



LUND
UNIVERSITY

Centre for Language and Literature

Master's Thesis

General Linguistics

June 2013

Supervisor: Gerd Carling

Co-supervisors: Arthur Holmer, Love Eriksen, Joost van de Weijer

Two Millennia of Lexical and Typological Change in Western Europe

- a quantitative geographical approach

Filip Larsson

Abstract

This thesis aims to examine the lexical and typological change found in the Western European language families of Germanic, Romance and Celtic over the last two millennia. The method used was to create one lexical and one typological database and the data was analysed according to etic grids. Tree models were generated from the results of the databases and the groups found in the tree models were mapped out on a map over Western Europe. The lexical results were similar to traditional classifications. The lexical results also showed that the changes appeared according to a pattern that could be described by the wave theory where lexical changes spread from the centre to the periphery. The results of the typological data were different from traditional classifications as it did not follow the boundaries of the three language families. In general the wave theory was applicable to a lesser extent to the typological data but it was relevant for the verbal morphology. Contrary to the lexical results the typological results indicated the existence of conservative centres with the peripheral languages being more typologically innovative. The conclusions drawn were that lexical change and typological change are two diametrically different and independent processes.

Keywords: Lexicology, typology, historical linguistics, Western Europe, Indo-European, Celtic, Germanic, Romance

Acknowledgements

I would like to thank my supervisor Gerd Carling for her highly appreciated support and my co-supervisor Arthur Holmer for his input and ideas. I would also like to especially thank and acknowledge the important role that my additional co-supervisors Joost van de Weijer and Love Eriksen have played in making this thesis possible. I am gratefully thanking you all for your support and help.

Table of Contents

List of Tables, Figures and Maps	3
1. Introduction	4
2. Background	5
2.1. The Germanic languages	5
2.2. The Italic languages	6
2.3. The Celtic languages	8
2.4. Standard Average European	9
2.5. Dead languages	9
3. Method	11
3.1. Cladistics	13
3.2. The etic grids	14
3.3. Lexical database	14
3.4. Typological database	17
3.5. Tree models	21
3.6. Mapping	22
4. Results	24
4.1. Lexical results	24
4.2. Typological results	30
5. Discussion	47
5.1. Lexical discussion	47
5.2. Typological discussion	50
5.3. Combined discussion	57
6. Conclusions	60
Bibliography	63
Electronic Dictionaries and Corpora	68
Appendix A: The Swadesh 100-list	70
Appendix B: The Etic Grid for the typological data	71
Appendix C: Sources for all languages	72
Appendix D: Lexical dendrogram (250 BC to 1000 AD)	74
Appendix E: Lexical dendrogram (1000 AD to 1500 AD)	75
Appendix F: Lexical dendrogram (1500 AD to 2000 AD)	76
Appendix G: The entire typological database	77

List of Tables, Figures and Maps

Table 3.1: All languages of the databases according to period	13
Figure 4.1: Dendrogram of the entire lexical database	24
Map 4.1: Lexical results mapped out according to cluster (dendrogram 3.1)	27
Map 4.2: Lexical results mapped out according to number of nodes (dendrogram 4.1)	28
Figure 4.2: Dendrogram for the entire typological database	30
Map 4.3: Typological results mapped out according to cluster in dendrogram 4.2	32
Map 4.4: Typological results mapped out according to number of nodes (dendrogram 4.2)	33
Figure 4.3: An unrooted tree model generated for the entire typological database	35
Figure 4.4: Dendrogram of the typological data for the ancient languages	37
Figure 4.5: Typological dendrogram for the medieval languages.	38
Figure 4.6: Typological dendrogram for the modern languages.	40
Map 4.5: Typological results for the modern languages mapped out (number of nodes)	41
Figure 4.7: Dendrogram for the typological database excluding the verbal data	42
Figure 4.8: Dendrogram of the verbal grids of the typological database	44
Map 4.6: The verbal results according to cluster in dendrogram 4.8.	46

1. Introduction

The Indo-European languages of Western Europe have experienced significant change since the first written records in antiquity. Two millennia have passed and with them new words and features have emerged while old forms have been lost. What was lost and what was found and how did all these languages evolve into the living languages we know today? To be able to answer these questions it is first of all important to define what languages we speak of and as this thesis deals with Western Europe it will study the Germanic, the Romance and the Celtic languages. These language families will therefore be presented briefly in chapter 2.

The aim of this thesis has been to try to answer some of the most important questions with regard to language change:

- How does lexical change occur and how do these changes spread?
- How does typological change occur and how do these changes spread?
- Do lexicology and typology change according to the same patterns?
- What could have caused the lexical and typological changes in Western Europe?
- Is it possible to find the geographical sources of these changes?
- What role does geography play in the spreading of changes?

To answer these questions a quantitative method has been used where data was collected to form one lexical database and one typological database. To quantify the data it was structured according to the etic grid method described in chapter 3.2. When this had been done the two databases were cladistically analysed with GNU/R to create computer-generated dendrograms, see chapter 3.5. These dendrograms were later mapped out using ArcGIS according to two methods of clustering, i.e. cluster analysis and node counting which is further described in chapter 3.6. The results will be presented in chapter 4 commencing with the lexical results followed by the typological results. In chapter 5 these results will be discussed leading up to the conclusions in chapter 6.

2. Background

2.1. The Germanic languages

The Germanic language family is among the largest language families in the world, mainly due to the colonisation of the New World which spread the Germanic languages, or at least English, all over the world. Prior to this the Germanic languages were only found in northern, north western and central Europe and as this thesis focuses on Europe these areas are the only areas of relevance for this thesis. The Germanic languages are divided into three branches, namely East, North and West Germanic. The first of these branches to separate itself from the others appears to be the East Germanic languages (Harbert 2007). The eastern branch is today extinct and is represented in this thesis only by Gothic, its most prominent member, mostly in the form of Biblical texts from the 6th century and as such it presents one of the earliest sources of texts in Germanic. Gothic became extinct at some point before the 10th century but its descendant Crimean Gothic survived into the 17th century (Harbert 2007), but as sources are scarce Crimean Gothic could not be included in this thesis.

The North and West Germanic branches were next to split off from each other and in the middle of the 6th century the North Germanic languages started to show characteristics that set them apart from the West Germanic languages (Haugen 1982 via Harbert 2007). The North Germanic branch has been more homogenous in comparison to the western branch which is still true as the Continental North Germanic languages, i.e. Danish, Norwegian and Swedish, show a high level of mutual intelligibility. The remaining languages of the northern branch, i.e. Icelandic, Faroese and also Elfdalian, have not adopted some of the innovations found in Danish, Swedish and Norwegian and are therefore, due to their conservatism, hardly mutually intelligible with the other languages. Historically the North Germanic languages have been divided into Eastern North Germanic, comprising Danish, Swedish and Elfdalian, and Western North Germanic, comprising Norwegian, Icelandic and Faroese, but as Norwegian has been so influenced by Danish and Swedish it has become more relevant to speak of the Insular North Germanic languages, i.e. Icelandic and Faroese, and Continental North Germanic languages, i.e. the rest (Harbert 2007). Old Norse, i.e. Old Icelandic and Old Norwegian (Harbert 2007), is an important source for earlier texts in North Germanic but it is not the earliest form of written North Germanic as runic inscriptions predate it, but these inscriptions can unfortunately not provide enough information to be relevant for this thesis.

The last branch is the West Germanic branch which is the most important as it includes the three largest Germanic languages, namely English, German and Dutch. The West Germanic branch can be divided into three historical sub-branches, namely Ingvaeonic, Irminonic and Istvaeonic (Harbert 2007). Ingvaeonic, also known as North Sea Germanic, is a highly innovative sub-branch (Harbert 2007) and it comprises Anglo-Frisian, i.e. English and all Frisian languages, and Low German. It should be said that the languages of this sub-branch are rather divergent as Low German, or its ancestor Old Saxon, split off quite early from Anglo-Frisian and has been heavily influenced by High German for centuries while English has due to both its isolated location and French, North Germanic and Celtic influences developed in a different direction. The Irminonic branch includes the languages usually referred to as Upper German, i.e. e.g. Bavarian and Alemannic, represented in this thesis by Swiss German, and the Istvaeonic branch comprises Dutch and also the non-European Afrikaans not included in this thesis. High German, commonly known as just German, is found somewhere between the Irminonic and Istvaeonic branches as it shows features of both groups and the boundary between the historical branches have developed into a continuum from historically Istvaeonic varieties such as Franconian in the north to the historically Irminonic varieties such as Alemannic and Bavarian in the south, where High German has emerged in between. This continuum becomes clear while examining how many of the changes of the High German consonant shift a variety has undergone (Harbert 2007).

2.2. The Italic languages

The Italic languages, or really the Romance sub-branch, are amongst the absolutely most spoken languages in the world, originally found in south-western Europe but as an effect of extensive colonialism they are now dominating Central and South America and they are also widely used in many parts of Africa. The Italic languages can be divided into two major groups, namely the Latin-Faliscan group, comprising all living Romance languages, and the Sabellic/Sabellian or Osco-Umbrian group, comprising, amongst others, the two extinct languages Oscan and Umbrian. In early antiquity these two groups were more or less equal in importance, with the Latin-Faliscan group being the smaller of the two only occupying modern Latium and the Sabellic group found in what is now central and southern Italy (Buck 1904). This changed however in the 2nd century BC as the Latin-speaking Romans conquered more and more of the Sabellic-speaking areas and thus incorporated the Oscans, the Umbrians and all other Italic-speaking peoples into the realm of Rome. This Roman dominance led to

the gradual extinction of the Sabellic languages and by the end of antiquity Latin was the sole survivor of the Italic languages.

The descendants of Latin, or more precisely Vulgar Latin, became the Romance languages, one of the most influential language families in the world from the New Age and onwards. The first language group to branch off is not that influential though as they are the Eastern Romance languages. The Eastern Romance or Balkan-Romance languages are all confined to the Balkans and they are only represented by Romanian in this thesis. There are other minor Eastern Romance languages such as Aromanian or Vlach and Megleno-Romanian but due to scarcity of data they were excluded. The Eastern Romance languages differ quite significantly from the other Romance languages as a result of the geographical distance, the Balkan Sprachbund and Slavic influences. Depending on how the classification is done the second branch to branch off is the Southern Romance branch which is found only on Sardinia and it is composed of the three varieties of what is collectively known as Sardinian, namely Campidanese, Logudorese and Nuorese. Sardinian is often described as the most conservative of all the Romance languages as it has preserved many archaic features of Latin that have been lost elsewhere (Harris 1997).

The remaining languages can be divided into three major sub-branches and they are the Italo-Dalmatian or Italo-Romance languages, the Gallo-Romance languages and the Ibero-Romance languages (Kabatek & Pusch 2011). The Italo-Dalmatian languages are mainly spoken on the Italian Peninsula from Tuscany in the north to Calabria in the south and on Corsica and Sicily. The Italo-Dalmatian language Dalmatian used to be spoken in northern Croatia but it is today extinct. Italian and Old Italian, more precisely Old Tuscan, are the only languages representing the Italo-Dalmatian languages in this thesis. The second sub-branch is the Gallo-Romance sub-branch stretching from the Gallo-Italian languages such as Lombard and Venetian in northern Italy to the Oil languages such as French and Walloon in the north western part of continental Europe. These languages are mainly found in areas that were dominated by Celtic speaking tribes in antiquity, hence the name. The extent of this sub-branch is not universally agreed upon though as for example Occitan (represented by Provençal in this thesis) is often classified as Gallo-Romance but it is also closely related to Catalan which is classified as Ibero-Romance in most cases. The classification of Occitan and Catalan should most likely be seen as some sort of continuum from the Gallo-Romance languages to the Ibero-Romance languages (Kabatek & Pusch 2011). Similarly the Rhaeto-

Romance languages, i.e. Romansh, Ladin and Friulian, are also sometimes included among the Gallo-Romance languages although they are classified as a separate sub-branch by others (e.g. Ruhlen 1991). Finally the third sub-branch comprises the Ibero-Romance languages all originally found on the Iberian Peninsula, namely Spanish, Portuguese, Catalan, Aragonese, Galician etc.

2.3. The Celtic languages

All three branches studied in this thesis have experienced great changes during the two last millennia, but in vastly different directions. The Germanic and Romance languages experienced significant expansion in this period when the Celtic languages suffered from the opposite, i.e. a dramatic decrease of both speakers and Celtic speaking areas. In antiquity the Celtic languages were found in Central Europe, Gaul (mainly modern Belgium and France), the Iberian Peninsula, the British Isles and even as far as Turkey (in what was known as Galatia). The Celtic languages are often divided into two main branches, the Insular Celtic languages spoken on the British Isles (including Breton) and the Continental Celtic languages comprising all other Celtic languages. The Continental Celtic languages became however victims of the Roman expansion and are today all extinct. Sources of these languages are in most cases fragmentary but a relatively large amount of material has been found on one of the more important Continental Celtic languages, i.e. Gaulish, and it could therefore be included in this study. There were other known Continental Celtic languages such as Celtiberian, Galatian, Lepontic and Noric but as the amount of material that could be found for these languages was too small they had to be excluded.

The Insular Celtic languages are today the only surviving Celtic languages and they have become marginalised by the Germanic and Romance languages. The Insular Celtic languages are usually divided into Brythonic languages, i.e. Welsh, Breton and Cornish, and Goidelic languages, i.e. Irish, Scottish Gaelic and Manx. Both of these branches had important migrations in late antiquity as Goidelic speakers from the north of Ireland migrated to Scotland laying the foundation of the future Scottish Gaelic language in 500 AD and Brythonic speaking migrants left the British Isles from the 5th to the 7th century to later become Bretons (Ronan 2011). In the Medieval Ages the Brythonic languages were spoken in most parts of Wales, Cornwall and Brittany and the Goidelic languages were spoken in all parts of Ireland, large parts of Scotland and on the Isle of Man. This changed gradually though as more and more parts of the British Isles were conquered by the English starting

with the conquest of Cornwall just before the Norman Conquest, the conquest of Wales in the 13th century, the conquest of Ireland around 1600 and last but not least the signing of the Treaty of Union incorporating Scotland into the new Kingdom of Great Britain in 1707.¹ This incorporation of the Celtic countries into the British realm resulted in a decline of the Insular Celtic languages leaving only pockets of Celtic speaking areas and leading to the extinction of both Cornish (in the 18th century) and Manx (in the 20th century). Breton has experienced a similar development since the incorporation of Brittany into France in the early 16th century.

2.4. Standard Average European

The term Standard Average European or SAE was first used by Benjamin Lee Whorf in 1939 but was not really a concept until the 1990's and it refers to a number of typological features shared by languages in Europe, especially amongst the languages of Western Europe (van der Auwera 2011). It was proposed as a Western European *Sprachbund* thus compiling a number of possible areal features. Some of these features include definite and indefinite articles, relative clauses with relative pronoun, perfects formed with the auxiliary verb to *have*, non-pro-drop etc. Almost all European languages have one of the features defined as SAE which is not that relevant but it becomes interesting when looking at the languages with high numbers of SAE traits. The core of this possible SAE is found in an area covering France, Switzerland, the Lowlands and Western Germany and possible also Italy with the features becoming progressively fewer the farther the languages are found from this core. It should be said however that there is research still to be done concerning SAE as it has not been explored to its *full potential* (van der Auwera 2011: 304).

2.5. Dead languages

This thesis includes both living and dead languages and it is therefore important to note the vast difference between working with living languages and working with dead languages. Dead languages can be divided into two groups, i.e. extinct languages with no living descendants and earlier stages of modern languages, both having in common that they lack living native speakers and are thus dead. This means that all information about these languages has to be retrieved from written sources such as literature and inscriptions, which limits our knowledge to what ever was deemed relevant to put into writing at that time and more importantly what has been preserved. As the quality and quantity of the written sources

¹ It should be noted though that Scots or Lowland Scots, a language or dialect closely related to English, has been spoken in southern Scotland since the Middle Ages.

vary considerably it is relevant to speak of three groups of dead languages, i.e. *Großcorpussprachen*, *Kleincorpussprachen* and *Restsprachen* (Mayrhofer 1980: 17). The *Großcorpussprachen*, or languages with large corpora, are the least problematic of the three groups as written records of these languages have been preserved with both a high quantity and quality thus giving us information about most aspects of the languages. Among the dead languages found in this thesis there are several *Großcorpussprachen* such as Latin, Old Norse, Old High German etc. The *Kleincorpussprachen*, or languages with small corpora, are slightly more problematic as the written sources of these languages are not as rich as the first group and it is therefore not possible to collect the same amount of information from them. They do offer an insight into the grammar and structure of the language though and they also contribute with basic lexical data. Languages belonging to this second group are e.g. Gothic, Old Welsh and Old Dutch. The last group contains the *Restsprachen* or the residual languages, i.e. languages that are only known from a small number of inscriptions, glosses, personal names and place names. The *Restsprachen* are in many cases not possible to describe due to the lack of significant data but when it is possible it can be highly interesting. Some *Restsprachen* have therefore been included in this thesis such as Oscan, Umbrian and Gaulish. Other *Restsprachen* were considered but they could not be included due to the scarce material as in the case of Ogham Irish and Runic Nordic.

3. Method

This thesis aims to examine the lexical and typological development of the Indo-European languages of Western Europe with the help of new methodologies made available by modern technology. The method used in this thesis has been borrowed from an article by Carling et al. (to appear 2013) that investigated development and change among the Arawak languages in South America. The method used in the above mentioned article consisted of gathering lexical, typological and archaeological data that was compiled into a database and later analysed using cladistic software. The results of the cladistic analysis were then mapped out to show the geographical distribution of the different features. The results of this method were highly interesting and the method used opted to be applied on the Indo-European languages. Due to the limitations of a master's thesis an analysis of all Indo-European languages was deemed impossible so the data was confined to the branches found in Western Europe, i.e. Celtic, Germanic and Italic. Not all aspects of the above mentioned article's method was borrowed though as e.g. no archaeological data was included.

One of the most influential cladistic studies made with a methodology similar to the one used in this thesis was Ringe et al. (2001). Their study had a larger amount of lexical items, i.e. 333 vs. 100 in this thesis, while they had only 24 languages covering all Indo-European sub-families and only 9 of these languages were Celtic, Italic or Germanic. They also included phonological and morphological aspects but no geographical data. A comparable study was conducted by Dunn et al. (2011) and it included all Indo-European sub-families but only lexical and word order data while also lacking geographical aspects. The results were presented in the form of dendrograms and the conclusion of the study was that *phylogenetic methods* and *large linguistic databases* are of high relevance when studying linguistic variation and language change (Dunn et al. 2011). Dunn et al. (2011) used a statistical method called the *Bayesian hypothesis test* which was criticised however along with the rest of the article by Levy and Daumé (2011). Greenhill et al. (2010) conducted another lexical and typological study of amongst others the Indo-European languages using similar methods based on lexical data from Dyen et al. (1992) covering 95 Indo-European languages. Only about 25 languages or less of these 95 languages were Western European and no dead languages were included. Greenhill et al. (2010) presented their results as so-called NeighbourNets which is paralleled in this thesis by unrooted tree models and as in the case of both Ringe et al. (2001) and Dunn et al. (2011) no geographical aspects were introduced. This

changed however when these articles were followed up by Bouckaert et al. (2012)² in which they introduce geographical aspects. Their article wished to infer the location of the root of the language families which is not at all the purpose of this thesis which sets it apart quite significantly from the geographical aspects found in this thesis however.

The first step that was undertaken was to create two separate databases, i.e. one lexical and one typological database. This was done because of the relevance of making a distinction between lexicon and typology as it is highly possible for a language to maintain most of its basic vocabulary while demonstrating significant typological changes, which is true for e.g. the transition from Old English to modern English. To make the creation of the two databases possible, especially for the typological data, etic grids were constructed to systematise the data found (see chapter 3.2). Then the languages included in the lexical and typological databases were split up according to which time period they were spoken in. The time periods first consisted of four periods of 500 years ranging from 1 AD to 2000 AD, but as the number of languages found in the period between 1 AD and 500 AD were few it was decided to differentiate between three periods, i.e. 1 AD to 1000 AD, 1000 AD to 1500 AD and 1500 AD to 2000 AD. The first period was further expanded to extend from 250 BC to 1000 AD to include Oscan and Umbrian as they were most likely extinct by 1 AD. Unfortunately some languages did not fit well into these periods but as the periods have been created to differentiate between different stages of language development the exact dating of these languages is therefore not that important. A table showing the sources for the languages in the databases is found in appendix C. To visualise both the languages in the database and their respective time period the table below has been included, thus visualising that in the first period Old English is found followed by Middle English in the second period and the last period is represented by modern English. This visualisation also helps to make it clearer which languages that could not be included as their slots are coloured gray, which is true for e.g. Old Danish and Old Catalan.

² NB. that both Dunn and Greenhill are found among the authors.

Table 3.1: All languages of the databases according to period

250 BC – 1000 AD	1000 AD - 1500 AD	1500 AD -2000 AD
Old English	Middle English	English
Old High German	Middle High German	German
		Swiss German
Old Dutch	Middle Dutch	Dutch
Old Saxon	Middle Low German	Low German
Old Frisian		Frisian
Old Norse		Norwegian (Bokmål)
		Norwegian (Nynorsk)
		Icelandic
		Faroese
	Old Swedish	Swedish
		Elfdalian
		Danish
Gothic		
Latin	Old Italian	Italian
	Old Spanish	Spanish
		Portuguese
		Romanian
		Catalan
		Romansh
		Friulian
		Ladin
Old French (and Latin)	Middle French	French
		Walloon
Old Provençal (and Latin)		Provençal
Oscan		
Umbrian		
Old Irish	Middle Irish	Irish
		Scottish Gaelic
		Manx
Old Welsh	Middle Welsh	Welsh
Old Breton	Middle Breton	Breton
Old Cornish	Middle Cornish	Cornish

3.1. Cladistics

Cladistics in historical linguistics are based on three assumptions, namely that that it is possible to identify groups of languages that descend from a common ancestor, that languages change and that languages split into daughter languages if the conditions demand it. These principles are also fundamental for evolutionary biology thus making cladistics a method of quantitative analysis utilised by both linguists and biologists (Johnson 2008: 184). To make cladistic analysis possible all data has to be categorised according to separate variables which in this thesis has been done through the *etic grid* methodology described in chapter 3.2. This needs to be done as cladisitic data is basically varying instances of variables (Ringe et al. 2001) where the instances shared by the largest amount of languages are found at the core and the more the languages have developed from this core the farther away from it they will be found. The cladistic method for *combining character-based trees* was used in this thesis as it

is both *pretty intuitively simple* and it does *yield sensible trees* (Johnson 2008) and was also used by e.g. Ringe et al. (2001). The *characters* refer in this thesis to what is called grids in chapter 3.2, i.e. the comparable separate feature in a language such as word order or the presence of a certain cognate.

3.2. The etic grids

The *etic grid* methodology was used in this thesis to make otherwise incomparable data comparable. For this thesis it meant that a number of lexical concepts and typological features were considered to be relevant for the analysis. The specific lexical concepts, e.g. *I* and *dog*, and typological features, e.g. word order in main clauses, constituted the *etic grids*. To make the collected data comparable all actual values of the features and concepts were given separate rows in the databases, and these separate rows are the *variants* of the etic grid. The variants of e.g. the typological feature of word order in main clauses were SVO, V2, VSO and SOV. This was done since much of the data was on a so called *emic* level, which is highly qualitative but difficult to quantify. The emic level ideally captures all nuances of a language but hinders commensurability as highly specific data makes more general comparisons less meaningful. Therefore there was a need to create an etic grid and thus transforming the emic data into etic data. Etic level data is data that has been adjusted to a limited number of possible outcomes which makes it easier to compare through categorical coding and data abstraction (Carling et al. to appear 2013). To do this it is inevitable to decrease the preciseness of the description of the data, but as long as the etic grid is conveying enough relevant differences and information this decrease is justified to make a comparative analysis meaningful.

3.3. Lexical database

The lexical database consisted of 100 lexical items translated into 57 living and dead Celtic, Germanic and Italic languages. These lexical items were based on the 100-item Swadesh list compiled by linguist Morris Swadesh, which can be found in appendix A. These 100 lexical items represent basic vocabulary as this is a category of vocabulary that is preferred in lexical comparison as it is more resistant to borrowing and thus making the chance of finding cognates among them from a common ancestor higher (Campbell & Poser 2008). The 200-item Swadesh list was used by Greenhill et al. (2010) for less than half of these languages but as this did not include any dead languages the 100-item Swadesh list was considered to be

more appropriate. The data was mainly collected from the Max Planck Institute's Indo-European Lexical Cognacy database (IELex), Blažek (2009), Kroonen (2013), Matasović (2013), de Vaan (2013) and Boutkan & Siebinga (2013) but as Swadesh 100-lists were not found for a number of relevant languages data for some languages had to be collected elsewhere. Lexical data for the following languages was therefore collected from dictionaries: Low German, Old French, Middle French, Old Provençal, Old Swedish and Nynorsk. The majority of the medieval languages were collected from Kroonen (2013), Matasović (2013) and Boutkan & Siebinga (2013) but as all 100 items were not found additional data had to be collected for the following languages: Middle High German, Middle Dutch, Old Frisian, Middle Low German and Middle Welsh. Data was also collected from the Intercontinental Dictionary Series (IDS) compiled by the Max Planck Institute in Leipzig for Middle English, Welsh and Irish. Lexical data for Old Italian and Old Spanish was mainly collected from Dante Alighieri's *Commedia* and *Cantar de mio Cid* respectively due to the difficulty of finding Old Italian and Old Spanish dictionaries. These two literary works were chosen because they were written in the Medieval Ages and therefore ought to contain medieval forms and they have both been highly influential. They were also easily accessible with translation online through Princeton University³ and University of Texas.⁴ A thorough presentation of the lexical sources used is presented in appendix C.

Each lexical concept of the Swadesh list was given a grid resulting in 100 lexical grids. These grids were then split up to create individual variants for each cognate group found, a method known as *cognate identification*. The method used to separate cognates was similar to the one used by Ringe et al. (2001) as the lexical data was grouped according to the Proto-Indo-European root the individual words were derived from, if such a root was found. If no Proto-Indo-European root was found Proto-Germanic or Proto-Celtic roots were found in most cases. Some of the lexical entries lacked roots altogether and they were given individual variants even though they probably would have been omitted by Ringe et al. (2001: 84), especially for the variants found only for one language. These variants were kept however as it could be seen as indications of those languages being innovative or conservative. The number of variants ranged from only one variant for e.g. *finger nail* up to 19 variants for the lexical item *walk*. The identification of cognates can seem unproblematic for the languages of this thesis as they are all Indo-European and historically well documented, but that was not

³ <http://etcweb.princeton.edu/dante/index.html>

⁴ <http://www.laits.utexas.edu/cid/>

always the case. There is a constant risk of lexical similarity due to chance as e.g. French *feu* and German *Feuer* both mean fire but they have developed from two different Proto-Indo-European stems (Campbell 2004). Most of the problems related to cognate identification have however been evaded as I have not made any cognate analyses myself but instead only based the cognate identification on etymological data from IELex, Blažek (2009), Kroonen (2013), Matasović (2013), de Vaan (2013) and Boutkan & Siebinga (2013) and etymological dictionaries. The cognate groupings were first entered according to the IELex data but they were later tested against the cognate data found in Blažek (2009), Kroonen (2013), Matasović (2013), de Vaan (2013) and Boutkan & Siebinga (2013) when possible. All conflicting data found were changed in favour of the groupings of Blažek (2009), Kroonen (2013), Matasović (2013), de Vaan (2013) and Boutkan & Siebinga (2013). Explicit or known loanwords were removed as they otherwise would contaminate the results, but the risk of failing to identify loanwords is always present (Ringe et al. 2001) The cognates were primarily coded according to the cognate codes used by the IELex database, but as some cognates were not present in the IELex database new codes had to be added. In the case of the Celtic languages most new cognate codes were based on Blažek (2009).

The aim of this thesis has been to include as many languages as possible to make the results more relevant than with a smaller quantity of languages. Lexical data was found for a number of languages that later were excluded from the databases as these languages lacked typological data or it was not possible to attain typological data within the time frame of this thesis. This was true for Luxemburgish, Flemish, Campidanese, Nuorese and Vlach (presumably the same as Aromanian). Most of these languages are fortunately not crucial to this thesis as they can be represented by closely related languages such as Dutch for Flemish and Logudorese for the other Sardinian variants. Vlach would have been interesting though but as it is deeply embedded in a Greco-Slavic environment the results for Vlach would probably not be that relevant for this thesis. A small number of languages also had to be excluded from the lexical database since they did not have sufficient data, such as Middle Irish and Old Dutch. To define what a sufficient amount of data is and what is not appeared to be problematic as general guidelines were difficult to set. At first a putative threshold was set to 50% of the lexical entries but was later raised to 70%, though this was more of a guideline than a threshold as some languages such as Oscan and Umbrian were never excluded even though they were far below 70% as they only had data for approximately 30% of the lexical entries, which naturally proved to be a mistake. There were yet other languages that were

excluded on the basis of them being found only outside of Europe, such as Afrikaans and Pennsylvania German. This was mainly done due to the problems they would have created for the mapping. Furthermore there were languages that were considered to be relevant but as no data could be collected they had to be left out of the thesis altogether, i.e. Galician, Corsican, Northern Frisian, Aragonese, Lombard, Venetian, Asturian and a number of dead languages. An example of a dead language that had to be excluded due to time restraints was Old Danish.

It is important to stress that, even though a significant amount of data has been collected from the IELex database, it has some problems when it comes to reliability. This is mainly due to some of the sources they have used as they are not always reliable as they have e.g. cited Wiktionary as a source in some cases and that the data has not always been thoroughly examined. This does not in any way mean that all their data is unreliable but the risk of finding faulty data is higher than what is normally expected for scientific databases. This has led to an awareness of the fact that data from IELex might not be accurate and I have therefore checked all dubious entries and corrected them if necessary. Problems relating to faulty data have therefore not been a major concern for the larger modern languages but it definitely became more problematic in regard to the smaller languages, such as all Celtic languages, Frisian, the Rhaeto-Romance languages and the dead languages. The Celtic data turned out to be less problematic though as it was, due to this uncertainty with IELex, almost entirely collected from Blažek (2009), Matasović (2013) and IDS. The lexical data for the minor Germanic languages were in many cases checked against dictionaries, as was done for Frisian, Icelandic, Faroese, Swiss German and Norwegian (Bokmål). This was unfortunately not done for the minor Romance languages, except to some degree for Logudorese (Sardinian), due to the limited amount of time. A small number of dead languages were checked against dictionaries as well, namely Old Norse, Gothic and Old High German. See appendix C for a complete table of the sources used for each language.

3.4. Typological database

The typological database consisted of a number of syntactical and morphological grids where all different forms found were given separate variants and the data was coded binarily. These grids were chosen by two criteria where the first were general grids relevant for all languages and the specific grids relevant for the separate branches. The general grids covered e.g. word order, personal inflection of verbs, case, gender and so on as these are grids that are relevant

for all languages thus being the foundation of the analysis. Some general aspects were excluded though as e.g. number but these aspects were not considered to be relevant as in the case of number the dual was so rare that it was not considered to be relevant enough. This also applies to aspects found in Dunn et al. 2011 such as adjective-noun order, adposition-noun phrase order and genitive-noun order but they were excluded as a result of both irrelevance for e.g. adposition-noun phrase order⁵ but also due to the limitations of a master thesis. The specific grids were included as they are relevant for the internal analysis of these branches as e.g. the grid for embedded V2 is only relevant for the North Germanic languages but it is relevant as it sets the North Germanic languages apart from the other Germanic languages that lack embedded V2 constructions. This is also true for the marking of definitiveness on adjectives which is only found among the Germanic languages and the agreement between preposition and pronouns found only in the Celtic languages. The typological data was only collected from grammatical descriptions of the languages thus not using data from WALS as was done by e.g. Dunn et al. (2011) and Greenhill et al. (2010). WALS data was not employed as it both has some issues when it comes to reliability and the expressed sparsity of the data (Greenhill et al. 2010: 2449). A table covering all sources used for the typological data is found in appendix C and the entire typological database is found in appendix G. An important difference between how the data was treated in Dunn et al. 2011 and in this thesis is linked to the etic grid methodology as in Dunn et al. (2011) the method induce the data to only have one value, in their case only one value for e.g. object-verb order (Dunn et al. 2011). This is not a problem found while using the etic grid methodology as a language can theoretically be represented by all variants in these databases, which is pointed out in Carling et al. (to appear 2013). This makes the database slightly more qualitative and it also make it easier to handle transitional languages (Carling et al. to appear 2013).

The typological grids could be grouped as pertaining to word order, nominal/pronominal morphology or verbal morphology. When it came to word order there were four grids: word order in main clauses, word order in subordinate clauses, placement of non-finite verbs and placement of clitic pronouns. A specific grid was also added to show whether or not the language allowed embedded V2 constructions. The inclusion of grids for word order in subordinate clauses was done as it is highly relevant for the Germanic languages due to the split between the languages with SOV and the languages with SVO or V2 in subordinate

⁵ Postpositions are not found in most parts of Western Europe which makes it rather redundant to study the position of the adposition as little or no variation would be found.

clauses (see appendix G). For languages with a relatively free word order the most frequent order was preferred if in existence. As subordinate clauses are found in a wide spectrum of clause types it was not easy to define what subordinate clauses to look for when investigating data relating to word order. The general guidelines were however that the data was primarily collected from adverbial clauses containing both an explicit subject and object and that relative clauses were to be avoided as they have a tendency to lack subject, object or both. Defining a non-finite verb proved to be slightly problematic for the Celtic languages as the role played by the infinitive in most Germanic and Romance languages has been filled by the frequently occurring verbal nouns. As these verbal nouns act like nouns in almost every aspect (King 1993, Ó Siadhail 1989) it is not meaningful to treat them as non-finite verbs even though they might at first appear to be what closest corresponds to an infinitive. The grids for the placement of clitic pronouns were included which is mainly relevant for the Romance languages, where there is a difference between the languages that place clitic pronouns after infinite verbs such as Italian and Spanish while other languages such as French that place the clitic pronoun before the infinite verb. The grid pertaining to clitic pronouns with finite verbs was added as it both shows possible variation among all languages but also due to the actual variation in word order found for many Romance languages as they are mostly SVO but they still place clitic pronouns before the finite verb showing some OV qualities (see appendix G for further information).

The nominal/pronominal morphological grids concerned case and gender focused on how many cases were found in the nominal and pronominal morphology and what genders were found in the nominal morphology. This was considered to be relevant as there has been a clear tendency among these languages to go from a nominal case system of four or more cases and three genders to a more limited case and gender system where English is the most extreme as it has lost both. As e.g. English and many of the Romance languages have lost the nominal case system but kept a pronominal case system it was deemed highly relevant to add a grid for pronominal cases (see appendix G). The nominal case morphology was based on the morphology of the entire noun phrase which for instance leads to the fact that the *s* marking genitive in many Germanic languages, such as English and Swedish, was not considered to be a true case as the rest of the noun phrase is not marked for genitive. It is also important to note that the case marking does not have to be found in all genders or, when it comes to pronouns, all persons, as e.g. German only makes a distinction between nominative and accusative on masculine nouns, as e.g. *der Mann, den Mann* but *die Frau* for both, but it still has four

nominal cases (Engel 1988). In a similar fashion French only differentiates indirect and direct object in the third person on pronouns, as in the difference between the direct object *le* and the indirect object *lui*, but it was still considered a three case pronominal system. A grid pertaining to how the nominal cases were marked, i.e. if the case marking is done on the noun, on the article or on both, was also added as this differs between e.g. the modern North Germanic languages and the modern West Germanic languages. Morphological grids for the combination of an attributive adjective and a definite noun (e.g. the difference between the Swedish *det stora huset* and the Danish *det store hus* both meaning *the big house*) was also added as it is relevant for the Germanic languages. Whether or not predicative adjectives are inflected for gender was also included as this differs between e.g. the North Germanic languages and most West Germanic languages and could be relevant for the other branches as well. Finally a variable for prepositional inflection when occurring with pronouns was also added as it is relevant for the Celtic languages as in Welsh e.g.: *drosta i* which means *for me* and *drostat ti* which means *for you* etc. (King 1993).

The verbal morphological grids focused on personal inflections on regular or weak verbs in present and past tense. Both present and past tense were included as there are inflectional differences between these tenses in all three branches thus making it interesting to look at both of them (see appendix G). As past tense is not expressed as just one tense in most of the languages in this study it has been operationalised as preterite in the Germanic languages, imperfect in the Romance languages and preterite or past tense in the Celtic languages. Separate variants for syncretism of certain verbal personal inflections have been added as it is highly interesting for the Germanic languages (Harbert 2007). It proved to be relevant for all three branches however as many languages displayed the same patterns of syncretism and therefore all syncretisms found in the data were added. How the languages express continuous present (e.g. *I am writing* in English) was also considered to be a relevant variable as this is done using the present in most of these languages but not in some of the Celtic languages and English (see appendix G). There were some problematic issues relating to the grids of verbal morphology as e.g. syncretism of personal inflections were given a higher number of grids than the grids of the ordinary verbal inflection, namely 16 vs. 12. This has possibly given syncretism a more important role than the ordinary personal inflection which could be seen as problematic. Even though this might be true it was the only possibility found to actually be able to convey all relevant information, as attempts were made to lower the number of syncretism related grids but it proved to be difficult. The fact that some of the languages e.g.

lacked a proper present or a preterite (as e.g. Swiss German (Baur 2002)) set them apart due to their non-values which could have resulted in a greater difference between them and the other languages. This might have been misleading but as absence of forms is relevant as well it was considered to be acceptable.

3.5. Tree models

The databases mentioned above were created to generate tree models showing the relationship between the languages' lexicon and typology. The tree models were in the form of computer-generated dendrograms with one unrooted tree model,⁶ generated with GNU/R (cladistic analysis) by researcher Joost van de Weijer at Lund University. The generating of tree models was preceded by a process of splitting up the databases into smaller databases to see whether or not that would yield any interesting results. This meant primarily that new databases were extracted from parts of the data found in the original databases. First the data for the three periods were extracted to create one ancient database, one medieval database and one modern database for the lexical and typological data respectively creating 6 new databases. As there were concerns regarding the size of the verbal part of the typological database two new typological databases were created where the first lacked all grids pertaining to verbal morphology and the second contained only verbal morphology. After the creation of the partial databases the total number of databases to be processed into tree models was 10.

When the dendrograms were generated only four of them were considered to be interesting enough to be mapped out as the remaining dendrograms conveyed more or less identical results thus making it redundant to map them out. These four dendrograms are the ones generated for the entire lexical database (figure 4.1), the entire typological database (figure 4.2), the modern typological database (figure 4.6) and the verbal typological database (figure 4.8). The dendrograms could not be directly transposed onto the map so the results had to be analysed and this was done using two different methods, namely divisive cluster analysis (Baayen 2008 via Carling et al. to appear 2013) and node counting. Divisive cluster analysis is a form of clustering achieved through calculating by means of GNU/R dependent on the number of cluster desired. It is therefore important to make an appropriate judgement of how many clusters that are needed to give the best representation of the dendrograms when mapped out. The guidelines have however been to try to have as many clusters as deemed

⁶ An unrooted tree model was only generated for the typological data as it was deemed as redundant for the lexical data.

relevant to give a more fine-grained result. To make it clearer when the results that have been mapped out have been analysed through cluster analysis the cluster groups have been named alphabetically, which is the case for maps 4.1, 4.3 and 4.6. The order of the clusters appeared to be relevant when they were mapped out as different orders lead to different similarities that in some cases were misleading. The preferred order has therefore been to follow the order found in the dendrograms thus lowering both arbitrariness and the risk of a misleading visualisation but it is difficult to avoid it all together. The other method used to analyse the dendrograms was to count the number of nodes from the root to the individual language. To make it easier to recognise when the results that have been mapped out have been achieved through node counting the groups have been named numerically according to the number of nodes found between the root and the individual languages for each group, which applies to maps 4.2, 4.4 and 4.5.

3.6. Mapping

The groupings of the tree models were later mapped out using the geographical mapping software ArcGIS, made possible by the assistance of Love Eriksen at the Human Ecology Division of Lund University. Before the results could be analysed in ArcGIS all languages had to be given individual coordinates to be properly visualised in ArcGIS. This was done in the form of a matrix where each language was assigned a coordinate. The coordinates were in most cases based on the historical centres for the languages but as many of the languages had more than one stage due to the inclusion of the historical languages the coordinates for some languages had to be chosen more arbitrarily, both due to the limited space and the need to make approximations to avoid too complex coordinates. This was e.g. the case for Old French, Middle French, French and Gaulish, as Gaulish was placed in the historical Lutetia, i.e. modern Paris, which meant that the French coordinates had to be set to other places, which in this case was Orléans for modern French, south of Tours for Middle French and somewhere outside Auxerre for Old French. All of these places are found in the heart of the French-speaking area but they are not the centres of these languages. The coordinates for some languages were highly symbolic as e.g. Gothic was placed at the mouth of Wisła even though the area it was spoken in stretched far into the southeast (and it probably survived longer there as indicated by the descendant Crimean Gothic) and Old Norse was placed in western Norway even though it was spoken in both Norway and on Iceland (Harbert 2007). The choice to map out the languages according to the principal of one language, one coordinate was preferred however over the alternative of assigning the actual ranges of the

individual languages used by e.g. Bouckaert et al. (2012). The employment of only one coordinate was favoured as the opposite would create significant problems for the dead languages as they would be found in approximately the same areas as their living descendants which would blur the results. When the coordinates were mapped out they were assigned different shapes according to their historical period giving all modern languages circles, all medieval languages squares and all ancient languages triangles. This was done to facilitate comparisons within the historical periods as well as making it easier to discern between different language stages found in vicinity to each other. To distinguish between the different clusters and groups they were colour-coded ranging from white for the first cluster or the least amount of nodes to black for the last cluster or the highest amount of nodes with the clusters and groups in between becoming gradually darker.

4. Results

The results will be presented commencing with the dendrograms later showing how the groupings found in the dendrograms translated onto the maps. The results for the lexical data will be presented first followed by the results for the typological data.

4.1. Lexical results

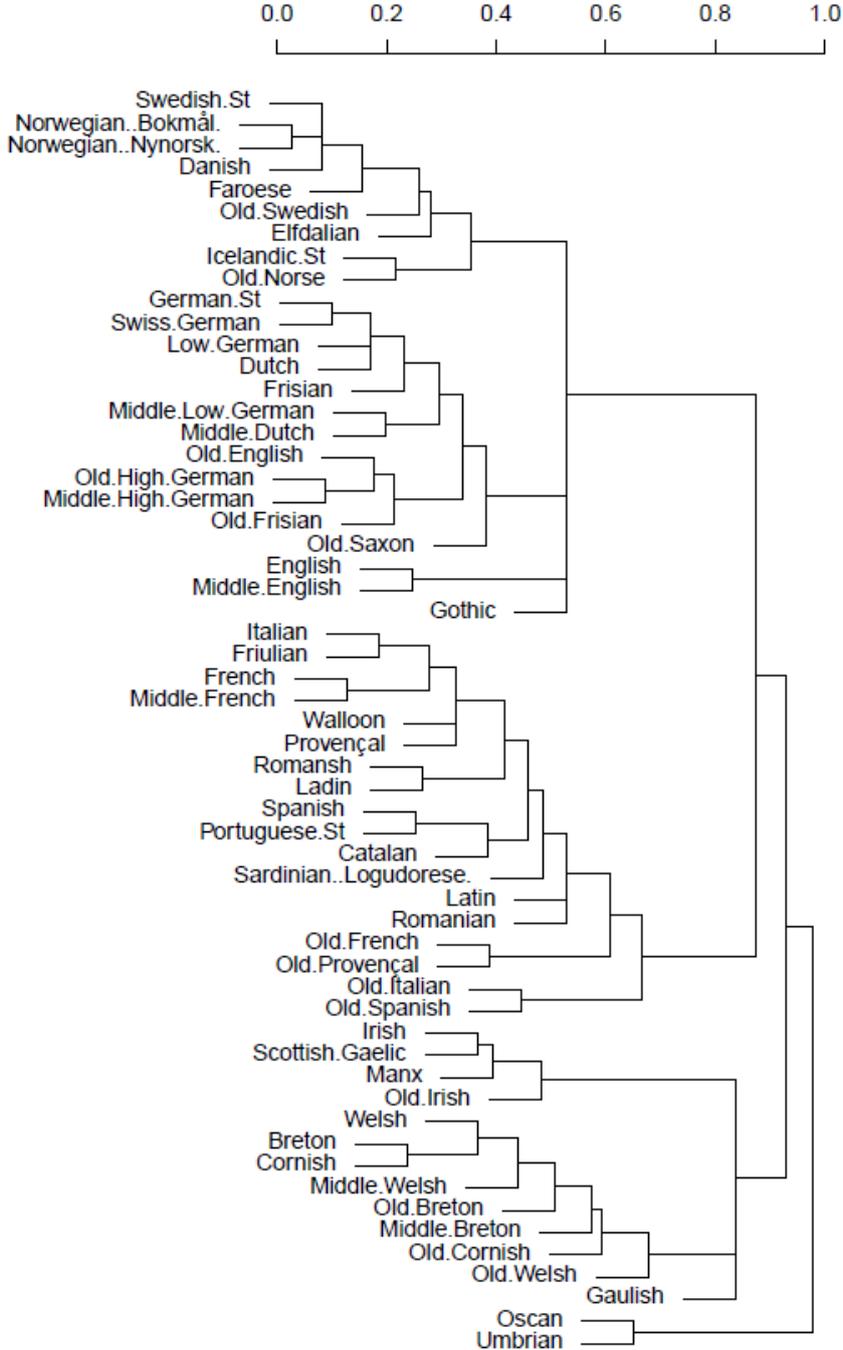


Figure 4.1: Dendrogram of the entire lexical database.

The dendrogram generated for the lexical database shows three principal branches that fully correlate to the Celtic, Romance and Germanic branches. Apart from these three branches there is also one separate branch for Oscan and Umbrian, which was caused by the scarcity of the Osco-Umbrian data and not an actual difference, indicating that Oscan and Umbrian should have been excluded from the lexical database. Returning to the three principal branches the first branch to branch off is the Celtic branch which was also true for Dunn et al. (2011) even though the Celtic languages were found among the Italic languages in Ringe et al. (2001). The first branch contains all Celtic languages studied in the thesis and divides itself into three sub-branches. The first sub-branch contains only Gaulish and thus the only Continental Celtic language in the database. It should be noted however that Gaulish also lacked some data as it had data for approximately 70% of the lexical items which might have affected the outcome, but the lack was a priori not considered significant enough to exclude Gaulish. The second Celtic branch comprises all Brythonic languages with the dead languages branching off one by one until the core at the end of the sub-branch where the living languages and Cornish are found. This gradual branching is most likely a result of the varying amount of data for these languages where Old Welsh and Old Cornish had the smallest amount of lexical entries and if they would have had a full set of data the results would probably have looked different. Welsh is the last language to branch off before the core of Cornish and Breton. The third Celtic branch is constituted by the Goidelic languages. Old Irish is the first language to branch off, but this cannot be an effect of insufficient data because Old Irish had entries for all lexical items. The modern Goidelic languages are found next with Irish and Scottish Gaelic forming some sort of core.

The two remaining principal branches share a common Romance-Germanic branch also found in Dunn et al. (2011)⁷ clearly splitting up into the Romance and Germanic branches. The first Romance languages to branch off are all dead Romance languages in this thesis except Latin and Middle French. First off are Old Italian and Old Spanish later followed by Old French and Old Provençal, highly connected to the amount of data found as Old Italian only had data for approximately 70% of the lexical entries and Old French had data for about 80%. This questions of course the inclusion of these dead languages in the database and considering figure 4.1 the last data would have been relevant. After this Latin and Romanian form two separate branches with Sardinian following shortly thereafter. The remaining Romance

⁷ This common Romance-Germanic branch is not found in Ringe et al. (2001) however.

languages branch off to form the Iberian branch (consisting of Spanish, Portuguese and Catalan) followed by the Rhaeto-Romance branch (comprising Romansh and Ladin), one Walloon branch and one Provençal branch. The core appears to be split between French and Middle French on one side and Italian and Friulian on the other side.

The last and third principal branch is the Germanic branch that is divided into four sub-branches, namely the Gothic sub-branch, the English sub-branch, the West Germanic sub-branch and the North Germanic sub-branch. It is interesting to note that in Dunn et al. (2011) all these sub-branches were more or less the same except for the English branch as Old English was not embedded within the West Germanic languages in their dendrogram and modern English is much more differentiated in this dendrogram. The sub-branches do however correspond to the ones found by Ringe et al. (2001) albeit in a different order. The Gothic sub-branch is the only representative of the East Germanic languages and is therefore in a situation similar to Gaulish and even though Gothic lacked some data as well it had data for approximately 85% of the lexical entries which was higher than the 70% of Gaulish. The English sub-branch consists of English and Middle English but not Old English. The West Germanic sub-branch can be divided into two sub-sub-branches (from now on only referred to as sub-branches) one containing all ancient languages and one containing all modern and medieval languages except Middle High German, English and Middle English. Old Saxon lies outside of these two sub-branches probably due to lack of data as it lacked data for almost 20% of the entries. The first language to branch off in the ancient sub-branch is Old Frisian most likely due to its lack of 10% of the lexical entries followed by Old English and last an Old and Middle High German core. Middle Low German and Middle Dutch are the first languages to branch off in the modern-medieval sub-branch followed by Low German, Dutch and last but not least a core made up of German and Swiss German. The North Germanic sub-branch contains all North Germanic languages where Old Norse and Icelandic are the first to branch off followed by Elfdalian, Old Swedish and Faroese. The Continental North Germanic languages form the last group being one of the most tightly knit groups of the entire dendrogram where Norwegian is the core. It is interesting to note however that Dunn et al. (2011) placed Norwegian⁸ somewhere in between Swedish-Danish and Faroese-Icelandic which is not at all the case in this dendrogram. Similarly Icelandic and Faroese were much closer in the dendrogram of Dunn et al. (2011).

⁸ NB. that the Norwegian referred to in Dunn et al. (2011) was Riksmål.



Map 4.1: Lexical results mapped out according to their cluster in dendrogram 3.1.

Map 4.1 displays the results found in dendrogram 4.1 according to what cluster they were assigned to. The name and colour of the clusters have been assigned with the first languages to branch off have been given the first letter of the alphabet and the lightest colour etc. The first clusters are Umbrian and Oscan due to their lacking data. The next cluster contains Gaulish and it is followed by clusters D to F for Old Welsh, Old Cornish and Middle Breton respectively. They are followed by the larger G cluster containing all modern Brythonic languages, Old Breton and Middle Welsh. The remaining Celtic languages are found in cluster H. The Celtic languages are followed by the clusters I and J for the dead Romance languages and thereafter all living Romance languages are found in cluster K. After the Romance languages the lexical outliers of the Germanic languages, namely English, Middle English and Gothic are found in cluster L. All other West Germanic languages are found in cluster M and the last cluster is composed of the North Germanic languages.



Map 4.2: Lexical results mapped out according to the number of nodes in dendrogram 4.1.

Map 4.2 shows the results of the lexical dendrogram analysed through node counting. The lighter colours signify languages that branch off early and therefore have a lower number of nodes and the darker colours are applied to those languages that are at the end of their branches and have a higher number of nodes. If the map is analysed according to region the Italian Peninsula has a range from the languages with the lowest number of nodes, i.e. two nodes for Oscan and Umbrian, to 5 for Old Italian, 7 for Latin and Sardinian and finally the maximal number of 12 nodes for Italian. The Iberian Peninsula has 9 or 10 nodes for the living languages and 5 nodes for Old Spanish. The languages of France have nodes ranging from 3 for Gaulish, 6 to 7 for Old French, Old Provençal, Old Breton and Middle Breton and 10 or 12 for French, Middle French, Breton and Provençal. In the Alps all languages have 10 or 12 nodes. The number of nodes for the languages in Germany ranges from 5 nodes for Old Saxon to 10 nodes for German with the rest having 8 or 9 nodes. The Lowlands display a gradual increase from 7 nodes for Old Frisian to 8 for Frisian and Middle Dutch to 9 nodes for

Dutch and finally 10 nodes for Walloon. Continuing to the British Isles the number of nodes is lowest for Old Irish and Old Welsh at 4 with most languages having 5 or 6 nodes with the exception of Old English and Middle Welsh at 8, Welsh at 9 and Cornish at 10. In the Nordic countries Old Norse, Icelandic and Elfdalian have the lowest node count at 6 nodes followed by Old Swedish at 7, Faroese at 8 and the rest at 9 or 10. The remaining languages are the outliers constituted by Gothic which has 4 nodes and Romanian with 6 nodes.

Three other lexical dendrograms were generated as well only showing the results for the ancient languages, the medieval languages and the modern languages respectively. They were not irrelevant but as they showed more or less the same results as the entire lexical dendrogram it was considered redundant to present them as well. However, as they can be of interest they have been included as appendixes D, E and F.

4.2. Typological results

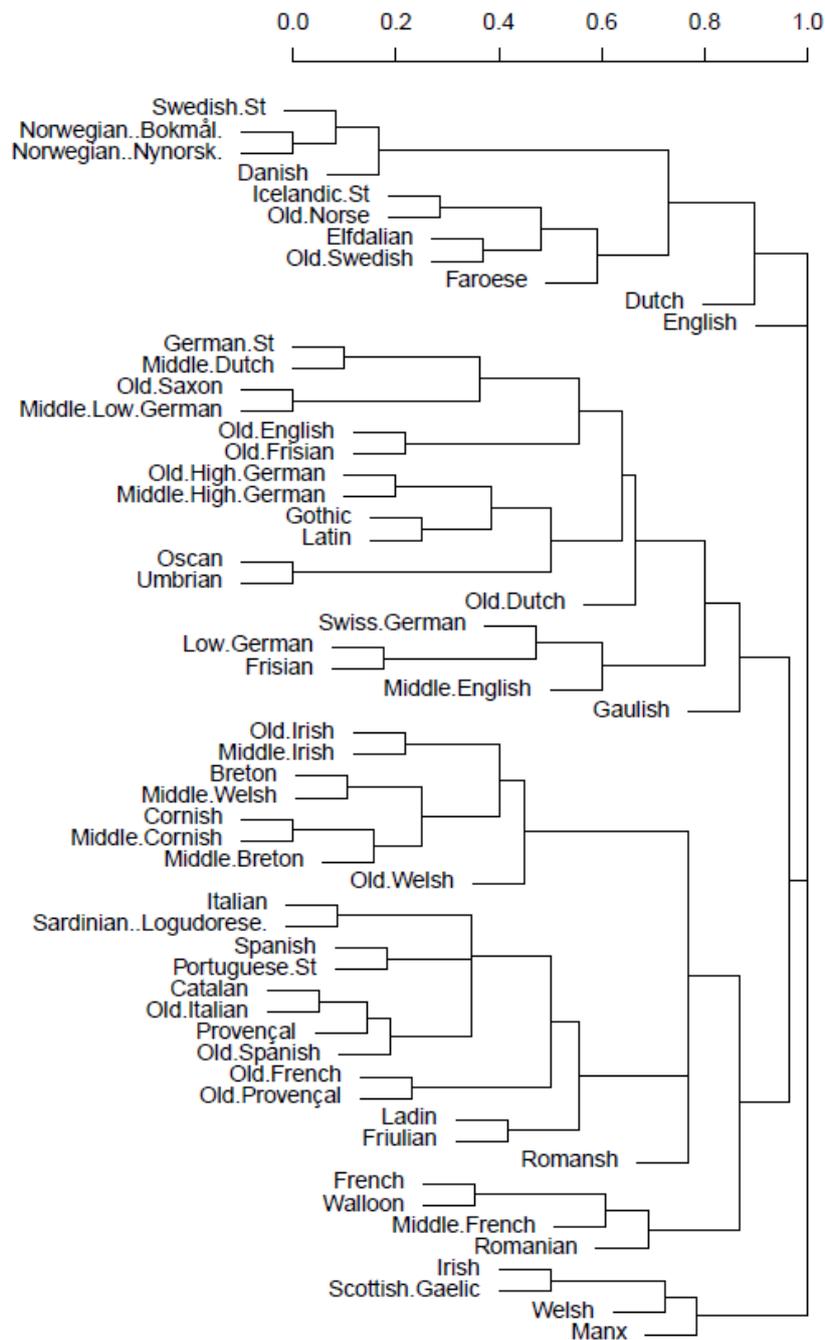


Figure 4.2: Dendrogram for the entire typological database.

The dendrogram generated for the entire typological database differs in many aspects from the lexical dendrogram. There are four separate branches originating from the root, namely the Insular Celtic branch, the Continental branch, the English branch and the Northern branch. The Insular Celtic branch consists of the modern Goidelic languages and Welsh, but none of the dead Insular Celtic languages. The Continental macro-branch contains two major sub-branches, i.e. the Celto-Romance sub-branch and the Italo-Germanic sub-branch. The Celto-

Romance sub-branch contains all Romance languages excluding Latin and the remaining Insular Celtic languages including all dead Insular Celtic languages. The first Celto-Romance languages to branch off are Romanian and the Oil languages, i.e. French, Middle French and Walloon but not Old French. They are followed by three branches comprising Romansh, all remaining Romance languages and the Celtic languages of the branch respectively. The remaining Romance languages group themselves as Rhaeto-Romance (except Romansh), Old Gallo-Romance (i.e. Old French and Old Provençal), Hispano-Portuguese, Italian-Sardinian and finally a last group consisting of Old Spanish, Old Provençal, Old Italian and Catalan. The Celtic sub-branch is divided into Old Welsh, historical Irish and Brythonic excluding Modern and Old Welsh. It should be noted that typological data for both Old Breton and Old Cornish was missing as well as some data for the verbal morphology for Old Welsh which might have misleadingly singled out Old Welsh.

The second major sub-branch of the Continental macro-branch is the Italo-Germanic branch comprising all West Germanic languages except English and Dutch, all Italic languages excluding Romance, Gothic and Gaulish. Gaulish is the first to branch off and should therefore maybe not be seen as a part of this sub-branch. It is followed by a split between a peculiar branch with Low German and Frisian at its core sided by Swiss German⁹ and Middle English and a second branch, i.e. the conservative Italo-Germanic branch, covering all dead West Germanic languages except Middle English, the Italic languages, Gothic and one mere living language, namely German. Old Dutch is the first language to branch off, probably due to the lack of word order data. The next branch consists of an Osco-Umbrian sub-branch, a Gothic-Latin sub-branch and an Old and Middle High German sub-branch. The last branch is divided into an Old Anglo-Frisian branch followed by an Old Saxon-Low German branch and finally a Modern German-Middle Dutch branch.

The English branch consists of modern English alone and the last Northern branch consists of all North Germanic languages and Dutch. Dutch splits off almost immediately indicating that it is not all that relevant to group North Germanic and Dutch. The North Germanic languages divide themselves into two major groups, i.e. the conservative group and the continental group. The conservative group consists of the dead North Germanic languages, the Insular North Germanic languages and Elfdalian. Faroese is the first language to branch off within

⁹ While considering Swiss German it is important to remember that their lack of a proper preterite might cause strange results.

the conservative group followed by a division between the Eastern North Germanic languages, i.e. Old Swedish and Elfdalian, and the Western North Germanic languages, i.e. Old Norse and Icelandic. The continental group comprises Danish, Swedish and Norwegian where Danish is the first to branch off followed by Swedish.



Map 4.3: Typological results mapped out according to cluster in dendrogram 4.2.

Map 4.3 presents the mapped out results of the typological database according to cluster. The first languages to split off are English and Dutch forming two separate clusters. The next cluster is constituted by Faroese and followed by cluster D for all the conservative North Germanic languages. The subsequent cluster E contains all remaining North Germanic languages. Manx and Welsh are the first Celtic languages forming clusters F and G respectively. They are followed by cluster H for Irish and Scottish Gaelic and later cluster I for Gaulish and cluster J for Middle English. Cluster K contains Low German, Swiss German and Frisian, showing that Middle English should not be grouped with them. The Italic

Middle Breton, Breton and Provençal. The Alpine languages have 4 to 6 nodes and Germany displays a spectrum of 7 nodes for Low German to 9 nodes for Old High German. Both the British Isles and the Lowlands have a situation similar to France with some few node languages such as English with just one node and Dutch with two nodes, some languages with a medium level of nodes such as Middle English and Old Dutch at 5 to Old Irish and Frisian at 7 and finally some high level languages such as Old English at 8 and Middle Dutch and Cornish at 9. It should however be noted that five of the thirteen British languages had four or less nodes and that five of the ten languages with the least amount of nodes were British languages showing that the low level languages are significant in the British Isles. The Nordic countries range from 4 to 6 nodes and Gothic has 9 nodes and Romanian has 4 nodes. The results point towards the Italian Peninsula, the Iberian Peninsula and Germany being the areas with the highest amount of nodes while the Nordic countries, the Alps and partially the British Isles are the areas with the lowest amount of nodes.

North Germanic. These two branches are also found in the dendrogram but here Faroese has been grouped with the continental North Germanic languages showing the double nature of Faroese. The second major branch is composed of mainly West Germanic languages, especially Ingvaemonic languages where English and Gaulish are first to split off. They are followed by Middle English and the last two branches, i.e. one historical Ingvaemonic branch consisting of Old English, Old Frisian, Old Saxon and Middle Low German and one modern Ingvaemonic branch, with Frisian and Low German at the core, with the addition of Dutch and Swiss German.

The third major branch of the unrooted tree model in figure 4.3 is identical to the conservative Italo-Germanic branch in the typological dendrogram with three sub-branches, i.e. one Italic-Gothic sub-branch, one Old Dutch sub-branch and one German-Middle Dutch sub-branch covering all forms of High German. The fourth major branch contains all Celtic languages except Gaulish and is split into two sub-branches. The first Celtic sub-branch is composed of all modern Celtic languages of the British Isles, thus excluding Cornish, and the second sub-branch contains all remaining Celtic languages which include all ancient and medieval languages. The last major branch is the Romance branch where Romanian is the first to split off, followed by the medieval-modern Oil languages, Romansh, Old French-Old Provençal, Ladin-Friulian, Italian-Sardinian and finally a core of the Iberian languages including Provençal. It is important to note that the Celto-Romance branch found in the typological dendrogram is not at all present in the unrooted tree showing that any conclusions drawn from the Celto-Romance branch shall be done with some caution. This is also true for the conservative Italo-Germanic branch and the Ingvaemonic languages and their typological affiliates Dutch and Swiss German as they are split up in the unrooted tree model while they are much closer in the dendrogram.

between the Romance sub-branch, only represented by Old French and Old Provençal and the Italo-Germanic sub-branch now without Gaulish. Old Dutch is again first of the West Germanic languages to split off but this is unfortunately due to a lack of word order data. The remaining languages are divided into an Ingvaemonic branch, comprising Old English, Old Frisian and Old Saxon, and an Italo-Germanic core consisting of an Osco-Umbrian branch and a last branch containing Old High German, Gothic and Latin, a core found in the entire typological dendrogram as well.

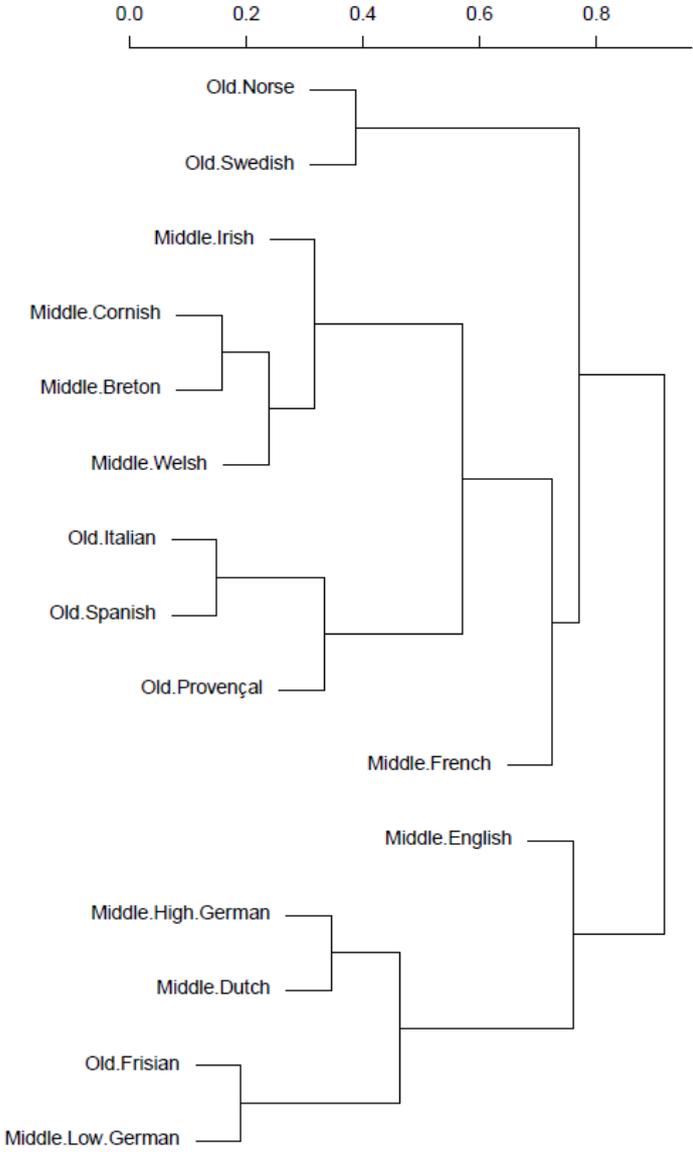


Figure 4.5: Typological dendrogram for the medieval languages.

In figure 4.5 the dendrogram generated for the medieval languages is shown. Once more the results have shifted somewhat to reveal a new situation for the medieval languages. First of all

the continental macro-branch is more or less non-existent leaving three major branches, i.e. the West Germanic branch, the Celto-Romance branch and the North Germanic branch. Beginning with the West Germanic branch Middle English is the first to branch off, which it does quite early, and the remaining languages divide themselves into an Ingvaeonic branch, comprising Old Frisian and Middle Low German, and an Istvaeonic-Irminonic branch, containing Middle High German and Middle Dutch. Continuing onto the Celto-Romance branch Middle French splits off in a fashion similar to Middle English and therefore locates itself apart from the rest of the branch which later splits up into two sub-branches. The first sub-branch comprises the other Romance languages namely Old Italian, Old Spanish and Old Provençal and the second sub-branch contains all medieval Celtic languages with Middle Irish being the first to branch off followed by Middle Welsh. The last branch is the North Germanic branch constituted by Old Norse and Old Swedish. It is finally interesting to note that the North Germanic branch shares a macro-branch with the Celto-Romance languages, also found to some extent in the ancient typological dendrogram but not at all in the dendrogram of the entire typological database.

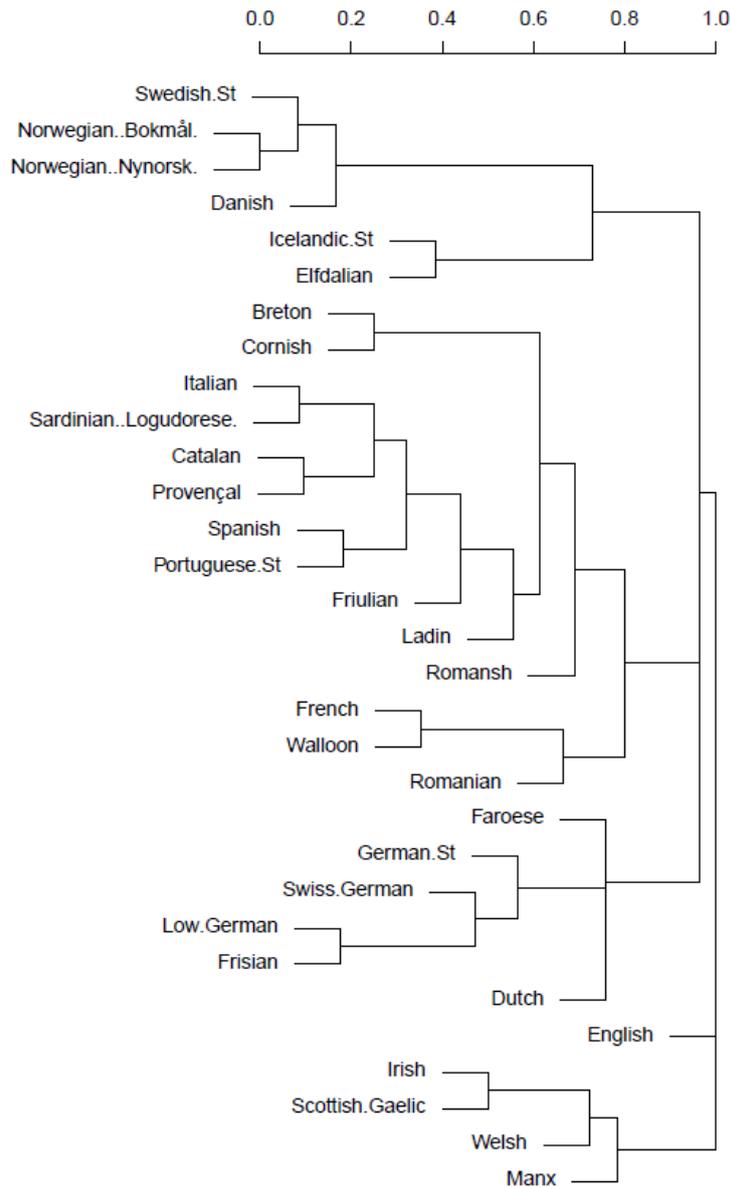


Figure 4.6: Typological dendrogram for the modern languages.

The typological dendrogram for the modern languages offers few surprises but it has some interesting differences from the typological dendrogram for the entire database. These mainly concern the former Italo-Germanic branch, now only Germanic, where the changes are apparent. First of all Dutch has migrated from a North Germanic vicinity to become closer to the other West Germanic languages. This is not all that relevant though as the differences are still significant. Faroese have for some reason done the same which can only be explained by the Faroese verbal inflection, which appears to be more or less identical to that of Dutch. Dutch and Faroese are outliers to the core formed by German, Swiss German and especially Low German and Frisian that are tightly knit together.



Map 4.5: Typological results for the modern languages mapped out according to number of nodes

Map 4.5 shows the distribution of the number of nodes for the modern languages which is interesting since the results are clearer for the modern languages than for the entire typological database. The languages of a continuous area covering the Italian Peninsula, France and the northern part of the Iberian Peninsula all have 10 nodes while Spanish and Portuguese have 9 nodes. This does not apply to Breton however as it has only 6 nodes. The Alpine languages range from Romansh with 4 nodes to Friulian with 7 nodes. In Germany German has 4 nodes and Low German has 6 nodes and in the Lowlands Dutch has 3 nodes and Walloon and Frisian have 5 and 6 nodes respectively. The languages of the British Isles range from English with 1 node to Irish and Scottish Gaelic with 4 nodes, excluding Cornish which has 6 nodes but it should actually not be considered a modern language. The Nordic languages all have 4 to 6 nodes except Faroese that has 3 nodes. Romanian has 4 nodes. This map shows that the Italian Peninsula, France and the Iberian Peninsula are the areas with the

highest number of nodes while all other areas range between 4 to 6 nodes except the British Isles where the node count is slightly lower at between 1 to 4 nodes.

The two last dendrograms generated were one where the verbal data had been excluded and one with only the verbal data. As these showed some rather unexpected results they were considered to be relevant enough to be discussed. The dendrogram excluding the verbal data will be presented first in figure 4.7.

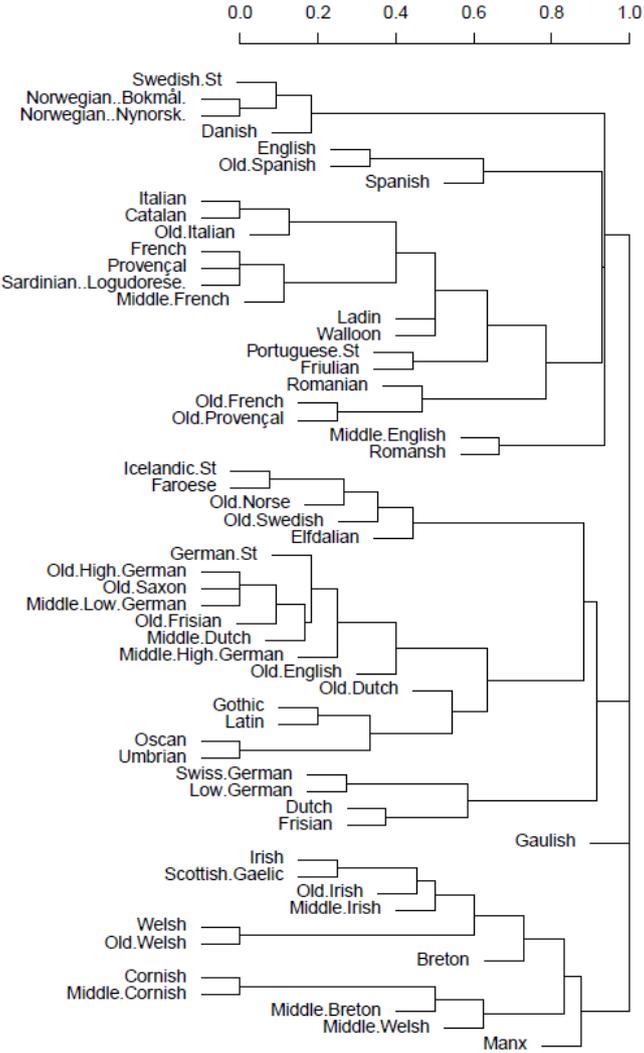


Figure 4.7: Dendrogram for the typological database excluding the verbal data

The dendrogram in figure 4.7 was generated from only the word order data and nominal data and it is clearly different from the entire typological dendrogram. The most striking difference is the formation of an exclusively Celtic branch comprising all Celtic languages except Gaelic, even though the internal differences are quite significant. The Goidelic languages, excluding Manx, form a separate entity within this Celtic branch which is far from the case

for the Brythonic languages. Cornish and the medieval Brythonic languages form one sub-branch, Breton branches off on its own and Welsh and Old Welsh branch off right before the Goidelic branch. The incohesion of the Brythonic languages is with all certainty due to the relative non-existence of nominal morphology (obviously excluding mutations) therefore giving the few grids pertaining to word order an overly important role. The Celtic branch is followed by something highly similar to the Italo-Germanic branch from the entire dendrogram but with some important additions, namely all languages of the conservative North Germanic branch and Dutch, while it has lost Middle English. The common denominator for all these languages is the presence of a nominal case system to some degree¹⁰ and this is the reason why the conservative North Germanic languages have been included. This new Italo-Germanic branch can be divided into three sub-branches where the first sub-branch contains all modern continental West Germanic languages found in this thesis except German, i.e. Frisian, Dutch, Low German and Swiss German. The second sub-branch contains all dead West Germanic languages (except Middle English), Gothic, Latin, Oscan, Umbrian and German. The second sub-branch can be divided further into an Italic-Gothic group and a West Germanic group with Old High German, Old Saxon and Middle Low German at its core. The third sub-branch comprises the conservative North Germanic languages, which differs from the entire dendrogram as there is no split between the Western and Eastern North Germanic languages, even though the core is Western North Germanic as it is constituted by Icelandic and Faroese.

Furthermore there are four remaining branches, namely a Middle English-Romanian branch, a Romance branch, a Spanish-English branch and a continental North Germanic branch. The Middle English-Romanian branch only shares three variants so it is highly questionable. The Romance branch can be divided into two sub-branches where the first sub-branch contains Romanian, Old French and Old Provençal. This slightly unexpected grouping is most certainly due to the fact that they are the only Romance languages with nominal case systems. The second branch contains the remaining Romance languages except Spanish and Old Spanish. It is important to note that French has been relocated to the centre of the Romance branch, which is definitely not the case in the entire dendrogram. It is also interesting to note that Italian and Catalan and French, Provençal and Logudorese have identical word order and nominal data while they are distinctly set apart in the entire typological dendrogram. The

¹⁰ Dutch and Frisian being exceptions of course.

Romance branch is followed by a quite peculiar branch of Spanish, Old Spanish and English. This is probably due to the scarcity of nominal morphology for both English and Spanish but why Spanish should be so differentiated from the rest of the Romance languages is perplexing, as Spanish and Old Spanish shares most of their variants with the other Romance languages (see appendix G). The last branch is the continental North Germanic branch which is identical to the one found in the entire typological dendrogram.

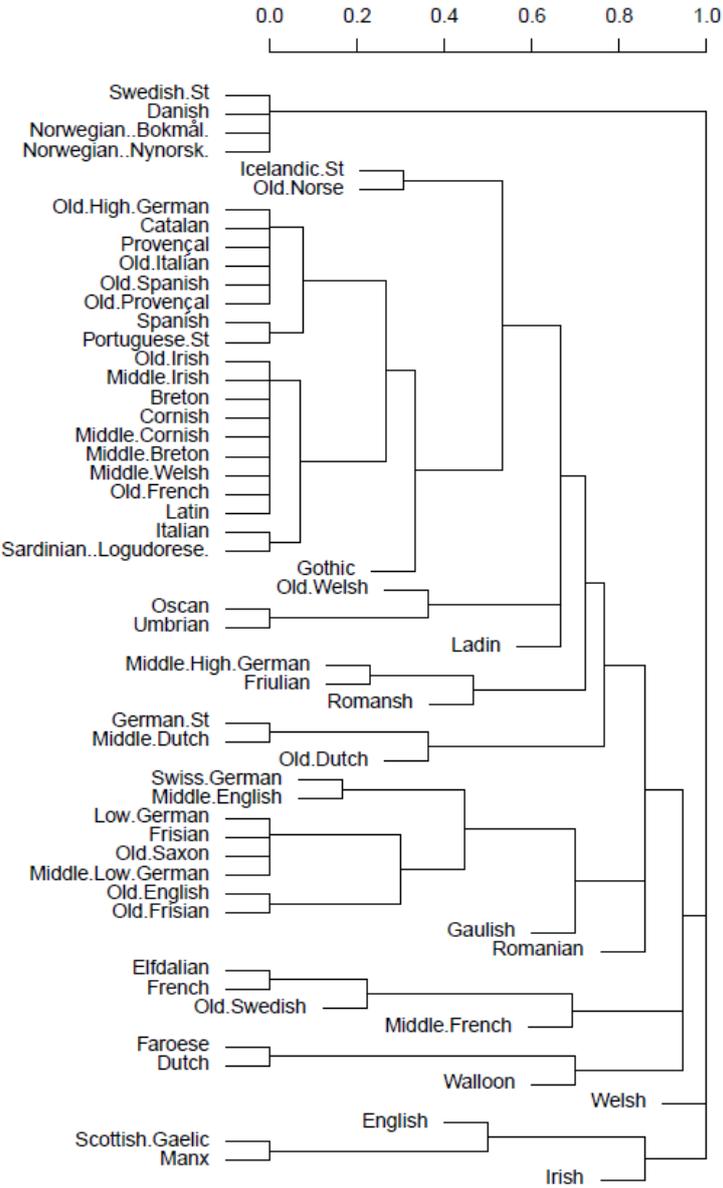


Figure 4.8: Dendrogram of the verbal grids of the typological database

The verbal dendrogram found in figure 4.8 is the most divergent of all the dendrogram and is therefore highly interesting. There are three branches and one macro-branch originating from the root, namely a British branch, a Welsh branch, a continental North Germanic branch and

the macro-branch covering all the rest. The British branch contains Irish, English, Scottish Gaelic and Manx all spoken in the British Isles. The branch is not immensely close though as Irish branches off quite early. This branch is characterised by a more or less non-existent personal verbal inflection and a progressive present such as the English *-ing* form. The fact that Welsh forms its own branch is misleadingly caused by the loss of the inflected present form similar to what has happened in the Goidelic languages but not in the other Brythonic languages. Welsh has at the same time retained an inflected past form and should therefore be somewhere in between the British branch and the rest of the Brythonic languages but that could not be showed in this dendrogram. The branches branch off from the macro-branch one by one and the two first groups to branch off are one branch consisting of Walloon, Faroese and Dutch where Walloon splits off at an early stage and the other branch comprising Middle French, French, Old Swedish and Elfdalian. The next languages to branch off are Romanian and Gaulish branching off separately followed by a major branch containing all Ingvaemonic languages (except modern English) and Swiss German. The next languages to branch off are a common branch of German, Old Dutch and Middle Dutch, a common branch of Romansh, Friulian and Middle High German, a Ladin branch, a common branch of Old Welsh and Osco-Umbrian and finally an Old Norse-Icelandic branch. These branches are followed by two major branches (excluding Gothic) that are internally more or less identical. The first of these major branches contains all dead Insular Celtic languages, Latin, Old French, Italian, Logudorese, Breton and Cornish where Italian and Logudorese form their own sub-branch. The second major branch contains all Ibero-Romance languages, Old Italian, Old Provençal and surprisingly Old High German, where Spanish and Portuguese form their own sub-branch.



Map 4.6: The verbal results according to cluster in dendrogram 4.8.

Map 4.6 shows the verbal results according to cluster. The first cluster contains only Irish and it is followed by cluster B for English, Scottish Gaelic and Manx and cluster C for Welsh. Cluster D contains Danish, Swedish and Norwegian and they are followed by cluster E for Walloon and F for Faroese and Dutch. Middle French forms its own cluster and is followed by the unexpected cluster of Old Swedish, Elfdalian and French. Romanian and Gaulish form their own clusters and cluster K contains all Ingvaenic languages except English with the addition of Swiss German. Cluster L is composed of German, Old and Middle Dutch and cluster M is composed of Middle High German, Romansh and Friulian. The subsequent cluster of N contains Old Welsh, Ladin, Oscan and Umbrian. All remaining languages are found in cluster O spanning over all families.

5. Discussion

5.1. Lexical discussion

This discussion will initially follow the same order as the results were presented with the lexical results preceding the typological results and afterwards the lexical and typological results will be compared to achieve an integrated and holistic discussion. Starting with the lexical results it is remarkable to see how well the lexical dendrograms correspond to traditional classifications of these language families, a fact that should not be taken for granted. This is not exceptional though as the tree models of Ringe et al. (2001) and Dunn et al. (2011) were also corresponding to traditional classification. This could question the relevance of the dendrogram in figure 4.1 but it should instead be seen as an indication of the validity of the method. In the tree model generated by Ringe et al. (2001) the internal relations of these three branches were even clearer but as they included such a small amount of languages it lowered the relevance somewhat. It would be interesting to see whether the results would be different if a larger amount of lexical items should be included, but as many if not most of the dead languages would be excluded then the quantity used in this thesis is apt for a diachronic study. This is apparent as the most important factor affecting the outcome of the dendrograms has become clear over and over again and it is the amount of data found for each language. As this has an impact on all the dendrograms it is relevant to discuss it before continuing with the actual results. There is an inherent struggle between the will to include as many languages as possible and making the analysis as relevant as possible. What should be the threshold for making a language relevant for inclusion then?

This thesis has showed a number of cases where the will to include a large amount of languages has triumphed when the inclusion should have been more conservative as the languages that lacked too much data just created disorder instead of adding something to the result. Where should the line be drawn and should this threshold be applied to all languages not regarding the circumstances? This becomes clear when comparing e.g. Old Italian and Gaulish as they both had entries for approximately 70% of the lexical items, which was set as an ideal minimum. Old Italian is a *Großcorpussprache* and the lack of data was caused by a difficulty to find data within the time frame given, but Gaulish is a *Restsprache* and as such 70% is quite impressive in comparison to other languages in the same situation. Old Italian did not contribute to the lexical dendrogram but it could have if a more complete set of data would have been collected which should be possible. Gaulish on the other hand gave more or

less all it could and the possibility of compiling a complete Gaulish set of data is highly implausible. Even though it is highly interesting to include Gaulish as it is a Continental Celtic language the results show that 70% is not enough to make a meaningful analysis. Even 80% as in the case of Old French appeared to be insufficient to make a relevant contribution to the result showing that the 90% of Old Frisian might be a more reasonable threshold. Old Saxon sheds a similar light on this as it had approximately 80% and it showed only that it was a West Germanic language and for Old French it failed to make a contribution worth mentioning. The most valuable lesson to learn from this is that it is obviously worth the while to complete the data set for the individual languages, as it becomes more important than the amount of languages.

With this said there are interesting results especially when it comes to the Germanic languages as they are the best documented languages in this thesis. The first is that the historical languages branch off gradually which is most prominent for the West Germanic languages. Whether this is due to an actual tendency or the nature of the data found for the historical languages is hard to say but the results appear to point towards the fact that it is a relevant tendency. As the West Germanic languages are not grouped primarily according to their genealogical branches but instead to their historical period it suggests that lexical change could be a result of periodical trends instead of branch internal changes. However there is a prerequisite for the spread to be possible and it is some sort of communication between the different areas. That mutual intelligibility should be this prerequisite is unlikely as loans can occur between unrelated languages but in the case of cognates as studied in this thesis it must obviously occur within the same language family. If an area becomes isolated from the rest of the language family they will not take part in these trends and therefore diverge from their related languages. This becomes clear when it comes to e.g. Icelandic and Romanian as they are geographically separated from the other languages in their respective families. The isolation does not have to be due to geographical reasons as it could just as well be due to cultural or political reasons (Campbell 2004).

Whether it is the isolation or the loss of mutual intelligibility that comes first is a complicated issue but theoretically it is plausible to hypothesise that the less contact a language has with the rest of its family the more it will develop independently thus leading to a greater chance of losing mutual intelligibility with its relatives. Where the lexical trends start is without intensive etymological research more or less impossible to say but it can be presumed that the

languages found at the end of a branch are the most probable proponents of the spread even though they might not be the source. This should be done with some caution though as e.g. in the entire lexical dendrogram Norwegian is at the heart of the North Germanic branch due to the fact that Bokmål and Nynorsk are almost identical which misleadingly would indicate that Norwegian is the most probable source of change in the North Germanic languages even though this is not necessarily the case. The fact that both Old High German and modern German are found at the end of their branches is probably not a coincidence as they have influenced the Germanic languages heavily. Traditionally the source has been accredited to a prestige centre (Campbell 2004) but there is no theoretical ground to make such an assumption even though prestige centres are definitely pivotal in the spread of features.

Isolation is one important factor but it is perhaps more relevant to talk of it in terms of periphery versus centre were the trends spread from the centre to the periphery with more and more aspects being lost the further out in the periphery the languages are found.¹¹ This is the basic idea of the *wave theory* postulated by both Schmidt and Schuchardt in the 1870's, even though it has been traditionally applied to sound change (Campbell 2004). The wave theory could also explain why the historical languages branch off earlier as it is an utter impossibility for the historical languages to have undergone the same amount of waves of change as the modern languages. If German is seen as the centre of the West Germanic languages then the dendrogram in figure 4.1 would suggest that Frisian is the most peripheral language followed by Low German and Dutch which is not far-fetched. It would imply that Swiss German is a central language though which is questionable but as German and Swiss German are as closely related as they are they maybe should be seen as parts of a common macro-language. It is of course important that this does not apply to the actual modern situation but instead the historical situation that has resulted in the situation found today.

If the same is done with the North Germanic languages Swedish, Norwegian and Danish maybe they too should be regard as one macro-language as they also are closely related were Icelandic is the most peripheral followed by Elfdalian and Faroese. Why Faroese is the language closest to the centre is perhaps due to a strong presence of Danish but why the presence of Danish in the Faroe Islands should be stronger than Swedish in Älvdalen (where Elfdalian is spoken) is hard to see. It could also indicate that Faroese became separated from

¹¹ NB that a peripheral language does not have to be geographically peripheral as it can be socially peripheral as well, i.e. due to political or religious separation from the other languages.

Norwegian later than Icelandic which is more plausible. It is of course also important to note that the 19th century “purification” of Icelandic distorts the results somewhat. The same pattern could be seen among the Romance languages with French and Italian as the centres and Romanian being the most peripheral followed by Sardinian, the Ibero-Romance languages the Rhaeto-Romance languages and closest to the centres Walloon and Provençal are found, both geographically close to French and Italian. It is interesting to note that the two centres found in map 4.2 correspond to the core of the SAE, but this should be regarded with some caution though as SAE only pertains to typology and the validity of the SAE could be questioned (Harbert 2007: 11-12). That the Ibero-Romance languages have their own branch is an indication of the relative separation of the Iberian cultures from the rest of the Romance languages having their own centre of Spanish and Portuguese. This is also the most likely reason to why Catalan and Provençal are set apart in the lexical dendrogram as Spanish was the centre for Catalan while French was the centre for Provençal. It is unfortunate that the ancient and medieval Romance languages were so dissatisfactorily covered in the database as it would have been interesting to know where they actually would have been located.

Last but not least it is interesting to note the internal differences found within the different families. Even though the birth of the Romance languages has happened in historical times the amount of shared vocabulary of the Romance languages is as large as for the Germanic languages even though the Germanic languages were separated in prehistoric times. This shows that the Germanic languages have either experienced little lexical change or that they have been in extensive contact with each other thus preserving a common lexical sphere. For the Romance languages it indicates instead that a significant amount of the Latin lexicon has been replaced either due to internal changes or external influences from the pre-Roman languages. The Celtic languages display much larger internal differences which might point towards an earlier split between the Goidelic and Brythonic languages.

5.2. Typological discussion

The typological results were in most aspects more interesting than the lexical results as they showed a situation that is quite different from the traditional classification of these languages. What was most striking was that the typological dendrogram grouped languages that were from different families but even more interesting is the fact that some families were clearly separated. The two most imperative examples are the Celtic and Germanic languages as there

were no common branch for either Celtic or Germanic. The independent branch for the North Germanic languages indicates that something must have happened in prehistoric time (i.e. before written accounts of North Germanic) to set them apart as much as they are. The fact that all non-continental groups of languages had separate branches is highly interesting but further conclusions regarding these languages are unfortunately difficult to make. The typology of the Italo-Germanic branch should reasonably be interpreted as the original Germanic typology as all dead West Germanic languages and Gothic share it. That they would have developed such similar systems simultaneously is implausible, especially if the fact that the Italic languages share the same system is taken into account as it indicates that the other languages have diverged from this system. Interestingly enough is that modern German is the only surviving language of this conservative group as all other Germanic languages have changed, but it is however important to point out that even German has changed quite a bit approaching the traditional Ingvaeonic typology. That German has changed in an Ingvaeonic direction could be explained by the fact that the political centre of Germany has moved northwards to Berlin and many important German cities are found in traditionally Low German areas, e.g. Hamburg and Hannover. This could however not explain why Swiss German, an Irminonic language found far from the Ingvaeonic areas, is closer to Low German than German is showing that it probably is more complex than that.

The Celtic languages seem to have experienced a development similar to the Germanic languages with regard to the Goidelic languages splitting off from the more conservative Celtic languages. As all dead Celtic languages are found together with Breton and Cornish it suggests that these two languages are closest to an original Insular Celtic typology while the remaining modern Insular Celtic languages all have diverged. Where this divergence started is impossible to tell from this data but there are two possible conjectures to be made. The first is that these languages all have been affected by English as the Celtic-speaking areas of the British Isles all have been dominated by English in the New Era, implying that the changes have an external source. It is worth noting in this case that the English verbal morphology is grouped together with the Goidelic languages in the verbal dendrogram in figure 4.8 showing that some sort of exchange and/or convergence might have taken place. The second conjecture is that the changes started in one of these languages and spread across the Irish Sea, but for this to be true it must have happened quite early as it must have predated the English supplantation of the Celtic languages in the area.

Continuing with the Celtic languages it is interesting to see how the Romance languages have set themselves apart from the original Italic languages and become closer to the Celtic languages. If this development is of relevance the question is what caused it. One possibility is that the Celtic languages came under extensive Vulgar Latin influence and consequently became more similar to the Romance languages. This is slightly problematic as e.g. Ireland was never conquered by the Romans, but the Latin influences were still present however due to the later advent of Catholicism even though this Latin was not Vulgar Latin. A second possibility is that the Celtic languages have influenced the Vulgar Latin as Celts were found to some extent in all Western European areas of the Roman Empire except Italy and the southern part of the Iberian Peninsula. The problem this time is of course that most of them were Continental Celtic languages and as such it is hard to draw any conclusions as the other languages are Insular Celtic. Gaulish was a Continental Celtic language and it is not among the conservative Insular Celtic languages in the typological dendrogram, but as Gaulish lacks a significant amount of data it does not really tell us anything relevant about the typology of the Continental Celtic languages, except the fact that it mainly had SVO word order (Lambert 2003) which is found among all modern Romance languages in the data.¹²

It should be noted however that nominal cases have been lost in all Romance and Celtic languages found in areas conquered by the Romans where Old Welsh is the first instance of non-existing nominal morphology found in this data. Outside the former Roman territory nominal cases are still found in the Celtic languages as is true for Irish and Scottish Gaelic. The third possibility is that it is due to a presumed Italo-Celtic branch supported by among others Ringe et al. (2001), which could also be strengthened by the shared Italo-Celtic verbal morphology in figure 4.8. The non-Romance Italic languages have however a nominal morphology that is different from the Insular Celtic languages as seen in the dendrogram excluding verbal morphology, but this is probably caused by the VSO word order and the loss of cases in the Brythonic languages, but e.g. Old Irish has more in common with Latin than what appears in the dendrograms, but the differences are still significant (see appendix G). Whatever the cause might have been there is interestingly enough a development where the Romance languages have abandoned the Latin nominal morphology while keeping most parts of the verbal system, something that has happened in the Celtic languages as well excluding the Goidelic languages and Welsh where the verbal system have been discarded instead. That

¹² With the exception of Romansh which appears to have V2 word order (see appendix G).

these two developments are connected to word order changes is plausible as the former languages all have SVO or V2 word order while the latter ones all have VSO word order.

The typological dendrogram does not directly display a development similar to the wave theory as seen in the lexical results. However the results found when mapping out the number of nodes present a different picture as there are two darker centres in map 4.4, namely northern Germany and the western part of the Mediterranean region. These two areas could be actual centres of typological spread as they are geographical centres of extensive contact with the surrounding areas. In the case of the Northern centre it comprises some of the earliest Germanic areas and later also the Ingvaemonic languages. This supports the notion that Ingvaemonic is a highly innovative sub-group as the Ingvaemonic languages are among the most embedded languages in the dendrogram. Moving farther from this centre the Germanic languages become less and less embedded thus sharing a smaller amount of features. Even though this is true it is slightly misleading for the North Germanic languages as they constitute their own branch and Low German has a lower node count than e.g. Old Saxon but it has instead partially branched off to create a new sub-branch. Similarly both Latin and Old Italian have high node counts but they are not found on the same sub-branch. High node counts shall not be seen as coherent areas of innovation but instead as central in their respective group. Once again these two centres are found at the core of the SAE which could be seen as some sort of support of the existence of the SAE but it is very important to not neglect that French is not a central language in the entire typological results, which it clearly is in the SAE. The node counting for the modern typological data found in map 4.5 does indicate however that French is part of a central area covering south-western Europe but as this excludes German and includes the Ibero-Romance languages the map in 4.5 also stands in contrast to the supposed SAE area. The relevance of the SAE might therefore be put into question.

It could be implied that if the wave theory is applied to typological changes the spread is determined by certain defining features, which means that if a language does not have these defining features they will probably not undergo the change. If true, these defining features are the equivalence to communication for the lexical changes. Some changes are either facilitated or hindered by a certain feature thus separating a group further from the rest. This appears to be the case for the Celtic languages with VSO word order as they have diverged from the SVO languages. The use of e.g. nominal suffixes in the North Germanic languages

has probably been decisive for the separation of North Germanic from the other Germanic languages. Perhaps even the change from SOV to SVO in the Romance languages lead to the demise of the case system, but it is certainly to go too far to claim that from this data. However it is interesting to consider a scenario where the Romance languages are influenced by the Continental Celtic languages such as Gaulish thus changing to SVO word order with the effect that the case system is rendered uneconomical and therefore lost. There is a tendency among the Western European SVO and V2 languages to lose their case system as almost all modern SVO and V2 languages have lost their nominal case system except German, the conservative North Germanic languages, Low German, Swiss German and Romanian. The Germanic languages that still have a nominal case system all have a word order for subordinate clauses that is SOV or V2.

Changes of word order are not easy to explain from these results as there are no clear tendencies. Word order is also quite elusive in comparison to many of the other features and it is therefore harder to understand how such a change takes place. There are two languages with a clear transition though and it is English and French. Both Old French and Old English have V2 word order but Old English had some SVO elements as well suggesting that Old English was in a transitional phase. This transitional phase is also found in Middle French where both V2 and SVO are found but in Middle English V2 has been superseded by SVO which also happened in French. The English transition from V2 to SVO could be attributed to a supposed creolisation or an internal change. Why Old French was a V2 language could be explained by Germanic influences but as Old Provençal was V2 as well it becomes more complicated. The change from V2 to SVO is perhaps not that dramatic but for the Celtic languages the changes are rather strange. Old Welsh appears to have had VSO which is true for Welsh as well but Middle Welsh was V2. Whether the word order of Middle Welsh actually was V2 or that this was just found in the written sources is a relevant question but this was not revealed from the data. Another possibility is of course also that both V2 and VSO were possible word orders but I did not find any information explicitly stating that. If Middle Welsh had V2 it is worth reflecting upon why it changed back to VSO. Breton experienced a similar change but from SVO in Middle Breton to V2 in modern Breton (Favereau 1997) thus making modern Breton the only V2 language in an area of SVO languages, including Cornish. This shows that changes in word order are complicated and sometimes incomprehensible, but these changes are highly important though as they can alter large parts of the typology as discussed above. Unfortunately nothing can be said about the

peculiarities of V2 in the Germanic languages and VSO in the Insular Celtic languages as these changes are found in the ancient languages and thus being prehistoric changes.

Even though the wave theory has some difficulties explaining the entire typological data it is highly relevant for the verbal results found in map 4.6. It is most noticeable among the North Germanic languages where Old Norse is the first language found in the last cluster of O. The Eastern North Germanic language of Old Swedish underwent a change and ended up in cluster H. This change spread to Elfdalian as well as they have more or less the same verbal system and Elfdalian is therefore also in cluster H. The Continental North Germanic languages changed again thus spawning cluster D which is the verbal system found in Swedish, Danish and Norwegian. This change did not find its way to Elfdalian however and leaving them in cluster H. The Insular North Germanic languages never underwent the change in Old Swedish as Icelandic remained in cluster O. The situation for Faroese is a bit complicated however as one would predict that they would change in accordance with the Continental North Germanic languages but even though the verbal system has been simplified it appears to have undergone a separate change setting it slightly apart. A similar situation can be found for High German as Old High German is found in cluster O while Middle High German has changed and therefore moves to cluster M and modern German goes through further changes and ends in cluster L. Also French and Middle French display a pattern comparable to this as Middle French has changed into cluster G while French has changed into cluster H and Walloon has changed into cluster E. Even though cluster H is highly bewildering it is not all that strange considering that both French (through Middle French) and Old Swedish are outcomes of the original system of cluster O converging as a result of the limited amount of possible systems. For the Celtic languages it is possible to conjecture that the changes in the verbal system originated in either Scottish Gaelic or Manx, i.e. cluster B, as they lack both personal inflection for present and past tense. Irish on the other hand retained some inflections especially for the present tense thus being only partially affected which is also true for Welsh as it retained the inflections for the past tense. This analysis could be misleading though as these changes could have occurred on their own as the only thing in common for these systems is the loss of personal inflection, even though they all share that continuous present is expressed with some sort of progressive form (see appendix G).

Furthermore there are two large languages that are rather perplexing, namely English and French. This is not a surprise in itself but the problem is how they shall be explained.

Unfortunately the Celtic languages spoken in central England have more or less been lost so it cannot be known with certainty if they have affected English. The loss of cases is the only aspect found in English and all Brythonic languages but this has happened in many other modern Germanic languages and is therefore not particularly relevant. Other substrates or possible superstrates might be suggested but whether this is true or not is impossible to say from this data, but the fact that English is as set apart as much as it is supports the notion of a possible creolisation. French on the other hand is hardly a creole but something has evidently happened when it comes to French typology. The presence of Celtic influences is not implausible as French is spoken in the heart of historical Gaul but the later development of a decimated verbal morphology is possibly caused by Germanic languages, as the Germanic-speaking Franks played an essential role in the creation of the later French culture. Syncretism in present verbal inflection is found in most of the Germanic languages but in the Romance family it is more or less only found among the Oil languages.

The notion of Germanic being the cause of change for the French verbal morphology is actually reinforced if an excluded feature is considered, namely pro-drop. Unfortunately there has been no data collected for pro-drop but it would have been relevant as the contrast is quite stark between the pro-dropping Romance languages and the Germanic languages where the subject pronouns cannot be omitted in verbal phrases (Harbert 2007).¹³ Most Romance languages have pro-drop (Harris & Vincent 1997) and they also have the original verb system found in cluster O. The modern Germanic languages do not have pro-drop and all except Icelandic are found outside of cluster O as they have a varying degree of inflectional syncretism. French is one of the few Romance languages with mandatory subject pronouns in verbal phrases (Helleland 2006) and it also has a high degree of syncretism. This development within French can be due to Germanic influences, but it could of course be the other way around that the personal inflections started to syncretise which lead to the necessity of a mandatory subject. In Old French, which had the original verbal system, the subject pronouns were not mandatory and Old French was predominantly pro-drop (Buridant 2000; 424) and the cluster analysis shows that the changes must have occurred in Middle French. It could be problematic however to say that mandatory subjects enables simpler verbal systems instead of saying that simplified verbal systems necessitate mandatory subjects even though the result is the same.

¹³ With the only exception being Gothic (Harbert 2007).

5.3. Combined discussion

If the lexical and typological results are put together there are some interesting outcomes. The first outcome is the difference between the ancient Celtic languages in the lexical dendrogram and in the typological dendrogram. The early separation of the Celtic languages in the lexical dendrogram would indicate that the Celtic languages first changed lexically and that the major typological changes did not occur until the Middle Ages. That the Celtic languages should be lexically divergent is also found in a similar lexical dendrogram generated by Dunn et al. (2011) for a larger number of Indo-European languages. The later change is most certainly related to the loss of important Celtic centres both on the continent but also later in the British Isles. The continuous Celtic-speaking area surrounding the Irish Sea was lost and the Celtic languages became marginalised. Continuing onto the Romance languages it is interesting to find that Romanian is the only Romance language that is divergent in both the lexical and typological results. This clearly shows that isolation or peripherality is of relevance for both lexical and typological features. Two groups of Romance languages show that typological divergence does not hinder lexical diffusion as both French and the Rhaeto-Romance languages branch off early in the typological dendrogram but they are found at a central level in the lexical dendrogram. This could be seen as an indication of the fact that lexical changes can spread as long as communication is possible where a divergent typology does not have to be an obstacle. One of the most remarkable examples of the opposite is Logudorese which is found at a central level in the typological dendrogram but it is also one of the first Romance languages to branch off in the lexical dendrogram. This reversed development is harder to understand as the lexical data suggests that it has been fairly isolated while the typological data suggests the exact opposite. The one possible explanation is the existence of Italian literature on the island while the endemic Sardinian has mainly been a spoken language thus creating a situation where Italian has been the written norm for the Sardinian language. It is questionable whether written Italian could have had such an impact as general literacy has not been present before the Industrial Era. If it is true however it could mean that written standard languages can significantly alter the typology of colloquial and minority languages.

The combined results also shed some new light on dilemmas relating to classification. One of these is the case of Occitan or Provençal mentioned in chapter 2.2. The closest relative of Occitan is sometimes considered to be Catalan even though they often are classified as belonging to two different branches, namely Gallo-Romance and Ibero-Romance

respectively. These results point instead towards the fact that Occitan has a Gallo-Romance lexicon but an Ibero-Romance typology. It is however important to note that there are no Gallo-Romance languages except the divergent Oïl languages in this thesis which might be misleading as there is a possibility that the non-Oïl Gallo-Romance languages are more similar to Occitan which would justify to classify it as Gallo-Romance language. The typological result could be interpreted as Occitan originally being an Ibero-Romance language that later came under Gallo-Romance influence. This becomes problematic when looking at Old Provençal as it constantly groups itself with Old French, even though there are significant similarities between the Old Provençal and the Old Spanish typology (see appendix G). It would therefore be highly relevant to investigate where Old Catalan would have placed itself in these dendrograms but as this was not possible we are left to speculate. The classification of the Rhaeto-Romance languages was also mentioned in chapter 2.2 and these languages display a similar situation as there are differences between the lexical results and the typological results. The Rhaeto-Romance languages are found close to the Gallo-Romance languages in the lexical dendrogram which could support that they are Gallo-Romance languages, but it is still questionable as Ladin and Romansh are still found on their own branch. Friulian is however close to Italian which shows that the fact that Ladin and Romansh branch off earlier could be a result of them being more isolated than Friulian. It is also hard to make any conclusions pertaining to the Gallo-Romance languages as such a large part of them are missing from the data, i.e. the Gallo-Italian languages. The typological results for the Rhaeto-Romance languages present a different scenario though as they are quite clearly separated from the other Romance languages which could support a separate Rhaeto-Romance branch, but they lack a common branch in the typological dendrogram however which could be due to the greater impact German has had on Romansh in comparison to the other Rhaeto-Romance languages. Another question pertaining to classification is the results for English as it is clearly differentiated from the rest of the West Germanic languages in both the lexical and typological dendrograms. The question is therefore if the changes English has undergone are significant enough to actually set it apart as its own branch. More aspects must of course be considered but the results show that it is a relevant question.

Even though speculations should be avoided there is one speculation that might be interesting to make concerning the relationship between these language families. As the Germanic languages and the Italic languages share more vocabulary than they respectively do with the Celtic languages it is possible that the Germanic and Italic languages separated at some point

after the Celtic languages had split off, which is found in the results of Dunn et al. (2011) as well. This stands in contrast to the traditional grouping of Italo-Celtic supported by e.g. Ringe et al. (2001) and should therefore be approached with some scepticism. It would be easy to discard it as irrelevant if not the Italic and Germanic languages are grouped once more in the typological results showing that there might be something to this combined Italo-Germanic grouping. There are some aspects that make it less plausible though as e.g. the geographical distance between the two language families, but as both the Italic languages and the Germanic languages must have predecessors originating outside their respective areas as they are not indigenous to either Scandinavia or Italy. This origin is in both cases most likely to the east of these languages making it slightly more plausible. It should be noted that a need for a larger amount of data covering all Indo-European branches would be necessary to make any further speculations.

The maps for the lexical and typological results mapped out according to cluster do not reveal any greater surprises. One pattern could be seen though and it is that it is possible for languages to undergo immense typological changes without changing the lexicon. This is e.g. the case for the Goidelic languages, Welsh, the Continental North Germanic languages, Dutch and last but not least English. Comparing the maps of the lexical and typological results according to node count presents two different patterns. The lexical map shows high node counts for most central languages with the peripheral languages having lower node counts. These central languages have the highest node count in the central parts of Western Europe with a high node area stretching from Italy to France with the Alps and Walloon being found just outside the core of this area. The high level of node counts is however not exclusive for these Romance languages as they are also found among the Germanic languages and Celtic languages. In the typological map there are only two or possible three areas with a high node count, namely Northern Germany and an area from Catalonia to northern Italy as mentioned in chapter 4.2. The possible third high node area is Celtic languages on each side of the English Channel, i.e. Cornish and Breton.

The two first areas are also among the areas with the longest presence of the respective languages, which might be a coincidence though. These areas could either be seen as innovative centres or conservative cores. That they should be innovative centres is questionable as some of the least conservative languages such as English and French are found outside of these areas, and innovation would most likely lead to new branches which

means fewer nodes as in the case of English. It is therefore more appropriate to regard them as conservative, even though it might be misleading, as it is true in most of these cases. The common branch of Italian and Logudorese could actually be explained as a result of this as it would indicate that they share the same archaisms instead of the same innovations. All dead West Germanic languages have high node counts and the only living language with a high number of nodes is German which has been shown to be typologically conservative. The results for the North Germanic languages questions this though as neither Icelandic nor Old Norse have high levels of nodes but this is probably due to the new branch that Old Norse embarked upon when leaving the Continental Germanic languages. The changes in the Continental North Germanic languages seem to have been significant enough to set them apart on a separate branch which once again blurs the result for the North Germanic languages.

6. Conclusions

The first conclusions that can be drawn from the results are that the lexical results were what could be expected as they mainly correlated with traditional classifications and Dunn et al. (2011) and were therefore not very interesting. The Celtic languages were the most differentiated of the language families, which was also true for Dunn et al. (2011) but not Ringe et al. (2001), which leads to a speculation if an Italo-Germanic group could be of relevance, which was later reinforced by the typological results. The lexical results did however display a pattern best explained by the *wave theory* as there were groupings both in time and space where the modern and central languages were found at the core. This could indicate that the wave theory also could function as an explanation to the general process of lexical change which is of interest to investigate further. The main proponents of these waves of changes ought to be political, cultural and prestige centres as the languages found in these centres were the last languages to branch off. Languages found far from these centres or that were isolated for some other reason branched off earlier than the other languages showing that they had not been affected by all the waves of change. Similarly the relative isolation of certain languages due to geographical distance have proved to be relevant as the more isolated languages split off earlier than non-isolated languages, something found in Carling et al. (to appear 2013) as well. Therefore it is of interest to see whether it is possible to predict lexical change from these sorts of centres and how far they will reach out into the periphery. It should of course be noted that the results found in this thesis is a product of more than two millennia

of lexical change which might be difficult to directly apply to a modern environment and that these patterns could not be seen in Dunn et al. (2011) but they could in Ringe et al. (2001).

The typological results were more relevant though as the groups did not follow the traditional classification as all three families were separated to some extent. All non-continental languages (including the North Germanic languages) were found on branches separated from the continental languages, which is interesting but impossible to draw further conclusions from in this thesis. These separate branches all had related languages found among the continental languages and these continental relatives formed the typologically conservative languages. What was even more interesting was the common Italo-Germanic branch shared by all the conservative West Germanic languages, Gothic and the Italic languages, i.e. Latin, Oscan and Umbrian. All Romance languages were instead found closest to the conservative Celtic languages indicating a possible affiliation between those two groups. The loss of case in the Romance languages has been postulated to be due to the change from SOV word order to SVO as it was found that while all Celtic languages with SVO lacked cases but those Celtic languages that had VSO also retained their case system to a higher degree. A general conclusion might be drawn concerning the fact that most SVO and V2 languages in this thesis had lost their case system with two exceptions, namely the languages with SOV or V2 word order in subordinate clauses. It would be relevant to see if there could be a significant pattern between differentiated main clause and subordinate clause word order and case systems.

Even though the wave theory appears to be applicable to lexical change it is not as relevant for the typological data with the exception of verbal morphology. The results found for e.g. the North Germanic languages showed both a difference in time and in space similar to the lexical results. Pro-drop was also concluded to be relevant for verbal syncretism, but whether it is the mandatory subject or the verbal syncretism that comes first was not possible to tell as no pro-drop data had been collected. The typological results visualised by the maps instead showed a process of change that appeared to function in a manner that was completely different from the lexical changes as there were two conservative centres for Germanic and Romance respectively. The Germanic centre was found in Northern Germany corresponding to one of the earliest Germanic areas, which was considered to be of low relevance however. Languages found outside of this area had branched off and experienced typological changes of a varying degree. The Romance centre was found around the Western Mediterranean from Catalonia to northern Italy. These conservative centres have undergone fewer changes than

the other languages and are therefore closer to the earlier languages. The typological changes should therefore not be seen as originating in these centres but instead in areas found farther away from the centre. Examples of languages that have experienced significantly more changes than these centres are French, Romanian, the North Germanic languages and English. These two parallel processes are interesting to look for in the entire Indo-European language family as well as other language families. A brief discussion whether these conservative centres could be seen as an indication of the SAE concluded that the relevance of the SAE was worth questioning as the results did not appear to support the SAE. It is also interesting to note that the results for the verbal morphology could be seen as contradicting the process found for the entire typological result, but the fact that the wave theory is applicable to verbal morphology does not oppose the notion of conservative centres as they too can be affected by the waves of change just to a lesser extent. If typological change is not primarily initiated in the centre as the conservative centres point towards it is highly interesting to examine the sources of typological change to see if innovative centres exist as well or if it is caused by other factors. Finally the most important conclusion that can be drawn from these results is that lexical change and typological change appear to function in two diametrically different ways and that even though a language diverges typologically it still can undergo the same lexical changes as its relatives showing that typological change and lexical change are two independent processes.

Bibliography

- ADGER, David (2010). Gaelic Morphology, in MACLEOD, Michelle & WATSON, Moray (ed.): 283-303.
- AUWERA, Johan van der & KORTMANN, Bernd (ed.) (2011). *The Languages and Linguistics of Europe: A Comprehensive Guide*. Berlin: de Gruyter Mouton
- AUWERA, Johan van der (2011). Standard Average European, in AUWERA, Johan van der & KORTMANN, Bernd (ed.): 291-306.
- AXEL, Katrin (2007). *Studies on Old High German Syntax: Left sentence periphery, verb placement and verb-second*. Amsterdam: John Benjamins
- BACHMANN, Albert (1960). *Mittelhochdeutsches Lesebuch*. 16 ed. Zürich: Beer & Cie.
- BARÐDAL, Jóhanna, JÖRGENSEN, Nils, LARSEN, Gorm, MARTINUSSEN, Bente (1997). *Nordiska: Våra språk förr och nu*. Lund: Studentlitteratur
- BAUR, Arthur (2002). *Schwyzertüütsch: Praktisch Sprachlehre des Schweizerdeutschen*. Winterthur: Gemsberg-Verlag
- BJELLERUP, Sven (1990). *Portugisisk språklära*. 3rd ed. Lund: Studentlitteratur
- BLAŽEK, Václav (2009). "On the position of Gaulish within Celtic from the point of view of glottochronology", *Indogermanische Forschungen*. 114. Berlin: de Gruyter
- BORCHLING, Conrad, CORDES, Gerhard, LASCH, Agathe, MÖHN, Dieter (1928-) *Mittelniederdeutsches Handwörterbuch*. Neumünster: Wachholtz Verlag
- BOUCKAERT, Remco, LEMEY, Philippe, DUNN, Michael, GREENHILL, Simon J., ALEKSEYENKO, Alexander V., DRUMMOND, Alexei J., GRAY, Russell D., SUCHARD, Marc A., ATKINSON, Quentin D. (2012). Mapping the Origins and Expansion of the Indo-European Language Family. *Science* 319: 588.
- BRAUNE, Wilhelm & MITZKA, Walther (1967). *Althochdeutsche Grammatik*. 12th ed. Tübingen: Max Niemeyer Verlag
- BREMMER, Rolf H. Jr. & QUAK, Arend (1992). *Zur Phonologie und Morphologie des Altniederländischen*. Odense: Odense University Press
- BREMMER, Rolf H. Jr. (2009). *An Introduction to Old Frisian: History, Grammar, Reader, Glossary*. Amsterdam: John Benjamins
- BRINTON, Laurel J. (2000). *The Structure of Modern English: A Linguistic Introduction*. Amsterdam: John Benjamins
- BRODERICK, George (1984). *A Handbook of Late Spoken Manx: Vol. 1 Grammar and Texts*. Tübingen: Max Niemeyer Verlag

- BRUYNE, Jacques de & POUNTAIN, Christopher J. (1995). *A Comprehensive Spanish Grammar*. Oxford: Blackwell Pub.
- BUCK, Carl Darling (1974[1904]). *A Grammar of Oscan and Umbrian*. Hildesheim: Georg Olms Verlag
- BURIDANT, Claude (2000). *Grammaire nouvelle de l'ancien français*. Paris: Sedes
- CAMPBELL, Lyle & POSER, William J. (2008). *Language Classification: History and Method*. Cambridge: Cambridge University Press
- CAMPBELL, Lyle (2011[2004]). *Historical Linguistics: An Introduction*. 2nd ed. Edinburgh: Edinburgh University Press
- CARLING, Gerd, ERIKSEN, Love, GOERGENSEN, Anne, HOLMER, Arthur, WEIJER, Joost van de (to appear 2013). The dispersal of linguistic and cultural features and their mutual relationship: an Arawakan case study. *Language Dynamics and Change*.
- DEVINE, A. M., STEPHENS, Laurence D. (2006). *Latin Word Order*. New York: Oxford University Press
- DI GIROLAMO, Costanzo & LEE, Charmaine (1998). *Avviamento alla filologia provenzale*. Rome: Carocci
- DUNN, Michael, GREENHILL, Simon J., LEVINSON, Stephen C., GRAY, Russell D. (2011). Evolved structure of languages shows lineage-specific trends in word-order universals. *Nature* 473: 79-82.
- ENGEL, Ulrich (1988). *Deutsche Grammatik*. Heidelberg: Julius Groos Verlag
- FABRY, Marcel (1951). *Grammaire pratique du wallon liégeois*. Liège: Paul Gothier
- FALILEYEV, Alexander (2000). *Etymological Glossary of Old Welsh*. Tübingen: Max Niemeyer Verlag
- FALILEYEV, Alexandre I. (2008). *Le Vieux-Gallois*. Potsdam: Universitätsverlag Potsdam
- FAVEREAU, Francis (1997). *Grammaire du breton contemporain*. Morlaix: Skol Vreizh
- FEIST, Sigmund (1939). *Vergleichendes Wörterbuch der gotischen Sprache mit Einschluss des kringotischen und sonstiger zerstreuter Überreste des gotischen*. 3rd ed. Leiden: Brill
- FRANCK, Johannes (1910). *Mittelniederländische Grammatik*. 2nd ed. Leipzig
- GALLÉE, Johan Hendrik (1993). *Altsächsische Grammatik*. Tübingen: Max Niemeyer Verlag
- GREENHILL, Simon J., ATKINSON, Quentin D., MEADE, A., GRAY, Russell D. (2010). The shape and tempo of language evolution. *Proceedings of the Royal Society B* 277: 2443-2450.
- HAIMAN, John (1997). Rhaeto-Romance, in HARRIS, Martin & VINCENT, Nigel (ed.): 351-390.

- HARBERT, Wayne (2007). *The Germanic Languages*. Cambridge: Cambridge University Press
- HARRIS, Martin & VINCENT, Nigel (ed.) (1997). *The Romance Languages*. London: Routledge
- HARRIS, Martin (1997). The Romance Languages in HARRIS, Martin & VINCENT, Nigel (ed.): 1-25.
- HAUGAN, Jens (2000). *Old Norse Word Order and Information Structure*. Diss. Trondheim: Norwegian University of Science and Technology
- HELLAND, Hans Petter (2006). *Ny fransk grammatikk: morfologi, syntaks og semantikk*. Oslo: Universitetsforlaget
- HOLTHAUSEN, Ferdinand (1925). *Altfriesisches Wörterbuch*. Heidelberg: Winter
- HOLTHAUSEN, Ferdinand (1934). *Gotisches etymologisches Wörterbuch: mit Einschluss der Eigennamen und der gotischen Lehnwörter im Romanischen*. Heidelberg: Winter
- HOROBIN, Simon & SMITH, Jeremy (2002). *An Introduction to Middle English*. Edinburgh: Edinburgh University Press
- JOHNSON, Keith (2008). *Quantitative Methods in Linguistics*. Malden, MA: Blackwell Pub.
- JOKINEN, Ulla & SIHVONEN-HAUTECCÆR, Päivi (ed.) (1988). *Approches du moyen français*. Jyväskylä: University of Jyväskylä
- KABATEK, Johannes & PUSCH, Claus D (2011). The Romance languages, in AUWERA, Johan van der & KORTMANN, Bernd (ed.): 69-96.
- KAHL, Heinrich. & THIES, Heinrich. (2002). *Plattdeutsches Wörterbuch: plattdeutsch-hochdeutsch ; hochdeutsch-plattdeutsch ; plattdeutsche Rechtschreibung*. 2nd ed. Neumünster: Wachholtz Verlag
- KING, Gareth (1993). *Modern Welsh: A Comprehensive Grammar*. London: Routledge
- LAMBERT, Pierre-Yves (2003). *La langue gauloise*. Paris: Editions Errance
- LEVY, Roger & DAUMÉ, Hal III (2011). Computational methods are invaluable for typology, but the models must match the questions. *Linguistic Typology* 15: 393-399.
- LEXER, Matthias (1930). *Mittelhochdeutsches Taschenwörterbuch*. 19th ed. Leipzig: Hirzel
- LINDOW, Wolfgang, MÖHN, Dieter, NIEBAUM, Hermann, STELLMACHER, Dieter, TAUBKEN, Hans, WIRRER, Jan (1998). *Niederdeutsche Grammatik*. Leer: Verlag Schuster
- LIVER, Ricarda (1999). *Rätoromanisch: Eine Einführung in das Bündnerromanische*. Tübingen: Gunter Narr Verlag
- LÜBBEN, August (1882). *Mittelniederdeutsche Grammatik*. Leipzig: T. O. Weigel
- Lyts frysk Wirdboek. 1, Frysk – Nederlânsk*. (1944) Bolsward/Boalsert: Osinga

- MACLEOD, Michelle & WATSON, Moray (ed.) (2010). *The Edinburgh Companion to the Gaelic Language*. Edinburgh: Edinburgh University Press
- MALLINSON, Graham (1997). Rumanian, in HARRIS, Martin & VINCENT, Nigel (ed.): 391-419.
- MAYRHOFER, Manfred (1980). *Zur Gestaltung des etymologischen Wörterbuches einer "Grosscorpus-Sprache"*. Wien: Österreichischen Akademie der Wissenschaften
- MERRILEES, Brian & SITARZ-FITZPATRICK, Beata (ed.) (1993). *Liber Donati: A Fifteenth-Century Manual of French*. London: Anglo-Norman Text Society
- MITCHELL, Bruce & ROBINSON, Fred C. (2007). *A Guide to Old English*. 7th ed. Oxford: Blackwell Pub.
- McCONE, Kim (2005). *A First Old Irish Grammar and Reader: Including an Introduction to Middle Irish*. Maynooth: Department of Old and Middle Irish, National University of Ireland
- NILSSON, Torbjörn & SVENSSON, Patrik (1997). *Gotiska: Grammatik, text och ordförklaringar*. Lund: Studentlitteratur
- Ó SIADHAIL, Mícheál (1989). *Modern Irish: Grammatical structure and dialectal variation*. Cambridge: Cambridge University Press
- PATOTA, Giuseppe (2007). *Nuovi lineamenti di grammatica storica dell'italiano*. Bologna: il Mulino.
- PITTAU, Massimo (1991). *Grammatica della lingua sarda: Varietà Logudorese*. Sassari: Carlo Delfino
- POERCK, G. de & MOURIN, L. (1961) *Introduction a la morphologie comparée des langues romanes. Tome I: Ancien portugais et ancien castillan*. 1st ed. Bruges: De Tempel
- POPKEMA, Jan (2006). *Grammatica Fries*. Utrecht: Prisma
- RINGE, Don, WARNOW, Tandy, TAYLOR, Ann (2001). Indo-European and Computational Cladistics. *Transactions of the Philological Society* 100:1: 59-129.
- RONAN, Patricia (2011). The Celtic languages, in AUWERA, Johan van der & KORTMANN, Bernd (ed.): 31-46.
- RUHLEN, Merritt (1991). *A guide to the world's languages*. Stanford, CA: Stanford University Press
- SCHRIJVER, Peter (2011). Middle and Early Modern Breton, in TERNES, Elmar (ed.): 359-430.
- SCHUMACHER, Stefan (2011). Mittel- und Frühneukymrisch, in TERNES, Elmar (ed.): 85-236.

- SCHÜTZEICHEL, Rudolf (1995). *Althochdeutsches Wörterbuch*. 5th ed. Tübingen: Max Niemeyer Verlag
- SHETTER, William Z. & HAM, Esther (2007). *Dutch: An Essential Grammar*. 9th ed. Abingdon: Routledge
- TERNES, Elmar (ed.) (2011). *Brythonic Celtic – Britannisches Keltisch: From Medieval British to Modern Breton*. Bremen: Hempen Verlag
- THRÁINSSON, Höskuldur, JACOBSEN, Jógvan í Lon, PETERSEN, Hjalmar P., SVABO HANSEN, Zakaris (2012). *Faroese: An Overview and Reference Grammar*. Tórshavn: Fróðskapur – Faroe University Press
- THURNEYSSEN, Rudolf (1946). *A Grammar of Old Irish*. Dublin: The Dublin Institute for Advanced Studies
- VERDAM, Jacob & EBBINGE WUBBER, C. H. (1949). *Middelnederlandsch handwoordenboek*. 's-Gravenhage: Nijhoff
- VINCENT, Nigel (1997). Italian, in HARRIS, Martin & VINCENT, Nigel (ed.): 279-313.
- VINCENT, Nigel (1997). Latin, in HARRIS, Martin & VINCENT, Nigel (ed.): 26-78.
- VRIES, Jan de (1977). *Altnordisches etymologisches Wörterbuch*. 3rd ed. Leiden: Brill
- WHEELER, Max W. (1997). Catalan, in HARRIS, Martin & VINCENT, Nigel (ed.): 170-208.
- WHEELER, Max W. (1997). Occitan, in HARRIS, Martin & VINCENT, Nigel (ed.): 246-278.
- WIDMARK, Gun (2001). *Det språk som blev vårt. Ursprung och utveckling i svenskan*. Uppsala: Gustav Adolfs akademien
- WIESE, Berthold (1928). *Altitalienisches Elementarbuch*. 2nd ed. Heidelberg: Carl Winter's Universitätsbuchhandlung
- WILLIAMS, Nicholas J. A. (2011). Middle and Late Cornish, in TERNES, Elmar (ed.): 237-358.
- ÅKERBERG, Bengt & NYSTRÖM, Gunnar (2012). *Älvdalsk grammatik*. Spånga: Bengt Åkerberg

Electronic Dictionaries and Corpora

Bokmålsordboka/Nynorskordboka.

Oslo: University of Oslo & Språkrådet:

<http://www.nob-ordbok.uio.no/perl/ordbok.cgi?> (2013-04-08 to 2013-04-29)

BOUTKAN, Dirk & SIEBINGA, Sjoerd Michiel (2013) *Old Frisian Etymological Dictionary*. Indo-European Etymological Dictionaries Online. Edited by

Alexander Lubotsky. Brill, 2013

Brill Online:

<http://iedo.brillonline.nl.ludwig.lub.lu.se/dictionaries/content/frisian/index.html>
(2013-03-06 to 2013-04-26)

Cantar de mio Cid.

Austin: University of Texas:

<http://www.laits.utexas.edu/cid/> (2013-03-20 to 2013-04-16)

Den Danske Ordbog.

Copenhagen: Det Danske Sprog- og Litteraturselskab:

<http://ordnet.dk/ddo> (2013-04-29)

Dictionnaire du Moyen Français.

Nancy: Analyse et traitement informatique de la langue française (ATILF):

<http://www.atilf.fr/dmf/> (2013-04-26)

Dictionnaire Électronique de Chrétien de Troyes.

Nancy: Analyse et traitement informatique de la langue française (ATILF):

<http://www.atilf.fr/dect/> (2013-04-26)

Fornsvensk lexikalisk databas.

Göteborg: Språkbanken:

<http://spraakbanken.gu.se/fsvldb/> (2013-03-08 to 2013-03-12)

Føroysk orðabók.

Orðabókagrunnurin:

<http://www.sleipnir.fo/obg/fob/fob.php> (2013-04-15 to 2013-04-17)

Indo-European Lexical Cognacy Database.

Nijmegen: Max Planck Institute: <http://ielex.mpi.nl/>

(2013-02-04 to 2013-04-26)

The Intercontinental Dictionary Series.

Leipzig: Max Planck Institute for Evolutionary Anthropology:

<http://lingweb.eva.mpg.de/ids/> (2013-03-06 to 2013-04-26)

ISLEX.

Reykjavík: Stofnun Árna Magnússonar í íslenskum fræðum:

<http://www.islex.hi.is/> (2013-03-11 to 2013-04-29)

KROONEN, Guus (2013) *Etymological Dictionary of Proto-Germanic*. Indo-European Etymological Dictionaries Online. Edited by Alexander Lubotsky. Brill, 2013.

Brill Online:

<http://iedo.brillonline.nl.ludwig.lub.lu.se/dictionaries/content/proto-germanic/index.html> (2013-03-06 to 2013-04-26)

MATASOVIĆ, Ranko (2013). *Etymological Dictionary of Proto-Celtic*. Indo-European Etymological Dictionaries Online. Edited by Alexander Lubotsky. Brill, 2013.

Brill Online:

<http://iedo.brillonline.nl.ludwig.lub.lu.se/dictionaries/content/proto-celtic/index.html> (2013-03-06 to 2013-04-26)

The Princeton Dante Project.

Princeton: Princeton University:

<http://etcweb.princeton.edu/dante/index.html> (2013-03-14 to 2013-04-16)

Svenska Akademiens ordbok.

Lund: Svenska Akademiens ordboksredaktion:

<http://g3.spraakdata.gu.se/saob/> (2013-03-08 to 2013-03-12)

VAAN, Michiel de (2013). *Etymological Dictionary of Latin and the other Italic Languages.*

Indo-European Etymological Dictionaries Online. Edited by Alexander

Lubotsky. Brill, 2013. Brill Online:

<http://iedo.brillonline.nl.ludwig.lub.lu.se/dictionaries/content/latin/index.html>
(2013-03-06 to 2013-04-26)

Appendix A: The Swadesh 100-list

1. I	26. root	51. breast	76. rain
2. thou	27. bark	52. heart	77. stone
3. we	28. skin	53. liver	78. sand
4. this	29. meat	54. drink (verb)	79. earth
5. that	30. blood	55. eat	80. cloud
6. who	31. bone	56. bite (verb)	81. smoke
7. what	32. fat	57. see	82. fire
8. not	33. egg	58. hear	83. ashes
9. all	34. horn	59. know	84. burn
10. many	35. tail	60. sleep (verb)	85. road
11. one	36. feather	61. die	86. mountain
12. two	37. hair	62. kill	87. red
13. big	38. head	63. swim	88. green
14. long	39. ear	64. fly (verb)	89. yellow
15. small	40. eye	65. walk	90. white
16. woman	41. nose	66. come	91. black
17. man	42. mouth	67. lie	92. night
18. person	43. tooth	68. sit	93. warm
19. fish	44. tongue	69. stand	94. cold
20. bird	45. fingernail	70. give	95. full
21. dog	46. foot	71. say	96. new
22. louse	47. knee	72. sun	97. good
23. tree	48. hand	73. moon	98. round
24. seed	49. belly	74. star	99. dry
25. leaf	50. neck	75. water	100. name

Appendix B: The Etic Grid for the typological data

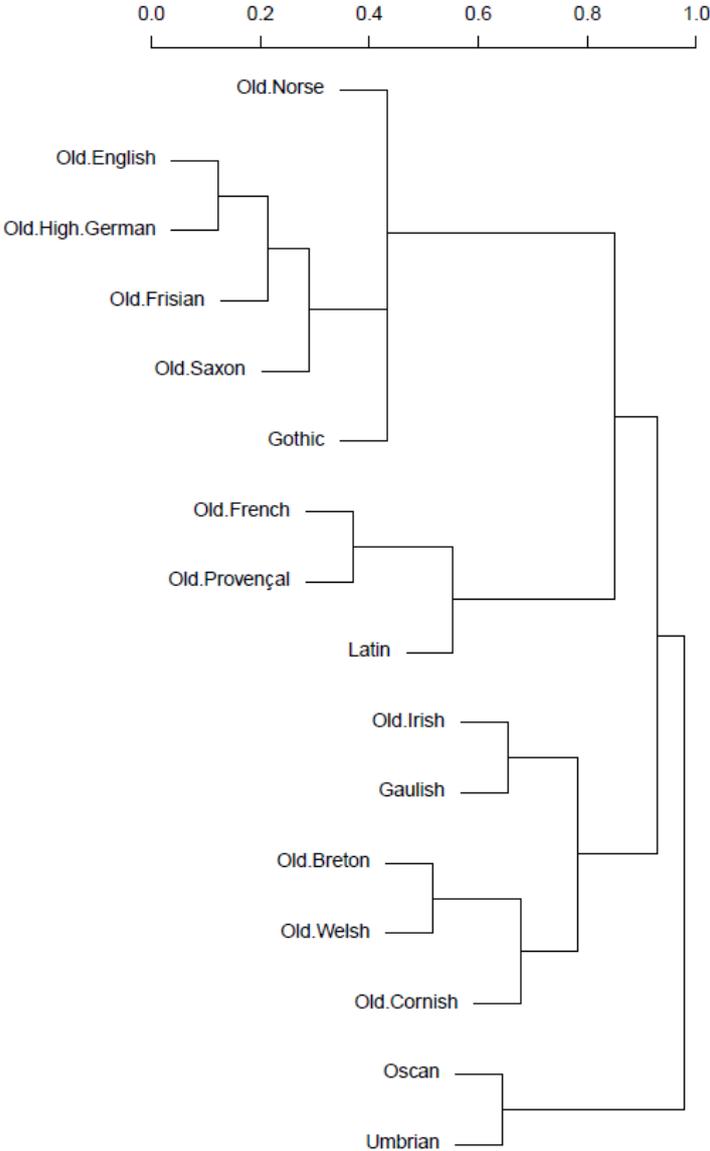
Word order	Main clause	SVO
		V2
		VSO
		SOV
	Subordinate clause	SVO
		V2
		SOV
		VSO
	Non-finite verb	VO
	Clitic pronouns	OV
VO		
Clitic pronouns infinite verbs	OV	
	VO	
Embedded clause	Embedded V2	
Nominal morphology	Nominal cases	2
		3
		4 or more
	Pronominal cases	2
		3
		4 or more
	Case marking	On nouns
		On articles
	Gender	Masculine/Feminine distinction
		Distinct neuter
Definitiveness marking	Suffix on the noun	
	Suffix on the adjective	
	Suffix on both noun and adjective	
Gender agreement	Predicative adjectives	
Preposition agreement	With pronouns	
Verbal morphology	Present personal inflection	1 st person singular
		2 nd person singular
		3 rd person singular
		1 st person plural
		2 nd person plural
		3 rd person plural
	Present syncretism	All singular
		All plural
		1 st singular & 3 rd singular
		1 st singular & 3 rd plural
		2 nd singular & 3 rd singular
		2 nd singular & 2 nd plural
		3 rd singular & 2 nd plural
		3 rd singular & 3 rd plural
		1 st plural & 3 rd plural
		2 nd plural & 3 rd plural
	Past personal inflection	1 st person singular
		2 nd person singular
		3 rd person singular
		1 st person plural
		2 nd person plural
		3 rd person plural
	Past syncretism	All singular
		All plural
		1 st singular & 3 rd singular
		2 nd singular & 2 nd plural
		3 rd singular & 3 rd plural
1 st plural & 3 rd plural		
Continuous present	Present	
	Progressive present	

Appendix C: Sources for all languages

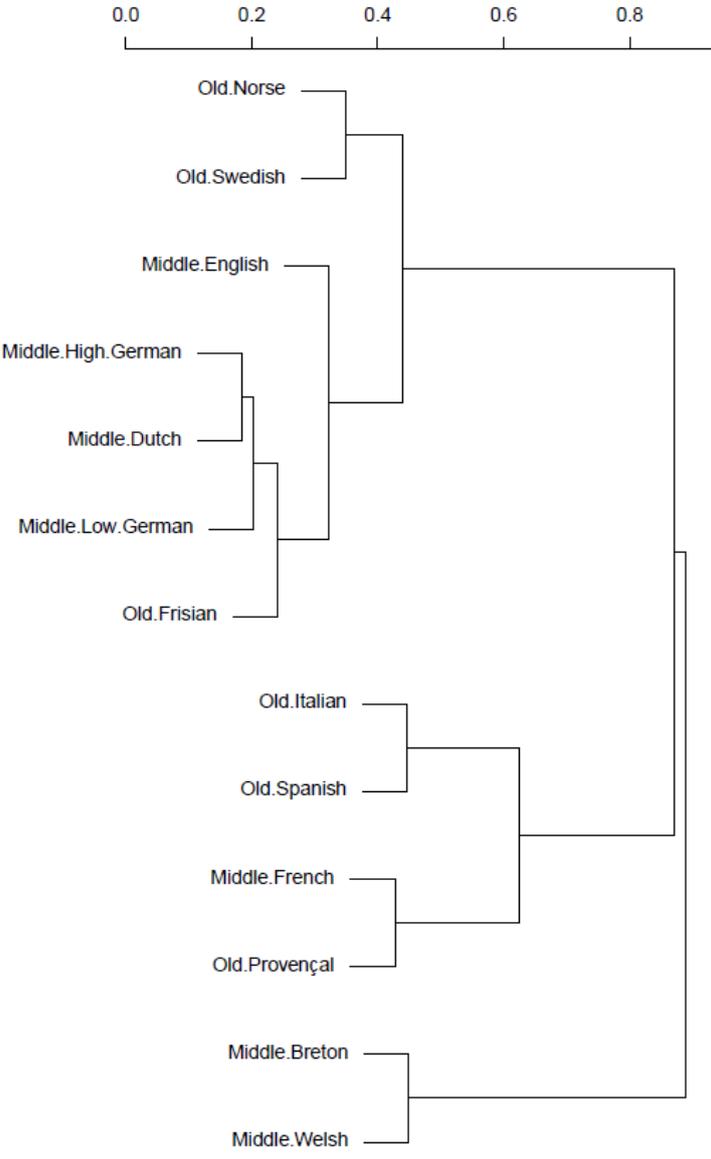
Language	Lexical data	Typological data
Breton	IELex, Blažek (2009), Matasović (2013)	Favereau (1997)
Catalan	IELex, IDS	Wheeler (1997)
Cornish	IELex, Blažek (2009), Matasović (2013)	Williams (2011)
Danish	IELex, Den Danske Ordborg	Barðdal et al. (1997)
Dutch	IELex, Kroonen (2013)	Shetter & Ham (2007)
Elfdalian	IELex, Kroonen (2013)	Åkerberg & Nyström (2012)
English	IELex, Kroonen (2013)	Brinton (2000)
Faroese	IELex, Kroonen (2013), Føroysk orðabók	Thráinsson et al. (2012)
French	IELex	Helland (2006)
Frisian	IELex, Kroonen (2013), Lyts frysk Wirdboek	Popkema (2006)
Friulian	IELex	Haiman (1997)
Gaulish	Blažek (2009), Matasović (2013)	Lambert (2003)
German	IELex	Engel (1988)
Gothic	IELex, Matasović (2013), Holthausen (1934), Feist (1939)	Nilsson & Svensson (1997)
Icelandic	IELex, ISLEX	Barðdal et al. (1997)
Irish	IELex, IDS, Blažek (2009)	Ó Siadhail (1989)
Italian	IELex	Vincent (1997), Patota (2007)
Ladin	IELex	Haiman (1997)
Latin	IELex, de Vaan (2013)	Vincent (1997), Devine & Stephens (2006)
Low German	Kahl & Thies (2002)	Lindow et al. (1998)
Manx	Blažek (2009)	Broderick (1984)
Middle Breton	Blažek (2009), Matasović (2013)	Schrijver (2011)
Middle Cornish	-	Williams (2011)
Middle Dutch	Kroonen (2013), Boutkan & Siebinga (2013), Verdam & Ebbinga Wubber (1949)	Bremmer & Quak (1992), Franck (1910)
Middle English	IDS, Kroonen (2013)	Horobin & Smith (2002)
Middle French	Dictionnaire du Moyen Français	Jokinen & Sihvonen (1988), Merrilees & Sitarz-Fitzpatrick (1993)
Middle High German	Kroonen (2013), Boutkan & Siebinga (2013), Lexer (1930)	Bachmann (1960)
Middle Irish	-	McCone (2005)
Middle Low German	Kroonen (2013), Boutkan & Siebinga (2013), Borchling et al. (1928-)	Lübber (1882)
Middle Welsh	Blažek (2009), Matasović (2013), Ternes (2011)	Schumacher (2011)
Norwegian (Bokmål)	IELex, Kroonen (2013), Bokmålsordboka	Barðdal et al. (1997)
Norwegian (Nynorsk)	Nynorskordboka	Barðdal et al. (1997)
Old Breton	Blažek (2009), Matasović (2013), Ternes (2011)	-
Old Cornish	Blažek (2009), Matasović (2013)	-
Old Dutch	-	Bremmer & Quak (1992)
Old English	IELex, Kroonen (2013), Boutkan & Siebinga (2013)	Mitchell & Robinson (2007)
Old French	Dictionnaire Électronique de Chrétien de Troyes, Buridant (2000)	Buridant (2000)
Old Frisian	Kroonen (2013), Boutkan & Siebinga (2013), Holthausen (1925)	Bremmer (2009)
Old High German	IELex, Kroonen (2013), Boutkan & Siebinga (2013), Schützeichel (1995)	Braune & Mitzka (1967), Axel (2007)
Old Irish	IELex, Blažek (2009), Matasović (2013)	McCone (2005), Thurneysen (1946)
Old Italian	The Princeton Dante Project	Wiese (1928)
Old Norse	IELex, Kroonen (2013), de Vries (1977)	Haugan (2000), Barðdal et al. (1997)

Old Provençal	Di Girolamo & Lee (1998)	Di Girolamo & Lee (1998)
Old Saxon	Kroonen (2013), Boutkan & Siebinga (2013)	Gallée (1993)
Old Spanish	Cantar de mio Cid	Poerck & Mourin (1961)
Old Swedish	Kroonen (2013), Svenska Akademiens ordbok, Fornsvensk lexikalisk databas	Widmark (2001)
Old Welsh	Blažek (2009), Matasović (2013), Falileyev (2000 & 2008)	Falileyev (2008)
Oscan	IELex, de Vaan (2013)	Buck (1904)
Portuguese	IELex, IDS	Bjellerup (1990)
Provençal	IELex	Wheeler (1997)
Romanian	IELex	Mallinson (1997)
Romansh	IELex	Liver (1999), Haiman (1997)
Sardinian (Logudorese)	IELex, Pittau (1991)	Pittau (1991)
Scottish Gaelic	IELex, Blažek (2009)	Adger (2010)
Spanish	IELex	de Bruyne & Pountain (1995)
Swedish	IELex	Barðdal et al. (1997)
Swiss German	IELex, Baur (2002)	Baur (2002)
Umbrian	IELex, de Vaan (2013)	Buck (1904)
Walloon	IELex	Fabry (1951)
Welsh	IDS, Blažek (2009), Matasović (2013)	King (1993)

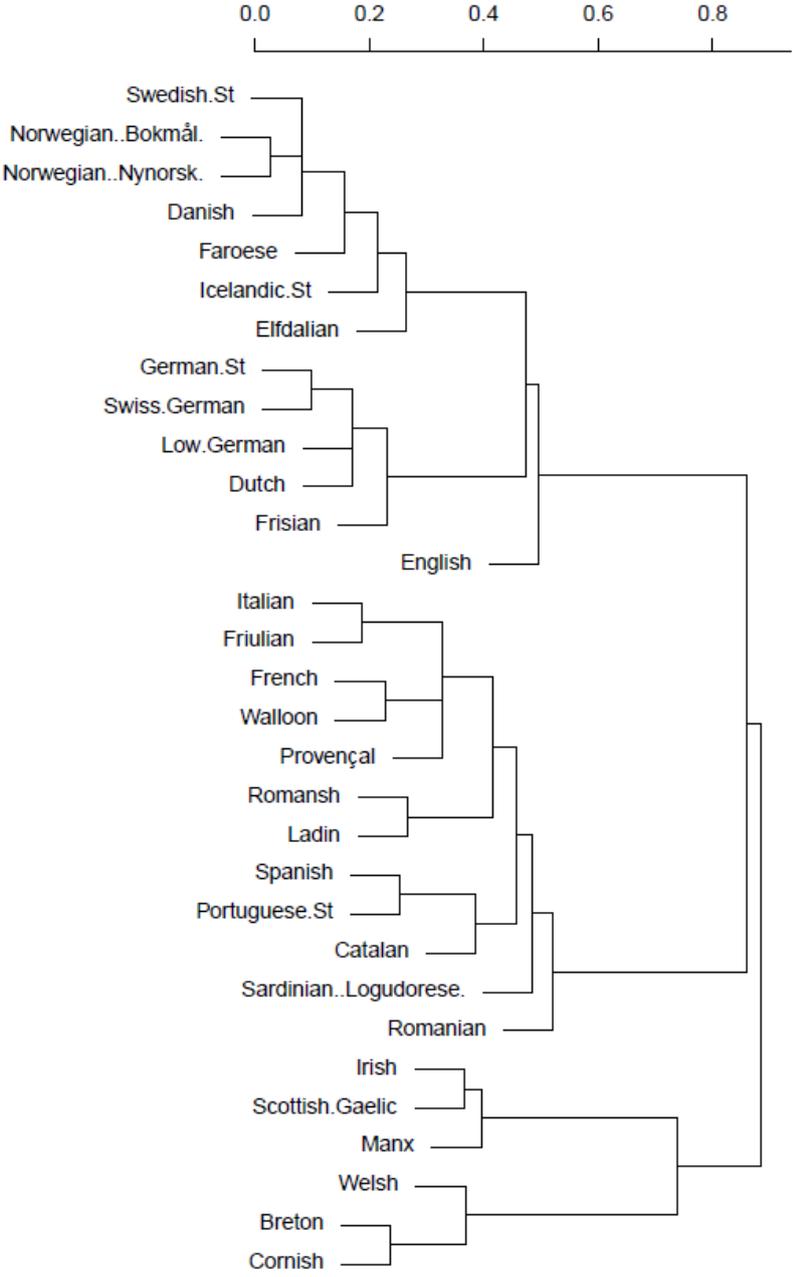
Appendix D: Lexical dendrogram (250 BC to 1000 AD)



Appendix E: Lexical dendrogram (1000 AD to 1500 AD)



Appendix F: Lexical dendrogram (1500 AD to 2000 AD)



Appendix G: The entire typological database

	Swedish St	Danish	Norwegian (Bokmål)	Norwegian (Nynorsk)	Icelandic St	Faroese	Elfdalian	German St	Swiss German	Low German	Dutch	English
WO Main SVO												1
WO Main V2	1	1	1	1	1	1	1	1	1	1	1	
WO Main VSO												
WO Main SOV												
WO Sub SVO	1	1	1	1		1						1
WO Sub V2					1	1	1					
WO Sub SOV								1	1	1	1	
WO Sub VSO												
Non-finite VO	1	1	1	1			1					1
Non-finite OV								1	1	1	1	
Embedded V2	1	1	1	1	1	1	1					
2 NP Cases									1	1		
3 NP Cases							1					
4+ NP Cases					1	1		1				
2 P Cases	1	1	1	1						1	1	1
3 P Cases									1			
4+ P Cases					1	1	1	1				
N case m.					1	1						
Art case m.								1	1	1		
M=F			1	1	1	1	1	1	1	1		
N≠M/F	1	1	1	1	1	1	1	1	1	1	1	
1 PS Pres					1	1		1	1	1	1	
2 PS Pres								1		1		
3 PS Pres									1	1		1
1 PP Pres					1		1					
2 PP Pres					1		1					
3 PP Pres					1		1					
2 PS + 3 PS Pres					1	1					1	
3 PS + 2 PP Pres								1				
1+2+3 PS Pres							1					
1+2+3 PP Pres						1			1	1	1	
1 PP + 3 PP Pres								1				
1 PS + 3 PP Pres												
3 PS + 3 PP Pres												
2 PP + 3 PP Pres												
1 PS + 3 PS Pres												
2 PS + 2 PP Pres												
1 PS Past												
2 PS Past					1			1		1		
3 PS Past												
1 PP Past					1		1					
2 PP Past					1		1	1				
3 PP Past					1		1					
1 PS + 3 PS Past					1			1		1		
1+2+3 PS Past						1					1	
1+2+3 PP Past						1				1	1	
1 PP + 3 PP Past								1				
2 PS + 2 PP Past												
3 PS + 3 PP Past												
Present	1	1	1	1	1	1	1	1	1	1	1	
Progressive present												1
N-def	1	1	1	1	1	1	1					
Attr-def	1	1	1	1	1	1	1	1	1	1		
Attr-def + N-def	1		1	1	1	1	1					
Pred Adj G	1	1	1	1	1	1	1					
Congr												
Prep Pron												
Congr												
Clitic pron OV												
Clitic pron VO									1		1	
Clitic pron OV inf											1	
Clitic pron VO inf												

NP cases = Noun phrase cases

PP cases = Pronominal cases

N case m = Case marking on nouns

Art case m = Case marking on articles

	Frisian	Old Norse	Middle English	Old English	Old High German	Middle High German	Old Frisian	Old Saxon	Middle Low German	Old Dutch	Middle Dutch
WO Main SVO			1	1							
WO Main V2	1	1		1	1	1	1	1	1		1
WO Main VSO											
WO Main SOV						1					
WO Sub SVO		1	1	1							
WO Sub V2		1									
WO Sub SOV	1		1	1	1	1	1	1	1		1
WO Sub VSO		1									
Non-finite VO		1		1							
Non-finite OV	1				1	1	1	1	1		1
Embedded V2		1									
2 NP Cases											
3 NP Cases											
4+ NP Cases		1		1	1	1	1	1	1	1	1
2 P Cases	1										1
3 P Cases			1							1	
4+ P Cases		1		1	1	1	1	1	1		1
N case m.		1		1	1	1	1	1	1	1	1
Art case m.				1	1	1	1	1	1	1	1
M≠F		1		1	1	1	1	1	1	1	1
N≠M/F	1	1		1	1	1	1	1	1	1	1
1 PS Pres	1	1	1	1	1	1	1	1	1	1	1
2 PS Pres	1		1	1	1	1	1	1	1	1	1
3 PS Pres	1		1	1	1	1	1	1	1		
1 PP Pres		1			1	1					
2 PP Pres		1			1	1					
3 PP Pres					1	1				1	
2 PS + 3 PS Pres		1									
3 PS + 2 PP Pres										1	1
1+2+3 PS Pres											
1+2+3 PP Pres	1		1	1			1	1	1		
1 PP + 3 PP Pres											1
1 PS + 3 PP Pres											
3 PS + 3 PP Pres											
2 PP + 3 PP Pres											
1 PS + 3 PS Pres											
2 PS + 2 PP Pres											
1 PS Past		1		1			1				
2 PS Past	1	1	1	1	1	1	1	1	1	1	1
3 PS Past		1		1			1				
1 PP Past		1			1						
2 PP Past		1			1	1					1
3 PP Past		1			1					1	
1 PS + 3 PS Past	1				1	1		1	1	1	1
1+2+3 PS Past											
1+2+3 PP Past	1			1			1	1	1		
1 PP + 3 PP Past						1				1	1
2 PS + 2 PP Past											
3 PS + 3 PP Past											
Present	1	1	1	1	1	1	1	1	1	1	1
Progressive present											
N-def		1									
Attr-def	1	1	1	1	1	1	1	1	1		1
Attr-def + N-def											
Pred Adj G		1				1					
Congr											
Prep Pron Congr											
Clitic pron OV											
Clitic pron VO							1				
Clitic pron OV inf											
Clitic pron VO inf											

M≠F = Differentiation between masculine and feminine gender
N≠M/F = Differentiation between neuter and masculine/feminine
PS = Person singular
PP = Person plural
Pres = present
N-def = Definitiveness is marked as a suffix on the noun
Attr-def = Attributive adjectives are marked for definitiveness

	Old Swedish	Gothic	Irish	Scottish-Gaelic	Manx	Old Irish	Middle Irish	Welsh	Breton	Cornish	Middle Cornish	Middle Breton	Old Welsh
WO Main SVO										1	1	1	
WO Main V2	1								1				
WO Main VSO			1	1	1	1	1	1					1
WO Main SOV		1											
WO Sub SVO	1												
WO Sub V2	1												
WO Sub SOV										1	1		
WO Sub VSO	1		1	1		1		1	1				1
Non-finite VO	1												
Non-finite OV	1	1											
Embedded V2													
2 NP Cases					1								
3 NP Cases			1										
4+ NP Cases	1	1		1		1	1						
2 P Cases												1	
3 P Cases													
4+ P Cases	1	1											
N case m.	1	1	1	1		1	1						
Art case m.	1	1	1	1		1							
M≠F	1	1	1	1		1	1	1	1	1	1	1	1
N≠M/F	1	1				1							
1 PS Pres		1	1			1	1		1	1	1	1	1
2 PS Pres		1				1	1		1	1	1	1	1
3 PS Pres		1				1	1		1	1	1	1	1
1 PP Pres	1	1	1			1	1		1	1	1	1	
2 PP Pres	1					1	1		1	1	1	1	
3 PP Pres	1	1				1	1		1	1	1	1	1
2 PS + 3 PS Pres			1										
3 PS + 2 PP Pres		1											
1+2+3 PS Pres	1												
1+2+3 PP Pres													
1 PP + 3 PP Pres													
1 PS + 3 PP Pres													
3 PS + 3 PP Pres													
2 PP + 3 PP Pres			1										
1 PS + 3 PS Pres													
2 PS + 2 PP Pres													
1 PS Past						1	1	1	1	1	1	1	
2 PS Past		1				1	1	1	1	1	1	1	
3 PS Past						1	1	1	1	1	1	1	1
1 PP Past	1	1	1			1	1		1	1	1	1	1
2 PP Past	1	1				1	1	1	1	1	1	1	1
3 PP Past	1	1				1	1		1	1	1	1	1
1 PS + 3 PS Past													
1+2+3 PS Past	1												
1+2+3 PP Past													
1 PP + 3 PP Past								1					
2 PS + 2 PP Past													
3 PS + 3 PP Past													
Present	1	1				1	1		1	1	1	1	1
Progressive present			1	1	1			1					
N-def	1												
Attr-def	1												
Attr-def + N-def	1												
Pred Adj G Congr	1					1							
Prep Pron Congr			1	1	1	1	1	1	1	1	1	1	1
Clitic pron OV													
Clitic pron VO						1				1	1		
Clitic pron OV inf													
Clitic pron VO inf													

Pred Adj G Congr = Predicative adjectives agree in gender

Prep Pron Congr = Prepositions and pronouns agree with each other

Clitic pron = Placement of clitic pronouns in verbal contexts

	Middle Welsh	Gaulish	Italian	French	Spanish	Portuguese St	Romanian	Catalan	Provençal	Walloon	Sardinian (Logudorese)
WO Main SVO		1	1	1	1	1	1	1	1	1	1
WO Main V2	1										
WO Main VSO											
WO Main SOV											
WO Sub SVO			1	1	1	1	1	1	1	1	1
WO Sub V2											
WO Sub SOV											
WO Sub VSO	1										
Non-finite VO			1	1			1	1	1		1
Non-finite OV											
Embedded V2											
2 NP Cases							1				
3 NP Cases											
4+ NP Cases		1									
2 P Cases	1				1						
3 P Cases			1	1		1	1	1	1	1	1
4+ P Cases											
N case m.		1					1				
Art case m.							1				
M≠F	1		1	1	1	1	1	1	1	1	1
N≠M/F							1				
1 PS Pres	1	1	1		1	1		1	1		1
2 PS Pres	1		1		1	1	1	1	1		1
3 PS Pres	1	1	1		1	1	1	1	1		1
1 PP Pres	1		1	1	1	1	1	1	1	1	1
2 PP Pres	1		1	1	1	1	1	1	1	1	1
3 PP Pres	1	1	1	1	1	1		1	1	1	1
2 PS + 3 PS Pres											
3 PS + 2 PP Pres											
1+2+3 PS Pres				1						1	
1+2+3 PP Pres											
1 PP + 3 PP Pres											
1 PS + 3 PP Pres							1				
3 PS + 3 PP Pres							1				
2 PP + 3 PP Pres											
1 PS + 3 PS Pres											
2 PS + 2 PP Pres											
1 PS Past	1		1								1
2 PS Past	1		1		1	1		1	1		1
3 PS Past	1	1	1								1
1 PP Past	1		1	1	1	1		1	1		1
2 PP Past	1		1	1	1	1		1	1		1
3 PP Past	1		1		1	1		1	1		1
1 PS + 3 PS Past					1	1		1	1		
1+2+3 PS Past										1	
1+2+3 PP Past										1	
1 PP + 3 PP Past											
2 PS + 2 PP Past											
3 PS + 3 PP Past											
Present	1	1	1	1	1	1	1	1	1	1	1
Progressive present			1		1	1					1
N-def							1				
Attr-def											
Attr-def + N-def											
Pred Adj G Congr			1	1	1	1	1	1	1		1
Prep Pron Congr	1										
Clitic pron OV			1	1	1		1	1	1	1	1
Clitic pron VO	1					1					
Clitic pron OV inf				1					1	1	1
Clitic pron VO inf			1		1	1		1			

	Romansh	Ladin	Friulian	Middle French	Old French	Old Italian	Old Spanish	Old Provençal	Latin	Oscan	Umbrian
WO Main SVO		1	1	1		1	1				
WO Main V2	1			1	1			1			
WO Main VSO											
WO Main SOV		1	1						1	1	1
WO Sub SVO	1			1		1	1	1			
WO Sub V2											
WO Sub SOV	1				1				1	1	1
WO Sub VSO											
Non-finite VO	1	1		1	1	1	1	1			
Non-finite OV									1		
Embedded V2											
2 NP Cases					1			1			
3 NP Cases											
4+ NP Cases									1	1	1
2 P Cases							1				
3 P Cases	1	1	1	1	1	1		1			
4+ P Cases									1		
N case m.					1			1	1	1	1
Art case m.					1			1	1	1	1
M≠F	1	1	1	1	1	1	1	1	1	1	1
N≠MF									1	1	1
1 PS Pres	1	1	1		1	1	1	1	1	1	1
2 PS Pres	1	1	1		1	1	1	1	1	1	1
3 PS Pres	1		1		1	1	1	1	1	1	1
1 PP Pres	1	1	1	1	1	1	1	1	1		
2 PP Pres	1	1	1		1	1	1	1	1		
3 PP Pres	1		1	1	1	1	1	1	1	1	1
2 PS + 3 PS Pres											
3 PS + 2 PP Pres											
1+2+3 PS Pres											
1+2+3 PP Pres											
1 PP + 3 PP Pres											
1 PS + 3 PP Pres											
3 PS + 3 PP Pres		1									
2 PP + 3 PP Pres											
1 PS + 3 PS Pres				1							
2 PS + 2 PP Pres				1							
1 PS Past	1	1	1	1	1				1	1	1
2 PS Past		1	1		1	1	1	1	1	1	1
3 PS Past	1		1	1	1				1	1	1
1 PP Past		1		1	1	1	1	1	1		
2 PP Past		1	1		1	1	1	1	1		
3 PP Past				1	1	1	1	1	1	1	1
1 PS + 3 PS Past						1	1	1			
1+2+3 PS Past											
1+2+3 PP Past											
1 PP + 3 PP Past	1		1								
2 PS + 2 PP Past	1			1							
3 PS + 3 PP Past		1									
Present	1	1	1	1	1	1	1	1	1	1	1
Progressive present	1										
N-def											
Attr-def											
Attr-def + N-def											
Pred Adj G		1	1	1	1			1	1		
Congr											
Prep Pron Congr											
Clitic pron OV		1	1	1	1	1	1	1			
Clitic pron VO	1										
Clitic pron OV inf		1		1							
Clitic pron VO inf			1		1	1					