Lexical and orthographic distances between Germanic, Romance and Slavic languages and their relationship to geographic distance

Wilbert Heeringa, Jelena Golubovic, Charlotte Gooskens, Anja Schüppert, Femke Swarte & Stefanie Voigt

Abstract

When reading texts of different but closely related languages, intelligibility is determined among others by the number of words which are cognates of words in the reader's language, and orthographic differences. Orthographic differences partly reflect pronunciation differences and therefore are partly a linguistic level. Dialectometric studies in particular showed that different linguistic levels may correlate with each other and with geography. This may raise the question of whether both lexical distance and orthographic distance need to be included in a model which explains written intelligibility, or whether both factors can even be replaced by geographic distance.

We study the relationship between lexical and orthographic variation among Germanic, Romance and Slavic languages to each other and to geography. The lexical distance is the percentage of non-cognate pairs, and the orthographic distance is the average of the Levenshtein distances of the cognate pairs.

For each language group we found a significant correlation between lexical and orthographic distances with a medium effect size. Therefore, when modelling written intelligibility preferably both factors are included in the model.

We considered several measures of geographic distance where languages are located at the center or capital of the countries where they are spoken. Both as-the-crow-flies distances and travel distances were considered. Largest effect sizes are obtained when correlating lexical distances with travel distances between capitals and when correlating orthographic distances with as-the-crow-flies distances between capitals. The results show that geographic distance may represent lexical and orthographic distance to some extent in a model of written intelligibility.

1. Introduction

Sometimes we are confronted with texts in an unknown but related language. An example may be Dutch people who visit Sweden and attempt to read a local newspaper in their hotel. Swedish is not standardly taught in Dutch elementary or high schools. Nevertheless, some words may be understood since both Dutch and Swedish belong to the Germanic language group. Other examples may be French people visiting Romania, or Polish people visiting Croatia. The extent to which the unknown language is understood may depend on the number of cognates, differences in orthography and syntax, etc.

The study presented in this paper was carried out in the context of a larger research program which aims at finding both linguistic and non-linguistic determinants of mutual intelligibility within the Germanic, Romance and Slavic language families. Within this research program, a large-scale web-based intelligibility test is performed, including cloze tests and word translation tasks. In a cloze test both texts and individual words are presented in either written or audible form, and subjects are asked to fill in missing words in the texts or to translate individual words. The higher the number of correctly produced or translated words, the better the intelligibility. Intelligibility scores are obtained on the basis of both written and spoken language. In this study an explanatory model will be developed in which several linguistic factors will be included such as lexical, orthographic, morphological and syntactic distances. The aim is to find out to what extent each of the factors contribute to the ease or difficulty in understanding a closely related language.

In this study we focus on lexical distance and orthographic distance being potential explanatory factors of the scores which will be obtained with the written word translation tasks. We focus on lexical and orthographic variation within the Germanic, Romance and Slavic language families. Lexical distance is a linguistic factor, and orthography is partly a linguistic factor. On the one hand, orthography reflects differences in spelling convention (e.g. German *Kontakt* versus Dutch *contact*); on the other hand, it reflects pronunciation differences (e.g. German *helfen* versus Dutch *helpen*).

Dialectometric studies show that linguistic levels may correlate with each other and with geography. Spruit et al. (2009) measured pronunciation, lexical and syntactic distances among Dutch dialects and calculated the correlations between the three linguistic levels on the basis of a subset of 70 local dialects. All of the correlations are significant. Additionally, they correlated each of the three measures with geography. The correlations are significant again. This agrees with Nerbonne & Kleiweg (2007: 154) who argue that it is a fundamental postulate of dialectology that geographically close varieties ought to be more linguistically similar than distant ones. They mention that the basic idea of this fundamental dialectological postulate has also been proposed in historical linguistics, and refer to Dyen (1956) and Campbell (1995). Theoretically the postulate is based on the wave model of language change (Schmidt 1872). From the earliest work in dialectometry it has been common to examine the dependence of dialect distance on geographic distance (Séguy 1971). Hinskens, Auer and Kerswill (2005) argue that dialects are primarily geographically defined,

but "geography as such does not influence language varieties, but does so through its social effects" (p. 28).

We measure lexical and orthographic distances within the Germanic, Romance and Slavic language groups. We answer the following questions for each of the language groups:

1 Do lexical and orthographic distances correlate with each other?
2 Do lexical and orthographic distances correlate with geographic distances?

When there is a strong correlation between lexical distance and orthographic distance, it would be sufficient to include either lexical distance or orthographic distance in the model of written intelligibility. When both factors would be included in a multiple regression model, the model indicates how well the entire bundle of predictors predicts the intelligibility scores, but it may not give valid results about any individual predictor, or about which predictors are redundant with respect to others.

The fact that dialect variation correlates with geographic distance does not necessarily mean that language variation also correlates with geographic distance. But in many cases languages are related to dialects and are derived from dialects. For example, Standard Dutch is especially related to the local dialect of Haarlem. All modern languages started as regional dialects. Therefore, we consider it as worthwhile to investigate whether lexical and orthographic distances between languages correlate with geographic distances. Relating language variation to geographic distance is less common than relating dialect variation to geographic distance, but is not new. An example is the work of Cisouw (to appear) on the relationship between cross-linguistic variation and geographic variation.

When we find that both lexical and orthographic distance strongly correlate with geographic distance and with each other, these two factors can be replaced by geographic distance in the model.

This research is particularly inspired by the *High Level Group on Multilingualism* (HLGM), a study group of the European Union. In 2007 this group noted a lack of knowledge about mutual intelligibility between closely related languages in Europe and the lack of knowledge about the possibilities for communicating through receptive multilingualism. Therefore, within each language group we consider languages of countries which belong to the European Union.

In Figure 1 the countries where the languages are spoken are colored in grey (Germanic), lighter grey (Romance) and darker grey (Slavic). The

Germanic group includes Danish, Dutch, English, German and Swedish. The Romance group includes Catalan, French, Italian, Portuguese, Romanian and Spanish. In principle we focus on national languages. However, in this paper Catalan is included, a regional language spoken in Catalonia.[1] The Slavic group includes Bulgarian, Czech, Polish, Slovak and Slovene.

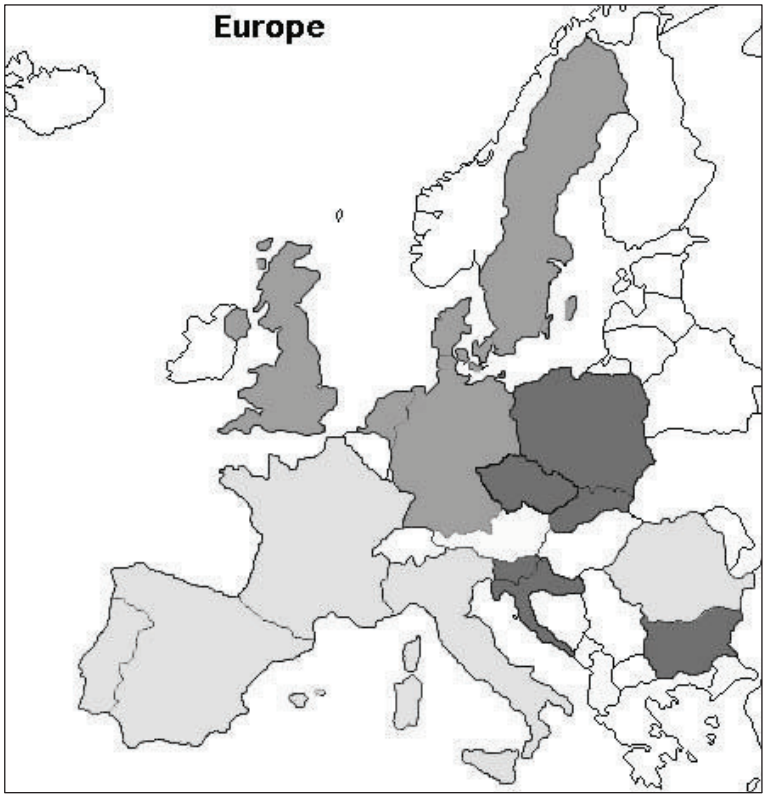

Figure 1: We consider (mainly national) languages spoken in countries which belong to the European Union. We focus on the Germanic language area (grey), the Romance language area (lighter grey) and the Slavic language area (darker grey). When a language is spoken in several countries, we choose the largest state in terms of surface.

In Section 2 we give a brief overview of related research concerning lexical and orthographic measurements. Section 3 describes the design of the data collection, and the way in which lexical and orthographic distances are measured. The results of the distance measurements are presented in Section 4. In Section 5 each of the research questions is addressed. Finally, some general conclusions will be drawn in Section 6.

1 Catalonia is an autonomous community of Spain, with the official status of a "nationality" according to the first article of the *Statute of Autonomy of Catalonia*. We thank Matthew Smith for providing us with the Catalan data.

2. Previous research

2.1 Lexis

Jean Séguy was director of the *Atlas linguistique de la Gascogne*. He and his associates published six atlas volumes containing maps in which single responses were plotted for 154 dialect locations. Séguy's major survey used five types of linguistic variables: 67 features from diachronic phonetics, 76 from phonology, 68 from morphosyntax, 44 from verb morphology and 170 lexical items, 425 variables in total. Séguy sought a way to analyze the maps in a more satisfying way than was possible with traditional analytic methods, and came up with the idea of counting "the number of items on which neighbors disagreed." (Chambers & Trudgill 1998: 138).

Hans Goebl continued and elaborated on Séguy's work, striking out independently in several respects. We base our sketch on Goebl (1993, 1984, 1982). Goebl most extensively analyzed *l'Atlas Linguistique de l'Italie et de la Suisse Méridionale* (AIS), compiled by Karl Jaberg and Jakob Jud in the first quarter of the twentieth century. He selected 251 varieties and 696 working maps from the AIS. Each working map represents a dialect feature and requires an assigned value at each of the 251 sites. 569 maps represent lexical variation, and 127 maps represent morphosyntactic variation. While Séguy calculated distances, Goebl calculated similarities. The similarity between two varieties is calculated as the percentage of items on which the two varieties agree.

The methodology of Séguy/Goebl has been frequently used for measuring lexical distances among local dialects. For example, Heeringa & Nerbonne (2006) measured lexical distances among 360 varieties in the Dutch dialect area, and Giesbers (2008) measured lexical distances among ten locations in the Kleverlands area, along the Dutch-German national border. The same methodology has also been used for measuring lexical distances between languages. Van Bezooijen & Gooskens (2005) investigated the intelligibility of Frisian and Afrikaans for speakers of Dutch, both in written and spoken form. They wrote: "A large proportion of cognates, i.e. words in two languages with a common root, may be expected to facilitate comprehension." They calculated lexical similarity between Afrikaans and Dutch and between Frisian and Dutch as the percentage of (paradigm-related) cognates and lexical distance as the percentage of non-cognates. The results suggest that both for the function words and the content words the relationship with the Dutch counterparts appears to be more direct for Frisian than for Afrikaans.

2.2 Orthography

Van Bezooijen & Gooskens (2005) also considered orthography as an explanatory factor of intelligibility between Afrikaans and Dutch and between Frisian and Dutch. They calculated orthographic distances by means of Levenshtein distance (Levenshtein 1966). The Levenshtein distance between two strings is calculated as the 'cost' of the total set of insertions, deletions and substitutions needed to transform one string into another (Kruskal 1999). When calculating orthographic distances the algorithm finds the minimum number of letters that need to be inserted, deleted or replaced when changing the spelled word of one language into the corresponding spelled word in another language. The authors found that the orthographic distances of cognates that are related directly or via a synonym is much smaller for Afrikaans than for Frisian.

Zulu, Botha & Barnard (2008) measured orthographic distance between 11 South African languages. Levenshtein distances were calculated using existing parallel orthographic word spellings in sets of 50 and 144 words from each of the 11 official languages of South Africa. This data was manually collected from various multilingual dictionaries and online resources. The authors concluded that statistical methods based solely on orthographic transcriptions are able to provide useful objective measures of language similarities.

Doetjes & Gooskens (2009) studied the role of orthography in the mutual intelligibility of Danish and Swedish spoken languages. They measured phonetic distances between the languages and took into consideration the help that the listeners can receive from the orthography when listening to the neighboring language. Both phonetic and orthographic distances were measured by means of Levenshtein distance. The authors concluded that Danish listeners indeed seemed to make use of the additional information that the orthography can provide.

3. Data source and measurement techniques

3.1 Compiling the database

The data set in the present study is compiled so that it models a written word translation task. The basic assumption we made is that a reader who is reading words spelled in a different but closely related language will understand the words relatively easily when cognates exist in his/her native language. In this study we adopt the definition of *cognates* as given by Van Bezooijen & Gooskens (2005): “words in two languages with a common

root." Cognates are words that have a common etymological origin. For example, English *child* is regarded as an etymological cognate of Dutch *kind*, and Swedish *rum* is considered as a cognate from the derived Dutch word *ruimte*. As will appear below, our notion of cognates covers real cognates - pairs of words which are mutual translations of each other, and partial cognates - pairs of words which have the same meaning in both languages only in some contexts. We make no difference between loan words and inherited words.

We compiled the data set in two steps, which are described below.

Step 1: compiling basic word lists

In this step the stimuli were collected. Since we were working with 16 different languages (see Section 1) we chose English as the baseline. Since we want to model written intelligibility it is important to use those words which a reader will most likely find in a written text. We considered using the Swadesh list, but we have to keep in mind that this list is a classic compilation of basic concepts for the purposes of historical-comparative linguistics, rather than for the purposes of modeling written intelligibility. Therefore, we preferred another approach, where we chose the most frequent nouns from the British National Corpus. A few of them were excluded if they were too similar to each other. For example, both *kind* and *sort* were found among the most frequent nouns, and one of them was left out because of the similar meanings of the two words. Finally a set of 100 words was left.

For each of the 100 words we provided a context to make sure the translators – native speakers of each of the languages who translated the words from English – knew the correct meaning. For example, the word *form* has different meanings such as 'questionnaire,' 'shape' or 'to shape something.' By putting the word in the sentence 'You have to fill out this form today,' the translator knows which meaning is used.

The 100 words were translated into each of the languages in the three language groups, where the context in which they appeared – the short sentences – was taken into account. If several translations were still possible, cognates of the English words were preferred to non-cognates. If there were several cognates, the orthographically closest cognate was chosen. When the cognates were orthographically equally similar to the cognates in the basic word list, the most commonly used word was chosen.

In this way we obtained basic word lists, five lists for the Germanic family (including the English list), six for the Romance family and six for the Slavic family. These lists were used for the word translation task.

Step 2: finding the cognates

We are trying to model comprehension of written language. The words in the basic word lists are the stimuli which will be presented to the readers who participate in a word translation task which we will conduct in the future. We expect that the reader will relatively easily understand stimuli for which cognates exist in his or her language. If no cognate of a stimulus exist, it will be hard for the reader to guess the meaning of that stimulus.

In order to model this we compiled the cognates of the stimuli which exist in the native language of the reader and which are synonyms. For each basic word list for each of the other languages in the same language group the cognates are found. The cognates should be synonyms, but not necessarily in the context of the 100 sentences. When in a particular language there are no cognates for some words in the basic list, the cells in the table remain empty.

Table 1: Schematical visualization of the database containing orthographic data of five Germanic languages. The database consists of five sheets. Each sheet consists of a basic word list (in gray) , and four cognate lists (white).

	Danish sheet 5 columns					Dutch sheet 5 columns					English sheet 5 columns					German sheet 5 columns					Swedish sheet 5 columns				
100 rows																									

The table for the Germanic language group is schematically shown as Table 1. For each of the languages there is one sheet, one for Danish, one for Dutch, one for English, one for German and one for Swedish. Each sheet consists of five columns. The first column contains the basic word list, consisting of 100 words, one word per row. When focusing on the English sheet, the first column is the basic word list consisting of English words. The second column contains the Danish cognates of the English

words, the third column contains the Dutch cognates, the fourth column contains the German cognates and the fifth column contains the Swedish cognates. The sheet is shown in more detail as Table 2.

Both for the Romance and Slavic languages we obtained a database which consists of six sheets, and each sheet consists of six columns.

Table 2: *Sheets 1 to 5, represented by five rows in Table 1, represent a sub-table each. Here we have a closer look at Sheet 3. A selection of ten out of 100 rows is shown. Column 1 contains the basic word list in English. Columns 2 to 5 contain the corresponding cognate translations in Danish, Dutch, German and Swedish, respectively. Empty cells appear when a cognate translation could not be found.*

	English	Danish	Dutch	German	Swedish
1	case		casus	Kasus	
2	child		kind	Kind	
3	house	hus	huis	Haus	hus
4	interest	interesse	interesse	Interesse	intresse
5	member				
6	question		kwestie		
7	week	uge	week	Woche	vecka
8	woman				
9	year	år	jaar	Jahr	år
10	law	lov			lag

3.2 Measuring lexical distances

Séguy (1973) measured the lexical distance between two dialects as the number of items upon which the two dialects disagree lexically because of heteronomy. We use basically the same method and measure the lexical distance as the percentage of non-cognates in the language of the reader compared to the stimulus language.

We illustrate this by using the example in Table 2, where English is the stimulus language, and the corresponding cognates are given in Danish, Dutch, German and Swedish. For each language the number of non-cognates – i.e. the number of empty cells in the table – is counted and divided by the total number of words. In our example there are ten words. The calculation is shown in Table 3. Danish and Swedish have the largest

percentage of non-cognates (50%) and Dutch has the smallest percentage of non-cognates (30%).

Table 3: The percentage of non-cognates of Danish, Dutch, German and Swedish with English is the number of non-cognates divided by the total number of words. Empty cells represent non-cognates. In this example Dutch is most similar to English (30%) and Danish and Swedish are most distant (50%).

	English	Danish	Dutch	German	Swedish
1	case		casus	Kasus	
2	child		kind	Kind	
3	house	hus	huis	Haus	hus
4	interest	interesse	interesse	Interesse	intresse
5	member				
6	question		kwestie		
7	week	uge	week	Woche	vecka
8	woman				
9	year	år	jaar	Jahr	år
10	law	lov			lag
	Number of non-cognates:	5	3	4	5
	Percentage of non-cognates:	50%	30%	40%	50%

3.3 Measuring orthographic distances

Orthographic distances between a stimulus language (data in the basic word list) and the language of the reader (cognates in one of the other columns in the same table) are measured with the aid of the Levenshtein distance metric. We illustrate this algorithm by comparing the English word *interest* with the Swedish word *intresse*:

	1	2	3	4	5	6	7	8	9
English	i	n	t	e	r	e	s	t	
Swedish	i	n	t		r	e	s	s	e
				1				1	1

In the fourth slot an *e* is deleted, in the eighth slot a *t* is replaced by an *s*, and in the ninth slot an *e* is inserted. As there are three operations and the alignment has nine slots, the distance is calculated as (3/9) x 100 = 33%. The word *interest* can be mapped to *intresse* in many different ways, but the Levenshtein distance always gives the cost of the cheapest mapping.

For each character we distinguish between a base and a diacritic. For example, the base of *é* is *e*, and the diacritic is the acute accent. Two characters may differ in the base and/or in their diacritics. We weigh differences in the base as 1; for example: *a* versus *e*, *p* versus *b*. If two characters have the same base, but different diacritics, we weigh this as 0.3. For example: *e* versus *é*, *è* versus *é*. We admit that the choice of this weight is not based on empirical measurements and may sound arbitrary, but our choice is motivated by the idea that diacritical differences should be weighed much less strongly than base differences, since differences in the base will usually confuse the reader to a much greater extent than diacritical differences. When the bases are different, the weight is 1, regardless of whether there are diacritical differences, since differences or similarities between diacritics are meaningless when the corresponding bases are different. Insertions and deletions are weighed as 1.

We assure that the minimum cost is based on an alignment in which a vowel matches with a vowel, and a consonant matches with a consonant.

We calculate the aggregated Levenshtein distance between the stimulus language and the language of a reader. We illustrate this by comparing English as stimulus language with Swedish as the language of a reader, using the data from Table 2. The comparison is made in Table 4. There are five English words for which a Swedish cognate exists (rows 3, 4, 7, 9 and 10); therefore, for each of the five word pairs the Levenshtein distance is calculated. The word pair distances are divided by the corresponding alignment lengths and multiplied by 100. The last column shows the percentage of different characters for each of the pairs. The average distance between the English stimuli and the cognates supposedly known by the hypothetical Swedish reader is 49.7% in this example. This percentage is

the hypothetical score of a Swedish reader who performs a word translation task in which English words are translated.

Table 4: Example of distance measurement between English (stimulus language) and Swedish (language of the reader) on the basis of ten words. The orthographic distances are obtained on the basis of five cognate pairs. Levenshtein distances are calculated for each of the pairs and divided by the corresponding alignment lengths and multiplied by 100. Thus, we obtain percentages which are given in the last column. In the last row the average percentage is given.

	English	Swedish	Levenshtein distance	Number of slots in the alignment	Percentage of different characters
1	case				
2	child				
3	house	hus	2	5	40.0%
4	interest	intresse	3	9	33.3%
5	member				
6	question				
7	week	vecka	4	6	66.7%
8	woman				
9	year	år	3	4	75.0%
10	law	lag	1	3	33.3%
	average				49.7%

3.4 Asymmetries

The advantage of the procedure described in Section 3.1 is that it takes into account that relationships between languages are not necessarily symmetric. Both lexical and orthographic distance may be asymmetric. We illustrate this by some examples.

The common Swedish translation of English *woman* is *kvinna*, and the common Dutch translation is *vrouw*. There exists no Dutch cognate for Swedish *kvinna*, but there does exist a Swedish cognate for Dutch *vrouw*, namely *fru*. The fact that Swedish has two synonyms – *kvinna* and *fru* – with preference for *kvinna*, and Dutch has just one word – *vrouw* – that is a cognate of Swedish *fru* causes an asymmetry. A Dutch-speaking reader will have difficulties understanding the word *kvinna* in a Swedish text,

while a Swedish-speaking person will be more likely to understand the word *vrouw* in a Dutch text. Table 5 shows more examples of lexical asymmetries.

Asymmetries at the lexical level may also affect the orthographic measurements. In Table 5 for 'Dutch stimuli versus Swedish reader' we find three complete cognate pairs, but for 'Swedish stimuli versus Dutch reader' we do not find any cognate pair. As a result, orthographic distances may be asymmetrical since the orthographic distance 'Swedish stimuli versus Dutch reader' is obtained on the basis of a different set of cognate pairs than the orthographic distance 'Dutch stimuli versus Swedish reader.'

Table 5: Swedish and Dutch stimuli are translated from English. There are no Dutch cognates of the Swedish stimuli, but there are Swedish cognates of the Dutch stimuli.

English	Swedish stimulus	Cognate known by Dutch reader	Dutch stimulus	Cognate known by Swedish reader
Woman	kvinna		vrouw	fru
Area	område		ruimte	rum
Voice	röst		stem	stämma

Table 6: German and Dutch stimuli are translated from English. The German and Dutch translations of the English words are not cognates of each other, but for the German words Dutch cognates can be found, and for the Dutch words German cognates can be found.

English	German stimulus	Cognate known by Dutch reader	Dutch stimulus	Cognate known by German reader
Thing	Sache	zaak	ding	Ding
Side	Seite	zijde	kant	Kante
Effect	Wirkung	werking	effect	Effekt

Asymmetric orthographic distances may also occur when the same English word is translated by different words in the different languages, where the words are not cognates of each other, although for each of the words cognates exist in the other language. For example the English word *thing* in the context of the sentence 'This thing I will never forget' is

translated as *Sache* in German and as *ding* in Dutch. The Dutch cognate of *Sache* is *zaak*, and the German cognate of *ding* is *Ding*. It is obvious that the orthographic distance between *Sache* and *zaak* is larger than between *ding* and *Ding*. Therefore we find an asymmetric orthographic distance. More examples are given in Table 6.

In short: asymmetry in orthographic distances may be found for any language pair and is the result of differences in cognate sets.

4. Lexical and orthographic distances between languages

In this section we present the results of lexical and orthographic measurements for each of the three language groups. Both lexical and orthographic distances are visualized by means of so-called beam maps (Inoue 1996). In the maps the countries are represented by their geographic centers (see Section 5.1). The centers are connected to each other by lines or 'beams,' where darker lines connect lexically or orthographically close languages and lighter lines more remote ones. Beam maps were introduced by Goebl (1993); in his maps only neighbouring locations are connected. We use the Groningen-style network maps where every location can in principle link to any other location in the network. These kind of maps were developed by Peter Kleiweg; examples can be found in Heeringa (2004).

In beam maps each pair of locations - geographic centers in our case - are connected by a line. In Section 3, and especially Section 3.4, we discuss that our methodology may reveal asymmetric relationships between languages. We illustrate this by an example. We found that the lexical distance between a native speaker of English who reads a Danish text is 41%, but the lexical distance between a native speaker of Danish who reads an English text is 30%. In a beam map both distances cannot be visualized. Therefore, we visualize the average distance: (41+30)/2 = 35.5%.

In each of the beam maps in this section the smallest distance is represented by a line which is nearly black, and the largest distance is represented by a white line. On a white background, however, white lines are not visible. By scaling color intensities between the smallest and largest distance per map the largest degree of detail is visualized.

4.1 Germanic

The lexical distances between the Germanic languages are given in Table A.1, and the orthographic distances in Table A.2 in Appendix A. The

averaged distances (see the introduction of this section) are geographically visualized in Figure 2.

At the lexical level we find English distinct from the other Germanic languages. In Table 3 English is lexically compared to the other Germanic languages on the basis of a subset of ten words. The percentage of cognates varies from 30% to 50%. For two words – *member* and *woman* – no cognate can be found in any of the other languages.

English originated from the fusion of closely related dialects, now collectively termed Old English, by Germanic settlers, and ultimately from their ancestral region of Angeln, presently known as Schleswig-Holstein. The language was influenced by the Old Norse language because of Viking invasions in the 8th and 9th centuries. The Norman conquest of England in the 11th century gave rise to heavy borrowings from Norman-French (Baugh & Cable 2002).

Close relationships are found between Danish and Swedish (Danish reader vs. Swedish stimulus: 4% / Swedish reader vs. Danish stimulus: 6%) and between Dutch and German (10%/14%). Danish and Swedish belong to the North Germanic group and Dutch and German belong to the West Germanic group.

At the orthographic level a close relationship between Danish and Swedish is also found, but relationships between other language pairs are weaker. German is most distant from Danish (34%/36%) and closest to Dutch (28%/32%). German is distinguished from the other languages as the result of the High German consonant shift or second Germanic consonant shift, probably beginning between the 3rd and 5th centuries AD, and was almost complete before the earliest written records in the High German language were made in the 8th century. In this shift the three Germanic voiceless plosives became fricatives in certain phonetic environments (Dutch *water* maps to German *Wasser*) and affricates in other positions (Danish *tid* versus German *Zeit*, where the *Z* is pronounced as /ts/). The three voiced plosives became voiceless. As far as Standard German is concerned, only /t/ became /d/ (Danish *del* versus German *Teil*) (Schwerdt 2000).

Dutch is also distinguished as the result of the *De Hollandsche Expansie* ("The expansion of linguistic characteristics from the West of the Netherlands to the non-peripheral regions," see Kloeke (1927)) which included the developments /u/→/y/→/œy/ and /i/→/ɛi/. These distinctions are reflected in the orthography; for example, Danish *hus* versus Dutch *huis* and Danish *tid* versus Dutch *tijd*. Although German also has

diphthongs here –/aʊ/ and /aɪ/ respectively – they differ in both pronunciation and spelling (Dutch *ui* and *ij* versus German *au* and *ei*).

English underwent the great vowel shift between 1350 and 1700 (Stockwell 2002). The two highest long vowels became diphthongs, and the other five underwent an increase in tongue height. The standard view in the Anglicist literature is that English orthography is a mess because it was standardized before the Great Vowel Shift was complete (Denham & Lobeck 2009). However, many words kept the old spelling (for example, compare the Danish pronunciation of *side* [si:ðə] (Hjort & Kristensen 2003) with the English pronunciation).

A distinguishing feature of Danish and Swedish (North Germanic) found in our data set is the loss of initial /j/ (example: Dutch *jaar* versus Danish *år*).

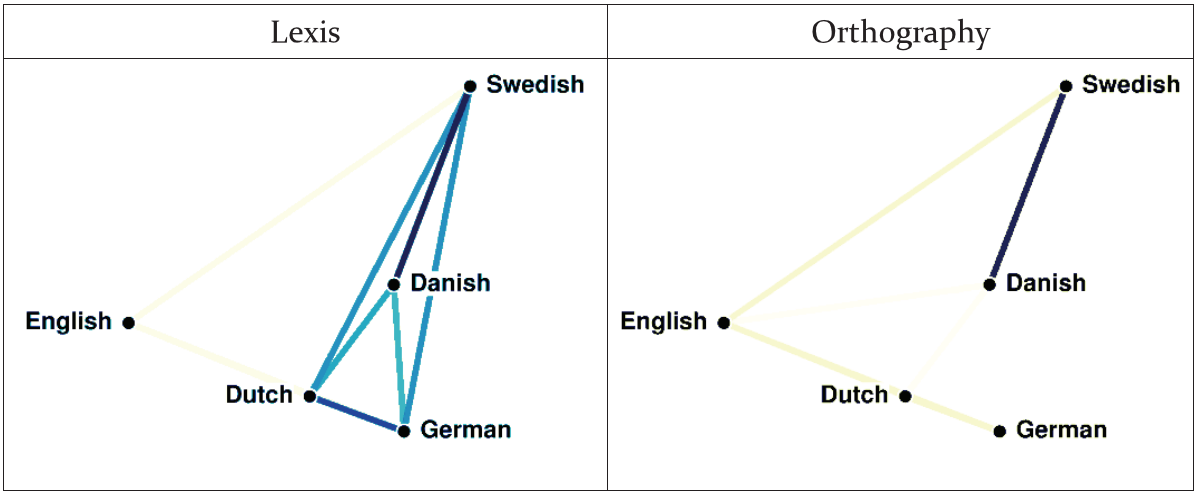

Figure 2: *Lexical and orthographic distances between Germanic languages. Darker lines connect lexically/orthographically close varieties, lighter lines more remote ones. The largest distances are represented by white lines which are invisible on the white background. At the lexical level the distances vary from 5% (between Danish and Swedish) to 36% (between Danish and English), and at the orthographic level distances vary from 17% (between Danish and Swedish) to 37% (between Danish and German).*

In Section 3.4 we explained that lexical and orthographic distances may be asymmetric. We checked whether there exists an asymmetric relationship between languages *A* and *B* by comparing distances obtained on the basis of the word pairs of stimulus language *A* and reader language *B* with the distances obtained on the basis of the word pairs of stimulus language *B* with the reader language *A*. We examined both lexical and

orthographic distances.[2] We did not find asymmetries at the lexical level, but we did find them at the orthographic level. The results are shown in Table 7. The results mean that in a written intelligibility test in which the same stimulus words are used as in this paper, Dutch readers understand German more easily than Germans understand Dutch, Danes will understand Swedish better than Swedes understand Danish, and Germans understand Swedish better than Swedes understand German.

Table 7: Asymmetric relationships within the Germanic language group. Asymmetries are found at the orthographic level only. P values are one-tailed.

Reader	Stimulus	Dist.		Reader	Stimulus	Dist.	*p*
German	Dutch	28%	<	Dutch	German	32%	< 0.001
Danish	Swedish	17%	<	Swedish	Danish	20%	< 0.05
German	Swedish	32%	<	Swedish	German	37%	< 0.05

4.2 Romance

The lexical distances between the Romance languages are given in Table B.1, and the orthographic distances in Table B.2. The averaged distances (see the introduction of this section) are geographically visualized in Figure 3. At the lexical level we find that Portuguese, Catalan and Spanish form a close group. Italian and French are relatively closely connected to this group and to each other (12%/13%). All of these languages belong to the Western Romance language group. Romanian belongs to the Eastern Romance language group. The Eastern Romance languages developed in Southeastern Europe from the local eastern variant of Vulgar Latin. Sometimes Italian is counted among eastern Romance languages as well. Romanian shares fewer lexical items with the Western Romance languages than the Western Romance languages do among each other.

At the orthographic level the smallest distance is found between Spanish and Portuguese (16%/16%). The two languages are relatively strongly related to Catalan and Italian. Romanian is closest to Catalan (31%/30%). French is most distinct within this set of Romance languages. French

2 Given 100 words we compare two sets of 100 word pair distances at the lexical level. Since the distances are either 0% (there is a cognate in the language of the reader) or 100% (there is not a cognate) we used a two-sample proportion test. At the orthographic level the number of word pair distances will maximally be equal to 100 and vary between 0% and 100%. Two sets of word pair distances are compared by means of a paired-samples *t*-test.

belongs to the Gallo-Romance subgroup in the Romance language family. This subgroup also includes Occitan, Catalan, Franco-Provençal, Piedmontese, Lombard, Emiliano-Romagnolo, Ligurian and Rhaeto-Romance. Gallo-Romance languages are generally considered the most innovatory among all the Romance languages, and are as a whole usually characterized by the loss of all unstressed final vowels other than /-a/; most significantly final /-o/ and /-e/ were lost. Examples are: French *enfant* versus Spanish *infante*, French *cas* versus Italian *caso*. On the other hand, modern French has a conservative spelling system which is still based on the pronunciation of Old French and despite changes in the pronunciation during the past centuries, orthography stayed more or less the same since the 12th century. For the other investigated Romance languages orthography reflects pronunciation to a high degree.

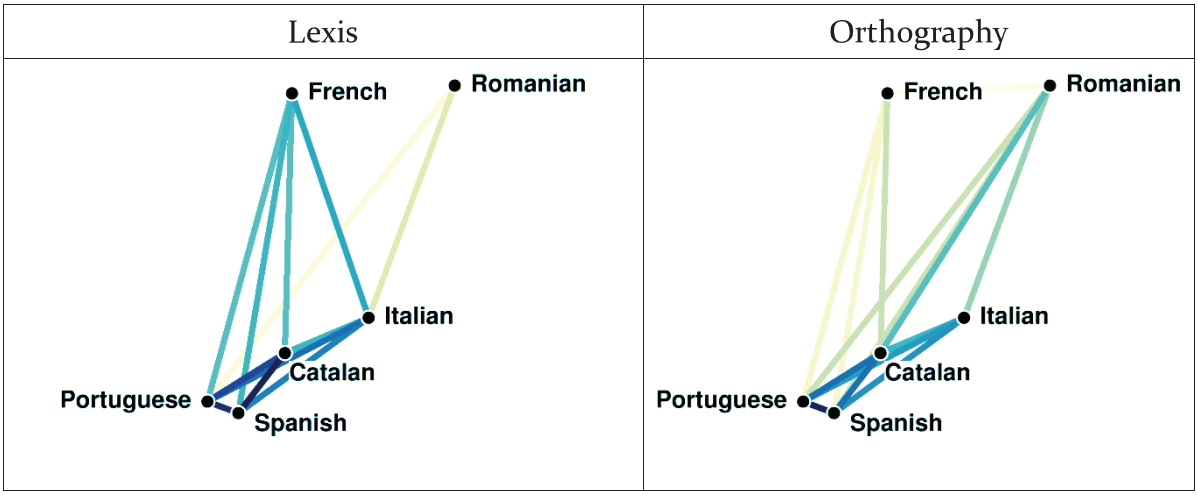

Figure 3: *Lexical and orthographic distances between Romance languages. Darker lines connect lexically/orthographically close varieties, lighter lines more remote ones. The largest distances are represented by white lines which are invisible on the white background. At the lexical level the distances vary from 3% (between Catalan and Spanish) to 26% (between Catalan and Romanian), and at the orthographic level distances vary from 16% (between Portuguese and Spanish) to 47% (between French and Italian).*

We examined whether there are asymmetric relationships within the Romance language group both at the lexical and the orthographic levels. The results are shown in Table 8. At the lexical level we find that Romanians will more easily understand Catalan and Portuguese than the Catalans and the Portuguese will understand Romanian. At the orthographic level

we find that the French will understand more easily Italian and Romanian, than the Italians and Romanians will understand French.

Table 8: Asymmetric relationships within the Romance language group. Asymmetries are found at both the lexical and the orthographic level. P values are one-tailed.

Level	Reader	Stimulus	Dist.		Reader	Stimulus	Dist.	*p*
Lexis	Romanian	Catalan	19%	<	Catalan	Romanian	32%	< 0.05
Lexis	Romanian	Portuguese	18%	<	Portuguese	Romanian	28%	< 0.01
Orth.	French	Italian	42%	<	Italian	French	51%	< 0.001
Orth.	French	Romanian	39%	<	Romanian	French	50%	< 0.01

4.3 Slavic

The lexical distances between the Slavic languages are given in Table C.1, and the orthographic distances in Table C.2. The averaged distances (see the introduction of this section) are geographically visualized in Figure 4. The graphs displaying lexical distances suggest a division in a northern and a southern group. In the northern group there is small distance between Czech and Slovak. Together with Polish they belong to the West Slavic languages. In the southern group the distances are larger. Slovene and Bulgarian are particularly distant from each other (38%/44%). Together with Croatian they belong to the South Slavic languages.

Within our set of Slavic languages, Bulgarian uses the Cyrillic alphabet and the other languages use the Latin alphabet. Since we want to model written intelligibility, we should use the spelling systems *as they are used* for each of the languages and therefore will be used in the web-based written intelligibility test. Both our distance measurements and the scores of the intelligibility test should be obtained on the basis of the same material. Of course, we may predict in advance that the orthographic distance between a language which uses the Cyrillic orthography and a language which uses the Latin alphabet will be relatively large. We find this when we look at the orthographic picture: Bulgarian is completely separated from the 'Latin' languages which apparently are a close group.

We also performed an analysis in which the Cyrillic transcriptions in Bulgarian were replaced by Latin transliterations. Bulgarian is not written in Latin; Latin cannot be used as an alternative as is the case with Serbian. But an analysis on the basis of a Latin transliteration will show how

Bulgarian is related to the other Slavic languages if a Latin orthography would have been adopted.

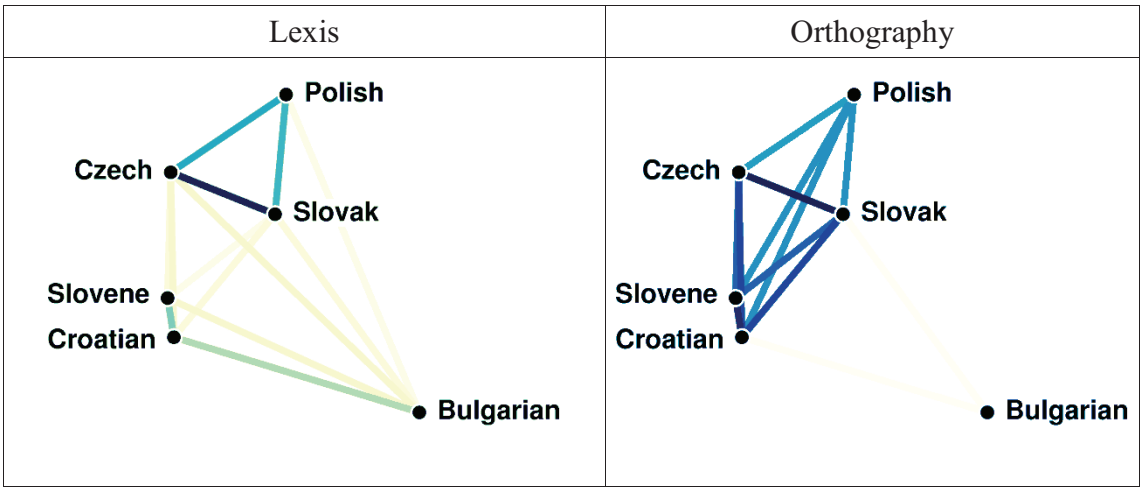

Figure 4: *Lexical and orthographic distances between Slavic languages. Darker lines connect lexically/orthographically close varieties, lighter lines more remote ones. The largest distances are represented by white lines which are invisible on the white background. For Bulgarian the original Cyrillic orthography is used. At the lexical level the distances vary from 4% (between Czech and Slovak) to 48% (between Polish and Slovene), and at the orthographic level distances vary from 11% (between Czech and Slovak) to 67% (between Bulgarian and Polish).*

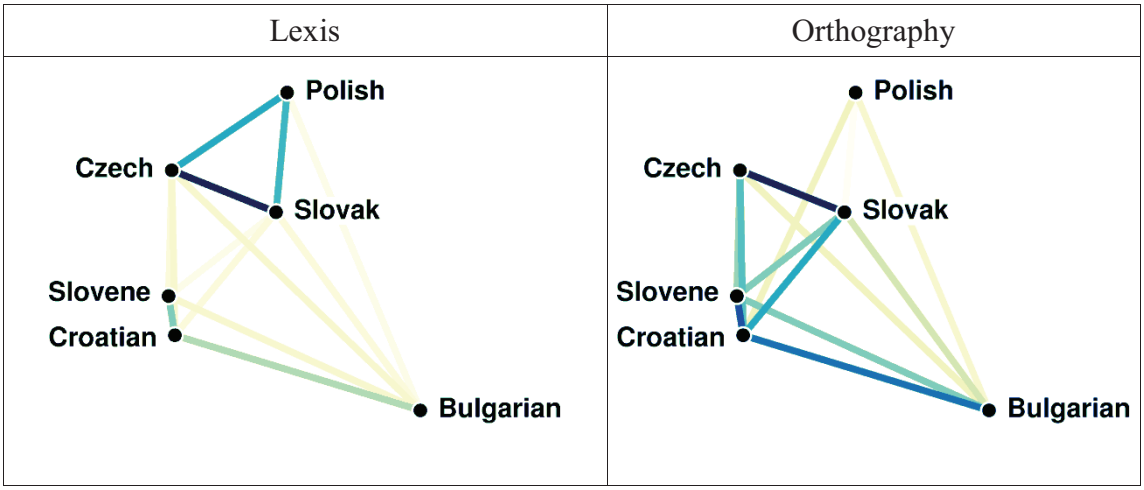

Figure 5: *Lexical and orthographic distances between Slavic languages. Darker lines connect lexically/orthographically close varieties, lighter lines more remote ones. The largest distances are represented by white lines which are invisible on the white background. For Bulgarian the Latin transliteration of the original Cyrillic orthography is used. At the lexical level the distances vary from 4% (between Czech and Slovak) to 48% (between Polish and Slovene), and at the orthographic level distances vary from 11% (between Czech and Slovak) to 32% (between Czech and Polish).*

The transliterations were made with the web application *Translit*.[3] The results can be seen in Figure 5. In this figure Bulgarian is no longer orthographically separated from the other languages; instead, Polish is now found to be very distinct. This may be explained by the fact that Polish is the only language in the Slavic group which uses digraphs. When the Latin alphabet is used in Slavic languages, characters may be modified by diacritics. However, the Polish orthography includes digraphs and trigraphs which are used instead of diacritics, for example *cz* (IPA: /tʃ/), *sz* (IPA: /ʃ/) and *rz (*IPA: /r̪̊/, a simultaneous combination of /r/and /ʒ/). Examples in our data set are: Polish *czas* is *čas* in Slovene, Czech and Slovak; Polish *szkoła* is *škola* in Croatian, Czech and Slovak; Polish *formularz* is *formulář* in Czech.

Within the Slavic group we do not find asymmetric relationships at the lexical level. At the orthographic level we find three asymmetries which are listed in Table 9. Czech readers will understand Polish texts better than Polish readers will understand Czech texts. Slovakians and Slovenes will understand Croatian texts more easily than Croatians will understand Slovakian and Slovene texts.

Table 9: Asymmetric relationships within the Slavic language group. Asymmetries are found at the orthographic level only. P values are one-tailed.

Reader	Stimulus	Dist.		Reader	Stimulus	Dist.	*p*
Czech	Polish	31%	<	Polish	Czech	32%	< 0.05
Slovak	Croatian	19%	<	Croatian	Slovak	20%	< 0.05
Slovene	Croatian	14%	<	Croatian	Slovene	15%	< 0.05

5. Lexical and orthographic distance in relation to each other and to geographic distance

In Section 4 we calculated lexical and orthographic distances between languages for each of the three language groups. In this section we answer the research questions formulated in the introduction. In Section 5.1 we correlate lexical and orthographic variation patterns to each other. In Section 5.2 we answer the question whether lexical and orthographic distances correlate with geographical distances.

3 The application can be found at: http://bg.translit.cc/.

5.1 Do lexical and orthographic distances correlate with each other?

The correlations between the two linguistic levels are shown in Table 10 per language group. For all language groups significant correlations are found, all of them being "medium" correlations (i.e. 0.30 ≤ |r| < 0.50, cf. Cohen 1988).

Table 10: Correlations between lexical and orthographic distances for Germanic, Romance and Slavic language groups. For Slavic languages two analyses are performed, one using the original Cyrillic orthography for Bulgarian (1) and another using a Latin transliteration of Bulgarian (2).

	r	*n*	*p* value
Germanic	0.44	20	0.02
Romance	0.47	30	< 0.01
Slavic (1)	0.34	30	0.03
Slavic (2)	0.42	30	0.01

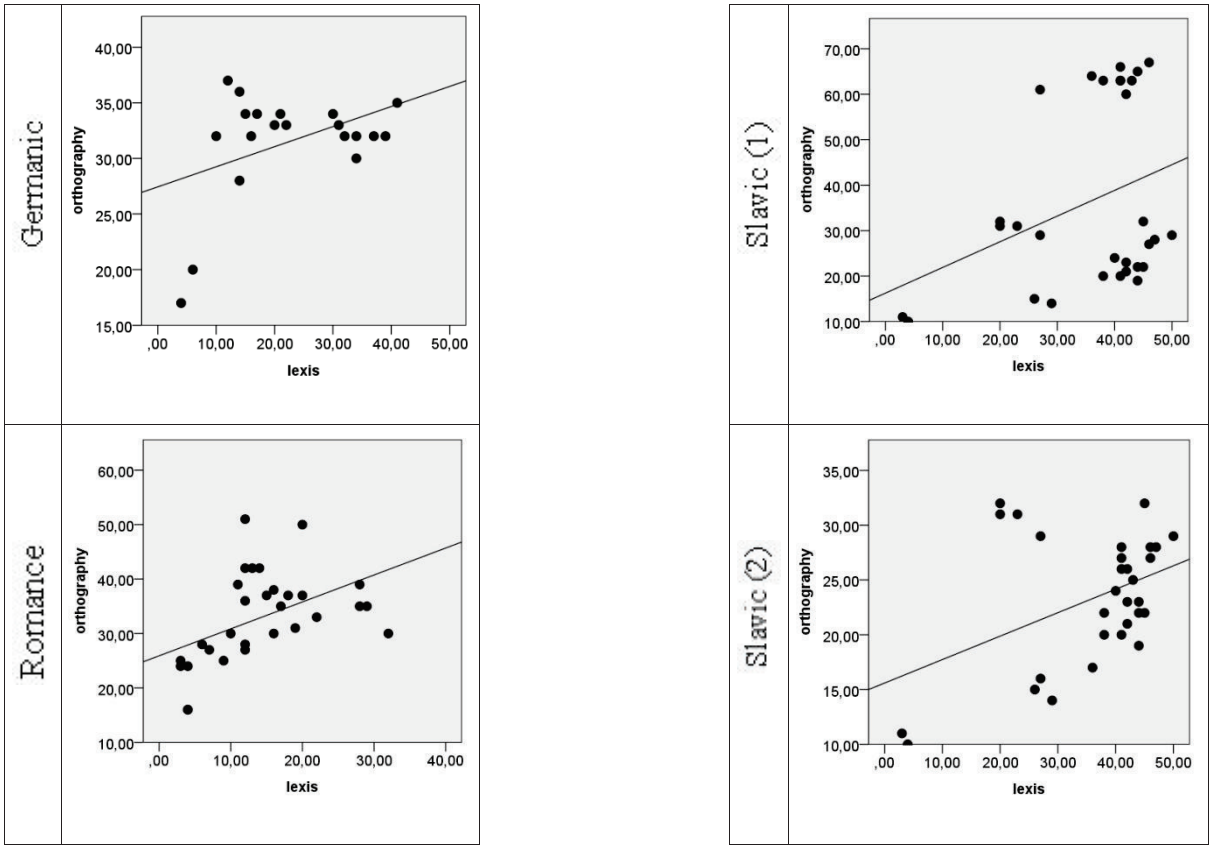

Figure 6: Scatterplots show lexical distances versus orthographic distances for each language group. For Slavic languages two scatterplots are given, one obtained on the basis of measurements using the original Cyrillic orthography for Bulgarian (1) and another using a Latin transliteration of Bulgarian (2). In each scatterplot the linear regression line is drawn.

In Figure 6 for each of the language groups a scatterplot is shown in which lexical distance is plotted against orthographic distance for every language pair. Except for Romance languages the plots do not suggest a linear relationship. We expect that both lexical and orthographic distances are potential predictors of written intelligibility scores. However, in a multiple linear regression analysis of Romance languages, lexical distance may rule out orthographic distance or the other way around due to collinearity.

5.2 Do lexical and orthographic distances correlate with geographic distances?

In Section 1 we mentioned the fundamental dialectological postulate: "Geographically proximate varieties tend to be more similar than distant ones" which was confirmed by Nerbonne & Kleiweg (2007: 154). We suggested the same may apply to (national) languages to some extent. Languages are often based on one or more dialects in the countries or areas where they are spoken. Therefore, we examine whether geographically proximate languages tend to be linguistically more similar than distant ones. In this section we correlate lexical and orthographic distances with geographic distances between the countries or areas where the languages are spoken. We are aware of the fact that orthographic distance is just partly a linguistic variable, since it partly represents pronunciational variation which is the linguistic component, and it partly represents spelling variation which is the non-linguistic component (see Section 1 and 4).

We consider four kinds of geographic distance measurements. First, we measure as-the-crow-flies distances between the geographical centers of the countries (Section 5.2.1) and between the capitals of the countries (Section 5.2.2). Next, we measure travel distances between the geographical centers of the countries (Section 5.2.3) and between the capitals of the countries (Section 5.2.4). In Section 5.2.5. we draw some conclusions.

5.2.1 As-the-crow-flies distances between the geographical centers

Since in many studies dialects are considered the 'language' as spoken in one particular village or town, geographical distances between dialects can simply be measured as the geographical distances between the corresponding villages or towns. When measuring geographic distances between languages it is not that easy because the area is much larger. In this study we measure as-the-crow-flies (or: in-a-beeline) distances between the geographical centers of the countries or areas where the languages are

spoken. When a language is spoken in several countries, we choose the largest state in terms of surface.

The geographical centers of the countries are taken from the NGA GEOnet Names Server (GNS).[4] For Catalan we take the coordinates of the city of Manresa, which is located in the geographic center of Catalonia. We found the coordinates on *Wikipedia*.[5] The centers are shown in Figure 7.

The correlations between geographic and linguistic distances are found in Table 11. In Germanic languages we do not find significant[6] correlations for lexis or for orthography. This also agrees with the pictures in Figure 2. For example, at the lexical level Dutch and Swedish are relatively close. Although the geographical distance between Dutch and English is smaller, the lexical distance is much larger. We find a small distance between Danish and Swedish in the orthographic map. Although the geographical distance between Danish and German is much smaller, the orthographic distance is much larger.

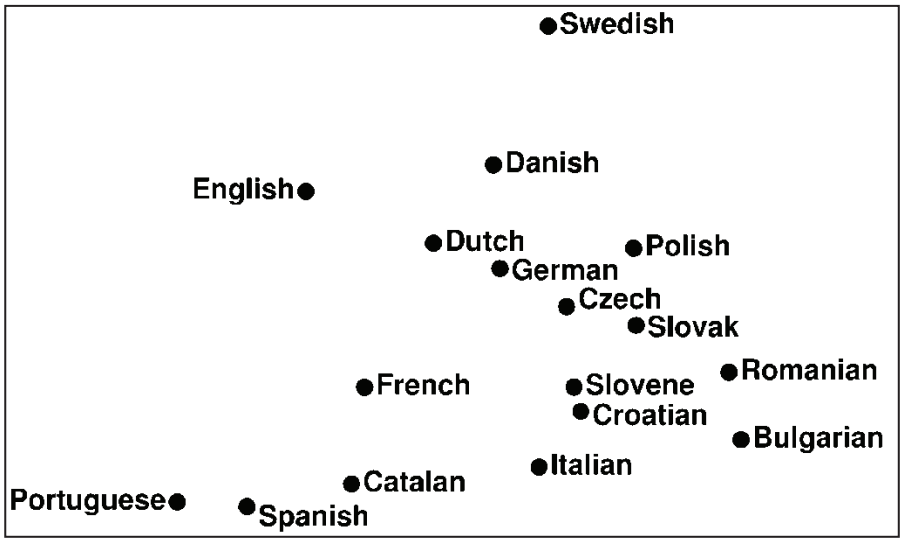

Figure 7: Geographical centers of the countries where the languages in this study are spoken. Catalan is spoken in Catalonia, an autonomous community of Spain with the official status of a "nationality." The x-axis represents the longitude coordinate and the y-axis the latitude coordinate.

4 See: http://earth-info.nga.mil/gns/html.

5 See: http://en.wikipedia.org/wiki/Manresa.

6 The significance of correlations is found by means of the Mantel test throughout this paper (Mantel 1967, see also Heeringa (2004), p. 74-75).

In Romance languages we find a significant correlation at the lexical level, but not at the orthographic level. We find significant correlations in the Slavic languages at both the lexical and the orthographic levels. As to the orthographic measurements, we have to keep in mind that most Slavic languages were inclined to adopt one letter-one sound policy. This means that they are generally read the way they are written, unlike, for example, French or English. Additionally many of the Slavic languages have similar conventions when it comes to diacritics.

We found that the correlation which we obtained when using the Cyrillic spelling (r = .84) is significant higher than the correlation which we obtained when using the Latin transliteration (r = .40) at the alpha = .05 level, but both correlations are significant. When we look at Figure 4 (which is obtained on the basis of an analysis where the Cyrillic spelling is used for Bulgarian), we find both large geographic and large orthographic distances between Bulgarian and the other Slavic languages. But in Figure 5 (where the Latin transliteration is used) the orthographic distances between Bulgarian and the other Slavic languages are relatively smaller, especially between Bulgarian and Croatian, which explains the lower correlation between orthographic and geographic distance.

Table 11: Correlations between lexical and geographical distances, and between orthographic and geographical distances. For Slavic two analyses are performed, one with the original Cyrillic orthography for Bulgarian (1) and another with Latin transliteration (2). Geographic distances are between capitals.

	lexis			orthography		
	r	*n*	*p* value	*r*	*n*	*p* value
Germanic	0.33	20	0.08	0.18	20	0.23
Romance	0.68	30	< 0.01	0.25	30	0.09
Slavic (1)	0.52	30	< 0.01	0.84	30	< 0.01
Slavic (2)	0.52	30	< 0.01	0.41	30	0.01

5.2.2 As-the-crow-flies distances between the capitals

In the previous section we measured geographic distances between the geographic centers of the countries. In Section 1 we mentioned that languages are usually derived from dialects. Instead of using the geographical centers, we measure geographical distances between the capitals of the countries. A standard language often has its roots in the dialects of the economically predominant area of the country, and often – but not always

– the capital is found in that area. Therefore, as an additional measurement, we consider geographic distances between capitals.

We motivate this by referring to the Dutch situation. In the Netherlands the provinces of Noord-Holland, Zuid-Holland and Utrecht are considered economically central. Smakman (2006) shows that this goes back to the 16th century:

> "Holland went on to develop into a strong sea-faring power and flourished economically … By the end of the 16th Century, the dominance of Hollands (the language of Holland) had tacitly been acknowledged across the higher social layers in the northern area." (p. 21)

Regarding the 17th Century he writes:

> "The most influential bourgeoisie of the Dutch-speaking resided in the Holland cities, and the power and wealth of these cities grew steadily. The supremacy of Holland was a fact, the new city varieties gradually started to share certain features … Van den Vondel (1650) wrote about Dutch being spoken most perfectly in the Hague and Amsterdam, by people of good upbringing, and it is true that Dutch pronunciation today is based largely on the speaking habits of the 17th Century wealthy middle classes of Amsterdam and the Hague." (p. 21, 22)

The strong relationship between the prestigious Holland dialects and standard Dutch is also shown by Heeringa (2004). He finds that standard Dutch pronunciation is closest to the pronunciation of the dialect of Haarlem, a smaller city to the west of Amsterdam, the Dutch capital (p. 274).

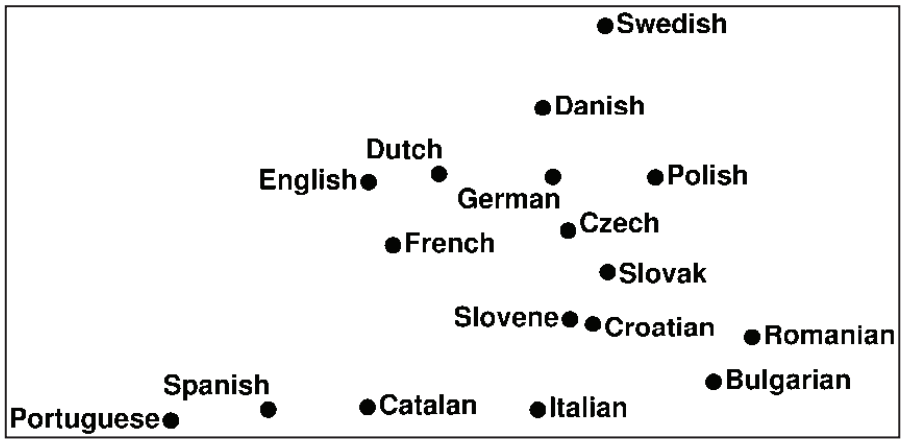

Figure 8: The languages located at the capitals. The x-axis represents the longitude coordinate and the y-axis the latitude coordinate.

Table 12: Capitals of the countries.

Germanic		Romance		Slavic	
Danish	Copenhagen	**Catalan**	Barcelona	**Bulgarian**	Sofia
Dutch	Amsterdam	**French**	Paris	**Croatian**	Zagreb
English	London	**Italian**	Rome	**Czech**	Prague
German	Berlin	**Portuguese**	Lisbon	**Polish**	Warsaw
Swedish	Stockholm	**Romanian**	Bucharest	**Slovak**	Bratislava
		Spanish	Madrid	**Slovene**	Ljubljana

Table 13: Correlations between lexical and geographical distances, and between orthographic and geographical distances. For Slavic two analyses are performed, one with the original Cyrillic orthography for Bulgarian (1) and another with Latin transliteration (2). Geographic distances are between capitals.

	lexis			orthography		
	r	*N*	*p* value	*r*	*n*	*p* value
Germanic	0.36	20	0.06	0.24	20	0.16
Romance	0.70	30	< 0.01	0.33	30	0.04
Slavic (1)	0.46	30	< 0.01	0.81	30	< 0.01
Slavic (2)	0.46	30	< 0.01	0.59	30	< 0.01

In Table 12 the capitals of the countries in our study are listed. The locations of the capitals are shown in Figure 8. The correlation coefficients between geographic distances between the capitals[7] and the lexical and orthographic distances are given per language group in Table 13. When we compare this table with Table 11 we now find a significant correlation of orthographic variation in the Romance languages and also lower *p* values. This indicates that geographic distances between capitals predict lexical and orthographic distances better than geographic distances between geographical centers. But we still do not find significant correlations of Germanic languages at either level. This may indicate that geographic

7 The geographic distances are obtained on the basis of the coordinates of the capitals. We used the coordinates provided by Global Gazetteer Version 2.2 which is available at: http://www.fallingrain.com/world/index.html .

distance is not a significant predictor of lexical and orthographic variation within the Germanic language family.

5.2.3 Travel distances between the geographical centers

Language change and diversity are largely determined by language contact or communication. Trudgill (1974) assumes that geographical distance is a predictor of communication. The closer two locations are geographically located to each other, the stronger the contact between these two locations. In the previous sections we measured geographic distances in the simplest way, namely, as-the-crow-flies distances. However, as-the-crow-flies distances may model language contact badly when countries are separated by mountains or water or other natural barriers or countries. Gooskens (2005) correlated perceptual distances and pronunciation distances between 15 Norwegian dialects with as-the-crow-flies distances and travel distances from around 1900 and from 2000. The perceptual distances were measured by means of an experiment in which high school pupils in each of 15 Norwegian locations listened to dialect recordings of the same 15 locations. They rated the distance to their own dialect in a scale ranging from 1 (dialect is similar) to 10 (dialect is not similar). In that way linguistic distances between the dialects were obtained as perceived by the language users themselves. The perceptual and pronunciation distances correlated significantly higher with the travel distances from 1900 than with the as-the-crow-flies distances. When using the travel distances from 2000 the correlation coefficients did not become significantly higher, probably because modern travel distances are very similar to as-the-crow-flies distances due to the modern infrastructure.

In this section we correlate the lexical and orthographic distances between languages with travel distances between the geographic centers of the corresponding countries. For each geographic center point we found the geographically closest town. The towns are listed in Table 14. Szmrecsanyi (2013) draws on *Google Maps*[8] to obtain non-linguistic correlates of linguistic distance. We also used *Google Maps* in order to find travel distances. The web application calculates the shortest route between two towns when the trip is made by car. Sometimes for a part of the route the car is transported by a ferry or train, especially between England and the continental European countries. The travel distances are given in either miles or kilometers.

8 See: http://maps.google.com/.

Table 14: Towns near the center points in the countries

Germanic		Romance		Slavic	
Danish	Skanderborg	**Catalan**	Manresa	**Bulgarian**	Sevlievo
Dutch	Dronten	**French**	Saint-Georges-la-Pouge	**Croatian**	Slunj
English	Douglas[9]	**Italian**	Campello sul Clitunno	**Czech**	Kácov
German	Bischofferode	**Portuguese**	Abrantes	**Polish**	Bełchów
Swedish	Remman	**Romanian**	Făgăraș	**Slovak**	Hroncok
		Spanish	Toledo	**Slovene**	Žalec

The correlations between the travel distances and the lexical and orthographic distances are given in Table 15. In general the correlations are lower than in Tables 11 and 13, except for the lexical distances between the Germanic languages. It is striking that the lexical distances of each of the language groups significantly correlate with the travel distances, but orthographic distances correlate significantly with the travel distances within the Slavic group only. This gives us the impression that geographical distance determines lexical variation rather than orthographic variation.

Table 15: Correlations between lexical and travel distances, and between orthographic and travel distances. For Slavic two analyses are performed, one using the original Cyrillic orthography for Bulgarian (1) and another using a Latin transliteration of Bulgarian (2). Travel distances are measured between the geographical centers of the countries.

	lexis			orthography		
	R	*n*	*p* value	*r*	*n*	*p* value
Germanic	0.39	20	< 0.05	0.07	20	0.40
Romance	0.48	30	< 0.01	0.03	30	0.43
Slavic (1)	0.46	30	< 0.01	0.69	30	< 0.01
Slavic (2)	0.46	30	< 0.01	0.33	30	0.04

The development of lexical variation and change is probably a more natural process than the existence of orthographic variation. Lexis is a

9 This town is located on the Isle of Man.

fully linguistic variable, driven by the speakers, while orthography is a combination of pronunciation and spelling. Variation in pronunciation is purely linguistic, but spelling is not; it is partly determined by political decisions, where each country is autonomous. Spelling differences do not always reflect pronunciation differences and may cause languages to look more different from each other than they actually are. For example, *sector* is written as *Sektor* in German and as *sector* in Dutch. The c in the Dutch orthographic form is pronounced in the same way as the k in the German orthographic form. Therefore, this difference does not reflect a difference in pronunciation but in spelling. This kind of difference may explain why orthographic variation correlates less well with travel distances than lexical variation does.

5.2.4 Travel distances between the capitals

In Section 5.2.2 we correlated lexical and orthographic distances with travel distances which were measured between the geographical centers of the countries. In Section 5.2.3 the two levels were correlated with as-the-crow-flies distances between the capitals of the countries. In this section we correlate the lexical and orthographic distances between languages with travel distances between the capitals of the corresponding countries. The results are shown in Table 16.

Table 16: Correlations between lexical and travel distances, and between orthographic and travel distances. For Slavic two analyses are performed, one using the original Cyrillic orthography for Bulgarian (1) and another using a Latin transliteration of Bulgarian (2). Travel distances are measured between the capitals.

	lexis			orthography		
	r	*n*	*p* value	*R*	*n*	*p* value
Germanic	0.25	20	0.15	0.26	20	0.13
Romance	0.73	30	< 0.01	0.28	30	0.07
Slavic (1)	0.50	30	< 0.01	0.79	30	< 0.01
Slavic (2)	0.50	30	< 0.01	0.61	30	< 0.01

In most cases we find significant correlations, except for Germanic languages (both levels) and Romance languages (orthography only). In general the correlations are higher than those which we found in Section

5.2.3, which suggests that travel distances should be measured between the capitals rather than between geographical centers.

5.2.5 Summary and findings

In this section we summarize the results of the previous sections over the three language groups. We performed an ANCOVA test for each of the four geographical measures. In each test the geographic distance is entered as covariate and language group is entered as fixed factor. By entering language group as a fixed factor we take into account that systematic differences between language groups may exist. Either lexical distance or orthographic distance is entered as dependent variable.

The partial eta squared values (η^2) for each of the four geographic measures are presented in Table 17. At the lexical level we get the larger η^2 for travel distances between capitals. 13.2% of the variance in the lexical distances is explained by travel distances between capitals. At the orthographic level the larger η^2 is obtained when using as-the-crow-flies distances between capitals. 15.1% (when using the Cyrillic spelling for Bulgarian) or 13.3% (when using the Latin transliteration for Bulgarian) of the variance in the orthographic distances is explained by as-the-crow-flies distances between capitals.

For all geographic measures the η^2's at the lexical level are higher than those at the orthographic level, regardless of whether the Cyrillic spelling or the Latin transliteration is used for Bulgarian.

*Table 17: Partial Eta Squared values for each of the four geographic measures. ANCOVA tests were performed where the geographic measures are entered as covariates and language group as fixed factor. Either lexical distance or orthographic distance is entered as dependent variable. For orthography (1) the Cyrillic spelling is used for Bulgarian, and for orthography (2) the Latin transliteration is used. *** means: p < .001, ** means: p < .01 and * means: p < .05.*

		lexis	orthography (1)	orthography (2)
as-the-crow-flies	centers	0.212***	0.126**	0.072*
as-the-crow-flies	capitals	0.215***	0.151***	0.133**
travel	centers	0.132**	0.028	0.006
travel	capitals	0.220***	0.144**	0.120**

6. Conclusions

In this paper we investigated whether lexical and orthographic distances correlate with each other and with geography.

We correlated lexical and orthographic distances with each other and found significant but relatively low correlations for each of the language families. The relatively low correlations show that lexis and orthography each have their own patterns of variation. We refer to a study of Heeringa & Hinskens (in press), who studied linguistic change in a set of 80 local Dutch dialects. They focused on lexis, morphology and sound components and found that the lexical level has been affected the most, and the morphological level is the most stable. This suggests that dialect change happens at different rates at the various linguistic levels. We expect that the same is the case for languages. When lexical and orthographic distance are examined as explanatory factors of the scores of written mutual intelligibility, both factors should be included in the model.

When correlating lexical and orthographic distances with geography, we were confronted with the problem of how to locate languages in geographic space. We located the countries at their center points on the one hand, and at their capitals on the other hand. We are aware that neither locating at the center points nor locating at the capitals is optimal. A better and more linguistically motivated approach may be to use the geographic distances between the source dialects of the standard varieties.[10] We considered two kinds of geographic measurements, namely, as-the-crow-flies distances and modern travel distances. At the lexical level the best results are obtained when correlating with travel distances between capitals, which explain 22.0% of the variance in the lexical distances. At the orthographic level the best results are obtained when correlating with as-the-crow-flies distances between capitals, which explain 13.3% (for Bulgarian the Cyrillic spelling is used) or 15.1% (a Latin transliteration is used instead) of the variance in the orthographic distances (see Table 15). For both lexical distance and orthographic distance the size of effect is large (Cohen 1988). Therefore, geography could represent lexical and orthographic variation in a model of written intelligibility to some extent. From Table 17 we will expect that geography will represent lexical variation better than orthographic variation.

In future work other linguistic levels – sound components, morphology and syntax – may be added into the analysis.

10 We received this suggestion from an anonymous reviewer.

Acknowledgments

We are grateful to Peter Kleiweg, whose RuG/L04 package was used to create the beam maps shown in this paper. We thank Renée van Bezooijen and three anonymous reviewers for their useful remarks.

Bibliography

Baugh, A. & Cable, T. (2002). *History of the English Language* (5th edition). Upper Saddle River, NJ: Prentice Hall.

Campbell, L. (1995). The Quechumaran hypothesis and lessons for distant genetic comparison. *Diachronica,* XII(2), 157–200.

Chambers, J.K. & Trudgill, P. (1998). *Dialectology*. Cambridge: Cambridge University Press.

Cisouw, M. (to appear). Disentangling geography from genealogy. In P. Auer, M. Hilpert, A. Stukenbrock & B. Szmrecsanyi (eds.). *Space in language and linguistics: Geographical, interactional, and cognitive perspectives*. New York: Walter de Gruyter.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Denham, K. & Lobeck, A. (2009). *Linguistics for everyone: An introduction*. Cengage Learning, Inc.

Doetjes, G. & Gooskens, C. (2009). Skriftsprogets rolle i den dansk-svenske talesprogsforståelse. *Språk och stil*, 19, 105-123.

Dyen, I. (1956). Language distribution and migration theory. *Language,* 32, 611–626.

Giesbers, C. (2008). *Dialecten op de grens van twee talen; een dialectologisch en sociolinguïstisch onderzoek in het Kleverlands dialectgebied*. Ph.D. dissertation, Nijmegen: Radboud University of Nijmegen.

Goebl, H. (1982). *Dialektometrie; Prinzipien und Methoden des Einsatzes der numerischen Taxonomie im Bereich der Dialektgeographie*. (*Philosophisch-Historische Klasse Denkschriften* 157). Vienna: Verlag der Österreichischen Akademie der Wissenschaften. With assistance of W.-D. Rase & H. Pudlatz.

Goebl, H. (1984). *Dialektometrische Studien. Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF*. (*Beihefte zur Zeitschrift für romanische Philologie* 191, 192, 193). Tübingen: Max Niemeyer Verlag.

Goebl, H. (1993). Probleme und Methoden der Dialektometrie: Geolinguistik in globaler Perspektive. In W. Viereck (ed.). *Proceedings of the International Congress of Dialectologists* 1. Stuttgart: Franz Steiner Verlag, 37-81.

Gooskens, C. (2005). Traveling time as a predictor of linguistic distance. *Dialectologia et Geolinguistica*, 13, 38-62.

Heeringa, W. (2004). *Measuring dialect pronunciation differences using Levenshtein distance*. Doctoral dissertation. Groningen: University of Groningen.

Heeringa, W. & Hinskens, F. (in press). Convergence between dialect varieties and dialect groups in the Dutch dialect area. In B. Szmrecsanyi & B. Wälchli (eds.). *Aggregating dialectology and typology: linguistic variation in text and speech, within and across languages* (working title). Series: *Linguae et Litterae: Publications of the School of Language and Literature*, Freiburg Institute for Advanced Studies. Berlin: De Gruyter.

Heeringa, W. & Nerbonne, J. (2006). De analyse van taalvariatie in het Nederlandse dialectgebied: Methoden en resultaten op basis van lexicon en uitspraak. *Nederlandse Taalkunde,* 11(3), 218-257.

Hinskens, F., Auer, P. & Kerswill, P. (2005). The study of dialect convergence and divergence: conceptual and methodological considerations. In P. Auer, F. Hinskens & P. Kerswill (eds.). *Dialect change. The convergence and divergence of dialects in contemporary societies*. Cambridge: Cambridge University Press, 1-48.

Hjort, E. & Kristensen, K. (2003). *Den Danske Ordbog*. Copenhagen: Det Danske Sprog- og Litteraturselskab/Gyldendal. Accessed via http://ordnet.dk/ddo on 16 november 2012.

Inoue, F. (1996). Computational dialectology (2). *Area and Culture Studies*, 53, 115-134.

Kloeke, G.G. (1927). *De Hollandsche expansie in de zestiende en zeventiende eeuw en haar weerspiegeling in de hedendaagsche Nederlandsche dialecten*. Den Haag: Martinus Nijhoff.

Kruskal, J. (1999). An overview of sequence comparison. In D. Sankoff & J. Kruskall (eds.). *Time Warps, String Edits, and Macromolecules. The Theory and Practice of Sequence Comparison*. Stanford: Center for the Study of Language and Information, 1-44.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory,* 10(8), 707-710.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27, 209-220.

Nerbonne, J. & Kleiweg, P. (2007). Toward a dialectological yardstick. *Journal of Quantitative Linguistics,* 14(2), 148-167.

Schmidt, J. (1872). *Die Verwandtschaftsverhältnisse der indogermanischen Sprachen.* Weimar: H. Böhlau.

Schwerdt, J. (2000). *Die 2. Lautverschiebung. Wege zu ihrer Erforschung*. Heidelberg: Winter.

Séguy, J. (1971). La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane*, 35, 335–357.

Séguy, J. (1973). La dialectométrie dans l'Atlas linguistique de la Gascogne. *Revue de Linguistique Romane,* 37, 1–24.

Smakman, D. (2006). *Standard Dutch in the Netherlands: a Sociolinguistic and phonetic description*. Doctoral dissertation. Utrecht: LOT, Netherlands Graduate School of Linguistics.

Szmrecsanyi, B. (2013). *Grammatical variation in British English dialects: a study in corpus-based dialectometry*. Cambridge: Cambridge University Press.

Spruit, M., Heeringa, W. & Nerbonne, J. (2009). Associations among linguistic levels. In J. Nerbonne (ed.). *Lingua*, 119(11), 1624–1642.

Stockwell, R. (2002). How much shifting actually occurred in the historical English vowel shift? In D. Minkova & R. Stockwell (eds.). *Studies in the history of the English language: A millennial perspective*. Berlin: Mouton de Gruyter, 267-281.

Trudgill, P. (1974). Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography. *Language in Society*, 2, 215-246.

Van Bezooijen, R. & Gooskens, C. (2005). How easy is it for speakers of Dutch to understand spoken and written Frisian and Afrikaans, and why? In J. Doetjes & J. van de Weijer (eds). *Linguistics in the Netherlands,* 22, 13-24.

Zulu, P.N., Botha, G. & Barnard, E. (2008). Orthographic measures of language distances between the official South African languages. *Literator: Journal of Literary Criticism, Comparative Linguistics and Literary Studies*, 29(1), 185-204.

Appendix A: Germanic distance matrices

Table A.1: Lexical distances between Germanic languages measured as the percentage of non-cognates

		reader				
		Danish	**Dutch**	**English**	**German**	**Swedish**
stimulus	**Danish**		22	41	21	6
	Dutch	17		37	14	15
	English	30	32		34	31
	German	14	10	39		12
	Swedish	4	20	34	16	

Table A.2: Orthographic distances between Germanic languages measured as the percentage of different letters with Levenshtein distance

		reader				
		Danish	**Dutch**	**English**	**German**	**Swedish**
stimulus	**Danish**		33	35	34	20
	Dutch	34		32	28	34
	English	34	32		30	33
	German	36	32	32		37
	Swedish	17	33	32	32	

Appendix B: Romance distance matrices

Table B.1: Lexical distances between Romance languages measured as the percentage of non-cognates

		reader					
		Catalan	**French**	**Italian**	**Portuguese**	**Romanian**	**Spanish**
stimulus	**Catalan**		15	10	4	19	3
	French	12		12	11	20	12
	Italian	16	13		12	17	7
	Portuguese	9	16	6		18	4
	Romanian	32	28	22	28		29
	Spanish	3	14	12	4	20	

Table B.2: Orthographic distances between Romance languages measured as the percentage of different letters with Levenshtein distance

		reader					
		Catalan	**French**	**Italian**	**Portuguese**	**Romanian**	**Spanish**
stimulus	**Catalan**		37	30	24	31	24
	French	36		51	39	50	42
	Italian	30	42		28	35	27
	Portuguese	25	38	28		37	16
	Romanian	30	39	33	35		35
	Spanish	25	42	27	16	37	

Appendix C: Slavic distance matrices

Table C.1: Lexical distances between Slavic languages measured as the percentage of non-cognates

		reader					
		Bulgarian	**Croatian**	**Czech**	**Polish**	**Slovak**	**Slovene**
stimulus	**Bulgarian**		27	41	41	43	38
	Croatian	36		42	47	44	29
	Czech	41	38		20	4	44
	Polish	46	46	23		27	50
	Slovak	42	41	3	20		45
	Slovene	44	26	40	45	42	

Table C.2: Orthographic distances between Slavic languages measured as the percentage of different letters with Levenshtein distance. For Bulgarian the original Cyrillic orthography is used.

		reader					
		Bulgarian	**Croatian**	**Czech**	**Polish**	**Slovak**	**Slovene**
stimulus	**Bulgarian**		61	63	66	63	63
	Croatian	64		21	28	19	14
	Czech	63	20		32	10	22
	Polish	67	27	31		29	29
	Slovak	60	20	11	31		22
	Slovene	65	15	24	32	23	

Table C.3: Orthographic distances between Slavic languages measured as the percentage of different letters with Levenshtein distance. For Bulgarian a Latin transliteration is used.

		reader					
		Bulgarian	**Croatian**	**Czech**	**Polish**	**Slovak**	**Slovene**
stimulus	**Bulgarian**		16	26	28	25	22
	Croatian	17		21	28	19	14
	Czech	27	20		32	10	22
	Polish	28	27	31		29	29
	Slovak	26	20	11	31		22
	Slovene	23	15	24	32	23	