A quantitative approach to social and geographical dialect variation

Wieling, Martijn

Martijn Wieling

A Quantitative Approach *to Social and Geographical* Dialect Variation

A Quantitative Approach to Social and Geographical Dialect Variation

Martijn Wieling

university of groningen

CLCG
Center for Language and Cognition Groningen

bcn

The work in this thesis has been carried out under the auspices of the Research School of Behavioural and Cognitive Neurosciences (BCN) and the Center for Language and Cognition Groningen (CLCG). Both are affiliated with the University of Groningen.

Groningen Dissertations in Linguistics 103
ISSN: 0928-0030
ISBN: 978-90-367-5521-4
ISBN: 978-90-367-5522-1 (electronic version)

Document prepared with LaTeX 2ε and typeset by pdfTeX (Minion Pro font)
Cover design: Esther Ris — www.proefschriftomslag.nl
Printed by: Off Page, Amsterdam, The Netherlands — www.offpage.nl

RIJKSUNIVERSITEIT GRONINGEN

A Quantitative Approach to Social and Geographical Dialect Variation

Proefschrift

ter verkrijging van het doctoraat in de
Letteren
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. E. Sterken,
in het openbaar te verdedigen op
donderdag 28 juni 2012
om 12.45 uur

door

Martijn Benjamin Wieling

geboren op 18 maart 1981
te Emmen

Promotores:	Prof. dr. ir. J. Nerbonne Prof. dr. R.H. Baayen
Beoordelingscommissie:	Prof. dr. J.K. Chambers Prof. dr. M.Y. Liberman Prof. dr. D. Speelman

To Aafke

Acknowledgements

During this enjoyable dissertation-writing journey I have had the privilege of being supported by many people. Not only by fellow-researchers, who helped me with the subjects covered in this thesis, but also by family and friends, who allowed me to keep in touch with the 'real world'. I am very grateful to all of them.

First and foremost, I thank John Nerbonne, who has been my PhD supervisor since the start of this project. We did not have many scheduled appointments, but as his door was always open (and his office opposite mine) there were few days when I did not ask his advice or discuss research ideas with him. I am certain this is one of the main reasons why I was able to write this thesis in only forty months. I am also thankful to John for allowing me to shape my own research project, and for not insisting that I conduct the research we initially envisioned. In addition, I thank Harald Baayen for agreeing to become my second supervisor in 2011. Since my visit to him in Edmonton in 2010, we have collaborated on several projects and I am really grateful for his patient explanations of various statistical techniques and his useful feedback, which were essential for much of the work presented in this dissertation.

Alongside my supervisors, there are several other people I am obliged to. I would like to offer my sincere thanks to the members of my reading committee: Jack Chambers, Mark Liberman and Dirk Speelman, for finding the time in their busy schedules to read this dissertation. Furthermore, I thank both Gerwin Blankevoort and Bob Shackleton for agreeing to be my paranymphs. As well as enjoying his company in both the Netherlands and Canada, I benefitted greatly from Bob's knowledge of English dialects during our collaboration. In addition, the antique book on English dialects (Ellis, 1889) which he gave me proved very useful when writing the introduction to this dissertation. Gerwin has been a good friend for many years, and sharing experiences about both our PhDs (and more important matters) made for many enjoyable lunches at Restaurant Academia. I have also enjoyed the pleasant evenings with him and his beautiful family (Ellen and Lise) very much.

During my PhD I have had the privilege of collaborating with many colleagues. I thank Esteve Valls for two intense weeks of research on Catalan dialects in Santiago de Compostela. I am very grateful for his hospitality during that time, including the (for a Dutchman) unique experience of having dinner at midnight. I also enjoyed collaborating with Simonetta Montemagni during

my two weeks in Pisa. I had a very enjoyable time sharing an office with her at the *Consiglio Nazionale delle Ricerche* and have learned much about Tuscan dialectology (as well as about the delicious Italian food). Finally, I thank Clive Upton for a very nice one-week collaboration in Leeds where we investigated contemporary English dialect variation (and I learned the diverse meanings of English 'tea').

There were various people in the Netherlands with whom I collaborated. I thank Boudewijn van den Berg, Jelke Bloem, Charlotte Gooskens, Wilbert Heeringa, Bob de Jonge, Hanneke Loerts, Eliza Margaretha, Kaitlin Mignella, Sandrien van Ommen, Jelena Prokić and Mona Timmermeister for their friendly collaboration. In particular, I thank Hanneke Loerts for involving me in the analysis of her eye-tracking data. I enjoyed our collaboration and interesting discussions very much.

During my PhD, I have met many nice people at this university. First of all I want to thank all current and former members of the Alfa-Informatica corridor for creating such a welcoming and unique atmosphere. I thoroughly enjoyed all social events, including Sinterklaas parties and Wii-nights. Thank you, Barbara, Çağrı, Daniël, Dicky, Dörte, Erik, Geoffrey, George, Gertjan, Gideon, Gosse, Harm, Hartmut, Henny, Ismail, Jelena, Johan, John K., John N., Jörg, Jori, Kilian, Kostadin, Leonie, Lonneke, Marjoleine, Noortje, Nynke, Peter M., Peter N. (also for your company in Uganda!), Proscovia, Tim, Valerio and Yan. Specifically, I want to thank Daniël for being my office mate for all these years. Despite us having very different opinions about a certain type of fruit, I really appreciated our conversations and his willingness to help whenever needed.

In addition, I have met many other friendly and interesting colleagues by visiting (or organizing) various CLCG activities. Alexandra, Diana, Gisi, Hanneke, Ildikó, Jacolien, Karin, Myrte, Ryan and Veerle are only a few of the people whose company made life in the CLCG as a PhD student very pleasant. In particular, I want to thank Myrte for our frequent lunches and coffee breaks. I've always enjoyed these very much.

Within the CLCG, I would like to thank a few other people. First, I thank Wyke van der Meer who always helped with any issues of an administrative nature. Second, I am grateful to Jonas Bulthuis and especially Adri Mathlener for responding quickly to any IT-related problems I encountered. Finally, I thank Peter Kleiweg for creating the L04 dialectometry software package, which has been very useful in my research.

During my PhD, I followed the training program of the Research School of Behavioural and Cognitive Neurosciences. I really appreciated the interdisciplinary nature of BCN, which allowed me to learn of interesting research outside of linguistics. The BCN Project Management Course (led by Erik Boddeke and Frans Cornelissen) was especially helpful in making me realize I needed to focus my research more. I also want to thank Janine Wieringa and Diana Koopmans (among others) who always made sure every BCN activity was perfectly organized.

In the past seven years I have been coordinating the *Examentraining VWO* of the University of Groningen. Being involved with this project enabled me to take a break from doing research whenever needed. Besides all teachers and other employees throughout the years, I thank Frits van Kouwenhove, Jan Batteram, Rashudy Mahomedradja, Maaike de Lange and Liesbeth Kerkhof for their pleasant collaboration.

During my time at the university I have met many people, both in- and outside academia, who I would like to thank for their support and interest. I thank Bianca, Dirk, Gerwin and Ellen, Hanna, Karin and Arjen, Laurens, Rachel, Rudolf and Ilse, and Vincent Paul and Lianne for their support and friendship throughout the years. In addition, I thank Adriaan, Alie, Arnold, Bill, Bart and Jolanda, Boyana, Carleyn and Richard, Esther, Gerard, Ira, Jack G., Jan de W., Jos, June, Lisa Lena, Louwarnoud, Maarten, Marie-Anne and Berry, Marieke, Markos, Marten, Michael, Sebastiaan, Vivian, Wilmer and Mandy, and all those that I might have forgotten, for their interest and nice conversations.

Besides friends, there is family. First and foremost I want to thank my loving mother and father for always having stimulated me to learn as much as possible. While my mother unfortunately passed away in 2004, I am grateful for my father's support in all these years. I also thank my dear sister Sarah for being there. I greatly respect that she has become such a compassionate person and a uniquely gifted teacher. In addition, I thank Joko, Sarah's soon-to-be husband. I've enjoyed his company and his humor very much. For her interest and unsalted opinion, I am grateful to Hennie, who has in a way also become part of our family. Having a partner in life also implies having a second family. I consider myself very lucky with Manny, Benn, William, Lysette, Maaike and Keimpe, and always enjoy spending time with them at their farm 'Ikemaheerd' in Kloosterburen. While not mentioning each of them personally, I also thank my other relatives for their interest and support.

My final word of thanks is reserved to my love, Aafke. For your unending understanding of all my flaws, your support and your affection, I dedicate this dissertation to you.

Groningen, May 5, 2012

Contents

List of Tables

List of Figures

Part I

Introduction

CHAPTER 1

INCREASING THE DIALECTOLOGY IN DIALECTOMETRY

This dissertation revolves around dialectology, the study of dialects. Dialectology is an attractive field to study, as dialectal speech signals the provenance of the speaker and allows us some insight in the speaker's social environment. Furthermore, similarity between different dialects are indications of contact, both in the present and in the past. Perhaps the attractiveness of studying dialects is best explained by Walt Wolfram and Natalie Schilling-Estes (2006, p. 20):

> "[O]ur natural curiosity is piqued when we hear speakers of different dialects. If we are the least bit interested in different manifestations of human behavior, then we are likely to be intrigued by the facets of behavior revealed in language."

This thesis attempts to extend the remit of dialectometry within dialectology. Dialectometry is a subfield of dialectology and dialectometrists have focused on a selection of the problems studied in dialectology. Dialectologists generally have characterized variation by identifying a small set of varying linguistic features in connection with the (social) characteristics of the speakers or with their geographical location (i.e. dialect geography). Dialectometry was developed as a quantitative method to investigate dialectal variation more objectively, by measuring dialect distances on the basis of (i.e. aggregating over) hundreds of linguistic features (instead of only a few). In contrast to researchers in social dialectology, dialectometrists have focused on dialect geography and have mainly disregarded the (social) characteristics of the speakers. The lack of focus on the social dimension, and the problem of identifying individual features in the results of an aggregate analysis have resulted in harsh criticism of the dialectometric approach by linguists (Schneider, 1988; Woolhiser, 2005). The main purpose of this thesis, therefore, is to introduce and evaluate several methods which combine the merits of dialectometry and social dialectology, and which should be attractive to researchers in both fields. First, however, we will give a brief introduction to dialectology and dialectometry. A much more detailed introduction to dialectology is provided by Chambers and Trudgill (1998) and Niebaum and Macha (2006), while dialectometry is covered in more detail by Goebl (2006) and Nerbonne (2009).

1.1 Dialectology

Dialectology originated as dialect geography in which the relationship between dialectal variants and their geographic location was documented using maps (Chambers and Trudgill, 1998, Ch. 2). Later dialectologists also took social factors into account by embracing a sociolinguistic perspective.

1.1.1 Dialect geography

The first systematic approach to investigate dialect geography is often attributed to Georg Wenker, who conducted a large-scale dialect survey among almost 45,000 schoolmasters starting in 1876 (e.g., Chambers and Trudgill, 1998, p. 15). The schoolmasters were asked to write down forty sentences in their local dialect, obviously resulting in an enormous amount of data. Wenker published only two sets of maps based on a limited set of features and geographical locations (Wenker, 1881). However, a larger part of Wenker's work was used in the *Deutscher Sprachatlas* which was published between 1927 and 1956 (Wrede et al., 1927–1956). More information about this atlas is provided by Niebaum and Macha (2006, Ch. 2).

In contrast to common belief, however, Alexander Ellis was the first to start a large-scale systematic dialect survey (even though it was published later). In his study, which already began in 1868 (Ellis, 1889, p. xviii), Ellis obtained information on English dialects in 1145 locations. Initially, he used several small paragraphs of text which had to be 'translated' into dialectal speech. In order to better obtain the desired idioms, he later employed a word list containing 971 separate words. A third approach using only 76 words, dubbed the 'dialect test', was specifically used to detect dialectal sound variation. All three approaches were indirect, in the sense that they relied on orthographically represented dialectal pronunciations by "educated people who did not speak dialect naturally, and hence had only more or less observed what was said, and imitated it as well as they could" (Ellis, 1889, p. 3). Fortunately, this method of data collection was enriched by fieldwork in which "old and if possible illiterate peasants" were interviewed (Ellis, 1889, p. 4). These pronunciations were transcribed following the 'dialectal paleotype' (a precursor of the International Phonetic Alphabet) devised by Ellis (1889, pp. 76*-88*). Based on Ellis' extensive data collection and analysis, he mapped the main dialectal divisions in England, Wales and Scotland. As an illustration of dialect geography, Figures 1.1 and 1.2 on pages 6 and 7 show the dialect areas (and borders) identified by Ellis in England and Wales, and Scotland, respectively.

While dialect geography was very popular in the first half of the 20th century, the interest in the field diminished around the 1950s. In the 1980s however, interest resurfaced. Chambers and Trudgill (1998, pp. 20–21) attribute this to the technological advancements which enabled the improved analysis of the

large amounts of data available, but also to the positioning of dialectology in a conceptual framework provided by sociolinguistics.

1.1.2 Social dialectology

When dialectologists recognized that focusing on the geographical dimension excluded attention to social factors, they started to take into account the social dimension. This new research field, sociolinguistics, was pioneered by William Labov. In his study on vowel use on Martha's Vineyard (Labov, 1972), he showed that differences in vowel pronunciation were related to social factors (i.e. a native versus tourist identity in Martha's Vineyard). In another well-known study, Labov (1966) showed that the type of /r/ pronunciation depended on the social stratification of the speakers (in three different department stores, catering to different layers of society). Since then, many researchers have investigated the influence of social factors on language variation. For example, Trudgill (1974b) showed that women used prestigious standard forms more frequently than men (in British English in Norwich), while Milroy and Margrain (1980) showed that the frequency of use of non-standard phonological features correlated with the social network of the speakers (in Belfast English). Obviously the new interest in social variation also entailed rethinking methods. Pre-sociolinguistic research used informants who varied as little as possible socially, the so-called non-mobile, older, rural males (NORMs). Once social variation was the focus, informants necessarily varied in age, gender, status, urban vs. rural residence, or background.

1.1.3 Individual linguistic variables

Modern dialectologists have generally focused on studying individual linguistic variables. For example, the type of /r/ pronunciation in the aforementioned study of Labov (1966), or the variability in /ŋg/ pronunciation across gender (Trudgill, 1974b). There are several reasons for focusing on individual linguistic variables as opposed to aggregating over many different variables. First, it allows researchers to study linguistic structure in variation. For example, Labov et al. (1972) reported a series of innovations in the vowels of English spoken in urban centers in the northern part of the U.S. (e.g., Chicago, Detroit, etc.), dubbed the Northern Cities Shift (i.e. a rotation of the low and mid vowels). Furthermore, much variation is phonemically organized. For instance, Kloeke (1927) (see also Bloomfield, 1933) found that the phoneme /ui/ (pronounced [uː], [yː], [øː] or [øy]) in the Dutch words *huis*, 'house', and *muis*, 'mouse' generally shared a single pronunciation in one location (although not exclusively). Finally, studying individual linguistic variables enables researchers to investigate perceived salient differences between dialects or social groups. For example, Van de Velde et al. (2010) investigated one of the most salient differences between speakers in the north and south of the Netherlands, the soft versus

Figure 1.1. Dialect division of England and Wales. The striped/dashed red lines indicate the main dialectal borders. The thick and thin red lines indicate major and minor dialect areas. The original image was taken from Ellis (1889).

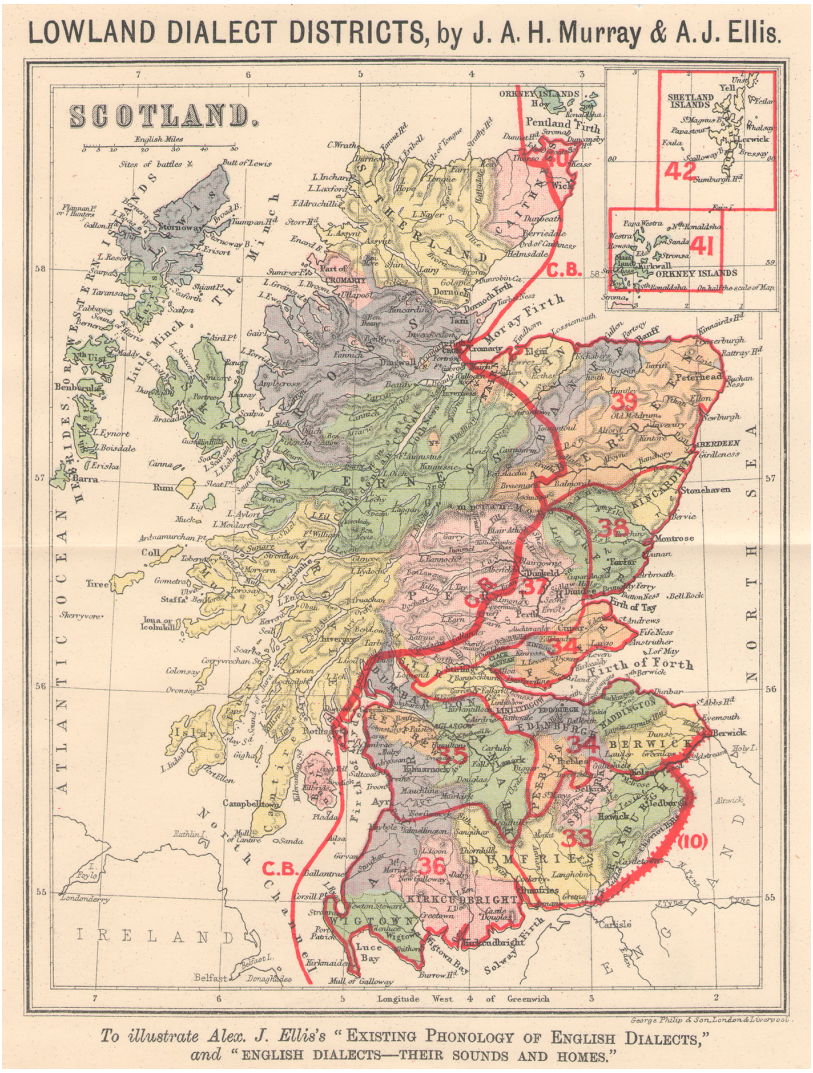


Figure 1.2. Dialect division of Scotland. The striped/dashed red lines indicate the main dialectal borders. The thick and thin red lines indicate major and minor dialect areas. The original image was taken from Ellis (1889).

hard /ɡ/. In the south of the Netherlands a palatal fricative is commonly used, while the north of the Netherlands uses a voiceless velar (or uvular) fricative instead.

1.2 Dialectometry

Dialectometry focuses on characterizing dialects by measuring their distances on the basis of a large number of features. This approach was pioneered by Jean Séguy (1971), whose goal was to find a more objective way of revealing dialect differences. Since then many researchers have embraced the dialectometric approach, including Goebl (1984), Nerbonne et al. (1996), Kretzschmar (1996), Heeringa (2004) and Szmrecsanyi (2011).

In his paper, Séguy (1971) calculated the linguistic distance between pairs of sites by counting the number of items (out of 100) for which the dialects used a different lexical form. By aggregating in this way over a large set of data, one prevents "cherry picking" the features which confirm the analysis one wishes to settle on (Nerbonne, 2009). Later, Séguy (1973) also took other types of linguistic items into account (i.e. with respect to pronunciation, phonology, morphology and syntax) in determining the linguistic distance between dialect pairs. Séguy (1973) visualized these dialect distances by drawing a line with a number (indicating the linguistic distance) between all pairs of neighboring sites. Goebl (1984) independently took a similar approach to that of Séguy, but he measured similarity instead of distance and proposed to weight infrequent items more heavily than frequent items. He also introduced more advanced mapping techniques (see Goebl, 2006 for an overview).

In contrast to the binary same-different distinctions as proposed by Séguy and Goebl, it is also possible to calculate more sensitive distances per item. When investigating transcribed pronunciations, the Levenshtein distance (Levenshtein, 1965) yields the minimum number of insertions, deletions and substitutions to transform one pronunciation into the other (and consequently also results in an alignment of corresponding sound segments). In this way pronunciations which only differ slightly (e.g., in a single sound) will be closer than those which differ more (e.g., in multiple sounds). Kessler (1995) was the first to use the Levenshtein distance for comparing dialectal pronunciations, and the method has been used frequently in dialectometry since then (e.g., see Heeringa, 2004 and Nerbonne, 2009). The Levenshtein distance (and its underlying sound alignments) also plays an important role in this thesis and is introduced in Chapter 2. An example of a dialectometric visualization identifying different dialect areas (in the Netherlands and Flanders) on the basis of aggregate Levenshtein distances is shown in Figure 1.3. The Dutch dialect data underlying this visualization is discussed in Chapter 4.

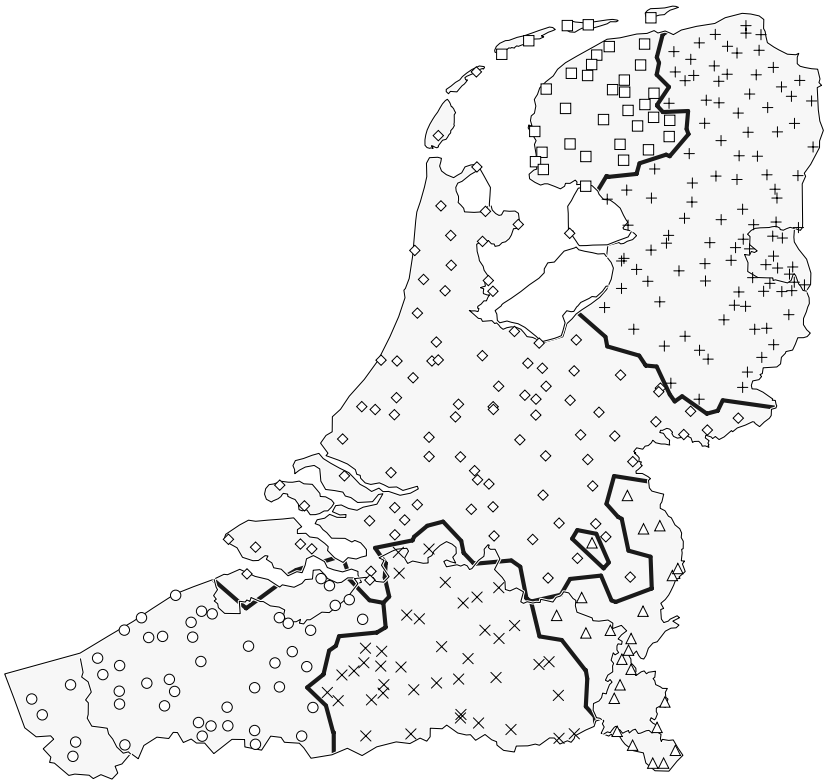


Figure 1.3. Visualization of six Dutch dialect areas. The image is based on Dutch dialect data available in the online dialect analysis application Gabmap (http://www.gabmap.nl; Nerbonne et al., 2011).

1.3 A new dialectometric approach

It is not surprising that modern social dialectologists have not embraced the dialectometric approach. Dialectometry mainly focuses on dialect geography, while generally disregarding social factors. There are, of course, some exceptions. For example, Montemagni et al. (accepted) and Valls et al. (accepted) considered speaker age, and Valls et al. (accepted) investigated the difference between urban and rural communities. However, they evaluated the effect of these social factors by visually comparing maps, as opposed to statistically testing the differences. Furthermore, the aggregation step, central in many dialectometric studies, effectively hides the contribution of individual linguistic variables (which are important in dialectology).

In this thesis, we proceed from the dialectometric view that investigating

a large set of data will provide a more reliable view of linguistic variation than selecting only a small set of linguistic features. We also follow dialectometry in calculating dialect distances (mainly on the basis of pronunciation data). In line with social dialectology, however, the methods we propose enable a focus on individual linguistic features, and also allow us to add a social perspective to dialectometry by taking various sociolinguistic factors into account.

The traditional dialectometric focus on aggregate differences freed its practitioners from the obligation of discriminating between linguistic items very finely. Since this dissertation aims to extend dialectometry to phenomena below the aggregate level (i.e. at the word and sound segment level), we need a more sensitive measure of difference between linguistic items. We therefore start by evaluating several methods to obtain pronunciation distances in Chapter 2. Based on those results we opt for an adapted Levenshtein distance algorithm employing automatically determined sensitive sound segment distances (i.e. similar sounds, such as [o] and [u] will be assigned a low distance, while the distance between [o] and [i] will be large). Chapter 3 evaluates these sound segment distances in detail, by comparing them to acoustic vowel distances.

In contrast to simply distinguishing dialect areas on the basis of aggregate dialect distances, Chapters 4 and 5 introduce and evaluate a novel method to determine geographical dialect areas while *simultaneously* yielding the linguistic basis (in terms of sound segment correspondences). Chapter 4 reports the results on the basis of Dutch dialect data, while Chapter 5 discusses the geographical clustering and linguistic basis of English dialects.

The final part of this thesis introduces a novel regression approach which allows us to predict linguistic distances (per word) with respect to a certain reference point (in our case, the standard language) on the basis of geography, but also various social factors. In addition, the benefit of considering a large set of words is that it enables us to investigate the influence of word-related factors (such as word frequency), which is especially interesting in the light of lexical diffusion (Wang, 1969; see also Chapter 6). We illustrate the usefulness of the regression approach by applying it to a Dutch, a Catalan and a Tuscan dialect dataset in Chapters 6, 7 and 8, respectively.

In summary, the main contribution of this dissertation is to integrate both social factors and a focus on individual linguistic features in the dialectometric approach, and consequently increasing the dialectology in dialectometry.

Part II

Obtaining linguistically sensitive pronunciation and sound distances

Chapter 2

Improving pronunciation alignments

Abstract. Pairwise string alignment is an important general technique for obtaining a measure of similarity between two strings. The current chapter focuses on introducing and evaluating several pairwise string alignment methods at the alignment level instead of via the distances it induces. About 3.5 million pairwise alignments of phonetically transcribed Bulgarian dialect pronunciations are used to compare four algorithms with a manually corrected gold standard. The algorithms include three variants of the Levenshtein distance algorithm, as well as the Pair Hidden Markov Model. Our results show that while all algorithms perform very well and align around 95% of all sequence pairs correctly, there are specific qualitative differences in the (mis)alignments of the different algorithms. Due to its good performance, efficiency and intuitive interpretation, the Levenshtein distance algorithm using automatically determined sound segment distances is our method of choice.[1]

2.1 Introduction

Our cultural heritage is not only accessible through museums, libraries, archives and their digital portals, it is alive and well in the varied cultural habits practiced today by the various peoples of the world. To research and understand this cultural heritage we require instruments which are sensitive to its signals, and, in particular sensitive to signals of common provenance. The present chapter focuses on speech habits which even today bear signals of common provenance in the various dialects of the world's languages, and which have also been recorded and preserved in major archives of folk culture internationally. We present work in a research line which seeks to develop digital instruments capable of detecting common provenance among pronunciation habits, focusing in this chapter on the issue of evaluating the quality of these instruments.

Pairwise string alignment (PSA) methods, such as the popular Levenshtein distance algorithm (or in short, Levenshtein algorithm; Levenshtein, 1965)

[1] This chapter is based on Wieling, Prokić and Nerbonne (2009) and Wieling and Nerbonne (2011b).

which uses insertions (alignments of a segment against a gap), deletions (alignments of a gap against a segment) and substitutions (alignments of two segments), often form the basis of determining the distance between two strings. Since there are many alignment algorithms, and specific settings for each algorithm influencing the distance between two strings (Nerbonne and Kleiweg, 2007), evaluation is very important in determining the effectiveness of these methods.

As indicated in Chapter 1, determining the distance (or similarity) between two phonetic strings is an important aspect of dialectometry, and alignment quality is important in applications in which string alignment is a goal in itself. For example, when determining if two words are likely to be cognate (Kondrak, 2003), detecting confusable drug names (Kondrak and Dorr, 2006), or determining whether a string is the transliteration of a name from another writing system (Pouliquen, 2008).

In this chapter we evaluate string distance measures on the basis of data from dialectology. We therefore explain a bit more of the intended use of the pronunciation distance measure.

Dialect atlases normally contain a large number of pronunciations of the same word in various places throughout a language area. All pairs of pronunciations of corresponding words are compared in order to obtain a measure of the aggregate linguistic distance between dialectal varieties (Heeringa, 2004). It is clear that the quality of the measurement is of crucial importance.

Almost all evaluation methods in dialectometry focus on aggregate results and ignore the individual word pair distances and individual alignments on which the distances are based. The focus on the aggregate distance of 100 or so word pairs effectively hides many differences between methods. For example, Heeringa et al. (2006) find no significant differences in the degrees to which several pairwise string distance measures correlate with perceptual distances when examined at an aggregate level. Wieling et al. (2007b) and Wieling and Nerbonne (2007) also report almost no difference between several PSA algorithms at the aggregate level. As it is important to be able to evaluate the different techniques more sensitively, this chapter examines alignment quality at the segment level.

Kondrak (2003) applies a PSA algorithm to align words in different languages in order to detect cognates automatically. Exceptionally, he does provide an evaluation of the string alignments generated by different algorithms. But he restricts his examination to a set of only 82 gold standard pairwise alignments and he only distinguishes correct and incorrect alignments and does not look at misaligned phones.

In the current chapter we introduce and evaluate several alignment algorithms more extensively at the alignment level. The algorithms we evaluate include the Levenshtein algorithm (with syllabicity constraint, which is explained below), which is one of the most popular alignment methods and has successfully been used in determining pronunciation differences in phonetic strings

(Kessler, 1995; Heeringa, 2004). In addition, we look at two adaptations of the Levenshtein algorithm. The first adaptation includes the swap-operation (Wagner and Lowrance, 1975), while the second adaptation includes sound segment distances which are automatically generated by applying an iterative pointwise mutual information (PMI) procedure (Church and Hanks, 1990). Finally, we include alignments generated with the Pair Hidden Markov Model (PHMM) as introduced to language studies by Mackay and Kondrak (2005). The PHMM has also successfully been applied in dialectometry by Wieling et al. (2007b).

2.2 Material

The dataset used in this evaluation consists of 152 words collected from 197 sites equally distributed over Bulgaria. The transcribed word pronunciations include diacritics and suprasegmentals (e.g., intonation). The total number of different phonetic types (or segments) is 98. The dataset was analyzed and discussed in detail by Prokić et al. (2009a) and is available online at the website http://www.bultreebank.org/BulDialects.

The gold standard pairwise alignments were automatically generated from a manually corrected gold standard set of multiple alignments (i.e. alignments of more than two pronunciations; see Prokić et al., 2009) in the following way:

- Every individual string (including gaps) in the multiple alignment was aligned with every other string of the same word. With 152 words and 197 sites and in some cases more than one pronunciation per site for a certain word, about 3.5 million pairwise alignments were obtained.
- If a resulting pairwise alignment contained a gap in both strings at the same position (a gap-gap alignment), these gaps were removed from the pairwise alignment. We justify this, reasoning that no alignment algorithm may be expected to detect parallel deletions in a single pair of words. There is no evidence for this in the single pair.

As an illustration, consider the multiple alignment of three Bulgarian dialectal variants of the word 'I' (as in 'I am'):[2]

j	ɑ	s	
	ɑ	z	i
j	ɑ		

Using the procedure above, the three generated pairwise alignments are:

j	ɑ	s		j	ɑ	s			ɑ	z	i
	ɑ	z	i	j	ɑ			j	ɑ		

[2] For presentation purposes, stress is not marked in the pronunciation examples.

2.3 Algorithms

We evaluated four algorithms with respect to the quality of their alignments, including three variants of the Levenshtein algorithm, as well as the Pair Hidden Markov Model.

2.3.1 The VC-sensitive Levenshtein algorithm

The Levenshtein algorithm (Levenshtein, 1965)[3] is a very efficient dynamic programming algorithm, which was first introduced by Kessler (1995) as a tool for computationally comparing dialects. The Levenshtein distance between two strings is determined by counting the minimum number of edit operations (i.e. insertions, deletions and substitutions) needed to transform one string into the other.

For example, the Levenshtein distance between [jɑs] and [ɑzi], two Bulgarian dialectal variants of the word 'I', is 3:

jɑs	delete j	1
ɑs	subst. z/s	1
ɑz	insert i	1
ɑzi		
		3

The corresponding alignment is:

j	ɑ	s	
	ɑ	z	i
1		1	1

The Levenshtein algorithm has been used frequently and successfully for measuring linguistic distances in several languages, including Irish (Kessler, 1995), Dutch (Heeringa, 2004) and Norwegian (Heeringa, 2004). Additionally, the Levenshtein algorithm has been shown to yield aggregate results that are consistent (Cronbach's $\alpha = 0.99$) when applied to about 100 transcriptions at each site and valid when compared to dialect speakers' judgements of similarity ($r \approx 0.7$; Heeringa et al., 2006).

Following Heeringa (2004), we have adapted the Levenshtein algorithm slightly, so that it does not allow alignments of vowels with consonants (i.e. the syllabicity constraint, mentioned above).

[3] The more famous Needleman-Wunsch algorithm (Needleman and Wunsch, 1970) maximizes the similarity between two strings and allows for a segment similarity matrix, whereas the original Levenshtein algorithm (Levenshtein, 1965) minimizes the distance between two strings and only uses a binary segment distinction (i.e. same and different). More recently, the Levenshtein algorithm has been adapted to allow for sensitive segment distances (Kessler, 1995; Heeringa, 2004), effectively making it equal to the Needleman-Wunsch algorithm.

2.3.2 The Levenshtein algorithm with the swap operation

Because metathesis occurs frequently in the Bulgarian dialect dataset (in 21 out of 152 words), we extend the VC-sensitive Levenshtein algorithm as described in Section 2.3.1 to include the swap-operation (Wagner and Lowrance, 1975) which allows two adjacent characters to be interchanged. The swap-operation is also known as a transposition, which was introduced with respect to detecting spelling errors by Damerau (1964). As a consequence, the Damerau distance refers to the minimum number of insertions, deletions, substitutions, and transpositions required to transform one string into the other. In contrast to Wagner and Lowrance (1975) and in line with Damerau (1964), we restrict the swap operation to be allowed only for string X and Y when $x_i = y_{i+1}$ and $y_i = x_{i+1}$ (with x_i being the token at position i in string X):

x_i	x_{i+1}
y_i	y_{i+1}
><	1

Consider the alignment of [vrɣ] and [vɣr], two Bulgarian dialectal variants of the word 'peak' (mountain). The alignment involves a swap and results in a total distance of 1:

v	r	ɣ
v	ɣ	r
	><	1

However, the alignment of the transcription [vrɣ] with another dialectal transcription [var] does not allow a swap and yields a total distance of 2:

v		r	ɣ
v	a	r	
	1		1

Including the option of swapping identical segments in the implementation of the Levenshtein algorithm is relatively easy. We set the cost of the swap operation to one[4] plus twice the cost of substituting x_i with y_{i+1} plus twice the cost of substituting y_i with x_{i+1}. In this way the swap operation will be preferred when $x_i = y_{i+1}$ and $y_i = x_{i+1}$, but not when $x_i \neq y_{i+1}$ or $y_i \neq x_{i+1}$. In the first case the cost of the swap operation is 1, which is less than the cost of the alternative of two substitutions. In the second case the cost is either 3 (if either $x_i \neq y_{i+1}$ or $y_i \neq x_{i+1}$) or 5 (if both $x_i \neq y_{i+1}$ and $y_i \neq x_{i+1}$), which is higher than the cost of only using insertions, deletions and substitutions.

Just as in the previous section, we do not allow vowels to align with consonants (except in the case of a swap).

[4] More specifically, the cost is set to 0.999 to prefer an alignment involving a swap over an alternative alignment involving only regular edit operations.

2.3.3 The Levenshtein algorithm with automatically generated PMI-based segment distances

As we are investigating dialectal pronunciations which are reasonably similar to each other, we expect that similar sounds, such as [i] and [ɪ], will align more frequently than more distant sounds, such as [i] and [a]. We use pointwise mutual information (PMI; Church and Hanks, 1990) to convert these frequencies to distances.

The PMI approach consists of obtaining initial string alignments by using the VC-sensitive Levenshtein algorithm. Based on these initial alignments, the algorithm collects correspondences such as [i]:[ɪ] in a large segment × segment contingency table (i.e. a VARIATION MATRIX, see also Chapter 3). For example, the ([i],[ɪ]) cell of the table records how often the [i] aligned with the [ɪ]. These counts are subsequently used in the pointwise mutual information formula to determine the association strength between every pair of sound segments:

$$\mathrm{PMI}(x, y) = \log_2 \left(\frac{p(x, y)}{p(x)\, p(y)} \right)$$

Where:

- $p(x, y)$ is estimated by calculating the number of times sound segments x and y occur in correspondence in two aligned pronunciations X and Y, divided by the total number of aligned segments (i.e. the relative occurrence of the aligned sound segments x and y in the whole dataset).
- $p(x)$ and $p(y)$ are estimated as the number of times sound segment x (or y) occurs, divided by the total number of segment occurrences (i.e. the relative occurrence of sound segments x or y in the whole dataset). Dividing by this term normalizes the correspondence frequency with respect to the frequency expected if x and y are statistically independent.

We always add a tiny fractional value (< 1) to the frequency of occurrence of x, y, and the correspondence frequency of x and y. This ensures the numerator is always larger than zero. As the addition is very small, the effect is negligible for sound correspondences which align together. In contrast, sounds which do not align together now obtain the lowest (negative) PMI score (penalizing them severely).

Positive PMI values indicate that sounds tend to correspond relatively frequently (the greater the PMI value, the more two sounds tend to align), while negative PMI values indicate that sounds tend to correspond relatively infrequently. Sound distances (i.e. sound segment substitution costs) are generated by subtracting the PMI value from zero and adding the maximum PMI value (to ensure that the minimum distance is zero), and finally scaling these values between zero and one. Note that the lack of a segment (a gap) is also treated as a segment in the PMI procedure. The PMI-based sound segment distance

(i.e. PMI distance) between identical segments is always set to zero, as from an alignment perspective no cost accrues to aligning identical sounds.

After the sound segment substitution costs have been calculated for the first time (using the procedure above), the pronunciations are aligned anew with the Levenshtein algorithm using the adapted sound segment distances. This process (i.e. calculating new sound segment distances on the basis of the alignments and then obtaining new alignments by using these new sound segment distances) is repeated until the pronunciation alignments and sound distances remain constant. In general, it takes fewer than ten iterations for the alignments to remain constant. When two sound segments are not aligned at all, their distance will be very high (due to the very low PMI score, see above) and this effectively ensures they will never align in subsequent iterations.

When introducing the PMI-based Levenshtein algorithm (Wieling et al., 2009), all segments were included in the calculations to determine the PMI distance (between non-identical sound segments; the distance between identical sound segments is always set to zero). From an alignment perspective, however, aligning identical sounds does not involve a cost. Therefore, we also experimented with a version where pairs of identical sounds did not contribute towards the counts in the PMI formula (Wieling and Nerbonne, 2011b; here the procedure played a supporting role in removing the effects of inconsistent transcription practices). In the remainder of this chapter we will refer to these two different versions as the *diagonal-inclusive* (DI) and the *diagonal-exclusive* (DE) version of the PMI-based Levenshtein algorithm, respectively.

The potential merit of using PMI-generated segment distances can be made clear by the following example. Consider the transcriptions [vɤn] and [vɤŋkə], two Bulgarian dialectal variants of the word 'outside'. The VC-sensitive Levenshtein algorithm yields the following (correct) alignment:

v	ɤ	n		
v	ɤ	ŋ	k	ə
		1	1	1

But also the alternative (incorrect) alignment:

v	ɤ		n	
v	ɤ	ŋ	k	ə
		1	1	1

The VC-sensitive Levenshtein algorithm generates the erroneous alignment because it has no way to identify that the consonant [n] is more similar to the consonant [ŋ] than to the consonant [k]. In contrast, the PMI-based Levenshtein algorithm only generates the correct (first) alignment, because the [n] and [ŋ] align with a higher relative frequency than [n] and [k]. Therefore, the distance between [n] and [ŋ] will be lower than the distance between [n]

and [k]. The alignment according to the PMI-based Levenshtein algorithm is shown below.

v	ɤ	n			
v	ɤ	ɳ		k	ə
		0.015		0.036	0.042

The idea behind this procedure is similar to the approach of Ristad and Yianilos (1998), who automatically generated string edit distance costs using an expectation maximization algorithm. Our approach differs from theirs as we only learn segment distances based on the alignments generated by the VC-sensitive Levenshtein algorithm, while Ristad and Yianilos (1998) obtained segment distances by considering all possible alignments of two strings.[5]

2.3.4 The Pair Hidden Markov Model

The Pair Hidden Markov Model (PHMM) also generates alignments based on automatically obtained segment distances and has been used successfully in language studies (Mackay and Kondrak, 2005; Wieling et al., 2007b).

A Hidden Markov Model (HMM) is a probabilistic finite-state transducer that generates an observation sequence by starting in an initial state, going from state to state based on transition probabilities, and emitting an output symbol in each state based on the emission probabilities in that state for the output symbol (Rabiner, 1989). The PHMM was originally proposed by Durbin et al. (1998) for aligning biological sequences and was first used in linguistics by Mackay and Kondrak (2005) to identify cognates. The PHMM differs from the regular HMM as it outputs two observation streams (i.e. a series of alignments of pairs of individual segments) instead of only a series of single symbols. The PHMM displayed in Figure 2.1 has three emitting states: the substitution ('match') state (M) which emits two aligned segments, the insertion state (Y) which emits a segment and a gap, and the deletion state (X) which emits a gap and a segment.

The following example shows the state sequence for the pronunciations [jɑs] and [ɑzi] (English 'I'):

j	ɑ	s	
	ɑ	z	i
X	M	M	Y

Before generating the alignments, all probabilities of the PHMM have to be estimated. These probabilities consist of the 5 transition probabilities shown in Figure 2.1: ε, λ, δ, τ_{XY} and τ_{M}. Furthermore, there are 98 emission probabilities

[5] We also experimented with using the Levenshtein algorithm without the vowel-consonant alignment restriction to generate the initial PMI segment distances, but this adversely affected performance.

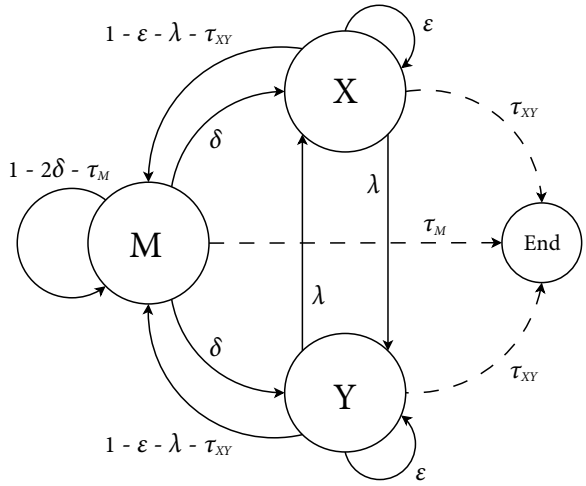


Figure 2.1. Pair Hidden Markov Model. Image courtesy of Mackay and Kondrak (2005).

for the insertion state and the deletion state (one for every segment) and 9604 emission probabilities for the substitution state. The probability of starting in one of the three states is set equal to the probability of going from the substitution state to that particular state. The Baum-Welch expectation maximization algorithm (Baum et al., 1970) can be used to iteratively reestimate these probabilities until a local optimum is found.

To prevent order effects in training, every word pair is considered twice (e.g., $w_a - w_b$ and $w_b - w_a$). Consequently, the resulting insertion and deletion probabilities are the same (for each segment), and the probability of substituting x for y is equal to the probability of substituting y for x, effectively yielding 4802 distinct substitution probabilities.

Wieling et al. (2007b) showed that sensible segment distances were obtained using Dutch dialect data for training; acoustic vowel distances on the basis of spectrograms correlated significantly ($r \approx -0.7$) with the vowel substitution probabilities of the PHMM. In addition, probabilities of substituting a segment with itself were much higher than the probabilities of substituting an arbitrary vowel with another non-identical vowel (*mutatis mutandis* for consonants), which were in turn much higher than the probabilities of substituting a vowel for a consonant.

After training, the well known Viterbi algorithm is used to obtain the best alignments (Rabiner, 1989).

2.3.5 Evaluation procedure

As described in Section 2.2, we use the generated pairwise alignments from a gold standard of multiple alignments for evaluation. In addition, we look at the alignment performance with respect to a baseline, which is constructed by aligning the strings according to the Hamming distance (i.e. only allowing substitutions and no insertions or deletions; Hamming, 1950).[6]

The evaluation procedure consists of comparing the alignments of the previously discussed algorithms (including the baseline algorithm) with the alignments of the gold standard. The evaluation proceeds as follows:

1. The pairwise alignments of the four algorithms, the baseline and the gold standard are generated and standardized (see below). When multiple equal-scoring alignments are generated by an algorithm, only one (i.e. the final) alignment is selected.

2. For each alignment, we convert every pair of aligned segments to a single token, so that every alignment of two strings is converted to a single string of segment pairs.

3. For every algorithm these transformed strings are aligned with the transformed strings of the gold standard using the standard Levenshtein algorithm.

4. The resulting Levenshtein distances are added together, resulting in the total distance between every alignment algorithm and the gold standard. Only if individual segments match completely, the segment distance is 0, otherwise it is 1.

To illustrate this procedure, consider the following gold standard alignment (detecting the swap) of [vlɣk] and [vɣlk], two Bulgarian dialectal variants of the word 'wolf':

v	l	ɣ	k
v	ɣ	l	k

Every aligned segment pair is converted to a single token by adding the symbol '/' between the segments and using the symbol '-' to indicate a gap. This yields the following transformed string:

v/v l/ɣ ɣ/l k/k

[6] Strictly speaking, the Hamming distance is defined as a measure of distance between strings of equal length. If they do not have equal length, we calculate the Hamming distance between the (sub)strings consisting of the first *n* segments (where *n* is the length of the shortest string) and add the difference in length (in segments) to the total Hamming distance.

Suppose another algorithm generates the following alignment (not detecting the swap):

v	l	ɤ			k
v			ɤ	l	k

The transformed string for this alignment is:

v/v l/- ɤ/ɤ -/l k/k

To evaluate this alignment, we align this string to the transformed string of the gold standard and obtain a Levenshtein distance of 3:

v/v	l/ɤ	ɤ/l		k/k
v/v	l/-	ɤ/ɤ	-/l	k/k
	1	1	1	

By repeating this procedure for all alignments and summing up all distances, we obtain the total distance between the gold standard and every individual alignment algorithm. Algorithms which generate high-quality alignments will have a low distance from the gold standard, while the distance will be higher for algorithms which generate low-quality alignments.

Standardization

The gold standard contains a number of alignments which have alternative equivalent alignments, most notably an alignment containing an insertion followed by a deletion (which is equal to the deletion followed by the insertion), or an alignment containing a syllabic consonant such as [ɻ], which in fact matches both a vowel and a neighboring /r/-like consonant and can therefore be aligned with either the vowel or the consonant. In order to prevent punishing the algorithms which do not match the exact gold standard in these cases, the alignments of the gold standard and all alignment algorithms are transformed to one standard form in all relevant cases.

For example, consider the correct alignment of [viɑ] and [vij], two Bulgarian dialectal variations of the English plural pronoun 'you':

v	i	ɑ	
v	i		j

Of course, the following alignment is just as reasonable:

v	i		ɑ
v	i	j	

To avoid punishing the latter, we transform all insertions followed by deletions to deletions followed by insertions, effectively scoring the two alignments the same.

For the syllabic consonants we transform all alignments to a form in which the syllabic consonant is followed by a gap and not vice versa. For instance, aligning [vɻx] with [vɑrx] (English: 'peak') yields:

v		ɻ	x
v	ɑ	r	x

Which is transformed to the equivalent alignment:

v	ɻ		x
v	ɑ	r	x

2.4 Results

We will report both quantitative results using the evaluation method discussed in the previous section, as well as the qualitative results, where we focus on the characteristic errors of the different alignment algorithms.

2.4.1 Quantitative results

Because there are two algorithms which use generated segment distances (or probabilities) in their alignments, we first check if these values are sensible and comparable to each other.

Comparison of segment distances

Similar to Wieling et al. (2007b), we found sensible PHMM substitution probabilities (convergence was reached after 1675 iterations, taking about 7 CPU hours): the probability of matching a segment with itself was significantly higher than the probability of substituting one vowel for another (similarly for consonants), which in turn was higher than the probability of substituting a vowel with a consonant (all t's > 9, $p < 0.001$).

These comparisons are not very informative for the PMI-based Levenshtein algorithms, as the distance between identical segments is always set to zero and no vowel-consonant alignments are allowed. The PMI-based Levenshtein algorithms, however, are much faster than the PHMM algorithm, as convergence was reached in less than 10 iterations, taking only a few minutes of CPU time.

Because the occurrence frequency of the sound segments influences the PHMM substitution probabilities, we do not compare these substitution probabilities directly to the PMI distances. To obtain comparable scores, the PHMM substitution probabilities are divided by the product of the relative frequencies

of the two sound segments used in the substitution. Since substitutions involving similar infrequent segments now get a much higher score than substitutions involving similar, but frequent segments, the logarithm is used to bring the respective scores into a comparable scale.

As the residuals of the linear regression between the PHMM similarities and PMI distances were not normally distributed, we used Spearman's rank correlation coefficient to assess the relationship between the two variables.

We found high significant correlations between the PHMM probabilities and the PMI sound segment distances (not taking same-segment distances and vowel-consonant distances into account as these remain fixed in the PMI approach). For the diagonal-inclusive version of the PMI-based algorithm (i.e. not ignoring same-segment alignments) Spearman's $\rho = -0.965$ ($p < 0.001$). For the diagonal-exclusive version (i.e. ignoring same-segment alignments) Spearman's $\rho = -0.879$ ($p < 0.001$). When looking at the insertions and deletions we also found a significant relationship between the PHMM and the PMI-based Levenshtein algorithms: Spearman's $\rho = -0.740$ ($p < 0.001$). These results indicate that the relationship between the PHMM similarities and the PMI distances is very strong. As the PHMM takes all sound correspondences into account, it is not surprising that the correlation is higher for the diagonal-inclusive version of the PMI-based Levenshtein algorithm than for the diagonal-exclusive version.

Evaluation against the gold standard

Using the procedure described in Section 2.3.5, we calculated the distances between the gold standard and the alignment algorithms. Besides reporting the total number of misaligned tokens, we also divided this number by the total number of aligned segments in the gold standard (about 16 million) to get an idea of the error rate. Note that the error rate is zero in the perfect case, but might rise to nearly two in the worst case, which is an alignment consisting of only insertions and deletions and therefore up to twice as long as the alignments in the gold standard. Finally, we also report the total number of alignments (word pairs) which are not completely identical to the alignments of the gold standard.

The results are shown in Table 2.1. We can clearly see that all algorithms beat the baseline and align about 95% of all string pairs correctly. While the Levenshtein PMI algorithm (diagonal-exclusive version) aligns most strings perfectly, it misaligns slightly more individual segments than the PHMM algorithm (i.e. it makes more errors per individual string alignment). The VC-sensitive Levenshtein algorithm performs significantly worse than all other non-baseline algorithms. It is clear that the (newer) diagonal-exclusive version of the PMI-based Levenshtein algorithm outperforms the (original) diagonal-inclusive version.

Algorithm	Alignment errors (%)	Segment errors (rate)
Baseline (Hamming)	726844 (20.92%)	2510094 (0.1579)
Levenshtein VC	191674 (5.52%)	490703 (0.0309)
Levenshtein Swap	161834 (4.66%)	392345 (0.0247)
PHMM	160896 (4.63%)	**362423 (0.0228)**
Levenshtein PMI (DI)	156440 (4.50%)	399216 (0.0251)
Levenshtein PMI (DE)	**152808 (4.40%)**	387488 (0.0244)

Table 2.1. Comparison to gold standard alignments. All differences are significant ($p < 0.01$). The diagonal-exclusive (DE) version of the PMI-based Levenshtein algorithm excludes pairs of identical sounds from the calculations needed to determine the sound distances, whereas the diagonal-inclusive (DI) version does not.

2.4.2 Qualitative results

Let us first note that it is almost impossible for any algorithm to achieve a perfect overlap with the gold standard, because the gold standard was generated from multiple alignments and therefore incorporates other constraints. For example, while a certain pairwise alignment could appear correct in aligning two consonants, the multiple alignment could show contextual support (from pronunciations in other varieties) for separating the consonants. Consequently, all algorithms discussed below make errors of this kind.

In general, the specific errors of the VC-sensitive Levenshtein algorithm can be separated into three cases. First, as illustrated in Section 2.3.3, the VC-sensitive Levenshtein algorithm has no way to distinguish between aligning a consonant with one of two neighboring consonants and sometimes chooses the wrong one (this also holds for vowels). Second, it does not allow alignments of vowels with consonants and therefore cannot detect correct vowel-consonant alignments such as correspondences of [u] with [v]. Third, the VC-sensitive Levenshtein algorithm is not able to detect metathesis of vowels with consonants (as it cannot align these).

The misalignments of the Levenshtein algorithm with the swap-operation can also be split into three cases. It suffers the same two problems as the VC-sensitive Levenshtein algorithm in choosing to align a consonant incorrectly with one of two neighboring consonants and not being able to align a vowel with a consonant. Third, even though it aligns some of the metathesis cases correctly, it also makes some errors by incorrectly applying the swap-operation. For example, in the following case the swap-operation is applied by the algorithm:

s	i	r^{j}	ɪ	n	i
s	i	r^{j}	n	ɪ	
			><	ɪ	ɪ

However, the two ɪ's are not related (according to historical linguistics) and should not be swapped, which is reflected in the gold standard alignment:

s	i	r^{j}	ɪ	n	i
s	i	r^{j}		n	ɪ

The incorrect alignments of (both versions of) the Levenshtein algorithm with the PMI-generated segment distances are mainly caused by its inability to align vowels with consonants and therefore, just as the VC-sensitive Levenshtein algorithm, it fails to detect metathesis and correct vowel-consonant alignments. On the other hand, using segment distances often solves the problem of selecting which of two plausible neighbors a consonant (or vowel) should be aligned with.

The two different versions of the PMI-based Levenshtein algorithm generated slightly different alignments. Same-segment alignments (included in the diagonal-inclusive version) increase the relative frequency of the segments involved, which results in a lower PMI score (and a higher segment distance). As gap-gap alignments cannot occur, gaps do not have that effect and alignments involving a gap (i.e. an insertion or a deletion) have a relatively low distance compared to substitutions. As the diagonal-exclusive version ignores same-segment alignments, it does not suffer from a preference for insertions and deletions (over substitutions), and consequently alignment quality is improved slightly.

Because the PHMM employs segment substitution probabilities, it also often solves the problem of aligning a consonant (or vowel) with one of two plausible neighbors. In addition, the PHMM often correctly aligns metathesis involving equal as well as similar segments, even realizing an improvement over the Levenshtein algorithm with the swap operation. Unfortunately, many wrong alignments of the PHMM are also caused by allowing vowel-consonant alignments.

2.5 Discussion

In this chapter we introduced several pairwise string alignment methods, and showed that they improve on the popular Levenshtein algorithm with respect to alignment accuracy.

While the results indicated that the PHMM misaligned the fewest segments, training the PHMM is a lengthy process which lasts several CPU hours. Considering that the Levenshtein algorithm with PMI distances is much quicker to apply, and that it has only slightly lower performance with respect

to the segment alignments, we prefer it over the PHMM. Another argument in favor of this choice is that it is *a priori* clearer what type of alignment errors to expect from the PMI-based Levenshtein algorithm, while the PHMM algorithm and its errors are less predictable and harder to comprehend.

It would be interesting to investigate if using the PMI-based segment distances while also including the swap operation is beneficial. One approach could be to use the Levenshtein algorithm with the swap operation to generate the initial alignments for the PMI procedure. In this case, two segments involved in a swap will obtain a lower PMI distance, and will be more likely to correspond in subsequent alignments. A drawback of this approach, however, is that these segments will also be more likely to align if no swap is involved (as no context is taken into account in the PMI-based Levenshtein algorithm).

We only looked at the PMI-based measure of association strength, but of course other measures are also possible. Evert (2005) evaluates various measures with respect to collocation identification (i.e. words which are commonly found together) and it would be interesting to evaluate the performance of these measures here as well.

In the next chapter we will see that the sound segment distances generated by the PMI-based Levenshtein algorithm (diagonal-exclusive version) are also acoustically sensible. Consequently, the algorithm is very suitable for determining individual sound correspondences (see Chapter 4) and, given the close link between alignment and distance (see Section 2.3.1), also for improving word pronunciation distances (see Chapters 6 and 7).

Chapter 3

Inducing phonetic distances from variation

Abstract. Structuralists famously observed that language is "un système où tout se tient" (Meillet, 1903, p. 407), insisting that the system of relations of linguistic units was more important than their concrete content. In this chapter we attempt to derive content from relations, in particular phonetic (acoustic) content from the distribution of alternative pronunciations used in different geographical varieties. We start from data documenting language variation, examining six dialect atlases each containing the phonetic transcriptions of the same sets of words at hundreds of sites. We then use the pointwise mutual information (PMI) procedure, introduced in the previous chapter, to obtain the phonetic distances and evaluate the quality of these distances by comparing them to acoustic vowel distances. For all dialect datasets (Dutch, German, Gabon Bantu, U.S. English, Tuscan and Bulgarian) we find relatively strong significant correlations between the induced phonetic distances and the acoustic distances, illustrating the usefulness of the method in deriving valid phonetic distances from distributions of dialectal variation. In addition, these results provide further support for using the PMI-based Levenshtein algorithm in pronunciation comparison.[1]

3.1 Introduction

In this chapter we evaluate the success of the pointwise mutual information (PMI) approach, introduced in Section 2.3.3, by comparing derived phonetic segment distances to independent acoustic characterizations. Since there is a consensus that formant frequencies characterize vowels quite well, we compare in particular the phonetic segment distances of vowels generated by our method to vowel distances in acoustic space. As we know of no well-accepted method to measure acoustic differences between consonants,[2] we cannot evaluate these, but we do examine them. In the following, we will

[1] This chapter is based on Wieling, Margaretha and Nerbonne (2011a) and Wieling, Margaretha and Nerbonne (2012).

[2] See, however, Mielke (2005) for an interesting recent attempt using Dynamic Time Warping, a continuous analogue of the Levenshtein distance.

elaborate on the motivations for automatically deriving phonetic segment distances.

3.1.1 Dialectology

As indicated in Chapters 1 and 2, improving the measure of similarity (or distance) between two phonetic strings is important in dialectometry, especially when focusing on word pronunciation distances. Kessler (1995) and especially Heeringa (2004, pp. 27–119) experimented with a large number of segment distance measures, which form an optional component of edit-distance measures such as the Levenshtein distance, seeking to validate the measures using the correlation between aggregate varietal distance as measured by the Levenshtein distance algorithm with dialect speakers' judgements (of overall similarity to their own dialect). Unfortunately, none of Kessler's or Heeringa's measures improved (much) on the very simple, binary measure which distinguishes only identical and non-identical segments (Kessler, 1995; Heeringa, 2004, pp. 27–119, 186). We suggest that the difficulty of demonstrating improvement arose because these researchers compared results at relatively high levels of aggregation.

By using automatically determined sound segment distances, we will refine the measure, but we are aware that the refinement may be better and still not lead to improvements at larger levels of aggregation. Consider improving on the precision of a centimeter-based measure of people's height by using more precise millimeters: this would be more accurate, but comparisons of average heights in different populations would be unlikely to benefit, indeed unlikely to even change. We do expect the new measure (if acoustically sensible) to prove more useful as one turns to more sensitive uses, such as validations involving pairs of individual pronunciations.

3.1.2 Historical linguistics

The Levenshtein distance is basically an inverse similarity measure, and historical linguists are clear about rejecting similarity as a basis for inference about historical relatedness (Campbell, 2004, p. 197). But an improvement in measuring segment distance also improves alignments (see Chapter 2), which means in turn an improved ability in (automatically) identifying the sound correspondences which historical linguistics *does* rely on (Hock and Joseph, 1996, Ch. 4 and 16). This is therefore a second linguistic area which may benefit from improved measures of segment distance. Indeed, historical examination normally relies on detecting regular sound correspondences such as the famous [p]:[f] correspondence between Romance and Germanic languages, seen in pairs such as Lat. *pater*, Eng. 'father'; Lat. *plenus*, Eng. 'full' or Lat. *pisces*, Eng. 'fish'. The step we imagine improving is the extraction of large numbers of correspondences, which may then be analyzed using phylogenetic inference

(Bryant et al., 2005). Covington (1996), Oakes (2000) and Kondrak (2003) have experimented with automatic alignment in service of historical linguistics, focusing on the ability to identify cognates. Prokić (2010, Ch. 6) has taken first steps to use multiple aligned data in phylogenetic inference.

3.1.3 Phonetics and phonology

As Laver (1994, p. 391) notes, there is no widely accepted procedure for determining phonetic similarity, nor even explicit standards: "Issues of phonetic similarity, though underlying many of the key concepts in phonetics, are hence often left tacit."

It is clear that there has nonetheless been a great deal of work on related topics in phonetics and laboratory phonology. In phonetics, Almeida and Braun (1986) developed a measure of segment distance in order to gauge the fidelity of phonetic transcriptions. It was used, e.g., to evaluate intra- and intertranscriber differences. Cucchiarini (1993) refined this work and Heeringa (2004) also experimented with Almeida and Braun's segment distance measure in dialectometry.

In laboratory phonology, Pierrehumbert (1993) experimented with a simple feature-overlap definition of similarity to which Broe (1996) added an information-theoretic refinement discounting redundant features. Frisch (1996) recast these definitions in terms of natural classes, rather than features, and Frisch et al. (2004) demonstrate that the Arabic syllable is best described as involving a gradient constraint against similar consonants in initial and final position, the so-called 'Obligatory Contour Principle'. Bailey and Hahn (2005) measure the degree to which the definitions of Frisch et al. (2004) predict the frequency of perceptual confusions in confusion matrices (see below), obtaining fair levels of strength ($0.17 \leq r^2 \leq 0.42$).

In general, the work from phonetics and (laboratory) phonology has experimented with theoretically inspired definitions of similarity as a means of explaining phonotactic constraints or potential confusions. Bailey and Hahn (2005) contrasted theoretically inspired definitions of phonetic similarity to empirical measures based on confusion matrices. A confusion matrix (Miller and Nicely, 1955) normally records the outcome of a behavioral experiment. It is a square matrix in which the rows represent sounds (or symbols) presented to subjects and the columns the sounds perceived. Each cell (r, c) records the number of times the signal in row r was perceived as the signal in column c. So cell (ɔ,o) records how often [ɔ] was perceived as [o], and the diagonal then represents the non-confused, correctly perceived signals.

As opposed to confusion matrices which record variants in speech perception, we introduce VARIATION MATRICES which record (dialectal) variants in speech production. In our case the variation matrix is initiated not with a behavioral experiment, but rather using distributional data available in dialect atlases. Based on alignments of dialectal pronunciations for a large set of words,

we obtain the frequency with which sound segments align. Continuing with the example above, cell (ɔ,o) in a variation matrix thus represents the number of times [ɔ] was used in the pronunciation of one variety, whereas [o] was used at the corresponding position in the pronunciation of another variety. We will use these variation matrices to directly extract information about sound segment similarity in a data-driven manner (as opposed to proceeding from a theoretical notion, see above). Specifically, we employ the information-theoretic PMI measure of association strength (introduced in the previous chapter) to determine the final sound segment distances.[3] Studies involving (data similar to) confusion matrices have often applied MDS as well (Fox, 1983), just as we will here.

3.1.4 Computational linguistics

Sequence alignment and sequence distance are central concepts in several areas of computer science (Sankoff and Kruskal, 1999; Gusfield, 1999), and the Levenshtein distance algorithm and its many descendants are used frequently, not only for phonetic transcriptions, but also for comparing computer files, macromolecules and even bird song (Tougaard and Eriksen, 2006). Computational linguists have also experimented with variable segment distances for various reasons.

Kernighan et al. (1990) induced segment distances from teletype data in order to better predict the intended word when faced with a letter sequence that did not appear in their lexicon. Ristad and Yianilos (1998) applied an expectation maximization algorithm to the problem of learning edit costs, and evaluated the results on their effectiveness in classifying phonetic transcriptions representing spoken words in the Switchboard corpus. The phonetic transcriptions were correctly classified if they corresponded to words that annotators understood (represented phonemically). Brill and Moore (2000) generalized earlier work to include many-to-one substitutions, testing their scheme on spelling correction, which Toutanova and Moore (2002) took as a basis for focusing on pronunciation modeling. As indicated in Section 2.3.4, Wieling and Nerbonne (2007) applied a Pair Hidden Markov Model (PHMM) to dialect data, demonstrating that the PHMM could likewise induce acoustic distances (for Dutch) fairly well ($r \approx -0.7$), but also that the runtime involved many hours of CPU time.

3.1.5 Additional motivation

It is clear that the notion 'phonetically similar' is often informally invoked, and not only when describing how deviant a given dialect pronunciation is with

[3] Ohala (1997) calls for an information-theoretic perspective on confusion matrices, but he is particularly interested in non-symmetric aspects of the matrices.

respect to a standard language or other dialects (the issue we are most interested in). Nor do the various research lines mentioned above exhaust the utility of the concept.

Phonetic similarity also plays a role when discussing the comprehensibility of foreigners' speech and how heavy their accents are (Piske et al., 2001; Flege et al., 1995), when assessing the success of foreign language instruction, or when discussing the quality of speech synthesizers (Van Heuven and Van Bezooijen, 1995). Sanders and Chin (2009) measure the intelligibility of the speech of cochlear implant bearers using a measure of phonetic similarity. Kondrak and Dorr (2006) apply a measure of pronunciation distance to identify potentially confusing drug names. And, although we will not attempt to make the argument in detail, we note that the many appeals to "natural" phonetic and phonological processes also seem to appeal to a notion of similarity, at least in the sense that the result of applying a natural process to a given sound is expected to sound somewhat like the original, albeit to varying degrees.

3.1.6 Structuralism

It was a major structuralist tenet that linguistics should attend to the relations (distributions) among linguistic entities more than to their substance proper (Meillet, 1903, p. 407). For example, a structuralist attends more to phonemic distinctions, to sounds which fall in the relation "potentially capable of distinguishing lexical meaning" than to the details of how the sounds are pronounced, but also to sounds that fall in the complementary distribution relation (not found in the same phonetic environment) or the free variation relation (found in the same phonetic environment, but without an effect on lexical meaning).

In the present case we attend to sounds which participate in the relation "potentially used as a dialect variant" and we do not privilege either phonemic or sub-phonemic variation. Some structuralists might well draw the line at considering variation outside a tightly defined variety, and in that sense we are perhaps not merely developing structuralist ideas. Other structuralists nonetheless recognized that the speech of "the whole community" was the proper concern of linguistics, in spite of the fact that "every person uses speech forms in a unique way" (Bloomfield, 1933, p. 75). They did not advocate attention to the idiolects of speakers in "completely homogeneous speech communities" (Chomsky, 1965, p. 3).

In suggesting a renewed focus on phonetic and phonological relations, i.e. distributions, we are aware that phonetics — and to some extent phonology (Cole, 2010) — has largely and successfully ignored the advice to concentrate on relations, in favor of examining the articulatory, acoustic and auditory basis of sounds, and we do not presume to question the wisdom of that development. It nonetheless remains scientifically interesting to see how much information is present in (cross-speaker) distributions. As we note above, the sort of dis-

tribution we examine below is perhaps of a different sort than the ones many structuralists had in mind, but its key property is that it is derived from a large number of alternative pronunciations.

3.2 Material

3.2.1 Dialect pronunciation datasets

To obtain a representative view of the quality of the sound segment distances generated by the PMI procedure (see Section 2.3.3; we use the diagonal-exclusive version), we apply the method to six independent datasets. In addition to the Bulgarian dataset introduced in Chapter 2, we will generate PMI-based sound segment distances (i.e. PMI distances) on the basis of a Dutch dataset, a German dataset, a U.S. English dataset, a Gabon Bantu dataset, and a Tuscan dataset.

We evaluate the quality of the vowel distances by comparing them to acoustic distances in formant space. In order to focus on segmental distances we ignore suprasegmentals, and in order to limit the number of distinct phonetic sounds in each dataset, we ignore diacritics. To obtain a reliable set of automatically generated vowel distances, however, we exclude vowels having a frequency lower than one percent of the maximum vowel frequency in each dataset.

As indicated in Section 2.2, the Bulgarian dataset consists of phonetic transcriptions of 152 words in 197 locations equally distributed over Bulgaria. The Bulgarian dataset is characterized by a relatively small number of vowels (11): /i, e, ɛ, u, ʊ, a, ɑ, o, ɔ, ɤ, ə/.

The Dutch dialect data set contains phonetic transcriptions of 562 words in 613 locations in the Netherlands and Flanders. Wieling et al. (2007a) selected the words from the Goeman-Taeldeman-Van-Reenen-Project (GTRP; Goeman and Taeldeman, 1996) specifically for an aggregate analysis of pronunciation variation in the Netherlands and Flanders. The Dutch dataset differentiates 18 vowels (excluding the low-frequency vowels): /a, ɑ, ɒ, ʌ, æ, e, ɛ, i, ɪ, y, o, ɔ, u, ʊ, ɵ, œ, ø, ə/.

The German dataset contains phonetic transcriptions of 201 words in 186 locations collected from the *Phonetischer Atlas der Bundesrepublik Deutschland* (Göschel, 1992) and was analyzed and discussed in detail by Nerbonne and Siedle (2005). The German dataset differentiates 21 vowels (excluding the low-frequency vowels): /a, ɑ, ɒ, ʌ, ɐ, æ, e, ɛ, i, ɪ, y, ʏ, o, ɔ, u, ʊ, ɯ, ɵ, œ, ø, ə/.

The U.S. English dataset contains phonetic transcriptions of 153 concepts in 483 locations (1162 informants) collected from the *Linguistic Atlas of the Middle and South Atlantic States* (Kretzschmar, 1994). We obtained the simplified phonetic data from http://www.let.rug.nl/~kleiweg/lamsas/download. The U.S. English dataset differentiates 17 vowels (excluding the low-frequency vowels): /i, ɪ, e, ɛ, u, ʊ, æ, a, ɑ, ɒ, ɜ, ɞ, o, ɔ, ʌ, ɐ, ə/.

The Bantu dataset consists of phonetic transcriptions of 160 words in 53 locations and is equal to the subset of the *Atlas Linguistique du Gabon* analyzed and discussed in detail by Alewijnse et al. (2007). The Bantu dataset is distinctive, because several different language varieties (e.g., Fang and Tsogo) are included. In contrast to the Dutch, German and U.S. English datasets which distinguish many vowels, the Bantu dataset differentiates only eight vowels (excluding the low-frequency vowels): /e, ɛ, i, o, ɔ, u, a, ə/.

The Tuscan dataset, finally, consists of 444 words in 213 locations. In every location on average 10 informants were interviewed. This dataset was analyzed and discussed by Montemagni et al. (in press) and is a subset of the *Atlante Lessicale Toscane* (Giacomelli et al., 2000). As this dataset was compiled with a view to identifying lexical variation (note that we focused on a single lexical form per word), transcriptions are quite crude and consequently only a limited number of vowels were included. The Tuscan dataset therefore only differentiates eight vowels (excluding the low-frequency vowels): /i, e, ɛ, u, o, ɔ, a, ə/.

3.2.2 Acoustic vowel measurements

For every dialect dataset, we obtained formant frequency measurements (in Hertz) of the first two formants, F1 and F2, of the vowels. We included measurements for all vowels which also occurred in the corresponding dialect dataset.

For Bulgarian, we used the formant frequency measurements of a single Bulgarian male speaker[4] (a radio commentator speaking standard Bulgarian) reported by Lehiste and Popov (1970) for six vowels: /i, e, ə, a, o, u/. Every measurement was based on 18 pronunciations of the (stressed) vowel. Unfortunately, no information was provided about where in the course of the vowel the measurements were taken and how many time points were sampled.

For Dutch, we used vowel formant frequency measurements of 50 male (Pols et al., 1973) and 25 female (Van Nierop et al., 1973) standard Dutch speakers. The formant frequency information was obtained from the initial (stable) part of the vowel waveform and was based on 10 sampling points (i.e. 10 periods generated as a continuous periodic waveform and input to the wave analyzer). We included the formant frequency measurements for 12 vowels: /i, ɪ, y, ʏ, e, ɛ, a, ɑ, o, ɔ, u, ø/. We averaged the mean frequencies of men and women in order to obtain a single set of frequencies.

For German, we used vowel formant frequency measurements of 69 male and 58 female standard German speakers (Sendlmeier and Seebode, 2006) for 14 vowels (stressed, except for the schwa): /i, ɪ, y, ʏ, e, ɛ, a, o, ɔ, u, ʊ, ʌ, ɐ, ə/. We averaged the mean frequencies of men and women in order to obtain a single set of frequencies. Unfortunately, no information was provided about

[4] We are aware of the variability in formant frequencies between different speakers. Unfortunately we were not able to find more formant frequency measurements of Bulgarian (and Bantu) speakers.

where in the course of the vowel the measurements were taken and how many time points were sampled.

For U.S. English, we used vowel formant frequency measurements of 45 men and 48 women speaking standard U.S. English (Hillenbrand et al., 1995). The formant frequency information was obtained from the initial (stable) part of the vowel waveform and was based on seven sampling points. We included acoustic measurements for 11 stressed vowels: /i, ɪ, e, ɛ, æ, ɑ, ɔ, o, ʊ, u, ʌ/ and we averaged the mean frequencies of men and women in order to obtain a single set of frequencies.

The Bantu dataset consisted of different languages, but we were only able to find vowel formant frequency measurements for the Fang language (Nurse and Philippson, 2003, p. 22). We included acoustic measurements for eight vowels: /i, e, ɛ, ə, a, ɔ, o, u/. Every measurement was based on six pronunciations of the vowel by a single speaker. Unfortunately, no information was provided about where in the course of the vowel the measurements were taken, if the vowels were stressed or not, and how many time points were sampled.

For Tuscan, we used the formant frequency measurements for two Tuscan dialects (the Pisan and Florentine varieties) reported by Calamai (2003). The formant frequency information was obtained from the (stable) vowel waveform and was based on three sampling points. For both dialects, recordings of two male speakers for seven stressed vowels (pronounced multiple times) were used: /a, ɛ, e, i, ɔ, o, u/.

3.3 Methods

3.3.1 Obtaining sound segment distances based on dialect pronunciations

The sound segment distances based on the dialect pronunciations were determined automatically using the PMI approach explained in Section 2.3.3. As the diagonal-inclusive version of the PMI-based Levenshtein algorithm was outperformed by the diagonal-exclusive version, we used the latter version to obtain the sound segment distances.

To appreciate how the present attention to relations is several magnitudes more encompassing than earlier structuralist work, we note that the procedure always involves a large number of correspondences. A word has four or five segments on average, so an aligned pronunciation pair yields about five correspondences. We work with word lists containing 152 – 562 words, meaning we obtain 760 – 2810 correspondences per pair of sites. As our datasets contain data from between 53 and 613 sites, there are between 1378 and 187,578 site pairs, and we collect between 10^6 and 5×10^8 correspondences per dataset.

	Pearson's r	Explained variance (r^2)	Significance
Dutch	0.672	45.2%	$p < 0.01$
Dutch w/o Frisian	0.686	47.1%	$p < 0.01$
German	0.630	39.7%	$p < 0.01$
German w/o /ə/	0.785	61.6%	$p < 0.01$
U.S. English	0.608	37.0%	$p < 0.01$
Bantu	0.642	41.2%	$p < 0.01$
Bulgarian	0.677	45.8%	$p < 0.01$
Tuscan	0.758	57.5%	$p < 0.01$

Table 3.1. Correlations between the acoustic and PMI distances for all datasets. Significance was assessed using the Mantel test (Mantel, 1967).

3.3.2 Calculating acoustic distances

To obtain the acoustic distances between vowels, we calculated the Euclidean distance of the average formant frequencies (in Bark, to correct for our non-linear perception of formant frequency; Traunmüller, 1990). Unfortunately, as we mainly obtained the average formant frequencies from published research, we were not able to apply speaker-based normalization (e.g., Lobanov, 1971).

We employ the acoustic distances to validate the corpus-based PMI procedure, but while the induced segmental distances are based on an entire language area, the acoustic differences have normally been measured using pronunciations according to the standard variety. One might object that we should compare with the acoustics of each of the varieties we examine, but we note that we induce distances from phonetic transcriptions which are used consistently across an entire language area. We therefore take it that we can use the acoustic pronunciations of the relevant IPA (International Phonetic Alphabet) vowels according to the standard variety as validation material.

3.4 Results

For all datasets, Table 3.1 shows the correlation between the acoustic and PMI distances. We assessed the significance of the correlation coefficients by using the Mantel test (Mantel, 1967), as our sound distances are not completely independent. It is clear that the acoustic and PMI distances match reasonably well, judging by the correlation coefficients ranging from 0.61 to 0.76 (for the complete datasets).

Given a matrix of vowel distances, we can use multidimensional scaling (MDS; Togerson, 1952) to place each vowel at the optimal position relative to all other vowels in a two-dimensional plane. Figure 3.1(a) shows the relative positions of the Bulgarian vowels on the basis of their acoustic distances (since these

are based on the first two formants, the complete variance is always visualized in two dimensions), while Figure 3.1(b) shows the relative positions of the Bulgarian vowels based on their PMI distances. Similarly, Figures 3.2 through 3.6 show the relative positions of the vowels based on the acoustic distances (a) as well as the PMI distances (b) for Dutch, German, U.S. English, Bantu and Tuscan. As the MDS calculations did not allow for missing distances, some sounds may be missing from the PMI distance visualizations. When the PMI method did not yield a distance between a pair of sounds (i.e. the two sounds did not align), we excluded one of these sounds from the MDS procedure.[5] Of course, all distances were included when calculating the correlation between the acoustic and PMI distances (shown in Table 3.1).

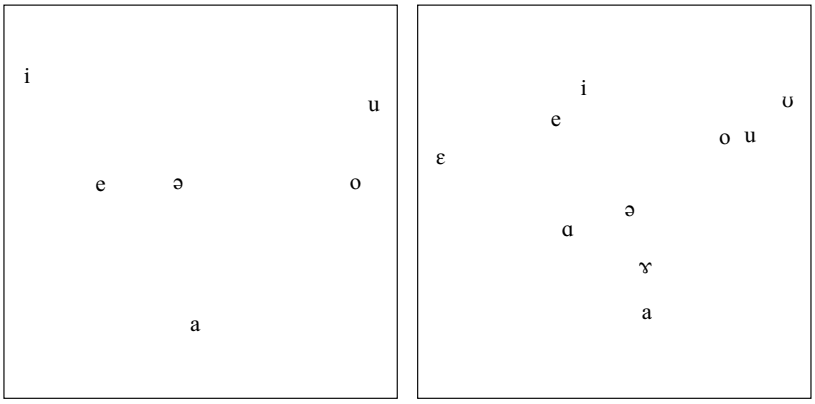
In examining the MDS visualizations of the vowels, one should keep in mind that they are visualizations of the relative distances of the vowels to each other, and not simply visualizations of vowels in any absolute coordinate system. So questions regarding the relative position of a certain vowel compared to other vowels can be answered, while those about the absolute position of a vowel (e.g., in the top-right) cannot.

It is clear that the visualizations on the basis of the acoustic distances resemble the IPA vowel chart (shown in Figure 3.7) quite nicely. The visualizations on the basis of the PMI distances are somewhat less striking and will be discussed for every figure separately.

The visualization of the Bulgarian data in Figure 3.1(b) (capturing 86% of the variation) reveals a deviating position of the [ɛ], presumably caused by its relatively large distance from (i.e. infrequent alignment with) [o] and [u]. If we ignore the [ɛ], the relative positions of the [i], [a] and [u] seem reasonable, however. Especially when considering the distances are based *only* on how frequently the sounds align in dialect data. Note that the [ɔ] was excluded from the MDS visualization, as this sound did not align with all other vowels (and no missing distances were allowed in the MDS procedure).

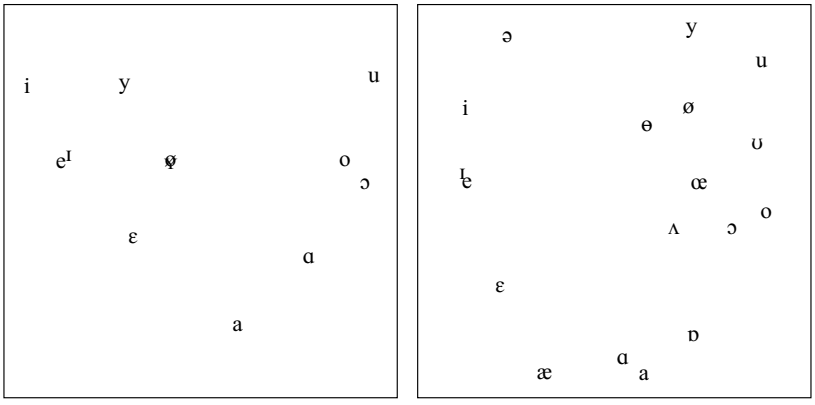
The visualization of the Dutch PMI distances in Figure 3.2(b) captures 76% of the variation and reveals quite acceptable relative positions of the [i], [u], [a] and similar sounds. However, the relative position of the [ə] (schwa) deviates significantly from the position on the basis of the acoustic distances. Investigating the underlying alignments revealed that the schwa was frequently deleted (i.e. aligned against a gap) and this resulted in relatively high distances between the schwa and the other vowels (which were deleted less frequently) compared to the other distances. Consequently, excluding the schwa increased the ability to visualize the distances between the vowels adequately in two dimensions: the explained variance increased from 76% to 85%. A second striking deviation for the Dutch dataset is the position of the front rounded vowels, which are surprisingly back (i.e. [y], [ø] and [œ]). Unfortunately, we do not have an immediate explanation for this, but it is likely that this reflects the frequency with

[5] We chose the sound to exclude in a way that maximized the number of sounds displayed.



(a) Acoustic distance visualization (b) PMI distance visualization

Figure 3.1. Relative positions of Bulgarian vowels based on their acoustic (a) and PMI distances (b). The visualization in (a) captures 100% of the variation in the original distances, while the visualization in (b) captures 86% of the variation in the original distances.

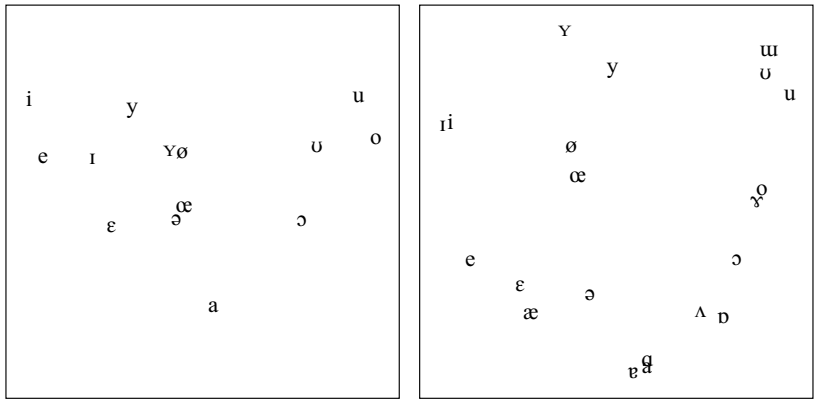


(a) Acoustic distance visualization (b) PMI distance visualization

Figure 3.2. Relative positions of Dutch vowels based on their acoustic (a) and PMI distances (b). The visualization in (a) captures 100% of the variation in the original distances, while the visualization in (b) captures 76% of the variation in the original distances.

which [u] and [y], etc. correspond, which may ultimately suggest a systematic limitation of the technique (i.e. sensitivity to umlaut).

The Dutch dataset also includes dialects where the Frisian language is spoken. We experimented with excluding the Frisian dialects from the Dutch



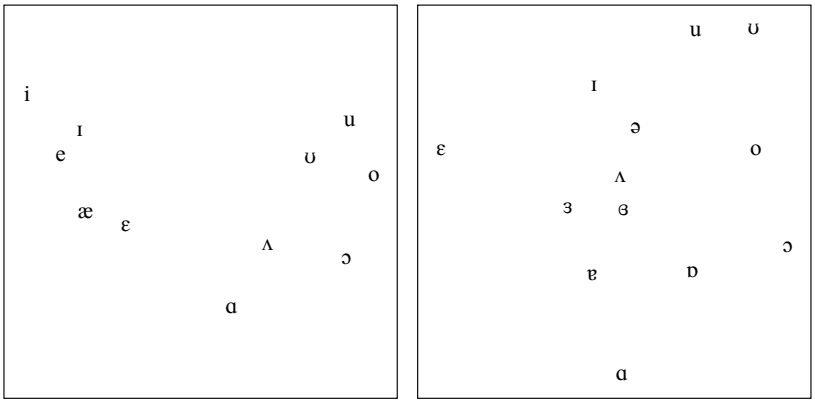
(a) Acoustic distance visualization (b) PMI distance visualization

Figure 3.3. Relative positions of German vowels based on their acoustic (a) and PMI distances (b). The visualization in (a) captures 100% of the variation in the original distances, while the visualization in (b) captures 70% of the variation in the original distances.

dataset as Frisian is recognized as a different language politically and is generally recognized as historically less closely related to Dutch than (for example) to English. In addition, Frisian and Dutch dialects have some sound correspondences consisting of rather dissimilar sounds (see Chapter 4), such as [a]:[i] (e.g., *kaas*, 'cheese': [kas] vs. [tsis]). As excluding the Frisian dialects did not have a serious effect on the correlation coefficient (an increase of only 0.014 to $r = 0.686$, see Table 3.1), we may conclude that the dissimilar sound correspondences are still outweighed in frequency by the phonetically similar sound correspondences. Only if dissimilar correspondences occurred more frequently than similar ones, would our method generate inadequate phonetic distances. However, as we generally include as much material as possible, it is unlikely that dissimilar sound correspondences will dominate.

The visualization of the German PMI distances in Figure 3.3(b) captures 70% of the variation and also reveals quite acceptable relative positions of the [i], [u], [a] and similar sounds. While the schwa was positioned better in Figure 3.3(b) than in Figure 3.2(b), the schwa was the most frequently deleted sound in the German dataset. Consequently, excluding the schwa from the visualization increased the explained variance from 70% to 82% and also resulted in a higher correlation between the acoustic and PMI distances (see Table 3.1).

The relative positions of the vowels based on the U.S. English PMI distances in Figure 3.4(b) (capturing 65% of the variation) are much more chaotic than the Dutch and German visualizations. If we ignore the [ɛ], the relative positions of the [ɪ], [ɒ] and [u] seem reasonable, however. Similar to the Bulgarian



(a) Acoustic distance visualization (b) PMI distance visualization

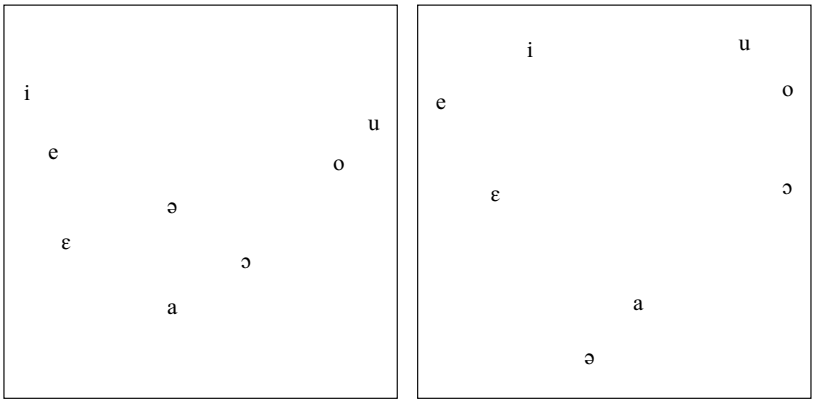
Figure 3.4. Relative positions of U.S. English vowels based on their acoustic (a) and PMI distances (b). The visualization in (a) captures 100% of the variation in the original distances, while the visualization in (b) captures 65% of the variation in the original distances.

visualization, the deviating position of the [ɛ] was likely caused by its relatively large distance from (i.e. infrequent alignment with) [o] and [u]. Note that the [i], [e], [a] and [æ] were excluded from the MDS visualization, as these sounds did not align with all other vowels (and no missing distances were allowed in the MDS procedure).

We turn now to the Bantu data. Similar to Dutch and German, the visualization of the Bantu PMI distances (capturing 90% of the variation) in Figure 3.5(b) reveals reasonable relative positions of the [i], [u] and [a]. The most striking deviation is the position of the schwa, caused by its low distance from the [a] and greater distance from [i] and [u].

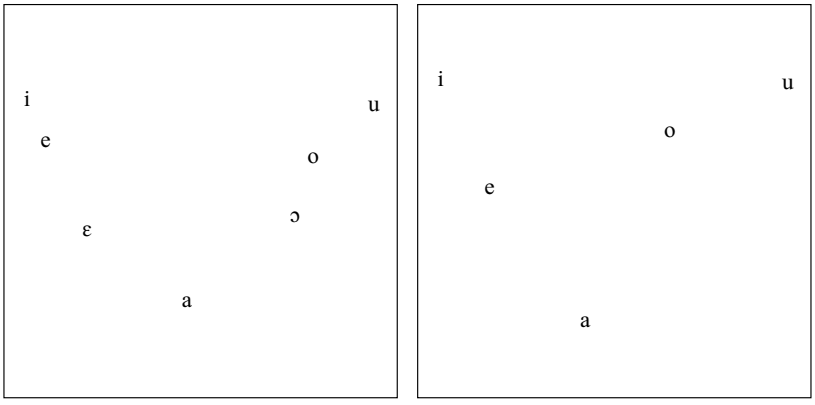
The visualization of the Tuscan PMI distances in Figure 3.6(b) captures 97% of the variation and shows a reasonably good relative placement of all sounds. Of course, this is not very surprising as there are only five sounds included in the visualization (i.e. the [ə], [ɔ] and [ɛ] were excluded as these sounds did not align with all other sounds, and the MDS procedure did not allow missing distances).

As we know of no well-accepted method to measure acoustic differences between consonants, we were not able to evaluate the quality of the automatically generated consonant distances explicitly. To illustrate that the consonant distances also seem quite sensible, Figure 3.8 shows the MDS visualization of several Dutch consonants (capturing 50% of the variation). Note that consonants having a frequency lower than one percent of the maximum consonant frequency were excluded, as well as consonants which did not align with all



(a) Acoustic distance visualization (b) PMI distance visualization

Figure 3.5. Relative positions of Bantu vowels based on their acoustic (a) and PMI distances (b). The visualization in (a) captures 100% of the variation in the original distances, while the visualization in (b) captures 90% of the variation in the original distances.



(a) Acoustic distance visualization (b) PMI distance visualization

Figure 3.6. Relative positions of Tuscan vowels based on their acoustic (a) and PMI distances (b). The visualization in (a) captures 100% of the variation in the original distances, while the visualization in (b) captures 97% of the variation in the original distances.

other consonants (no missing distances are allowed in the MDS procedure). Figure 3.8 clearly shows sensible groupings of the velar consonants [x], [χ], [ɣ], [g], [ŋ] in the upper-left, the rhotic consonants [ʀ], [r], [ɾ] in the upper-right, the alveopalatal consonants [j], [s], [n], [t], [d] in the center, the laterals [l],

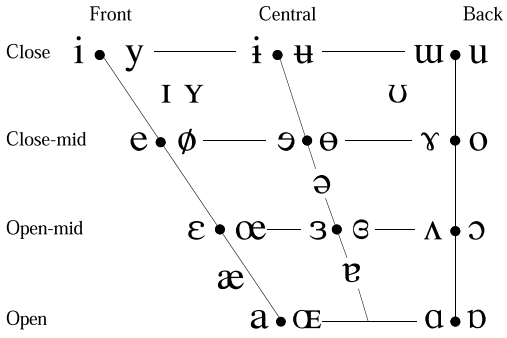


Figure 3.7. Vowel chart of the International Phonetic Alphabet

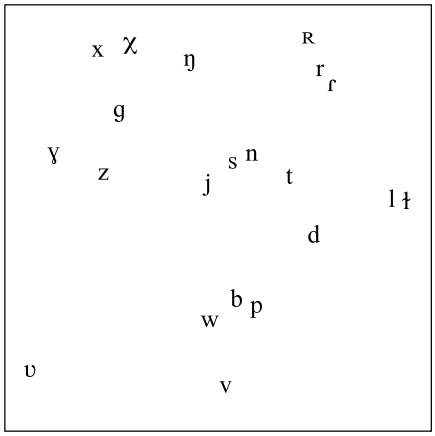


Figure 3.8. Relative positions of Dutch consonants based on their PMI distances. Fifty percent of the variation in the original distances is captured.

[ɬ] to the right and the bilabial and labiodental consonants [ʋ], [w], [b], [p], [v] at the bottom. (Even though the [ʋ] is fairly distant from the other members of this group, it is the closest group.) In contrast, the position of the [z] close to the velars is not easy to explain. Also note that the visualization of the consonant distances seems to indicate that place and manner characteristics dominate over voicing.

3.5 Discussion

In this chapter we have shown that the structure of variation matrices reflects phonetic distance. We used acoustic distances in formant space as an evaluation of the claim that we could derive information about phonetic distances

from cross-speaker distributions of variation. The level of correlation between the automatically determined phonetic distances and acoustic distances was similar in six independent dialect datasets and ranged between 0.61 and 0.76, a good indication that the relation between functioning as an alternative pronunciation and being similar in pronunciation is neither accidental nor trivial.

Of course, one might argue that these results are perfectly in line with what one would expect. Indeed, it is more likely that dialectal pronunciations will be similar to each other, rather than completely different and, consequently, similar sounds will align more frequently than dissimilar sounds. However, we would like to emphasize that our results have quantified how much information is implicit in (cross-speaker) distributions, something which has largely been lacking. Whether or not one is surprised at how much information is found in these distributions undoubtedly depends on one's theoretical convictions, but the present chapter has quantified this.

In line with this, the MDS visualizations of the automatically obtained segment distances were never completely identical to the visualizations based on the acoustic data. In some cases, this was caused by the frequency with which a particular sound segment (e.g., the schwa in Dutch and German) was deleted in the alignments (which consequently affected the other distances), but in other cases acoustically similar sounds simply aligned infrequently. So, while there is a clear connection between acoustic distances and the information about phonetic distances present in the distribution of alternative pronunciations, it is by no means perfect. It would be interesting to see if there is some kind of structure in these deviations. Unfortunately, we do not yet have a clear approach toward investigating this.

We excluded diacritics in order to limit the number of distinct phonetic sounds in each dataset. It would be insightful to investigate if including diacritics is possible and would still yield reliable sound distances, given the limited amount of data available. Another extension of the research in this chapter would be to investigate sound distances obtained using other measures of association strength instead of PMI (see also Section 2.5).

Clearly, we first need phonetic events in order to study their distributions. In this sense, this chapter has demonstrated how to *detect* phonetic relations from (cross-speaker) distributions, but we concede that it would be overeager to imagine that distributions *cause* the phonetics. We would like to note, however, that there have been demonstrations that distributions within acoustic space do influence children's learning of categories (Maye et al., 2002). There is room in linguistic theory to imagine that distributions in fact do influence phonetics.

We emphasize that we tested our inductive procedure against the ground truth of acoustics, and that we restricted our test to comparisons of vowels only because there is phonetic consensus about the characterization of vowels in a way that supports a measure of distance. While we did not investigate the automatically generated consonantal distances in this chapter extensively (as these

cannot be validated easily), a visual inspection of Dutch consonantal distances (see Figure 3.8) suggests that the method also yields adequate results for consonants.

Of course, we have not tried to demonstrate that improved segment distance measures lead to genuine improvements in all the various areas (besides dialectometry) discussed in the introduction, such as second-language learning (foreign accents), spelling correction, and the study of speech disorders. We note merely that there is a broad interest in measures of phonetic segment similarity, the focused issue to which we contribute. We are well aware that potential and genuine improvements are two very different matters.

We suggest that the results be viewed as vindicating the structuralists' postulate that the sound *system* of a language is of central importance, as this is reflected in the relations among variant pronunciations. We have shown that distributions (of alternative dialectal pronunciations) contain enough information to gauge content (i.e. phonetic similarity) to some extent. The only phonetic content made available to the algorithm was the distinction between vowels and consonants, and yet the algorithm could assign a phonetic distance to all pairs of vowel segments in a way that correlates fairly well with acoustic similarity. We know of no work in the strict structuralist tradition that attempted to analyze corpora of 10^8 segment pairs, nor of attempts to analyze entire tables reflecting pronunciation relations. We nonetheless find it appropriate to emphasize that our focus in this chapter is very much in the structuralist tradition of understanding the systems by studying relations within it.

In conclusion, it is promising that the good alignment performance of the PMI-based Levenshtein algorithm introduced in Chapter 2 is also supported by a relatively strong correlation between the PMI segment distances and acoustic distances. The opportunity to exploit phonetic segment distances in string alignment and string distance algorithms will allow us to assess word (string) distances more accurately and improve pronunciation alignments. As indicated in Chapter 1, this is a result we will take advantage of extensively in many of the following chapters.

Part III

Identifying dialect regions and their characteristic features

Chapter 4

Clustering Dutch dialects and their features

Abstract. In this chapter we use hierarchical bipartite spectral graph partitioning to simultaneously cluster varieties and identify their most characteristic linguistic features in Dutch dialect data. While clustering geographical varieties with respect to their features, e.g., pronunciation, is not new, the simultaneous identification of the features which give rise to the geographical clustering presents novel opportunities in dialectometry. Earlier methods aggregated sound differences and clustered on the basis of aggregate differences. The determination of the significant features which co-vary with cluster membership was carried out on a *post hoc* basis. Hierarchical bipartite spectral graph partitioning simultaneously seeks groups of individual features which are strongly associated, even while seeking groups of sites which share subsets of these same features. We show that the application of this method results in clear and sensible geographical groupings and discuss and analyze the importance of the concomitant features.[1]

4.1 Introduction

Dialect atlases contain a wealth of material that is suitable for the study of the cognitive, but especially the social dynamics of language. Although the material is typically presented cartographically, we may conceptualize it as a large table, where the rows are the sampling sites of the dialect survey and the columns are the linguistic features probed at each site. Table 4.1 illustrates the sort of information we wish to analyze. We use constructed data to keep the point maximally simple.

The features in the first two columns are intended to refer to the cognates, frequently invoked in historical linguistics. The first three varieties (dialects) all have lexicalizations for the concept 'hoe' (a gardening instrument), the fourth does not, and the question does not have a clear answer in the case of the fifth. The first two varieties use the same cognate for the concept 'eel', as do the last three, although the two cognates are different. More detailed material

[1] This chapter is based on Wieling and Nerbonne (2009), Wieling and Nerbonne (2010), and Wieling and Nerbonne (2011a).

	HOE?	EEL	...	night	hog	...	p^h	t^h	asp.	...
Appleton	+	A	...	[naɪt]	[hɔg]	...	+	+	+	...
Brownsville	+	A	...	[nat]	[hɑg]	...	+	+	+	...
Charleston	+	B	...	[nat]	[hɑg], [hɔg]	...	-	-	-	...
Downe	-	B	...	[nat]	[hɑg]	...	-	-	-	...
Evanston	?	B	...	[naɪt]	[hɔg]	...	+	+	+	...

Table 4.1. Example of dialect atlas data

is also collected, e.g., the pronunciations of common words, shown above in the fourth and fifth columns, and our work has primarily aimed at extracting common patterns from such transcriptions. As a closer inspection will reveal, the vowels in the two words suggest geographical conditioning (assuming that the first and last sites are near each other). This illustrates the primary interest in dialect atlas collections: they constitute the empirical basis for demonstrating how geography influences linguistic variation. On reflection, the influential factor is supposed to be not geography or proximity *simpliciter*, but rather the social contact which geographical proximity facilitates. Assuming that this reflection is correct, the atlas databases provide us with insights into the social dynamics reflected in language.

More abstract characteristics such as whether initial fortis consonants like [p, t] are aspirated (to be realized then as [p^h, t^h]) is sometimes recorded, or, alternatively, the information may be extracted automatically (see Chapters 2 and 3). Note, however, that we encounter here two variables, aspiration in /p/ and aspiration in /t/, which are strongly associated irrespective of geography or social dynamics. In fact, in all languages which distinguish fortis and lenis plosives /p, b/, /t, d/, etc., it turns out that aspiration is invariably found on *all* (initial) fortis plosives (in stressed syllables), or on none at all, regardless of social conditioning. We thus never find a situation in which /p/ is realized as aspirated ([p^h]) and /t/ as unaspirated (Lisker and Abramson, 1964). This is exactly the sort of circumstance for which cognitive explanations are generally proposed, i.e. explanations which do not rely on social dynamics. The work we discuss below does not detect or attempt to explain cognitive dynamics in language variation, but the datasets we study should ultimately be analyzed with an eye to cognitive conditioning as well.[2] The present chapter focuses exclusively on the social dynamics of variation.

Exact methods have been applied successfully to the analysis of dialect variation for over three decades (see Chapter 1), but they have invariably functioned

[2] Wieling and Nerbonne (2007) explore whether the perception of dialect differences is subject to a bias toward initial segments in the same way spoken word recognition is, an insight from cognitive science.

by first probing the linguistic differences between each pair of a range of varieties (sites, such as Whitby and Bristol in the UK) over a body of carefully controlled material (say the pronunciation of the vowel in the word 'put'). Second, the techniques *aggregate* over these linguistic differences, in order, third, to seek the natural groups in the data via clustering or multidimensional scaling (MDS; see Nerbonne, 2009).

Naturally, techniques have been developed to determine which linguistic variables weigh most heavily in determining affinity among varieties. But all of the following studies separate the determination of varietal relatedness from the question of its detailed linguistic basis. Kondrak (2002) adapted a machine translation technique to determine which sound correspondences occur most regularly. His focus was not on dialectology, but rather on diachronic phonology, where the regular sound correspondences are regarded as strong evidence of historical relatedness. Heeringa (2004, pp. 268–270) calculated which words correlated best with the first, second and third dimensions of an MDS analysis of aggregate pronunciation differences. Shackleton (2005) used a database of abstract linguistic differences in trying to identify the British sources of American patterns of speech variation. He applied principal component analysis (see also Chapter 5) to his database to identify the common components among his variables. Nerbonne (2006) examined the distance matrices induced by each of two hundred vowel pronunciations automatically extracted from a large U.S. English dataset, and subsequently applied factor analysis to the covariance matrices obtained from the collection of vowel distance matrices. Prokić (2007) analyzed Bulgarian pronunciation using an edit distance algorithm and then collected commonly aligned sounds. She developed an index to measure how characteristic a given sound correspondence is for a given site.

To study varietal relatedness and its linguistic basis in parallel, we apply bipartite spectral graph partitioning. Dhillon (2001) was the first to use spectral graph partitioning on a bipartite graph of documents and words, effectively clustering groups of documents and words simultaneously. Consequently, every document cluster has a direct connection to a word cluster; the document clustering implies a word clustering and vice versa. In his study, Dhillon (2001) also demonstrated that his algorithm identified sensible document and word clusters.

The usefulness of this approach is not only limited to clustering documents and words simultaneously. For example, Kluger et al. (2003) used a somewhat adapted bipartite spectral graph partitioning approach to successfully cluster microarray data simultaneously in clusters of genes and conditions.

There are two main contributions of this chapter. The first contribution is to apply a graph-theoretic technique, hierarchical bipartite spectral graph partitioning, to dialect pronunciation data in order to solve an important problem (see Chapter 1), namely how to recognize groups of varieties while simultaneously characterizing the linguistic basis (in terms of sound segment correspondences; see Chapters 2 and 3) of the group. The second contribution is the

application of a ranking procedure to determine the most important sound correspondences (with respect to a reference variety) in each cluster of varieties. This approach is an improvement over the procedure of ranking the most important elements in a cluster based only on their frequency (Dhillon, 2001), because it also takes differences between clusters into account.

4.2 Material

In this chapter we use the same Dutch dialect dataset as briefly introduced in Chapter 3. The Dutch dialect dataset originated from the Goeman-Taeldeman-Van Reenen-Project (GTRP; Goeman and Taeldeman, 1996; Van den Berg, 2003). The GTRP consists of digital transcriptions for 613 dialect varieties in the Netherlands (424 varieties) and Belgium (189 varieties), gathered during the period 1980–1995. For every variety, a maximum of 1876 items was narrowly transcribed according to (a variant of) the International Phonetic Alphabet. The items consist of separate words and phrases, including pronominals, adjectives and nouns. A detailed overview of the data collection is given by Taeldeman and Verleyen (1999).

Because the GTRP was compiled with a view to documenting both phonological and morphological variation (De Schutter et al., 2005) and our purpose here is the analysis of pronunciation, we ignore many items of the GTRP. We use the same 562-item subset as introduced and discussed in depth by Wieling et al. (2007a). In short, the 1876-item word list was filtered by selecting only single-word items, plural nouns (the singular form was sometimes preceded by an article and therefore not included), base forms of adjectives instead of comparative forms, and the first-person plural verb instead of other forms. In general, the same lexeme was used for a single item.

Since the GTRP transcriptions of Belgian varieties are fundamentally different from transcriptions of the Netherlandic varieties (i.e. they were not based on the same number of phonetic segments; Wieling et al., 2007a), we will restrict our analysis to the 424 Netherlandic varieties. The geographical distribution of these varieties is shown in Figure 4.1. Furthermore, note that we will ignore diacritics (as these are transcribed less reliably; Goeman, 1999) concentrating on the 82 distinct base sound segments present in the dataset. The average length of every item in the GTRP (without diacritics) is 4.7 segments (i.e. sound segments in a phonetic transcription).

4.3 Methods

To obtain a clear signal of varietal differences in phonology, we ideally want to compare the pronunciations of each variety with a single reference point. We might have used the pronunciations of a proto-language for this purpose, but these are not available in the same transcription system. We settled on using

Figure 4.1. Geographical distribution of the Dutch GTRP varieties. The province names are indicated.

the sound segment correspondences of one reference variety with respect to all other varieties as a means of comparison. These sound correspondences form a general and elaborate basis of comparison for the varieties. The use of the correspondences as a basis of comparison is general in the sense that we can determine the correspondences for each variety, and it is elaborate since it results in nearly 1000 points of comparison (sound correspondences).

But this strategy also leads to the question of what to use as a reference point. There are no pronunciations of standard Dutch in the GTRP and transcribing the standard Dutch pronunciations ourselves would likely have introduced between-transcriber inconsistencies. Especially at the segment level, it is likely that these transcriber differences will be detrimental to the results, more so than at a higher level of aggregation (i.e. word or dialect distances; see

Chapter 6) where transcriber differences are likely to be smoothed out more. Heeringa (2004, pp. 274–276) identified pronunciations in the variety of Haarlem as being the closest to standard Dutch. Because Haarlem was not included in the GTRP varieties, we chose the transcriptions of Delft (also close to standard Dutch) as our reference point.

4.3.1 Obtaining sound correspondences

To obtain the sound correspondences for every site in the GTRP with respect to the reference variety Delft, we used the PMI-based Levenshtein algorithm (diagonal-exclusive version) as explained in Section 2.3.3. In this chapter, however, we incorporated some additional linguistic information in the initialization step of the PMI-based Levenshtein algorithm by allowing, e.g., the alignment of the central vowel [ə] with sonorant consonants (e.g., [m] and [n]).

After obtaining the final string alignments, we used a matrix to store the presence or absence of each segment substitution for every variety (with respect to the reference variety). We thus obtained a binary $m \times n$ matrix $\boldsymbol{A}$ (matrices and vectors are denoted in boldface) of m varieties (in our case 423; Delft was excluded as it was our reference site) by n segment substitutions (in our case 957; not all possible segment substitutions occurred). A value of one in $\boldsymbol{A}$ (i.e. $A_{ij} = 1$) indicates the presence of segment substitution j in variety i (compared to the reference variety), while a value of zero indicates the absence. To alleviate the effect of noise, we only regarded a sound correspondence as present in a variety when it occurred in at least three aligned pronunciations. Consequently, we reduced the number of sound correspondences (columns of $\boldsymbol{A}$) by more than half to 477.

4.3.2 Bipartite spectral graph partitioning

An undirected bipartite graph can be represented by $G = (R, S, E)$, where R and S are two sets of vertices and E is the set of edges connecting vertices from R to S. There are no edges between vertices in a single set, e.g., connecting vertices in R. In our case R is the set of varieties, while S is the set of sound segment substitutions (i.e. sound correspondences). An edge between r_i and s_j indicates that the sound segment substitution s_j occurs in variety r_i. It is straightforward to see that matrix $\boldsymbol{A}$ is a representation of an undirected bipartite graph. Figure 4.2 shows an example of an undirected bipartite graph consisting of four varieties and three sound correspondences.

If we represent a graph such as that in Figure 4.2 using a binary adjacency matrix in which a cell (a,b) has the value one just in case there is an edge from a to b, and zero otherwise, then the spectrum of the graph is the set of eigenvalues of its adjacency matrix. Note that the adjacency matrix (having $(m+n)\times(m+n)$ elements) is larger than $\boldsymbol{A}$ (having $m \times n$ elements), as it contains values for all possible vertex combinations.

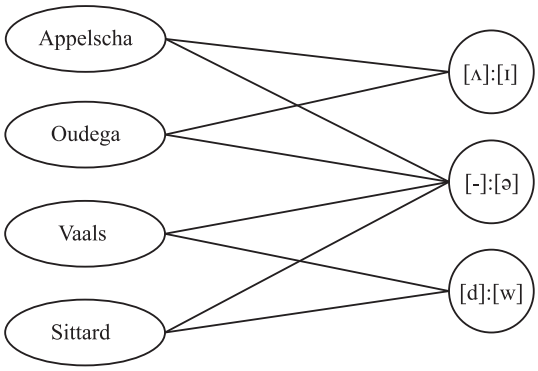


Figure 4.2. Example of a bipartite graph of four varieties and three sound correspondences. Note that [-] represents a gap and therefore the insertion of a schwa is indicated by [-]:[ə].

Spectral graph theory is used to find the principal properties and structure of a graph from its graph spectrum (Chung, 1997). Dhillon (2001) applied spectral graph partitioning to a bipartite graph of documents and words, resulting in a simultaneous clustering of documents and words. In similar fashion, we would like to obtain a clustering of varieties and corresponding sound segment substitutions.

The algorithm of Dhillon (2001) is based on the fact that finding the optimal bipartition having balanced clusters is solved by finding the eigenvector corresponding with the second-smallest eigenvalue of the adjacency matrix. Using linear algebra, it turns out that this solution can also be found by computing the left and right singular vectors corresponding to the second (largest) singular value of the normalized word-by-document matrix (or in our case variety-by-segment-correspondence matrix). Because the latter matrix is smaller than the adjacency matrix needed for the first method, the second method is computationally cheaper.

Instead of repeatedly finding two clusters to obtain a hierarchical clustering, it is also possible to find k clusters in a single step (i.e. a flat clustering), using $l = \lceil \log_2 k \rceil$ singular vectors.[3] However, Shi and Malik (2000) indicated that a hierarchical clustering, obtained by repeatedly grouping in two clusters, should be preferred over the flat clustering approach as approximation errors are reduced. More importantly, genealogical relationships between languages (or dialects) are generally expected to have a hierarchical structure due to the dynamics of language change in which early changes result in separate varieties

[3] See Wieling and Nerbonne (2009) and Wieling and Nerbonne (2011a) for the application of the flat hierarchical spectral graph clustering approach to the Dutch data.

which then undergo subsequent changes independently (Jeffers and Lehiste, 1979). We therefore apply the hierarchical approach in this chapter.

The hierarchical bipartite spectral partitioning algorithm, following Dhillon (2001), proceeds as follows:

1. Given the $m \times n$ variety-by-segment-correspondence matrix $\boldsymbol{A}$ as discussed previously, form the normalized matrix

$$\boldsymbol{A_n} = \boldsymbol{D_1}^{-1/2} \boldsymbol{A} \boldsymbol{D_2}^{-1/2}$$

   with $\boldsymbol{D_1}$ and $\boldsymbol{D_2}$ diagonal matrices such that $D_1(i,i) = \Sigma_j A_{ij}$ and $D_2(j,j) = \Sigma_i A_{ij}$

2. Calculate the singular value decomposition (*SVD*) of the normalized matrix $\boldsymbol{A_n}$ to obtain the singular values ($\boldsymbol{\Lambda}$) and the left (the columns of $\boldsymbol{U}$) and right (the columns of $\boldsymbol{V}$) singular vectors

$$SVD(\boldsymbol{A_n}) = \boldsymbol{U} * \boldsymbol{\Lambda} * \boldsymbol{V}^{\mathrm{T}}$$

   and extract the second singular vector $\boldsymbol{u_2}$ from $\boldsymbol{U}$ and $\boldsymbol{v_2}$ from $\boldsymbol{V}$

3. Compute $\boldsymbol{z_2} = \begin{bmatrix} \boldsymbol{D_1}^{-1/2} \boldsymbol{u_2} \\ \boldsymbol{D_2}^{-1/2} \boldsymbol{v_2} \end{bmatrix}$

4. Run the k-means algorithm with $k = 2$ on $\boldsymbol{z_2}$ to obtain the bipartitioning[4]

5. Repeat steps 1 to 4 on both clusters separately to create the hierarchical clustering

To illustrate this procedure, we will co-cluster the following variety-by-segment-correspondence matrix $\boldsymbol{A}$ in two groups (note that this matrix is visualized by Figure 4.2).

	[ʌ]:[ɪ]	[-]:[ə]	[d]:[w]
Appelscha (Friesland)	1	1	0
Oudega (Friesland)	1	1	0
Vaals (Limburg)	0	1	1
Sittard (Limburg)	0	1	1

We first construct matrices $\boldsymbol{D_1}$ and $\boldsymbol{D_2}$. $\boldsymbol{D_1}$ contains the total number of edges from every variety (in the same row) on the diagonal, while $\boldsymbol{D_2}$ contains the total number of edges from every segment substitution (in the same column) on the diagonal. Both matrices are shown below.

[4] As the initialization of the k-means algorithm is random, we repeat the clustering procedure 100 times to ensure a stable bipartitioning.

$$\boldsymbol{D}_1 = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix} \qquad \boldsymbol{D}_2 = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

We can now calculate $\boldsymbol{A}_n$ using the formula displayed in step 1 of the hierarchical bipartite spectral partitioning algorithm:

$$\boldsymbol{A}_n = \begin{bmatrix} 0.5 & 0.35 & 0 \\ 0.5 & 0.35 & 0 \\ 0 & 0.35 & 0.5 \\ 0 & 0.35 & 0.5 \end{bmatrix}$$

Applying the *SVD* to $\boldsymbol{A}_n$ yields:

$$\boldsymbol{U} = \begin{bmatrix} -0.5 & 0.5 & 0.71 & 0 \\ -0.5 & 0.5 & -0.71 & 0 \\ -0.5 & -0.5 & 0 & -0.71 \\ -0.5 & -0.5 & 0 & 0.71 \end{bmatrix} \qquad \boldsymbol{\Lambda} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.71 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\boldsymbol{V}^{\mathrm{T}} = \begin{bmatrix} -0.5 & -0.71 & -0.5 \\ 0.71 & 0 & -0.71 \\ -0.5 & 0.71 & -0.5 \end{bmatrix}$$

For clustering, we use the second singular vector of $\boldsymbol{U}$ (second column) and $\boldsymbol{V}^{\mathrm{T}}$ (second row; i.e. second column of $\boldsymbol{V}$) and compute the 1-dimensional vector $\boldsymbol{z}_2$:

$$\boldsymbol{z}_2 = \begin{bmatrix} 0.35 & 0.35 & -0.35 & -0.35 & 0.5 & 0 & -0.5 \end{bmatrix}^{\mathrm{T}}$$

Note that the first four values correspond with the places (Appelscha, Oudega, Vaals and Sittard) and the final three values correspond to the sound segment substitutions ([ʌ]:[ɪ], [-]:[ə] and [d]:[w]).

After running the k-means algorithm (with random initialization and k = 2) on $\boldsymbol{z}_2$, the items are assigned to one of two clusters as follows:

$$\begin{bmatrix} 1 & 1 & 2 & 2 & 1 & 1 & 2 \end{bmatrix}^{\mathrm{T}}$$

The clustering shows that Appelscha and Oudega are clustered together (corresponding to the first and second item of the vector, above) and linked to the clustered segment substitutions of [ʌ]:[ɪ] and [-]:[ə] (cluster 1). Similarly, Vaals and Sittard are clustered together and linked to the clustered segment

substitution [d]:[w] (cluster 2). Note that the segment substitution [-]:[ə] (an insertion of [ə]) is actually not meaningful for the clustering of the varieties (as can also be observed in $\boldsymbol{A}$), because the middle value of $\boldsymbol{V}^{\mathrm{T}}$ corresponding to this segment substitution equals zero. It could therefore just as likely be grouped in cluster 2. Nevertheless, the k-means algorithm always assigns every item to a single cluster.[5]

This result also illustrates that the connection between a cluster of varieties and sound correspondences does not necessarily imply that those sound correspondences only occur in that particular cluster of varieties. In general, however, sound correspondences will occur less frequently outside of the cluster.

The procedure to determine the importance of sound correspondences in a cluster is discussed next.

4.3.3 Ranking sound correspondences

Before deciding how to calculate the importance of each sound correspondence, we need to consider the characteristics of important sound correspondences.

Note that if a variety contains a sound correspondence, this simply means that the sound correspondence (i.e. two aligned segments) occurs at least three times (our threshold) in any of the aligned pronunciations (with respect to the reference variety Delft).

In the following, we will discuss two characteristics of an important sound correspondence: representativeness and distinctiveness. Representativeness (R) indicates the proportion of varieties v in the cluster c_i which contain the sound correspondence. A value of zero indicates that the sound correspondence does not occur in any of the varieties, while a value of one indicates that the sound correspondence occurs in all varieties in the cluster. This is shown in the formula below for sound correspondence [a]:[b] and cluster c_i:

$$R(\mathrm{a}, \mathrm{b}, c_i) = \frac{|v \text{ in } c_i \text{ containing [a]:[b]}|}{|v \text{ in } c_i|}$$

The second characteristic of an important sound correspondence is distinctiveness. This characteristic indicates how prevalent a sound correspondence [a]:[b] is in its own cluster c_i as opposed to other clusters.

Suppose sound correspondence [a]:[b] is clustered in group c_i. We can count how many varieties v in c_i contain sound correspondence [a]:[b] and how many varieties in the complete dataset contain [a]:[b]. Dividing these two values yields the relative occurrence O of [a]:[b] in c_i:

[5] Note that we could also have decided to drop this sound correspondence. However using our ranking approach (see Section 4.3.3) already ensures that the uninformative sound correspondences are ranked very low.

$$O(\text{a}, \text{b}, c_i) = \frac{|v \text{ in } c_i \text{ containing [a]:[b]}|}{|v \text{ containing [a]:[b]}|}$$

For instance, if [a]:[b] occurs in 20 varieties and 18 belong to c_i, the relative occurrence is 0.9. We intend to capture in this measure how well the correspondence signals the area represented by c_i. While it may seem that this number can tell us if a sound correspondence is distinctive or not, this is not the case. For instance, if c_i consists of 95% of all varieties, the sound correspondence [a]:[b] is not very distinctive for c_i (i.e. we would expect [a]:[b] to occur in 19 varieties instead of 18). To correct for this, we also need to take into account the relative size S of c_i:

$$S(c_i) = \frac{|v \text{ in } c_i|}{|\text{all } v\text{'s}|}$$

We can now calculate the distinctiveness of a sound correspondence by subtracting the relative size from the relative occurrence. Using the previous example, this would yield $0.90 - 0.95 = -0.05$. A positive value indicates that the sound correspondence is distinctive (the higher the value, the more distinctive), while a negative value identifies values which are not distinctive. To ensure the maximum value equals 1, we use a normalizing term in the formula to calculate the distinctiveness D:

$$D(\text{a}, \text{b}, c_i) = \frac{O(\text{a}, \text{b}, c_i) - S(c_i)}{1 - S(c_i)}$$

Values of D below zero (which are unbounded) indicate sound correspondences which are not distinctive, while positive values (≤ 1) signify distinctive sound correspondences.

To be able to rank the sound correspondences based on their distinctiveness and representativeness, we need to combine these two values. A straightforward way to determine the importance I of every sound correspondence based on the distinctiveness and representativeness is to take the average of both values:

$$I(\text{a}, \text{b}, c_i) = \frac{R(\text{a}, \text{b}, c_i) + D(\text{a}, \text{b}, c_i)}{2}$$

It is clear that we might explore more complicated combinations, but for this dataset we regard both representativeness and distinctiveness as equally important.

Because it is essential for an important sound correspondence to be distinctive, we will only consider sound correspondences having a non-negative distinctiveness. As both representativeness and distinctiveness will therefore range between zero and one, the importance will also range between zero and

one. Higher values *within a cluster* indicate more important sound correspondences for that cluster. Since we take the cluster size into account in calculating the distinctiveness, we can also compare the clusters with respect to the importance values of their sound correspondences. Even though — after the first round of partitioning — the hierarchical spectral graph partitioning method is applied repeatedly to a subset of the data (belonging to a cluster, which subsequently has to be split up), the importance of a sound segment correspondence is always calculated with respect to the complete dataset.

Wieling and Nerbonne (2010) reported that the values of the singular vector $\boldsymbol{v}_2$ could be used as an alternative method to determine the most important sound correspondences. This would obviate the need for an external ranking method (as the one introduced above). However, experiments on another dataset (see Chapter 5) revealed that while high (positive) values of $\boldsymbol{v}_2$ were indicative of important sound correspondences for one cluster, low (negative) values did *not* signify important sound correspondences for the other cluster. Rather, it indicated these sound correspondences were unimportant for the first cluster (which does not necessarily imply importance with respect to the other cluster). Consequently, we use the external ranking method to determine the most important sound correspondences in this chapter.[6]

Connection to precision and recall

If we regard the varieties grouped in a certain cluster as the 'target' elements we are seeking and the varieties in which the correspondence occurs as the 'selected' elements, then we clearly see that representativeness (R) is similar to *recall* in information retrieval (Manning and Schütze, 1999, p. 268). Using the same analogy as above, we see relative occurrence (O) is similar to *precision*.[7]

Following this comparison, we note that distinctiveness (D) is a precision measure corrected for chance effects. The numbers in information retrieval would hardly change at all if one corrected for chance, but we examine much smaller sets in dialectology.

4.4 Results

In this section, we will report the results of applying the hierarchical spectral partitioning method to our Dutch dialect dataset.

We will only focus on the four main clusters, each consisting of at least ten varieties. While our method is able to detect smaller clusters in the data, we do not believe these to be stable. We confirmed this by applying three well-known distance-based clustering algorithms (i.e. UPGMA, WPGMA and

[6] Coincidentally, the ranking approach using the values of $\boldsymbol{v}_2$ does work for the hierarchy reported in this chapter (see Wieling and Nerbonne, 2010). This is due to the specific hierarchy (see Figure 4.3), in which for every split only a single cluster is of interest.

[7] We thank Peter Kleiweg for pointing out the parallels to the information retrieval concepts.

Ward's Method; Prokić and Nerbonne, 2009) to our data which also only agreed on four main clusters.

4.4.1 Geographical clustering

Figure 4.3 shows a geographical visualization of the clustering as well as the hierarchy. The first thing to note is that we obtain a sensible geographical grouping. The first split clearly separates the Frisian language area (in the province of Friesland) from the Dutch language area. This is the expected result as Heeringa (2004, pp. 227–229) also measured Frisian as the most distant of all the language varieties spoken in the Netherlands and Flanders, and Frisian is closely related to Anglo-Saxon dialects (Van Bree, 1987, p. 68). In addition, Frisian has the legal status of a different language rather than a dialect of Dutch. Note that the separate 'islands' in the Frisian language area (see Figure 4.3) correspond to the Frisian cities which are generally found to deviate from the rest of the Frisian language area (Heeringa, 2004, pp. 235–241).

Similar to Heeringa (2004, pp. 227–229) we also identify both Limburg and the Low Saxon area (Groningen, Drenthe and Overijssel) as separate groups. Note, however, that both the Limburg and Low Saxon area contain fewer varieties in our clustering than according to traditional dialectology and Heeringa's results (Heeringa, 2004, p. 231; see also Figure 1.3). Our method was also not able to detect additional dialect areas (e.g., the dialect areas of Zeeland or Brabant), accepted in traditional Dutch dialectology (Heeringa, 2004, Ch. 9). However, these dialect areas are less distinctive than the three areas our method does detect (Heeringa, 2004, p. 229). Especially since our method uses simplified data (i.e. binary values indicating if a sound correspondence occurs in a location with respect to a non-ideal reference variety), it might be hard to achieve complete overlap with the regions identified in traditional dialectology.

In the following, we will discuss the three distinctive geographical clusters (i.e. Frisian, Low Saxon and Limburg) together with their simultaneously derived sound correspondences. We will not discuss the group of remaining varieties, as this group is linguistically much less interesting. For brevity, we will only focus on explaining the five most important sound correspondences for each geographical group. The main point to note is that besides a sensible geographical clustering, we also obtain linguistically sensible results.

4.4.2 Characteristic sound correspondences

We report the most important sound correspondences on the basis of their calculated importance score (see Section 4.3.3). The most important sound correspondences reported here, therefore, deviate slightly from the sound correspondences reported by Wieling and Nerbonne (2010) which were based on the values of the second singular vector $\boldsymbol{v}_2$ (see Section 4.3.3 for additional information).

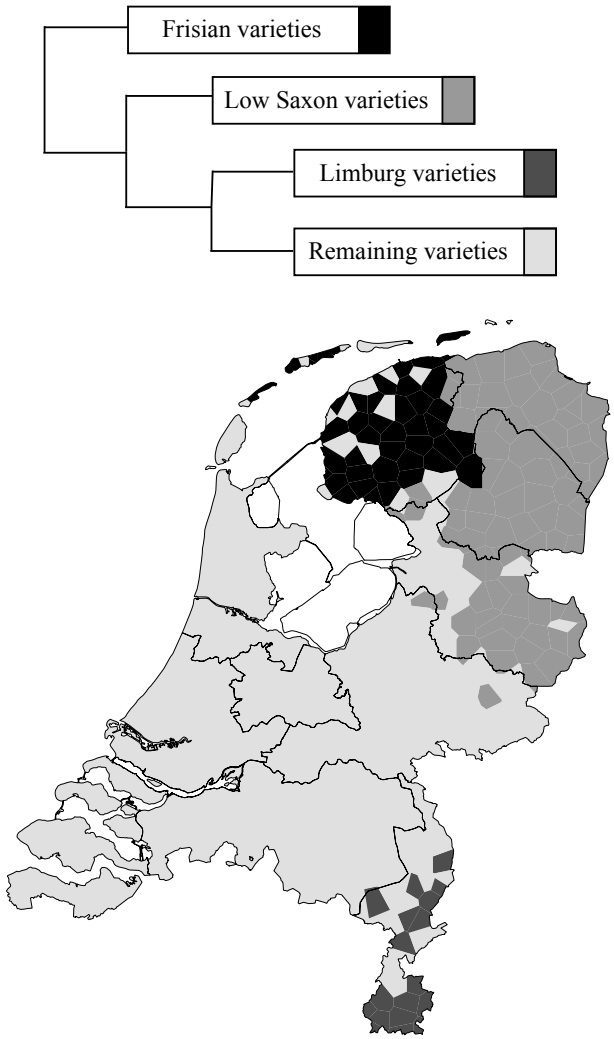


Figure 4.3. Geographical visualization of the clustering with hierarchy. The shades of gray in the hierarchy correspond with the map (e.g., the Limburg varieties can be found in the southeast).

Frisian area

Table 4.2 shows the five most important sound correspondences of the Frisian area. We commonly find the correspondence [-]:[ʃ] (an insertion of [ʃ]) in the infinitive form of verbs such as *wachten* 'wait', Dutch [waxtə], Frisian

Rank	1	2	3	4	5
Reference	[-]	[x]	[f]	[x]	[f]
Frisian	[ʃ]	[j]	[-]	[z]	[w]
Importance	0.95	0.95	0.95	0.94	0.88
Representativeness	0.90	1.00	1.00	0.88	0.85
Distinctiveness	1.00	0.90	0.90	1.00	0.91

Table 4.2. Most important sound correspondences of the Frisian area. The [-] indicates a gap and consequently marks an insertion or a deletion.

[waxtʃə]; *vechten* 'fight', Dutch [vɛxtə], Frisian [vɛxtʃə]; or *spuiten* 'spray', Dutch [spœytə], Frisian [spoʏtʃə], but it also appears (e.g.,) in *tegen* 'against', Dutch [teixə], Frisian [tʃɪm].

The second sound correspondence we identify is [x]:[j] where Dutch /x/ has been lenited, e.g., in *geld* 'money', Dutch [xɛlt], Frisian [jɪlt] (variety of Grouw), but note that [x]:[g] as in [gɛlt] (Franeker) also occurs, illustrating that sound correspondences from another cluster can also occur in the Frisian area.

The [f]:[-] correspondence (a deletion of [f]) is found in words such as *sterven* 'die', Dutch [stɛrfə], Frisian [stɛrə]. The sound correspondence [x]:[z] occurs, e.g., in *zeggen* 'say', Dutch [zɛxə], Frisian [sizə]. Finally, the [f]:[w] sound correspondence occurs in words such as *wrijven* 'rub', Dutch [frɛifə], Frisian [vrywə] (Appelscha).

Note that the previously reported characteristic sound correspondence [a]:[i] between Dutch and Frisian (in words such as *zwaar* 'heavy', [swar] in Dutch and pronounced [swiər] in Holwerd, Friesland; Wieling and Nerbonne, 2010) is not shown in the table as it is ranked sixth.

Low Saxon area

Table 4.3 shows the most important sound correspondences of the Low Saxon area. The first correspondence is interesting because it is not commented on often. We notice the [k]:[ʔ] correspondence where /k/ was historically initial in the weak syllable /-kən/, for example in *planken* 'boards, planks', Dutch [plɑŋkə], pronounced [plɑŋʔŋ] in Low Saxon. Similar examples are provided by *denken* 'think', pronounced [dɛŋʔŋ] in Low Saxon, and *drinken* 'drink', pronounced [drɪŋʔŋ] in Low Saxon.

The [v] corresponds with the [b] in words such as *leven* 'live', Low Saxon [lebm][8] (Aduard); *bleven* 'remain', Low Saxon [blibm] (Aduard); *doven* 'deaf',

[8] Because we only focus on the 82 distinct base sound segments without diacritics, syllabic markers (e.g., [m̩]) were ignored as well and are not shown in the transcriptions in this section.

Rank	1	2	3	4	5
Reference	[k]	[v]	[f]	[p]	[ə]
Low Saxon	[ʔ]	[b]	[b]	[ʔ]	[m]
Importance	0.72	0.71	0.71	0.68	0.68
Representativeness	0.71	0.55	0.53	0.53	1.00
Distinctiveness	0.74	0.87	0.89	0.84	0.35

Table 4.3. Most important sound correspondences of the Low Saxon area

Low Saxon [dubm] (Aduard); *graven* 'dig', Low Saxon [xrabm] (Anloo); and *dieven* 'thieves', Low Saxon [dibm] (Westerbork). In all these cases we encounter a [v] in the reference variety in place of the [b].

The correspondence just discussed involves the lenition, but no devoicing of the stop consonant /b/, but we also have examples where the /b/ has been lenited and devoiced, and this is the third correspondence, to which we now turn. The [f]:[b] correspondence appears in words such as *proeven* 'test', Dutch [prufə], pronounced [proybm] in Low Saxon, in e.g., Barger-Oosterveld and Bellingwolde; *schrijven* 'write', [sxrəɪfə], pronounced [sxribm] in Low Saxon, e.g., in Anloo and Aduard; or *wrijven* 'rub', [frəɪfə], pronounced [vribm] in Low Saxon. Similar examples involve *schuiven* 'shove' and *schaven* 'scrape, plane'. Note that the diphthong in [sxrəɪfə] and [frəɪfə] is not standard Dutch, but rather the pronunciation in our reference variety Delft.

The correspondence [p]:[ʔ] occurs in words such as *lampen* 'lamps', Dutch [lampə], Aduard (Low Saxon) [lamʔm], but also postvocalically, as in *gapen* 'yawn', Dutch [xapə], Aduard (Low Saxon) [xoʔm]. It is obviously related to the [k]:[ʔ] correspondence discussed above.

The best-known characteristic of the Low Saxon area (Goossens, 1977), the so-called *slot-n* ('final-n'), shows up as the fifth sound correspondence. It is instantiated strongly in words such as *strepen*, 'stripes', realized as [strepm] in the northern Low Saxon area. It would be pronounced [strepə] in standard Dutch, so the difference shows up as an unexpected correspondence of [ə] with [m] (but also with [ŋ], rank 6, and with [n], rank 11).

When comparing the importance values of the sound correspondences characteristic of the Low Saxon area to those characteristic of the Frisian area, we clearly see the latter are higher. When including all sound segment correspondences in each cluster, we observe a significant difference between the importance values of the two areas ($t = 3.7$, $p < 0.001$), indicating that the Frisian area is characterized by more distinctive and representative sound correspondences than the Low Saxon area.

Rank	1	2	3	4	5
Reference	[n]	[r]	[s]	[n]	[o]
Limburg	[x]	[x]	[ʒ]	[ʁ]	[-]
Importance	0.75	0.68	0.64	0.63	0.61
Representativeness	0.50	0.67	0.61	0.39	0.67
Distinctiveness	1.00	0.69	0.67	0.87	0.55

Table 4.4. Most important sound correspondences of the Limburg area. The [-] indicates a gap and consequently marks a deletion.

Limburg area

Table 4.4 shows the most important sound correspondences of the Limburg area. The [n]:[x] correspondence appears in words such as *bladen* 'magazines', Dutch [bladən], pronounced [blajəx] in (e.g.,) Roermond and Venlo in Limburg; *graven* 'graves', Dutch [xravən], pronounced [xʁavəx] in Gulpen, Horn and elsewhere in Limburg; and *kleden* '(table)cloths', Dutch [kledən], pronounced [klɛɪdəx] in (e.g.,) Kerkrade and Gulpen in Limburg.

The sound correspondence [r]:[x] can be found in words like *vuur* 'fire', Dutch [fyr], pronounced [vyəx] in Limburg. The third sound correspondence [s]:[ʒ] occurs when comparing the standard-like Delft variety to Limburg varieties in words such as *zwijgen* 'to be silent', [sweixə], Limburg [ʒwiɣə], or *zwemmen* 'swim', [swɛmə], Limburg [ʒwɵmə].

Some regular correspondences merely reflect other, and sometimes more fundamental differences. For instance, the correspondence between [n] and [ʁ] turned out to be a reflection of the older plurals ending in /-r/. For example, in the word *kleden* '(table)cloths', plural *kleden* in Dutch, *kleder* in Limburg.

The final sound correspondence [o]:[-] (a deletion of [o]) can be found in *wonen* 'living', pronounced [wounə] in our reference variety Delft and [wunə] in Limburg. As the standard Dutch pronunciation is actually [wonə], this correspondence is caused by the choice of our reference variety, which is similar, but unfortunately not identical to standard Dutch.

When comparing the importance values of the sound correspondences characteristic of the Limburg area to those characteristic of the Frisian and Low Saxon area, we observe a significant difference between Frisian and Limburg (t = 4.1, p < 0.001), but not between Limburg and Low Saxon (t = 1.2, p = 0.217). Consequently, the Frisian area is characterized by more distinctive and representative sound correspondences than the Low Saxon and Limburg area.

4.5 Discussion

In this chapter we have introduced a novel dialectometric method which simultaneously identifies groups of varieties together with their linguistic basis (i.e. sound segment correspondences). We demonstrated that the hierarchical bipartite spectral graph partitioning method introduced by Dhillon (2001) gave sensible clustering results in the geographical domain as well as for the concomitant linguistic basis.

In line with our discussion in Chapter 1, we are optimistic that the use of techniques such as the one presented in this chapter can be more successful in engaging traditional dialectologists exactly, because the relation between the proposed division into dialect areas and the linguistic basis of the division is directly accessible. This is not the case in dialectometric studies in which aggregate relations form the basis for the division into dialect areas. In those approaches, the sum of linguistic differences is used as an indicator of the relations between varieties. This perspective has its advantages (Nerbonne, 2009), but it has not converted large numbers of dialectologists to the use of exact techniques. While the method introduced in the present chapter is more complicated, its linguistic basis is more accessible.

As mentioned above, we did not have transcriptions of standard Dutch, but instead we used transcriptions of a variety (Delft) close to the standard language. While the pronunciations of most items in Delft were similar to standard Dutch, there were also items which were pronounced differently from the standard. Even though we do not believe that this will influence the detection of the three main geographical clusters (although their shape might change somewhat), using standard Dutch transcriptions produced by the transcribers of the GTRP corpus would make the interpretation of sound correspondences more straightforward.

We already indicated that the geographical clusters detected by the hierarchical bipartite spectral graph partitioning method do not overlap perfectly with the insights from traditional dialectology and on the basis of other dialectometric methods (see Heeringa, 2004, Ch. 9). This might be a problem of the method, which necessarily operates on simplified data (and in this case, using a non-ideal reference variety). However, it might also point to localities which might be similar to their surrounding areas in terms of aggregate linguistic distance, but not in terms of exact shared sound correspondences.

The important sound correspondences found by our procedure are not always the historical correspondences which diachronic linguists build reconstructions on. Instead, they may reflect entire series of sound changes and may involve elements that do not correspond historically at all. We suspect that dialect speakers likewise fail to perceive such correspondences as *general* indicators of another speaker's provenance, except in the specific context of the words such as those in the dataset from which the correspondences are drawn. This means that some manual investigation is still necessary to analyze the charac-

teristic elements of the dialects as well.

This study has improved the techniques available for studying the social dynamics of language variation. In dialect geography, social dynamics are operationalized as geography, and hierarchical bipartite spectral graph partitioning has proven itself capable of detecting the effects of social contact, i.e. the latent geographical signal in the data. In the future, techniques that attempt not just to detect the geographical signal in the data, but moreover to incorporate geography as an explicit parameter in models of language variation (see Chapters 6, 7 and 8) may be in a position to overcome weaknesses inherent in current models. The work presented here aims only at detecting the geographical signal. Other dialectometric techniques have done this as well, but linguists (e.g., Schneider, 1988) have rightly complained that linguistic factors have been neglected in dialectometry (see Chapter 1). This chapter has shown that bipartite spectral graph clustering can detect the linguistic basis of dialectal affinity, and thus provide the information that Schneider and others have missed.

The applicability of this method is not only restricted to Dutch dialects, as Montemagni et al. (accepted) have successfully used it to investigate the spreading of consonantal weakening in Tuscany (i.e. *Gorgia Toscana*). They also showed that the method was useful when sound correspondences were distinguished on the basis of their (left and right) context.

In the next chapter, we will further strengthen the support for the method by applying the hierarchical bipartite spectral graph clustering approach to an English dataset, and also compare it to other, more traditional approaches.

Chapter 5

Clustering English dialects and their features

Abstract. This chapter applies the hierarchical bipartite spectral graph partitioning method introduced in Chapter 4 to phonetic data from the traditional English dialects. We compare the results using this approach to previously published results on the same dataset using cluster and principal component analysis (Shackleton, 2007). While the results of the spectral partitioning method and Shackleton's approach overlap to a broad extent, the three analyses offer complementary insights into the data. The traditional cluster analysis detects some clusters which are not identified by the spectral partitioning analysis, while the reverse also occurs. Similarly, the principal component analysis and the spectral partitioning method detect many overlapping, but also some different linguistic variants. The main benefit of the hierarchical bipartite spectral graph partitioning method over the alternative approaches is its ability to *simultaneously* identify sensible geographical clusters of localities with their corresponding linguistic features.[1]

5.1 Introduction

A great deal of language variation is conditioned geographically, giving rise to geographic dialects, which have been studied in dialectology for well over a century. Dissatisfaction with dialectology's tendency to focus on details gave rise in the 1970s to dialectometry, which systematizes procedures and obviates the need for feature selection, at least to some extent. Nerbonne (2009) argues that dialectometry has been successful because of its emphasis on measuring aggregate levels of differentiation (or similarity), strengthening the geographic signals in the linguistic data, which are often complex and at times even contradictory. As indicated in Chapter 1, the professional reception of dialectometry has been polite but less than enthusiastic, as some scholars express concern that its focus on aggregate levels of variation ignores the kind of linguistic detail that may help uncover the linguistic structure in variation. For this reason there have been several recent attempts to supplement (aggregate) dialectometric techniques with, on the one hand, techniques

[1] This chapter is based on Wieling, Shackleton and Nerbonne (accepted).

to identify linguistic variables which tend to be strongly associated throughout geographic regions and, on the other hand, techniques to extract prominent linguistic features that are especially indicative of aggregate differentiation.

Ruette and Speelman (submitted) introduced a type of three-way multidimensional scaling (i.e. individual differences scaling) to variationist studies. Just as the standard two-way multidimensional scaling technique (Nerbonne, 2010), it operates on a distance matrix to group similar varieties. In addition, however, it also reveals the structure of the underlying linguistic variables.

Building on the ranking approach introduced in Chapter 4, Prokić et al. (2012) examined each item in a dataset seeking those that differ minimally within a candidate area and maximally with respect to sites outside the area.

Grieve et al. (2011) analyzed a large dataset of written English with respect to lexical variation. They used spatial autocorrelation to detect significant geographical patterns in 40 individual lexical alternation variables, and subsequently applied factor analysis to obtain the importance of individual lexical alternation variables in every factor (which can globally be seen as representing a geographical area). In the following step, they applied cluster analysis to the factor scores in order to obtain a geographical clustering.

Shackleton (2007) used cluster analysis and principal component analysis (PCA) to identify linguistic variables which tend to correlate when compared across many localities. We illustrate the basic idea with an example: if the localities in which a standard /æ/ is raised to [ɛ] tend to be the same as those in which /e/ is also raised (to [eɪ]), then a good cluster analysis should identify a cluster of localities that share those variables, while PCA should identify a principal component which is common to the two linguistic variables. Shackleton (2007) identified several interesting clusters and components, which we discuss below at greater length.

In the previous chapter, the hierarchical bipartite spectral graph partitioning (BiSGP) method was introduced, which clusters localities on the basis of the features they share, and features on the basis of the localities in which they co-occur. To continue with the example from the last paragraph, a good BiSGP would identify the two variables as associated and also the sites in which this and other associations are evident. From a dialectometric point of view, BiSGP is attractive in attributing a special status to features as well as to the localities, but like all procedures for seeking natural groups in data, it needs to be evaluated empirically.

In the present chapter we provide more support of the general applicability of the BiSGP analysis by applying it to Shackleton's (2007) data. We compare these results to those on the basis of cluster analysis and PCA reported by Shackleton (2007).

5.2 Material

In this chapter we use the dataset described by Shackleton (2007), derived mainly from Anderson's (1987) *A Structural Atlas of the English Dialects* (henceforth *SAED*). The *SAED* contains more than 100 maps showing the geographical distribution and frequency of occurrence of different phonetic variants in groups of words found in the *Survey of English Dialects* (Orton et al., 1962–1971; henceforth *SED*), the best broad sample of traditional dialect forms that were still in use in 313 rural localities throughout England in the mid-20th century. The dataset assembled from the *SAED* maps classifies over 400 responses from the *SED* by assigning each to one of 39 groups. All of the words in a given group include a segment or combination of segments that is believed to have taken a single uniform pronunciation in the 'standard' Middle English dialect of the Home Counties of southeastern England. The segments include all of the Middle English short and long vowels, diphthongs, and most of the relatively few consonants that exhibit any variation in the English dialects. For each idealized Middle English pronunciation, in turn, the responses may take any of several 20th-century pronunciations, and, in any given location, may take different pronunciations for different words in the group. The dataset thus tabulates frequencies of use for a total of 199 different variant pronunciations of the 39 idealized phonemes. For example, one group includes a number of words, such as *root* and *tooth*, all of which included a segment /oː/ in Middle English. Several maps are associated with that group, one for each modern variant. One of the maps shows the frequency with which /oː/ has become [uː] (that is, the percentage of the words with the vowel articulated as [uː]) in each locality in the *SED*, another shows the frequency with which /oː/ has become [ʏː], and so on. (Throughout this chapter, we write the Middle English form considered common to the group as /x/ and the variants recorded in the *SED* as [x].) The complete list of variants is given by Shackleton (2010, pp. 180–186).[2]

In a few cases, Anderson classified localities from geographically separate regions as having 'different' variants, even though the variants are actually the same, on the grounds that the variant is likely to have arisen independently in the two regions. Moreover, many maps actually show a range of distinguishable pronunciations that Anderson somewhat arbitrarily took to be similar enough to be classified into a single variant. Although it tends to understate the true range of variation in the speech it characterizes, the dataset summarizes a large body of phonetic information in a tractable form that enables straightforward quantitative analyses of phonetic variation in the traditional English dialects.

Most variants have a relatively unique distribution among and frequency of use within localities, and very few large geographic correlations with others. Variants with a large number of high geographic correlations with each other

[2] This list (containing 209 variants) includes ten variants which were not included in this chapter as they did not occur in any of the locations.

are found either in the far southwest or in the far north of England, suggesting that those regions tend to have relatively distinctive speech forms with several features that regularly co-occur in them (exemplified by the very similar geographic distributions of voiced fricatives in the southwest). The comparative lack of geographic correlation raises challenges for analytic techniques, such as the hierarchical bipartite spectral graph partitioning presented here, that seek to identify groups of linguistic features characterizing regional dialects.

5.3 Methods

In this chapter, the bipartite graph is represented by a geographic locality × linguistic variant matrix where every position in the table marks the relative variant frequency (i.e. ranging between zero and one) as used by Shackleton (2007) in his analysis. To ensure every variant carries comparable weight in the analysis, we scaled all individual columns of the matrix (relative variant frequency) between zero and one; that is, for each variant, all of the relative frequencies are divided by the highest relative frequency for that variant.[3] This approach potentially places greater emphasis than other approaches on regionally distinctive but comparatively uncommon variants.[4] After applying the hierarchical bipartite spectral graph partitioning method (explained in Section 4.3.2) to the scaled input matrix, we obtain a hierarchical clustering where localities are clustered together with the linguistic variants.

In line with Section 4.3.3, we rank the importance of a variant in a cluster based on the linear combination of its distinctiveness and representativeness. Normally (see Chapter 4) representativeness and distinctiveness are averaged to obtain the importance score for every variant, but it is also possible to assign different weights to representativeness and distinctiveness. When the input matrix contains many variants which occur almost everywhere, representativeness will be very high for these (non-informative) sound correspondences. In that case it makes sense to weight distinctiveness more heavily than representativeness. Alternatively, if there are many variants occurring only in a few localities, the distinctiveness of these (non-informative) variants will be very high. In that situation, it makes sense to weight representativeness more heavily than distinctiveness. As our matrix contained many frequent variants, we weighted distinctiveness twice as heavily as representativeness.

[3] In Chapter 4 we used a binary matrix (with a threshold), but here we opted to use the scaled values as the *SAED* input matrix already included an aggregation step by having grouped several words, e.g., *root* and *tooth*, and we did not wish to add another aggregation step.

[4] Note that when using the raw frequencies, results were generally similar to those using the scaled frequencies (as most columns already had a maximum value of one).

Figure 5.1. Bipartite spectral graph partitioning in two groups

5.4 Results

Applied to the data from Shackleton (2007), the BiSGP analysis initially distinguishes northern and southern English dialects along a line that roughly traces the border separating northern from Midlands and southern dialects in Shackleton's (2007) cluster analysis. Figure 5.1 shows this division. The southern region includes 198 (63%) of the localities and 123 (62%) of the variants, while the northern region includes the remaining 115 (37%) localities and 76 (38%) variants. A few of the roughly 70 high-scoring southern variants are widely found throughout the region, including those reflecting movements or lengthening of the Middle English short vowels, but most are members of several groups that, on closer inspection, tend to be restricted to areas of the south; these include upgliding diphthongization of the Middle English long vowels (e.g., [lɛin] or [læin] for *lane*) occurring mainly in the southeast, voicing of fricatives and retention of rhoticity (e.g., [vɑrm] for *farm*) largely in the southwest, and fronting of many vowels (e.g., [nʏːn] for *noon*) in Devon. The roughly 50 high-scoring northern variants are similar in that some reflect widely distributed conservative retentions of the Middle English short vowels (e.g., [man] for *man*) along with more restricted ingliding diphthongs for some Middle English long vowels (e.g., [liən] for *lane*), limited retention of rhoticity, and fronting of some vowels (e.g., [bøːn] for *bone*) in the far north.

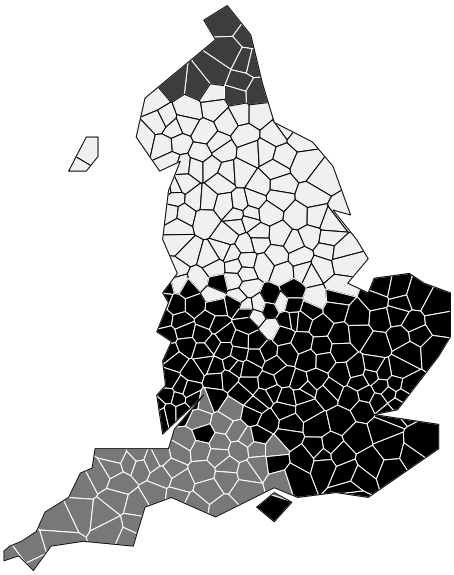


Figure 5.2. Bipartite spectral graph partitioning in four groups

The second round of partitioning divides England into four separate regions (shown in Figure 5.2) , somewhat more clearly reflecting regionally coherent distributions of variants. A small region in the far north emerges, restricted mainly to Northumberland, with 11 localities and 21 variants including retained rhoticity ([ʀ] and, in one location in Cumberland, [r]) and aspirates as well as fronting or ingliding of Middle English long vowels (e.g., [liən] for *lane*), while the rest of the north — 104 localities with 55 variants — includes a number of other features irregularly distributed throughout that region. The southwest — 51 localities with 42 variants — includes the voiced fricatives characteristic of the entire region (e.g., [vɑrm] for *farm*) as well as the fronted vowels characteristic only of Devon (e.g., [nʏːn] for *noon*), while the southeast — 147 localities with 81 variants — includes the upgliding diphthongization of the Middle English long vowels characteristic of much of the southeast (e.g., [lein] or [lain] for *lane*) as well as a number of more sporadically occurring variants.

A further round of partitioning into eight regions (shown in Figure 5.3) yields yet more coherent distributions in the north and south. In the far north, a single locality in Cumberland is distinguished by its alveolar trill [r] (marked with number 1 in Figure 5.3), with the rest of the far north (marked with number 2 in Figure 5.3) characterized by the 'Northumbrian burr' (a uvular trill [ʀ]), retained aspirates, and fronting or ingliding of Middle English long vowels. Most of the remaining northerly localities — 82 localities with 44 variants (marked

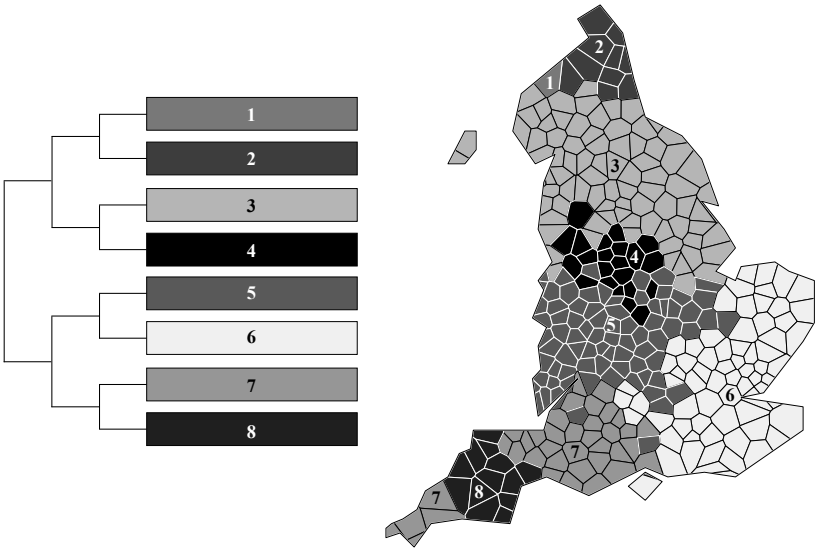


Figure 5.3. Bipartite spectral graph partitioning in eight groups with hierarchy

with number 3 in Figure 5.3) — have irregular distributions of variants, but an irregularly shaped region of 22 localities (marked with number 4 in Figure 5.3) centered on Staffordshire and Derbyshire is associated with 9 unusual variants that, on closer examination, include 5 of the 8 variants that Trudgill (1999) associates together as characteristics of a regional "Potteries" dialect: [tiːl] for *tail*, [bɛɷt] for *boot*, [dain] for *down*, [ʃɛip] for *sheep*, and [koɷt] for *caught*. In the southwest, 13 localities in Devon and Cornwall (marked with number 8 in Figure 5.3) are associated with 14 variants, mainly fronting of Middle English back vowels (e.g., [nʏːn] for *noon*) and the development of a low monophthong for Middle English /iː/ (e.g., [naːf] for *knife*), while the remaining 38 localities (two areas marked with number 7 in Figure 5.3) are associated with 28 other variants, the highest scoring of which nearly all involve the voicing of fricatives and the retention of a retroflex rhotic (e.g., [vaɻm] for *farm*). Much of the southeast (marked with number 6 in Figure 5.3) — 69 localities with 38 variants — is associated with the upgliding diphthongization of the Middle English long vowels (e.g., [lɛin] or [lain] for *lane*) and particularly strong movements of Middle English short vowels (e.g., [mɛn] for *man*), as well as a number of less extensively distributed variants such as those restricted mainly to East Anglia. The rest of the south, including most of the West Midlands (marked with number 5 in Figure 5.3) — 78 localities with 43 variants — is associated only quite loosely with a wide variety of variants that are generally distributed throughout a much wider region, or are found only in more isolated areas. The highest-

scoring variants in this region, for example, include the development (mainly in the Severn Valley) of a back unrounded vowel [ɑː] in *daughter*, *law*, and *cough*, [faiv] for *five* mainly in Shropshire, and the palatalization of Middle English /ɛː/ (e.g., [bjʌnz] for *beans*) in the Southwest Midlands.

5.4.1 Comparison to traditional cluster analysis

The results from the BiSGP analysis can be usefully compared with those that emerge from Shackleton's (2007) cluster analysis of the same data, thus illustrating the comparative strengths of the two approaches. In contrast to the hierarchical bipartite spectral graph partitioning approach described here, cluster analysis may use a variety of techniques to group localities on the basis of some measure of the aggregate similarity of the localities' patterns of usage, rather than optimizing over a balance of representativeness and distinctiveness. Shackleton (2007) applied several different clustering techniques to the English dataset and combined them into a single site × site table of mean cophenetic differences (i.e. distances in dendrograms). He then applied multidimensional scaling to the cophenetic distances in order to reduce the variation in the results to a relatively small, arbitrary number of dimensions that summarize fundamental relationships in the data. For visualization purposes the variation is reduced to three dimensions, which can be mapped onto the RGB color spectrum. The resulting pattern, shown in Figure 5.4, shows many similarities to the eight regions resulting from the BiSGP analysis.

As mentioned above, the demarcation of northern and southern dialect regions is similar to Shackleton's delineation of northern dialects from Midlands and southern dialects, except that the BiSGP analysis classifies a few localities in Shackleton's transitional Central Midlands region into the north. The BiSGP analysis distinguishes almost exactly the same southeastern and southwestern regions as Shackleton on the basis of highly similar sets of dialect features, and does the same for the Northumberland region, except that the BiSGP analysis isolates the single locality in Cumberland by its rhotic trill [r]. Those peripheral regions of the English dialect landscape tend to be distinguished by distinct sets of variants that have comparatively coherent geographic distributions — rhoticity and aspirates in the far north, voicing of fricatives in the southwest, fronting in Devon, and the particularly strong upgliding diphthongization found in the southeast — and that are therefore relatively straightforward to identify.

Differences arise in the two analyses' delineation of dialect regions in the lower north and much of the Midlands, where the various traditional dialect developments tend to be less coherently or much more locally distributed. For example, the BiSGP analysis groups together Shackleton's Upper Southwest with most of his Central Midlands region and consequently does not detect a region corresponding with the Central dialect region as identified by Trudgill (1999), whereas the cluster analysis (partly) does (Shackleton, 2007).

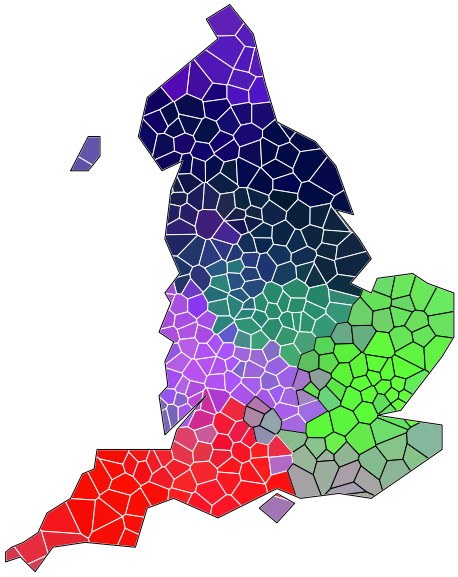


Figure 5.4. Multidimensional scaling visualization of cluster analyses. The original image was taken from Shackleton (2010).

Interestingly, however, the BiSGP analysis identifies a region in the Northwest Midlands, centered on Staffordshire and Derbyshire, on the basis of a number of variants associated by Trudgill (1999) with the "Potteries" region, that Shackleton's cluster analysis consistently fails to distinguish.

5.4.2 Comparison to principal component analysis

The regions resulting from the BiSGP analysis can also be usefully compared to those isolated by Shackleton's varimax principal component analysis of the data, illustrating the comparative strengths of both approaches. In contrast to the BiSGP's focus on representativeness and distinctiveness of variant usage in localities, principal component analysis identifies groups of variants that are strongly positively or negatively correlated — that is, that tend to occur together or that always occur separately — and combines them into principal components that are essentially linear combinations of the correlated variables. A principal component typically has two 'poles', one involving large positive values for a group of variables that tend to be found together, and another involving large negative values for a different group of variables that are also found together but never with the first group. (Varimax rotation tends to sharpen the focus and concentration of each component by increasing the loading on its

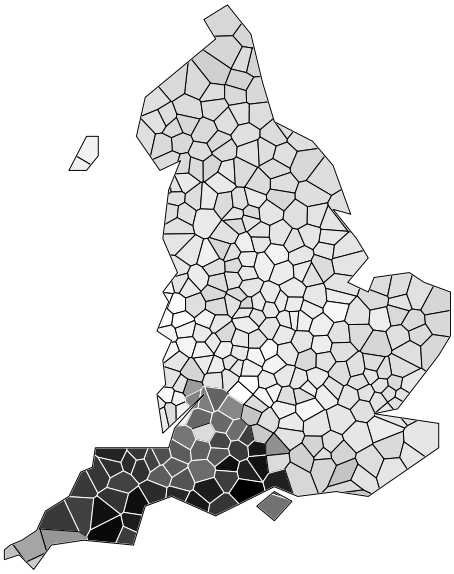


Figure 5.5. Visualization of the component scores for the first varimax rotated principal component. Darker shades of gray indicate higher component scores. The original image was taken from Shackleton (2010).

most highly correlated variants, and when applied to linguistic data, tends to yield groups of variants that are more readily interpretable in linguistic terms.)

Localities can be assigned component scores that indicate the extent to which the variants in a given principal component appear in that particular locality, and in many cases a group of localities may have sufficient geographic cohesion to suggest a dialect region identified by the variants with high scores in that component. Indeed, principal component analysis of the English dataset finds groups of identifying variants for about a dozen regions of England, accounting in the process for roughly half of the variation in the dataset. In some cases, the principal components appear to provide a fairly objective method for characterizing traditional English dialect regions on a quantitative basis. However, unlike the BiSGP analysis, principal component analysis does not comprehensively divide England into regions; moreover, it often isolates variants that are unique to fairly small areas or include variants that are not unique to the relevant region; and few localities in any of the identified regions even use most of the variants identified by the relevant principal component.

For example, as illustrated in Figure 5.5, the high-scoring localities of the first component largely overlap with the broad southwest dialect region identified by the spectral partitioning analysis (regions 7 plus 8 in Figure 5.3), while

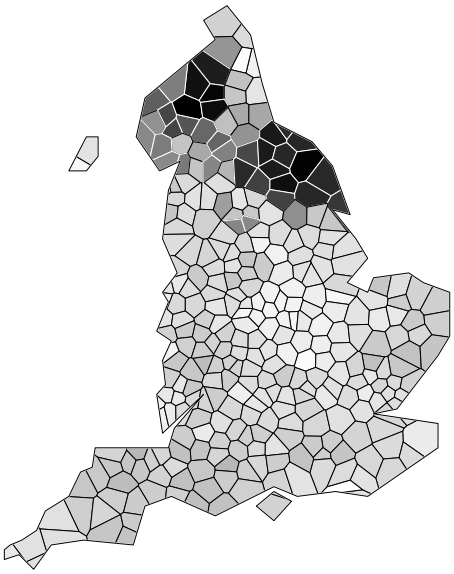


Figure 5.6. Visualization of the component scores for the second varimax rotated principal component. Darker shades of gray indicate higher component scores. The original image was taken from Shackleton (2010).

the loadings indicate that the features most closely associated with the principal component are the voicing of fricatives (also linked to this region by the BiSGP analysis) and occasionally the voicing of medial dentals (e.g., [vɪst] for *fist* and [bʌder] for *butter*), the plausibly related voicing and dentalizing fortition of medial fricative /s/ (e.g., [ɪrnt] for *isn't*), and lowering and unrounding of /u/ (e.g., [bʌt] for *but*). (The principal component also assigns comparatively high loadings to strong rhoticity, as well as to a set of vocalic features that nearly fully describe a nonstandard regional dialect system of vowels, but the rhoticity is not unique to the southwest while the vocalic features appear only sporadically.) Nonetheless, the variants associated with the principal component are never found all together in any single southwestern locality and can only rather loosely be thought of as representing a southwestern dialect. Instead of strictly delineating a dialect region in the manner of the BiSGP analysis, the principal component analysis is at best suggestive of where the region's boundaries might lie.

The second rotated principal component (shown in Figure 5.6) appears to be strongly associated with a large region of the upper north that is not identified by the BiSGP analysis. The defining variants in this principal component all involve the development of ingliding from a high front onset for low long

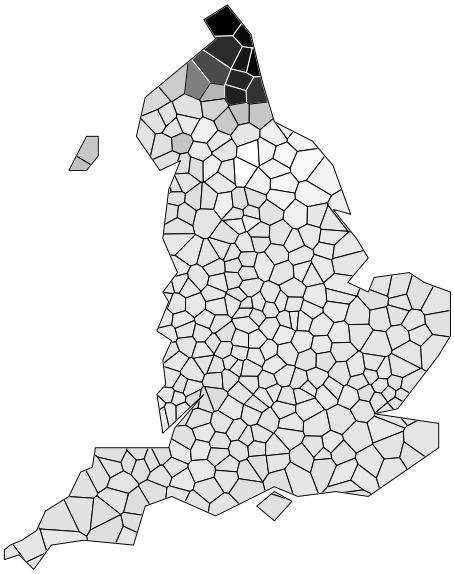


Figure 5.7. Visualization of the component scores for the third varimax rotated principal component. Darker shades of gray indicate higher component scores. The original image was taken from Shackleton (2010).

vowels (e.g., [lɛən] for *lane* and [koəl] for *coal*). The component also includes nearly all of the variants that are the most common regional pronunciations of Middle English long vowels.

The third rotated principal component (shown in Figure 5.7) assigns high component scores to localities in the far north, and assigns high positive loadings to the same variants associated with that region by the BiSGP analysis. In this case, the variants are sufficiently highly correlated with each other and also sufficiently unique to the region to allow both approaches to arrive at essentially the same classification. Note that the area shows some overlap with that of the second principal component (see Figure 5.6).

The fourth principal component (shown in Figure 5.8) rather weakly delineates most of East Anglia on the basis of the development of [w] for /v/ (e.g., [wɪnɪgə] for *vinegar*) and the development of a centered, unrounded onset in /iː/ (e.g., [ʌis] for *ice*). This classification has no counterpart in the BiSGP analysis, but it does appear (and is also similar in terms of characteristic variants) when the number of regions distinguished by the BiSGP approach is increased. We did not discuss this finer-grained division in this chapter, as this also resulted in many additional (uninformative) regions consisting of only a single locality.

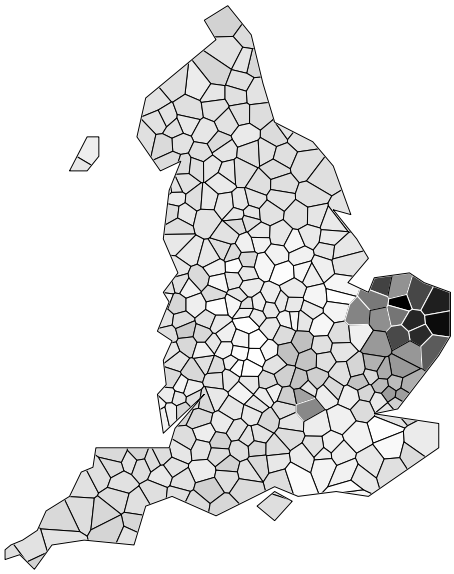


Figure 5.8. Visualization of the component scores for the fourth varimax rotated principal component. Darker shades of gray indicate higher component scores. The original image was taken from Shackleton (2010).

Several other principal components (not shown) match dialect regions identified in the BiSGP analysis. For instance, the sixth principal component distinguishes Devon (region 8 in Figure 5.3) from the rest of the southwest by its unique fronting of back vowels, the same features associated with that region by the BiSGP analysis. Somewhat similarly, the seventh principal component, to a limited extent, distinguishes the "Potteries" zone (region 4 in Figure 5.3) by the use of [iː] and [uː] for /aː/ and /ɔː/, respectively (e.g., [giːt] for *gate* and [guːt] for *goat*). The seventeenth principal component isolates the single locality (region 1 in Figure 5.3) on the Scottish border in Cumberland that uses the alveolar trill [r].

5.5 Discussion

Hierarchical bipartite spectral graph partitioning complements other exact approaches to dialectology by simultaneously identifying groups of localities that are linguistically similar and groups of linguistic variants that tend to co-occur. This introduces an inductive bias in which the linguistic and the geographic dimensions reinforce one another. In a *post hoc* step we identified the most important variants associated with a dialect region, by examining a linear combi-

nation of the variant's distinctiveness (usage frequency in the region as opposed to outside of the region) and representativeness (comparative frequency within the region). That approach contrasts with and complements one-dimensional clustering techniques, which identify regions as groups of localities with similar aggregate patterns of variant frequencies, and principal component analysis, which identifies correlated groups of variants.

Applied to the English dialect dataset used in this chapter, the BiSGP analysis identifies dialect regions that are broadly similar to those identified by cluster analysis and PCA, and isolates sets of variants distinctive for those regions that are also broadly similar to many of the sets identified by principal component analysis. In some cases, however, the BiSGP analysis failed to identify such well-accepted clusters as the Central dialect region (Trudgill, 1999), which was detected (in part) using cluster analysis (Shackleton, 2007). In other cases, most notably in the "Potteries" region, the BiSGP analysis distinguishes regionally distinctive combinations of variants that the other methods largely fail to identify.

Principal component analysis applied to linguistic material identifies groups of variables whose values tend to co-occur with one another. It applies primarily to numerical values, but also works well with frequency counts, once these attain a substantial size. PCA attaches no special value to solutions which privilege finding coherent groups of sites, i.e. finding groups of sites which tend to share strong values for one or more principal components. It is remarkable that (rotated) principal components normally do identify (suggestive) regions, i.e. geographically coherent groups of sites where variables tend to co-vary.

BiSGP seeks partitions of an input matrix that simultaneously identify co-varying linguistic variants (just as PCA does) and also co-varying sites, i.e. sites which share linguistic variants. It is more broadly applicable than PCA, even supporting the analysis of binary data (see Chapter 4). As we are interested in identifying common structure in both the geographic and linguistic dimensions, hierarchical bipartite spectral graph partitioning is intuitively appealing and provides unique insights into the relationships between dialects.

Based on the results reported in Chapter 4 and the present chapter, we hope to have provided adequate support for the suitability of the hierarchical bipartite spectral graph partitioning method in countering one of the main criticisms of dialectometry (see Chapter 1), namely the lack of attention to linguistic factors.

In the following three chapters, we will integrate not only the geographical signal in the data, but also investigate the role of social factors which have been lacking from the clustering approach outlined here.

Part IV

Integrating social, geographical and lexical influences

Chapter 6

Determinants of Dutch dialect variation

Abstract. In this chapter we examine linguistic variation and its dependence on both social and geographical factors. We follow dialectometry in applying a quantitative methodology and focusing on dialect distances, and dialectology in the choice of factors we examine in building a model to predict word pronunciation distances from the standard Dutch language to 424 Dutch dialects. We combine linear mixed-effects regression modeling with generalized additive modeling to predict the pronunciation distance of 559 words. Although geographical position is the dominant predictor, several other factors emerged as significant. The model predicts a greater distance from the standard for smaller communities, for communities with a higher average age, for nouns (as contrasted with verbs and adjectives), for more frequent words, and for words with relatively many vowels. The impact of the demographic variables, however, varied from word to word. For a majority of words, larger, richer and younger communities are moving towards the standard. For a smaller minority of words, larger, richer and younger communities emerge as driving a change away from the standard. Similarly, the strength of the effects of word frequency and word category varied geographically. The peripheral areas of the Netherlands showed a greater distance from the standard for nouns (as opposed to verbs and adjectives) as well as for high frequency words, compared to the more central areas. Our findings indicate that changes in pronunciation have been spreading (in particular for low frequency words) from the Hollandic center of economic power to the peripheral areas of the country, meeting resistance that is stronger wherever, for well-documented historical reasons, the political influence of Holland was reduced. Our results are also consistent with the theory of lexical diffusion, in that distances from the Hollandic norm vary systematically and predictably on a word-by-word basis.[1]

6.1 Introduction

In this chapter, we attempt to integrate the approaches of dialectometry and (social) dialectology. As indicated in Chapter 1, dialectologists often focus on the social dimension of language variation (using a small number of

[1] This chapter is based on Wieling, Nerbonne and Baayen (2011b).

features), whereas researchers in dialectometry mainly investigate dialect geography (aggregating over hundreds of features). We follow dialectometry in viewing the pronunciation distance between hundreds of individual words as our primary dependent variable. In contrast to other dialectometric studies, however, we apply a mixed-effects regression approach,[2] allowing us to simultaneously assess the effect of various sociolinguistic and lexical factors. Furthermore, the strength of the present analysis is that it focuses on individual words in addition to aggregate distances predicted by geography. As a consequence, this quantitative social dialectological study is the first to investigate the effect of a range of social and lexical factors on a large set of dialect distances.

In the following we will focus on building a model to explain the pronunciation distance between dialectal pronunciations (in different locations) and standard Dutch for a large set of distinct words. Of course, choosing standard Dutch as the reference pronunciation is not historically motivated, as standard Dutch is not the proto-language. However, the standard language remains an important reference point for two reasons. First, as noted by Kloeke (1927), in the sixteenth and seventeenth centuries individual sound changes have spread from the Hollandic center of economic and political power to the more peripheral areas of the Netherlands. Furthermore, modern Dutch dialects are known to be converging to the standard language (Wieling et al., 2007a; Van der Wal and Van Bree, 2008, pp. 355–356). We therefore expect geographical distance to reveal a pattern consistent with Kloeke's 'Hollandic Expansion', with greater geographical distance correlating with greater pronunciation distance from the Hollandic standard.

Kloeke (1927) also pointed out that sound changes may proceed on a word-by-word basis. The case for lexical diffusion was championed by Wang (1969) and contrasts with the Neogrammarian view that sound changes are exceptionless and apply to all words of the appropriate form to undergo the change. The Neogrammarian view is consistent with waves of sound changes emanating from Holland to the outer provinces, but it predicts that lexical properties such as a word's frequency of occurrence and its categorial status as a noun or verb should be irrelevant for predicting a region's pronunciation distance to the standard language.

In order to clarify the extent to which variation at the lexical level co-determines the dialect landscape in the Netherlands, we combine generalized additive modeling (which allows us to model complex non-linear surfaces) with mixed-effects regression models (which allow us to explore word-specific variation). First, however, we introduce the materials and methods of our study.

[2] Dialectologists have frequently used logistic regression designs to investigate social influence on linguistic features (Paolillo, 2002). More recently, however, they have also used mixed-effects regression modeling (Johnson, 2009; Tagliamonte and Baayen, in press).

6.2 Material

6.2.1 Pronunciation data

The Dutch dialect dataset (i.e. GTRP) was introduced in detail in Section 4.2 and contains phonetic transcriptions of 562 words in 424 locations in the Netherlands. The transcriptions in the GTRP were made by several transcribers between 1980 and 1995, making it currently the largest contemporary Dutch dialect dataset available. The word categories include mainly verbs (30.8%), nouns (40.3%) and adjectives (20.8%). The complete list of words is presented by Wieling et al. (2007a). For the present study, we excluded three words of the original set (i.e. *gaarne*, *geraken* and *ledig*) as it turned out these words also varied lexically.[3] The standard Dutch pronunciation of all 559 words was transcribed by one of the authors based on Gussenhoven (1999).

Because the set of words included common words (e.g., 'walking') as well as less frequent words (e.g., 'oats'), we included word frequency information, extracted from the CELEX lexical database (Baayen et al., 1996), as an independent variable.

6.2.2 Sociolinguistic data

Besides the information about the speakers recorded by the GTRP compilers, such as year of recording, gender and age of the speaker, we extracted additional demographic information about each of the 424 places from Statistics Netherlands (CBS Statline, 2010). We obtained information about the average age, average income, number of inhabitants (i.e. population size) and male-female ratio in every location in the year 1995 (approximately coinciding with the end of the GTRP data collection period). As Statistics Netherlands uses three measurement levels (i.e. neighborhood, district and municipality), we manually selected the appropriate level for every location. For large cities (e.g., Rotterdam), the corresponding municipality (generally having the same name) was selected as it mainly consisted of the city itself. For smaller cities, located in a municipality having multiple villages and/or cities, the district was selected which consisted of the single city (e.g., Coevorden). Finally, for very small villages located in a district having multiple small villages, the neighborhood was selected which consisted of the single village (e.g., Barger-Oosterveld).

[3] These words were not excluded in Chapters 3 and 4, but they represent only a very small fraction of the complete dataset and are unlikely to affect the results of those chapters significantly.

6.3 Methods

6.3.1 Obtaining pronunciation distances

For all 424 locations, the pronunciation distance between standard Dutch and the dialectal pronunciations (for every individual word) was calculated by using the PMI-based Levenshtein algorithm (diagonal-exclusive version) as explained in Section 2.3.3. In line with Chapter 4, we incorporated some additional linguistic information in the initialization step of the PMI-based Levenshtein algorithm by allowing the alignment of the central vowel [ə] with sonorant consonants (e.g., [m] and [n]), as well as the alignment of semivowels (i.e. [j] and [w]) with both vowels and consonants. Because longer words will likely have a greater pronunciation distance (as more sounds may change) than shorter words, we normalized the PMI-based word pronunciation distances by dividing by the alignment length.

6.3.2 Modeling the role of geography: generalized additive modeling

Given a fine-grained measure capturing the distance between two pronunciations, a key question from a dialectometric perspective is how to model pronunciation distance as a function of the longitude and latitude of the pronunciation variants. The problem is that for understanding how longitude and latitude predict pronunciation distance, the standard linear regression model is not flexible enough. The problem with standard regression is that it can model pronunciation distance as a flat plane spanned by longitude and latitude (by means of two simple main effects) or as a hyperbolic plane (by means of a multiplicative interaction of longitude by latitude). A hyperbolic plane, unfortunately, imposes a very limited functional form on the regression surface that for dialect data will often be totally inappropriate.

We therefore turned to a generalized additive model (GAM), an extension of multiple regression that provides flexible tools for modeling complex interactions describing wiggly surfaces. For isometric predictors such as longitude and latitude, thin plate regression splines are an excellent choice. Thin plate regression splines model a complex, wiggly surface as a weighted sum of geometrically simpler, analytically well defined, surfaces (Wood, 2003). The details of the weights and smoothing basis functions are not of interest for the user, they are estimated by the GAM algorithms such that an optimal balance between undersmoothing and oversmoothing is obtained, using either generalized cross-validation or relativized maximum likelihood (see Wood, 2006 for a detailed discussion). Due to the integration of cross-validation in the GAM fitting procedure, the risk of overfitting is reduced. The significance of a thin plate regression spline is assessed with an F-test evaluating whether the estimated degrees of freedom invested in the spline yield an improved fit of the model

to the data. Generalized additive models have been used successfully in modeling experimental data in psycholinguistics; see Tremblay and Baayen (2010) for evoked response potentials, and see Baayen et al. (2010), Baayen (2010) and Baayen et al. (2011) for chronometric data. They are also widely used in biology, see, for instance, Schmidt et al. (2011) for spatial explicit modeling in ecology.

For our data, we use a generalized additive model to provide us with a two-dimensional surface estimator (based on the combination of longitude and latitude) of pronunciation distance using thin-plate regression splines as implemented in the `mgcv` package for `R` (Wood, 2006). Figure 6.1 presents the resulting regression surface using a contour plot. The (solid) contour lines represent aggregate distance isoglosses. Darker shades of gray indicate smaller distances, lighter shades of gray represent greater distances from the standard language.

The general geographic pattern fits well with Kloeke's hypothesis of a Hollandic expansion: As we move away from Holland, pronunciation distances increase (Kloeke, 1927). Kloeke showed that even in the sixteenth and seventeenth centuries the economic and political supremacy of the provinces of North and South Holland led to the spread of Hollandic speech norms to the outer provinces.

We can clearly identify the separation from the standard spoken in the provinces of North and South Holland (in the central west) of the province of Friesland (in the north), the Low Saxon dialects spoken in Groningen and Drenthe (in the northeast), and the Franconian dialects of Zeeland (in the southwest) and Limburg (in the southeast). The 28.69 estimated degrees of freedom invested in the thin plate regression spline were supported by an F-value of 1051 ($p < 0.0001$). The local cohesion in Figure 6.1 makes sense, since people living in nearby communities tend to speak relatively similar (Nerbonne and Kleiweg, 2007).

6.3.3 Mixed-effects modeling

A problem with this GAM is that the random-effects structure of our dataset is not taken into account. In mixed-effects regression modeling (for introductions, see, e.g., Pinheiro and Bates, 2000; Baayen et al., 2008; Baayen, 2008), a distinction is made between fixed-effect and random-effect factors. Fixed-effect factors are factors with a small number of levels that exhaust all possible levels (e.g., the gender of a speaker is either male or female). Random-effect factors, by contrast, have levels sampled from a much larger population of possible levels. In our data, there are three random-effect factors that are likely to introduce systematic variation that is ignored in our generalized additive model.

A first random-effect factor is location. Our observations are made at 424 locations where speakers were interviewed. Since these 424 locations are a sample of a much larger set of communities that might have been sampled, location is a random-effect factor. Because we used the pronunciations of a single

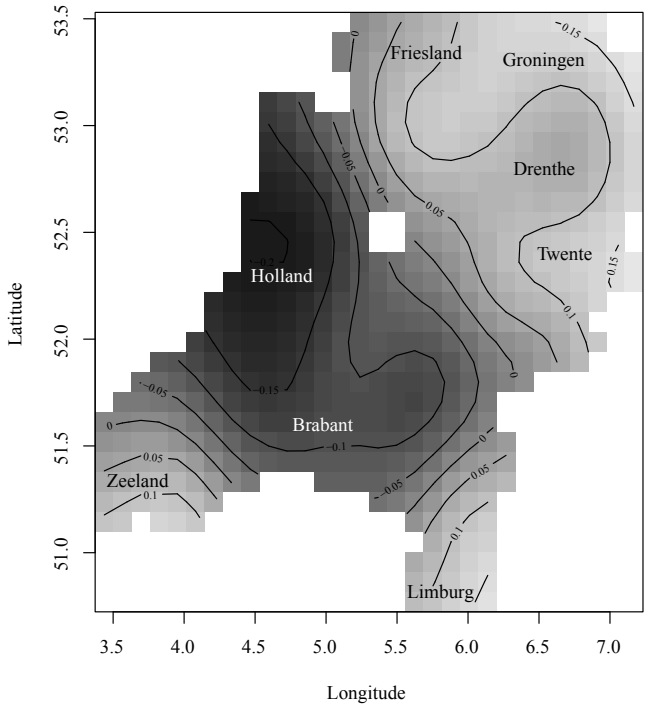


Figure 6.1. Contour plot obtained with a generalized additive model. The contour plot shows a regression surface of pronunciation distance (from standard Dutch) as a function of longitude and latitude obtained with a generalized additive model using a thin plate regression spline. The (black) contour lines represent aggregate distance isoglosses, darker shades of gray indicate smaller distances closer to the standard Dutch language, lighter shades of gray represent greater distances. Note that the empty square indicates the location of the IJsselmeer, a large lake in the Netherlands.

speaker at a given location, location is confounded with speaker. Hence, our random-effect factor location represents both location and speaker.

The data obtained from the 424 locations were coded phonetically by 30 different transcribers. Since these transcribers are themselves a sample of a larger set of possible transcribers, transcriber is a second random-effect factor in our model. By including transcriber in our model, we can account for biases in how individuals positioned the data that they listened to with respect to the standard language.

The third random-effect factor is word. Each of the 559 words was pronounced in most of the 424 locations. The words are also sampled from a much

larger population of words, and hence constitute a random-effect factor as well.

In mixed-effect models, random-effect factors are viewed as sources of random noise that can be linked to specific observational units, in our case, locations, transcribers, and words. In the simplest case, the variability associated with a given random-effect factor is restricted to adjustments to the population intercept. For instance, some transcribers might be biased towards the standard language, others might be biased against it. These biases are assumed to follow a normal distribution with mean zero and unknown standard deviation to be estimated from the data. Once these biases have been estimated, it is possible to adjust the population intercept so that it becomes precise for each individual transcriber. We will refer to these adjusted intercepts as — in this case — by-transcriber random intercepts.

It is possible, however, that the variation associated with a random-effect factor affects not only the intercept, but also the slopes of other predictors. We shall see below that in our data the slope of population size varies with word, indicating that the effect of population size is not the same for all words. A mixed-effects model will estimate the by-word biases in the slope of population size, and by adding these estimated biases to the general population size slope, by-word random slopes are obtained that make the estimated effect of population size as precise as possible for each word.

Whether random intercepts and random slopes are justified is verified by means of likelihood ratio tests, which evaluate whether the increase in the number of parameters is justified given the increase in goodness of fit.

Statistical models combining mixed-effects regression and generalized additive modeling are currently under development. We have explored the `gamm4` package for `R` developed by Simon Wood, but this package proved unable to cope with the rich random-effects structure characterizing our data. We therefore used the generalized additive model simply to predict the pronunciation distance from longitude and latitude, without including any further predictors. We then used the fitted values of this simple model (see Figure 6.1) as a predictor representing geography in our final model. (The same approach was taken by Schmidt et al. (2011), who also failed to use the `gamm4` package successfully.) In what follows, we refer to these fitted values as the GAM distance.

In our analyses, we considered several other predictors in addition to GAM distance and the three random-effect factors location, transcriber, and word. We included a contrast to distinguish nouns (and adverbs, but those only occur infrequently) from verbs and adjectives. Other lexical variables we included were word frequency, the length of the word, and the vowel-to-consonant ratio in the standard Dutch pronunciation of each word. The demographic variables we investigated were average age, average income, male-female ratio and the total number of inhabitants in every location. Finally, the speaker- and transcriber-related variables we extracted from the GTRP were gender, year of birth, year of recording and gender of the fieldworker (not necessarily being the same person as the transcriber). Unfortunately, for about 5% of the locations

the information about gender, year of birth and year of recording was missing. As information about the employment of the speaker or speaker's partner was missing even more frequently (in about 17% of the locations), we did not include this variable in our analysis.

A recurrent problem in large-scale regression studies is collinearity of the predictors. For instance, in the Netherlands, communities with a larger population and higher average income are found in the west of the country. In order to facilitate interpretation, and to avoid enhancement or suppression due to correlations between the predictor variables (Friedman and Wall, 2005), we decorrelated such predictors from GAM distance by using as predictor the residuals of a linear model regressing that predictor on GAM distance. For average age as well as for population size, the resulting residuals correlated highly with the original values ($r \geq 0.97$), indicating that the residuals can be interpreted in the same way as the original values. Because average income and average population age were also correlated ($r = 0.44$) we similarly corrected the variable representing average age for the effect of average income.

In order to reduce the potentially harmful effect of outliers, various numerical predictors were log-transformed (see Table 6.1). We scaled all numerical predictors by subtracting the mean and dividing by the standard deviation in order to facilitate the interpretation of the fitted parameters of the statistical model. Our dependent variable, the pronunciation distance per word from standard Dutch (averaged by alignment length), was also log-transformed and centered. The value zero indicates the mean distance from the standard pronunciation, while negative values indicate a distance closer and positive values a distance farther away from standard Dutch.

The significance of fixed-effect predictors was evaluated by means of the usual t-test for the coefficients, in addition to model comparison likelihood ratio tests and AIC (Akaike Information Criterion; Akaike, 1974). Since our dataset contains a very large number of observations (a few hundred thousand items), the t-distribution approximates the standard normal distribution and factors will be significant ($p < 0.05$) when they have an absolute value of the t-statistic exceeding 2 (Baayen et al., 2008). A one-tailed test (only applicable with a clear directional hypothesis) is significant when the absolute value of the t-statistic exceeds 1.65.

6.4 Results

The total number of cases in our original dataset was 228,476 (not all locations have pronunciations for every word). To reduce the effect of noise in the transcriptions, we eliminated all items in our dataset with a pronunciation distance from standard Dutch larger than 2.5 standard deviations above the mean pronunciation distance for each word. Because locations in the province of Friesland are characterized by having a separate official language (Frisian) with a

	Estimate	Std. err.	*t*-value	Effect size
Intercept	-0.0153	0.0105	-1.4561	
GAM distance (geography)	0.9684	0.0274	35.3239	0.3445
Population size (log)	-0.0069	0.0026	-2.6386	-0.0402
Average population age	0.0045	0.0025	1.8049	0.0295
Average population income (log)	-0.0005	0.0026	-0.1988	-0.0042
Word frequency (log)	0.0198	0.0060	3.2838	0.1205
Noun instead of Verb/Adjective	0.0409	0.0122	3.3437	0.0409
Vowel-consonant ratio (log)	0.0625	0.0059	10.5415	0.3548

Table 6.1. Fixed-effect coefficients of a minimally adequate model fitted to the pronunciation distances from standard Dutch. Effect size indicates the increase or decrease of the dependent variable when the predictor value increases from its minimum to its maximum value (i.e. the complete range).

relatively large distance from standard Dutch, we based the exclusion of items on the means and standard deviation for the Frisian and non-Frisian area separately. After deleting 2610 cases (1.14%), our final dataset consisted of 225,866 cases.

We fitted a mixed-effects regression model to the data, step by step removing predictors that did not contribute significantly to the model fit. In the following, we will discuss the specification of the resulting model including all significant predictors and verified random-effect factors. This model explains approximately 44.5% of the variance of our dependent variable (i.e. the pronunciation distance from standard Dutch).

The coefficients and associated statistics of the fixed-effect factors and covariates are shown in Table 6.1 (note that most values in the table are close to zero as we are predicting average PMI distances, which are relatively small). To allow a fair comparison of the effect of each predictor, we included a measure of effect size by specifying the increase or decrease of the dependent variable when the predictor increased from its minimum to its maximum value (following the approach of Baayen et al., 2008). Clearly geography and the word-related predictors have the greatest influence on the pronunciation distance from standard Dutch.

The random-effects structure is summarized in Table 6.2. The residuals of our model followed a normal distribution, and did not reveal any non-uniformity with respect to location. Table 6.3 summarizes the relation between the independent variables and the distance from standard Dutch. A more detailed interpretation is provided in the sections below on demographic and lexical predictors.

The inclusion of the fixed-effect factors (except average population income)

Factor	Rnd. effect	Std. dev.	Cor.	
Word	Intercept	0.1394		
	Population size (log)	0.0186		
	Average population age	0.0086	-0.856	
	Average population income (log)	0.0161	0.867	-0.749
Location	Intercept	0.0613		
	Word frequency (log)	0.0161	-0.084	
	Noun instead of Verb/Adjective	0.0528	-0.595	0.550
Transcriber	Intercept	0.0260		
Residual		0.2233		

Table 6.2. Random-effect parameters of the minimally adequate model fitted to the pronunciation distances from standard Dutch. The columns Cor. contain the correlations between the random effects. The first column contains the correlations with the by-word random slope for population size (top two values), and the by-location random intercept (bottom two values). The second column contains the correlation with the by-word random slope for average age (-0.749), and the by-location random slope for word frequency (0.550). See the text for interpretation.

and random-effect factors shown in Table 6.1 and 6.2 was supported by likelihood ratio tests indicating that the additional parameters significantly improved the goodness of fit of the model. Tables 6.4 and 6.5 show the increase in goodness of fit for every additional factor measured by the increase of the log-likelihood and the decrease of the Akaike Information Criterion (Akaike, 1974). To assess the influence of each additional fixed-effect factor, the random effects were held constant, including only the random intercepts for word, location and transcriber. The baseline model, to which the inclusion of the first fixed-effect factor (geography) was compared, only consisted of the random intercepts for word, location and transcriber. Subsequently, the next model (including both geography and the vowel-to-consonant ratio per word), was compared to the model including geography (and the random intercepts) only. This is shown in Table 6.4 (sorted by decreasing importance of the individual fixed-effect factors). Log-likelihood ratio tests were carried out with maximum likelihood estimation, as recommended by Pinheiro and Bates (2000).

Similarly, the importance of additional random-effect factors was assessed by restricting the fixed-effect predictors to those listed in Table 6.1. The baseline model in Table 6.5, to which the inclusion of the random intercept for word was compared, only consisted of the fixed-effect factors listed in Table 6.1. The next model (also including location as a random intercept) was compared to the model with only word as a random intercept. In later steps random slopes were added. For instance, the sixth model (including random slopes for pop-

Predictor	Interpretation
GAM distance (geography)	Peripheral areas in the north, east and south have a higher distance from standard Dutch than the central western part of the Netherlands (see Figure 6.1).
Population size (log)	Locations with a larger population have a pronunciation closer to standard Dutch, but the effect varies per word (see Figure 6.4).
Average population age	Locations with a younger population have a pronunciation closer to standard Dutch, but the effect varies per word (see Figure 6.4).
Average population income (log)	There is no significant general effect of average income in a population, but the effect varies per word (see Figure 6.4).
Word frequency (log)	More frequent words have a higher distance from standard Dutch, but the effect varies per location (see Figure 6.5).
Noun instead of Verb/Adjective	Nouns have a higher distance from standard Dutch than verbs and adjectives, but the effect varies per location (see Figure 6.6).
Vowel-consonant ratio (log)	Words with relatively more vowels have a higher distance from standard Dutch.

Table 6.3. Interpretation of significant predictors

ulation size and average population age, as well as their correlation) was compared to the fifth model which only included population size as a random slope. Log-likelihood ratio tests evaluating random-effect parameters were carried out with relativized maximum likelihood estimation, again following Pinheiro and Bates (2000).

Due to the large size of our dataset, it proved to be computationally infeasible to include all variables in our random-effects structure (e.g., the vowel-to-consonant ratio was not included). As further gains in goodness of fit are to be expected when more parameters are invested in the random-effects structure, our model does not show the complete (best) random-effects structure. However, we have checked that the fixed-effect factors remained significant when additional uncorrelated by-location or by-word random slopes were included in the model specification. In other words, we have verified that the *t*-values of

	LL+	AIC-	LL ratio test
Random intercepts			
+ GAM distance (geography)	270.6	539.2	$p < 0.0001$
+ Vowel-consonant ratio (log)	50.9	99.8	$p < 0.0001$
+ Noun instead of Verb/Adjective	5.6	9.2	$p = 0.0008$
+ Population size	3.8	5.7	$p = 0.0056$
+ Word frequency (log)	3.8	5.7	$p = 0.0056$
+ Average population age	2.5	3.1	$p = 0.0244$
+ Average population income (log)	0.0	-2.0	$p = 0.9554$

Table 6.4. Goodness of fit of the fixed-effect factors of the model. Each row specifies the increase in goodness of fit (in terms of log-likelihood increase and AIC decrease) obtained by adding the specified predictor to the model including all preceding predictors (as well as the random intercepts for word, location and transcriber). Note that the final row indicates that including average income does not improve the model.

	LL+	AIC-	LL ratio test
Fixed-effect factors			
+ Random intercept word	32797.8	65593.6	$p < 0.0001$
+ Random intercept location	5394.2	10786.4	$p < 0.0001$
+ Random intercept transcriber	14.0	26.1	$p < 0.0001$
+ Population size (word)	490.3	978.6	$p < 0.0001$
+ Average population age (word)	96.0	188.0	$p < 0.0001$
+ Average population income (word)	443.9	881.8	$p < 0.0001$
+ Word frequency (location)	220.1	436.3	$p < 0.0001$
+ Noun instead of Verb/Adj. (location)	1064.4	2122.8	$p < 0.0001$

Table 6.5. Goodness of fit of the random-effect factors of the model. Each row specifies the increase in goodness of fit of the model (in terms of log-likelihood increase and AIC decrease) resulting from the addition of the specified random slope or intercept to the preceding model. All models include the fixed-effect factors listed in Table 6.1.

the fixed-effect factors in Table 6.1 are not anti-conservative and therefore our results remain valid.

Interpreting fixed and random effects

Because mixed-effects regression is not (yet) a common technique in linguistic analysis, we will attempt to explain the technique in more detail on the basis of

the results shown in Tables 6.1 and 6.2. While this section focuses on the technique, the interpretation of the results is discussed in Section 6.4.1 and 6.4.2.

We start with the explanation of fixed-effect factors and covariates, as these may be interpreted in the same way as standard linear regression results and should be most familiar. Table 6.1 shows that the intercept is not significant (the absolute t-value is lower than 2). This indicates that, in general, the (centered) distance from the standard language for a certain word in a certain location does not differ from zero (which is unsurprising, as the mean of the *centered* distance from the standard is zero). If we write this in a simple regression formula, $y = ax + b$, b represents the intercept and is set to zero. Consequently the value of y (in our case the pronunciation distance from the standard) is only dependent on the coefficient a and the actual value of the predictor x. If x represents population size, the value of a equals −0.0069, and indicates that if a population in location A is one (log) unit larger than in location B, the distance (for individual words) from standard Dutch will be 0.0069 lower in location A than in location B. It is straightforward to include more predictors in a linear regression formula by simply adding them (e.g., $y = a_1x_1 + a_2x_2 + \ldots + a_nx_n + b$). When considering that the coefficient for population income does not significantly differ from zero and y represents the pronunciation distance from standard Dutch for an individual word in a certain location, the general regression formula for this model (based on Table 6.1) equals:

$$y = 0.9684x_g - 0.0069x_s + 0.0045x_a + 0x_i + 0.0198x_f + 0.0409x_n + 0.0625x_v + 0$$

This formula can also be used to predict the pronunciation distance for a novel word. In that case, we would substitute the lexical variables with the specific values for that word (e.g., x_f is substituted by the logarithm of the frequency of the word). The pronunciation distance of the novel word in every individual location can then be calculated by substituting the location-specific variables as well.

Random-effect factors may be conceptualized as a tool to make precise regression lines for every individual factor included in the random-effects structure (e.g., every individual word). In the following we will focus on only four words: *voor* 'in front of', *wijn* 'wine', *slim* 'smart' and *twee* 'two'. While the intercept of the general model does not significantly differ from 0, some words might have a higher pronunciation distance from standard Dutch (and consequently a higher intercept) than others, even after taking into account various lexical predictors. Indeed, Table 6.2 shows that the intercepts for individual words vary to a large extent (a standard deviation of 0.1394). What might this mean linguistically? There are several possible interpretations, but they all come down to supposing that certain words simply lend themselves more to variation than others.

Figure 6.2 visualizes the varying (i.e. random) intercepts for each of the four words. We clearly see that the word *voor* has a higher intercept than the intercept of the general model (the dashed line; note that it shows the exact value

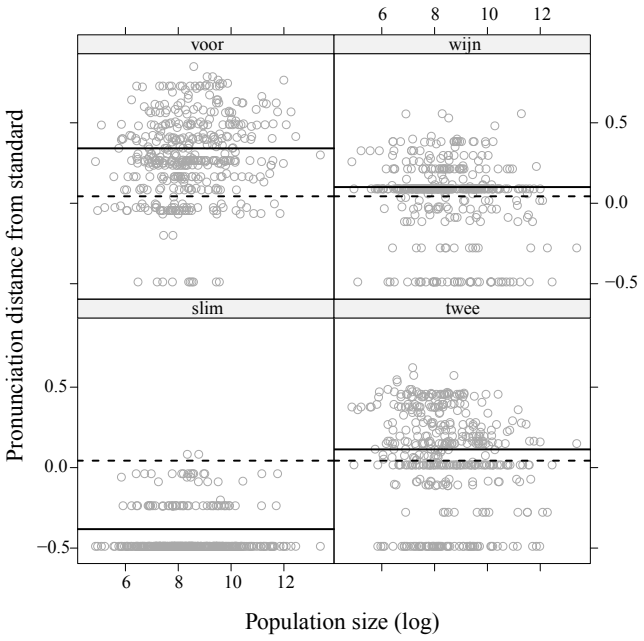


Figure 6.2. Example of random intercepts per word. The dashed line represents the general model estimate (non-significantly different from zero), while the solid lines represent the estimates of the intercept for each individual word.

of the intercept, even though it does not significantly differ from zero). Consequently, *voor* is more different from standard Dutch than the average word. The words *wijn* and *twee* each have an intercept which is only slightly higher than the intercept of the general model. The word *slim* shows the opposite pattern, as its intercept is lower than the intercept of the general model. When looking at the word-specific regression formulas, the intercept (the b in $y = ax + b$) will not be 0, but rather 0.1 (*wijn* and *twee*), 0.4 (*voor*), and -0.4 (*slim*). It is straightforward to see that this explanation generalizes to the random intercepts for location and transcriber.

Similar to the intercept, the regression coefficients, or slopes (i.e. a in $y = ax + b$) may also vary to make the regression formula as precise as possible for every individual word. When looking at the regression formula for individual words, only coefficients which are *not* connected to word-related predictors may vary (i.e. word frequency itself already varies per word, so it makes no sense to vary the effect of word frequency per word as well). For example, the effect (i.e. coefficient) of population size might vary from word to word. Some

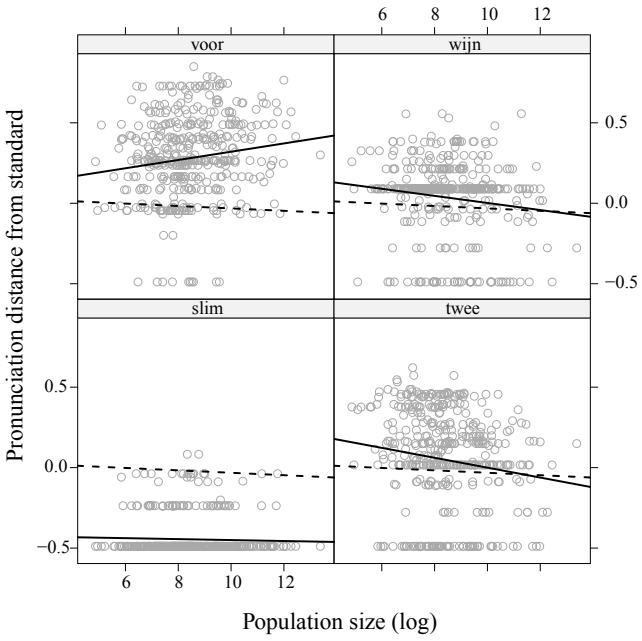


Figure 6.3. Example of random slopes for population size per word. The dashed line represents the general model estimate (of the intercept and the coefficient for population size), while the solid lines represent the estimates of the intercept and the slope for each individual word.

words may be influenced strongly by the population size of a location (e.g., there is a large difference in how close the word is to standard Dutch in a large city, as opposed to a small village), while others might not be influenced at all by population size. The second line of Table 6.2 indeed shows that the effect of population size varies significantly per word (the standard deviation equals 0.0186). Figure 6.3 visualizes the varying coefficients (i.e. random slopes) for each of the four words. We clearly see that the word *slim* has a slightly negative coefficient (i.e. slope), which is somewhat less negative than the coefficient of the general model, while the words *wijn* and *twee* have a more negative slope. In contrast, the word *voor* has a positive slope, which is completely opposite to the pattern of the general model (i.e. the word *voor* is more different from the standard in larger cities than in smaller villages, rather than the other way around). When looking at the regression formulas specific for these words, the coefficient of population size will not be -0.0069, but rather -0.015 (*wijn* and *twee*), 0.015 (*voor*), and -0.005 (*slim*). It is straightforward to see that this

explanation easily generalizes to other predictors and random-effect factors.

In the following, we will use this information to discuss the demographic and lexical results in more detail.

6.4.1 Demographic predictors

The geographical predictor GAM distance (see Figure 6.1) emerged as the predictor with the smallest uncertainty concerning its slope, as indicated by the huge *t*-value. As GAM distance represents the fitted values of a generalized additive model fitted to pronunciation distance from standard Dutch (adjusted $R^2 = 0.12$), the strong statistical support for this predictor is unsurprising. Even though GAM distance accounts for a substantial amount of variance, location is also supported as a significant random-effect factor, indicating that there are differences in pronunciation distances from the standard language that cannot be reduced to geographical proximity. The random-effect factor location, in other words, represents systematic variability that can be traced to the different locations (or speakers), but that resists explanation through our demographic fixed-effect predictors. To what extent, then, do these demographic predictors help explain pronunciation distance from the standard language over and above longitude, latitude, and the location (speaker) itself?

Table 6.1 lists two demographic predictors that reached significance. First, locations with many inhabitants (a large population size) tend to have a lower distance from the standard language than locations with few inhabitants. A possible explanation for this finding is that people tend to have weaker social ties in urban populations, which causes dialect leveling (Milroy, 2002). Since the standard Dutch language has an important position in the Netherlands (Smakman, 2006; Van der Wal and Van Bree, 2008), and has been dominant for many centuries (Kloeke, 1927), conversations between speakers of different dialects will normally be held in standard Dutch and consequently leveling will proceed in the direction of standard Dutch. The greater similarity of varieties in settlements of larger size is also consistent with the predictions of the gravity hypothesis which states that linguistic innovation proceeds first from large settlements to other large nearby settlements, after which smaller settlements adopt the innovations from nearby larger settlements (Trudgill, 1974a).

The second (one-tailed) significant demographic covariate is the average age of the inhabitants of a given location. Since younger people tend to speak less in their dialect and more in standard Dutch than the older population (Heeringa and Nerbonne, 1999; Van der Wal and Van Bree, 2008, pp. 355–356), the positive slope of average age is as expected.

Note that Table 6.1 also contains average income as a demographic covariate. This variable is not significant in the fixed-effects structure of the model (as the absolute *t*-value is lower than 1.65), but is included as it is an important predictor in the random-effects structure of the model.

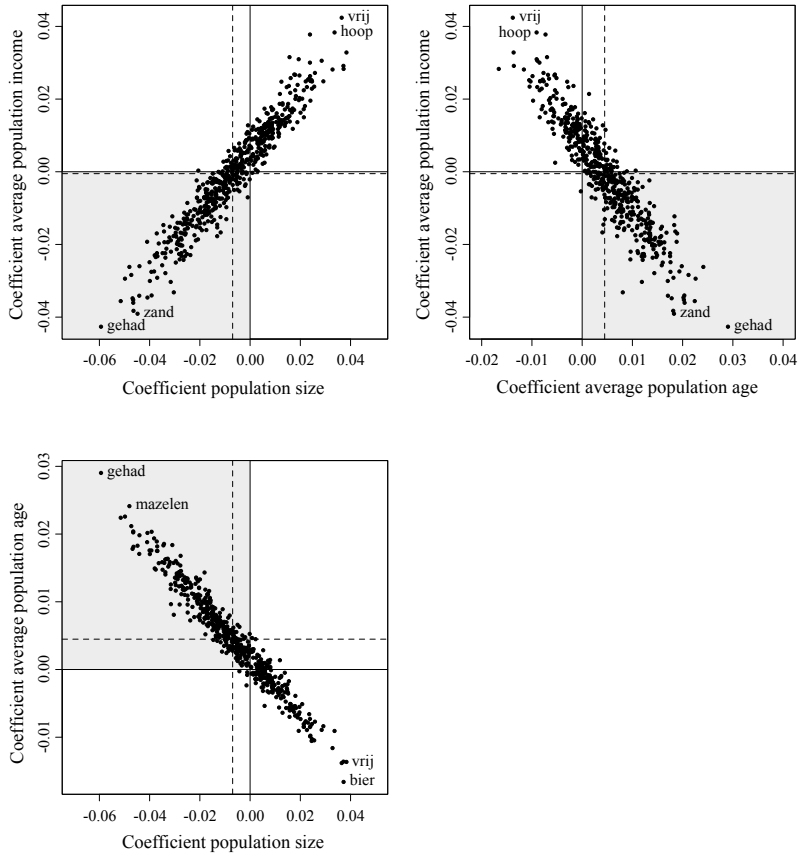


Figure 6.4. By-word random slopes in a mixed-effects regression model fitted to pronunciation distances from standard Dutch. All three possible combinations of by-word random slopes (i.e. the word-specific coefficients) for population size, age and income are shown. The gray quadrant in every graph marks where most words (dots) are located. The dashed lines indicate the model estimates of every predictor.

Interestingly, all three demographic predictors required by-word random slopes (i.e. varying coefficients per word, as explained above). Figure 6.4 shows the by-word random slopes for all combinations of population size (i.e. the number of inhabitants), average age and average income. At the extremes in every graph, the words themselves have been added to the scatter plot (*gehad*, 'had'; *zand*, 'sand'; *hoop*, 'hope'; *vrij*, 'free'; *mazelen*, 'measles'; *bier*, 'beer'). The

gray quadrant in every graph marks where most words are located. Words in this quadrant have slopes consistent with the general model (the model estimates shown in Table 6.1 are indicated by the dashed lines).

When looking at the top-left graph, we see that most words (represented by dots) are located in the lower-left quadrant, consistent with the negative slope of population size (-0.0069) and the (non-significant) negative slope of average income (-0.0005; see Table 6.1). Words in this quadrant have negative slopes for population size, indicating that these words will tend to be more similar to the standard in larger communities (the more to the left the dot is located, the more similar it will be to the standard language). At the same time, the same words also have negative slopes for average income, indicating that these words will tend to be more similar to the standard in richer communities (the lower the dot is located, the more similar it will be to the standard language). This pattern reverses for the words in the opposite quadrant. A word such as *vrij* (free) has a large positive coefficient for population size, indicating that in larger communities this word will differ more from the standard language. The word *vrij* also has a positive coefficient for average income. Therefore, speakers in poorer communities will pronounce the word closer to the standard, while speakers in richer communities will pronounce it more differently. The correlation parameter of 0.867 in Table 6.2 quantifies the strong connection between the by-word random slopes for average income and population size.

The top-right graph illustrates that the coefficients of average age and average income are also closely linked per word (indicated by the high correlation parameter of -0.749 in Table 6.2). Words in the gray quadrant behave in accordance with the general model (e.g., the word *gehad* will be more similar to the standard language in a richer community as well as in a younger community), while words in the opposite quadrant behave in a reversed fashion (e.g., the word *vrij* will differ more from the standard in a richer community as well as in a younger community).

Finally, the bottom-left graph shows that the coefficients of population size and average age are also closely connected per word (indicated by the high correlation parameter of -0.856 in Table 6.2). Words in the gray quadrant behave in accordance with the general model (e.g., the word *gehad* will be more similar to the standard language in a larger community as well as in a younger community), while words in the opposite quadrant behave in a reversed fashion (e.g., the word *bier* will differ more from the standard in a larger community as well as in a younger community).

Two important points emerge from this analysis. First, the effects of the three demographic variables, population size, average age and average income, differ dramatically depending on what word is being considered. Second, words tend to be influenced by all three demographic variables similarly. If a word is influenced more strongly by one variable than predicted by the general model, it will likely also be influenced more strongly by the other two variables (e.g., the word *gehad*). Alternatively, if a word is influenced in the reverse di-

rection by one variable compared to the general model, it will likely also be influenced in the reverse direction by the other two variables (e.g., the word *vrij*).

Besides these significant variables, we investigated several other demographic predictors that did not reach significance. One variable we considered was the male-female ratio at a given location. While the gender of the speaker is likely to play an important role, we are uncertain if the ratio of men versus women in a location should play a significant role. With other predictors in the model, it did not prove significant. We expected a negative influence of average income, since standard Dutch has a relatively high prestige (Van der Wal and Van Bree, 2008, Ch. 12). However, as shown in Table 6.1, this effect did not reach significance, possibly due to the large collinearity with geography; the highest average income in the Netherlands is earned in the western part of the Netherlands (CBS Statline, 2010), where dialects are also most similar to standard Dutch (Heeringa, 2004, p. 274). Average income was highly significant when geography was excluded from the model.

No speaker-related variables were included in the final model. We were surprised that the gender of the speaker did not reach significance, as the importance of this factor has been reported in many sociolinguistic studies (Cheshire, 2002). However, when women have a limited social circle (e.g., the wife of a farmer living on the outskirts of a small rural community), they actually tend to speak more traditionally than men (Van der Wal and Van Bree, 2008, p. 365). Since such women are likely present in our dataset, this may explain the absence of a gender difference in our model. We also expected speaker age to be a significant predictor, since dialects are leveling in the Netherlands (Heeringa and Nerbonne, 1999; Van der Wal and Van Bree, 2008, pp. 355–356). However, as the speakers were relatively close in age (e.g., 74% of the speakers were born between 1910 and 1930) and we only used pronunciations of a single speaker per location, this effect might have been too weak to detect.

The two fieldworker-related factors (gender of the fieldworker and year of recording) were not very informative, because they suffered from substantial geographic collinearity. With respect to the year of recording, we found that locations in Friesland were visited quite late in the project, while their distances from standard Dutch were largest. Regarding the gender of the fieldworkers, female fieldworkers mainly visited the central locations in the Netherlands, while the male fieldworkers visited the more peripheral areas (where the pronunciation distance from standard Dutch is larger).

6.4.2 Lexical predictors

Table 6.1 lists three lexical predictors that reached significance: the vowel-to-consonant ratio (having the largest effect size), word frequency, and the contrast between nouns and verbs. Unsurprisingly, the length of the word was not

a significant predictor, as we normalized pronunciation distance by the alignment length.

The first significant lexical factor was the vowel-to-consonant ratio. The general effect of the vowel-to-consonant ratio was linear, with a greater ratio predicting a greater distance from the standard. As vowels are much more variable than consonants (Keating et al., 1994), this is not a very surprising finding.

The second, more interesting, significant lexical factor was word frequency. More frequent words tend to have a higher distance from the standard. We remarked earlier that Dutch dialects tend to converge to standard Dutch. A larger distance from the standard likely indicates an increased resistance to standardization. Indeed, given the recent study of Pagel et al. (2007), where they showed that more frequent words were more resistant to change, this seems quite sensible.

However, the effect of word frequency is not uniform across locations, as indicated by the presence of by-location random slopes for word frequency in our model (see Table 6.2). The parameters for these random slopes (i.e. the standard deviation, and the parameter measuring the correlation with the random intercepts) jointly increase the log-likelihood of the model by no less than 220 units, compared to 3.8 log-likelihood units for the fixed-effect (population) slope of frequency. Interestingly, although the by-location random slopes for frequency properly follow a normal distribution, they are not uniformly distributed across the different regions of the Netherlands, as illustrated in the upper right panel of Figure 6.5. In this panel, contour lines link locations for which the slope of the frequency effect is the same. The two dark gray areas (central Holland and Groningen and Drenthe) are characterized by slopes close to zero, while the white area in Friesland indicates a large positive slope (i.e. the Frisian pronunciations become more distinct from standard Dutch for higher-frequency words).

To clarify how geography (GAM distance) and frequency jointly predict distance from the standard language, we first calculated the fitted GAM distance for each location. We then estimated the predicted distance from the standard language using GAM distance and word frequency as predictors, weighted by the coefficients estimated by our mixed-effects model. Because the fitted surfaces vary with frequency, we selected the minimum frequency (Q0), first (Q1) and third (Q3) quartiles, as well as the maximum frequency (Q4) for visualization (see the lower panels of Figure 6.5). Panel Q0 shows the surface for the words with the lowest frequency in our data. As frequency increases, the surface gradually morphs into the surface shown in the lower right panel (Q4). The first thing to note is that as frequency increases, the shades of gray become lighter, indicating greater differences from the standard. This is the main effect of frequency: higher-frequency words are more likely to resist assimilation to the standard language. The second thing to note is that the distances between the contour lines decrease with increasing frequency, indicating that the differences between regions with respect to the frequency effect

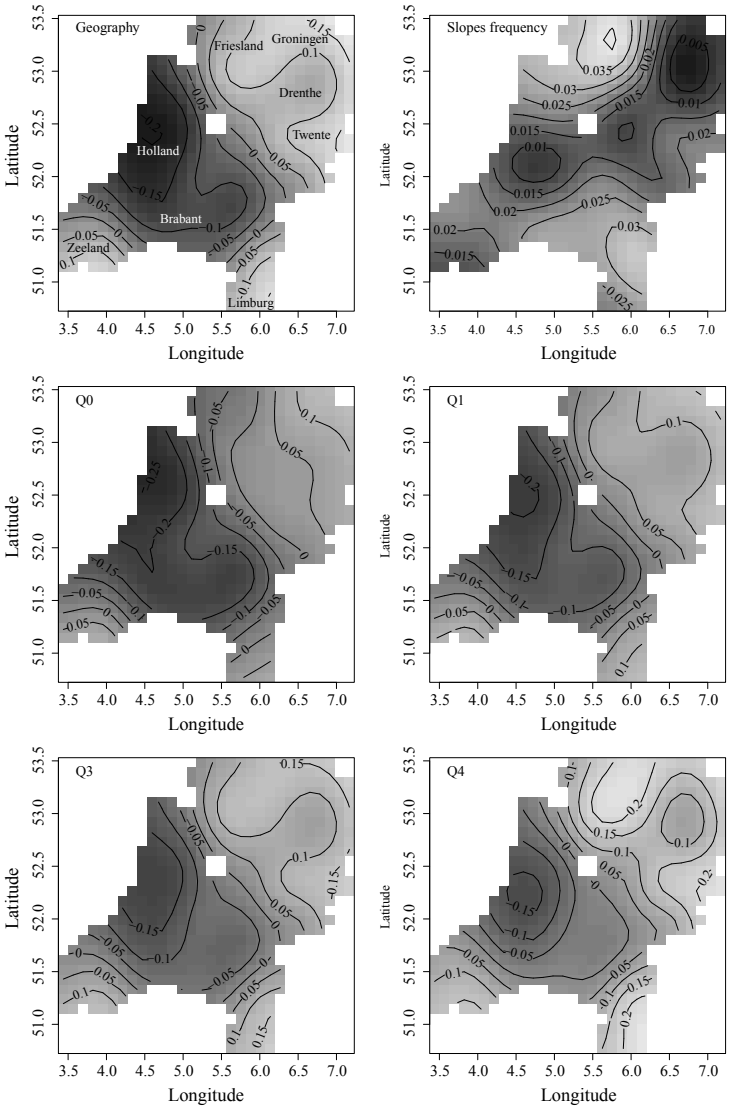


Figure 6.5. Geography, word frequency and distance from standard Dutch. Upper left: distance predicted only from longitude and latitude. Upper right: the geographical distribution of random slopes for word frequency. Lower four panels: the combined effect of geography and word frequency on pronunciation distance for the minimum frequency (Q_0), the first quartile (Q_1), the third quartile (Q_3), and the maximum frequency (Q_4). Darker shades of gray denote smaller values, lighter shades of gray indicate larger values.

become increasingly more pronounced. For instance, the Low Saxon dialect of Twente on the central east border with Germany, and the Frisian varieties in the north profile themselves more clearly as different from standard Dutch for the higher-frequency words (Q4) than for the lower-frequency words (Q0).

For the lowest-frequency words (panel Q0), the northeast separates itself from the Hollandic sphere of influence, with pronunciation distances slowly increasing towards the very northeast of the country. This area includes Friesland and the Low Saxon dialects. As word frequency increases, the distance from standard Dutch increases, and most clearly so in Friesland. For Friesland, this solid resistance to the Hollandic norm, especially for high frequency words, can be attributed to Frisian being recognized as a different official language.

Twente also stands out as highly resistant to the influence of the standard language. In the sixteenth and seventeenth centuries, this region was not under firm control of the Dutch Republic, and Roman Catholicism remained stronger here than in the regions towards its west and north. The resistance to protestantism in this region may have contributed to its resistance to the Hollandic speech norms (see also Van Reenen, 2006).

In Zeeland (in the southwest) and Limburg (in the southeast), we find Low Franconian dialects that show the same pattern across all frequency quartiles, again with increased distance from Holland predicting greater pronunciation distance. The province of Limburg has never been under firm control of Holland for long, and has a checkered history of being ruled by Spain, France, Prussia, and Austria before becoming part of the kingdom of the Netherlands. Outside of the Hollandic sphere of influence, it has remained closer to dialects found in Germany and Belgium. The province of Zeeland, in contrast, has retained many features of an earlier linguistic expansion from Flanders (in the Middle Ages, Flanders had strong political influence in Zeeland). Zeeland was not affected by an expansion from Brabant (which is found in the central south of the Netherlands as well as in Belgium), but that expansion strongly influenced the dialects of Holland. This Brabantic expansion, which took place in the late Middle Ages up to the seventeenth century, clarifies why, across all frequency quartiles, the Brabantic dialects are most similar to the Hollandic dialects.

Our regression model appears to conflict with the view of Kloeke (which was also adopted by Bloomfield) that high frequency words should be more likely to undergo change than low frequency words (Kloeke, 1927; Bloomfield, 1933). This position was already argued for by Schuchardt (1885), who discussed data suggesting that high frequency words are more profoundly affected by sound change than low frequency words. Bybee (2002) called attention to language-internal factors of change that are frequency-sensitive. She argued that changes affecting high frequency words first would be a consequence of the overlap and reduction of articulatory gestures that comes with fluency. In contrast, low frequency words would be more likely to undergo analogical lev-

eling or regularization.

Our method does not allow us to distinguish between processes of articulatory simplification and processes of leveling or regularization. Moreover, our method evaluates the joint effect of many different sound changes for the geographical landscape. Our results indicate that, in general, high frequency words are most different from the standard. However, high frequency words can differ from the standard for very different reasons. For instance, they may represent older forms that have resisted changes that affected the standard. Alternatively, they may have undergone region-specific articulatory simplification. Furthermore, since higher-frequency forms are better entrenched in memory (Hasher and Zacks, 1984; Baayen, 2007), they may be less susceptible to change. As a consequence, changes towards the standard in high frequency words may be more salient, and more likely to negatively affect a speaker's in-group status as a member of a dialect community. Whatever the precise causes underlying their resistance to accommodation to the standard may be, our data do show that the net outcome of the different forces involved in sound change is one in which it is the high frequency words that are most different from the standard language.

The third lexical factor that reached significance was the contrast between nouns as opposed to verbs and adjectives. Nouns have a greater distance from the standard language than verbs and adjectives. (Further analyses revealed that the effects of verbs and adjectives did not differ significantly.) This finding harmonizes well with the results of Pagel et al. (2007), where they also observed that nouns were most resistant to change, followed by verbs and adjectives.

Similar to word frequency, we also observe a non-uniform effect of the contrast between nouns as opposed to verbs and adjectives across locations, indicated by the presence of the by-location random slopes for the word category contrast in our model (see Table 6.2). The parameters for these random slopes (i.e. the standard deviation, and the parameter measuring the correlation with the random intercepts) jointly increase the log-likelihood of the model by 1064 units, compared to 5.6 log-likelihood units for the fixed-effect (population) slope of this contrast. These by-location random slopes are not uniformly distributed across the geographical area, as shown by the upper right panel of Figure 6.6. This panel clearly shows that the word category in the northeast of the Netherlands does not influence the distance from the standard language (i.e. the slope is zero), while in Friesland nouns have a much higher distance from the standard than verbs or adjectives.

To clarify how geography (GAM distance) and the word category contrast jointly predict distance from the standard language, we first calculated the fitted GAM distance for each location. We then estimated the predicted distance from the standard language using GAM distance, a fixed (median) word frequency, and the word category contrast as predictors, weighted by the coefficients estimated by our mixed-effects regression model. Because the fitted surfaces are different for nouns as opposed to verbs and adjectives, we visualized

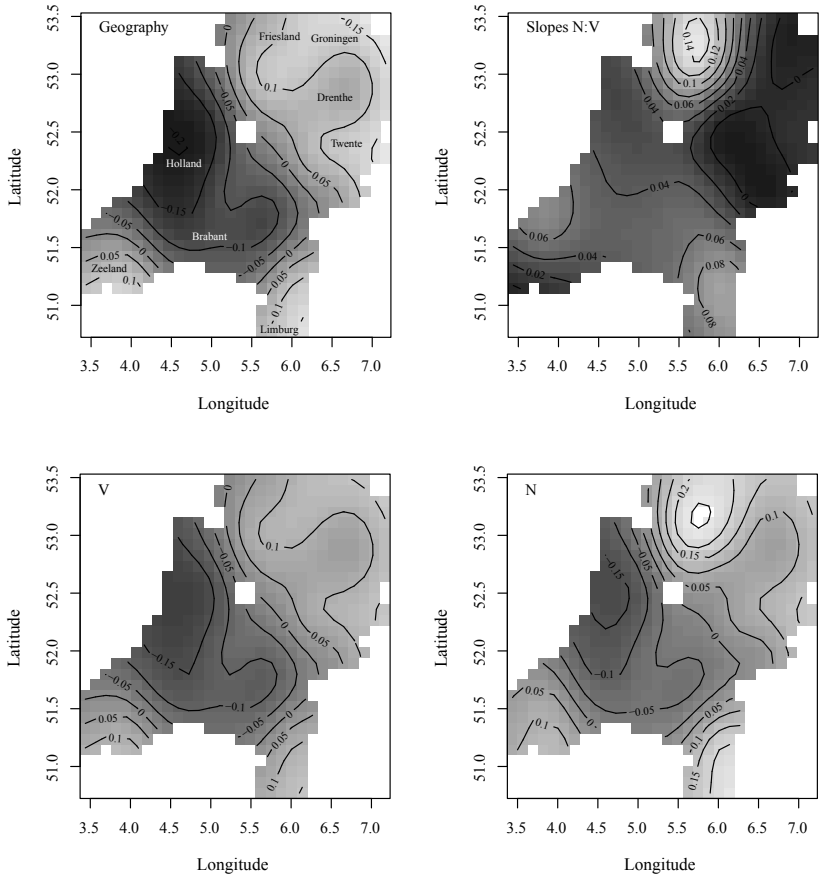


Figure 6.6. Geography, word category and distance from standard Dutch. Upper left: distance predicted only from longitude and latitude. Upper right: the geographical distribution of random slopes for the contrast between nouns as opposed to verbs and adjectives. Bottom panels: the combined effect of geography and word category on pronunciation distance for verbs/adjectives (panel V) and nouns (panel N). Darker shades of gray denote smaller values, lighter shades of gray indicate larger values.

are more pronounced for nouns than for verbs.

As the pattern of variation at the periphery of the Netherlands is quite similar to the pattern reported for high frequency words (i.e. the peripheral areas are quite distinct from the standard), we will not repeat its discussion here. The similarity between high frequency words and nouns (as opposed to verbs and adjectives) is also indicated by the correlation parameter of 0.550 in Table 6.2.

6.5 Discussion

In this chapter we have illustrated that several factors play a significant role in determining dialect distances from the standard Dutch language. Besides the importance of geography, we found clear support for three word-related variables (i.e. the contrast between nouns as opposed to verbs and adjectives, word frequency, and the vowel-to-consonant ratio in the standard Dutch pronunciation), as well as two variables relating to the social environment (i.e. the number of inhabitants in a location, and the average age of the inhabitants in a location). These results clearly indicate the need for variationists to consider explanatory quantitative models which incorporate geographical, social and word-related variables as independent variables.

We did not find support for the importance of speaker-related variables such as gender and age. As we only had a single pronunciation per location, we cannot exclude the possibility that these speaker-related variables do play an important role (especially considering the results of Chapters 7 and 8, where speaker age is found to be significant). It would be very informative to investigate dialect change in a dataset containing pronunciations of both young and old speakers in the same location, using the apparent time construct (Bailey et al., 1991). In addition, being able to compare male and female speakers in a single location would give us more insight into the effect of gender.

It is important to note that the contribution of the random-effects structure to the goodness of fit of the model tends to be one or two orders of magnitude larger than the contributions of the fixed-effect predictors, with GAM distance (geography) as sole exception. This indicates that the variation across locations and across words is huge compared to the magnitude of the effects of the socio-demographic and lexical predictors.

Our model also provides some insight into lexical diffusion. While we did not focus on individual sound changes, it is clear that the resistance to change at the word level is influenced by several word-related factors, as well as a number of socio-demographic factors of which the precise effect varies per word. Consequently, it is sensible to presume that a sound in one word will change more quickly than the same sound in another word (i.e. constituting a lexically gradual change). However, to make more precise statements about lexical diffusion as opposed to the lexically abrupt sound changes posited in the Neogrammar-

ian hypothesis (e.g., see Labov, 1981 for a discussion of both views), it is necessary to look at the level of the individual sound correspondences.

It would, therefore, be rewarding to develop a model to predict if an individual sound in a dialectal pronunciation is equal to or different from the corresponding sound in the standard Dutch pronunciation. As the Levenshtein distance is based on the alignments of sounds, these sound correspondences are already available (see Chapters 2 and 3). Using a logistic mixed-effects regression model would enable us to determine which factors predict the (dis)similarity of individual sounds with respect to the corresponding sounds in the standard Dutch pronunciations. Of course, this would also increase the computational effort, but since on average every word consists of about four to five sounds, this potential study should remain tractable (at least, when focusing on a subset of the data).

In the present study, we connected a larger distance from standard Dutch with a greater resistance to change (i.e. standardization). While this might be true, it is also possible that words do not only change in the direction of the standard language. Ideally this should be investigated using pronunciations of identical words at different moments in time. For example, by comparing our data to the overlapping older pronunciations in the *Reeks Nederlandse Dialectatlassen* (Blancquaert and Pée, 1925–1982).

Instead of using standard Dutch as our reference point, we could also use proto-Germanic, following the approach of Heeringa and Joseph (2007). It would be rewarding to see if smaller distances from the proto-language correspond to larger distances from the standard language. Alternatively, we might study the dialectal landscape from another perspective, by selecting a dialectal variety as our reference point. For example, dialect distances could be calculated with respect to a specific Frisian dialect.

In summary, our quantitative sociolinguistic analysis has found support for lexical diffusion in Dutch dialects and has clearly illustrated that convergence towards standard Dutch is most likely in low-frequent words. Furthermore, we have shown that mixed-effects regression modeling in combination with a generalized additive model representing geography is highly suitable for investigating dialect distances and its (sociolinguistic) determinants.

To demonstrate the wide applicability of this approach, which combines the merits of both dialectometry and dialectology (see Chapter 1), we will also apply an improved version of this method to a Catalan dialect dataset in Chapter 7 and a Tuscan dialect dataset in Chapter 8.

Chapter 7

Catalan language policies and standardization

Abstract. In this chapter we investigate which factors are involved in determining the pronunciation distance of Catalan dialectal pronunciations from standard Catalan. We use pronunciations of 320 speakers of varying age in 40 places located in three regions where the Catalan language is spoken (the autonomous communities Catalonia and Aragon in Spain, and the state of Andorra). In contrast to Aragon, Catalan has official status in both Catalonia and Andorra, which is likely to have a different effect on the standardization of dialects in these regions. We fitted a generalized additive mixed-effects regression model to the pronunciation distances of 357 words from standard Catalan. In the model, we found clear support for the importance of geography (with higher distances from standard Catalan in Aragon as opposed to Catalonia and Andorra) and several word-related factors. In addition, we observed a clear distinction between Aragon and the other two regions with respect to speaker age. Speakers in Catalonia and Andorra showed a clear standardization pattern, with younger speakers having dialectal pronunciations closer to the standard than older speakers, while this effect was not present in Aragon. These results clearly show the important effect of language policies on standardization patterns and border effects in dialects of a single language. In addition, this chapter provides further support for the usefulness of generalized additive modeling in analyzing dialect data.[1]

7.1 Introduction

In this chapter we investigate a Catalan dialect dataset using a generalized additive mixed-effects regression model in order to identify sociolinguistic and word-related factors which play an important role in predicting the distance between dialectal pronunciations and the Catalan standard language. Using this approach also allows us to investigate border effects caused by different policies with respect to the Catalan language quantitatively. We use Catalan dialect pronunciations of 320 speakers of varying age in 40 places located in three regions where the Catalan language is spoken (i.e. the autonomous communities Catalonia and Aragon in Spain, and the state of Andorra). As

[1] This chapter is based on Wieling, Valls, Baayen and Nerbonne (submitted-b).

the Catalan language has official status in Andorra (as the only language) and Catalonia (where both Catalan and Spanish are the official languages; Woolard and Gahng, 2008), but not in Aragon (Huguet et al., 2000), we will contrast these two regions in our analysis.

7.1.1 Border effects

Border effects in European dialectology have been studied intensively (see Woolhiser, 2005 for an overview). In most of these studies, border effects have been identified on the basis of a qualitative analysis of a sample of linguistic features. In contrast, Goebl (2000) used a dialectometric approach and calculated aggregate dialect distances based on a large number of features to show the presence of a clear border effect at the Italian-French and Swiss-Italian borders, but only a minimal effect at the French-Swiss border. This approach is arguably less subjective than the traditional approach in dialectology (see Chapter 1), as many features are taken into account simultaneously and the measurements are very explicit. However, Woolhiser (2005) is very critical of this study, as Goebl does not discuss the features he used and also does not consider the sociolinguistic dynamics, as well as the ongoing dialect change (i.e. he uses dialect atlas data).

Several researchers have offered hypotheses about the presence and evolution of border effects in Catalan. For example, Pradilla (2008a, 2008b) indicated that the border effect between Catalonia and Valencia might increase, as both regions recognize a different variety of Catalan as the standard language (i.e. the unitary Catalan standard in Catalonia and the Valencian Catalan substandard in Valencia). In a similar vein, Bibiloni (2002, p. 5) discussed the increase of the border effect between Catalan dialects spoken on either side of the Spanish-French border in the Pyrenees during the last three centuries. More recently, Valls et al. (accepted) conducted a dialectometric analysis of Catalan dialects and found on the basis of aggregate dialect distances (i.e. average distances based on hundreds of words) a clear border effect contrasting Aragon with Catalonia and Andorra. Their dialectometric approach is an improvement over Goebl's (2000) study, as Valls et al. (accepted) measure dialect change by including pronunciations for four different age groups (i.e. using the apparent-time construct; Bailey, 1991) and also investigate the effect of community size. Unfortunately, however, their study did not investigate other sociolinguistic variables.

7.1.2 Regression models to study pronunciation variation

In this chapter we use the same Catalan dialect dataset as studied by Valls et al. (accepted). We measure the pronunciation distances of a large number of words (following dialectometry; see Chapter 1), and we will use a generalized additive mixed-effects regression model to predict these distances per word for

all individual speakers (eight per location) on the basis of geography, word-related features and several sociolinguistic determinants.

In our regression analysis we will contrast the area where Catalan is an official language (Catalonia and Andorra) with the area where this is not the case (Aragon). Based on the results of Valls et al. (accepted), we expect to observe larger pronunciation distances from standard Catalan in Aragon than in the other two regions. Furthermore, we expect that the regions will differ with respect to the importance of the sociolinguistic factors. Mainly, we expect to see a clear effect of speaker age (i.e. with younger speakers having pronunciations closer to standard Catalan) in the area where Catalan has official status, while we do not expect this for Aragon, as there is no official Catalan language policy which might 'attract' the dialect pronunciations to the standard.

In contrast to the exploratory visualization-based analysis of Valls et al. (accepted), the regression analysis allows us to assess the significance of these differences. For example, while Valls et al. (accepted) state that urban communities have pronunciations more similar to standard Catalan than rural communities, this pattern might not be significant.

7.2 Material

7.2.1 Pronunciation data

The Catalan dialect dataset contains phonetic transcriptions (using the International Phonetic Alphabet) of 357 words in 40 dialectal varieties and the Catalan standard language. The same lexical form was always used for a single word (i.e. the dataset did not contain lexical variation). The locations are spread out over the state of Andorra (two locations) and two autonomous communities in Spain (Catalonia with thirty locations and Aragon with eight locations). In all locations, Catalan has traditionally been the dominant language. Figure 7.1 shows the geographical distribution of these locations. The locations were selected from twenty counties, and for each county the (urban) capital as well as a rural village was chosen as a data collection site. In every location eight speakers were interviewed, two per age group (F1: born between 1991 and 1996; F2: born between 1974 and 1982; F3: born between 1946 and 1960; F4: born between 1917 and 1930). All data was transcribed by a single transcriber, who also did the fieldwork for the youngest (F1) age-group between 2008 and 2011. The fieldwork for the other age groups was conducted by another fieldworker between 1995 and 1996. The complete dataset contains 357 words, consisting of 16 articles, 81 clitic pronouns, 8 demonstrative pronouns, 2 neuter pronouns, 2 locative adverbs, 220 verbs (inflected forms of five verbs), 20 possessive pronouns and 8 personal pronouns. The original dataset consisted of 363 words, but six words were excluded as they did not have a pronunciation in the standard Catalan language. A more detailed description of the dataset is given by Valls et al. (accepted).

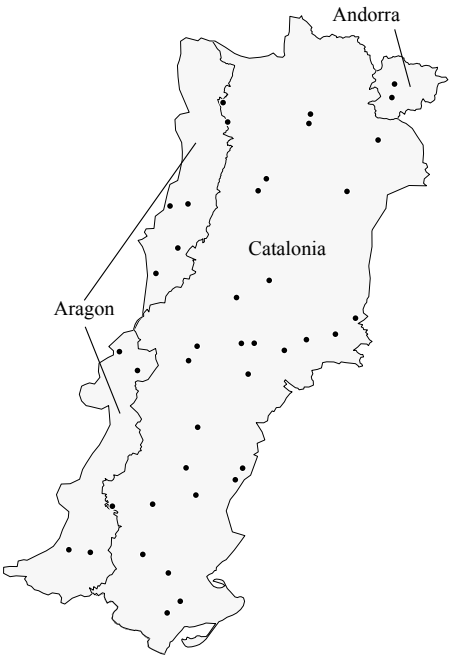


Figure 7.1. Geographical distribution of the Catalan varieties. Two locations are found in Andorra, eight in Aragon and the remaining thirty locations are found in Catalonia.

Given the significant effect of word frequency in the previous chapter, we also investigated if we could obtain Catalan word frequencies. While we were able to find a dictionary with frequency information (Rafel, 1996–1998), it did not contain contextual information which was necessary to assess the frequency of clitics and articles (representing almost one third of our data). Consequently, we did not include word frequency information.

7.2.2 Sociolinguistic data

Besides the information about the speakers present in the corpus (i.e. gender, age, and education level of the speaker), we extracted additional demographic information about each of the 40 locations from the governmental statistics institutes of Catalonia (Institut d'Estadística de Catalunya, 2008, 2010), Aragon (Instituto Aragonés de Estadística, 2007, 2009, 2010) and Andorra (Departament d'Estadística del Govern d'Andorra, 2010). The information we extracted for each location was the number of inhabitants (i.e. community size), the average community age, the average community income, and the relative number

of tourist beds (i.e. per inhabitant; used as a proxy to measure the influence of tourism) in the most recent year available (ranging between 2007 and 2010). There was no location-specific income information available for Andorra, so for these two locations we used the average income of the country (Cambra de Comerç - Indústria i Serveis d'Andorra, 2008).

As the data for the older speakers (age groups F2, F3 and F4) was collected in 1995, the large time span between the recordings and measurement of demographic variables might be problematic. We therefore obtained information on the community size, average community age, and average community income for most locations in 2000 (which was the oldest data available online). Based on the high correlations between the data from the year 2000 and the most recent data (in all cases: $r > 0.9$, $p < 0.001$), we decided to use the most recent demographic information. No less recent information about the number of tourist beds was available for Catalonia and Aragon, but we do not have reason to believe that this correlation strength should be lower than for the other variables.

7.3 Methods

7.3.1 Obtaining pronunciation distances

For every speaker, the pronunciation distance between standard Catalan and the dialectal word pronunciations was calculated by using the PMI-based Levenshtein algorithm (diagonal-exclusive version) as explained in Section 2.3.3. In line with the previous chapter, we incorporated some additional linguistic information in the initialization step of the PMI-based Levenshtein algorithm by allowing the alignment of semivowels (i.e. [j] and [w]) with both vowels and consonants. We normalized the PMI-based word pronunciation distances by dividing them by the alignment length.

7.3.2 Generalized additive mixed-effects regression modeling

Similar to Chapter 6, we use a thin plate regression spline combining longitude and latitude in a generalized additive model (GAM; see Section 6.3.2) to represent geography. However, in contrast to creating a separate linear mixed-effects regression model and including the fitted values of the GAM as a predictor, we include all other factors and covariates, as well as the random-effects structure (see Section 6.3.3 for an explanation) in the generalized additive model. This methodological advancement was possible due to improvements in the software package we used (i.e. the `mgcv` package in `R`; Wood, 2006).

In this study, we identify three random-effect factors, namely word, speaker and location. The significance of the associated random slopes and intercepts in the model was assessed by the Wald test. Besides the smooth combining longitude and latitude representing geography, we considered several other predic-

tors. Based on our initial analysis which showed that articles, clitic pronouns and demonstrative pronouns had a significantly larger distance from the corresponding standard Catalan pronunciations than the other word categories, we included a factor to distinguish these two groups. Other lexical variables we included were the length of the word (i.e. the number of sounds in the standard Catalan pronunciation), and the relative frequency of vowels in the standard Catalan pronunciation of each word. In addition, we included several location-specific variables: community size, the average community age, the average community income, and the relative number of tourist beds (as a proxy for the amount of tourism). The speaker-related variables we took into account were the year of birth, the gender, and the education level of the speaker. Finally, we used a factor to distinguish speakers from Catalonia and Andorra as opposed to Aragon.

As already remarked in Section 6.3.3, collinearity of predictors is a general problem in large-scale regression studies. In our dataset, communities with a larger population tend to have a higher average income and lower average age and also show a specific geographical distribution, somewhat similar to Figure 7.2 (e.g., the largest communities appear mainly in the east). To be able to assess the pure effect of each predictor, we took out the effect of other correlated variables, by instead using as predictor the residuals of a linear model regressing that predictor on the correlated variables (i.e. one way only, so we took out the effect of community size from average income, but not the other way around). In this context, geography was represented by the fitted values of a GAM predicting the pronunciation distance from standard Catalan only based on longitude and latitude. As the new predictors all correlated positively with the original predictors, they can still be interpreted in the same way as the original predictors.

Following Chapter 6, a few numerical predictors (i.e. community size and the relative number of tourist beds) were log-transformed in order to reduce the potentially harmful effect of outliers. To facilitate the interpretation of the fitted parameters of our model, we scaled all numerical predictors by subtracting the mean and dividing by the standard deviation. In addition, we log-transformed and centered our dependent variable (i.e. the pronunciation distance per word from standard Catalan, averaged by alignment length). Consequently, the value zero represents the mean distance, negative values a smaller distance, and positive values a larger distance from the standard Catalan pronunciation. The significance of each fixed-effect factor and covariate was evaluated by means of the Wald test (reporting an F-value).

7.4 Results

For the purpose of another study, a multiple alignment of the sound segments in every pronunciation was made. This multiple alignment did not reveal any

transcription errors and therefore signifies the high quality of the dataset. For this reason, we did not remove pronunciations with a large distance from standard Catalan, as these are genuine distances instead of noise. As not all words are pronounced by every speaker, the total number of cases (i.e. word-speaker combinations) is 112,608.

We fitted a generalized additive mixed-effects regression model, step by step removing predictors that did not contribute significantly to the model. We will discuss the specification of the model including all significant predictors (shown in Table 7.1) and verified random effects (shown in Table 7.2). The model explained 75% of the variation in pronunciation distances from standard Catalan. This indicates that the model is highly capable of predicting the individual distances (for specific speaker and word combinations), providing support for our approach of integrating geographical, social and lexical variables. The main contributor (63%) for this good fit was the variability associated with the words (i.e. the random intercepts for word). Without the random-effects structure, the fixed-effect factors and covariates explained 20% of the variation. To compare the relative influence of each of these (fixed-effect) predictors, we included a measure of effect size by specifying the increase or decrease of the dependent variable when the predictor increased from its minimum to its maximum value (following Chapter 6). The effect size of the geographical smooth was calculated by subtracting the minimum from the maximum fitted value (see Figure 7.2). Similar to the results of Chapter 6, we observe that geography and the word-related predictors have the greatest influence on the pronunciation distance from the standard language.

As our initial analysis (investigating the random intercepts for word, location and speaker) revealed that the inclusion of a random intercept for location was not warranted given its limited improvement in goodness of fit, location is not included as a random-effect factor.

7.4.1 Geography

Figure 7.2 shows the resulting regression surface (represented by the final line of Table 7.1) for the complete area under study using a contour plot. The thin plate regression spline was highly significant as the invested 23.9 estimated degrees of freedom were supported by an F-value of 24.9 ($p < 0.001$). The (solid) contour lines represent aggregate distance isoglosses connecting areas which have a similar distance from standard Catalan. Darker shades of gray indicate smaller distances, lighter shades of gray represent greater distances from the standard Catalan language. We can clearly identify the separation between the dialects spoken in the east of Catalonia compared to the Aragonese varieties in the west. The local cohesion in Figure 7.2 is sensible, as people living in nearby communities tend to speak relatively similar (Nerbonne and Kleiweg, 2007).

	Estimate	Std. err.	p-value	Effect size
Intercept	-0.1018	0.0209	< 0.001	
Word length	0.1302	0.0218	< 0.001	0.4411
Vowel ratio per word	0.1050	0.0137	< 0.001	0.6491
Word category is A/D/C	0.3051	0.0478	< 0.001	0.3052
Community size (log)	-0.0074	0.0038	0.051	-0.0282
Speaker year of birth	-0.0114	0.0031	< 0.001	-0.0342
Location in Aragon	0.0470	0.0372	0.206	0.0470
Loc. in Aragon × Sp. yr. of birth	0.0168	0.0063	0.008	0.0490
s(longitude,latitude) [23.9 edf]			< 0.001	0.2716

Table 7.1. Fixed-effect factors and covariates of the final model. Effect size indicates the increase or decrease of the dependent variable when the predictor value increases from its minimum to its maximum value (i.e. the complete range). Community size was included as it neared significance. The factor distinguishing locations in Aragon from those in Catalonia and Andorra was included as the interaction with year of birth was significant. The geographical smooth (see Figure 7.2; 23.9 estimated degrees of freedom) is represented by the final row. Its effect size equals the minimum value subtracted from the maximum value of the fitted smooth.

Factor	Random effect	Std. dev.	p-value
Word	Intercept	0.2547	< 0.001
	Relative nr. of tourist beds	0.0216	< 0.001
	Average community age	0.0138	< 0.001
	Community size (log)	0.0150	< 0.001
	Average community income	0.0140	< 0.001
	Speaker year of birth	0.0259	< 0.001
	Speaker education level	0.0137	< 0.001
	Location in Aragon	0.1559	< 0.001
	Location in Aragon × Speaker yr. of birth	0.0245	< 0.001
Speaker	Intercept	0.0369	< 0.001
	Word length	0.0291	< 0.001
	Vowel ratio per word	0.0168	< 0.001
	Word category is A/D/C	0.0590	< 0.001
Residual		0.1725	< 0.001

Table 7.2. Significant random-effect parameters of the final model. The standard deviation indicates the amount of variation for every random intercept and slope.

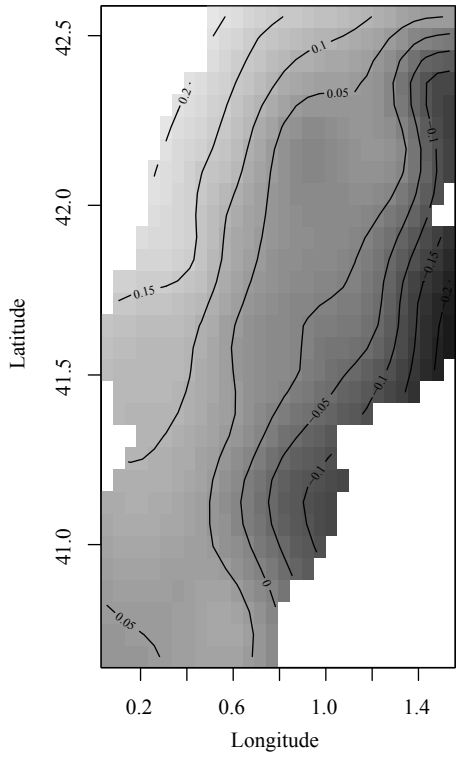


Figure 7.2. Contour plot for the regression surface of pronunciation distance (from standard Catalan) as a function of longitude and latitude obtained with a generalized additive model using a thin plate regression spline. The (black) contour lines represent aggregate distance isoglosses, darker shades of gray (lower values, negative in the east) indicate smaller distances from the standard Catalan language, while lighter shades of gray (higher values) represent greater distances.

7.4.2 Demographic predictors

Of all location-based predictors (i.e. the relative number of tourist beds, community size, average community income, and average community age), only community size was close to significance ($p = 0.051$) as a main effect in our general model (see Table 7.1). All location-based predictors, however, showed significant word-related variation. For example, while there is no main effect of average community income, the pronunciation of individual words might be influenced (positively or negatively) by community income.

It might seem strange that the factor distinguishing the locations in Aragon

from those in Catalonia and Andorra was not significant, but the smooth function representing geography (see Figure 7.2) already shows that the Aragonese varieties have a higher distance from standard Catalan than the other varieties. Note that the contour lines in Figure 7.2 all run roughly north-south, and the distances increase monotonically as one looks further west. In fact, when we exclude the smooth function the factor is highly significant ($p < 0.001$) and assigns higher distances from standard Catalan to the Aragonese varieties.

With respect to the speaker-related predictors, only year of birth was a significant predictor indicating that younger speakers use pronunciations which are more similar to standard Catalan than older speakers. However, the significant interaction (i.e. Location in Aragon × Speaker year of birth) in Table 7.1 indicates that this pattern does not hold for speakers from Aragon. In line with our hypothesis, there is no effect of speaker age for the Aragonese speakers. This result suggests the existence of a border effect between Aragon on the one hand, and Catalonia and Andorra on the other.

We did not find an effect of gender (in both the fixed-effects and the random-effects structure), despite this being reported in the literature frequently (see Cheshire, 2002 for an overview). However, in the previous chapter, we also did not find a gender effect with respect to the pronunciation distance from the standard language. It might be that this phenomenon is more strongly linked to individual sounds (e.g., see Chambers and Trudgill, 1998, Section 5.3) than to pronunciation distances between complete words.

We also did not find support for the inclusion of education level as a covariate in our model. The reason for this might be similar to the reason for the absence of a gender effect, but the education measure alone (without any other social class measures) might simply have too little power to discover social class effects (Labov, 2001, Ch. 5; but see Gorman, 2010 for a new analysis of Labov's data suggesting that education does have sufficient power).

Interestingly, we do see that the effect of (some of) these speaker-related variables varies significantly *per word* (see Table 7.2).

7.4.3 Lexical predictors

All lexical variables we tested were significant predictors of the pronunciation distance from standard Catalan and also showed significant by-speaker random slopes.

It is not surprising that the factor distinguishing articles, clitic pronouns and demonstratives from the other words was highly significant, since we grouped these word categories on the basis of their higher distance from the standard language (according to our initial analysis). Articles and clitic pronouns are relatively short (in many cases only having a length of one or two sounds), and when they are different from the standard, the relative distance will be very high. While the demonstratives are not so short, they tend to be ei-

ther completely identical to the standard pronunciation, or almost completely different, explaining their larger distance.

We were somewhat surprised that the number of sounds in the reference pronunciation contributed significantly to the distance from the standard, as we normalized dialect distances by dividing them by the alignment length (which correlates highly, $r > 0.95$, with the number of sounds in the reference pronunciation). This result indicates that longer words have a higher average distance from the standard pronunciation than shorter words. While we do not have a clear explanation for this, including this predictor in our model allows us to more reliably assess the effect of the more interesting (sociolinguistic) predictors.

Finally, the proportion of vowels in the reference pronunciation was a highly significant predictor (having the largest effect size). This is not surprising (and similar to the result reported in Chapter 6) as vowels are much more variable than consonants (e.g., Keating et al., 1994).

Besides playing a significant role as fixed-effect factors, all lexical predictors showed significant variation across speakers in their effect on the pronunciation distance from standard Catalan. For example, while some speakers might pronounce longer words more different from standard Catalan than shorter words, other speakers might show the opposite pattern.

7.5 Discussion

In this chapter we have used a generalized additive mixed-effects regression model to provide support for the existence of a border effect between Aragon (where the Catalan language has no official status) and Catalonia and Andorra (where Catalan is an official language). Our analysis clearly revealed a greater distance from standard Catalan for speakers in Aragon, as opposed to those in Catalonia and Andorra. Furthermore, our analysis identified a significant effect of speaker age (with younger speakers having pronunciations closer to standard Catalan) for Catalonia and Andorra, but not for Aragon. This provides strong evidence for the existence of a border effect in these regions caused by different language policies, and is in line with the results of Valls et al. (accepted). Also, our analysis revealed the importance of several word-related factors in predicting the pronunciation distance from standard Catalan, and confirms the utility of using generalized additive mixed-effects regression modeling to analyze dialect distances, with respect to traditional dialectometric analyses.

In contrast to the conclusion of Valls et al. (accepted) that the older speakers in urban communities use pronunciations closer to standard Catalan than the older speakers in rural communities, we did not find a clear significant effect of community size (nor a significant interaction between speaker age and community size). In fact, when using the binary distinction they based their conclusion on (i.e. distinguishing urban and rural communities in twenty dif-

ferent counties), the results do not even approach significance ($p = 0.18$). This clearly illustrates the need for adequate statistical tests to prevent reaching statistically unsupported conclusions.

We did not find support for the general influence of the other demographic variables. This contrasts with Chapter 6 where we found a significant effect of community size (larger communities use pronunciations closer to the standard) and average community age (older communities use pronunciations closer to the standard language). However, the number of locations studied here was only small and might have limited our power to detect these effects (i.e. in the previous chapter more than ten times as many locations were included).

We see two promising extensions of this study. First, it would be interesting to compare the dialectal pronunciations to the Spanish standard language instead of the Catalan standard language. In our dataset there are clear examples of the usage of a dialectal form closer to the standard Spanish pronunciation than to the standard Catalan pronunciation, and it would be rewarding to investigate which word- and speaker-related factors are related to this.

As already suggested in Chapter 6, another extension involves focusing on the individual sound correspondences between Catalan dialect pronunciations and pronunciations in standard Catalan (or in another language, such as Vulgar Latin for a more historically motivated reference point). These sound correspondences can easily be extracted from the alignments generated by the Levenshtein distance algorithm (see Section 2.3.3). When focusing on a specific set of locations (e.g., the Aragonese varieties), it would be computationally feasible to create a generalized additive mixed-effects regression model to investigate which factors determine when a sound in a certain dialectal pronunciation is different from the corresponding sound in the standard Catalan (or Vulgar Latin) pronunciation. Of course, this approach is also possible for all locations, but due to the larger size of the dataset, much more patience will be required before all parameters are successfully estimated.

In conclusion, this chapter has improved the methodology introduced in Chapter 6 by constructing a single generalized additive mixed-effects regression model instead of combining two separate models. We also illustrated the wide applicability of this approach by investigating the influence of various social and linguistic factors on language variation, thereby providing support for the existence of a border effect (caused by different language policies) in the Catalan-speaking area investigated in this study. In the next chapter, we will investigate the applicability of the method with respect to lexical variation.

Chapter 8

Determinants of Tuscan lexical variation

Abstract. In this chapter we use a generalized additive mixed-effects regression model to predict lexical differences for 170 concepts in 213 Tuscan dialects with respect to standard Italian. In our model, geographical position was found to be an important predictor, with locations more distant from Florence having lexical forms more likely to differ from standard Italian. In addition, the geographical pattern varied significantly for low versus high frequency concepts and old versus young speakers. Several other factors emerged as significant. The model predicts that lexical variants used by older speakers and in smaller communities are more likely to differ from standard Italian. The impact of community size, however, varied from concept to concept. For a majority of concepts, lexical variants used in smaller communities are more likely to differ from the standard Italian form. For a smaller minority of concepts, however, lexical variants used in larger communities are more likely to differ from standard Italian. Similarly, the effect of average community income and average community age varied per concept. These results clearly identify important factors involved in dialect variation at the lexical level in Tuscany. In addition, this chapter illustrates the potential of generalized additive mixed-effects regression modeling applied to lexical dialect data.[1]

8.1 Introduction

In this chapter we investigate a Tuscan lexical dialect dataset using a generalized additive mixed-effects regression model (as introduced in Chapter 7) in order to identify sociolinguistic and concept-related factors which play an important role in predicting lexical differences with respect to standard Italian.

8.1.1 Relationship between Tuscan and standard Italian

Standard Italian is unique among modern European standard languages. Although Italian originated in the fourteenth century, it was not consolidated as a spoken national language until the twentieth century. For centuries, Italian

[1] This chapter is based on Wieling, Montemagni, Nerbonne and Baayen (submitted-a).

was a written literary language, acquired through literacy when one learned to read and write, and was therefore only known to a minority of (literate) people. During this period, people spoke only their local dialect. A good account of the rise of standard Italian is provided by Migliorini and Griffith (1984). The particular nature of Italian as a literary language, rather than a spoken language was recognized since its origin and widely debated from different (i.e. socio-economic, political and cultural) perspectives as the *questione della lingua* or 'language question'.

At the time of the Italian political unification in 1860 only a very small percentage of the population was able to speak Italian, with estimates ranging from 2.5% (De Mauro, 1963) to 10% (Castellani, 1982). Only during the second half of the twentieth century real native speakers of Italian started to appear, as Italian started to be used by Italians as a spoken language in everyday life. Mass media (newspapers, radio and TV) and education played a central role in the diffusion of the Italian language throughout the country. According to the most recent statistics of ISTAT (*Istituto Nazionale di Statistica*) reported by Lepschy (2002), 98% of the Italian population is able to use their national language. However, dialects and standard Italian appear to coexist. For example, ISTAT data show that at the end of the twentieth century (1996) 50% of the population used (mainly or exclusively) Italian to communicate with friends and colleagues, while this percentage decreased to 34% when communication with relatives was taken into account.

To see the reason for the coexistence of standard Italian and local dialects, the origin of the standard language has to be taken into account. Italian has its roots in one of the speech varieties that emerged from spoken Vulgar Latin (Maiden and Parry, 1997), namely that of Tuscany, and more precisely the variety of Tuscan spoken in Florence. The importance of the Florentine variety in Italy was mainly determined by the prestige of the Florentine culture, and in particular the establishment of Dante, Petrarch and Boccaccio, who wrote in Florentine, as the 'three crowns' (*tre corone*) of the Italian literature. Consequently, Italian dialects do not represent varieties of the Italian language, but they are simply 'sisters' of the Italian language (Maiden, 1995).

In contrast to other Italian regions where a sort of 'sisterhood' relationship holds between the standard language and local dialects, in Tuscany this relationship is complicated by the fact that standard Italian originated from the Florentine dialect centuries ago. This also causes the frequent overlap between dialectal and standard Italian forms in Tuscany, which occurs much less frequently in other Italian regions (Giacomelli, 1978). However, since the Florentine dialect has developed (for several centuries) along its own lines and independently of the (literary) standard Italian language, its vocabulary does not always coincide with standard Italian. Following Giacomelli (1975), the types of mismatch between standard Italian and the dialectal forms can be partitioned into three groups. The first group consists of Tuscan words which are used in literature throughout Italy, but are not part of the standard language (i.e. these

terms usually appear in Italian dictionaries marked as 'Tuscanisms'). The second group consists of Tuscan words which *were* part of old Italian and are also attested in the literature throughout Italy, but have fallen into disuse as they are considered old-fashioned (i.e. these terms may appear in Italian dictionaries marked as 'archaisms'). The final group consists of Tuscan dialectal words which have no literary tradition and are not understood outside of Tuscany.

This chapter investigates the relationship between standard Italian and the Tuscan dialects from which it originated on the basis of the data collected through fieldwork for a regional linguistic atlas, the *Atlante Lessicale Toscano* ('Lexical Atlas of Tuscany', henceforth, ALT; Giacomelli et al., 2000). The ALT is a specially designed lexical atlas in which the dialectal data both have a diatopic (geographic) and diastratic (social) characterization. In particular, the advanced regression techniques we apply make it possible to keep track of the sociolinguistic and lexical factors at play in the complex relationship linking the Tuscan dialects with standard Italian.

The ALT data appear to be particularly suitable to explore the *questione della lingua* from the Tuscan point of view. As the compilation of the ALT questionnaire was aimed at capturing the specificity of Tuscan dialects and their relationships, concepts whose lexicalizations were identical to Italian (almost) everywhere in Tuscany were programmatically excluded (Giacomelli, 1978; Poggi Salani, 1978). This means that the ALT dataset was collected with the main purpose of better understanding the complex relationship linking the standard language and local dialects in the case the two did not coincide.

Previous studies have already explored the ALT dataset by investigating the relationship between Tuscan and Italian from the lexical point of view. Giacomelli and Poggi Salani (1984) based their analysis on the dialect data available at that time. Montemagni (2008), more recently, applied dialectometric techniques to the whole ALT dialectal corpus to investigate the relationship between Tuscan and Italian. In both cases it turned out that the Tuscan dialects overlap most closely with standard Italian in the area around Florence, expanding in different directions and in particular towards the southwest. Obviously, this observed synchronic pattern of lexical variation has the well-known diachronic explanation that the standard Italian language originated from the Florentine variety of Tuscan.

Montemagni (2008) also found that the observed patterns varied depending on the speaker's age: only 37 percent of the dialectal answers of the old speakers overlapped with standard Italian, while this percentage increased to 44 for the young speakers. In addition, words having a larger geographical coverage (i.e. not specific to a small region), were more likely to coincide with the standard language than words attested in smaller areas. These first, basic results illustrate the potential of the ALT dataset (which will be discussed in more detail the following section) to shed light on the widely debated *questione della lingua* from the point of view of Tuscan dialects.

8.2 Material

8.2.1 Lexical data

The lexical data used in this chapter was taken from the *Atlante Lessicale Toscano* (ALT; see also Section 3.2.1). ALT interviews were carried out between 1974 and 1986 in 224 localities of Tuscany, with 2193 informants selected with respect to a number of parameters ranging from age and socio-economic status to education and culture. It is interesting to note that only the younger ALT informants were born in the period when standard Italian was used as a spoken language. The interviews were conducted by a group of trained fieldworkers who employed a questionnaire of 745 target items, designed to elicit variation mainly in vocabulary and semantics.

In this chapter we focus on Tuscan dialects only, spoken in 213 out of the 224 investigated locations (see Figure 8.1; Gallo-Italian dialects spoken in Lunigiana and in small areas of the Apennines were excluded). We used the normalized lexical answers to a subset of the ALT onomasiological questions (i.e. those looking for the attested lexicalizations of a given concept). Out of 460 onomasiological questions, we selected only those which prompted 50 or fewer normalized lexical answers (the maximum in all onomasiological questions was 421 unique lexical answers). We used this threshold to exclude questions having many hapaxes as answers which did not appear to be lexical (a similar approach was taken by Montemagni, 2007). For instance, the question looking for denominations of 'stupid' included 372 different normalized answers, 122 of which are hapaxes. These either represent productive figurative usages (e.g., metaphors such as *cetriolo* 'cucumber' and *carciofo* 'artichoke'), productive derivational processes (e.g., *scemaccio* and *scemalone* from the lexical root *scemo* 'stupid'), or multi-word expressions (e.g., *mezzo scemo* 'half stupid', *puro locco* 'pure stupid' and similar). From the resulting 195-item subset, we excluded a single adjective and twelve verbs (as the remaining concepts were nouns) and all twelve multi-word concepts. Our final subset, therefore, consisted of 170 concepts (the complete list can be found in Wieling et al., submitted-a).

The normalized lexical forms in the ALT dataset still contained some morphological variation. In order to assess the pure lexical variation we abstracted away from variation originating in, e.g., assimilation, dissimilation, or other phonological differences (e.g., the dialectal variants *camomilla* and *capomilla*, meaning 'chamomile', have been treated as instantiations of the same normalized form), as well as from both inflectional and derivational morphological variation (e.g., inflectional variants such as singular and plural are grouped together). We compare these more abstract forms to the Italian standard.[2]

[2] The effect of the morphological variation was relatively limited, as the results using the unaltered ALT normalized lexical forms were highly similar to the results based on the lexical forms where morphological variation was filtered out.

Figure 8.1. Geographical distribution of the 213 Tuscan locations investigated in this chapter. The *F*, *P* and *S* mark the approximate locations of Florence, Pisa and Sienna, respectively.

The list of standard Italian forms for the 170 concepts was extracted from the online ALT corpus (ALT-Web; available at http://serverdbt.ilc.cnr.it/altweb), where it had been created for query purposes. This list, originally compiled on the basis of lexicographic evidence, was carefully reviewed by members of the *Accademia della Crusca*, the leading institution in the field of research on the Italian language in both Italy and the world, in order to make sure that it contained real Italian, and not old-fashioned or literary words originating in Tuscan dialects.

In every location multiple speakers were interviewed (between 4 and 29) and therefore each normalized answer is anchored to a given location, but also to a specific speaker. While we could have included all speakers separately (a total of 2081), we decided against this, as this would be computationally infeasible (logistic regression is computationally quite slow). Consequently, we grouped the speakers in an older age group (born in 1930 or earlier; 1930 was the median year of birth) and a younger age group (born after 1930). For both age groups, we used the lexical form pronounced by the majority of the speakers in the respective group. As not all concepts were attested in every location, the total number of cases (i.e. concept-speaker group combinations) was 69,259. Given the significant effect of word frequency on Dutch dialect distances reported in Chapter 6, we obtained the concept frequencies (of the standard Italian lexical form) by extracting the corresponding frequencies from a large corpus of 8.4

million Italian unigrams (Brants and Franz, 2009). While the frequencies of other lexical forms are likely somewhat different, these frequencies should give a good idea about the relative frequencies of different concepts.

As concrete concepts may be easier to remember (Wattenmaker and Shoben, 1987) and stored differently in the brain than abstract concepts (Crutch and Warrington, 2005), we also investigated if the concreteness of a concept played a role in lexical differences with respect to standard Italian. We obtained concreteness scores (for the English translations of the Italian concepts) from the MRC Psycholinguistic Database (Coltheart, 1981). The concreteness scores in this database ranged from 100 (abstract) to 700 (concrete). Our most abstract concept ('cheat') had a score of 329, while our most concrete concepts (e.g., 'cucumber' and 'grasshopper') had a score of about 660.

8.2.2 Sociolinguistic data

Besides the classification of the speaker group (old or young) and the year of recording for every location, we extracted additional demographic information about each of the 213 locations from a website with statistical information about Italian municipalities (Comuni Italiani, 2011). We extracted the number of inhabitants (in 1971 or 1981, whichever year was closer to the year when the interviews for that location were conducted), the average income (in 2005; which was the oldest information available), and the average age (in 2007; again the oldest information available) in every location. While the information about the average income and average age was relatively recent and may not precisely reflect the situation at the time when the dataset was constructed (between 1974 and 1986), the global pattern will probably be similar (i.e. in line with Section 7.2.2).

8.3 Methods

The method we employ is relatively similar to that of the previous chapter (explained in Section 7.3.2). Instead of predicting pronunciation distances, however, we will predict lexical differences (for 170 concepts in 213 Tuscan varieties, for both old and young speakers) with respect to standard Italian. In addition, there are some other methodological changes which are discussed in the following paragraphs.

8.3.1 Frequency-dependent geographical modeling

In this chapter we will take a more sophisticated approach to modeling geography. Given that the effect of word frequency varied geographically in Chapter 6, we allow the effect of geography to vary depending on concept frequency. Since the generalized additive model can combine an arbitrary number of predictors to represent a smoothing (hyper)surface, we created a three-dimensional

smooth (longitude × latitude × concept frequency), allowing us to assess the concept frequency-specific geographical pattern of lexical variation with respect to standard Italian. For example, similar to the Dutch results (shown in Figure 6.5), the geographical pattern may differ for low as opposed to high frequency concepts. Furthermore, we will investigate whether these patterns also vary for old speakers as opposed to young speakers (i.e. we create two three-dimensional smooths, one for old speakers and one for young speakers). We represent these three-dimensional smooths by a tensor product which allows combinations of non-isotropic predictors (i.e. measurements of the predictors are not on the same scale: e.g., longitudinal degrees versus frequency; Wood, 2006, p. 162). In the tensor product, we model both longitude and latitude with a thin plate regression spline as this is suitable for combining isotropic predictors and also in line with the approach of Chapters 6 and 7, while the concept frequency effect is modeled by a cubic regression spline, which is computationally more efficient than the thin plate regression spline. More information about these tensor product bases (which are implemented in the `mgcv` package for `R`) is provided by Wood (2006, Ch. 4).

8.3.2 Logistic regression modeling

In contrast to the previous two chapters, our dependent variable is binary (0: the lexical form is identical to standard Italian; 1: the lexical form is different from standard Italian) and this requires logistic regression.[3] Logistic regression does not model the dependent variable directly, but it attempts to model the probability (in terms of logits) associated with the values of the dependent variable. A logit is the natural logarithm of the odds of observing a certain value (in our case, a lexical form different from standard Italian). When interpreting the parameter estimates of our regression model, we should realize these need to be interpreted with respect to the logit scale. More detailed information about logistic regression is provided by Agresti (2007).

In our analysis, we include two random-effect factors, namely location and concept. In line with the previous chapter, the significance of the associated random slopes and intercepts in the model was assessed using the Wald test. In addition to the (concept frequency and speaker age group-specific) geographical variation we considered several other predictors. The only lexical variables we included were concept frequency (based on the frequency of the standard Italian lexical form) and the concreteness rating of each concept. The demographic variables we investigated were community size, average community age, average community income, and the year of recording. The only speaker-related variable we took into account was the age group (old: born in 1930 or earlier; young: born after 1930).

[3] In sociolinguistics, logistic regression is widely used and known as the VARBRUL analysis (Paolillo, 2002).

Similar to Chapters 6 and 7, our dataset contains some predictor collinearity. In our dataset, communities with a higher average age tend to have a lower average income. To be able to assess the pure effect of each predictor, we decorrelated average age from average income by using as predictor the residuals of a linear model regressing average age on average income. Since the new predictor correlated highly (r = 0.9) with the original average age values, we can still interpret the new predictor as representative of average age (but now excluding the effect of average income).

Following the previous two chapters, several numerical predictors were log-transformed (i.e. community size, average age, average income, and concept frequency) in order to reduce the potentially harmful effect of outliers. We scaled all numerical predictors by subtracting the mean and dividing by the standard deviation in order to facilitate the interpretation of the fitted parameters of the statistical model. The significance of fixed-effect factors and covariates was evaluated by means of the Wald test (reporting a z-value) for the coefficients in a logistic regression model.

8.4 Results

We fitted a generalized additive mixed-effects logistic regression model, step by step removing predictors that did not contribute significantly to the model. In the following we will discuss the specification of the final model including all significant predictors and verified random-effect factors.

Our dependent value was binary with a value of one indicating that the lexical form was different from the standard Italian form and a value of zero indicating that the lexical form was identical to standard Italian. Intuitively it is therefore easiest to view these values as a distance measure from standard Italian. The coefficients and the associated statistics of the significant fixed-effect factors and linear covariates are presented in Table 8.1. To allow a fair comparison of the effect of both predictors, we included a measure of effect size by specifying the increase or decrease of the likelihood of having a non-standard Italian lexical form (in terms of logits) when the predictor increased from its minimum to its maximum value (in line with Chapters 6 and 7).

Table 8.2 presents the significance of the three-dimensional smooth terms (modeling the concept frequency-dependent geographical pattern for both the old and young speaker group)[4] and Table 8.3 lists the significant random-effects structure of our model.

To evaluate the goodness of fit of the final model (see Tables 8.1 to 8.3), we used the index of concordance C. This index is also known as the receiver operating characteristic curve area 'C' (see, e.g., Harrell, 2001). Values of C exceed-

[4] We verified the necessity of including concept frequency and the contrast between old and young speakers in the geographical smooth (all p's < 0.001).

	Estimate	Std. err.	z-value	p-value	Eff. size
Intercept	-0.4129	0.1395	-2.960	0.003	
Old instead of yng. speakers	0.4407	0.0193	22.896	< 0.001	0.4407
Community size (log)	-0.1154	0.0266	-4.343	< 0.001	-0.7152

Table 8.1. Significant parametric terms of the final model. A positive estimate indicates that a higher value for this predictor increases the likelihood of having a non-standard Italian lexical form, while a negative estimate indicates the opposite effect. Effect size indicates the increase or decrease of the likelihood of having a non-standard Italian lexical form when the predictor value increases from its minimum to its maximum value (i.e. the complete range).

	Est. d.o.f.	Chi. sq.	p-value
Geography × concept frequency (old)	53.88	221.5	< 0.001
Geography × concept frequency (young)	59.48	326.6	< 0.001

Table 8.2. Significant smooth terms of the final model. For every smooth the estimated degrees of freedom is indicated, as well as its significance in the model. See Figure 8.2 for a visualization of these smooths.

Factor	Random effect	Std. dev.	p-value
Location	Intercept	0.2410	< 0.001
Concept	Intercept	1.7748	< 0.001
	Average community income (log)	0.3127	< 0.001
	Average community age (log)	0.2482	< 0.001
	Community size (log)	0.1166	0.006

Table 8.3. Significant random-effect parameters of the final model. The standard deviation indicates the amount of variation for every random intercept and slope.

ing 0.8 are generally regarded as indicative of a successful classifier. According to this measure, the model performed well with $C = 0.85$.

8.4.1 Geographical variation and lexical predictors

Inspecting Table 8.2, it is clear that the geographical pattern is a very strong predictor, and it varied significantly with concept frequency (which was not significant by itself in our general model) and speaker age group. Figure 8.2 visualizes the geographical variation related to concept frequency (for low fre-

quency: two standard deviations below the mean, mean frequency, and high frequency: two standard deviations above the mean) and speaker age group. Lighter shades of gray indicate a greater likelihood of having a lexical form different from standard Italian.

The three graphs to the left present the geographical patterns for old speakers, while those to the right present the geographical patterns for young speakers. In general, the graphs for the younger speakers are somewhat darker than those for the older speakers, supporting the finding (discussed in Section 8.4.2, below) that older speakers have a greater likelihood of using a lexical form different from standard Italian than younger speakers.

The first thing to note is that in the top graphs Florence (indicated by the star) is located in (approximately) the area with the smallest likelihood of having a non-standard Italian lexical form. This clearly makes sense as standard Italian originates from the Florentine variety.

The second observation is that, going from the top to the bottom graphs, we see a strong effect of concept frequency, both for older speakers and (to a slightly reduced extent) for younger speakers.[5] More frequent concepts in the central Tuscan area, including Florence, are more likely to differ from standard Italian than lower frequent concepts (i.e. the values in the bottom maps in the central area are higher than the values in the central and top maps).

Third, we observe a reverse pattern in the more peripheral areas (in the Tuscan archipelago in the west, but also in the north and east), with a greater likelihood of having a non-standard Italian lexical form for low frequency concepts than for high frequency concepts.

When looking in more detail at the data, high frequency concepts typically include cases for which standard Italian and Tuscan dialects diverge (e.g., standard Italian *angolo* 'angle', in Tuscany *canto*, *cantonata* or *cantone*; or standard Italian *pomeriggio* 'afternoon', in Tuscany *sera* 'evening' or multi-word expressions such as *dopo mangiato/pranzo/desinare* meaning 'after lunch', but also *dopo mezzogiorno* 'after noon'). For mean frequency concepts, the standard Italian and dialectal words share the same etymology, with the latter frequently (but not always) representing analogical variants of the former (e.g., for 'ivy', the standard Italian form is *edera*, whereas the set of dialectal forms includes *ellera*, *ellora*, *lellera*, *lallera*, etc.).[6] Finally, the low frequency concepts belong to an obsolete, progressively disappearing rural world and include (for example) *bigoncia* 'vat', *seccatoio* 'squeegee', and *stollo* 'haystack pole'.

To understand the pattern of results, we need to distinguish between three dimensions of change. First, as one moves out from the heartland of Tuscany, it is more likely that different words are used for a certain concept. This is the

[5] While the general effect of concept frequency appeared to be stronger for old speakers as opposed to young speakers, this interaction was not significant.

[6] Note that the normalization process we have employed in this chapter still distinguishes these forms, which represent lexical variants in their own right, in spite of their origin from the same etymon.

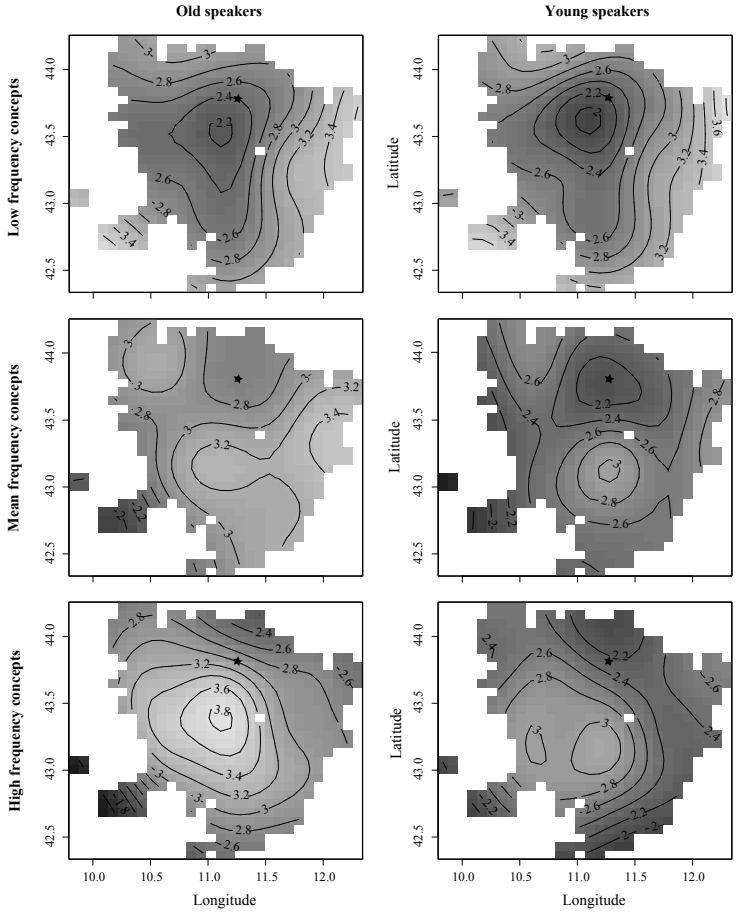


Figure 8.2. Contour plots for the regression surface of predicting lexical differences from standard Italian as a function of longitude, latitude, concept frequency, and speaker age group obtained with a generalized additive model. The (black) contour lines represent aggregate isoglosses, darker shades of gray (lower values) indicate a smaller lexical 'distance' from standard Italian, while lighter shades of gray (higher values) represent locations with a larger lexical 'distance' from standard Italian. The black star marks the approximate location of Florence.

well-known effect of (increasing) geographical distance (e.g., see Nerbonne and Kleiweg, 2007). We see this effect most clearly for the low frequency concepts, and reversed for the high frequency concepts.

Second, the standard literary Italian language was ill-equipped for use in

everyday discourse (very likely involving high frequency concrete concepts), and consequently the lexical gaps of the standard Italian (literary) language were filled with dialectal forms whose origin was not necessarily from Tuscany. Furthermore, during the evolution of the standard Italian language in the past centuries, alternative cognitively well-entrenched forms of high frequency concepts might have been preferred over the Florentine forms, and thus became part of the standard language instead.

Third, with the relatively recent emergence of the standard spoken language, a new wave of change is affecting Tuscany, causing Tuscan speakers to adopt the new standard Italian norm. This process is clearly documented in Figure 8.2, which shows that younger speakers (right panels) have moved closer to standard Italian than the older speakers (left panels).

Considering these three dimensions, the high frequency concepts in central Tuscany are more different (than low frequency concepts) from standard Italian for two reasons. First, the high frequency Florentine forms likely did not contribute as prominently to the standard Italian language because of the competition of alternative non-Florentine (well-entrenched) high frequency forms. Second, high frequency lexical forms are most resistant to replacement by the current equivalents in standard Italian because they are cognitively well entrenched in the central Tuscan (high prestige) speakers' mental lexicons. This explanation is in line with the finding reported in Chapter 6, where we observed that high frequency words were more resistant to standardization than low frequency words (see also Pagel et al., 2007).

With respect to the peripheral areas, the low frequency concepts (mainly belonging to an obsolete, progressively disappearing rural world) differ most from standard Italian as these are either represented by non-Tuscan dialectal forms (especially in the north and east, which border to other dialect areas), or by original Tuscan forms that were different from the Florentine norm due to geographical distance (or separation from the mainland). For medium and high frequency concepts in these peripheral areas, the pattern reverses, and the lexical forms are more likely to match the standard Italian form. With no close cultural ties to central Tuscany, and no prestige of its own, these dialects have been more open to accepting the standard Italian forms (spread via education and mass media).

We also investigated the effect of concept concreteness, but we did not find support for the significance of this predictor (both in the fixed- and random-effects structure). As the majority of our most abstract concepts were only mildly abstract (according to the categorization of Crutch and Warrington, 2005), this might have limited our ability to investigate the effect of abstract versus concrete concepts on lexical differences with respect to standard Italian.

8.4.2 Demographic predictors

When inspecting Table 8.1, it is clear that the contrast between the age groups was highly important, judging by its high z-value. Old speakers were much more likely to have a lexical form different from standard Italian. This result is not surprising as younger speakers tend to converge to standard Italian. There was no support for the inclusion of this predictor as a by-concept random slope, indicating that the effect of age was similar across concepts.

Of all demographic predictors (i.e. the community size, the average community income and the average community age) only the first was a significant predictor in the general model. Larger communities were more likely to have a lexical variant identical to standard Italian (i.e. the estimate in Table 8.1 is negative). A possible explanation for this finding is that people tend to have weaker social ties in urban communities, which causes dialect leveling (Milroy, 2002). As the standard Italian language is more prestigious than dialectal forms (Danesi, 1974), conversations will be normally held in standard Italian and leveling will proceed in the direction of standard Italian.

The other demographic predictors, average age and average income, were not significant in the general model. In Chapter 6, average age was identified as a significant predictor of pronunciation distance from standard Dutch, while average income was not. The effect of average community age may be less powerful here, as we have two age groups per location (which are much more suitable to detect the influence of age). In line with Chapter 6, the effect of average income pointed to a negative influence (with richer communities having lexical variants closer to the standard), but not significantly so (p = 0.3). Also note that year of recording was not significant in the general model, which is likely caused by the relatively short time span (with respect to lexical change) in which the data was gathered.[7]

All demographic variables showed significant by-concept variation. Figure 8.3, illustrating the effect of community size, shows some concepts (i.e. *ovile* 'sheepfold', *scoiattolo* 'squirrel', *orecchio* 'ear', and *mirtillo* 'blueberry') which are more likely to be identical to standard Italian in larger communities (i.e. consistent with the general pattern; the model estimate is indicated by the dashed line), while others behave in completely opposite fashion (i.e. *castagnaccio* 'chestnut cake', *melone* 'melon', *neve* 'snow', and *ditale* 'thimble') and are more likely to be different from standard Italian in larger communities. Many of these latter concepts (e.g., *castagnaccio*, but also *verro* 'boar, male swine', and *stollo* 'haystack pole', which are not marked in the graph) involve very old-fashioned rural concepts which may have fallen into disuse in larger cities, but not in smaller, more traditional, villages. As a consequence, people in larger cities may have forgotten the (old-fashioned) standard Italian lexical form, and use multi-word phrases or more general terms instead (e.g., 'pig' instead of 'boar').

[7] Year of recording was significant as a by-concept random slope, but as the other results were not altered significantly by its inclusion, we report the results of the simpler model.

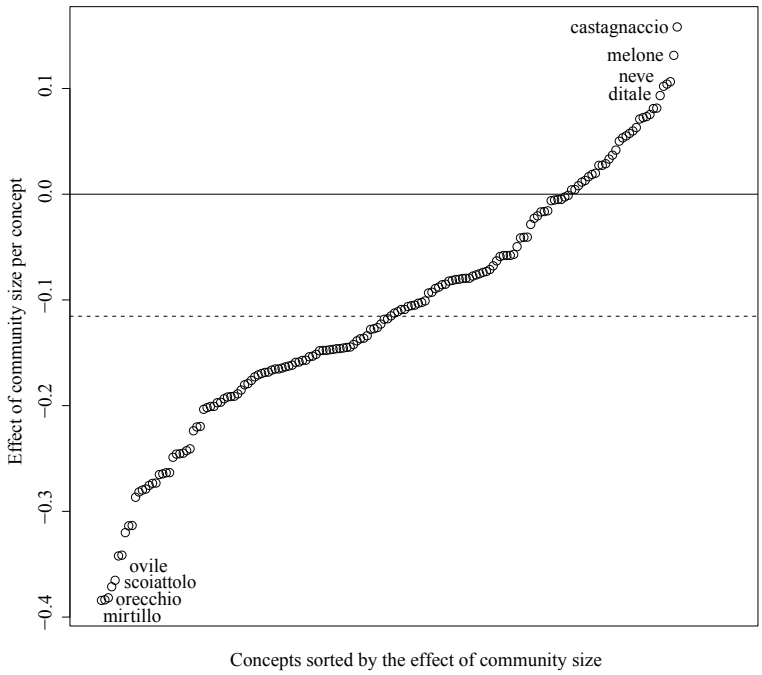


Figure 8.3. By-concept random slopes of community size. The concepts are sorted by the value of their community size coefficient (i.e. the effect of community size). The strongly negative coefficients (bottom-left) are associated with concepts that are more likely to be identical to standard Italian in larger communities, while the positive coefficients (top-right) are associated with concepts that are more likely to be different from standard Italian in larger communities. The model estimate (see Table 8.1) is indicated by the dashed line.

It is interesting to note that the set of latter concepts also includes *melone* and *ditale*, which represent two of the few well-known cases in which all Tuscan dialects diverge from standard Italian.

Figure 8.4 illustrates the by-concept random slopes for average age and average income (both were not significant as a fixed-effect factor). In the left part of the graph we see concepts which are more likely to have a standard Italian lexical form in richer communities (with concepts *caprone* 'goat', *cocca* 'corner', e.g., of a handkerchief, and *grattugia* 'grater' being close to the extreme), while the concepts in the bottom-right quadrant (e.g., *pimpinella* 'pimpernel', *stollo*

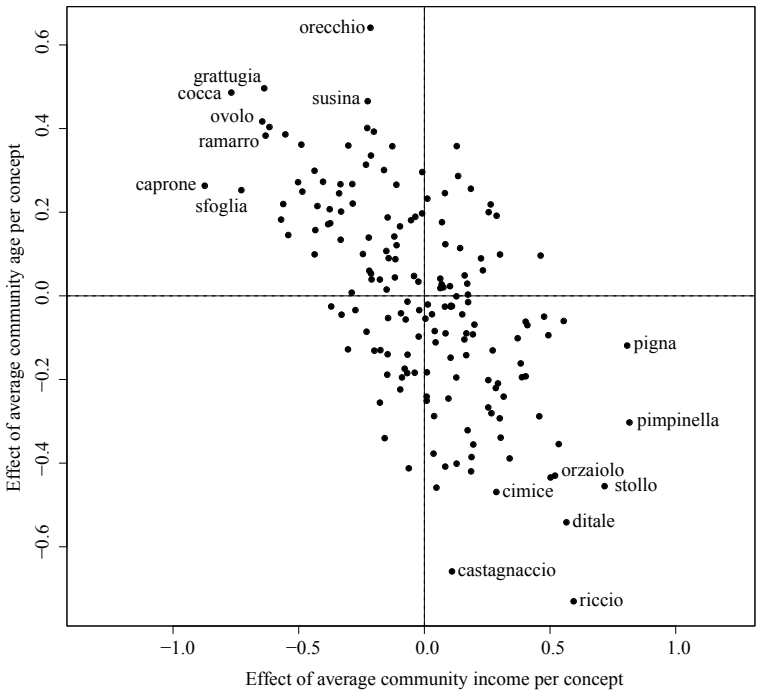


Figure 8.4. By-concept random slopes of average community age and income. The correlation between the by-concept random slopes is $r = -0.623$ ($p < 0.001$).

'haystack pole', and *ditale* 'thimble') demonstrate the opposite pattern and are more likely to have a lexical form different from standard Italian in richer communities.

The by-concept random slopes of average age and average income are closely linked (their correlation is $r = -0.623$, $p < 0.001$), which is reflected by the general negative trend in the scatter plot (see Figure 8.4). Concepts that are more likely to differ from the standard Italian form in poorer communities are also more likely to differ from the standard Italian form in communities with a higher average age (e.g., *cocca* and *grattugia* in the top-left). Similarly, concepts which follow the opposite pattern and are more likely to differ from standard Italian in richer communities, are also more likely to differ from the standard Italian lexical form in younger communities (e.g., *pimpinella* and *stollo* in the bottom-right). We observed a comparable result in Chapter 6, where by-word random slopes of average age, average income, as well as community size were

closely linked. In this case, however, there was no support for a link between the by-concept random slopes for community size and the other by-concept random slopes. Despite this, we again observe some old-fashioned rural concepts (e.g., *castagnaccio* and *stollo*) in the bottom-right quadrant, suggesting that also in younger and richer communities, people are less likely to remember these forms.

8.5 Discussion

In this chapter we have shown that the lexical variation in Tuscan dialects with respect to standard Italian can be adequately modeled by a generalized additive mixed-effects logistic regression model. We found clear support for the importance of speaker age, community size, as well as geography, which varied significantly depending on concept frequency and speaker age.

The results which have emerged from our analysis of the ALT corpus also shed new light on the widely debated *questione della lingua* from the point of view of Tuscan dialects. Previous studies, based both on individual words (Giacomelli and Poggi Salani, 1984) and on aggregated data (Montemagni, 2008), provided a flat view according to which Tuscan dialects overlap most closely with standard Italian in the area around Florence, with expansions in different directions and in particular towards the southwest. Montemagni's (2008) aggregate analysis also showed that a higher likelihood of using standard Italian was connected with speaker age and geographical coverage of words. The results of the analysis introduced in this chapter, however, provide a much more finely articulated picture in which new factors (such as concept frequency) are shown to play a significant role. In addition, these results allowed us to speculate about the spread of standard Italian with a particular emphasis on its relationship to the Florentine variety from which it originated.

For example, we observed that in the central Tuscan area, including Florence, high frequency concepts are more likely to differ from standard Italian than low frequency concepts, whereas in the marginal areas in the north, east and southwest a reverse pattern was observed. There, infrequent concepts are more likely to differ from standard Italian and frequent concepts are more likely to be identical to the standard. Frequency of concepts thus shows markedly different effects based on the history of the Italian language (originating from Florence) and the status attributed to the dialects in the specific area: the standard Italian language diverged more from the Florentine variety it originated from for high frequency concepts than for low frequency concepts, and also dialects from the central Tuscan area, including Florence, are accorded higher prestige than the dialects spoken in marginal areas of Tuscany, and are therefore able to counterbalance the rising of standard Italian.

On the demographic side, besides finding a significant effect of speaker age (with younger speakers using lexical forms more likely to be identical to stan-

dard Italian), we observed that larger communities are more likely to use standard Italian vocabulary than smaller communities. In addition, the effect of community size, but also average community age and income (even though these two were not significant as main effects), shows significant by-concept variation: concepts belonging to an obsolete disappearing rural world, such as 'haystack pole' and 'boar', are more likely to differ from the standard in larger, richer and younger communities, due to the fact that they are no longer part of everyday life.

Given the general features of the ALT dataset, it would be feasible to investigate other speaker-related characteristics by creating a different grouping (than the present age-based split). For example, it would be interesting to investigate the importance of speaker education or profession in this way.

In this chapter we used a binary lexical difference measure with respect to standard Italian. It would also be possible to use a more sensitive distance measure such as the Levenshtein distance introduced in Chapter 2. In that case, lexical differences which are closely related (i.e. in the case of lexicalized analogical formations) can be distinguished from more rigorous lexical differences. As this would not require the time-consuming logistic regression analysis, it would be possible to analyze all individual speakers (and incorporate their speaker-specific characteristics in the model specification) instead of simply grouping them.

In conclusion, this chapter has shown that the methodology developed in the previous two chapters can be usefully applied to lexical data. The next and final chapter of this thesis will provide some general conclusions about the work presented in this dissertation.

Part V

Conclusions

Chapter 9

A more comprehensive dialectometry

In this thesis we have aimed at making dialectometry more appealing to dialectologists. In Part I of this dissertation, we have argued that dialectometry has lacked a focus on individual linguistic features, as well as on the social domain. We therefore proposed new dialectometric methods taking into account social factors, and we have introduced techniques allowing us to investigate individual linguistic features.

Given the emphasis in this thesis on individual linguistic items (i.e. words and sound correspondences), a more sensitive discrimination between these items was needed than required for an aggregate analysis. Consequently, Part II of this dissertation focused on this aspect. Chapter 2 evaluated various string alignment algorithms with respect to their alignment quality. Due to its good performance, efficiency and intuitive interpretation, the Levenshtein distance algorithm incorporating automatically determined sound segment distances (on the basis of the information-theoretic pointwise mutual information measure) was selected to provide the dependent measures used in most of the subsequent chapters.

To demonstrate the linguistic sensitivity of the automatically determined segment distances used in the Levenshtein distance algorithm, Chapter 3 evaluated their quality by comparing them to acoustic vowel distances. The relatively strong significant correlations between the two (for several independent dialect datasets) confirmed that the algorithm also makes sense from a phonetic perspective.

Part III aimed at alleviating the lack of linguistic detail in most dialectometric work. We introduced hierarchical bipartite spectral graph partitioning to dialectometry in order to *simultaneously* identify geographical clusters of similar dialects together with their linguistic basis (in terms of sound correspondences). Chapter 4 evaluated the approach on a Dutch dialect dataset and also proposed a method to find the most characteristic sound correspondences in each cluster. Further support for the method was provided in Chapter 5, where it was applied to an English dataset and shown to complement traditional clustering approaches and principal component analysis.

Part IV was the most ambitious, in that it aimed at incorporating not only the influence of geography and individual linguistic features (i.e. words), but

also various sociolinguistic factors which are generally disregarded in dialectometric work. Chapter 6 used generalized additive modeling in combination with a mixed-effects regression analysis to predict dialect distances from standard Dutch for hundreds of words in more than 400 dialects. In addition to the importance of geography, we found support for the importance of several sociolinguistic factors. Communities with a small number of inhabitants or a high average age had pronunciations more distant from standard Dutch than those with a large number of inhabitants or a low average age. However, the precise effect of the community-related variables varied from word to word. Whereas most words followed the general pattern, some words even showed the opposite pattern with a higher distance from standard Dutch for larger and younger communities. In addition, we found support for the importance of word frequency and word category. The model predicted a greater distance from standard Dutch for more frequent words as well as for nouns (as opposed to verbs and adjectives), but the precise effect of these variables varied geographically. Besides providing some insight into lexical diffusion (Wang, 1969), these results also highlight the importance of examining a large set of items (following the dialectometric approach) in order to obtain a more comprehensive view of dialectal variation.

Chapter 7 improved on the methodology of Chapter 6 and showed that Catalan dialects are converging towards the standard language, but only in regions where Catalan is recognized as an official language. Chapter 8 illustrated the final application of the generalized additive mixed-effects regression approach, by predicting (binary) lexical differences between Tuscan dialects and standard Italian. The results of that chapter again revealed (in line with Chapter 6) that the geographical pattern of variation varied depending on concept frequency, and that more frequent concepts show more resistance to change (in the heartland of Tuscany). Besides identifying several significant concept-related characteristics, we also found a significant general effect of speaker age and population size (with younger speakers and speakers in larger communities using lexical forms more likely to match the standard language). Similar to the Dutch results, the effect of the community-related variables varied from concept to concept.

The approaches introduced in Part III and IV of this dissertation both allow a focus on individual linguistic features as well as geography. Despite this, they should be considered complementary. The bipartite spectral graph partitioning approach excludes social factors and always results in a (geographical) clustering on the basis of individual features, but does not require the use of a reference point. The regression approach, which does require a reference point, provides a more gradual view of the influence of geography on the basis of pronunciation distances or lexical differences, and allows the inclusion of social (and other) factors. Of course, the geographical results of both methods can be usefully compared. For example, Figure 4.3 shows that the peripheral areas of the Netherlands are clustered separately from the central area. In line with this,

Figure 6.1 shows that the peripheral areas have a greater distance from standard Dutch than the central area.

As we observed in Sections 4.4.1 and 5.4.1, the (geographical) results of our new methods generally corroborated those of the older aggregate-only methods (despite small differences) and identified the same main dialect areas. The advantage of the new methods, however, is that they provide additional (social or linguistic) information besides the general geographical pattern of dialect variation.

Of course, there are also other methods which tap into the linguistic basis of aggregate variation (see Chapter 5), such as principal component analysis (Shackleton, 2007), factor analysis (Grieve et al., 2011), or three-way multidimensional scaling (Ruette and Speelman, submitted), and these can adequately compete with the bipartite spectral graph partitioning method. Furthermore, Section 1.3 reports several dialectometric studies taking social variation into account (albeit to a limited extent). However, we do not know of any studies, other than those reported in Part IV of this dissertation, allowing a focus on individual linguistic items, while simultaneously taking into account geography, lexical factors, as well as various sociolinguistic variables.

The results of this thesis also give rise to further research questions. Our generalized additive mixed-effects regression approach focused on individual pronunciation distances (or lexical differences) per word. However, as dialectologists have frequently studied variation at the sound segment level, it would be very informative to create a model focusing on individual sound segment differences instead. Especially for datasets consisting of a limited set of words or locations, this would be computationally feasible.

A drawback of the analyses presented in this thesis is that they have focused exclusively on analyzing dialect atlas data. Consequently, the individual variation has effectively been eliminated as only a single pronunciation per speaker was present. It would be very interesting to apply the methods introduced in Part IV of this thesis to data containing within-speaker variation. In this way, we might investigate the effect of the specific social environment (Labov, 1972), while not restricting the analysis to only a few preselected linguistic variables.

Finally, with this dissertation we hope that we have taken adequate first steps in making dialectometry more comprehensive and interesting to dialectologists, by introducing methods which allow a focus on individual linguistic features and enable the integration of sociolinguistic factors.

Bibliography

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*. John Wiley & Sons, Hoboken, NJ, 2nd edition.

Akaike, H. (1974). A new look at the statistical identification model. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Alewijnse, B., Nerbonne, J., Van der Veen, L., and Manni, F. (2007). A computational analysis of Gabon varieties. In Osenova, P., editor, *Proceedings of the RANLP Workshop on Computational Phonology*, pages 3–12.

Almeida, A. and Braun, A. (1986). 'Richtig' und 'Falsch' in phonetischer Transkription: Vorschläge zum Vergleich von Transkriptionen mit Beispielen aus deutschen Dialekten. *Zeitschrift für Dialektologie und Linguistik*, LIII(2):158–172.

Anderson, P. M. (1987). *A Structural Atlas of the English Dialects*. Croom Helm, London.

Baayen, R. H. (2007). Storage and computation in the mental lexicon. In Jarema, G. and Libben, G., editors, *The Mental Lexicon: Core Perspectives*, pages 81–104. Elsevier.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge University Press.

Baayen, R. H. (2010). The directed compound graph of English. An exploration of lexical connectivity and its processing consequences. In Olson, S., editor, *New Impulses in Word-Formation (Linguistische Berichte Sonderheft 17)*, pages 383–402. Buske, Hamburg.

Baayen, R. H., Davidson, D., and Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412.

Baayen, R. H., Kuperman, V., and Bertram, R. (2010). Frequency effects in compound processing. In Scalise, S. and Vogel, I., editors, *Compounding*, pages 257–270. Benjamins, Amsterdam / Philadelphia.

Baayen, R. H., Milin, P., Durdević, D. F., Hendrix, P., and Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, 118(3):438–481.

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1996). CELEX2. Linguistic Data Consortium, Philadelphia.

Bailey, G., Wikle, T., Tillery, J., and Sand, L. (1991). The apparent time construct. *Language Variation and Change*, 3:241–264.

Bailey, T. M. and Hahn, U. (2005). Phoneme similarity and confusability. *Journal of Memory and Language*, 52(3):339–362.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov Chains. *The Annals of Mathematical Statistics*, 41(1):164–171.

Bibiloni, G. (2002). Un estàndard nacional o tres estàndards regionals? In Joan, B., editor, *Perspectives sociolingüístiques a les Illes Balears*. Res Publica, Eivissa.

Blancquaert, E. and Pée, W. (1925–1982). *Reeks Nederlandse Dialectatlassen*. De Sikkel, Antwerpen.

Bloomfield, L. (1933). *Language*. Holt, Rhinehart and Winston, New York.

Brants, T. and Franz, A. (2009). *Web 1T 5-gram, 10 European Languages. Version 1*. Linguistic Data Consortium, Philadelphia.

Brill, E. and Moore, R. C. (2000). An improved error model for noisy channel spelling correction. In *Proceedings of ACL 2000*, pages 286–293, Shroudsburg, PA. ACL.

Broe, M. (1996). A generalized information-theoretic measure for systems of phonological classification and recognition. In *Computational Phonology in Speech Technology: Proceedings of the Second Meeting of the ACL Special Interest Group in Computational Phonology*, pages 17–24, Santa Cruz. ACL.

Bryant, D., Filimon, F., and Gray, R. D. (2005). Untangling our past: Languages, trees, splits and networks. In Mace, R., Holden, C. J., and Shennan, S., editors, *The Evolution of Cultural Diversity: A Phylogenetic Approach*, pages 53–66. UCL Press, London.

Bybee, J. (2002). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change*, 14:261–290.

Calamai, S. (2003). Vocali fiorentine e vocali pisane a confronto. *Quaderni del Laboratorio di Linguistica, Scuola Normale Superiore di Pisa*, 3:40–71.

Cambra de Comerç - Indústria i Serveis d'Andorra (2008). Informe economic 2008. Available at http://www.ccis.ad/. Accessed: February 28, 2011.

Campbell, L. (2004). *Historical Linguistics: An Introduction*. Edinburgh University Press, Edinburgh.

Castellani, A. (1982). Quanti erano gli italofoni nel 1861? *Studi Linguistici Italiani*, 8:3–26.

CBS Statline (2010). Kerncijfers wijken en buurten 1995. Available at http://statline.cbs.nl. Accessed: August 9, 2010.

Chambers, J. and Trudgill, P. (1998). *Dialectology*. Cambridge University Press, Cambridge, 2nd edition.

Cheshire, J. (2002). Sex and gender in variationist research. In Chambers, J., Trudgill, P., and Schilling-Estes, N., editors, *The Handbook of Language Variation and Change*, pages 423–443. Blackwell Publishing Ltd.

Chomsky, N. A. (1965). *Aspects of the Theory of Syntax*. MIT Press, Cambridge.

Chung, F. (1997). *Spectral Graph Theory*. American Mathematical Society.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Cole, J. (2010). Editor's note. *Laboratory Phonology*, 1(1):1–2.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505.

Comuni Italiani (2011). Informazioni e dati statistici sui comuni in Italia, le province e le regioni italiane. Sito ufficiale, CAP, numero abitanti, utili link. Available at http://www.comuni-italiono.it. Accessed: May 23, 2011.

Covington, M. (1996). An algorithm to align words for historical comparison. *Computational Linguistics*, 22(4):481–496.

Crutch, S. and Warrington, E. (2005). Abstract and concrete concepts have structurally different representational frameworks. *Brain*, 128(3):615–627.

Cucchiarini, C. (1993). *Phonetic Transcription: A Methodological and Empirical Study*. PhD thesis, Katholieke Universiteit Nijmegen.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7:171–176.

Danesi, M. (1974). Teaching standard Italian to dialect speakers: A pedagogical perspective of linguistic systems in contact. *Italica*, 51(3):295–304.

De Mauro, T. (1963). *Storia linguistica dell'Italia unita*. Laterza, Bari-Roma.

De Schutter, G., Van den Berg, B., Goeman, T., and De Jong, T. (2005). *Morfologische Atlas van de Nederlandse Dialecten (MAND) Deel 1*. Amsterdam University Press, Meertens Instituut - KNAW, Koninklijke Academie voor Nederlandse Taal- en Letterkunde, Amsterdam.

Departament d'Estadística del Govern d'Andorra (2010). Societat i població. Available at http://www.estadistica.ad. Accessed: February 28, 2011.

Dhillon, I. (2001). Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 269–274. ACM New York, NY, USA.

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, United Kingdom.

Ellis, A. J. (1889). *On Early English Pronunciation. Part V*. Trübner, London.

Evert, S. (2005). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. PhD thesis, Universität Stuttgart.

Flege, J., Munro, M., and MacKay, I. (1995). Factors affecting strength of perceived foreign accent in a second language. *Journal of the Acoustical Society of America*, 97(5):3125–3134.

Fox, R. A. (1983). Perceptual structure of monophthongs and diphthongs in English. *Language and Speech*, 26(1):21–60.

Friedman, L. and Wall, M. (2005). Graphical views of suppression and multicollinearity in multiple regression. *The American Statistician*, 59:127–136.

Frisch, S. (1996). *Similarity and Frequency in Phonology*. PhD thesis, Northwestern University.

Frisch, S., Pierrehumbert, J. B., and Broe, M. B. (2004). Similarity avoidance and the OCP. *Natural Language & Linguistic Theory*, 22(1):179–228.

Giacomelli, G. (1975). Dialettologia toscana. *Archivio Glottologico Italiano*, 60:179–191.

Giacomelli, G. (1978). Come e perchè il questionario. In Giacomelli, G., editor, *Atlante Lessicale Toscano: Note sul questionario*, pages 19–26. Facoltà di Lettere e Filosofia, Firenze.

Giacomelli, G., Agostiniani, L., Bellucci, P., Giannelli, L., Montemagni, S., Nesi, A., Paoli, M., Picchi, E., and Poggi Salani, T., editors (2000). *Atlante Lessicale Toscano*. Lexis Progetti Editoriali, Roma.

Giacomelli, G. and Poggi Salani, T. (1984). Parole toscane. *Quaderni dell'Atlante Lessicale Toscano*, 2(3):123–229.

Goebl, H. (1984). *Dialektometrische Studien: Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF*. Max Niemeyer, Tübingen.

Goebl, H. (2000). Langues standards et dialectes locaux dans la France du Sud-Est et l'Italie septentrionale sous le coup de l'effet-frontière: Une approche dialectomètrique. *International Journal of the Sociology of Language*, 145:181–215.

Goebl, H. (2006). Recent advances in Salzburg dialectometry. *Literary and Linguistic Computing*, 21(4):411–435.

Goeman, A. (1999). *T-deletie in Nederlandse dialecten. Kwantitatieve analyse van structurele, ruimtelijke en temporele variatie*. Holland Academic Graphics.

Goeman, A. and Taeldeman, J. (1996). Fonologie en morfologie van de Nederlandse dialecten. Een nieuwe materiaalverzameling en twee nieuwe atlasprojecten. *Taal en Tongval*, 48:38–59.

Goossens, J. (1977). *Inleiding tot de Nederlandse dialectologie*. Wolters-Noordhoff, Groningen.

Gorman, K. (2010). The consequences of multicollinearity among socioeconomic predictors of negative concord in Philadelphia. In Lerner, M., editor, *University of Pennsylvania Working Papers in Linguistics*, volume 16, issue 2, pages 66–75.

Göschel, J. (1992). Das Forschungsinstitut für Deutsche Sprache "Deutscher Sprachatlas". Wissenschaftlicher Bericht, Das Forschungsinstitut für Deutsche Sprache, Marburg.

Grieve, J., Speelman, D., and Geeraerts, D. (2011). A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change*, 23:193–221.

Gusfield, D. (1999). *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge Univesity Press, Cambridge.

Gussenhoven, C. (1999). Illustrations of the IPA: Dutch. In *Handbook of the International Phonetic Association*, pages 74–77. Cambridge University Press, Cambridge.

Hamming, R. (1950). Error detecting and error correcting codes. *Bell System Technical Journal*, 29:147–160.

Harrell, F. (2001). *Regression Modeling Strategies*. Springer, Berlin.

Hasher, L. and Zacks, R. T. (1984). Automatic processing of fundamental information. The case of frequency of occurrence. *American Psychologist*, 39:1372–1388.

Heeringa, W. (2004). *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. PhD thesis, Rijksuniversiteit Groningen.

Heeringa, W. and Joseph, B. (2007). The relative divergence of Dutch dialect pronunciations from their common source: An exploratory study. In Nerbonne, J., Ellison, T. M., and Kondrak, G., editors, *Proceedings of the Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*, pages 31–39, Stroudsburg, PA. ACL.

Heeringa, W., Kleiweg, P., Gooskens, C., and Nerbonne, J. (2006). Evaluation of string distance algorithms for dialectology. In Nerbonne, J. and Hinrichs, E., editors, *Linguistic Distances*, pages 51–62, Shroudsburg, PA. ACL.

Heeringa, W. and Nerbonne, J. (1999). Change, convergence and divergence among Dutch and Frisian. In Boersma, P., Breuker, P. H., Jansma, L. G., and Van der Vaart, J., editors, *Philologia Frisica Anno 1999. Lêzingen fan it fyftjinde Frysk Filologekongres*, pages 88–109. Fryske Akademy, Ljouwert.

Hillenbrand, J., Getty, L., Clark, M., and Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, 97(5):3099–3111.

Hock, H. H. and Joseph, B. D. (1996). *Language History, Language Change, and Language Relationship: An Introduction to Historical and Comparative Linguistics*. Walter de Gruyter, Berlin.

Huguet, A., Vila, I., and Llurda, E. (2000). Minority language education in unbalanced bilingual situations: A case for the linguistic interdependence hypothesis. *Journal of Psycholinguistic Research*, 3:313–333.

Institut d'Estadística de Catalunya (2008, 2010). Territori. Available at http://www.idescat.cat. Accessed: February 28, 2011.

Instituto Aragonés de Estadística (2007, 2009, 2010). Población i territorio. Available at http://www.aragon.es. Accessed: February 28, 2011.

Jeffers, R. and Lehiste, I. (1979). *Principles and Methods for Historical Linguistics*. MIT Press, Cambridge.

Johnson, D. E. (2009). Getting off the GoldVarb standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass*, 3(1):359–383.

Keating, P., Lindblom, B., Lubker, J., and Kreiman, J. (1994). Variability in jaw height for segments in English and Swedish VCVs. *Journal of Phonetics*, 22:407–422.

Kernighan, M., Church, K., and Gale, W. (1990). A spelling-correction program based on the noisy channel model. In Kahlgren, H., editor, *Proceedings of COLING '90*, pages 205–210, Helsinki.

Kessler, B. (1995). Computational dialectology in Irish Gaelic. In *Proceedings of the Seventh Conference on European Chapter of the Association for Computational Linguistics*, pages 60–66, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Kloeke, G. G. (1927). *De Hollandse expansie in de zestiende en zeventiende eeuw en haar weerspiegeling in de hedendaagsche Nederlandse dialecten*. Martinus Nijhoff, The Hague.

Kluger, Y., Basri, R., Chang, J., and Gerstein, M. (2003). Spectral biclustering of microarray data: Coclustering genes and conditions. *Genome Research*, 13(4):703–716.

Kondrak, G. (2002). Determining recurrent sound correspondences by inducing translation models. In *Proceedings of the Nineteenth International Conference on Computational Linguistics (COLING 2002)*, pages 488–494, Taipei. COLING.

Kondrak, G. (2003). Phonetic alignment and similarity. *Computers and the Humanities*, 37:273–291.

Kondrak, G. and Dorr, B. (2006). Automatic identification of confusable drug names. *Artificial Intelligence in Medicine*, 36(1):29–42.

Kretzschmar, Jr., W., editor (1994). *Handbook of the Linguistic Atlas of the Middle and South Atlantic States*. The University of Chicago Press, Chicago.

Kretzschmar, Jr., W. (1996). Quantitative areal analysis of dialect features. *Language Variation and Change*, 8:13–39.

Labov, W. (1966). *The Social Stratification of English in New York City*. Center for Applied Linguistics, Washington.

Labov, W. (1972). *Sociolinguistic Patterns*. University of Pennsylvania Press, Philadelphia.

Labov, W. (1981). Resolving the Neogrammarian controversy. *Language*, 57:267–308.

Labov, W. (2001). *Principles of Linguistic Change, Volume 2. Social Factors*. Blackwell Publishers Inc, Malden, MA.

Labov, W., Yaeger, M., and Steiner, R. (1972). *A Quantitative Study of Sound Change in Progress*. U.S. Regional Survey, Philadelphia.

Laver, J. (1994). *Principles of Phonetics*. Cambridge Univeristy Press, Cambridge.

Lehiste, I. and Popov, K. (1970). Akustische Analyse bulgarischer Silbenkerne. *Phonetica*, 21:40–48.

Lepschy, G. (2002). *Mother Tongues & Other Reflections on the Italian Language*. University of Toronto Press, Toronto.

Levenshtein, V. (1965). Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk SSSR*, 163:845–848. In Russian.

Lisker, L. and Abramson, A. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 20(3):384–422.

Lobanov, B. (1971). Classification of Russian vowels spoken by different speakers. *The Journal of the Acoustical Society of America*, 49:606–608.

Mackay, W. and Kondrak, G. (2005). Computing word similarity and identifying cognates with Pair Hidden Markov Models. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL)*, pages 40–47, Morristown, NJ, USA. Association for Computational Linguistics.

Maiden, M. (1995). *A Linguistic History of Italian*. Longman, London.

Maiden, M. and Parry, M. (1997). *The Dialects of Italy*. Routledge, London.

Manning, C. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press, Cabridge, Massachusetts.

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

Maye, J., Werker, J., and Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82:B101–B111.

Meillet, A. (1903). *Introduction à l'étude comparative des langues indo-européennes*. Librairie Hachette et Cie, Paris.

Mielke, J. (2005). Modeling distinctive feature emergence. In Alderete, J., Han, C., and Kochetov, A., editors, *Proceedings of the 24th West Coast Conference on Formal Linguistics*, pages 281–289, Somerville, MA. Cascadilla Proceedings Project.

Migliorini, B. and Griffith, T. (1984). *The Italian Language*. Faber and Faber, London.

Miller, G. A. and Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America*, 27:338–352.

Milroy, L. (2002). Social networks. In Chambers, J., Trudgill, P., and Schilling-Estes, N., editors, *The Handbook of Language Variation and Change*, pages 549–572. Blackwell Publishing Ltd.

Milroy, L. and Margrain, S. (1980). Vernacular language loyalty and social network. *Language in Society*, 9:43–70.

Montemagni, S. (2007). Patterns of phonetic variation in Tuscany: Using dialectometric techniques on multi-level representations of dialectal data. In Osenova, P., editor, *Proceedings of the RANLP Workshop on Computational Phonology*, pages 49–60.

Montemagni, S. (2008). Analisi linguistico-computazionali del corpus dialettale dell'Atlante Lessicale Toscano. Primi risultati sul rapporto toscano-italiano. In Nesi, A. and Maraschio, N., editors, *Discorsi di lingua e letteratura italiana per Teresa Poggi Salani (Strumenti di filologia e critica, vol. 3)*, pages 247–260. Pacini, Pisa.

Montemagni, S., Wieling, M., De Jonge, B., and Nerbonne, J. (accepted). Synchronic patterns of Tuscan phonetic variation and diachronic change: Evidence from a dialectometric study. *LLC. The Journal of Digital Scholarship in the Humanities*.

Montemagni, S., Wieling, M., De Jonge, B., and Nerbonne, J. (in press). Patterns of language variation and underlying linguistic features: A new dialectometric approach. In De Blasi, N., editor, *La variazione nell'italiano e nella sua storia. Varietà e varianti linguistiche e testuali. Proceedings of the Congresso della Società Internazionale di Linguistica e Filologia Italiana*.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453.

Nerbonne, J. (2006). Identifying linguistic structure in aggregate comparison. *Literary and Linguistic Computing*, 21(4):463–476. Special Issue, J. Nerbonne

and W. Kretzschmar, Jr., editors, Progress in Dialectometry: Toward Explanation.

Nerbonne, J. (2009). Data-driven dialectology. *Language and Linguistics Compass*, 3(1):175–198.

Nerbonne, J. (2010). Mapping aggregate variation. In Lameli, A., Kehrein, R., and Rabanus, S., editors, *Language and Space. International Handbook of Linguistic Variation. Volume 2: Language Mapping*, pages 476–495, 2401–2406. Mouton de Gruyter, Berlin.

Nerbonne, J., Colen, R., Gooskens, C., Kleiweg, P., and Leinonen, T. (2011). Gabmap — a web application for dialectology. *Dialectologia*, Special Issue II:65–89.

Nerbonne, J., Heeringa, W., Van den Hout, E., Van de Kooi, P., Otten, S., and Van de Vis, W. (1996). Phonetic distance between Dutch dialects. In Durieux, G., Daelemans, W., and Gillis, S., editors, *CLIN VI: Proceedings of the Sixth CLIN Meeting*, pages 185–202, Antwerp. Centre for Dutch Language and Speech.

Nerbonne, J. and Kleiweg, P. (2007). Toward a dialectological yardstick. *Journal of Quantitative Linguistics*, 14:148–167.

Nerbonne, J. and Siedle, C. (2005). Dialektklassifikation auf der Grundlage aggregierter Ausspracheunterschiede. *Zeitschrift für Dialektologie und Linguistik*, 72:129–147.

Niebaum, H. and Macha, J. (2006). *Einführung in die Dialektologie des Deutschen*. Max Niemeyer Verlag, Tübingen, 2nd edition.

Nurse, D. and Philippson, G. (2003). *The Bantu Languages*. Routledge, London.

Oakes, M. P. (2000). Computer estimation of vocabulary in a protolanguage from word lists in four daughter languages. *Journal of Quantitative Linguistics*, 7(3):233–243.

Ohala, J. J. (1997). Comparison of speech sounds: Distance vs. cost metrics. In *Speech Production and Language. In Honor of Osamu Fujimura*, pages 261–270. Mouton de Gruyter, Berlin.

Orton, H., Dieth, E., Halliday, W., Barry, M., Tilling, P., and Wakelin, M., editors (1962–1971). *Survey of English Dialects B: The Basic Material*. E.J. Arnold, Leeds.

Pagel, M., Atkinson, Q., and Meade, A. (2007). Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature*, 449:717–720.

Paolillo, J. C. (2002). *Analyzing Linguistic Variation: Statistical Models and Methods*. Center for the Study of Language and Information, Stanford, California.

Pierrehumbert, J. B. (1993). Dissimilarity in the Arabic verbal roots. In *Proceedings of the North East Linguistics Society*, volume 23, pages 367–381. GLSA, Amherst, MA.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. Statistics and Computing. Springer, New York.

Piske, T., MacKay, I. R., and Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics*, 29(2):191–215.

Poggi Salani, T. (1978). Dialetto e lingua a confronto. In Giacomelli, G., editor, *Atlante Lessicale Toscano: Note sul questionario*, pages 51–65. Facoltà di Lettere e Filosofia, Firenze.

Pols, L., Tromp, H., and Plomp, R. (1973). Frequency analysis of Dutch vowels from 50 male speakers. *The Journal of the Acoustical Society of America*, 43:1093–1101.

Pouliquen, B. (2008). Similarity of names across scripts: Edit distance using learned costs of n-grams. In Nordström, B. and Ranta, A., editors, *Proceedings of the Sixth International Conference on Natural Language Processing*, pages 405–416.

Pradilla, M.-À. (2008a). *La tribu valenciana. Reflexions sobre la desestructuració de la comunitat lingüística*. Onada, Benicarló.

Pradilla, M.-À. (2008b). *Sociolingüística de la variació i llengua catalana*. Institut d'Estudis Catalans, Barcelona.

Prokić, J. (2007). Identifying linguistic structure in a quantitative analysis of dialect pronunciation. In *Proceedings of the ACL 2007 Student Research Workshop*, pages 61–66, Prague. Association for Computational Linguistics.

Prokić, J. (2010). *Families and Resemblances*. PhD thesis, Rijksuniversiteit Groningen.

Prokić, J., Çöltekin, Ç., and Nerbonne, J. (2012). Detecting shibboleths. In Butt, M., Prokić, J., Mayer, T., and Cysouw, M., editors, *Visualization of Linguistic Patterns and Uncovering Language History from Multilingual Resources*, pages 72–80.

Prokić, J., Nerbonne, J., Zhobov, V., Osenova, P., Simov, K., Zastrow, T., and Hinrichs, E. (2009a). The computational analysis of Bulgarian dialect pronunciation. *Serdica Journal of Computing*, 3:269–298.

Prokić, J., Wieling, M., and Nerbonne, J. (2009b). Multiple sequence alignments in linguistics. In Lendvai, P. and Borin, L., editors, *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*, pages 18–25.

Rabiner, L. R. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Rafel, J. (1996–1998). *Diccionari de freqüències. Corpus textual informatitzat de la llengua catalana*. Institut d'Estudis Catalans, Barcelona.

Ristad, E. S. and Yianilos, P. N. (1998). Learning string-edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:522–532.

Ruette, T. and Speelman, D. (submitted). Transparant aggregation of linguistic variables with individual differences scaling.

Sanders, N. and Chin, S. B. (2009). Phonological distance measures. *Journal of Quantitative Linguistics*, 16(1):96–114.

Sankoff, D. and Kruskal, J., editors (1999). *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*. CSLI, Stanford.

Schmidt, M., Kiviste, A., and von Gadow, K. (2011). A spatially explicit height-diameter model for Scots pine in Estonia. *European Journal of Forest Research*, 130(2):303–315.

Schneider, E. (1988). Qualitative vs. quantitative methods of area delimitation in dialectology: A comparison based on lexical data from Georgia and Alabama. *Journal of English Linguistics*, 21:175–212.

Schuchardt, H. (1885). *Über die Lautgesetze: Gegen die Junggrammatiker*. Oppenheim, Berlin.

Séguy, J. (1971). La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane*, 35(138):335–357.

Séguy, J. (1973). La dialectométrie dans l'Atlas linguistique de Gascogne. *Revue de Linguistique Romane*, 37(145):1–24.

Sendlmeier, W. and Seebode, J. (2006). Formantkarten des deutschen Vokalsystems. TU Berlin. Available at http://www.kgw.tu-berlin.de/forschung/Formantkarten. Accessed: November 1, 2010.

Shackleton, Jr., R. G. (2005). English-American speech relationships: A quantitative approach. *Journal of English Linguistics*, 33(2):99–160.

Shackleton, Jr., R. G. (2007). Phonetic variation in the traditional English dialects. *Journal of English Linguistics*, 35(1):30–102.

Shackleton, Jr., R. G. (2010). *Quantitative Assessment of English-American Speech Relationships*. PhD thesis, Rijksuniversiteit Groningen.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905.

Smakman, D. (2006). *Standard Dutch in the Netherlands. A Sociolinguistic and Phonetic Description*. PhD thesis, Radboud Universiteit.

Szmrecsanyi, B. (2011). Corpus-based dialectometry: A methodological sketch. *Corpora*, 6(1):45–76.

Taeldeman, J. and Verleyen, G. (1999). De FAND: Een kind van zijn tijd. *Taal en Tongval*, 51:217–240.

Tagliamonte, S. A. and Baayen, R. H. (in press). Models, forests and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change*.

Togerson, W. (1952). Multidimensional scaling. I. Theory and method. *Psychometrika*, 17:401–419.

Tougaard, J. and Eriksen, N. (2006). Analysing differences among animal songs quantitatively by means of the Levenshtein distance measure. *Behaviour*, 143(2):239–252.

Toutanova, K. and Moore, R. C. (2002). Pronunciation modeling for improved spelling correction. In *Proceedings of ACL 2002*, pages 144–151, Shroudsburg, PA. ACL.

Traunmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. *The Journal of the Acoustical Society of America*, 88:97–100.

Tremblay, A. and Baayen, R. H. (2010). Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In Wood, D., editor, *Perspectives on Formulaic Language: Acquisition and Communication*, pages 151–173. The Continuum International Publishing Group, London.

Trudgill, P. (1974a). Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography. *Language in Society*, 3(2):215–246.

Trudgill, P. (1974b). *The Social Differentiation of English in Norwich*. Cambridge University Press, Cambridge.

Trudgill, P. (1999). *The Dialects of England*. Blackwell, Oxford, 2nd edition.

Valls, E., Wieling, M., and Nerbonne, J. (accepted). Linguistic advergence and divergence in north-western Catalan: A dialectometric investigation of dialect leveling and border effects. *LLC. The Journal of Digital Scholarship in the Humanities*.

Van Bree, C. (1987). *Historische grammatica van het Nederlands*. Foris Publications, Dordrecht.

Van de Velde, H., Kissine, M., Tops, E., Van der Harst, S., and Van Hout, R. (2010). Will Dutch become Flemish? Autonomous developments in Belgian Dutch. *Multilingua*, 29:385–416.

Van den Berg, B. (2003). *Phonology and Morphology of Dutch and Frisian Dialects in 1.1 million transcriptions*. Goeman-Taeldeman-Van Reenen project 1980–1995, Meertens Instituut Electronic Publications in Linguistics 3. Meertens Instituut (CD-ROM), Amsterdam.

Van der Wal, M. and Van Bree, C. (2008). *Geschiedenis van het Nederlands*. Spectrum, Utrecht, 5th edition.

Van Heuven, V. J. and Van Bezooijen, R. (1995). Quality evaluation of synthesized speech. In Paliwal, K., editor, *Speech Coding and Synthesis*, pages 707–738. Elsevier Science, Amsterdam.

Van Nierop, D., Pols, L., and Plomp, R. (1973). Frequency analysis of Dutch vowels from 25 female speakers. *Acoustica*, 29:110–118.

Van Reenen, P. (2006). *In Holland staat een 'Huis'. Kloekes expansietheorie met speciale aandacht voor de dialecten van Overijssel*. Stichting Neerlandistiek VU & Nodus Publikationen, Amsterdam & Münster.

Wagner, R. and Lowrance, R. (1975). An extension of the string-to-string correction problem. *Journal of the ACM*, 22(2):177–183.

Wang, W. (1969). Competing changes as a cause of residue. *Language*, 45(1):9–25.

Wattenmaker, W. and Shoben, E. (1987). Context and the recallability of concrete and abstract sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13(1):140–150.

Wenker, G. (1881). *Sprach-Atlas von Nord- und Mitteldeutschland. Auf Grund von systematisch mit Hülfe der Volksschullehrer gesammeltem Material aus circa 30000 Orten. Abtheilung I, Lieferung 1 (6 karten und Textheft)*. Trübner, Strasbourg.

Wieling, M., Heeringa, W., and Nerbonne, J. (2007a). An aggregate analysis of pronunciation in the Goeman-Taeldeman-Van Reenen-Project data. *Taal en Tongval*, 59:84–116.

Wieling, M., Leinonen, T., and Nerbonne, J. (2007b). Inducing sound segment differences using Pair Hidden Markov Models. In Nerbonne, J., Ellison, M., and Kondrak, G., editors, *Computing and Historical Phonology: Ninth Meeting of the ACL Special Interest Group for Computational Morphology and Phonology*, pages 48–56.

Wieling, M., Margaretha, E., and Nerbonne, J. (2011a). Inducing phonetic distances from dialect variation. *Computational Linguistics in the Netherlands Journal*, 1:109–118.

Wieling, M., Margaretha, E., and Nerbonne, J. (2012). Inducing a measure of phonetic similarity from dialect variation. *Journal of Phonetics*, 40(2):307–314.

Wieling, M., Montemagni, S., Nerbonne, J., and Baayen, R. H. (submitted-a). Lexical differences between Tuscan dialects and standard Italian: A sociolinguistic analysis using generalized additive mixed modeling.

Wieling, M. and Nerbonne, J. (2007). Dialect pronunciation comparison and spoken word recognition. In Osenova, P., editor, *Proceedings of the RANLP Workshop on Computational Phonology*, pages 71–78.

Wieling, M. and Nerbonne, J. (2009). Bipartite spectral graph partitioning to co-cluster varieties and sound correspondences in dialectology. In Choudhury, M., Hassan, S., Mukherjee, A., and Muresan, S., editors, *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, pages 26–34.

Wieling, M. and Nerbonne, J. (2010). Hierarchical spectral partitioning of bipartite graphs to cluster dialects and identify distinguishing features. In Banea, C., Moschitti, A., Somasundaran, S., and Zanzotto, F. M., editors, *Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing*, pages 33–41.

Wieling, M. and Nerbonne, J. (2011a). Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features. *Computer Speech and Language*, 25(3):700–715.

Wieling, M. and Nerbonne, J. (2011b). Measuring linguistic variation commensurably. *Dialectologia*, Special Issue II:141–162.

Wieling, M., Nerbonne, J., and Baayen, R. H. (2011b). Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PLoS ONE*, 6(9):e23613.

Wieling, M., Prokić, J., and Nerbonne, J. (2009). Evaluating the pairwise alignment of pronunciations. In Borin, L. and Lendvai, P., editors, *Proceedings of*

the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education, pages 26–34.

Wieling, M., Shackleton, Jr., R. G., and Nerbonne, J. (accepted). Analyzing phonetic variation in the traditional English dialects: Simultaneously clustering dialects and phonetic features. *LLC. The Journal of Digital Scholarship in the Humanities*.

Wieling, M., Valls, E., Baayen, R. H., and Nerbonne, J. (submitted-b). The effects of language policies on standardization of Catalan dialects: A sociolinguistic analysis using generalized additive mixed modeling.

Wolfram, W. and Schilling-Estes, N. (2006). *American English: Dialects and Variation*. Wiley-Blackwell, Malden, MA, 2nd edition.

Wood, S. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):95–114.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC.

Woolard, K. and Gahng, T.-J. (2008). Changing language policies and attitudes in autonomous Catalonia. *Language in Society*, 19:311–330.

Woolhiser, C. (2005). Political borders and dialect divergence/convergence in Europe. In Auer, P., Hinskens, F., and Kerswill, P., editors, *Dialect Change. Convergence and Divergence in European Languages*, pages 236–262. Cambridge University Press, New York.

Wrede, F., Martin, B., and Mitzka, W., editors (1927–1956). *Deutscher Sprachatlas auf Grund des von Georg Wenker begründeten Sprachatlas des deutschen Reiches*. Elwert, Marburg. 23 fascicles.

Summary

This dissertation focuses on dialect variation. In the field of dialectology, researchers study the influence of various factors on variation in dialectal speech. Initially, dialectologists were interested in identifying geographical dialect regions. Later, however, they became more interested in the influence of social characteristics (such as age and status). In general, (social) dialectologists have restricted their studies to a small set of linguistic variables (e.g., word pronunciations) which they selected themselves.

Dialectometry, a subfield of dialectology, is characterized by a less subjective approach, in which a large set of linguistic variables is analyzed simultaneously. By counting how many linguistic variables differ between one speaker and another, a measure of linguistic distance between the two speakers is obtained. This measure can be made more precise by taking into account how much individual pronunciations differ (e.g., the pronunciation of 'can' is closer to 'cat' than to 'bat').

In contrast to dialectology, dialectometric studies have lacked a focus on the social dimension and have almost exclusively investigated the connection between dialect variation and geography. By calculating a single distance between two speakers on the basis of a large set of linguistic variables, researchers in dialectometry have also been criticized for their lack of attention to the contribution of individual linguistic variables. Consequently, in this thesis we propose new dialectometric methods integrating social factors as well as allowing a focus on individual linguistic variables.

Given the importance of individual linguistic variables (in our case word pronunciations), our distance measure needs to be as precise as possible. Besides considering the number of different sounds in two pronunciations, it also makes sense to look at which sounds are involved. For example when only a single sound differs between two pronunciations, it should matter if the corresponding sounds are relatively similar (e.g., [o] versus [u]) or very different (e.g., [o] versus [i]). In Chapter 2 we introduce a novel method to *automatically* obtain sensitive sound distances. Chapter 3 shows that these sound distances make sense acoustically (e.g., the obtained distance between [o] and [u] is lower than the obtained distance between [o] and [i]), and this suggests that our pronunciation distances will also be more precise. In addition, by using these sensitive sound distances we improve our ability to identify which sounds correspond in different pronunciations (such as [w] and [v] in two pronunciations

of 'vinegar', [wɪnɪgə] and [vɪnɪgə]). This is demonstrated in Chapter 2.

In Chapter 4 we propose a novel method to identify groups of linguistically similar dialects while simultaneously identifying their underlying linguistic basis (in terms of sound correspondences). Besides applying the method to a Dutch dialect dataset in Chapter 4, we also apply it to an English dialect dataset in Chapter 5. In both cases, we find sensible geographical clusters together with their most characteristic sound correspondences.

In Chapters 6, 7 and 8 we propose an integrative approach to simultaneously investigate the effect of geography, several word-related factors (such as word frequency), and various social factors (such as speaker age) on dialect variation at the word level. The wide applicability of this approach (combining mixed-effects regression and generalized additive modeling) is illustrated by applying the method to three different dialect datasets.

Chapter 6 investigates a Dutch dialect dataset. In addition to the importance of geography, we find clear support for the importance of several demographic factors. Communities with a small number of inhabitants or a high average age have dialectal pronunciations more distant from standard Dutch than those with a large number of inhabitants or a low average age. In addition, we observe that nouns (as opposed to verbs and adjectives) and more frequent words are more resistant to standardization.

In Chapter 7 we investigate a Catalan dialect dataset containing dialectal pronunciations of speakers from Catalonia, Andorra and Aragon. As Catalan is not recognized as an official language in Aragon (in contrast to Catalonia and Andorra), this dataset allows us to study the effect of the standard language on dialectal pronunciations. The results clearly show that Catalan dialects are converging towards the standard language (i.e. younger speakers use pronunciations more similar to standard Catalan than older speakers), but only in regions where Catalan is recognized as an official language and taught in school. Consequently, the presence of an official standard language influences dialectal variation.

While Chapters 6 and 7 focus on pronunciation distances, our method also allows us to study lexical differences (e.g., using 'car' as opposed to 'automobile'). Chapter 8 investigates lexical differences between Tuscan dialects and standard Italian. In line with Chapter 6, we observe that more frequent words (in the heartland of Tuscany) are more resistant to standardization. In addition, younger speakers and speakers in larger communities are more likely to use standard Italian lexical forms.

In conclusion, the novel dialectometric methods proposed in this dissertation should be more appealing to dialectologists, as they incorporate social factors and allow a focus on individual linguistic variables. Furthermore, examining a large set of linguistic variables allows us to obtain a more comprehensive view of dialectal variation than by using only a small set of linguistic variables.

Samenvatting

Het centrale thema van dit proefschrift is dialectvariatie. Onderzoekers binnen de dialectologie bestuderen de invloed van verschillende factoren op deze variatie. In eerste instantie waren dialectologen enkel geïnteresseerd in het bepalen van geografische dialectgebieden, maar later verlegden ze hun interesse steeds meer naar de invloed van sociale kenmerken (zoals leeftijd en status). Dialectologen hebben eigenlijk altijd hun onderzoeken beperkt tot een kleine set van zelfgekozen taalkundige variabelen (bijvoorbeeld de uitspraak van enkele woorden).

Dialectometrie is een deelgebied van de dialectologie waarbij een grote en dus minder subjectieve set van taalkundige variabelen tegelijkertijd wordt geanalyseerd. Door te tellen hoeveel taalkundige variabelen verschillen wanneer de ene spreker met de ander wordt vergeleken, wordt een maat verkregen van de dialectafstand tussen de twee sprekers. Kijken we naar verschillen in uitspraak, dan kan deze maat preciezer worden gemaakt door te bepalen hoeveel klanken verschillend zijn (de afstand tussen de woorden 'lat' en 'lot' is dan bijvoorbeeld kleiner dan de afstand tussen 'lat' en 'lof').

In tegenstelling tot dialectologen besteden onderzoekers in de dialectometrie weinig aandacht aan sociale factoren, maar richten ze zich voornamelijk op de verbinding tussen dialectvariatie en geografie. Daarnaast ontbreekt er in de dialectometrie vaak aandacht voor de rol van de individuele taalkundige variabelen (waarin dialectologen juist wel geïnteresseerd zijn), omdat deze samengenomen worden bij het bepalen van de dialectafstanden. Vanuit die kritiek ontwikkelen we in dit proefschrift nieuwe dialectometrische methodes, met aandacht voor zowel de rol van individuele taalkundige variabelen als die van sociale factoren.

Vanwege de nadruk op de rol van individuele taalkundige variabelen (in ons geval woorduitspraken) is het belangrijk dat onze maat voor het bepalen van de uitspraakverschillen zo precies mogelijk is. Naast het meenemen van het aantal verschillende klanken in twee uitspraken, is het goed om naar de specifieke klanken te kijken. Het zou bijvoorbeeld uit moeten maken voor de afstandsmaat of de klanken op elkaar lijken (bijvoorbeeld [o] en [u]), of juist sterk van elkaar verschillen (bijvoorbeeld [o] en [i]). We introduceren daarom in hoofdstuk 2 een nieuwe methode om *automatisch* gevoelige klankafstanden te bepalen. In hoofdstuk 3 laten we zien dat deze klankafstanden akoestisch zinvol zijn (de gevonden afstand tussen [o] en [u] is bijvoorbeeld kleiner dan

de gevonden afstand tussen [o] en [i]) en dit suggereert dat onze uitspraakafstanden hierdoor ook preciezer worden. Tevens zien we in hoofdstuk 2 dat we door het gebruiken van deze gevoelige klankafstanden beter kunnen bepalen welke klanken corresponderen in verschillende uitspraken (zoals [s] en [z] in twee uitspraken van het woord 'zon', [sɔn] en [zɔn]) .

In hoofdstuk 4 introduceren we een nieuwe methode om groepen van vergelijkbare dialecten te vinden, waarbij we tegelijkertijd de onderliggende taalkundige basis (gebaseerd op klankcorrespondenties) bepalen. In hoofdstuk 4 passen we de methode toe op Nederlandse en in hoofdstuk 5 op Engelse dialecten. In beide gevallen vinden we aannemelijke geografische dialectgebieden en identificeren we hun meest karakteristieke klankcorrespondenties.

In hoofdstuk 6, 7 en 8 introduceren en evalueren we een integrale aanpak waarbij we de effecten onderzoeken van diverse factoren op dialectvariatie per woord. Niet alleen nemen we geografie mee, maar ook onderzoeken we de rol van verschillende sociale en woord-gerelateerde factoren (zoals leeftijd van de spreker en woordfrequentie). De brede toepasbaarheid van deze (regressie) aanpak wordt geïllustreerd door de methode op drie verschillende dialect dataverzamelingen toe te passen.

In hoofdstuk 6 onderzoeken we een dataverzameling van Nederlandse dialecten. Naast het belang van geografie vinden we duidelijke steun voor diverse demografische factoren. De dialectuitspraken van een gemeenschap met een klein aantal inwoners of een hoge gemiddelde leeftijd wijken meer af van de Nederlandse standaardtaal dan die van een gemeenschap met een groot aantal inwoners of een lage gemiddelde leeftijd. Ook zien we dat zelfstandige naamwoorden (in vergelijking met werkwoorden en bijvoeglijke naamwoorden) en meer frequente woorden meer resistent zijn tegen standaardisatie.

In hoofdstuk 7 onderzoeken we een dataverzameling van Catalaanse dialecten waarin de uitspraken van sprekers uit Catalonië, Andorra en Aragon zijn opgenomen. Doordat Catalaans niet als officiële taal in Aragon erkend wordt en ook op school niet onderwezen wordt (in tegenstelling tot Catalonië en Andorra), kunnen we deze dataverzameling gebruiken om het effect van het wel of niet hebben van een standaardtaal te onderzoeken. De resultaten laten duidelijk zien dat Catalaanse dialecten standaardiseren (jongere sprekers hebben uitspraken die meer lijken op standaard Catalaans dan oudere sprekers), maar alleen in die regio's waar Catalaans erkend wordt als officiële taal (en op school onderwezen wordt). Hieruit blijkt duidelijk dat het hebben van een standaardtaal dialectvariatie beïnvloedt.

Terwijl hoofdstukken 6 en 7 zich richten op uitspraakverschillen, is onze methode ook geschikt voor het analyseren van lexicale verschillen (zoals het gebruik van 'bolide' in plaats van 'auto'). In hoofdstuk 8 onderzoeken we lexicale verschillen tussen Toscaanse dialecten en standaard Italiaans. Vergelijkbaar met de resultaten van hoofdstuk 6 vinden we ook hier dat meer frequente woorden (in het centrale gebied van Toscane) resistenter zijn tegen standaardisatie. Daarnaast zien we dat jongere sprekers, maar ook sprekers die in grotere

gemeenschappen wonen een grotere kans hebben om een lexicale vorm te gebruiken die overeenkomt met die van de standaardtaal.

Samenvattend hebben we in dit proefschrift dialectometrische methodes ontwikkeld die aantrekkelijker zouden moeten zijn voor dialectologen, aangezien deze methodes niet alleen aandacht hebben voor de invloed van geografie, maar juist ook voor de rol van individuele taalkundige variabelen en sociale factoren. Daarnaast geeft het onderzoeken van een groot aantal taalkundige variabelen een vollediger beeld van dialectvariatie dan wanneer alleen een klein aantal taalkundige variabelen wordt onderzocht.

About the author

Martijn Wieling was born on March 18, 1981 in Emmen, the Netherlands. After graduating High School with an interest in both mathematics and language, he enrolled at the Computing Science programme of the University of Groningen. In 2005, he obtained his BSc degree with distinction and was admitted to the Research Master Behavioural and Cognitive Neurosciences. In 2007, he graduated with distinction for both MSc programmes, Computing Science and Behavioural and Cognitive Neurosciences. During his studies, Martijn was actively involved in teaching, PR activities, and he was elected for one year in the university council.

After his graduation and a one-year period of doing research in Educational Technology, Martijn started his PhD programme in Computational Linguistics (Dialectometry), under the supervision of John Nerbonne and later also Harald Baayen. During his PhD programme, he collaborated with researchers in Canada, Spain, Italy and England. Besides conducting research, Martijn taught a one-week course at Makerere University, Uganda and he was co-organizer of TABU Dag 2009, a two-day international linguistics conference in Groningen. Between 2005 and 2012, Martijn also coordinated the *Examentraining VWO*, a yearly event where hundreds of High School students are prepared for their national exams in three-day courses. In his spare time, Martijn enjoys playing squash, reading, and flying a microlight aircraft.

From September 2012 to August 2013, Martijn will collaborate with Harald Baayen in Tübingen, Germany on a project investigating language variation by means of analyzing tongue and lip movement during speech. This project is funded by a Rubicon grant of the Netherlands Organisation for Scientific Research (NWO).

Publications

Submitted

1. Martijn Wieling, Clive Upton and Ann Thompson. Analyzing the BBC Voices data: Contemporary English dialect areas and their characteristic lexical variants.

2. Hanneke Loerts, Martijn Wieling and Monika Schmid. Neuter is not common in Dutch: Eye movements reveal asymmetrical gender processing.

3. Martijn Wieling, Simonetta Montemagni, John Nerbonne and R. Harald Baayen. Lexical differences between Tuscan dialects and standard Italian: A sociolinguistic analysis using generalized additive mixed modeling.

4. Martijn Wieling, Esteve Valls, R. Harald Baayen and John Nerbonne. The effects of language policies on standardization of Catalan dialects: A sociolinguistic analysis using generalized additive mixed-effects regression modelling.

Published or accepted for publication

1. Martijn Wieling, Robert G. Shackleton, Jr. and John Nerbonne (accepted, 2012). Analyzing phonetic variation in the traditional English dialects: Simultaneously clustering dialects and phonetic features. *LLC. The Journal of Digital Scholarship in the Humanities*.

2. Simonetta Montemagni, Martijn Wieling, Bob de Jonge and John Nerbonne (accepted, 2012). Synchronic patterns of Tuscan phonetic variation and diachronic change: Evidence from a dialectometric study. *LLC. The Journal of Digital Scholarship in the Humanities*.

3. Esteve Valls, Martijn Wieling and John Nerbonne (accepted, 2012). Linguistic advergence and divergence in north-western Catalan: A dialectometric investigation of dialect leveling and border effects. *LLC. The Journal of Digital Scholarship in the Humanities*.

4. Simonetta Montemagni, Martijn Wieling, Bob de Jonge and John Nerbonne (accepted, 2012). Patterns of language variation and underlying linguistic features: a new dialectometric approach. In: Nicola de Blasi (ed.) *La variazione nell'italiano e nella sua storia. Varietà e varianti linguistiche e testuali. Proceedings of the Congresso della Società Internazionale di Linguistica e Filologia Italiana (XI Congresso SILFI)*.

5. John Nerbonne, Sandrien van Ommen, Charlotte Gooskens and Martijn Wieling (accepted, 2011). Measuring socially motivated pronunciation differences. In: Lars Borin and Anju Saxena (eds.) *Proceedings of the Gothenburg Workshop on comparing approaches to measuring linguistic differences*.

6. Esteve Valls, John Nerbonne, Jelena Prokić, Martijn Wieling, Esteve Clua, and Maria-Rosa Lloret (accepted, 2011). Applying Levenshtein dis-

tance to Catalan dialects. A brief comparison of two dialectometric approaches. *Verba. Anuario Galego de Filoloxía*.

7. Martijn Wieling, Eliza Margaretha and John Nerbonne (2012). Inducing a measure of phonetic similarity from pronunciation variation. *Journal of Phonetics*, 40(2), 307–314.

8. Martijn Wieling, Eliza Margaretha and John Nerbonne (2011). Inducing phonetic distances from dialect variation. *Computational Linguistics in the Netherlands Journal*, 1, 109–118.

9. Martijn Wieling, John Nerbonne and R. Harald Baayen (2011). Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PLoS ONE*, 6(9), e23613.

10. Martijn Wieling and John Nerbonne (2011). Measuring linguistic variation commensurably. In: John Nerbonne, Stefan Grondelaers, Dirk Speelman and Maria-Pilar Perea (eds.), *Dialectologia*, Special Issue II: Production, Perception and Attitude, 141–162.

11. Martijn Wieling and John Nerbonne (2011). Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features. *Computer Speech and Language*, 25(3), 700–715.

12. John Nerbonne, Jelena Prokić, Martijn Wieling and Charlotte Gooskens (2010). Some further dialectometrical steps. In: G. Aurrekoexea and J. L. Ormaetxea (eds.) *Tools for Linguistic Variation*. Bilbao: Supplements of the *Anuario de Filologia Vasca "Julio de Urquijo"*, LIII, pp. 41–56.

13. Martijn Wieling and John Nerbonne (2010). Hierarchical spectral partitioning of bipartite graphs to cluster dialects and identify distinguishing features. In: Carmen Banea, Alessandro Moschitti, Swapna Somasundaran and Fabio Massimo Zanzotto (eds.) *Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing*, ACL, Uppsala, Sweden, July 16, 2010, pp. 33–41.

14. Martijn Wieling and Adriaan Hofman (2010). The impact of online video lecture recordings and automated feedback on student performance. *Computers and Education*, 54(4), 992–998.

15. Kevin Williams, Justin Park and Martijn Wieling (2010). The face reveals athletic flair: Better National Football League quarterbacks are better looking. *Personality and Individual Differences*, 48, 112–116.

16. Wilbert Heeringa, Martijn Wieling, Boudewijn van den Berg and John Nerbonne (2009). A quantitative examination of variation in Dutch Low Saxon morphology. In: Alexandra Lenz, Charlotte Gooskens and

Siemon Reker (eds.) *Low Saxon Dialects across Borders - Niedersächsische Dialekte über Grenzen hinweg* (ZDL-Beiheft 138), Franz Steiner Verlag, 2009, pp. 195–216.

17. Martijn Wieling and John Nerbonne (2009). Bipartite spectral graph partitioning to co-cluster varieties and sound correspondences in dialectology. In: Monojit Choudhury, Samer Hassan, Animesh Mukherjee and Smaranda Muresan (eds.) *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-4)*, ACL-IJCNLP, Singapore, 7 August 2009, pp. 14–22.

18. Martijn Wieling, Jelena Prokić and John Nerbonne (2009). Evaluating the pairwise string alignment of pronunciations. In: Lars Borin and Piroska Lendvai (eds.) *Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH – SHELT&R 2009)*, Workshop at the 12th Meeting of the European Chapter of the Association for Computational Linguistics. Athens, 30 March 2009, pp. 26–34

19. Jelena Prokić, Martijn Wieling and John Nerbonne (2009). Multiple sequence alignments in linguistics. In: Lars Borin and Piroska Lendvai (eds.) *Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH - SHELT&R 2009)*, Workshop at the 12th Meeting of the European Chapter of the Association for Computational Linguistics. Athens, 30 March 2009, pp. 18–25

20. Justin Park, Martijn Wieling, Bram Buunk, and Karlijn Massar (2008). Sex-specific relationship between digit ratio (2D:4D) and romantic jealousy. *Personality and Individual Differences*, 44, 1039–1045.

21. Martijn Wieling and John Nerbonne (2007). Dialect pronunciation comparison and spoken word recognition. In: Petya Osenova (ed.) *Proceedings of the RANLP Workshop on Computational Phonology*, Borovetz, pp. 71–78.

22. Martijn Wieling, Therese Leinonen and John Nerbonne (2007). Inducing sound segment differences using Pair Hidden Markov Models. In: John Nerbonne, Mark Ellison and Greg Kondrak (eds.) *Computing and Historical Phonology, 9th Meeting of ACL Special Interest Group for Computational Morphology and Phonology Workshop*. Prague, pp. 48–56.

23. Martijn Wieling, Wilbert Heeringa and John Nerbonne (2007). An aggregate analysis of pronunciation in the Goeman-Taeldeman-van Reenen-Project data. *Taal en Tongval*, 59, 84–116.

24. Justin Park, Bram Buunk and Martijn Wieling (2007). Does the face reveal athletic flair? Positions in team sports and facial attractiveness. *Personality and Individual Differences*, 43, 1960–1965.

25. Martijn Wieling, Mark-Jan Nederhof and Gertjan van Noord (2006). Parsing partially bracketed input. In: Khalil Sima'an, Maarten de Rijke, Remko Scha and Rob van Son (eds.) *Proceedings of the Sixteenth Computational Linguistics in the Netherlands*. Amsterdam, pp. 1–16.

26. Nicolai Petkov and Martijn Wieling (2004). Gabor filtering augmented with surround inhibition for improved contour detection by texture suppression, *Perception*, 33 supplement, 68c.

Groningen Dissertations in Linguistics (GRODIL)

1. Henriëtte de Swart (1991). *Adverbs of Quantification: A Generalized Quantifier Approach.*
2. Eric Hoekstra (1991). *Licensing Conditions on Phrase Structure.*
3. Dicky Gilbers (1992). *Phonological Networks. A Theory of Segment Representation.*
4. Helen de Hoop (1992). *Case Configuration and Noun Phrase Interpretation.*
5. Gosse Bouma (1993). *Nonmonotonicity and Categorial Unification Grammar.*
6. Peter Blok (1993). *The Interpretation of Focus: An Epistemic Approach to Pragmatics.*
7. Roelien Bastiaanse (1993). *Studies in Aphasia.*
8. Bert Bos (1993). *Rapid User Interface Development with the Script Language Gist.*
9. Wim Kosmeijer (1993). *Barriers and Licensing.*
10. Jan-Wouter Zwart (1993). *Dutch Syntax: A Minimalist Approach.*
11. Mark Kas (1993). *Essays on Boolean Functions and Negative Polarity.*
12. Ton van der Wouden (1994). *Negative Contexts.*
13. Joop Houtman (1994). *Coordination and Constituency: A Study in Categorial Grammar.*
14. Petra Hendriks (1995). *Comparatives and Categorial Grammar.*
15. Maarten de Wind (1995). *Inversion in French.*
16. Jelly Julia de Jong (1996). *The Case of Bound Pronouns in Peripheral Romance.*
17. Sjoukje van der Wal (1996). *Negative Polarity Items and Negation: Tandem Acquisition.*
18. Anastasia Giannakidou (1997). *The Landscape of Polarity Items.*
19. Karen Lattewitz (1997). *Adjacency in Dutch and German.*
20. Edith Kaan (1997). *Processing Subject-Object Ambiguities in Dutch.*
21. Henny Klein (1997). *Adverbs of Degree in Dutch.*
22. Leonie Bosveld-de Smet (1998). *On Mass and Plural Quantification: The Case of French 'des'/'du'-NPs.*
23. Rita Landeweerd (1998). *Discourse Semantics of Perspective and Temporal Structure.*
24. Mettina Veenstra (1998). *Formalizing the Minimalist Program.*
25. Roel Jonkers (1998). *Comprehension and Production of Verbs in Aphasic Speakers.*
26. Erik F. Tjong Kim Sang (1998). *Machine Learning of Phonotactics.*
27. Paulien Rijkhoek (1998). *On Degree Phrases and Result Clauses.*
28. Jan de Jong (1999). *Specific Language Impairment in Dutch: Inflectional Morphology and Argument Structure.*

29. H. Wee (1999). *Definite Focus.*
30. Eun-Hee Lee (2000). *Dynamic and Stative Information in Temporal Reasoning: Korean Tense and Aspect in Discourse.*
31. Ivilin Stoianov (2001). *Connectionist Lexical Processing.*
32. Klarien van der Linde (2001). *Sonority Substitutions.*
33. Monique Lamers (2001). *Sentence Processing: Using Syntactic, Semantic, and Thematic Information.*
34. Shalom Zuckerman (2001). *The Acquisition of "Optional" Movement.*
35. Rob Koeling (2001). *Dialogue-Based Disambiguation: Using Dialogue Status to Improve Speech Understanding.*
36. Esther Ruigendijk (2002). *Case Assignment in Agrammatism: A Cross-Linguistic Study.*
37. Tony Mullen (2002). *An Investigation into Compositional Features and Feature Merging for Maximum Entropy-Based Parse Selection.*
38. Nanette Bienfait (2002). *Grammatica-onderwijs aan allochtone jongeren.*
39. Dirk-Bart den Ouden (2002). *Phonology in Aphasia: Syllables and Segments in Level-Specific Deficits.*
40. Rienk Withaar (2002). *The Role of the Phonological Loop in Sentence Comprehension.*
41. Kim Sauter (2002). *Transfer and Access to Universal Grammar in Adult Second Language Acquisition.*
42. Laura Sabourin (2003). *Grammatical Gender and Second Language Processing: An ERP Study.*
43. Hein van Schie (2003). *Visual Semantics.*
44. Lilia Schürcks-Grozeva (2003). *Binding and Bulgarian.*
45. Stasinos Konstantopoulos (2003). *Using ILP to Learn Local Linguistic Structures.*
46. Wilbert Heeringa (2004). *Measuring Dialect Pronunciation Differences using Levenshtein Distance.*
47. Wouter Jansen (2004). *Laryngeal Contrast and Phonetic Voicing: A Laboratory Phonology Approach to English, Hungarian and Dutch.*
48. Judith Rispens (2004). *Syntactic and Phonological Processing in Developmental Dyslexia.*
49. Danielle Bougaïré (2004). *L'approche communicative des campagnes de sensibilisation en santé publique au Burkina Faso: Les cas de la planification familiale, du sida et de l'excision.*
50. Tanja Gaustad (2004). *Linguistic Knowledge and Word Sense Disambiguation.*
51. Susanne Schoof (2004). *An HPSG Account of Nonfinite Verbal Complements in Latin.*
52. M. Begoña Villada Moirón (2005). *Data-Driven Identification of Fixed Expressions and their Modifiability.*
53. Robbert Prins (2005). *Finite-State Pre-Processing for Natural Language Analysis.*
54. Leonoor van der Beek (2005). *Topics in Corpus-Based Dutch Syntax.*
55. Keiko Yoshioka (2005). *Linguistic and Gestural Introduction and Tracking of Referents in L1 and L2 Discourse.*
56. Sible Andringa (2005). *Form-Focused Instruction and the Development of Second Language Proficiency.*
57. Joanneke Prenger (2005). *Taal telt! Een onderzoek naar de rol van taalvaardigheid en tekstbegrip in het realistisch wiskundeonderwijs.*

58. Neslihan Kansu-Yetkiner (2006). *Blood, Shame and Fear: Self-Presentation Strategies of Turkish Women's Talk about their Health and Sexuality.*
59. Mónika Z. Zempléni (2006). *Functional Imaging of the Hemispheric Contribution to Language Processing.*
60. Maartje Schreuder (2006). *Prosodic Processes in Language and Music.*
61. Hidetoshi Shiraishi (2006). *Topics in Nivkh Phonology.*
62. Tamás Biró (2006). *Finding the Right Words: Implementing Optimality Theory with Simulated Annealing.*
63. Dieuwke de Goede (2006). *Verbs in Spoken Sentence Processing: Unraveling the Activation Pattern of the Matrix Verb.*
64. Eleonora Rossi (2007). *Clitic Production in Italian Agrammatism.*
65. Holger Hopp (2007). *Ultimate Attainment at the Interfaces in Second Language Acquisition: Grammar and Processing.*
66. Gerlof Bouma (2008). *Starting a Sentence in Dutch: A Corpus Study of Subject- and Object-Fronting.*
67. Julia Klitsch (2008). *Open your Eyes and Listen Carefully. Auditory and Audiovisual Speech Perception and the McGurk Effect in Dutch Speakers with and without Aphasia.*
68. Janneke ter Beek (2008). *Restructuring and Infinitival Complements in Dutch.*
69. Jori Mur (2008). *Off-Line Answer Extraction for Question Answering.*
70. Lonneke van der Plas (2008). *Automatic Lexico-Semantic Acquisition for Question Answering.*
71. Arjen Versloot (2008). *Mechanisms of Language Change: Vowel Reduction in 15th Century West Frisian.*
72. Ismail Fahmi (2009). *Automatic Term and Relation Extraction for Medical Question Answering System.*
73. Tuba Yarbay Duman (2009). *Turkish Agrammatic Aphasia: Word Order, Time Reference and Case.*
74. Maria Trofimova (2009). *Case Assignment by Prepositions in Russian Aphasia.*
75. Rasmus Steinkrauss (2009). *Frequency and Function in WH Question Acquisition. A Usage-Based Case Study of German L1 Acquisition.*
76. Marjolein Deunk (2009). *Discourse Practices in Preschool. Young Children's Participation in Everyday Classroom Activities.*
77. Sake Jager (2009). *Towards ICT-Integrated Language Learning: Developing an Implementation Framework in terms of Pedagogy, Technology and Environment.*
78. Francisco Dellatorre Borges (2010). *Parse Selection with Support Vector Machines.*
79. Geoffrey Andogah (2010). *Geographically Constrained Information Retrieval.*
80. Jacqueline van Kruiningen (2010). *Onderwijsontwerp als conversatie. Probleemoplossing in interprofessioneel overleg.*
81. Robert G. Shackleton, Jr. (2010). *Quantitative Assessment of English-American Speech Relationships.*
82. Tim Van de Cruys (2010). *Mining for Meaning: The Extraction of Lexico-Semantic Knowledge from Text.*
83. Therese Leinonen (2010). *An Acoustic Analysis of Vowel Pronunciation in Swedish Dialects.*
84. Erik-Jan Smits (2010). *Acquiring Quantification. How Children Use Semantics and Pragmatics to Constrain Meaning.*

85. Tal Caspi (2010). *A Dynamic Perspective on Second Language Development.*
86. Teodora Mehotcheva (2010). *After The Fiesta is Over. Foreign Language Attrition of Spanish in Dutch and German Erasmus Students.*
87. Xiaoyan Xu (2010). *English Language Attrition and Retention in Chinese and Dutch University Students.*
88. Jelena Prokić (2010). *Families and Resemblances.*
89. Radek Šimík (2011). *Modal Existential WH-Constructions.*
90. Katrien Colman (2011). *Behavioral and Neuroimaging Studies on Language Processing in Dutch Speakers with Parkinson's Disease.*
91. Siti Mina Tamah (2011). *Student Interaction in the Implementation of the Jigsaw Technique in Language Teaching.*
92. Aletta Kwant (2011). *Geraakt door prentenboeken. Effecten van het gebruik van prentenboeken op de sociaal-emotionele ontwikkeling van kleuters.*
93. Marlies Kluck (2011). *Sentence Amalgamation.*
94. Anja Schüppert (2011). *Origin of Asymmetry: Mutual Intelligibility of Spoken Danish and Swedish.*
95. Peter Nabende (2011). *Applying Dynamic Bayesian Networks in Transliteration Detection and Generation.*
96. Barbara Plank (2011). *Domain Adaptation for Parsing.*
97. Çağrı Çöltekin (2011). *Catching Words in a Stream of Speech: Computational Simulations of Segmenting Transcribed Child-Directed Speech.*
98. Dörte Hessler (2011). *Audiovisual Processing in Aphasic and Non-Brain-Damaged Listeners: The Whole is More than the Sum of its Parts.*
99. Herman Heringa (2012). *Appositional Constructions.*
100. Diana Dimitrova (2012). *Neural Correlates of Prosody and Information Structure.*
101. Harwintha Anjarningsih (2012). *Time Reference in Standard Indonesian Agrammatic Aphasia.*
102. Myrte Gosen (2012). *Tracing Learning in Interaction. An Analysis of Shared Reading of Picture Books at Kindergarten.*
103. Martijn Wieling (2012). *A Quantitative Approach to Social and Geographical Dialect Variation.*

Grodil
Secretary of the Department of General Linguistics
Postbus 716
9700 AS Groningen
The Netherlands