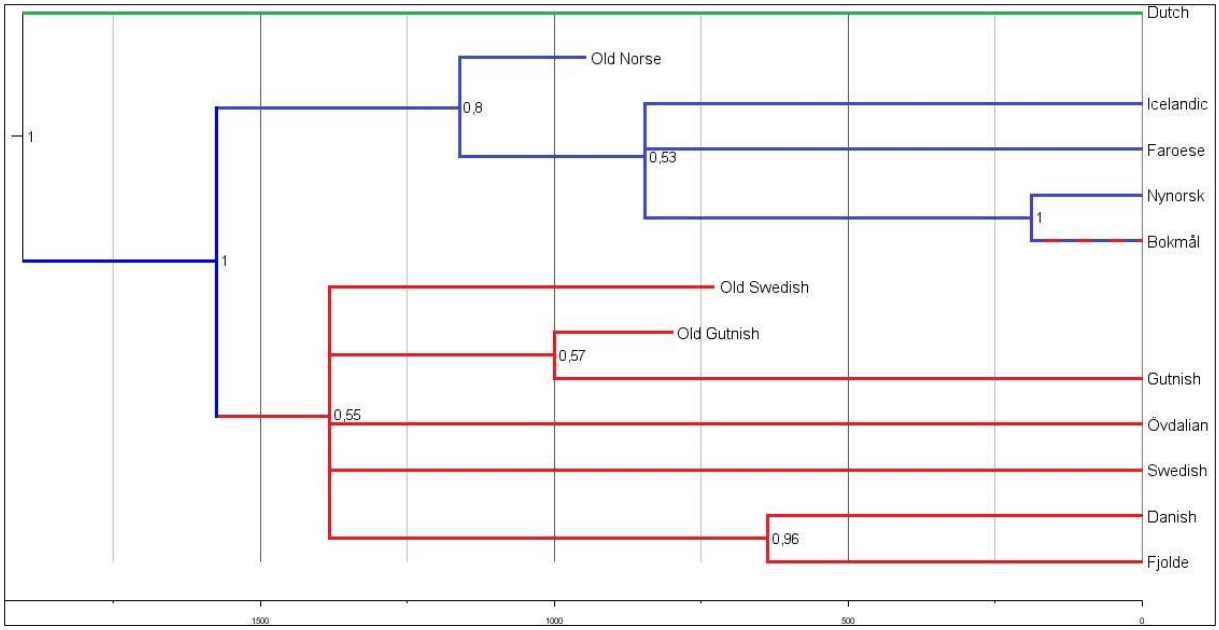

FIGURE 15. The consensus tree for the domestic animal lexical data. Green edges denote West Germanic languages, blue West Scandinavian and red East Scandinavian. A time scale is added below the tree. The numbers at the nodes are the posterior probability values.

The terminology for domestic animals shows a less unequivocal picture in the East Scandinavian languages. Old East Norse evolves into five different branches according to the MCC tree, three of which only consist of one language. The other two clades both contain two languages and represent the Gutnish and Danish branches. This is in fact the only tree that clusters Old Gutnish directly together with Gutnish. Old Swedish is not recognised as the ancestor language of Swedish. The dating for the first split in the East Scandinavian branch is much earlier than generally accepted. The same goes for the first split in the North Germanic branch as a whole and the split between North and West Germanic.

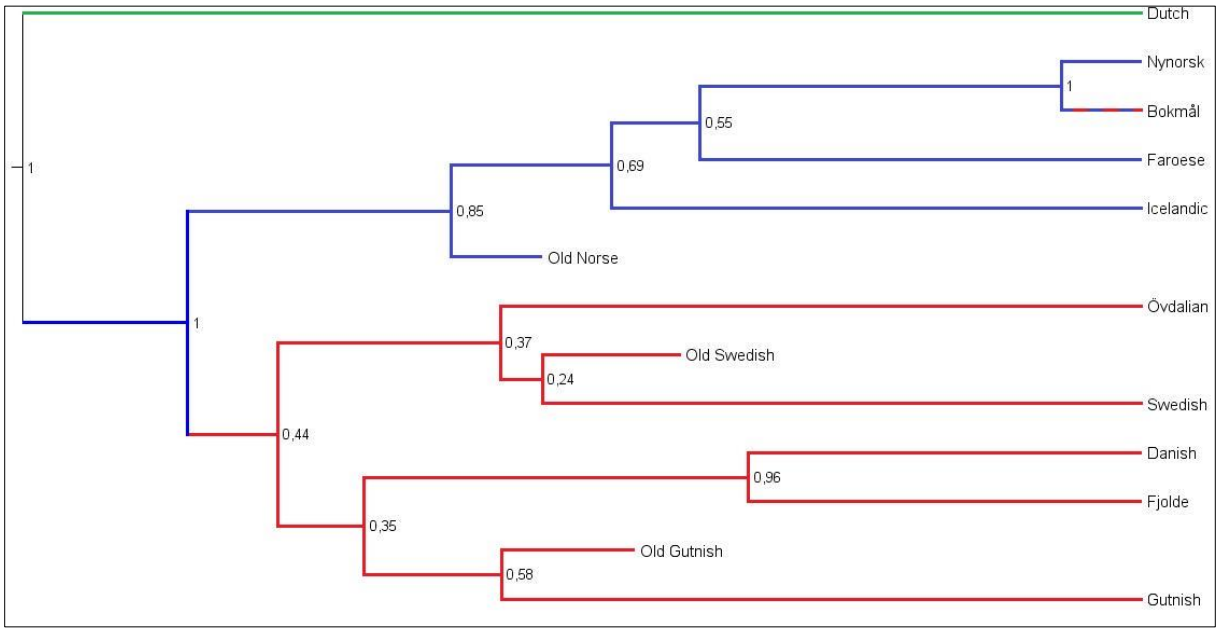

FIGURE 16. The MCC tree for the domestic animal lexical data. Green edges denote West Germanic languages, blue West Scandinavian and red East Scandinavian. The numbers at the nodes are the posterior probability values.

The MCC tree for this dataset shows a quite different picture than the consensus tree. The division between West and East Scandinavian is preserved, but the internal structure of both branches is much more detailed. Old Norse is still considered the ancestor language of the West Scandinavian branch, but the subsequent partition occurs in three steps now. First, Icelandic splits up from Faro-Norwegian and approximately 150 years later, the Faro-Norwegian branch splits up into a Faroese and a Norwegian branch.

The East Scandinavian clade consists of a Swedish and a Dano-Gutnish sub-clade. It should be noted, however, that the posterior probability value for either sub-clade is rather low. In other words, the evidence for the existence of exactly these two sub-clades is quite weak. This might explain the surprising clustering of Danish and Gutnish.

In the Swedish sub-clade, Old Swedish is considered to be only the ancestor language of Swedish and not that of Övdalian. This implies that Övdalian split off at very early stage in time, even before Old Swedish had developed from Proto-East Scandinavian. The Dano-Gutnish sub-clade splits off into a distinct Danish and Gutnish sub-branch. Old Gutnish is correctly recognised as the ancestor language of modern Gutnish. The dating for the splits in this tree are comparable to those in the consensus tree.

The phonological tree

The consensus tree inferred from the phonological data is shown in Figure 7. As becomes immediately clear from the figure, the fundamental split between West and East Scandinavian languages is perfectly represented.

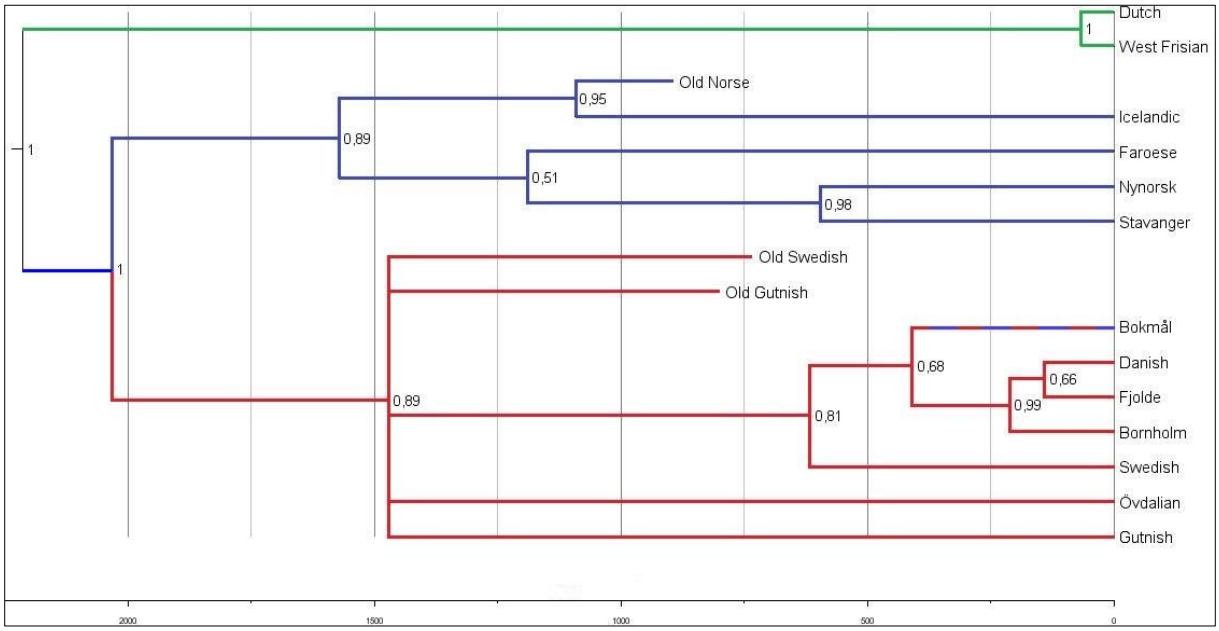

FIGURE 17. The consensus tree for the phonological data. Green edges denote West Germanic languages, blue West Scandinavian and red East Scandinavian. A time scale is added below the tree. The numbers at the nodes are the posterior probability values.

The root of the tree is dated to about 200 CE, which is rather early compared to estimations from historical linguistics. The split between West and East Scandinavian, which is put around 50 CE, occurs over 700 years earlier in this tree than is generally agreed upon. If we then zoom in on the West Scandinavian clade, we see the first divergence take place around 400 AD between (Old Norse and Icelandic) and (Faroese and Norwegian). Again, this split is dated very

early compared to conventional estimations and even entirely impossible as the speakers of West Scandinavian did not leave for Iceland before 800 AD. Moreover, Old Norse is classed as the ancestor of only Icelandic, despite it being the ancestor language of all West Scandinavian languages.

The first diversification among the East Scandinavian languages is estimated to have occurred around 550 AD. Even though this is quite early as well it is more sensible than the dating of the first split within West Scandinavian. The internal classification of the East Scandinavian languages turns out to be more complicated and problematic than in its western counterpart. The first split divides the family into five subgroups of which four only contain one language. Two of these sub groups are Old Swedish and Old Gutnish, which are not grouped together with their descendant languages Swedish and Gutnish. This is probably due to the fact that Swedish and Gutnish share some innovations with the other East Scandinavian tongues which were not yet present in Old Swedish and Old Gutnish.

The core cluster in the East Scandinavian branch, consisting of Swedish, Bokmål Norwegian and Danish and its dialects is also very remarkable. The first split within this clade is estimated to have occurred about 1400 AD, which is quite late as Old Danish and Old Swedish, spoken prior to 1400, were already showing variation. Another interesting observation is the position of Swedish in this tree. Intuitively, one would expect Swedish in a branch together with Övdalian and Gutnish, as they are generally considered dialects of Swedish or are at least spoken within the borders of Sweden. The statistics, however, prefer to group Swedish together with the other standard languages and the dialects of Danish. This suggests there was a considerable amount contact between the peoples of the most populated parts of Scandinavia which facilitated the transfer of phonological innovations beyond extant linguistic boundaries. These innovations never reached the fringes of the East Scandinavian continuum as is indicated by the positions of Övdalian and Gutnish in the tree.

A similar history of the East Scandinavian varieties was put forward by Hesselman (1936). He noticed that varieties from northern Sweden, Finland, the Baltic, Gotland and southern Denmark exhibited a number of similarities which were not found in the core area. He argued that it was the result of the existence of a prestigious language, which he initially called *birkasvenska* ‘Birka Swedish’, through which innovations spread within an economical core area. Later, he revised his theory and suggested the innovations spread from Hedeby. Hedeby was an economical centre, perched on the east coast of Jutland, close to the present day city of Schleswig in Germany. Widmark (1994) endorses Hesselman’s theory and provided additional evidence for the existence of what is now called Hedeby Nordic.

The dialects of Danish are also found in the area where Hedeby Nordic would have been spoken. This probably affected the accuracy of the dating, as the first diversification within Danish is estimated to have occurred 200 years ago in the tree. Runic inscriptions and early Danish texts already showed some fundamental differences between insular and mainland Danish. A-mutation, for example, was much more common in mainland Danish than in insular Danish, compare Fjolde *golər* and Danish *gulv* from Proto-Germanic **gulba* ‘floor’. Breaking, on the other hand, was mostly confined to the eastern varieties, compare Danish *jeg* and Fjolde *æ* from Proto-Nordic *eka* ‘I’. These innovations sprang and spread both before 1000 AD, so the existence of Hedeby Nordic might have obscured the early split of the dialect of Fjolde from insular Danish and a similar role might be assigned to standard Danish in the last centuries. As a result, the diversification within Danish is estimated to be more recent than it in fact is.

Worth noting as well is the classification of Gutnish. Instead of constituting a branch of its own, Gutnish is considered a member of East Scandinavian. The position of modern Gutnish within this clade could be explained by transfer of phonological material from standard Swedish, but that is not relevant for Old Gutnish and does not give an explanation of why Old

Gutnish is classified as East Scandinavian. The phonological data suggest a separate Gutnish branch has never existed.

In conclusion, the consensus tree built on phonological data presents an internal classification which very much resembles the conventional representation of the linguistic history of the Scandinavian languages. The estimations of dates of splitting events are completely different in most cases, however. A conceivable cause for the inaccuracies in the dating of the splits are continuous situations of language contact mainly between the mainland varieties and especially in the core area which comprises Denmark, the coastal areas of the southern half of Sweden and southeastern Norway. The extensive amount of contact between people from these areas caused the languages they spoke to become more similar even though they had already diverged from a common ancestor centuries before. Early splits might therefore seem older than they are and more recent splits even more recent than they in reality are.

Figure 18 shows the MCC tree for the phonological data.

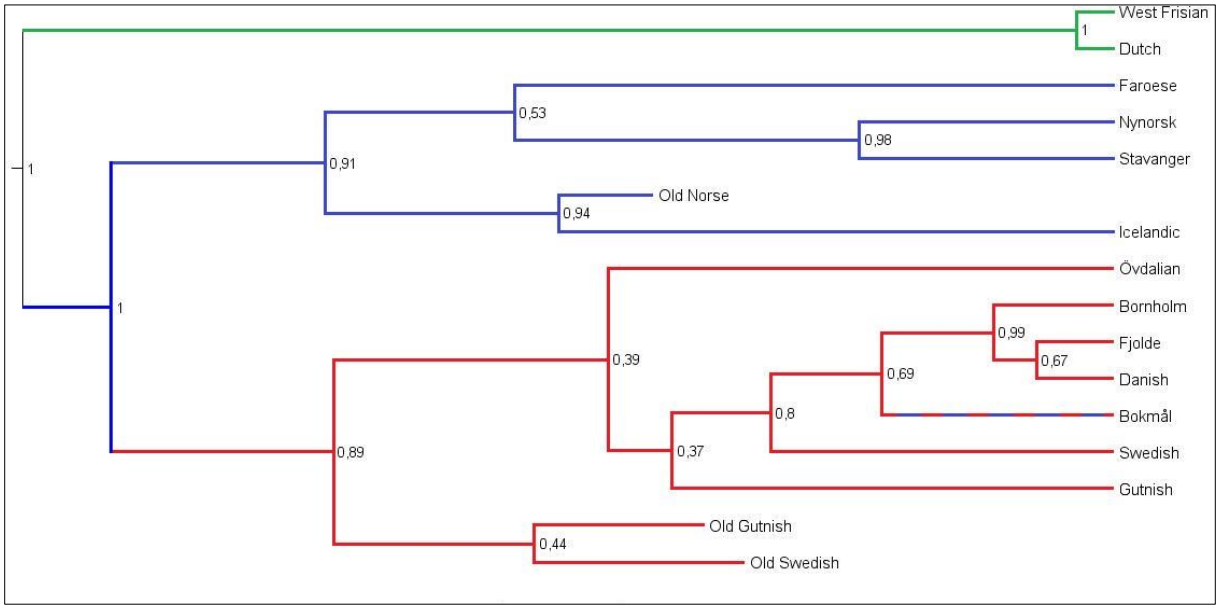

FIGURE 18. The MCC tree for the phonological data. Green edges denote West Germanic languages, blue West Scandinavian and red East Scandinavian. The numbers at the nodes are the posterior probability values.

The West-East division is still accounted for in this tree and the internal structure of the western branch is even identical to that in the consensus tree. The East Scandinavian branch shows some irregularities, though. The first split in that branch is the one between Old Swedish and Old Gutnish versus the other East Scandinavian varieties. This is probably due to later phonological innovations in East Scandinavian that were not yet present in Old Swedish and Old Gutnish. The first language to split off in the modern East Scandinavian sub-clade is Övdalian, followed by Gutnish, leaving a core group consisting of Swedish, Danish and eastern Norwegian. Again, this is compelling evidence for the Hedeby Nordic theory, especially when the high posterior probability value for the core group of 0.80 is taken into account. The earlier splits of Gutnish and Övdalian had much lower posterior probabilities, 0.37 and 0.39 respectively. Again, the dating in the MCC tree does not deviate remarkably from the dating in the consensus tree. Only for East Scandinavian, the first split is estimated to a more recent date, 1000 AD, if the split between Old Swedish, Old Gutnish and the modern varieties is ignored.

The morphosyntactic tree

The consensus tree built on morphosyntactic data is shown in Figure 19. The internal structure of the Scandinavian branch turned out quite undifferentiated. Again, the languages traditionally considered to be East Scandinavian are indicated by red edges and the West Scandinavian varieties are highlighted in blue. Clearly, the traditional division between West and East Scandinavian is not attested by the morphosyntactic data as the North Germanic branch branches off in eight branches instead of two.

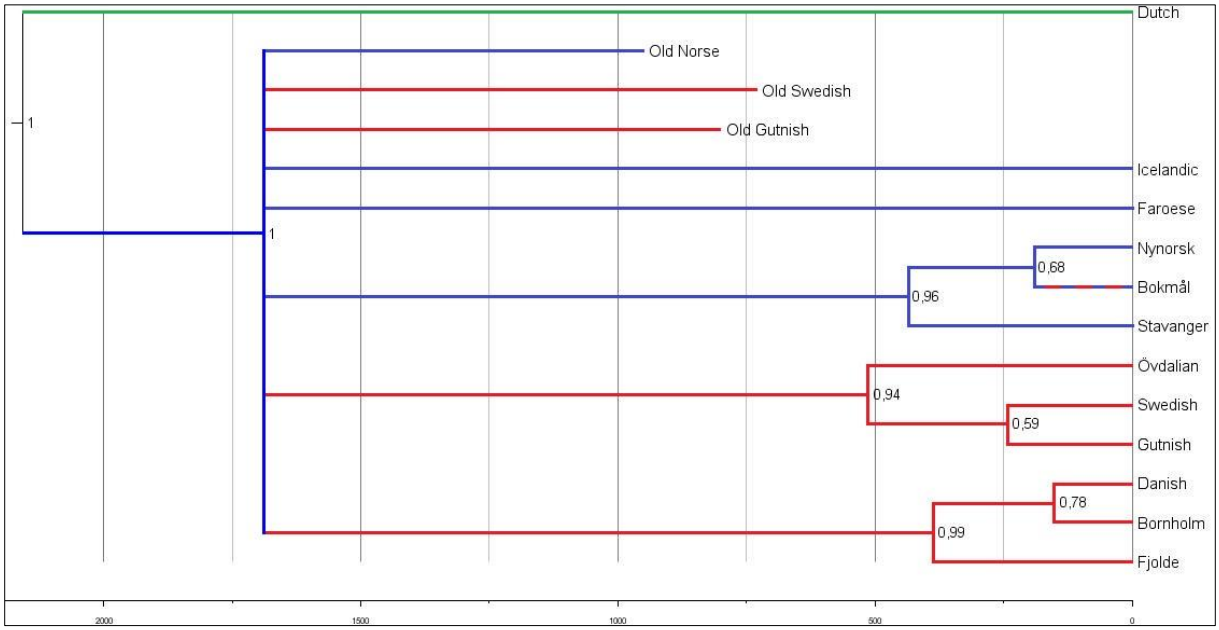

FIGURE 19. The consensus tree for the morphosyntactic data. Green edges denote West Germanic languages, blue West Scandinavian and red East Scandinavian. A time scale is added below the tree. The numbers at the nodes are the posterior probability values.

Although the tree does not show a picture in accordance with conventional historical descriptions, the internal structure of sub-branches containing more than one member does comply with what is expected. The three sub-branches at the lower half of the tree show separate Norwegian, Swedish and Danish clades. This suggests that the standard language might steer the morphological and syntactic development of all other varieties spoken within its "jurisdiction".

The one-member clades contain all the older languages as well as Icelandic and Faroese. The older languages might have ended up here as many of the innovations had not occurred yet at the time they were spoken or because the innovations they do have are common to North Germanic as a whole. Icelandic and Faroese are probably not clustered as the morphology and syntax of Faroese has been changed considerably under the influence of Danish. Danish has had even more impact on the morphology and syntax of Norwegian, so that might explain why a separate West Scandinavian branch could not be detected.

When it comes to the dating of the splits, the first splits are dated quite early. The split between West and North Germanic is put at 150 CE, which is more than half a millennium earlier than is usually thought. The first split within North Germanic is dated to 300 AD, about four centuries earlier than generally believed. The splits within Norwegian, Swedish and Danish are all concentrated in the last 500 years. Contrary to the other splits, these splits are dated quite late. This is probably caused by the influence of the standard language on the dialects.

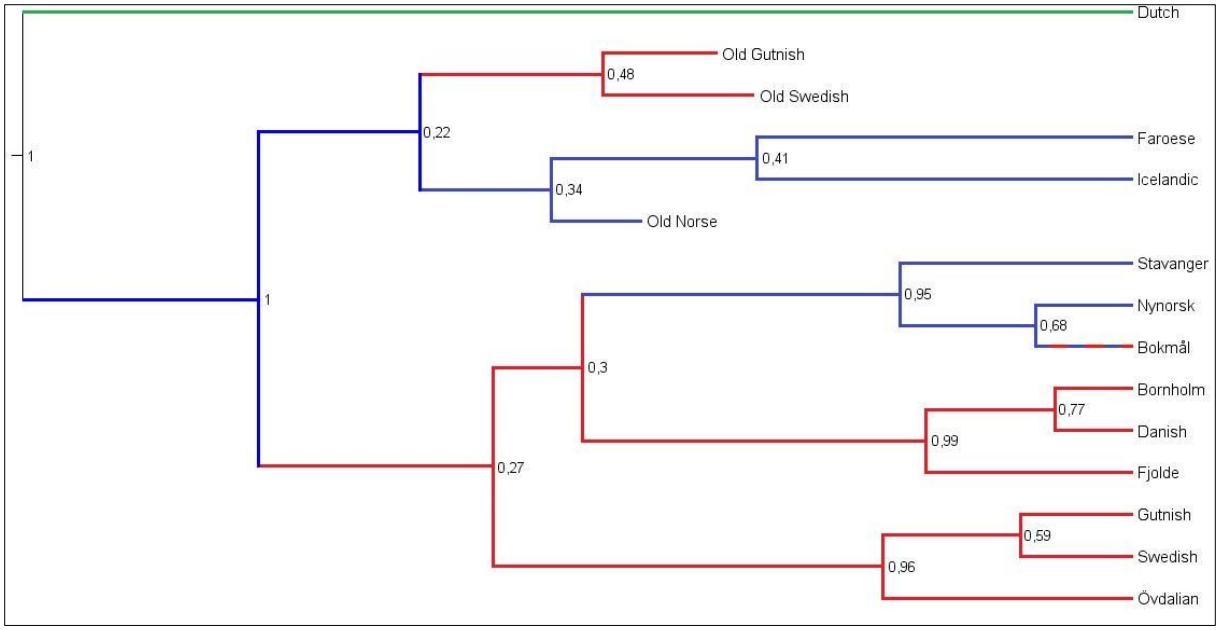

FIGURE 20. The MCC tree for the morphosyntactic data. Green edges denote West Germanic languages, blue West Scandinavian and red East Scandinavian. The numbers at the nodes are the posterior probability values.

The MCC tree for the morphosyntactic data in Figure 20 shows a rather different structure. It puts the first split within North Germanic at 300 AD. The branches sprouting from this split consist both of both West and East Scandinavian varieties, which means that genetics are unable to account for this split. The two branches are, however, not completely produced by chance, as the higher clade contains all the archaic varieties whereas all the innovative ones are found in the lower one. The evidence for these splits is very weak though, both have a posterior probability value below 0.3. The innovative branch still shows three separate clusters for the varieties of Norway, Denmark and Sweden. In the lower clade, Danish and Norwegian are classified as more closely related to each other than to the languages of Sweden. This is probably the result of imposition from Danish onto Norwegian through extensive language contact.

The overall tree

The consensus tree for the combined data is shown in Figure 21. As the colours of the edges indicate, the traditional West-East split is well represented in this tree. The internal organisation of the sub-branches do not surprise much either in comparison to the conventional family tree. The splits in the tree are dated somewhat earlier than generally accepted. The split between West and North Germanic is estimated to have occurred in the first century of our era. The first splits within Scandinavian are estimated to around 400 AD. Both a few centuries earlier than is generally accepted. The first split within West Scandinavian is put around 650 AD, predating the settlement of the Norwegian Vikings on the Atlantic Islands, hence earlier than possible. Also, the first split in East Scandinavian, circa 700 AD, is rather early. Again, the split between Danish and the dialect of Fjolde is considered to have occurred more recently than historical documents show.

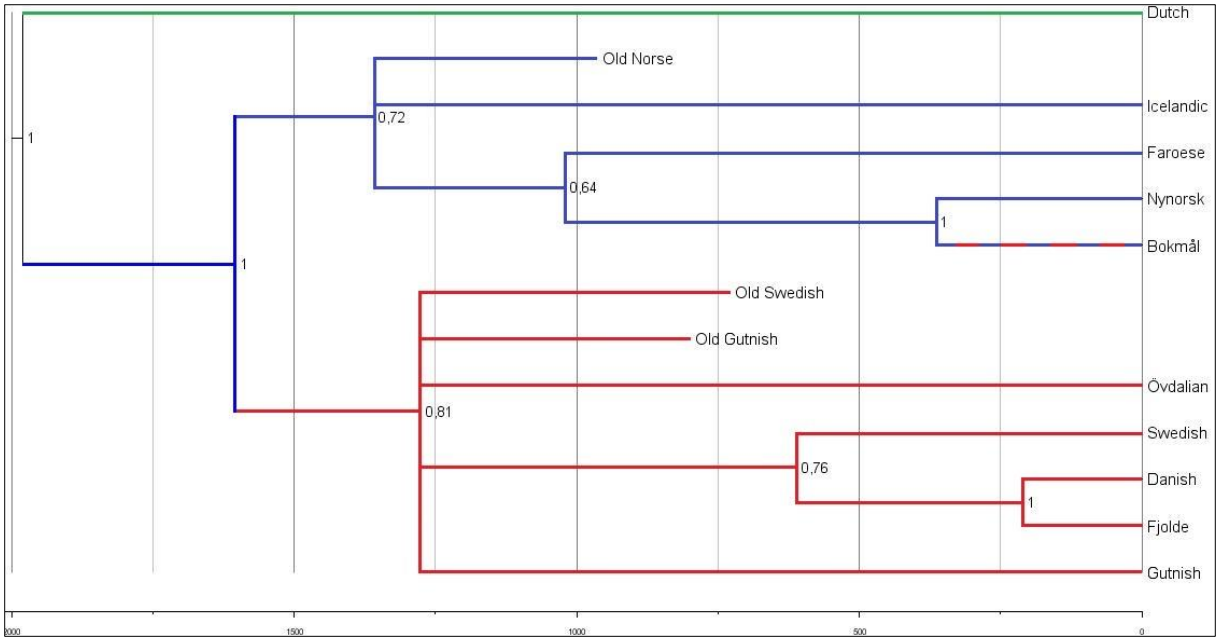

FIGURE 21. The consensus tree for the combined data. Green edges denote West Germanic languages, blue West Scandinavian and red East Scandinavian. A time scale is added below the tree. The numbers at the nodes are the posterior probability values.

The MCC tree for the overall data also shows the West-East split flawlessly. This tree resembles the consensus tree for the most part. The only really conspicuous difference is the split in the East Scandinavian branch between Danish and Swedish and the *Randmundarten* Övdalian and Gutnish. The classification in this tree also suggests a language like Hedeby Nordic might have existed.

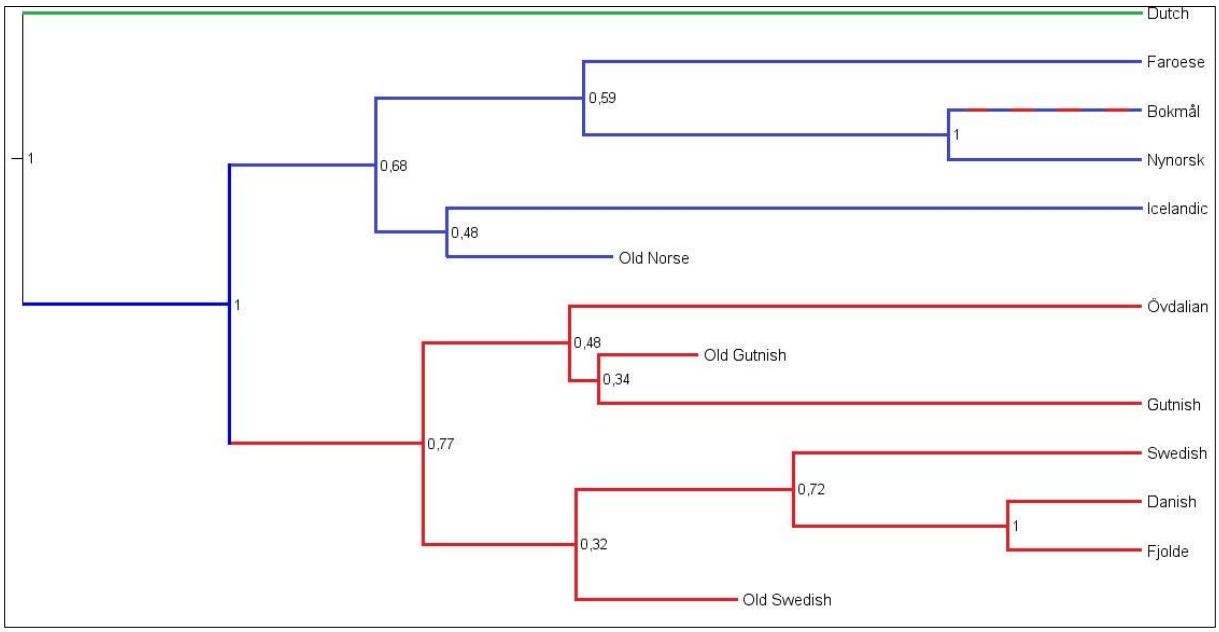

FIGURE 22. The MCC tree for the combined data. Green edges denote West Germanic languages, blue West Scandinavian and red East Scandinavian. The numbers at the nodes are the posterior probability values.

Discussion

The discussion of the results is divided into three sections. In the first section, the best dataset is presented accompanied with a discussion of why this particular dataset is exceptionally well-suited for the reconstruction of a language phylogeny. The central question in the second section is what this best tree actually tells about the history of the Scandinavian languages. In the third section the question whether trees or networks are better representations of language history is discussed. The last section is an exposition of the ramifications of this study for phylogenetic analysis of language in general.

The best dataset

Three of the five datasets used in this study returned a tree that shows the fundamental West-East split in Scandinavian. The two datasets that did not show this split, the Swadesh and morphosyntactic ones, do result in interesting trees, however. These trees do not show the genetic history of the North Germanic language family, but rather the convergence of the mainland varieties during the second half of the Middle Ages, which blurred the original West-East distinction and are thus good representations of the mutual intelligibility of the modern languages of Scandinavia. As they fail to show the fundamental split and are hardly compatible with conventional trees based on historical linguistics, these datasets are deemed unsuitable for phylogenetic analysis.

The inconsistencies between the trees has a number of implications. The inconsistencies might indicate that different parts of language do not evolve according the same evolutionary process. This means that not every part of language contains information about the genetic origin of the language. Perhaps phonological innovations are driven by other factors than morphosyntactic developments. Maybe morphosyntactic innovations are rather the result of geographical proximity or assimilation to a prestigious variety than common ancestry. The inconsistencies might also be explained by different rates of change of different parts of language. The phonological differences between the Nordic languages are much more significant than the syntactic ones for example. So, perhaps morphosyntactic change happens at such a slow rate that it is not capable of capturing the genetic signal and is consequently unsuitable data for reconstructions of such shallow time depths as that of North Germanic.

The disqualification of the Swadesh and morphosyntactic datasets leaves three at least rather good datasets: the domestic animal terminology, phonological and combined ones. However, the fact that the combined dataset also contains morphosyntactic characters, which apparently are no good indicators of the genetic history of the language family, disqualifies this dataset as well. The remaining domestic animal terminology and phonological datasets yield very comparable trees. There are some differences in the internal structure of mostly the East Scandinavian branch, and in the domestic animal terminology tree Bokmål Norwegian is considered to be West Scandinavian while it is classified as East Scandinavian in the phonological tree. Since all sources are inconclusive about the resolution of East Scandinavian after the Common East Scandinavian period and rather show the several East Scandinavian languages branch off contemporaneously from the common ancestor and as Bokmål Norwegian is often viewed as a hybrid of West and East Scandinavian, these differences are not very significant. A merit of the domestic animal terminology tree however, is that it correctly matches the old languages with their modern descendants.

The question then is what makes domestic animal terminology more suitable for phylogenetic analyses than phonological data? An advantage of lexical data over other types of data is that it is relatively easy to detect instances of horizontal transfer, i.e. loanwords, and

exclude them from the study. For identical phonological and morphosyntactic innovations it is much harder to be certain whether they are the results of common ancestry or horizontal transfer, especially in adjacent areas. The exclusion of loanwords turned out to be very essential in the analysis, as a second analysis that was run on the same dataset but with loanwords included resulted in a tree separating insular Scandinavian from mainland Scandinavian, exactly like the Swadesh and morphosyntactic data. Another advantage of lexical data is that a certain meaning or concept can in principle be expressed by an almost infinite number of possible words, whereas the number of possible phonological and morphosyntactic developments are much lower. This means that a certain phonological or morphosyntactic innovation is more likely to occur independently in different parts of the family.

This explains why domestic animal terminology contains fewer instances of horizontal transfer, but it does not reveal why the lexicon is a good predictor of the genetic language history. It has been pointed out by critics that phylogenetic analyses on lexical data basically measure the amount of shared retentions in the lexicon and that these do not tell anything about language history (Pereltsvag & Lewis 2015). Nonetheless, the method seems to work quite well. Heggarty (2010) explains the success of the method in the following way. He distinguishes three scenarios in the history of linguistic evolution. The first scenario is *shared retention*. Shared retention means that two or more language preserved for example a cognate which they inherited from a common ancestor. In this scenario, no change whatsoever occurred in this cognate class since the development from the common ancestor. Another scenario is *parallel innovation*. A parallel innovation is an innovation that happens independently in two or more languages and is consequently not the result of common ancestry. An example of this is the indefinite article, which has been developed in many of the Indo-European languages long after Indo-European was spoken. The last scenario is *shared innovation*. When an ancestor language introduced a new word for a given concept, then the presence of this word in the descendant languages is explained as being a shared innovation.

Heggarty shares the opinion that neither shared retentions nor parallel innovations are of any help in reconstructing a language phylogeny. He underlines, however, that the distribution of shared retentions and parallel innovations is random. So, all shared retentions and parallel innovations ideally balance each other out. The distribution of the shared innovations, on the contrary, is not random. If a language introduces a new lexeme for a certain meaning, all its daughter languages will have it as well, as long as they have not replaced it again. This means that shared retentions show the same patterns over and over again, which makes these recurrent patterns outweigh the random patterns of the shared retentions and parallel innovations. This also proves that the exclusion of loanwords is necessary, as they do not tend to be randomly distributed, because they usually stem from the same or a number of common sources for all languages.

A remark about shared retentions that Heggarty does not make is that some shared retentions do form evidence for common ancestry. Consider for example the PGmc word **hanhistaz* or **hangistaz* which meant ‘horse’ or ‘stallion’. The meaning ‘horse’ has been lost in West-Germanic languages, while all of the North Germanic ones, except for the Danish dialect of Fjolde, have retained this meaning. So, in cases like this the shared retention strengthens the pattern signalled by the shared innovations.

One could still question if the shared retentions are really balanced out in Scandinavian, as it is well known that Icelandic is extremely archaic and retentive whereas Danish at the other end has changed considerably and lost much of its North Germanic character. Considering this, one might argue that the split between West and East Scandinavian in the tree might have mainly been caused by the considerably larger number of shared retentions in the western varieties. In order to check this, the domestic animal terminology data were analysed again, but now all the cognate classes already found in the older languages, Old Norse, Old Gutnish and

Old Swedish were removed. The resultant tree still exhibited the split between West and East, though, and the posterior probability values in the West Scandinavian branch were even higher than those in the East Scandinavian clade. The structure of the tree on the whole dataset has thus not been significantly affected by the presumptive uneven spread of shared retentions in both branches.

Therefore, lexical data seem to be capable of signalling genetic ancestry but that does not explain why the Swadesh data did not yield such good results. I think this has to do with the nature of the meanings and corresponding lexemes on both the Swadesh and domestic animal list. Changes over time occurred in both datasets, but in the Swadesh dataset borrowing was much more common and probably the mechanism behind most of the changes. In the domestic animal terminology dataset borrowing is quite rare and most changes are the result of semantic shifts or new derivations. This difference might be explained by the way the lists are constructed. The domestic animal terminology list consists of meanings for ten species and some general animal terms. For each species usually distinct words are used for different ages and genders and sometimes even other characteristics. Within the semantic field of one species, new distinctions can be made and old distinctions can be lost. This means that new words are formed or old words alter meaning. It is not uncommon that more specific terms replace the generic term for a species. Examples of this are Gutnish *lamb*, originally 'lamb', which has come to mean sheep in general or Övdalian *takka* 'sheep' which originally meant 'ewe'. The reversed process also occurred a few times. Old Nors *hross* 'horse' which has become *hors* in Norwegian does no longer mean horse in general, but is only used for 'mare'.

Considering this, another dataset consisting of entities which also can easily acquire an either more or less specific meaning would give a similar result. Body parts seem to be a good candidate for such a dataset. A list of names of 67 body parts in ten Germanic languages was compiled to check this. This list also contained a number of words that originally had a more specific term, but which now describes a larger part of the body. Examples of this are the words *fótur* and *hönd* in Icelandic, which are cognates to English 'foot' and 'hand' which were the original meanings. In modern Icelandic, *fótur* can be used for either just the foot or the whole limb, so meaning foot plus leg. The same goes for *hönd* which means either just the hand or the arm and hand as a whole. Another good example of how easily words for body parts can alter their meaning are the lexemes used for 'cheek', which are tabularised in Table 9.

TABLE 9. The words for 'cheek' in four Germanic languages and their origin.

Language	Lexeme	Proto-Germanic form	original meaning
Swedish	kind	*kinnuz	cheek
Dutch	wang	*wanga/ōn	cheek, jaw
English	cheek	*kaukōn	jaw
Övdalian	annlit	ON: andlit	face

The Swedish word *kind* has continued denoting the same part of the face since Proto-Germanic. The reflexes of this word in West Germanic, English *chin* for example, have come to denote the adjacent part of the face, the chin. The Dutch word for cheek, *wang*, is a reflex of a Proto-Germanic word that referred to both the cheek and jaw, but which has become more restrictive in its meaning as it nowadays only means cheek. The English word *cheek* originally meant jaw and has through an intermediate stage in which it meant both cheek and jaw come to be used exclusively for the cheek. It is a cognate of Dutch *kaak* which still means jaw. In Övdalian, the word for cheek *annlit* used to denote the whole face but has been assigned a more restrictive meaning.

This would imply that the body part terminology dataset would produce a tree similar to that for the domestic animal terminology and one that tallies with received trees. The MCC

tree for the body part data is shown in Figure 23. Note that only modern standard Scandinavian languages and Övdalian were included so not all of the language in the sample of the domestic animal terminology dataset are analysed.

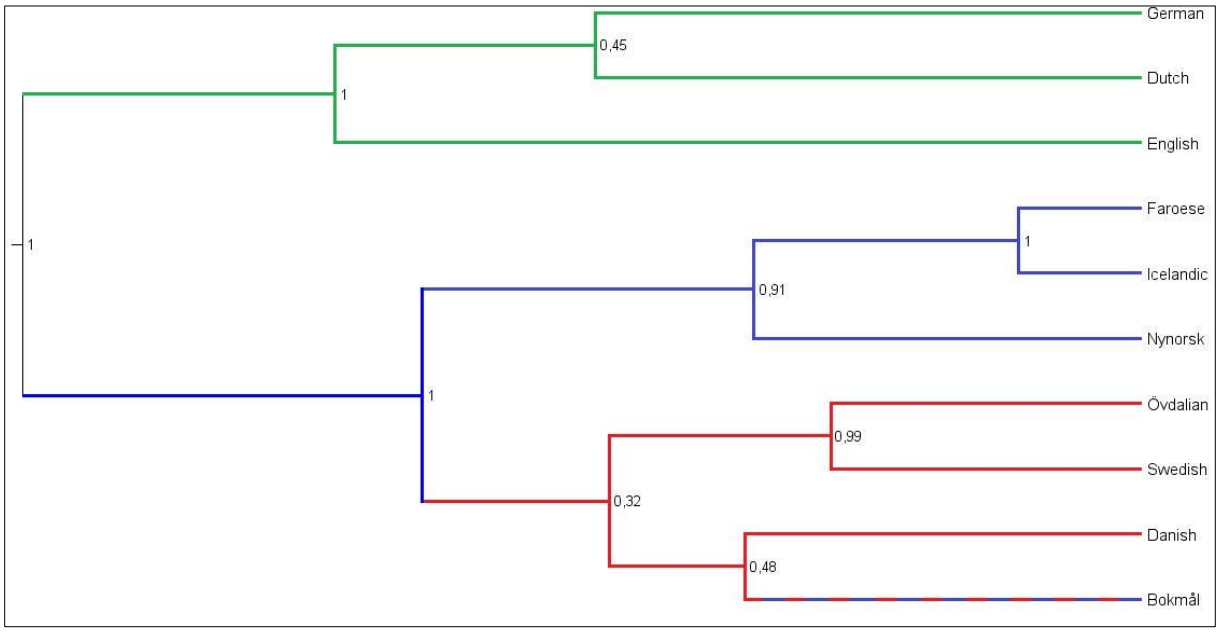

FIGURE 23. The MCC tree for the body part terminology data. Green edges denote West Germanic languages, blue West Scandinavian and red East Scandinavian. The numbers at the nodes are the posterior probability values.

The body part tree seems to be compatible with the domestic animal terminology tree. They mainly only differ in the classification of Bokmål Norwegian. In the domestic animal tree, Bokmål Norwegian turns up in the West Scandinavian clade, while it is considered to be East Scandinavian in the tree in Figure 23. However, as Bokmål is more or less a constructed language combining West Scandinavian and East Scandinavian features, it is not very surprising that it emerges in different clades in different trees. Another inconsistency with the domestic animal terminology tree is the internal structure of the West Scandinavian branch. Instead of Icelandic being the first language to split from the rest, Nynorsk Norwegian is the first one to branch off in this tree. This reversed order is probably caused by the exclusion of Old Norse, as the domestic animal terminology shows the same order when Old Norse is left out.

As some dialects and all of the older languages are missing in the body part tree, it is not possible to state conclusively that this tree is compatible with the domestic animal one, but the pattern it shows is promising and corresponds well to the received tree for the North Germanic language family. It is a good indication that semantic fields such as domestic animals and body parts are good indicators of genetic ancestry and are to be preferred to other parts of language as datasets for phylogenetic analysis. Moreover, both semantic groups, domestic animals and body parts, are also overrepresented in the preservation of irregular noun morphology in the North Germanic languages, together with kinship terms (Eekman 2013). This strengthens the assumption that these semantic groups carry a stronger ancestral signal than other ones. One should be cautious, though, with language families that have a longer history, as terms for domestic animals and body parts might change at such a fast rate, that they are not suitable for signalling deeper genetic relationships. However, kinship terms might be more stable over time, but further research is needed to confirm this.

What does the tree tell us?

It has been made clear now that domestic animal terminology is a suitable type of data to perform a phylogenetic analysis on. The next question is what the tree actually tells us about the history of Scandinavian. For practical purposes, the MCC tree for the domestic animal terminology is presented below again.

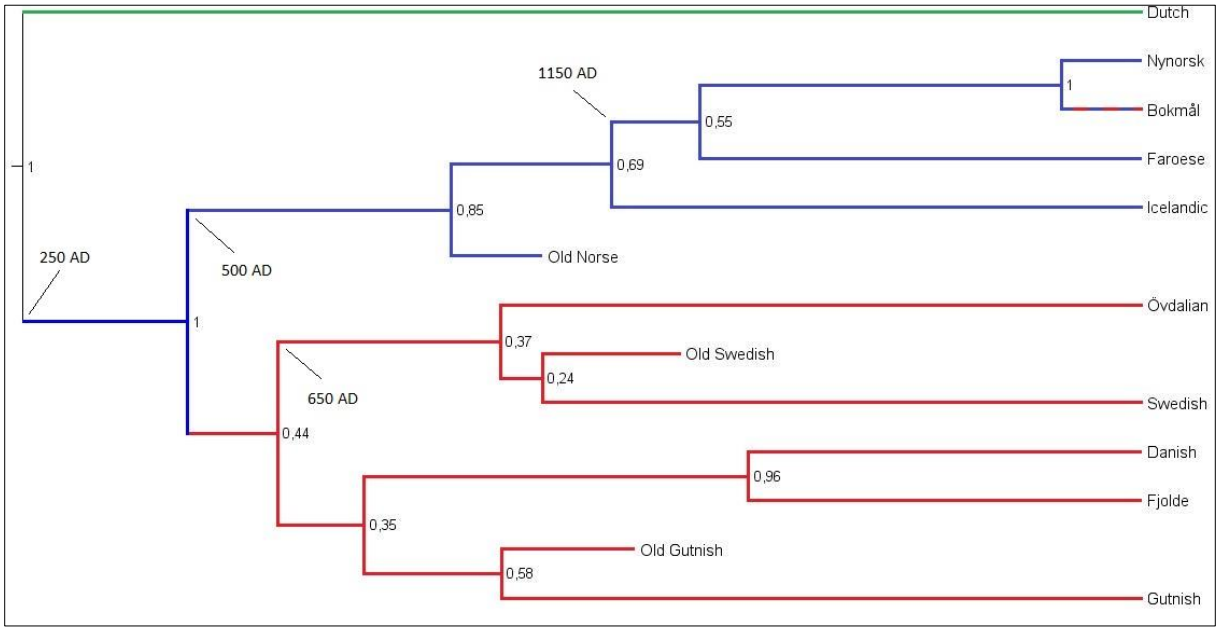

FIGURE 24. The MCC tree for the domestic animal lexical data. Green edges denote West Germanic languages, blue West Scandinavian and red East Scandinavian. The numbers at the nodes are the posterior probability values.

So, what does this tree tell about the dispersal of the North Germanic languages? Firstly, it confirms that a splitting event occurred in Proto-Nordic circa 500 AD, giving rise to Old West Norse and Old East Norse. The posterior probability values for both branches, 0.85 for West Norse and 0.44 for East Norse, show that there is stronger evidence for the clustering of the West Scandinavian languages. So, the signal for genetic ancestry is more unequivocal in the clade of West Scandinavian. That means that the eastern varieties have fewer shared innovations which indicate they began diverging rather quickly after the split between West and East Scandinavian which is estimated to have occurred by 500 AD. Evidence for this hypothesis is also found in the tree, as the first divergence within East Scandinavian is dated to 650 AD, whereas the first splitting event in West Scandinavian is put at 1150 AD.

The subsequent splits in the MCC tree have slightly lower posterior probability values and the consensus tree for these data hardly shows any further resolutions in the tree, apart from the most obvious ones: Bokmål and Nynorsk, Danish and Fjoldemål, Old Gutnish and Gutnish. This is indicative of conflicting signals in the data, which suggests a considerable amount of horizontal transfer, in particular in the East Scandinavian clade. Translated into real world events, this shows that after the initial split between West and East Scandinavian, innovations in especially East Scandinavian rather spread wave-like instead of tree-like. In other words, the East Scandinavian clade is a dialect continuum. Any real splitting events after the Old East Norse period are therefore not expected in the East Scandinavian clade. The splits in Figure 24 are mostly the result of the bifurcating structure forced on the data and most likely do not represent real-life splitting events in this part of the tree which is also indicated by the low posterior probability scores.

In the West Scandinavian languages, the evolution seems to have occurred much more in a tree-like fashion, which is indicated by the higher posterior probability values for the splits within this clade. This seems quite logical, as the West Scandinavian languages are not geographically adjacent as the East Scandinavian varieties are. Instead hundreds of kilometres of ocean separate the main West Scandinavian languages Icelandic, Faroese, Norwegian and Norn in former times. This impeded wave-like spread of features as the geographical distance limited the contact between the speakers of the various West Scandinavian tongues. This limited contact also paved the way for splitting events in this part of the Scandinavian language family.

In other words, the presence of natural barriers favours splitting events, while their absence leads to wave-like evolution and dialect continua. Natural barriers might also explain the first split of Proto-Nordic into Old West and Old East Norse. Maps of Scandinavia show that the landscape becomes much rougher and more mountainous on the west side of the border between the present-day Norwegian counties of Vest-Agder and Rogaland, which more or less coincides with the border between West and East Scandinavian. Although there was a continuous line of settlements along the coast of Norway, this suddenly much rougher terrain probably formed an obstacle for the contact between the peoples on either side of it. Furthermore, more inland the Scandinavian mountain range was even more of an obstacle for the contact between west and east.

In conclusion, a natural barrier probably caused the split between West and East Scandinavian. This in itself does not however tell which of the groups actually split off. From archaeology we know that the first speakers of Germanic lived in the southern parts of Scandinavia and from there spread to other parts of Europe, amongst which to the fjords of west Norway. The question is if this can also be inferred from the data or the tree and the answer is yes. Atkinson et al. (2008) found that languages evolve more quickly directly after a splitting event. Applied to Scandinavian, this means that the West Scandinavian languages should have undergone more change than the East Scandinavian ones. The rapid evolutional burst right after the splitting event would also imply that the West Scandinavian languages have more shared innovations as they evolved at a higher rate than the East Scandinavian ones. This is suggested by the higher posterior probability value for the West Scandinavian clade in the MCC tree. The more rapid evolution of the West Scandinavian branch is also shown in the NeighborNet split graph, in which the longer edges of the West Scandinavian languages represent a higher degree of innovation in that part of the language family. It should be noted, however, that part of the edge length might be the result of shared retentions in West Scandinavian as well.

Evolutionary splitting events often coincide with what biologists call genetic bottlenecks (Junk & Scherer 2006; Gould 2007). A genetic bottleneck is a sudden reduction in population size due to environmental catastrophes or human interference. In the case of the split between West Scandinavian and East Scandinavian, the migration of a small part of the population over the Scandinavian mountain ridge caused a genetic bottleneck, as the new population was much smaller than the original population of which they used to be a part. The assertion that the West Scandinavian community was much smaller is indeed confirmed by archaeological findings. Hans Andersson (2003) conducted research on the early beginnings of urbanisation in Scandinavia. The earliest towns were founded in eastern Scandinavia as is shown in Figure 25. The social organisation in Norway was different at that time. Instead of one town with all central functions, Norway consisted of central areas in which various central functions were spread over several places.

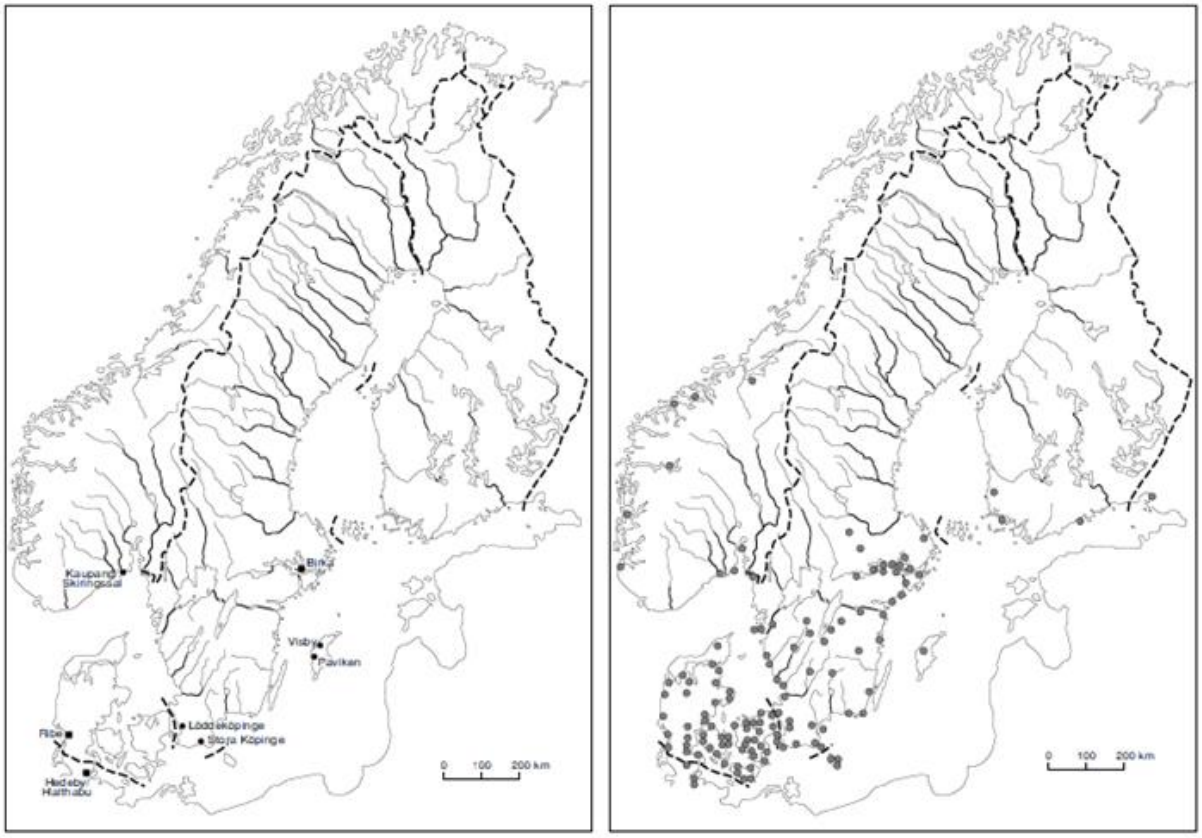

FIGURE 25. Left: Examples of early tendency towards urbanisation. Right: Density of towns 1150-1350 (Andersson 2003: 318 & 324).

An inspection of the map of the oldest towns in Scandinavia, reveals that the Norwegian fjords were still townless at that time. The first town to be established in the West Scandinavian language area was Trondheim around 990, much later than the split between West and East Scandinavian, which is estimated to have occurred by 500 according to the MCC tree. The other map which shows the density of towns between 1150 and 1350, shows that the number of towns in western Norway is considerably lower than in the eastern parts of Scandinavia. Both maps indicate that the population in the fjords of Norway was very small, much smaller than the population of southeast Scandinavia.

The sudden reduction of speaker population in west Norway also affected the language spoken by it. It has already been pointed out that the splitting event coincided with a punctuational burst of evolution, but apart from that, languages spoken by small communities are also more likely to develop more complex systems (Trudgill 2011). Hence, one would expect more irregularity in the language of the West Scandinavians compared to that of the East Scandinavians. As the oldest West and East Norse texts are runic inscriptions, which are few in number, we perforce have to compare later stages of the languages, such as Old Norse for West Scandinavian and the little younger Old Swedish for East Scandinavian (Barnes 2012).

In Table 10 the roots of the paradigm for the verb 'to fly' are given and counted. In Proto-Germanic, the paradigm contained four different roots. In Old Norse this number had increased to six. The two extra forms are the result of two phonological innovations. The first new root is found in the past participle. Proto-Germanic **fluganaz* had turned into *floginn* through a-mutation. A-mutation is a lowering of /u/ to /o/ caused by /a/ in the subsequent syllable. The other new root is found in the past tense of the subjunctive, where Proto-Germanic **flug*- had developed into *flyg*- as a result of i-mutation. I-mutation caused the rounding of the back vowels /u/, /o/, /a/ to /y/, /ø/ and /æ/ or /e/ before /i/ in the next syllable. Icelandic even

created a seventh root in the plural of the past tense subjunctive, which is also the result of a phonological process.

In Old Swedish, however, the original four roots had been reduced to three. The two Proto-Germanic roots for the present tense **fliug-* and **fleug* had merged into *flygh-* in Old Swedish. The root for the past participle in Old Swedish, then, is just as in Proto-Germanic identical to the root for the plural of the past tense indicative. A-mutation apparently did not affect this root in Old Swedish, which is not very surprising as a-mutation was much less frequent in East Scandinavian. The subjunctive had mostly disappeared in Old Swedish, so no new roots are to be found there. Even if we exclude the roots for the subjunctive, the West Scandinavian varieties still had five roots where Old Swedish had only three. Considering that Proto-Germanic had only four, it is safe to conclude that the verbal morphology of West Scandinavian had become more complex in at least this case.

TABLE 10. The paradigm for the verb 'to fly' in Proto-Germanic, Old Norse, Icelandic and Old Swedish. The numbers show how many stem changes each language has in its paradigm.

	PGmc	#	ON	#	IC	#	OS	#
PRS.SG	*fliug-	1	flýg-	1	flýg-	1	flygh-	1
PRS.PL	*fleug-	2	fljúg-	2	fljúg-	2	flygh-	1
PST.SG	*flaug-	3	flaug-	3	flaug-	3	flögh-	2
PST.PL	*flug-	4	flug-	4	flug-	4	flugh-	3
PP	*flug-	4	flog-	5	flog-	5	flugh-	3
SJV.PRS.SG	*fleug-	2	fljúg-	2	fljúg-	2	-	
SJV.PRS.PL	*fleug-	2	fljúg-	2	fljúg-	2	-	
SJV.PST.SG	*flug-	4	flyg-	6	flyg-	6	-	
SJV.PST.PL	*flug-	4	flyg-	6	flygj-	7	-	
Total		4		6		7		3
Total - SJV		4		5		5		3

The same tendency is found in the declension of nouns. Table 11 shows the declension for the word 'day' in Proto-Nordic, Old Norse, Old Swedish and Övdalian. The Proto-Nordic declension only has one root for all cases in both numbers: **dag-*. The same applies to Old Swedish which has *dagh-* throughout its paradigm. The Old Norse paradigm shows some variation, however. Although *dag-* is the most common stem, the dative singular has *deg-* and the dative plural *dǫg-*. Both stem alternations were caused by phonological processes. I-mutation had led to the rounding and raising from /a/ to /e/ in the dative singular. The form *dǫg-* is the reflex of u-mutation, the rounding of /a/ to /ɔ/ as a means of assimilation to /u/ in the following syllable.

Övdalian, the smallest East Scandinavian language with about 2000 speakers, also developed irregularity in this declension. Just as Old Norse Övdalian has three stems throughout the paradigm: *dag-*, *da-* and *dåg-*. These alternations are also the result of phonological innovations. The deletion of /g/ in *dag* occurs in front of certain vowels and syllabic /n/. The stem *dåg-* is the result of vowel balance.

TABLE 11. The declension of the word 'day' in Proto-Nordic, Old Norse, Old Swedish and Övdalian in the nominative, accusative, dative and genitive singular and plural.

	Proto-Nordic	ON	OS	OEV
NOM.SG	*dag-az	dag-r	dagh-er	dag-ø
ACC.SG	*dag-a	dag-ø	dagh-ø	dag-ø
DAT.SG	*dag-ē	deg-i	dagh-i	da-e
GEN.SG	*dag-as	dag-s	dagh-s	-
NOM.PL	*dag-ōz	dag-ar	dagh-ar	da-er
ACC.PL	*dag-an	dag-a	dagh-a	dåg-å
DAT.PL	*dag-umz	dǫg-um	dagh-om	dag-um
GEN.PL	*dag-ō	dag-a	dagh-a	-

The absence of mutated roots in Old Swedish does not mean that mutation processes never affected East Scandinavian. Mutation did occur in East Scandinavian, but at a much more moderate level. Reflexes of i-, a- and u-mutations are found, but interparadigmatic irregularities had already been resolved in Old Swedish, except for some instances of i-mutation in the plural of some nouns, such as *hand-händer* 'hand-hands'. An inspection of the Rök stone, a rune stone in Östergötland, Sweden from about 800 AD, shows that u-mutation was no longer used in the dative plural if it ever was. Instead of *sǫgum*, the inscription reads *sakum*, and we find *nabnum* instead of *nǫfnum* and *altum* instead of *ǫldum*.

I-mutation seemed to have been productive a bit longer, as in some Old Swedish texts, the usual *gar* 'go/walk' and *far* 'get' sometimes appear in rhyme with words ending in *-är* which suggests the original forms were *gär* and *fär*. This alternation is not found in any verbs with /u/ or /o/ in their stem. This does, however, not mean that i-mutation only occurred for /a/. In the Danish dialects of Jutland, the use of i-mutation in the present tense of verbs is currently declining, but the decline is gradual and mutated forms of /u/ and /o/ tend to disappear before mutated forms of /a/ in those dialects (Feilberg 1893-1914). Consequently, it is not inconceivable that i-mutation was originally more widespread in East Scandinavian, but that interparadigmatic irregularities were already lost at a very early stage.

This implies that additional complexity might have been formed in both the West and East Scandinavian languages, but more severely in the languages of the west as mutation operated on a larger scale there. The small West Scandinavian community could easily cope with this new complexity, while the larger East Scandinavian community could not and got rid of most of the irregular forms. This process of simplification probably intensified even more in the late Middle Ages, when the heyday of the Hanseatic league led to the migration of considerable numbers of Low Germans into Scandinavia. The high number of L2 speakers caused serious simplifications of the mainland Scandinavian languages, also of west Norwegian (Harbert 2007). Trading also brought the speakers of various mainland Scandinavian varieties into contact with each other which probably resulted in the convergence of the languages spoken in the main towns and cities.

In summary, the tree tells us a splitting event occurred in Proto-Nordic around 500 AD, which resulted a separate West and East Scandinavian branch. The higher rate of change in the West Scandinavian clade indicates that the speakers of that branch split off from the other speakers of Proto-Nordic. This also presupposes that the West Scandinavian population was small, which is supported by archaeological and linguistic evidence. Further splitting events in the West Scandinavian clade resulted in separate branches for Icelandic, Faroese and Norwegian. East Norse developed into a dialect continuum in which no further splits occurred. Gutnish is also part of this continuum, which means that a distinct Gutnish branch in opposition to the West and East Scandinavian branches probably never existed.

Trees or networks?

In this study, two types of models were shown for each dataset: trees and NeighborNet split graphs. Both are visualisations of the development of the North Germanic languages, but they visualise this history in different ways. As is generally known, languages diverge according to two discrete patterns. The first pattern, a splitting event, requires a split within a population. The most common case is that a group within the population migrates over a long distance and loses all contact with the rest of the population. The other pattern, wave-like development, occurs when a population expands over a continuous area. The contact situation is maintained in this pattern, but the increased distances between the groups of speakers makes it harder for innovations to spread through the entire population. Innovations, therefore, become more local, and the differences between the languages of the groups within the population become gradual.

Trees presuppose that language only evolves along the first pattern. This is quite a disadvantage of trees as the second pattern is more common in language evolution (Heggarty et al. 2010). As trees force a bifurcating structure on the data, the outcome only shows splitting events even when the languages in reality form a continuum which means that the evolution proceeded in a wave-like fashion. To some extend this argument is completely legitimate, but trees do give some indications that the splits it shows have no anchoring in the real world. In MCC trees, this is mainly shown by low posterior probability scores and in consensus trees also by unresolved branches in the tree.

Networks such as NeighborNet split graphs are capable of visualising both patterns of evolution. Splitting events are denoted by long parallel edges that are not intersected while reticulations are indicative of wave-like innovations. Even though networks are capable of expressing both patterns, they do have some limitations. Consider the innovations in the dialect continuum in Figure 26 for example.

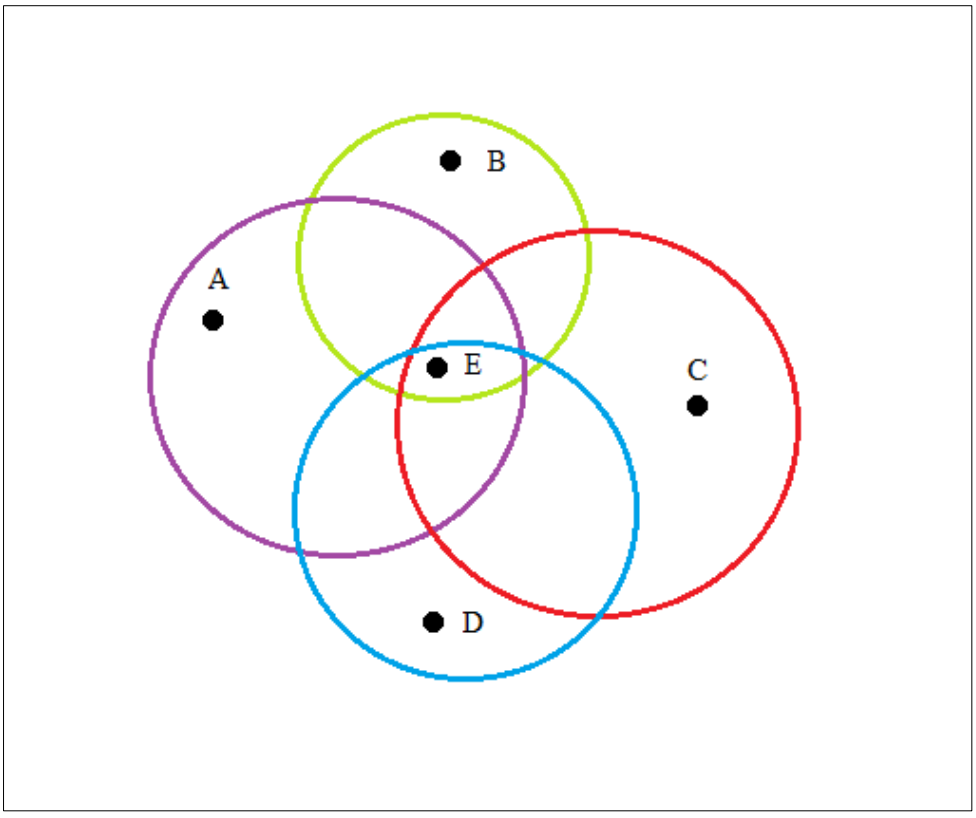

FIGURE 26. A fictive dialect continuum. The circles symbolise linguistic innovations. The dots are villages.

Four innovations occurred in this fictive dialect continuum. All of these innovations spread to village E. This means that village E stands linguistically in the middle of villages A, B, C and D. In a network, the variety of village E can only stand between two other varieties, while in reality it is in the middle of four varieties. Hence, a case like this cannot be expressed by either a network nor by a tree.

Another shortcoming of networks is that they are not capable of dating splits, whereas trees are. However, dating only makes sense if the splits in the tree also represent a split in the real world. Dating estimations of forced splits that do not have a real world equivalent, are completely meaningless, of course. So, in conclusion both models have their merits and demerits and the best results are probably attained if both visualisation models are analysed. Networks are good indicators of the tree-likeness of the data and detectors of horizontal transfer or wave-like evolutionary patterns in the data. The merit of trees is that the methods behind them are capable of dating the splits and creating a timeline of the history of the language phylogeny.

What are the ramifications for phylogenetic analysis of language in general?

Although this study focuses exclusively on the Scandinavian languages, its findings can be extended to phylogenetic analysis of language in general. The most serious finding for phylogenetic practice is that the Swadesh list, at least for Scandinavian, is a quite poor predictor of genetic history. This is not very astonishing considering that its validity has never been really tested. Nevertheless, it is a quite serious observation since the majority of phylogenetic analyses are based on a Swadesh list of some sort. To establish its validity and maybe find a better alternative, the Swadesh list and other lists, such as the Leipzig-Jakarta list (Tadmor 2009), the domestic animal list used in this study or a combinatory list of domestic animal, body part and kinship terminology should be tested for language families for which the histories are well documented, such as some the sub-branches of Indo-European. The unreliability of the Swadesh list has also serious consequences for the phylogenies proposed on the basis of phylogenetic analysis for language families of which the history is unknown. The results presented in such studies might be further away from reality than is generally thought.

Another observation that poses a problem for language families without documented history is the influence of loanwords on the genetic signal. Analyses in which loanwords were included revealed that a low number of loan words in the data can already blur subtle distinctions in the history of a language family. Therefore, it is of utmost importance to filter out loanwords before running an analysis. This might not always be possible for language families of which we have no historical records. So, when working with such languages one should bear in mind that certain finesses or sub-groupings might have been caused by the noise of loanwords and on should be very cautious about drawing conclusions from such analyses. A good alternative for such language families might be using phonological developments as input data, as phonological data turned out to be almost as good a predictor of ancestral relationships as lexical data.

Lastly, one should not simply present a tree as the sole and only correct phylogeny. Every time one deduces a phylogenetic tree, it is important to check if the language family and every subdivision of it really evolved in a tree-like fashion. As I pointed out in this study, unresolved trees and low posterior probability values are indicators of a wave-like evolutionary pattern instead of a tree-like one. Also, creating and analysing NeighborNet split graphs might help one to discover what pattern was dominant in every part of the history of a language family. Holden and Gray (2006) provide a good example on how one can detect dialect continua and

rapid radiation in NeighborNet split graphs with material from Bantu languages and how phylogenetic analyses are not capable of detecting such events and showing them in a tree. Syrjänen et al. (2013) also take branch lengths into account in their analysis of the Uralic languages and question the position of some splits where the branch lengths are very short and the posterior probability values rather low. Such phenomena should be checked for in order to capture the genetic history of a phylogeny as good as possible, before one draws too bold conclusions from a phylogenetic tree.

Appendix

Swadesh

#	lexeme	description
1	all	plural adjective “all”
2	and	conjunction
3	animal	generic term for animal
4	ashes	result of combustion of wood, grass or dry dung
5	at	generic locative, no spatial specification
6	back	of human
7	bad	deleterious, unsuitable
8	bark	of a tree; applicable to as many trees as possible
9	because	infers causality
10	belly	part of the human located directly above the pelvis
11	big	of size
12	bird	the most neutral term for bird
13	bite	of animal or people
14	black	the natural colour of coal
15	blood	of human blood
16	blow	transitive
17	bone	general term for human and animal bones (anatomical)
18	breast	frontal part of the human body, located between the neck and the upper belly
19	breathe	take air into the lungs and expel it
20	burn	both intransitive and transitive
21	child	a young human being
22	cloud	white cumulous clouds
23	cold	the regular antonym of hot
24	come	of humans: to reach a destination
25	count	enumerate
26	cut	as in “to cut one’s finger”
27	day	period of 24 hours, reckoned from one midnight to the next
28	die	non-violent human death
29	dig	as in “to dig a hole”
30	dirty	not clean
31	dog	domesticated dog
32	drink	to take a liquid into the mouth and swallow
33	dry	antonym of wet; applicable to as many objects as possible
34	dull	of a knife that is not sharp
35	dust	dust resulting from dry earth
36	ear	the organ of hearing and balance
37	earth	earth as soil
38	eat	put food into the mouth and chew and swallow it
39	egg	an oval or round object laid by a female bird, reptile, fish or invertebrate, usually containing a developing embryo
40	eye	each of a pair of globular organs of sight in the head of humans
41	fall	generic falling, i.e. inadvertent transfer of an object from point A to point B
42	far	antonym of near
43	fat	animal fat as general substance or used for eating or various economic needs
44	father	male parent, avoid baby talk
45	fear	be afraid
46	feather	most general term for feather of a bird
47	few	antonym of many
48	fight	to be in a physical confrontation

49	fingernail	the flattish horny part on the upper surface of the tip of each finger
50	fire	a process in which substances combine chemically with oxygen from the air and typically give out bright, light, heat and smoke; combustion or burning
51	fish	living fish
52	five	cardinal number five
53	float	rest or move on or near the surface of a liquid without sinking, non-human
54	flow	(of a liquid, gas or electricity) move steadily and continuously in a current or stream
55	flower	generic term for flower
56	fly	a neutral movement through the sky, of birds
57	fog	a thick cloud of tiny water droplets suspended in the air at or near the earth's surface
58	foot	the part of the leg below the ankle joint
59	four	cardinal number four
60	freeze	to turn into ice
61	fruit	the sweet and fleshy product of a tree or other plants that contains seeds and can be eaten as food
62	full	full state of a vessel
63	give	freely transfer the possession of something to someone
64	good	combinable with as many objects as possible
65	grass	generic term for grass
66	green	the natural colour of fresh, newly sprung grass
67	guts	generic term for all the organs in the abdomen
68	hair	collective, applicable to hair of humans
69	hand	the body part on the arm below the wrist
70	he	third person singular masculine pronoun
71	head	the upper part of the human separated from the rest of the body by the neck
72	hear	perceive sound with the ear
73	heart	the hollow muscular organ that pumps blood through the circulatory system, of humans
74	heavy	of large mass
75	here	antonym of there
76	hit	bring one's hand into contact with someone or something, quickly and forcefully
77	hold	as in "hold one's hand"
78	horn	applied to the prevailing/endemic species in the region
79	how	interrogative pronoun, in which way
80	hunt	pursue and kill a wild animal
81	husband	a married man considered in his relationship to his spouse
82	I	first person singular pronoun
83	ice	frozen water
84	if	introducing a conditional clause
85	in	located inside
86	kill	cause the death of something, by an animate agent
87	knee	the knee joint, outside part
88	know	expresses previous knowledge of a situation
89	lake	a large area of water surrounded by land
90	laugh	make the spontaneous sounds and movements of the face and body that are the instinctive expressions of lively amusement and sometimes also of derision
91	leaf	applicable to as many plants as possible
92	left	antonym of right
93	leg	each of the limbs on which a human walks and stands
94	lie	a verb of continuative/durative action
95	live	to inhabit

96	liver	body organ
97	long	of physical objects
98	louse	as a human parasite
99	man	male human
100	many	applicable to countable objects
101	meat	meat as food
102	moon	the natural satellite of the earth
103	mother	female parent, avoid baby talk
104	mountain	a large elevation of the earth's surface rising abruptly from the surrounding level
105	mouth	as a part of the face
106	name	a word or set of words by which a person is known
107	narrow	antonym of wide
108	near	a short distance away
109	neck	body part between the head and the rest of the body
110	new	just made, just acquired
111	night	the period devoid of sunshine
112	nose	the part projecting above the mouth on the face of a person, containing nostrils
113	not	the basic negation of an assertion
114	old	antonym of young
115	one	cardinal number one
116	other	used to refer to a person or a thing that is different or distinct from one already mentioned or known about
117	person	basic term for a human being
118	play	engage in activity for enjoyment and recreation
119	pull	exert force on someone or something so as to cause movement towards oneself
120	push	exert force on someone or something so as to cause movement away from oneself
121	rain	the condensed moisture of the atmosphere falling visibly in separate drops
122	red	the natural colour of arterial blood
123	right	true or correct as a fact
124	rightside	antonym of left
125	river	general term for a natural waterway
126	road	most typical kind of walkway for the region
127	root	generic term referring to the roots of trees
128	rope	a length of thick strong cord
129	rotten	of trees
130	round	two-dimensional round
131	rub	apply firm pressure to the surface of something, using a repeated back and forth motion
132	salt	as a natural product
133	sand	a loose granular substance as found on beaches
134	say	a neutral designation of a single speech act
135	scratch	with fingernails to relieve an itch
136	sea	the expanse of salt water that covers most of the earth's surface
137	see	durative process of perception with eyes
138	seed	generic term for seeds of any plant
139	sew	join, fasten or repair something by making stitches with a needle
140	sharp	antonym of dull
141	short	of physical objects
142	sing	make musical sounds with the voice
143	sit	static verb
144	skin	human skin
145	sky	the region of the atmosphere
146	sleep	static verb

147	small	antonym of big
148	smell	transitive with experiencer object
149	smoke	smoke as result of the combustion of wood, grass or dry dung
150	smooth	having an even and regular surface
151	snake	generic term for snake
152	snow	atmospheric water vapour frozen into ice crystals and falling in white light flakes
153	some	a couple of
154	spit	eject saliva forcibly from one's mouth
155	split	of wood
156	squeeze	firmly press something soft
157	stab	thrust a knife or other pointed weapon into someone as to wound or kill
158	stand	static verb
159	star	a fixed luminous point in the night sky
160	stick	a long, thin piece of wood used for support in walking or as a weapon
161	stone	an average sized stone that can be taken into one's hand
162	straight	not bent
163	suck	draw into the mouth by making a partial vacuum
164	sun	the star round which the earth orbits
165	swell	become larger in size as a result of an accumulation of fluid
166	swim	move through water, of humans
167	tail	the hindmost part of an animal, prolonged beyond the rest of the body
168	that	demonstrative pronoun with adjectival functions
169	there	antonym of here
170	they	third person plural masculine pronoun
171	thick	with opposite sides or surface that are far or relatively far apart
172	thin	of objects
173	think	direct one's mind towards someone or something
174	this	antonym of that
175	thou	second person singular
176	three	cardinal number three
177	throw	propel something with force through the air by a movement of the arm and hand
178	tie	fasten something by the means of strings
179	tongue	the fleshy muscular organ in the mouth of humans
180	tooth	generic term for human teeth
181	tree	generic term, applicable to as many trees as possible
182	turn	move in a circular direction
183	two	cardinal number two
184	vomit	throw up
185	walk	move at a regular pace never having both feet off the ground at once
186	warm	of or at a fairly or comfortably high temperature
187	wash	clean with water
188	water	sweet water, good for drinking
189	we	first person plural pronoun
190	wet	antonym of dry
191	what	interrogative pronoun referring to an inanimate object
192	when	interrogative pronoun, at what time
193	where	interrogative pronoun, at what place
194	white	the natural colour of snow
195	who	interrogative pronoun referring to an animate subject
196	wide	antonym of narrow
197	wife	married woman in relationship to her husband
198	wind	generic term for wind as movement of air

199	wing	in a bird, a limb used for flying
200	wipe	clean or dry something by rubbing a cloth
201	with	in the sense of accompaniment
202	woman	adult human being of the female sex
203	woods	forest
204	worm	earthworm
205	you	second person plural pronoun
206	year	the time taken by the earth to make one revolution around the sun
207	yellow	the natural colour of egg yolk

Domestic animals

#	species	specification
1	goat	generic term
2		female
3		male
4		generic newly born
5		female newly born
6	sheep	generic term
7		female
8		male
9		generic young
10		female young
11		male young
12		generic newly born
13		female newly born
14	bovine cattle	generic term
15		female
16		male
17		hornless
18		not producing milk
19		generic young
20		female young
21		male young
22		generic newly born
23	pig	generic term
24		female
25		male
26		female young
27		male young
28		generic newly born
29		male newly born
30	horse	generic term
31		female
32		male
33		intractable
34		old
35		mount
36		generic newly born
37		female newly born
38		male newly born
39		herd
40	chicken	generic term
41		female
42		male
43		generic newly born

44	duck	generic term
45		female
46		male
47		generic newly born
48	goose	generic term
49		male
50		generic newly born
51	dog	generic term
52		female
53		male
54		generic newly born
55	cat	generic term
56		female
57		male
58		newly born
59	cattle	generic term
60		small
61		one year old
62	castrated animal	generic term
63	beast of burden	generic term

Phonology

#	item	example
1	ww > ggw in PGmc **hawwan*	ON: *hǫggva*
2	a-mutation (u > o) in PGmc **hurna*	ON: *horn*
3	a-mutation (u > o) in PGmc **fulka*	ON: *fólk*
4	i-mutation (a > e) in PGmc **gastiz*	ON: *gestr*
5	z > r in PGmc **maizo*	ON: *meiri*
6	consonant gemination before /j/ in PGmc **ligjanan*	ON: *liggja*
7	loss of initial /j/ before vowels in PGmc **jera*	ON: *ár*
8	loss of initial /w/ before velar vowels in PGmc **wurða*	ON: *orð*
9	assimilation of ht > tt in PGmc **nahts*	ON: *nátt/nótt*
10	assimilation of lþ > ll in PGmc **gulþa*	ON: *goll/gull*
11	assimilation of nþ > nn in PGmc **anþara*	ON: *annarr*
12	assimilation of zd > dd in PGmc **uzdaz*	ON: *oddr*
13	assimilation of zn > nn in PGmc **raznaz*	ON: *rannr*
14	u-mutation (a > ǫ/ø/ö) in PGmc **arnuz*	ON: *ǫrn*
15	u-mutation (a > ǫ/ø/ö) in PGmc **galtuz*	ON: *gǫltr*
16	i-mutation (o > ǫ/ø/ö) in 3.SG present tense of PGmc **kwemanan*	ON: inf - *koma* | 3.SG.PRS *kemr/kømr*
17	unrounding of i-mutated /ø/ in ON *kømr* (ǫ/ø/ö > e)	ON: *kemr*
18	R-mutation (a > e) in PGmc **hasō*	ON: *heri*
19	g/k-mutation (a > e) in PGmc **ganganaz*	ON: *genginn*
20	nasal assimilation nt > tt in PGmc **brantaz*	ON: *brattr*
21	nasal assimilation nt > tt in PGmc **wintruz*	ON: *vetr*
22	nasal assimilation nd > nt > tt in PGmc **band*	ON: *batt*
23	i > e in PGmc **wintruz*	ON: *vetr*
24	nasal assimilation nk > kk in PGmc **drinkanan*	ON: *drekka*
25	nasal assimilation nk > kk in PGmc **drunknanan*	ON: *drukkna*
26	nasal assimilation mb > mp > pp in PGmc **swambō*	ON: *svǫppr*

27	hiatus contraction type a (ea > ja) in PN **sea*	ON: *sjá*
28	hiatus contraction type b (ea > e) in PN **sea*	SE: *se*
29	hiatus contraction type c (ea > íggja) in PN **sea*	FA: *síggja*
30	loss of initial /w/ before /r/ in PGmc **wraiþaz*	ON: *reiðr*
31	assimilation of rs > ss in PGmc **fursaz*	ON: *foss*
32	assimilation of rs > ss in PGmc **furistaz*	OEV: *fuäst*
33	assimilation of rs > ss in PGmc **þurskaz*	ON: *þoskr*
34	loss of /ð/ after /r/ (rzd > rð > r) in PGmc **burzda*	NO_S: *bor*
35	dissimilation of /rl/ into /dl/ in PGmc **karlaz*	IC: *karl* /kadl/
36	dissimilation of /rn/ into /dn/ in PGmc **barna*	NO_S: *badn*
37	/u/ > /o/ in final position in a word or syllable, as in PGmc **buanan*	OS: *boa*
38	breaking of /e/ (e > ja) in PGmc **helpanan*	ON: *hjalpa*
39	breaking of /e/ (e > ja) in PGmc **stelanan*	FA: *stjala*
40	lowering of e > æ/ä in PGmc **fehu*	SE: *fä*
41	w-breaking (i > iu) in PGmc **singwanan*	OS: *siunga*
42	progressive j-mutation (ja > je/jæ/jä) in PGmc **helpanan*	OG: *hielpa*
43	assimilation of bn > mn in PGmc **ebnaz*	NO_N: *jamn*
44	loss of initial h before /l/ in PGmc **hlaupanan*	SE: *löpa*
45	loss of initial h before /n/ in PGmc **hnuts*	SE: *nöt*
46	loss of initial h before /r/ in PGmc **hringaz*	SE: *ring*
47	loss of initial h before /j/ in PGmc **helpanan* > ON: *hjalpa*	SE: *hjälpa* /ˊjæl:pa/
48	loss of initial h before /v/ in PGmc **hwitaz*	SE: *vit*
49	monophthongisation of ai > ei > e in PGmc *stainaz* > ON *steinn*	SE: *sten*
50	monophthongisation of au > ey > ø/ö in PGmc **auzō*	SE: *öra*
51	monophthongisation of au > ø/ö in PGmc **augō*	SE: *öga*
52	stød	DK: *hund* /hunˀ/
53	lenition of /p/ in PGmc **drepanan*	DK: *dræbe*
54	lenition of /t/ in PGmc **setjanan*	DK: *sidde*
55	lenition of /k/ in PGmc **drinkanan*	DK: *drikke* /ˈdʁɛgə/
56	assimilation of tan > tn > nn in PGmc **watar/watan*	DK: *vand* /vanˀ/
57	opening of spirant g > u in PGmc **skogaz*	DK: *skov*
58	a > o/å before /ld/ in PGmc **haldanan*	DK: *holde*
59	a > o/å before /nd/ in PGmc **banda*	DK: *bånd*
60	a > o/å before /rd/ in PGmc **gardaz*	DK: *gård*
61	loss of /j/ between stem final /g/ and unstressed /a/ in PGmc **ligjanan*	DK: *ligge*
62	loss of /j/ between stem final /k/ and unstressed /a/ in PGmc **kirikō* > ON *kirkja*	DK: *kirke*
63	/þ/ > /t/ in PGmc **þinga*	DK: *ting*
64	/g/ > /j/ before front vowel in PGmc **gebanan*	SE: *giva* /ˊji:va/
65	/sk/ > /ʂ/ or /ʃ/ before front vowel in PGmc **skeranan*	SE: *skära* /ˊɧæ:ra/ or /ˊʂæ:ra/
66	/k/ > /ʧ/ or /ɕ/ before front vowel in PGmc **kirikō*	SE: *kyrka* /ˊɕyr:ka/
67	/stj/ > /ʂ/ or /ʃ/ in PGmc **sternō* > ON *stjarna*	SE: *stjärna* /ˊɧæ:ɳa/ or /ˊʂæ:ɳa/
68	/dj/ > /j/ in PGmc **deupaz*	SE: *djup* /jʉ:p/
69	/lj/ > /j/ in PGmc **leuhsaz* > ON *ljóss*	SE: *ljus* /jʉ:s/
70	/gj/ > /j/ in PGmc **geutanan* > ON *gjóta*	SE: *gjuta* /ˊjʉ:ta/
71	/g/ > /ʤ/ before ending in front vowel in PGmc **wargaz*	OEV: *warg - wardʂ-in*
72	retroflex in PGmc **stertaz* (rt > ʈ)	SE: *stjärt* /ɧæʈ:/ or /ʂæʈ:/

73	tones	SE: *anden* /ˈán:dən/ ’the duck’ /ˈàn:dən/ ‘the ghost’
74	level stress	OEV: *stäðer* /stɛːˈðər/
75	loss of /h/ in PGmc **hurna*	OEV: *uonn*
76	loss of /n/ preceding /s/ in PGmc **gans*	ON: *gás*
77	loss of /ð/ or /t/ in neutral definiteness ending	NO_B: *huset* /ˈhʉ̟:sə/
78	loss of /ɾ/ in the 3.SG.PRS ending of PGmc **kwemanan*	ON: *kemr* > NO_N: *kjem*
79	thick l /ɽ/ in PGmc **gardaz*	NO_B: *gard* /gɑ:ɽ/
80	vowel balance	OEV: *fara* (light syllable) > *fårå* *kasta* (heavy syllable) > *kast*
81	vowel harmony	OEV: *yfir* > *yvyr*
82	skerpning in PGmc **saiwiz*	FA: *sjógvur*
83	initial /sl/ > /ʃl/ in PGmc **slahanan*	NO_B: *slå* /ʃlo:/
84	assimilation of ld > ll in PGmc **kwelda*	SE: *kväll*
85	assimilation of ld > ll in PGmc **ailidaz* > ON *eldr*	NO_S: *ell*
86	assimilation of mb > mm in PGmc **lambaz*	SE: *lamm*
87	assimilation of lj > ll in PGmc **huljanan*	GUT: *hyllä*
88	/ald/ > /ɑud/ in PGmc **haldanan*	NL: *houden* /ˈhɑudə/
89	assimilation of final rn > nn in PGmc **hurna*	OEV: *uonn*
90	breaking (e > ja) in PGmc **eka*	SE: *jag*
91	a-mutation (u > o) in PGmc **wulfaz*	NL: *wolf*
92	/uld/ > /ɑud/ in PGmc **gulþa*	NL: *goud* /gɑut/
93	contraction of verb and in ending in PGmc **ganganan*	NL: *gaan*
94	gemination of /k/ in the sequence /kr/ in PGmc **akraz*	NL: *akker*
95	gemination of /p/ in the sequence /pl/ in PGmc **aplaz* or *apalja* > ON *epli*	NL: *appel*
96	loss of /m/ in PGmc **fimfe*	NL: *vijf*
97	loss of second /f/ in PGmc **fimfe*	ON: *fimm*
98	/s/ > /r/ in PGmc **fraus*	NL: *vroor*
99	initial /s/ > /z/ before vowel in PGmc **sunnōn* or **sōel*	NL: *zon*
100	initial /f/ > /v/ before vowel in PGmc **fiskaz*	NL: *vis*
101	dissimilation of mm > mb in the present tense of PGmc **kwemanan*	OG: *cumbr*
102	/i/ > /ɛ/ in the past tense of PGmc **ganganan*	ON: *gekk*
103	/ɛ/ > /i/ in PGmc **brestanan*	SE: *brista*
104	/ɛ/ > /ɑ/ in PGmc **brestanan*	NL: *barsten*

Morphosyntax

#	item	example
1	passive in -sk/-st	kalla-sk/st ‘call-PAS’
2	passive in -s	kalla-s ‘call-PAS’
3	*munna* as future tense auxiliary	
4	*skulu* as future tense auxiliary	
5	*fara* as future tense auxiliary	
6	*vilja* as future tense auxiliary	
7	*koma* as future tense auxiliary	
8	present perfect with both have and be as auxiliary for unaccusative verbs	Ég hef kom-ið ‘I have come-PRF’ Ég er kom-inn ‘I am come-PRF.M.NOM.SG’

9	present perfect with either have and be as auxiliaries	Jeg har tag-et ‘I have take-PRF’ Jeg er komm-et ‘I am come-PRF’
10	different form for the supine and the neuter singular form of the past participle	tag-it ‘take-PRF’ tag-et ‘take-PRF.N.SG’
11	present tense (3rd person singular) of *koma* with i-mutation	kom ‘come’ æ kømer ‘I come’
12	number system: 20-(&)-1	tjugo-ett ‘twenty-one’
13	number system: 1-(&)-20	en-og-tyve ‘one-and-twenty’
14	number system based ten	fem-tio ‘five-ten’ sex-tio ‘six-ten’ sju-ttio ‘seven-ten’
15	number system based twelve	hundrað ‘hundred = 120’
16	number system based twenty	halv-treds ‘half-third = 50’ treds ‘third = 60’
17	earthen X	een aarden pot ‘an earthen pot’
18	earth-X	en ler-kruka ‘an earth-pot’
19	woollen X	en ulden trøje ‘a woollen sweater’
20	wool-X	ein ull-troyggja ‘a wool-sweater’
21	golden X	en gyllene ring ‘a golden ring’
22	gold-X	en guld-ring ‘a gold-ring’
23	*smár* as plural form of *lítill*	et litet hus ‘a little house’ små hus ‘little houses’
24	*vesle* as definite singular form of *lítill*	et litet hus ‘a little house’ det vesle hus-et ‘the little house-the’
25	*sämre* as comparative of bad	
26	development of irregular plural for *auga* ‘eye’	
27	development of irregular plural for *kró/krá* ‘corner’	
28	development of n-plural	Old Swedish: hiärta hiärta-ø ‘heart heart-PL’ Swedish: hjärta hjärta-n ‘heart heart-PL’
29	*fingr* ‘finger’ has become neuter	
30	development of indefinite article	
31	development of suffixed definite article	hús ‘house’ hús-ið ‘house-the’
32	double definiteness	den långa väg-en ‘the long road-the’
33	development of indefinite pronoun from *maðr* ‘man’	
34	development of indefinite pronoun from *einn* ‘one’	
35	koma úr Englandi ‘come PREP England’	
36	koma frá Englandi ‘come PREP England’	
37	inversion of verb and sentential adverb in subclause	Eg veit ikkje ‘I know not’ att eg ikkje veit ‘that I not know’
38	no inversion of verb and sentential adverb in subclause	Ég veit ekki ‘I know not’ að ég veit ekki ‘that I know not’
39	V1	Gingu teir ‘Went they’
40	V1+V2	Jobbar gör jag ‘Work do I’
41	verb particle before the complement	Jag ringer upp honom ‘I call up him’
42	verb particle after the complement	Jeg ringer ham op ‘I call him up’
43	*begin* with infinitive marker	Han begynte å synge ‘He started to sing’
44	*begin* without infinitive marker	Han började sjunga ‘He started sing’
45	N POSS preference	hesten min ’horse-the my’
46	POSS N preference	min häst ’my horse’
47	*der* and *sem* as relative pronoun	mand-en der/som kommer ‘man-the who/who comes’
48	*ekki* not sentence initially	*Ikke ved jeg ‘Not know I’
49	a such	en sådan bil ‘a such car’
50	such a	sådan en bil ‘such a car’
51	Who their	Kvem er det sin bil ‘Who is this their car’
52	indefinite noun after *samr*	samma dag ‘the.same day’
53	definite noun after *samr*	sama slag-e ‘same sort-the’

54	*hræddr* without preposition	Eg er redd bjørn-en ‘I am scared bear-the’
55	development of irregular past tense of *gnaga* ‘gnaw’	
56	development of progressive aspect	Ég er að borða ‘I am to eat’
57	infinite-finite verb order in subclause	hver genginn væri ‘who gone were’
58	NEG IMP order	Ikke gå! ‘Not go!’
59	NEG INF as negated imperative	Ikki fara til hús! ‘Not go to house!’
60	perfect has prefixed *ge-*	ge-lop-en ‘PRF-walk-PRF’
61	OV order in main clauses	
62	VO order in main clauses	
63	nose only in plural	nas-ar ‘nose-PL = nose’

Bibliography

ÅKERBERG, B. 2012. *Älvdalsk grammatik*. Älvdalen: Ulum Dalska.

ALLAN, K. (ed.). 2013. *The Oxford Handbook of the History of Linguistics*. Oxford: Oxford University Press.

ALLAN, R., HOLMES, P. & LUNDSKÆR-NIELSEN, T. 1995. *Danish: a comprehensive grammar*. London: Routledge.

ANDERSSON, H. 2003. Urbanisation. In*: The Cambridge history op Scandinavia (K. Helle ed.)*, 312-342. Cambridge: Cambridge University Press.

ÁRNASON, M., JÓNSDÓTTIR, H. & BÖÐVARSSON, Á. 2002. *Íslensk orðabók*. Reykjavik: Edda.

ATKINSON, Q. D., MEADE, A., VENDITTI, C., GREENHILL, S. J. & PAGEL, M. 2008. Language Evolve in Punctuational Bursts. *Science, 2008; 319*, 588.

BAECHLER, R. & SEILER, G. 2011. Simplification, complexification, and microvariation: Towards a quantification of inflectional complexity in closely related varieties. *Proceedings of the 8th Mediterranean Morphology Meeting.*

BANDLE, O. 1967. *Studien zur westnordischen Sprachgeographie: Haustierterminologie im Norwegischen, Isländischen und Färöischen*. Copenhagen: Munksgaard.

BANDLE, O. 1973. *Die Gliederung des Nordgermanischen*. Basel: Helbing und Lichtenhahn.

BARNES, M. P. 2012. *Runes: A Handbook*. Woodbridge: The Boydell Press.

BAYLDON, G. 1870. *An elementary grammar of the Old Norse or Icelandic Language*. London/Edinburgh: Williams and Norgate.

BERGSLAND, K. & VOGT, H. 1962. On the validity of glottochronology. *Current Anthropology, 1962; 3*, 115-153.

BERNTSEN, M & LARSEN, A. B. 1925. *Stavanger Bymål*. Oslo: H. Aschehoug & Co.

BJERRUM, A. 1944. *Fjoldemålets lydsystem*. Copenhagen: Ejnar Munksgaard.

BJERRUM, M. & BJERRUM, A. 1974. *Ordbog over Fjoldemålet 1-2*. Copenhagen: Akademisk forlag.

BOUCKAERT, R., LEMEY, P., DUNN, M., GREENHILL, S. J., ALEKSEYENKO, A. V., DRUMMOND, A. J., GRAY, R. D., SUCHARD, M. A., & ATKINSON, Q. D. 2012. Mapping the origins and expansion of the Indo-European language family. *Science, 2012; 337(6097),* 957-960.

BRYANT, D. & MOULTON, V. 2002. NeighborNet: an agglomerative method for the construction of planar phylogenetic networks. *Proceedings of the Workshop in Algorithms for Bioinformatics.*

CHANG, W., CATHCART, C., HALL, D. & GARRETT, A. 2015. Ancestry-constrained Phylogenetic Analysis Supports Indo-European Steppe Hypothesis. *Language, 2015; 91*, 194-244.

CHRETIEN, D. 1962. The Mathematical Models of Glottochronology. *Language, 1962; 38*, 11-37.

CLEASBY, R. & VIGFUSSON, G. 1957 [1874]. *An Icelandic-English dictionary*. Oxford: Oxford University Press.

DE VRIES, J. 1961. *Altnordisches Etymologisches Wörterbuch*. Leiden: E. J. Brill.

DE VRIES, J. 1997 [1971]. *Nederlands Etymologisch Woordenboek*. Leiden: Koninklijke Brill.

DUNN, M. 2014. Language Phylogenies. In: *The Routledge Handbook of Historical Linguistic*s (C. Bowern & B. Evans eds.), 190-211. London: Routledge.

EEKMAN, S. J. 2013. *The Diachronic Development of Irregular Noun Morphology in North Germanic Languages - A Logistic Regression Analysis*. [unpublished].

ESPERSEN, J. C. 1994 [1908]. *Bornholmsk ordbog*. Hvidovre: Hardy Larsen.

FAARLUND, J. T., LIE, S. & VANNEBO K. I. 1997. *Norsk referansegrammatikk*. Oslo: Universitetsforlaget.

FAARLUND, J. T. 2004. *The syntax of Old Norse: with a survey of the inflectional morphology and a complete bibliography*. Oxford: Oxford University Press.

FEILBERG, H. F. 1893-1914. *Bidrag til en ordbog over Jyske almuesmål*. Copenhagen: Thiele.

FODOR, I. 1961. The validity of glottochronology on the basis of the Slavonic languages, *Studia Slavica, 1961; 7*, 295-346.

GOULD, S. J. 2007 [1977]. Bushes and Ladders in Human Evolution. In: Gould, S. J. *Ever Since Darwin - Reflections in Natural History*. New York/London: W. W. Norton & Company.

GRAY, R. D. & ATKINSON, Q. D. 2003. Language-tree divergence times support the Anatolian theory of Indo-European. *Proceedings of the National Academy of Sciences of the USA, 2003; 100(15)*, 9079-84.

GRAY, R. D., BRYANT, D. & GREENHILL, S. J. 2010. On the shape and fabric of human history. *Philosophical Transactions of the Royal Society B: Biological Sciences, 2010; 365*, 3923-3933.

GREENBERG, J. H. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In: *Universals of Language* (J. H. Greenberg ed.), 73-113. London: MIT Press.

GROENKE, U. 1998. *Die Sprachenlandschaft Skandinaviens*. Berlin: Weidler Buchverlag.

GUSTAVSON, H. 1940. *Gutamålet: en historisk-deskriptiv översikt*. Uppsala: Appelbergs Boktryckeriaktiebolag.

GUSTAVSON, H. (red.). 1972-1986. *Orbok över Laumålet på Gotland*. Uppsala: AB Lundequistska Bokhandeln.

GUY, J. 1980. *Experimental glottochronology: Basic methods and results*. Canberra: Australian National University, Research School of Pacific Studies.

HARBERT, W. 2007. *The Germanic Languages*. Cambridge: University Press.

HAUGEN, O. E., CHAPMAN, K. G. & GUNDERSEN, D. 1965. *Norwegian-English dictionary*. Oslo: Universitetsforlaget.

HAUGEN, O. E. 2012. *Frå urnordisk till norrønt språk*. http://folk.uib.no/hnooh/handbok/ekstra/Urnordisk/Urnordisk-OEH-2012-12-30-v-1-0.pdf (retrieved June 6th 2015).

HEGGARTY, P. 2006. Interdisciplinary Indiscipline? Can Phylogenetic Methods Meaningfully Be Applied to Language Data - and to Dating Language? In: *Phylogenetic methods and the prehistory of languages* (P. Forster & C. Renfrew eds.). Cambridge: McDonald Institute for Archaeological Research.

HEGGARTY, P., MAGUIRE, W. & MCMAHON, A. 2010. Splits or waves? Trees or webs? How divergence measures and network analysis can unravel language histories. *Philosophical Transactions of the Royal Society B: Biological Sciences, 2010; 365*, 3829-3843.

HEGGSTAD, L., HØDNEBØ F. & SIMENSEN, E. 2008. *Norrøn ordbok*. Oslo: Det Norske Samlaget.

HESSELMAN, B. 1936. Några nynordiska dialektformer och vikingatidens historia. En undersökning i svensk och dansk språkutveckling. In: *Ordgeografi och språkhistoria*. Stockholm-Köpenhamn.

HJORTH, E. & KRISTENSEN, K. (eds.). 2003-2005. *Det danske ordbog*. Copenhagen: Det danske sprog- og litteraturselskab.

HOLDEN, C. J. & GRAY, R. D. 2006. Rapid Radiation, Borrowing and Dialect Continua in the Bantu Languages. *Phylogenetic methods and the prehistory of languages* (P. Forster & C. Renfrew eds.), 19-31. Cambridge: MacDonald Institute Press, University of Cambridge.

HOVDENAK, M. 2006. *Nynorskordboka: definisjons- og rettskrivningsordbok*. Oslo: Det Norske Samlaget.

HUSBY, O. (ed.). 2008. *An introduction to Norwegian dialects*. Trondheim: Tapir Academic Press.

Í SKÁLA, A. & MIKKELSEN, J. (eds.). 2007. *Ensk-føroysk/Føroysk-ensk orðabók*. Tórshavn: Sprotin.

JAHR, E., H. (ed.). 1990. *Den store dialektboka*. Oslo: Novus Forlag AS.

JUNKER, R. & SCHERER, S. 2006. *Evolution - ein kritisches Lehrbuch*. Gießen: Lehrmittelverlag Weyel.

KALKAR, O. 1976 [1881-1918]. *Ordbog til det ældre danske Sprog*. Copenhagen: Akademisk Forlag.

KAUSEN, E. 2012. *Die indogermanischen Sprachen*. Hamburg: Helmut Buske Verlag GmbH.

KOENEMAN, O. & ZEIJLSTRA, H. 2010. Resurrecting the Rich Agreement Hypothesis: Weak isn't strong enough. In: *Movement in minimalism: proceedings of the 12th Seoul International Conference on Generative Grammar* (D. H. An & S. Y. Kim eds.), 289-304. Seoul: Hankuk.

KRESS, B. 1982. *Isländische Grammatik*. Leipzig: Verlag Enzyklopädie.

KROONEN, G. 2013. *Etymological Dictionary of Proto-Germanic*. Leiden/Boston: Brill.

LEVANDER, L. 1985 [1909]. *Älvdalsmålet i Dalarna - ordböjning ock syntax*. Älvdalen: Rege-Tryck.

LEVINSEN, C. 2015. Scandinavian semantics and the human body: an ethnolinguistic study in diversity and change. *Language Sciences, 2015; 49*, 51-66.

NAKHLEH, L., WARNOW, T., RINGE, D. & EVANS, S. N. 2005. A Comparison of Phylogenetic Reconstruction Methods on an Indo-European dataset. *Transactions of the Philological Society, 2005; 103(2),* 171-192.

NORDIC ATLAS OF LANGUAGE STRUCTURES. http://www.tekstlab.uio.no/nals#/ (retrieved June 2015).

NICHOLLS, G. K. & GRAY, R. D. 2006. Quantifying Uncertainty in a Stochastic Model of Vocabulary Evolution. In: *Phylogenetic methods and the prehistory of languages* (P. Forster & C. Renfrew eds.), 161-171. Cambridge: McDonald Institute for Archaeological Research.

NICHOLS, J. & WARNOW, T. 2008. Tutorial on Computational Linguistic Phylogeny. *Language and Linguistics Compass, 2008; 2/5*, 760-820.

NIELSEN, H. F. 2000. *The early runic language of Scandinavia: studies in Germanic dialect geography*. Heidelberg: Winter.

NOREEN, A. G. 1904. *Altschwedische Grammatik, mit Einschluss des Altgutnischen*. Halle: Max Niemeyer.

PAGEL, M., ATKINSON, Q. D. & MEADE, A. 2007. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature, 2007; 449*, 717-720.

PALM, R. 2004. *Vikingarnas språk: 750-1100*. Stockholm: Norstedts Förlag.

PERELTSVAIG, A. & LEWIS, M. W. 2015. *Indo-European Controversy: Facts and Fallacies in Historical Linguistics*. Cambridge: Cambridge University Press.

PETTERSON, G. 2005 [1996]. *Svenska språket under sjuhundra år: en historia om svenskan och dess utforskande*. Lund: Studentlitteratur.

PHILLIPA, M. (ed.). 2003-2009. *Etymologisch Woordenboek van het Nederlands*. Amsterdam: Amsterdam University Press.

PIPPING, H. 1907. *Guta Lag och Guta Saga jämte ordbok*. Copenhagen: S.L. Møllers bogtrykkeri.

POULSEN, J. H. W. 1998. *Føroysk orðabók*. Tórshavn: Føroya Fróðskaparfelag.

RAMAT, P. 1998 [1993]. The Germanic Languages. In: *The Indo-European Languages* (A. G. Ramat & P. Ramat eds.). London/New York: Routledge.

RINGGAARD, K. 1970 [1969]. *Danske dialekter: en korftfattet oversigt*. Copenhagen: Akademisk Forlag.

RINGMAR, M. 2005. Älvdalska. En önordisk språkö på fastlandet? In: *Proceedings of the 1st Conference on Övdalian, Älvdalen, June 18th-19th, 2004.*

SÖDERWALL, K. F. 1884. *Orbok öfver svenska medeltids-språket*. Lund: Berlingska boktryckeri- och stilgjuteri-aktiebolag.

SPEYER, A. 2007. *Germanische Sprachen: ein historischer Vergleich*. Göttingen: Vandenhoeck & Ruprecht.

STEENSLAND, L. 2010. *Material till en älvdalsk ordbok*. Falun: ScandBook AB.

SVENSKA AKADEMIEN. 2009. *Ordbok utgiven av Svenska Akademien*. Stockholm: Svenska Akademien.

SWADESH, M. 1952. Lexicostatistic Dating of Prehistoric Ethnic Contacts. *Proceedings of the American Philosophical Society, 1952; 96,* 452-463.

SWADESH, M. 1971. *The Origin and Diversification of Language* (J. Scherzer ed.). Chicago: Aldine.

SYKES, B. 2001. *The seven daughters of Eve*. London: Corgi.

SYRJÄNEN, K., HONKOLA, T., KORHONEN, K. LEHTINEN, J., VESAKOSKI, O, & WAHLBERG, N. 2013. Shedding more light on language classification using basic vocabularies and phylogenetic methods - A case study of Uralic. *Diachronica, 2013; 30(3),* 323-352.

TADMOR, U. 2009. Loanwords in the world's languages: Findings and results. In: *Loanwords in the world's languages: A comparative handbook* (M. Haspelmath & U. Tadmor (eds.), 55-75. Berlin: Walter de Gruyter.

TELEMAN, U., HELLBERG, S. & ANDERSSON, E. 1999. *Svenska Akademiens grammatik*. Stockholm: Svenska Akademien/Norstedts Ordbok.

THRÁINSSON, H., PETERSEN, H. P., Í LÓN JACOBSEN, J. & SVABO HANSEN, Z. 2004. *Faroese: an overview and reference grammar*. Tórshavn: Føroya Fróðskaparfelag.

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

VAN STERKENBURG, P. (ed.). 2002 [1984]. *Van Dale Groot woordenboek hedendaags Nederlands*. Utrecht/Antwerpen: Van Dale Lexicografie.

WANGENSTEEN, B. (ed.). 2005. *Bokmålsordboka: definisjons- og rettskrivningsordbok*. Oslo: Kunnskapsforlaget.

WIDMARK, G. 1994. Birkasvenskan - fanns den? *Arkiv för nordisk filologi, 1994; 109,* 173-216.

WIKANDER, O. 2006. *I döda språks sällskap: en bok om väldigt gamla språk*. Stockholm: Wahlström & Widstrand.

ZETTERHOLM, D. O. 1953. *Dialektgeografiska undersökningar III. Orne. Råne. Galt. Fargalt. So. Sugga. Purka - IV. Lockrop till får. Lockrop till höns (och svin)*. Uppsala: Almqvist & Wiksells Boktryckeri AB.

ZOLA CHRISTENSEN, R. 2007. *Dansk for svensktalende*. Lund: Studentlitteratur.