Orthographic Typology

- a comparative study of the Latin orthographies of Europe

Filip Larsson

BA Thesis
General Linguistics
Autumn Semester 2011

Lund University
Center for Language and Literature
Supervisor: Gerd Carling

Abstract

This thesis aims to explain the variation that is found among the Latin orthographies of Europe. The main question is if it can be explained as genealogical, areal or social. The hypothesis presented in this thesis is that genealogical factors are the most important. Orthographies are relevant to study in their own right since they are autonomous from spoken languages. Since orthographies basically express the relation between phonemes and graphemes the study has been done as a comparative analysis by comparing the amount of shared combinations of phonemes and graphemes in 45 orthographies. These shared combinations constituted the basis for a tree model of the relation of the studied orthographies. The results of the tree model and the database showed that orthographical variation is not random and that genealogical factors were the most important but historical factors were also important. The tree model also showed that the variation is greater among vowels than among consonants. Another conclusion that was made was that political dominance is a relevant factor when new orthographies are created.

Key words: Orthography, grapheme, Latin alphabet, genealogical, areal

Acknowledgments

I would like to thank my supervisor Gerd Carling for her support and Joost van der Weijer for the help with the tree models.

Contents

1. Introduction

We in the literate part of the world are confronted by orthographies not just on a daily basis but almost constantly. All that is written and all that is being written, regardless of language and context, relates somehow to an orthography whether one chooses to follow it or not. Orthographies do not solely relate to their users as they relate just as much to other orthographies. This relationship between different orthographies is what this thesis aims to investigate and explain. I will exclusively focus on the orthographies of the Latin alphabet in a European context, as Europe is the birth place and has been the principal scene of diffusion and development of the Latin alphabet since medieval times. How do these Latin orthographies of Europe relate to each other and are the causes of these differences possible to generalise?

1.1. Orthographic units

Alphabetic orthographies mainly postulate the relationship between *phonemes* and *graphemes*. These two linguistic phenomena are the basic units of the Latin alphabet. Phonemes can't be found in the orthography itself as they constitute the inventory of phones that a specific language differentiates between, i.e. differences of meaning between using different phones. In a similar fashion we also find *allophones* which are different phonetic values of one phoneme, but I will come back to these later. Furthermore we have graphemes, which are the basic constituents of orthographies and the main focus of this study. Graphemes are units whose meaning is differentiated from other graphemes of the orthography, i.e. letters and combination of letters which differentiate one word from another. Orthographies create a relationship between the spoken and the written language which requires some sort of structural correspondence between the spoken and the written norm as the opposite would lead to significant difficulties when switching between the two norms (Vachek 1976, p. 136). This structural correspondence does not necessarily have to be based on the relationship between phonemes and graphemes though, but it is necessary that the codification of the spoken language is not too complex (Vachek 1976, p. 136), as orthographies do not just depend on the relationship between phonemes and graphemes. This relationship is however basically the fundamental principal of the Latin orthographies.

Graphemes can be split into two major groups, which are either called *free graphemes* and *complex symbols* (Rogers 2005, p. 11) or *simple* and *complex characters* (Sgall 1987, p. 8). I will use the terms simple and complex characters as these terms are more specific than the term symbol. Simple characters are units which are composed of one single letter, e.g. *a, b, c* etc. and complex characters are units which are composed of one letter with some sort of diacritic sign, as *á, ć, č* etc. Note that graphemes can be composed of combinations of several simple and/or complex characters.

Graphemes can be further split into subgroups where the most relevant groups for this study are *polygraphs, polyphones, homographs* and *homophones* (Rogers 2005, p. 16).[1] Polygraphs are graphemes which are composed of two or more letters, e.g. the English grapheme *th* which represents the phoneme /θ/. A special case of polygraphs are digraphs which the polygraph *th* is an example of since a digraph is a grapheme composed of two letters. N.B. that the combination of *t* and *h* is not a grapheme in a word like *foothill* since this is merely the graphemes *t* and *h* in a sequence. Polyphones on the other hand are two phonemes which are represented by one grapheme, as the grapheme *x* constitutes the two phonemes /ks/ and is thus a polyphone. Homographs are graphemes which represent more than one phoneme, i.e. the same letters or letter combinations are pronounced differently in different contexts, e.g. the *ow* in *cow* /kaʊ/ and *low* /loʊ/. Homophones are phonemes that are represented by more than one grapheme, i.e. that one sound is spelled in more than one way, e.g. the phonematic /baɪ/ is realised in English as *by*, *bye* and *buy*.

Returning to the allophones that were mentioned briefly in the beginning of this chapter we established that they are varieties of phonemes. This means that they are not differentiated in writing as graphemes mainly relate to phonemes. If allophones were differentiated orthographically the question whether they remain allophones would arise. The relationship between allophones and orthographies is not unproblematic but in an orthographic context I think it would be reasonable to regard them as a special case of homography, i.e. a grapheme that is pronounced as different phones. To claim that allophones are a case of homography might be a bit misleading but when it comes to orthographies I find it suitable, as the grapheme *n* expresses both the phones [n] as in *ban* /bæn/ and [ŋ] as in *bank* /bæŋk/. Even though both of these phones constitute the same phoneme, i.e. /n/, there is still a difference in

[1] N.B. that the terminology might vary.

pronunciation which means that *n* is somewhat homographic. The question is if it is relevant to speak of homography when it comes to sub-phonematic units such as allophones. Any further discussions regarding the relation between orthographies and allophones will not be continued though since this study does not regard allophones.

Apart from these orthographic units we also find allographs, varieties of a grapheme that do not change the meaning of the grapheme, e.g. the fact that a grapheme can be written as a capital letter, in italics or by hand, which creates graphic variation (Rogers 2005, p. 10). Since this does not affect the meaning of the grapheme it is not relevant for this study and therefore allographs will not be considered in this study.

It becomes obvious that the statement that orthographies express the relationship between phonemes and graphemes is not valid when it comes to non-alphabetic writing systems such as the Chinese system which is *morphographic*, which means that it expresses the relationship between *morphemes* and graphemes (Rogers 2005, p. 14). Morphographic writing systems are not present in Europe but there are morphographic tendencies in some European orthographies. Rogers claims that English in particular has a relatively high amount of morphographic constructions in its orthography, e.g. the presence of homophones like *by*, *bye* and *buy* which are pronounced the same way but they are different morphemes (Rogers 2005, p. 275). This is certainly a relevant orthographical aspect but since this study does not consider morphological aspects of orthographies I will not go further into morphography.

1.2. Research question

This thesis aims to answer the question whether the orthographic variation in Europe is mainly due to genealogical, areal or social factors.

Presupposing that orthographic variation is not utterly random we are encountered by the question what might cause it. I postulate three relevant factors, namely genealogical linkage, areal features and social linkage, where the social factor could be divided into historical factors and factors such as identity.[2] Another social factor that might be relevant is religion since it is the principal factor in the split between Latin, Cyrillic and Greek orthographies.

[2] Further discussions about the factors are found under the chapter about method.

1.3. Hypothesis

My hypothesis is that orthographic variation in Europe is mainly due to genealogical linkage which means that languages which are related tend to have similar orthographies and that the closer the languages are related the more similar the orthographies are. Even if I claim below in chapter 2 that written language is autonomous from spoken language I also think that despite the autonomous nature of written language it is strongly linked to the spoken language's genealogical relation to other languages. I base this upon two assumptions, namely that orthographies try to be as unique as possible and languages choose orthographies which are similar to the orthographies of related languages. The assumption that orthographies try to be as unique as possible is based on the fact that I could not find any identical orthographies in Europe except Bosnian, Croatian and Serbian, which used to constitute Serbo-Croatian.[3] The consequences are that specific orthographic constructions and differing usages of characters tend to spread between languages in the same family or subfamily which results in orthographic phenomena that are unique for a group of languages instead of just one language.

2. Background

This thesis is based on the view that written language and its norms are *autonomous* from spoken language (Sgall 1987, p. 4) and that the function of the written language is not identical to the function of the spoken language (Vachek 1976, p. 134). With this said it its important to point out that written language cannot be studied independently from spoken language as written language and spoken language are two parallel ways of expressing the same underlying linguistic system (Sgall 1987, p. 5). Spoken language and written language could be separated through the distinction of the spoken language being unmarked and the written language being marked (Sgall 1987, p. 4, Vachek 1976, p. 136), as written language is *relatively artificial* and needs to be taught through education while spoken language is possible to acquire without any form of education (Sgall 1987, p. 4-5, referring to Bazell 1956 and Nauclér 1983, Liberman 1992, p. 167). The acquisition of spoken language could even be

[3] Slovene is highly similar as well, but not identical since it lacks *ć* and *đ* (Daniels & Bright 1995). Please do note that Serbian is also written in the Cyrillic alphabet.

seen as something that without any extraordinary circumstances is more or less unpreventable (Liberman 1992, p. 167). Written language is furthermore a target for conscious interventions, something that hardly applies to spoken language (Sgall 1987, p. 5). This makes it interesting to study written languages and their concrete codification, i.e. orthographies, as orthographies often convey information regarding the history of the language and they are not seldom a way to consciously or unconsciously set languages apart from each other or the opposite to make languages look more similar.

Orthographies are usually divided into *shallow* and *deep* orthographies. Shallow orthographies are orthographies in which graphemes represent phonemes and in deep orthographies graphemes represent morphophonemes (Rogers 2005, p. 177). Morphophonemes are something in between morphemes and phonemes, e.g. the English past marker *–ed* which conveys a specific morphological aspect but has two different phonematic values, namely /d/ and /t/ (Rogers 2005, p. 284). Orthographies are not either shallow or deep though as it is a matter of gradation where we find the Finnish orthography amongst the shallowest and amongst the deepest we find the English orthography. The most extreme case of a shallow orthography would be what could be referred to as a phonematic orthography, i.e. an orthography where all phonemes are represented by one grapheme, but it does not seem to exist any orthography that could be classed as phonematic in the narrow sense (Rogers 2005, p. 13).

Finally I would like to introduce a new aspect of orthographic variation, namely *monographic* versus polygraphic. Monographic orthographies are orthographies which mainly express their graphemes with *monographs*, i.e. graphemes consisting of only one character. This is in contrast with polygraphic orthographies which prefer to create graphemes from polygraphs. Do remember though that the majority of graphemes in most orthographies are monographs, which leads to the fact that all orthographies are somewhat monographic. This makes polygraphic orthographies interesting to focus on as the question is not whether an orthography is monographic or not but simply how polygraphic it is instead. Why I choose to introduce this aspect is due to the lack of information conveyed by the categories shallow/deep regarding the actual form of the graphemes in orthographies.

3. Method

This study has been carried out as a comparative analysis of variables where the variables are the phonemes of the studied 45 Latin orthographies and the values of the variables are their graphemic representation. The languages studied (and the sources used) were:

Table 3.1: Table of the 45 orthographies and their sources.

Albanian (Daniels & Bright 1995)	**Azerbaijani** (Öztopçu et al. 1996)	**Basque** (Saltarelli 1988)
Bosnian (Mønnesland 2002)	**Breton** (Ball & Fife 1993)	**Catalan** (Hualde 1992)
Corsican (Giacomo-Marcellesi 1997)	**Croatian** (Daniels & Bright 1995, Mønnesland 2002)	**Czech** (Comrie & Corbett 1993, Daniels & Bright 1995)
Danish (Herslund 2002)	**Dutch** (Daniels & Bright 1995)	**English** (Brinton & Brinton 2010)
Estonian (Daniels & Bright 1995)	**Faroese** (Árnason 2011)	**Finnish** (Karlsson 2009)
French (Blanche-Benveniste & Yaguello 2003)	**High German** (http://rechtschreibrat.ids-mannheim.de/download/regeln2006.pdf)	**Hungarian** (Daniels & Bright 1995)
Icelandic (Árnason 2011)	**Irish** (Ball & Fife 1993)	**Italian** (Serianni 2006)
Latvian (Daniels & Bright 1995)	**Lithuanian** (Daniels & Bright 1995, Mathiassen 1996)	**Low German** (Kahl & Thies 2002, Möhn & Lindow 1998)
Lower Sorbian (Daniels & Bright 1995)	**Luxembourgish** (Newton 1996)	**Maltese** (Daniels & Bright 1995)
Northern Sami (Nickel 1994)	**Norwegian** (bokmål) (Daniels & Bright 1995)	**Occitan** (classical norm) (Nouvel 1975)
Polish (Comrie & Corbett 1993, Daniels & Bright 1995)	**Portuguese** (Daniels & Bright 1995, Bjellerup 1990)	**Rhaeto-Romance** (Romansh) (Liver 1999)
Romanian (Daniels & Bright 1995)	**Scottish Gaelic** (Ball & Fife 1993, Rogers 1995)	**Serbian**[4] (Daniels & Bright 1995, Mønnesland 2002)
Slovak (Comrie & Corbett 1993, Daniels & Bright 1995)	**Slovene** (Daniels & Bright 1995)	**Southern Sami** (Bergsland 1994)
Spanish (de Bruyne & Pountain 1995)	**Swedish** (Bruce 2010)	**Turkish** (Daniels & Bright 1995, Öztopçu et al. 1996)
Upper Sorbian (Daniels & Bright 1995)	**Welsh** (Daniels & Bright 1995)	**West Frisian** (Popkema 2006)

After determining all relevant phonemes in the studied languages I set up a table where the phonemes constituted one axis and the languages constituted the other axis. The values in this table were the graphemes which represented the phonemes in the specific orthographies. As phonemes could be expressed by more than one phoneme the values of certain phonemes were more than one. I systematically discarded graphemes only found in loanwords such as

[4] Serbian is usually written in the Cyrillic alphabet but since a Latin orthography exists and is in widespread use (Mønnesland 2002) I thought it was reasonable to include it in this study.

the digraph *ch* in the Swedish word *choklad* (chocolate) and the digraph *ti* as in the English word *nation*.[5] This I did to prevent misleading results as graphemes like the ones mentioned above are examples of orthographical borrowing which do not affect the orthography as such but if included orthographies with the same borrowings would appear to be more similar than they actually are. This thesis aims to study the basic language specific orthographic systems and as borrowed graphemes do not exist outside of loanwords these borrowed graphemes do not tell us anything about the basic orthographies, rather the contrary. Another problematic aspect of graphemes only found in loanwords is that I could not find material concerning this for the majority of the languages. If I would have included borrowed graphemes just in the languages where I had material, i.e. Swedish, English and German, the data would have been unbalanced and the results would have been less reliable so I decided to discard all graphemes found only in loanwords. Apart from just phonemes three other variables were added, namely if the orthographies express vowel length, nasality and stress, since these aspects are found among several of the orthographies. Tone could also have been a relevant variable but since none of the orthographies studied expressed tone this variable became redundant.

Apart from the phonemes I also added some variables which are differentiated in a large number of languages but are graphemic rather than phonemic. The most important of these variables were the variables for /k/ and /g/ before a front or a back vowel. As this is differentiated in all the studied Romance languages and English I considered them to be relevant variables.[6] A variable for a sequence of two /k/ was also added as this is relevant for the Germanic languages,[7] but this does not consider consonant length as the variable is just when the grapheme *k* is duplicated. I also merged some variables such as graphemic or phonemic distinctions only found in one language without its own IPA symbol such as the Icelandic voiceless alveolar nasal /n̥/ with its grapheme *hn* (Árnason 2011) so that *hn* became a graphemic value for voiced /n/ instead. This might not be the optimal way to handle these phonemes but if I should have included aspects such as these the material might have been too specific to the phonology of the studied languages, which might have lead to serious implications when comparing the orthographies.

[5] The digraph *ti* is found in several of the languages via loanwords from Latin but the pronunciation differs.
[6] Cf. Italian, *che/casa*, French *que/courir*, Spanish, *queso/casa*, English *cake/king* etc.
[7] Cf. Swedish *backe*, Danish *bakke*, German *Bäckerei*, English *back*, Dutch *bakkerij* etc.

The last greater change that was made to the variables was to create general variables for diphthongs, i.e. for example diphthongs with the form /Vɪ/ are generally expressed as grapheme x, where V stands for any vowel. For example Finnish has amongst others the diphthongs /aɪ/, /eɪ/ and /oɪ/ expressed as *ai*, *ei* and *oi* and therefore the general variable of /Vɪ/ is V*i*. At first there were variables for all diphthongs that were found in the specific languages, but since this created differences between orthographies that expressed diphthongs in the same way but had different inventories of diphthongs I decided to create these general variables instead. Otherwise it would create an unequal relationship between languages with many diphthongs and languages with few diphthongs even though they expressed them the same way. Similarly there were some problems with similar but not identical diphthongs such as /oɪ/ and /ɔɪ/ which became differentiated even though it might be misleading. This has surely been caused by the fact that the material was at times not accurate enough. After creating these general variables I also kept some specific variables for diphthongs to display certain marked diphthongs such as the Dutch *ij* [ɛɪ] and the High German *eu* and *äu* both pronounced [ɔʏ].

Thereafter, when all the material was collected, I started to code the material binarily, i.e. the material was transferred into a binary database where all phonemes with all the specific graphemic values constituted one axis and the languages the other. The value became 1 if the language had the certain combination of phoneme and grapheme and if it didn't the value was blank, e.g. English had the value 1 for the variables ʃ.SH and b.B but blank values for the variables ʃ.Š and b.V. After completing the binary database I split it into two new databases, so that there were the old database consisting of all phonemes, one database consisting of only consonants and one database consisting of vowels, diphthongs, vowel length marking, nasality marking and stress marking. This split was made due to the fact that there are more consonants than vowels and since vowels do not convey as much information as consonants their information value is lower. I decided to split them up because the vocalic variation is in general noticeably larger than the consonantal variation which in relation to the fact that vowels have a lower information value makes the risk of the vowels distorting the results quite significant. This problem becomes clear when we look at English, since English has several graphemes for not one but several phonemes, which sets English apart from the rest of

the group. If we instead just look at the consonants the relationship between English and other orthographies might give a dramatically different result. This is one of several reasons to why I decided to split consonants and vowels so that the result would become more nuanced.

Finally the data from the binary database was put into a computer program to generate a tree model known as a dendrogram over the relationship between the orthographies. This was made as an explorative cluster analysis which means that it looked for patterns in the data and that those patterns are the basis for the clusters. I also set up a map with the distribution of the various orthographical branches and a table showing the orthographies and their branches according to their genealogical relations.

Finally I would have wanted to include another historical aspect, namely the age of the orthographies. It would have been relevant to see if there are any correlations between the age of the orthography and its position in the tree model, but I was forced to disregard this aspect. This was due to two reasons, the first being the fact that I could not find information regarding the age of most of the orthographies which made any meaningful comparisons impossible and the second being the fact that the information I could find mainly dealt with the age of the current orthography, i.e. when the last changes had been done and not for how long the orthographies had been more or less in their present form.

3.1. Tree model analysis

I have analysed the branches of the tree model starting from five plausible factors which might influence orthographies: *Genealogical relationship*, *areal closeness*, *history*, *identity* and *religion*. I will explain how I have used them below.

Genealogical relationship is the first factor that I have looked for when I have analysed the tree model. It is perhaps the least problematic as it is most often clear in Europe if languages are related or not and to what extent. The problem is instead how to determine whether or not a branch could be classified as truly genealogical or how to detect if it coincides with another factor such as areal closeness. Even though there have been some problems I have tried to solve them by analysing every branch as individually as possible. I have had some general guidelines, which were that genealogical relation trumps historical connection as they

inevitably are intertwined. Genealogical relationship is trumped by areal closeness in the case that the languages are only distantly related and do not belong to a common group or subgroup.

Areal closeness is of course easiest to point out if unrelated or relatively unrelated languages in vicinity to each other belong to the same orthographic branch. The problem is of course how to determine how unrelated they need to be and how close languages have to be to be considered in each others vicinity. I have tried to answer these questions when it has been necessary but the general guidelines are that they ought to be at least from two different subfamilies, e.g. Germanic and Romance, and that they are spoken next to each other. But it should be pointed out that this is basically up to the given languages, as this is not universally applicable. Last but not least there are correlations between genealogical and areal factors and in those cases it has been important to convey the duality of some branches.

The main historical factor that I have taken into consideration was political dominance, i.e. if the language of the orthography has been dominated by another language through the government of a state or the political elite. The focus has been on states and what state a language has been spoken within, for how long and how recently this has been. I have disregarded political dominance that has not been present for the last thousand years and in most cases I have only looked at the last five hundred years. Explanations that are found earlier than this are likely to be irrelevant. As related languages might have been spoken in the same state I have judged genealogical links to be more important than links due to political dominance in the past. Similarly historical explanations have trumped areal explanations since the presence of a common state is a more relevant factor to why orthographies converge than that they are spoken in the same area. This is caused by the tendency of the dominating orthography to become the model for which the dominated orthographies relate to.

Identity is a relevant factor when speaking of orthographic similarities and dissimilarities, but it is also very vague. As I can’t define nor measure identity it is a problematic factor but furthermore I can’t imagine how I could find a branch that is solely caused by identity. This makes identity both utterly difficult to observe amongst the orthographies but just as much in the tree model which leaves me with no choice but to discard identity as a factor as I can’t work with it in a meaningful matter.

Religion is also a relevant factor in orthographic variation, especially between the Latin alphabet and other writing systems. Something that ought to be remembered though is that the situation in Europe actually isn't that simple, as Romanian, Albanian, Turkish, Bosnian, Azerbaijani are all languages that are predominantly spoken by either Orthodox Christians or Muslims. Furthermore one could argue that since most languages written in the Cyrillic alphabet in Europe are Slavic[8] it might not be simply a matter of religion. The other problematic aspect of religion is that it tends to separate languages rather than making them more similar, as in the case of Serbian which is written in both Cyrillic and Latin (Mønnesland 2002). The Cyrillic Serbian separates itself from the neighbouring Latin orthographies of Bosnian, Croatian, Albanian, Romanian and Hungarian which the Latin Serbian orthography does not.[9] Moreover it is slightly troublesome to define where the boundaries of religious unity actually should be set, as it could be set as Roman Catholic versus Protestant, Lutheran versus Calvinist etc. However I have tried to refer from explaining branches through religion since all the factors above trump religion.

Finally there are of course branches that have coincided and would be rather pointless to classify according to the factors above. I have at least tried to find another explanation to the branches that can't be explained by these factors by returning to the database, which I will present at the end of the discussion. This means that I have compared the orthographies that are grouped together by coincidence to see if any of the languages have more similarities to either a related language or a language spoken in its vicinity. If this has been true I have questioned the group. Last but not least there are some branches that are pure coincidence as all orthographies have to start somewhere and the options are not infinite.

[8] The obvious exceptions to this, i.e. non-Slavic languages with Cyrillic orthographies, are minority languages in Russia.
[9] The Latin Serbian separates itself from the Cyrillic orthographies of Macedonian and Bulgarian though.

4. Results

Below in figure 4.1 you will find a dendrogram generated for all the phonemes of the 45 orthographies.

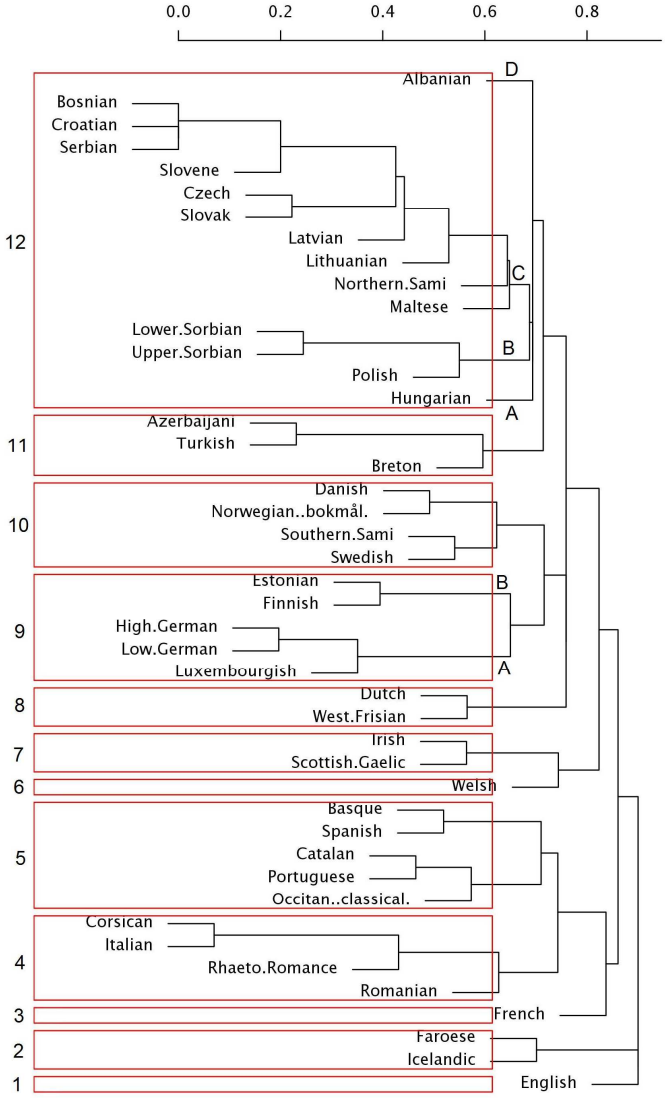

Figure 4.1: Dendrogram of 45 Latin orthographies in Europe grouped in 12 branches (in squares).

As figure 4.1 shows the computer program generated 12 branches for the 45 orthographies. They are numbered according to the order of their separation from the rest of the orthographies. English is the first branch to split off and it is followed by the second branch consisting of Faroese and Icelandic and these orthographies are the most divergent of the 45 orthographies. All the remaining 42 orthographies are a part of a shared branch which splits up into two major branches, the first is the common branch of branches 3, 4 and 5 and the

other one consists of branches 6 to 12. The sub-branches of the first major branch are branch 3, which is constituted by French, branch 4, consisting of Romanian, Rhaeto-Romance, Italian and Corsican and finally branch 5 which covers the languages of the Iberian Peninsula and Occitan.

The first two groups to split off from the second major branch are branches 6, consisting of Welsh, and 7, consisting of Irish and Scottish Gaelic. They are followed by branch 8, comprised of Dutch and West Frisian, and thereafter branch 9, in turn divided into one sub-branch which covers High German, Low German and Luxembourgish and one sub-branch consisting of Finnish and Estonian. Branch 9 shares a common branch with branch 10 which is comprised of Danish, Norwegian (Bokmål), Southern Sami and Swedish. Afterwards we find the common branch of branches 11 and 12, where branch 11 consists of Turkish, Azerbaijani and Breton and branch 12 is constituted by all the studied Balto-Slavic languages, Hungarian, Albanian, Northern Sami and Maltese.

Some of the branches mentioned above could be divided further into sub-branches. Since this is more relevant for certain branches they will receive defined sub-branch names. Firstly branch 9 shall be divided into sub-branch 9A consisting of High German, Low German and Luxembourgish and sub-branch 9B consisting of Finnish and Estonian. Similarly branch 12 shall also be divided into sub-branches 12A comprising Hungarian, 12B comprising Polish and Upper and Lower Sorbian, 12C comprising Maltese, Northern Sami, the Baltic languages, Czech, Slovak and the Southern Slavic languages and 12D comprising Albanian.

It is finally important to note that when a similar dendrogram was generated for only 44 of the orthographies there were some significant differences, e.g. that German was grouped amongst the Slavic languages. In another dendrogram consisting of only 27 of the orthographies branch 4 was further away from the other Romance languages. This shows that the amount of orthographies is a relevant factor for the results and therefore the results might have become different if other European or non-European orthographies would have been added. Below in figure 4.2 you will find the dendrogram covering only the consonants of the 45 orthographies.

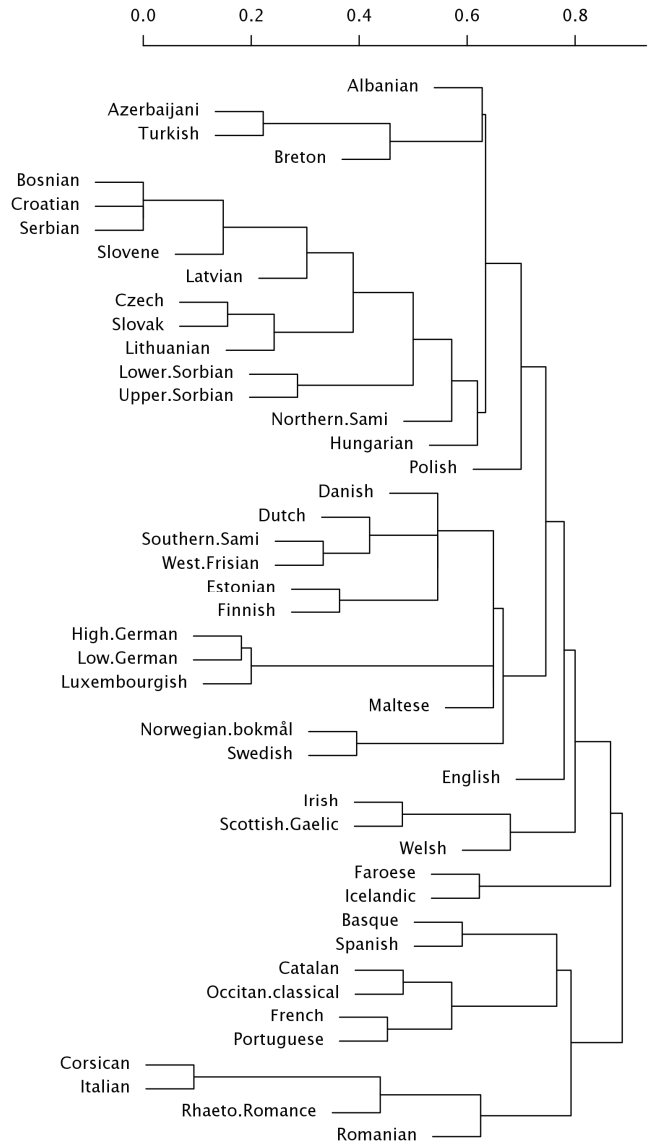

Figure 4.2: Dendrogram of the consonants of the 45 orthographies without further grouping.

The results of the consonantal dendrogram are mainly the same as figure 4.1 but with some important exceptions. The first is that the Romance languages and Basque are the first to split off from the rest of the orthographies and French is no longer set apart since it shares a consonantal sub-branch with Portuguese. In a similar fashion English is not the first to diverge nor is it set apart since it splits off after both the Faroese-Icelandic branch and the Celtic languages but somewhat earlier than the rest of the Germanic languages. It is also interesting to note that all the Germanic languages except English share a common branch together with Finnish, Estonian, Southern Sami and Maltese. Branch 12 is more or less intact with the exception of Polish which splits off into its own branch and the Sorbian

orthographies are closer to the rest of the Slavic orthographies. The last great difference is that Turkish-Azerbaijani and Breton are consonantally more similar to Albanian than they are in figure 4.1. Figure 4.3 below is a dendrogram covering only the vowels of the 45 orthographies.

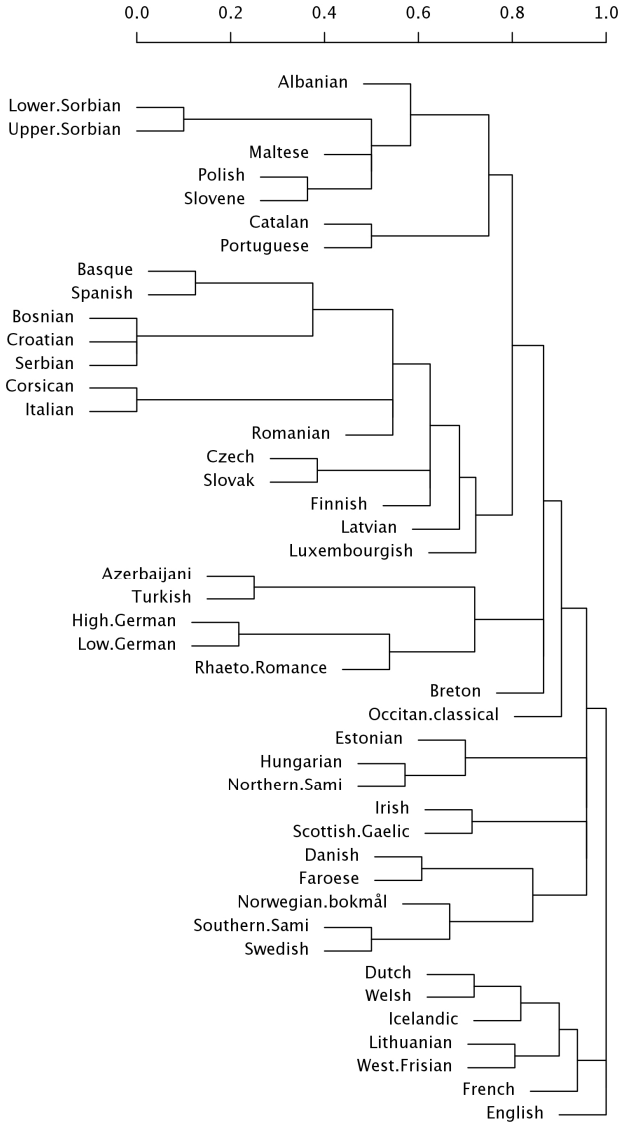

Figure 4.3: Dendrogram of the vowels of the 45 orthographies without further grouping.

The results of the vocalic dendrogram are significantly different to both the results of the dendrogram covering both vowels and consonants and the dendrogram covering only consonants. In contrast to figure 4.2 but similar to figure 4.1 both English and French split off quite early. What is more interesting is the branch that French is a part of since it also consists of orthographies from groups 2, 6, 8 and 12 which are not at all grouped together in figures

4.1 and 4.2. Thereafter there are three branches splitting off simultaneously, where the first branch comprises branch 10 but with the insertion of Faroese, the second branch is branch 7 and the third covers the Hungarian, Northern Sami and Estonian. One of the more interesting branches of the vocalic dendrogram is the shared branch of Rhaeto-Romance, High German-Low German and Turkish-Azerbaijani which is a grouping not found in the other two. The remaining orthographies are split into two major branches where the first constitutes half of the orthographies of branch 12, all of branch 4 except Rhaeto-Romance, Luxembourgish, Finnish and Spanish-Basque and the second branch covers the rest of branch 12 and Portuguese-Catalan. Below you will find the geographical and genealogical distribution.

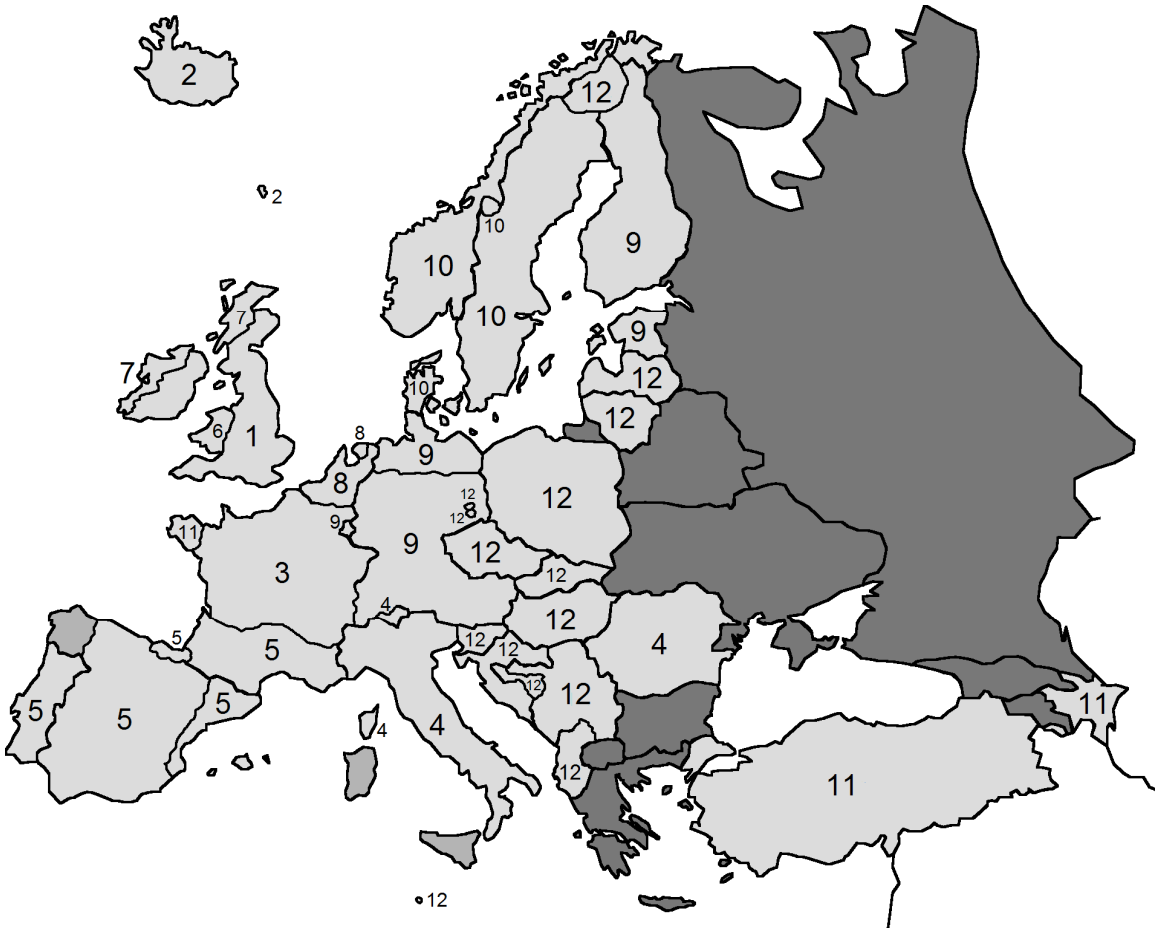

Figure 4.3: Geographical distribution of the 12 branches. The languages in light grey are the 45 orthographies. Languages in slightly darker grey are languages with Latin orthographies that are not studied and the dark grey are languages not written in the Latin alphabet.[10]

[10] N.B. that the distribution of some languages is rather arbitrary, as for example Low German, Occitan, Southern and Northern Sami, Hungarian and the Celtic languages. Some languages have been intentionally left out to simplify the map, e.g. the regional languages of northern Italy, Spain and Germany and minority languages in Russia, Romania and Ukraine.

Table 4.1: A genealogical tree of the studied languages. Separate languages are found to the right followed by the number of their orthographic branch. The genealogical tree is according to Ruhlen 1991.

<table>
<tr><td rowspan="20">Indo-European</td><td colspan="4">Albanian 12</td></tr>
<tr><td rowspan="4">Romance</td><td colspan="2">Eastern Romance</td><td>Romanian 4</td></tr>
<tr><td rowspan="3">Western Romance</td><td rowspan="2">Gallo-Ibero-Romance</td><td>Gallo-Romance:
French 3</td></tr>
<tr><td>Ibero-Romance:
Catalan 5
Occitan 5
Portuguese 5
Spanish 5</td></tr>
<tr><td>Italo-Romance</td><td>Corsican 4
Italian 4</td></tr>
<tr><td>Rhaeto-Romance[11]</td><td>Rhaeto-Romance 4</td></tr>
<tr><td rowspan="2">Celtic</td><td colspan="2">Brythonic</td><td>Breton 11
Welsh 6</td></tr>
<tr><td colspan="2">Goidelic</td><td>Irish 7
Scottish Gaelic 7</td></tr>
<tr><td rowspan="4">Germanic</td><td rowspan="2">North Germanic</td><td>East North Germanic</td><td>Danish 10
Swedish 10</td></tr>
<tr><td>West North Germanic</td><td>Faroese 2
Icelandic 2
Norwegian (Bokmål) 10</td></tr>
<tr><td rowspan="2">West Germanic</td><td>Continental West Germanic</td><td>Dutch 8
High German 9
Low German 9
Luxembourgish 9</td></tr>
<tr><td>North Sea (Anglo-Frisian)</td><td>English 1
West Frisian 8</td></tr>
<tr><td rowspan="4">Balto-Slavic</td><td colspan="2">Baltic</td><td>Latvian 12
Lithuanian 12</td></tr>
<tr><td rowspan="3">Slavic</td><td>South Slavic</td><td>Bosnian 12
Croatian 12
Serbian 12
Slovene 12</td></tr>
<tr><td rowspan="2">West Slavic</td><td>Central West Slavic:
Lower Sorbian 12
Upper Sorbian 12
North West Slavic:
Polish 12
South West Slavic:
Czech 12
Slovak 12</td></tr>
<tr><td rowspan="3">Finno-Ugric</td><td rowspan="2">Finnic</td><td rowspan="2">North Finnic</td><td>Baltic Finnic</td><td>Estonian 9
Finnish 9</td></tr>
<tr><td>Samic</td><td>Northern Sami 12
Southern Sami 10</td></tr>
<tr><td colspan="2">Ugric</td><td>Hungarian 12</td></tr>
<tr><td>Turkic</td><td colspan="2">Southern Turkic</td><td>Azerbaijani 11
Turkish 11</td></tr>
<tr><td>Semitic</td><td>West Semitic</td><td colspan="2">Central West Semitic</td><td>Maltese 12</td></tr>
<tr><td colspan="4">Basque 5</td></tr>
</table>

[11] N.B. that Rhaeto-Romance is classified as Gallo-Romance by the International Encyclopedia of Linguistics.

5. Discussion

Well how do I explain these results? First of all it was roughly what I expected, with some apparent surprises that will be discussed later. Leaving English for the moment we will focus on the second branch, or the Icelandic-Faroese branch. Since both Icelandic and Faroese are West North Germanic languages, more specifically Insular West North Germanic languages, this branch is clearly genealogical. Possible explanations to why these two languages split off this early will be given later. Thereafter we have the common group of 3, 4 and 5 which comprises all the Romance languages in this study, which is clear in table 4.1. Even though all of these languages are related it also has some amount of areality since the non-Romance orthography Basque is included.

Branch 3 with French is set apart from the rest of the common branch of 3, 4 and 5 entirely on the basis of the French vowel system, as becomes obvious when comparing figure 4.2 and figure 4.3. Furthermore the genealogical links of branch 4 are strong enough to actually make the common branch of 3, 4 and 5 genealogical rather than areal. This is true for branch 4 since there has been no historical links between its languages since the Roman period but a clear genealogical relationship (Ruhlen 1991). This is of course not true for Italian and Corsican, which have had strong historical ties. Therefore this branch must be genealogical, even though some question marks could be raised regarding Italian and Rhaeto-Romance. They are spoken in the proximity to each other and Italian has had an important role in the region. Even though this might be true Graubünden, where Rhaeto-Romance is spoken, has been under German dominance during its recent history (Schweizer Lexikon 1991, vol. 3). This makes a non-genealogical historical explanation less valid. However, this might not be as true if you do not follow Ruhlen's classification since the genealogical relation is quite different if Rhaeto-Romance is Gallo-Romance instead. Nevertheless, branch 4 should primarily be regarded as genealogical though.

If we continue with branch 5 it appears to be an areal branch as it is only found on and around the Iberian Peninsula and it covers two different language families, four Romance languages and the isolate Basque. This is true but as the four Romance languages are closely related, as they are all Ibero-Romance languages, the branch has an important amount of genealogical relations as well. Even though Basque is a part of this branch it is basically a genealogical

branch that Basque borrowed from when the Basque orthography was created. The majority of the Basque Country has been a part of Spain since the 16th century (Darby & Fullard 1970) which would explain why the Basque orthography has significant similarities with the Spanish. It is interesting to note that the Basque orthography is as close as it is to the Spanish orthography since written Basque appears at first sight to be highly divergent from the neighbouring Spanish orthography due to the use of e.g. *k* for /k/ in Basque. This is apparently misleading since the Basque orthography is highly similar to the Spanish one. This leads me to say that the explanation to why Basque is in the group is historical but the branch itself is basically genealogical with some areality.

The common branch of 6 and 7 would be easy to classify as purely genealogical but even though the languages in these branches are all Celtic some areal aspects should be regarded. This is especially true when it comes to the consonants as the branch that splits off first after the languages of branch 6 and 7 is English. Due to dominant role of the English language in the British Isles its influence should be taken in consideration. But since the overall orthographic systems of branches 6 and 7 and English are not that close and the similarities between the consonant systems are not necessarily that great. The only reasonable classification that can be made is that the common branch of 6 and 7 actually first of all is genealogical. Thereafter we find branch 8 which ought to be areal with a high amount of genealogy since even though Dutch and West Frisian are both West Germanic languages they are not that closely related. But if branch 8 is put into its context where it is closest to the common branch of 9 and 10 the two branches form sister branches of Germanic languages, namely Dutch, West Frisian, High German, Low German, Luxembourgish and the North Germanic orthographies except Icelandic and Faroese. The fact that branch 8 has this related sister branch makes it more genealogical.

Furthermore there are not just Germanic languages in branch 9, a fact that becomes obvious if you look at sub-branch 9B which has only Finno-Ugric languages. However 9A is clearly genealogical since it consists of three closely related West Germanic languages. Sub-branch 9B is also genealogical since it consists of the closely related Baltic Finnic languages of Finnish and Estonian. But if we look at the whole of branch 9 it cannot be anything else than coincidence since they are not found in the same area nor do they share historical ties, except that Estonia was controlled by the state of the German-speaking Teutonic Order from the Medieval Ages up until about 1500 (Darby & Fullard 1970). The reason to why these two

sub-branches are grouped together is likely to be found in its sister branch 10. The Scandinavian languages of branch 10 have dominated Scandinavia and the northern Baltic Sea for centuries. Swedes have been present in Finland since at least the 13th century and Finland was a part of Sweden until 1809 (Barðdal et al. 1997). Swedish has consequently had a significant impact on Finnish, not to mention the presence of Swedish still found in Finland today. Even though Estonia also has been a part of Sweden it was more or less only during the 17th century (Darby & Fullard 1970) which makes it more likely that the similarity between Finnish and Estonian is due to orthographical borrowing. If we look at branch 10 itself it is a clearly genealogical branch since Danish, Norwegian and Swedish are all closely related languages. Southern Sami has then borrowed its orthography from these languages which is not that surprising since Southern Sami is spoken in an area that has been politically dominated by Scandinavians for more than thousand years (Barðdal et al. 1997, p. 64).

Branch 11 is the most bewildering of all branches as it stretches from Breton in Western Europe to Azerbaijani on the border to Asia and covers two non-related language families. Therefore branch 11 must be a case of pure coincidence. Nevertheless it contains the closely related Turkish and Azerbaijani orthographies, which on their own form a genealogical branch, but since they for some reason share this branch with Breton it is not genealogical.

Thereafter we have the largest branch, branch 12, which could be summed up as the Balto-Slavic languages, Hungarian, Albanian, Northern Sami and Maltese. First of all Balto-Slavic could not form a relevant orthographically genealogical branch as they are too distantly related. It is interesting to note that branch 12 is found in a continuous area, excluding Maltese and the Sorbian orthographies, from Serbia in the south to Latvia in the north, which makes branch 12 an areal branch. These areas have all except Serbia, Albania, most of Lithuania and Latvia been under German-speaking dominance in the 19th century as parts of either Austria/Austria-Hungary or Prussia/Germany (Darby & Fullard 1970), which makes this branch a combined areal and historical branch. So the orthographies of branch 12 have likely evolved in relation to the German orthography because of the historical dominance of German in Central Europe.

Sub-branch 12C has a lot of important differences to the other sub-branches of branch 12 since the orthographies of 12C are all highly monographic and all but Maltese use the set of *c*,

č, *š* and *ž* for the phonemes /ʦ/, /ʧ/, /ʃ/ and /ʒ/.[12] These graphemes are interestingly enough also found in both Upper and Lower Sorbian but not in Polish. Polish, Hungarian and Albanian are all noticeably more polygraphic than the other orthographies of branch 12, since they represent /ʧ/ as *cz*, *cs* and *ç*, /ʃ/ as *sz*, *s* and *sh* and /ʒ/ as *ż* or *rz*, *zs* and *zh* respectively. The status of sub-branches 12A and 12D will be discussed below but 12B is clearly genealogical. If you exclude Maltese and Northern Sami from 12C it is what I can see two intertwined genealogical branches, namely the South Slavic-South West Slavic branch and the Baltic branch. Even though the entire branch of 11 is found in a continuous area the languages of branch 12C are not. The South Slavic-South West Slavic branch is areal as well though, as all of these languages except Serbian were under Austrian and later Austrian-Hungarian dominance from the 18th century to 1918 and some of them like Czech even earlier (Darby & Fullard 1970). It is also interesting to note that there has been a presence of German in the Baltics since the Medieval Ages due to crusades and the Teutonic Order (Darby & Fullard 1970).

The sub-branches 12A and 12D are in branch 12 due to areal factors, since they are found in areas surrounded by branch 12C. But the fact that both Albanian and Hungarian diverge from the rest of branch 12 is interesting since they are also genealogically diverging as Albanian constitutes its own genealogical branch amongst the Indo-European languages and Hungarian is the only Ugric language in the area. I therefore question the relevance of branch 12A and 12D amongst the rest of branch 12. It is also relevant to look at the consonantal branches as Albanian then is significantly closer to Turkish-Azerbaijani-Breton than it is in figure 4.1. This is interesting since Albania was a part of the Turkish-speaking Ottoman Empire from before 1500 to the beginning of the 20th century (Darby & Fullard 1970).

When it comes to branch 1, i.e. English, it is surely on its own due to its strong historical stance and since it has had its own sphere of influence it has not been influence by others to the same degree. But this does not explain why English is set apart as much as it is. The English orthography displays a relatively high amount of morphography and it is also a rather deep orthography (Rogers 2005, p. 275) This relatively widespread morphography is not

[12] Northern Sami uses the first three similarly to the others of 12C but *ž* stands for /ʤ/ (Nickel 1994).

found to the same extent in other European orthographies[13] which might be a contributing factor to why English is set so far apart from the rest of the orthographies.

Returning to branch 2 I will try to explain why it splits off so early. According to figure 4.2 Icelandic and Faroese have the most divergent consonant systems among the non-Romance orthographies. This is probably the result of the lack of a couple of common voiced phonemes such as /b/, /d/ and /g/ even though they have their graphemic counterparts *b*, *d* and *g* which are pronounced voicelessly. If this is combined with the fact that both Icelandic and Faroese have several graphemes which they do not share with each other or any other orthography this lack of voiced plosives becomes significant.

Finally I tested the tree model against the database myself since I was slightly sceptical to some of the branches. This I did by comparing the amount of graphemes they shared and then related this to their total amount of graphemes as well. One of the results this gave me was that Breton had more graphemes in common with French than it had with Turkish which shows us that Breton is an orthography that is hard to place in this tree. Another result concerning branch 11 was that Turkish had as many graphemes in common with Albanian as it had with Breton, and the common graphemes shared by Turkish and Hungarian were marginally less than the ones shared with Breton. Remarkably, Turkish had about as many graphemes in common with High German, i.e. 22, as it had with Breton, 21, or Hungarian, 20. This both questions branch 11 and the presence of Breton amongst these orthographies. Similarly Hungarian and High German shared approximately the same amount of graphemes as Hungarian and Turkish, which shows that there are relevant similarities between the common branch of 11 and 12 and High German. Turkish and Azerbaijani shared on the other hand approximately 90% of their graphemes when the other orthographies in branch 11 shared just above 50%, which makes Turkish and Azerbaijani a highly relevant grouping. Furthermore I found that High German shared exactly as many graphemes, i.e. 29, with Dutch as Dutch shared with West Frisian, which tells us that there are some interesting similarities between the Continental West Germanic orthographies.

The possible explanation to the similarity of the Celtic orthographies as due to influence from English was proven to be unlikely as they only shared 12 graphemes, which is about the same

[13] With some reservation for the Celtic languages.

amount of graphemes that are shared between all orthographies. The common group of 6 and 7 had almost twice as many graphemes in common but as it only constituted about 40% of all graphemes it actually does not make the branches 6 and 7 all that close. When it comes to branch 2 it should be noted that Faroese shares slightly more graphemes with Norwegian than it shares with Icelandic, and Faroese actually shares the same amount of graphemes with Icelandic as it does with Swedish. Interestingly enough Icelandic, which is less similar to the languages of sub-branch 10 than Faroese is, shares as many graphemes with 10 as the whole of branch 9 shares with each other. This shows that branch 2 is not as divergent as it seems in figure 4.1. Branch 9 was also proven to be rather irrelevant as they just shared 18 graphemes, which is only a few more than what is shared between all the Germanic languages except English, Icelandic and Faroese. It is also important to notice that Swedish and Finnish only shared 20 graphemes, which is less than both the amount of graphemes shared by Finnish and High German and Finnish and Dutch. This means that Branch 9B is more likely an independent branch amongst the Germanic orthographies than a sub-branch of either German or the Scandinavian orthographies, which questions the relevance of branch 9. It is also interesting to note that although branch 10 had 21 graphemes in common Danish, Norwegian and Swedish had 25 graphemes in common while Danish and Norwegian shared 31 graphemes and Norwegian and Swedish shared 36 graphemes. Something that was rather remarkable was that Swedish shared 26 graphemes with High German and 25 with Danish.

The absolutely most important finding was that branch 12 shared just 13 graphemes which is about 35% and is therefore not at all relevant, which I predicted. The sub-branches are relevant though, as sub-branch 12B shared 25 graphemes or almost 60%. Sub-branch 12C is in its entirety not as relevant as it only shared 16 graphemes or slightly more than 40% but without Maltese that number rises to 19 graphemes or about 50%. Without Northern Sami it shares 23 graphemes or more than 60% and without the Baltic languages it reaches 27 graphemes or almost 75%. It is also interesting to note that the Baltic orthographies are roughly as similar to each other as all the South and South West Slavic orthographies are at almost 75%. This means that there are considerable differences between Latvian and Lithuanian, even though they share more graphemes with each other than the Slavic orthographies of 12C do. The fact that Northern Sami is a part of 12C could be challenged since Southern Sami and Northern Sami share 21 graphemes while Northern Sami and 12C without Maltese share 19 graphemes. Furthermore the Slavic orthographies of sub-branch 12C share approximately as many graphemes with Romanian as they share with Turkish,

which is likely due to some areal features amongst the orthographies of South Eastern Europe. The similarities between Romanian and the rest of branch 4 were higher though and even more so in comparison to Italian, as they shared 25 graphemes or about 70% when Romanian shared 18 to 19 graphemes with Turkish and the Slavic languages of the Balkans.

6. Conclusions

After going through these results one conclusion is readily made: The orthographical variation of these European languages is neither arbitrary nor random. Even though this is true it is not a simple task to make any conclusions whether this variation is genealogical, areal or social. Nevertheless the tendencies are quite strong and a relevant conclusion is that the variation is based on genealogical factors to a higher extent than any other factors. As the results gave us shared branches for the Romance languages, the Baltic languages, the Slavic languages, the Turkic languages, the Baltic Finnic languages, the Continental Northern Germanic languages and the Insular Northern Germanic languages I find it hard to explain this with any other relevant and comprehensive cause than their genealogical linkage. This is not the case for vowel systems though as their distribution is arbitrary to a significantly higher extent.

Even though the Germanic orthographies did not share a common branch they had a significant amount of graphemes in common which lead to three Germanic sister branches. This argues further for the importance of the genealogical factor. In a similar fashion Icelandic and Faroese were proven to be significantly closer to the rest of the Northern Germanic languages than it appears with the computer generated tree model.

The conclusion that the Germanic orthographies are rather close tells us something about all the orthographies that they share their branch with, i.e. the Slavic orthographies, the Baltic orthographies, the Finno-Ugric orthographies and the Turkic orthographies. Since all but the Turkic languages have been dominated by Germanic languages it is reasonable to conclude that these orthographies are either based on or influenced by the Germanic orthographies, with some exceptions for English. This would arguably lead to a more significant role for areal and historical factors but since the diffusion of these derived orthographies mainly stay within groups of languages that are genealogically related the factor of genealogy still prevails as the most important.

The question of derived orthographies leads us to the next significant conclusion, namely that when an orthography is created it will most likely be based on the orthography of a politically dominant language, especially if the non-dominant language does not have its own state. This leads to what could be seen as direct or indirect orthographically colonisation by the official

or dominant language of the state. This is the case for Basque, Finnish, Southern Sami, to some extent Breton and perhaps the Slavic orthographies. It would therefore be interesting to see if this phenomenon of orthographical colonisation is found in languages outside of Europe that uses the Latin alphabet and have a colonial past. Will they be more similar to their colonising languages or will they have entirely different orthographic systems? This calls for us to look beyond the borders of Europe into the Latin orthographies of Asia, Africa, Oceania and last but not least the Americas.

References

Árnason, Kristján (2011). *The Phonology of Icelandic and Faroese*. Oxford: Oxford University Press

Ball, Martin John & **F**ife, James (ed.) (1993). *The Celtic languages*. London: Routledge

Barðdal, Jóhanna, **J**örgensen, Nils, **L**arsen, Gorm & **M**artinussen, Bente (1997). *Nordiska: våra språk förr och nu*. Lund: Studentlitteratur

Bergsland, Knut (1994). *Sydsamisk grammatikk*. 2. utg. Karasjok: Davvi girji

Bjellerup, Sven (1990). *Portugisisk språklära*. 3. uppl. Lund: Studentlitteratur

Blanche-Benveniste, Claire & **Y**aguello, Marina (ed.) (2003). *Le grand livre de la langue française*. Paris: Seuil

Brinton, Laurel J. & **B**rinton, Donna M. (2010). *The linguistic structure of modern English*. Rev. ed. Amsterdam: John Benjamins Pub. Co.

Bruce, Gösta, *Vår fonetiska geografi: om svenskans accenter, melodi och uttal*, 1. uppl., Studentlitteratur, Lund, 2010

Bruyne, Jacques de & **P**ountain, Christopher J. (1995). *A comprehensive Spanish grammar*. Oxford: Blackwell

Comrie, Bernard & **C**orbett, Greville G. (ed.) (1993). *The Slavonic languages*. London: Routledge

Daniels, Peter T. & **B**right, William (ed.) (1995). *The world's writing systems*. Oxford: Oxford Univ. Press

Darby, H. C. & **F**ullard, Harold (ed.) (1970). *The New Cambridge Modern History. 14, Atlas*. Cambridge: Cambridge U.P.

Frawley, William J. (ed.) (2003). *International Encyclopedia of Linguistics*. 2. ed. Oxford: Oxford University Press

Giacomo-Marcellesi, Mathée (1997). *Corse*. München: LINCOM Europa

Herslund, Michael (2002). *Danish*. München: LINCOM Europa

Hualde, José Ignacio (1992). *Catalan*. London: Routledge

Kahl, Heinrich. & **T**hies, Heinrich. (2002). *Plattdeutsches Wörterbuch: plattdeutsch-hochdeutsch ; hochdeutsch-plattdeutsch ; plattdeutsche Rechtschreibung.* 2., übearbeitete Aufl. Neumünster: Wachholtz

Karlsson, Fred (2009). *Finsk grammatik*. 9., utök. och rev. uppl. Helsinki: Suomalaisen Kirjallisuuden Seura

Liberman, Alvin M. (1992). The Relation of Speech and Reading and Writing. In: Frost, Ram & Katz, Leonard (ed.) (1992). *Orthography, phonology, morphology, and meaning*. Amsterdam: North-Holland

Liver, Ricarda (1999). *Rätoromanisch: eine Einführung in das Bündnerromanische*. Tübingen: Narr

Mathiassen, Terje (1996). *A short grammar of Lithuanian*. Columbus, Ohio: Slavica Publishers

Möhn, Dieter & Lindow, Wolfgang (ed.) (1998). *Niederdeutsche Grammatik*. 1. Aufl. Leer: Schuster

Mønnesland, Svein (2002). *Bosnisk, kroatisk, serbisk grammatikk*. Oslo: Sypress

Newton, Gerald (red.) (1996). *Luxembourg and Lëtzebuergesch: language and communication at the crossroads of Europe*. Oxford: Clarendon Press

Nickel, Klaus Peter (1994). *Samisk grammatikk*. 2. utg., [rev.] Karasjok: Davvi Girji

Nouvel, Alain (1975). *L'occitan sans peine*. Chennevières sur Marne: Assimil

Popkema, J. (2006). *Grammatica Fries*. Utrecht: Prisma Woordenboeken en Taaluitgaven

Rogers, Henry (2005). *Writing systems: a linguistic approach*. Malden, Mass.: Blackwell

Ruhlen, Merritt (1991). *A guide to the world's languages*. Stanford, Calif.: Stanford University Press

Saltarelli, Mario (1988). *Basque*. London: Croom Helm

***S**chweizer Lexikon 91*.. (1991). Luzern: Schweizer Lexikon Mengis+Ziehr

Serianni, Luca (2006). *Grammatica italiana: italiano comune e lingua letteraria*. Torino: UTET

Sgall, Petr (1984). Towards a theory of phonemic orthography. In: Luelsdorff, Philip (ed.) (1987). *Orthography and phonology*. Amsterdam: J. Benjamins Pub. Co.

Vachek, Josef (1976). *Selected writings in English and general linguistics*. Praha: Academia

Öztopçu, Kurtuluş, **A**buov, Zhoumagaly, **K**ambarov, Nasir & **A**zemoun, Youssef (1996). *Dictionary of the Turkic languages: English: Azerbaijani, Kazakh, Kyrgyz, Tatar, Turkish, Turkmen, Uighur, Uzbek*. London: Routledge

Electronical sources

Deutsche Rechtschreibung. Regeln und Wörterverzeichnis:
http://rechtschreibrat.ids-mannheim.de/download/regeln2006.pdf (2011-10-19)

Appendix A: The binary database

	Alban.	Azer.	Basq.	Bosn.	Bret.	Cat.	Cors.	Croat.	Cz.	Dan.	Dutch	Engl.	Est.	Faroe.	Finn.	French	H.Ger.	Hungar.	Icel.	Irish	Ital.	Latv.	Lit.	L.Ger.	L.Sorb.	Lux.	Malt.	N.Sami	Norw.	Occ.	Pol.	Port.	Rh.-R.	Rom.	Sc. Gael.	Serb.	Slovak	Slovene	S.Sami	Span.	Swed.	Turk.	U.Sorb.	Welsh	W.Fris
p.P	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
p._B											1						1									1					1				1										
p.B													1	1					1																										
b.B	1	1	1	1	1	1	1	1	1	1	1	1	1			1	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
b.BP																				1																									
b.V						1																								1									1					
t.T	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
t._D											1						1									1					1				1										
t.D													1	1					1																										
d.D	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
d.DT										1										1																									
c.Ķ																						1																							
c.Q	1																																											
c.Ť									1																								1										
c.K(v)																			1																										
c.G(v)																			1																										
c.KJ																			1																										
c.GJ																			1																										
c.TG																																	1										
c.CH																																	1										
c.TY																		1																										
c.CHJ							1																																				
ɟ.G																																	1										
ɟ.Ģ																						1																						
ɟ.Ď									1																								1										
ɟ.GJ	1																																											
ɟ.GY																		1																										
ɟ.GHJ							1																																				