Families and resemblances

Prokic, Jelena

Families and Resemblances

Jelena Prokić

rijksuniversiteit
groningen

The work presented here was carried out under the auspices of the Netherlands National Graduate School of Linguistics (LOT—Landelijke Onderzoekschool Taalwetenschap) and the Center for Language and Cognition Groningen of the Faculty of Arts of the University of Groningen.

The work was carried out within the project *Buldialect—Measuring Linguistic Unity and Diversity in Europe* financed by the Volkswagen Stiftung.

Groningen Dissertations in Linguistics 88
ISSN 0928-0030

Document prepared with LaTeX 2ε and typeset by pdfTeX
Cover design by Jelena Prokić & Jo-Ann Snel
Printed by Wöhrmann Print Service, Zutphen

RIJKSUNIVERSITEIT GRONINGEN

Families and Resemblances

Proefschrift

ter verkrijging van het doctoraat in de
Letteren
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. F. Zwarts,
in het openbaar te verdedigen op
maandag 29 november 2010
om 13.15 uur

door

Jelena Prokić
geboren op 26 augustus 1975
te Smederevo, Servië

Promotor:	Prof.dr.ir. J. Nerbonne
Copromotor:	Dr. H.P. Houtzagers
Beoordelingscommissie:	Prof.dr. E. Hinrichs Prof.dr. B. Joseph Prof.dr. D. Klein

ISBN: 978-90-367-4636-6

Acknowledgments

I am deeply grateful for the help I have received in writing this thesis. Help was direct as well as indirect.

First of all I would like to thank my promotor and supervisor John Nerbonne for all his advice and guidance during the past four years. I was really lucky to be able to learn from him, not only about dialectometry, but about academic life in general. I am especially grateful to him for the enormous energy he has invested in reading and correcting my thesis in the past few months. Without him, I wonder if I would have been able to finish the dissertation. I would also like to thank my co-promotor Peter Houtzagers for his help with all the matters related to the Bulgarian language and traditional dialectology.

This thesis has been written within the project *Buldialect—Measuring Unity and Diversity in Europe*, sponsored by the Volkswagen Foundation. I would like to thank Erhard Hinrichs, the principal investigator of the project, and John Nerbonne for inviting me to work in the project and for running it successfully in the last four years. I am also pleased to thank Petya Osenova and Vladimir Zhobov for their help with the *Buldialect* data set, Bulgarian language and dialectology. They were always fast and efficient in answering my numerous questions. I am also grateful to all other people who worked on this project for their advice and discussions that we had at our meetings: Georgi Kolev, Kiril Simov, Petar Shishkov, and Thomas Zastrow.

Erhard Hinrichs, Dan Klein and Brian Joseph agreed to be in my reading committee and to read my thesis. I would like to thank them, and especially Brain Joseph for his detailed comments that have improved the final version.

Groningen is one of the best places to work in the field of dialectometry and I am very happy to have had the chance to study there. I'd like to thank the many inspiring colleagues who made up the Groningen group, including: Charlotte Gooskens, Wilbert Heeringa, Sebastian Kürschner, Therese Leinonen, John Nerbonne, and Martijn Wieling. I am grateful to Therese for all our discussions and for reading my thesis. Therese's comments were always very valuable. I would also like to thank Martijn for our collaboration on two papers and numerous talks we had on dialectometry. Work with L04 program would not have been possible without Peter Kleiweg, who wrote the software, and helped me many times to run it successfully. I am very grateful for all his prompt

reactions.

Chapter 7 of this thesis is on an experiment I conducted while visiting the University of Auckland in New Zealand. I would like to thank Alexei Drummond and Russell Gray for being good hosts and giving me a lot of advice on how to set up and run the experiment.

The computational linguistics group at the University of Groningen, better known as Alfa-informatica, was a wonderful place to work at, and it was also a lot of fun. My thanks goes to all people who are, in a way, part of the group: Barbara, Begoña, Çağrı, Daniël, Dörte, Erik, Francisco, Geert, Geoffrey, George, Gerlof, Gertjan, Gideon, Giorgos, Gosse, Harm, Hartmut, Henny, Ismail, Jacky, Jelle, Jens, John, Jori, Jörg, Kostadin, Leonie, Lonneke, Martijn, Nynke, Peter M., Peter N., Proscovia, Sveta, Tim, Wilbert, Yan and all those that I might have forgotten. I would especially like to thank Barbara, Çağrı and Tim, who were my office mates for three years. Kostadin was always there for me to give me valuable help with the Bulgarian language and I want to thank him for being such a wonderful colleague. I want to express my gratitude to Wyke who was always there, not only for me, but for all the PhD students when we needed any kind of help, from paper work to a nice word. I also want to thank Adri for taking care of our computers and, especially, for long morning chats.

Special thanks goes to Barbara and Çağrı who are my paranimfs, and good friends. Thank you for the nice time we had in the past three years and for agreeing to stand next to me on the day of the defense.

Овом приликом желим посебно да захвалим Александру за подршку и разумевање током протеклих десет година.

My biggest support was Jelle, who shared a lot of good and bad moments with me, and helped me finish this thesis. Tige tank foar alles! Bedankt voor alles!

But most of all, I want to thank my family, the most essential part of my being. *Највећу захвалност дугујем мојој породици на безграничној љубави и разумевању.*

Munich, October 15, 2010

Contents

1 Introduction **1**

1.1 Background 1

1.2 The main research questions 5

2 Data **9**

2.1 Data collection 9

2.2 Selection of words and features 10

2.3 Traditional scholarship 16

3 Distance-based methods **23**

3.1 Levenshtein distance 23

3.2 Multidimensional scaling 24

3.3 Clustering 25

3.3.1 Hierarchical agglomerative clustering 26

3.3.2 K-means 28

3.4 Neighbor-joining and neighbor-net 29

3.4.1 Neighbor-joining 30

3.4.2 Neighbor-net 31

3.5 Visual inspection 35

3.5.1 MDS 35

3.5.2 Clustering 38

3.5.3 Neighbor-joining and neighbor-net 41

3.6 Evaluation of the results of distance-based methods 44

3.6.1 External validation 44

3.6.2 Internal validation 46

3.6.3 Results 47

3.7 Discussion and conclusions 50

4 Comparison to the traditional maps **55**
4.1 East-west division . . . 56
4.2 Western dialects . . . 57
4.2.1 Northwest-southwest split . . . 58
4.2.2 Transitional zone . . . 62
4.3 Eastern dialects . . . 63
4.3.1 Rupian dialects . . . 64
4.3.2 Moesian dialects . . . 65
4.4 Discussion . . . 67

5 Segment distances **71**
5.1 Pointwise mutual information . . . 72
5.2 Evaluation of the pairwise alignments . . . 75
5.2.1 Results . . . 77
5.3 Analysis of segment distances . . . 78
5.4 PMI and the aggregate analysis of dialects . . . 80
5.5 Discussion . . . 87

6 Multiple string alignments in linguistics **89**
6.1 Multiple sequence aligning . . . 89
6.1.1 Example of multiple sequence alignment . . . 91
6.2 Iterative pairwise alignment . . . 92
6.3 Gold standard and baseline . . . 94
6.3.1 Simple baseline . . . 94
6.3.2 Advanced baseline . . . 95
6.3.3 Gold standard . . . 95
6.4 Evaluation . . . 95
6.4.1 Order dependent method . . . 96
6.4.2 Modified Rand index . . . 98
6.4.3 Results . . . 99
6.5 Discussion . . . 101

7 Bayesian phylogenetic inference **103**
7.1 Phylogenetic inference . . . 103
7.2 Character-based methods . . . 105
7.3 Phylogenetic inference in linguistics . . . 106
7.4 Bayesian inference of phylogeny . . . 107
7.5 Experiment . . . 114
7.5.1 Different models of sound change . . . 117
7.6 Results . . . 120
7.7 Discussion . . . 134

8 Conclusions and discussion **139**

List of abbreviations **147**

Bibliography **149**

Short summary **157**

Samenvatting **159**

A List of words **161**

B List of phones **167**

C List of sites **169**

D Clustering results **173**

Chapter 1

Introduction

This book presents the results of four years of research on Bulgarian dialects and methods for dialectological analysis. It will present advances in techniques in several areas, namely application of clustering techniques in the detection of dialect groups, automatic extraction of phone distances using pointwise mutual information, improved pairwise alignment of word transcriptions obtained by employing automatically induced phone distances within the Levenshtein algorithm, multiple alignments of strings in linguistics, and application of methods taken from computational phylogenetics on dialect pronunciation data. It will also reexamine the geographic and historical organization of Bulgarian linguistic variation and suggest modifications in the traditional view. The rest of this chapter sketches the history of scholarship, first on diachronic linguistics, then on dialectology, with a particular focus on quantitative techniques.

1.1 Background

The question of how language has evolved has attracted the attention of scientists for the past few centuries, and the first speculations on the origin of language can be traced back 3.000 years (Crystal, 1987, 290). In linguistics, the first scientific attempts to discover the history of language started at the end of the 18th century, when Sir William Jones lectured on the resemblance between Sanskrit and ancient Greek and Latin. He suggested that all three languages have a common root, and that the common root can be the only explanation of the similarities among these languages. His lecture inspired the idea that language similarities such as those holding among Latin, Greek, and Sanskrit (etc.) could be due to common descent from a language no longer spoken and led to further inquiry by many prominent scholars who tried to compare different languages in a systematic way. This resulted in the development of the *comparative method*, a method for determining language relationships and the nature of the common source for related

languages that involves detailed feature-by-feature comparison of languages looking for recurring corresponding elements. One of the best-known scholars to use this method in order to prove the relatedness among the Indo-European languages was German linguist August Schleicher. He was the first one to illustrate the relatedness between languages using the figure of a tree. This representation of language relatedness suggests that the innovations occur in the process of transmission from a mother language to the daughter languages. In the late 19th century a group of German linguists, known as the *Neogrammarians*, proposed the hypothesis of the regularity of sound change. According to *the Neogrammarian hypothesis* sound change occurs regularly and uniformly whenever the appropriate phonetic environment is encountered (Campbell, 2004). Ever since then the understanding of sound change has played a major role in the comparative method. The method proceeds from the simultaneous comparison of different languages, i.e. lists of cognate terms from the related languages. The method has also been used, with great success, in regard to morphology, syntax, semantics, poetics, even cultural constructs like legal systems (Joseph, 2004). A few years after the *Neogrammarian hypothesis* was proposed, a student of Schleicher, Johannes Schmidt, proposed the so-called *wave theory* of language development, according to which new features of a language are spread from the center to the neighboring languages similar to the waves in continuously weakening concentric circles. Unlike in the competing *tree theory* the innovations in languages spread through borrowing. The *wave theory* was also directed against the *Neogrammarian hypothesis* of a sound change.

Quantitative methods were first introduced into comparative linguistics with the work of Alfred Kroeber and Charles Chrétien (Kroeber and Chrétien, 1939), although work of American linguist Morris Swadesh in 1950s received much more attention in linguistic circles. He suggested an approach in comparative linguistics called *lexicostatistics* that is based on the quantitative comparison of the cognates. In this approach the similarity between two languages is the proportion of the cognates from a fixed list of cognates, the so-called *Swadesh list*, that two languages share. Swadesh also suggested an approach in historical linguistics called *glottochronology* that can be used to calculate the divergence times of languages. It is based on the assumption that the basic vocabulary in every language is replaced at a steady rate. By counting the number of words that have been replaced from the basic vocabulary, we can estimate the time when two languages diverged from a common proto-language. This approach to historical linguistics has been heavily criticized mostly because of its assumption that the vocabulary changes at a constant rate. For a detailed discussion see for example Campbell (2004, 201-210).

Observations about dialect variation were recorded already by the ancient Greeks who had verbs that meant ‘to speak in a particular dialect way’, for instance, as well as by the ancient Indians, who had remarks in early Sanskrit texts about what happens when one uses forms other than those that the Brahmins use. However, a more scientific approach came only in the 19th century as a response to the advances in the research on the history of languages and particularly the *Neogrammarian hypothesis* that claimed

that sound change is regular (Chambers and Trudgill, 2007). Interest in a systematic approach to dialectology was the hope that apparent anomalies in language history might be explained once geographic conditioning was investigated and understood. The first systematic study of dialects started with the work of German linguist Georg Wenker. In 1876 he began collecting dialect data from the northern Germany. He collected around 45.000 questionnaires and made maps that were published as the first dialect atlas *Sprachatlas des Deutschen Reichs*. The results of Wenker's project, contrary to the primary expectations, has shown that sound changes are much more irregular than suggested by the *Neogrammarians*. Following this project, similar projects for many languages in Europe and Northern America were established: in 1898 for Danish, in 1896 for French, in 1930 *The Linguistic Atlas of the United States and Canada* for English (Chambers and Trudgill, 2007, 16-17). Traditional dialectology made great use of the isogloss: a line drawn between two regions that have different realizations of a certain feature. If there are many isoglosses that coincide, they form an isogloss bundle, which is an indication of a major dialect division. Many maps found in traditional dialect atlases are based only on one feature that is indicative of a certain dialect variation, but groups of similar division were always sought.

Introduction of the quantitative methods in dialectology came in 1971 with the work of French linguist Jean Séguy, who developed the first technique for measuring the distances between the dialects (Séguy, 1971). This branch of dialectology became known as *dialectometry*. Séguy aggregated over the individual differences between sites by counting the overlapping features between any two sites. In this way he introduced an aggregate view of language variation, as opposed to the traditional division of sites based on the individual linguistic features. Further improvement in the development of dialectometry came with the work of Hans Goebl (Goebl, 1982; Goebl, 1984), who also introduced a weighting of the features. He was also the first one to use clustering techniques in dialectometry. Brett Kessler (Kessler, 1995) was the first to use Levenshtein distance in order to calculate the pronunciation distance between the Irish Gaelic dialects. Levensthein distance was later successfully applied to many other languages: Dutch (Nerbonne et al., 1996; Heeringa, 2004), Sardinian (Bolognesi and Heeringa, 2002), Norwegian (Gooskens and Heeringa, 2004), German (Nerbonne and Siedle, 2005), American English (Nerbonne, 2005), and Bulgarian (Osenova, Heeringa, and Nerbonne, 2009). This thesis attempts to contribute to this line of work in Chapters 3-5.

In the past ten years there has been an increasing interest in the application of the methods taken from computational phylogenetics to the study of language history and change. Phylogenetics is a branch of biology that studies the evolutionary relatedness among various groups of organisms, especially among entire species. In the past few decades it has been a very active field of research, which has led to the development of many new methods that enable us to have better insight into the evolution and relatedness of species. In linguistics, these methods have been used to address the problems of the origins of Indo-European (Gray and Jordan, 2000) and Bantu languages (Holden, 2002;

Holden and Gray, 2006). They were also applied to the problems of the subgrouping of Indo-European (Ringe, Warnow, and Taylor, 2002; Nakhleh, Ringe, and Warnow, 2005), as well as to test various hypotheses about human prehistory (Dunn et al., 2005; Greenhill and Gray, 2005; Gray, Drummond, and Greenhill, 2009).

In dialectology, there are a few studies that apply methods taken from computational phylogenetics. In Hamed (2005) and Hamed and Wang (2006) phylogenetic techniques were exploited in research on Chinese dialects, while McMahon et al. (2007) used them to explore the phonetic similarity between English varieties. All these works address the old problem of branching vs. wave-like diffusion by testing their data with the help of the programs developed for inferring phylogenetic networks. This thesis attempts to contribute to the phylogenetic research on language history in Chapter 7.

In this thesis we apply and develop various quantitative methods to the Bulgarian phonetic dialect data. Bulgarian dialectology scholarship has a very old tradition that dates back to 1848 (Grigorovich, 1848). The most significant period of the development of the dialectology in Bulgaria came in 1950s and is related to work of Prof. Stoyko Stoykov. His study of Bulgarian dialects is the most widely known and the most authoritative until today (Stoykov, 2002). We use his classification of Bulgarian dialects in order to evaluate our computational methods and to compare the traditional and the quantitative approach to dialect diversity. All our experiments are done on the data set which contains most of the features that Stoykov uses as basis for his phonetically-based division of Bulgarian dialect area, which allows us to directly compare our computational methods to the traditional scholarship (see Chapter 4).

We analyze Bulgarian data taking two alternative approaches. One approach is based on the similarity among the varieties with the focus on geographic organization of Bulgarian dialects. We use the Levenshtein algorithm to aggregate over the numerous features found in the data and infer the similarities/distances among the groups of dialects. We also test an alternative approach to dialect variation that is more historically motivated. We employ methods taken from phylogenetics that focus on systematic shared innovations as a signal of common ancestry and reexamine the relatedness among the Bulgarian dialect varieties. The results of applying different quantitative techniques on the Bulgarian dialect data have shown that some of the traditional divisions of this area have to be questioned if only pronunciation data is taken into account. We do not examine other linguistic levels, nor do we attend to non-linguistic influences. The comparison of the divisions resulting from the geographic and historical approaches has shown that these two different perspectives gave very similar picture of the Bulgarian dialect variation.

Apart from reexamining Bulgarian dialect variation using new techniques, we also try to improve methods for dialectological research. We present advances in several techniques, related both to the Levenshtein approach to dialect variation and to the application of phylogenetic methods in linguistics as well. Although all experiments are performed on the Bulgarian data, none of the methods are language specific, nor are they

applicable only to the dialect data. The fact that we have tested our methods against a very well studied dialect area, helped us evaluate better our computational methods and improve them. However, these methods can be used to examine relationships between any language families by exploiting resemblances that they share. In this sense, language family is seen as a group of varieties that are related by the features that they have in common. While some methods in this thesis treat shared features as a sign of a common origin of the varieties, some others are based on the counting of the overlapping features regardless of the genetic relationship. Methods presented can help us split varieties into smaller groups, but also look into the mechanisms of language change. They investigate different aspects of language families and their resemblances.

In the next section we present the outline of the thesis and develop the main research questions addressed.

1.2 The main research questions

This thesis was written as a part of the project *Buldialect—Measuring Linguistic Unity and Diversity in Europe*. It was a joint project between the University of Tübingen, the University of Groningen and the Institute of Parallel Processing at the Bulgarian Academy of Sciences. The project was sponsored by the Volkswagen Stiftung, as part of the funding initiative *Unity and Diversity in Europe* . The aim of the *Buldialect* project was to develop machine-readable data on Bulgarian dialects and to analyze it using the methods from computational dialectometry in order to get better insight into the cultural unity and diversity of this region. The data was collected and digitalized in Sofia as a cooperation between Petya Osenova and Kiril Simov from the Bulgarian Academy of Sciences and Prof. Vladimir Zhobov from the University of Sofia. It consists of both phonetic and lexical data, although in this thesis we base all our experiments solely on the phonetic data. The data set used in this thesis is presented in Chapter 2.

In Chapters 3, 4 and 5 we rely on the Levenshtein distance to quantify the differences between the dialect varieties. In Chapter 3 we look into the problem of using clustering methods in order to detect dialect groups. In too many previous studies in dialectometry the common practice was to try as many clustering algorithms as possible and later pick the one whose results coincide the most with the traditional dialect division of the area or were attractive for other reasons. The comparison of the clustering results and the traditional maps was usually done by simply visually inspecting the similarities and the differences between the two. However, the aim of the research done in dialectometry is not to replicate the traditional dialect maps, but to quantify large amounts of data and to characterize general tendencies in linguistic variation that are missing in the traditional feature-by-feature approaches.

In this chapter we try to answer the following questions:

- Which exact methods can we use to compare the divisions done by traditional

dialectologists and computational methods?

- Is clustering, i.e. automatic determination of groups, an appropriate technique for the investigation of the dialect data that is, in most of the cases, continuous data? If so, which clustering techniques are most reliable?

- Can development of dialects better be described using the *tree* model or *wave* model of change? Which methods taken from computational phylogenetics can help us address this problem?

In Chapter 4 we compare the traditional and computational classifications on a level of very fine detail that proceeds from the aggregate varietal distances down to the specific segments in the words. We examine how different phonetic features are projected in the traditional and the computational divisions of the dialects. By examining the differences between the two classification in this manner we are hoping to answer the following questions:

- Does our data set contain the same features that traditional dialectologists have used to classify Bulgarian dialects?

- Are the distances obtained using the Levenshtein method, with our specific settings, capturing dialect diversity insightfully?

- Do clustering techniques identify the significant groups in the data?

- Are all the dialect groupings proposed by traditional linguists based on purely linguistic data? Or are they perhaps based on other criteria?

In Chapter 5 we apply a technique called pointwise mutual information (PMI) to automatically infer the distances between the phones in the data set. In many studies, including Chapter 3 of this thesis, the Levenshtein algorithm is used only with the constraint that vowels and consonants cannot be aligned. In that setting, all vowels are equally distant from each other. The same holds for the consonants. We employ the distances between the segments obtained using the PMI technique within the Levenshtein algorithm hoping to improve on the alignments produced by the Levenshtein algorithm and to get a better measure of the distances between the language varieties. We address the following questions in this chapter:

- Can we improve the quality of the alignments by using the PMI inferred segment distances with the Levenshtein algorithm?

- Are any phonetic (articulatory) features reflected in the PMI induced phone distances?

- Are there any improvements on the aggregate level of the dialect divisions if we incorporate PMI induced distances in our analysis?

In the Levenshtein approach all word transcriptions are pairwise aligned, compared to each other and the distances between each two strings are turned into a single number. In Chapters 6 and 7 we take a different approach to string alignment and to the data analysis. It is an alternative, historically motivated, approach that proceeds from the assumption that all our examined varieties are genetically related and share common ancestry. We adopt the methods from computational phylogenetics that can simultaneously perform the analysis on all transcriptions for a given word. First we multi-align all the transcriptions to get the desired format for our data. We do so by adopting an algorithm specifically designed to multi-align strings in linguistics. We present it in Chapter 6 and evaluate the quality of the produced alignments using two novel techniques. In Chapter 7 we analyze automatically multi-aligned phonetic transcriptions using a Bayesian inference method. Unlike in the earlier approaches, this technique enables us to test various hypotheses about the evolution of sounds and the evolution of dialects. In this chapter we address the following questions:

- Can we directly use phonetic segments as a basis for Bayesian phylogenetic inference? What are the problems?
- Which models developed for the evolution of species can be applied to the phonetic data?
- Are phones equally likely to change into any other phone?
- Do phones in some word positions change more frequently than in some other?

In the last chapter we summarize the results and provide a discussion on the solutions to the questions addressed in this thesis.

Chapter 2

Data

The data used in this thesis is part of the project *Buldialect—Measuring linguistic unity and diversity in Europe*.[1] The data set developed during this project consists of the pronunciations of 157 words collected at 197 places equally distributed all over Bulgaria (Figure 2.1).[2] The data was collected and digitalized as a joint work between the University of Sofia and the Institute for Parallel Processing, Bulgarian Academy of Sciences. The main source of the data was the large dialect archive at the University of Sofia. The word pronunciations that are part of this archive started to be gathered in 1950s, and this work continues till now. During the *Buldialect* project part of this data was selected and converted into X-SAMPA encoding for further computer processing and into IPA encoding for human usage. For some missing concepts and/or sites, additional expeditions were organized as a part of the project.

In this chapter we give the description of the data set, with special emphasis on the data collection and the selection of words. We also provide the extensive list of phonetic features present in the data set, since feature distribution can significantly influence the results obtained in the analyses performed. More detailed description of the data set can be found in Prokić et al. (2009). Parts of this chapter were published as Prokić et al. (2009) and Houtzagers, Nerbonne, and Prokić (2010).

2.1 Data collection

The dialect archive at the University of Sofia contains pronunciation data from various sources. They include supervised students' theses, published monographs, dictionaries,

[1] The project is sponsored by Volkswagen Stiftung. More information about the project can be found at http://www.sfs.uni-tuebingen.de/dialectometry.

[2] For the word живели /ʒiˈveli/ 'live - past 3rd pl' pronunciations from many villages were not recorded. For that reason, we do not use it in our experiments and work with 156 words at most.

and the archive of the *Ideographic Dictionary of Bulgarian Dialects*. The largest source for the pronunciation data are theses written by graduate students of Bulgarian language at the University of Sofia. The collection of these descriptions began at the end of 1950s and intensified significantly in the following decades. The majority of the theses used for the pronunciation data were written in the period 1960–1985, very few of them earlier or later.

Published dialect descriptions and dictionaries are another important source. There are two series of such publications. *Българска диалектология. Проучвания и материали* [Bulgarian Dialectology. Investigations and Data] is non-periodical collection of papers published by the Publishing House of Bulgarian Academy of Science in the period 1962-1981 (10 volumes). *Трудове по българска диалектология* [Studies in Bulgarian Dialectology] is a collection of monographs published by the Publishing House of Bulgarian Academy of Science in the period 1965-1979 (10 volumes). Some standalone books were also used as a source for the dialect pronunciation data.

Part of the material comes from the archive of the *Ideographic Dictionary of Bulgarian Dialects*. This project was launched by Prof. Stoyko Stoykov in the middle of the 1950s. The material for the dictionary was collected from all possible sources: theses and term papers written on the bases of a questionnaire composed by Stoyko Stoykov (Stoykov, 1954); abundant material from field work expeditions, which were regularly organized in the summers; all published dialect descriptions and dictionaries; and the personal archives of other scholars.

Tape recordings of dialect speech are another important source. A collection of phono-archives started in 1981. Till now there are over 250 hours of recorded dialect speech from around 100 villages from all parts of the Bulgarian language territory.

The basic methods for the collection of dialect material were the observation of natural dialect speech and some work with questionnaires. Direct questioning was greatly disfavored, and in some cases even prohibited. The informants were selected among the oldest inhabitants of the village who were born locally. Preference was given to women because they were socially and otherwise less mobile at the time. The conversations were centered on traditional rural life — customs, religious practices, agricultural work, surrounding nature.

2.2 Selection of words and features

For the *Buldialect* project pronunciations of 157 words from 197 sites were selected from the *Archive* and further processed. The first criterion for word selection was the words' availability. The words included are frequent words that were collected from all, or almost all of the 197 sites. In Figure 2.1 we present the distribution of all the sites present in the *Buldialect* project. The sites are more or less evenly distributed throughout the country, with the exception of the northeastern part where the concentration of the sites is much smaller. For villages in this area no data was available in the *Archive*.

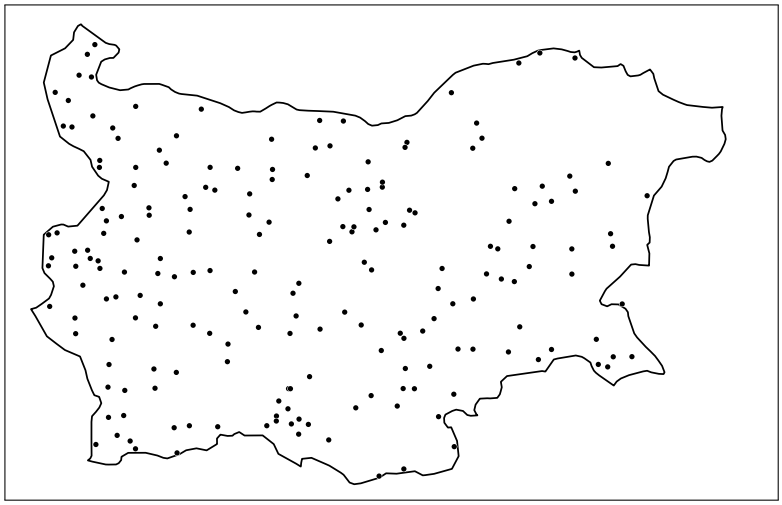

Figure 2.1: Distribution of 197 sites from the data set. Concentration of the sites is much smaller in the northeast than in the rest of the country.

During the data collection for the *Archive*, Prof. Stoykov included only villages that were dialectologically homogeneous. For example, villages with mixed Turkish-Bulgarian, or predominantly Turkish population were excluded.

Regarding the choice of words in *Buldialect* project, only words which are expected to show some degree of phonetic variation were included. Another important criterion for word selection was the balance between various phonetic features present in the data set. For example, the reflexes of Old Bulgarian vowels are represented with the same or nearly the same number of words. The complete list of words can be found in Appendix A. In total, there are 39 different dialectal features which have been represented in the chosen 157 words. Below is a list of the underlying linguistic features described in Prokić et al. (2009) and Houtzagers, Nerbonne, and Prokić (2010). With each feature we also provide a list of words in which the feature is present.

1. **Reflexes of** *yat*: In traditional dialectology, this is the most important dialect border in Bulgaria that divides the country into west and east. It represents different reflexes of the Old Bulgarian vowel **ě* (*yat*). In the west it is always pronounced as [e], while in the east it is pronounced either as [a], [æ], or [ɛ]. For more detailed explanation on the reflexes of *yat* see Section 2.3.[3]

[3]Throughout this thesis we use Cyrillic script to represent words in their standard orthography and phonemic transcriptions to refer to their pronunciation in Standard Bulgarian. Pronunciations of the words in various dialects are represented with phonetic transcriptions. The examples in the list (1)-(39) are presented in

Example: [xlʲap] vs. [xlep] vs. [xlɛp] ‘bread’

Words in the data set: бели /ˈbeli/ ‘white - pl.’, беше /ˈbeʃe/ ‘be - past 2nd sg, 3rd sg’, бяхме /ˈbʲaxme/ ‘be - past 1st pl’, вежда /ˈveʒda/ ‘eyebrow’, видях /viˈdʲax/ ‘see - aorist 1st sg.’, време /ˈvreme/ ‘time’, вътре /ˈvɤtre/ ‘inside’, вятър /ˈvʲatɤr/ ‘wind’, две /dve/ ‘two’, дете /deˈte/ ‘child’, добре /doˈbre/ ‘well’, горе /ˈgore/ ‘up’, желязо /ʒeˈlʲazo/ ‘iron’, живели /ʒiˈveli/ ‘live - past pl’, звезда /zvezˈda/ ‘star’, къде /kɤˈde/ ‘where’, месец /ˈmesɛ͡ts/ ‘month’, млякото /ˈmlʲakoto/ ‘the milk’, неделя /neˈdelʲa/ ‘Sunday’, нещо /ˈneʃto/ ‘something’, няма /ˈnʲama/ ‘there is no’, онези /oˈnezi/ ‘those’, орех /ˈorex/ ‘walnut’, петел /peˈtel/ ‘rooster’, пясък /ˈpʲasɤk/ ‘sand’, понеделник /poneˈdelnik/ ‘Monday’, река /reˈka/ ‘river’, ръце /rɤˈtse/ ‘hand - pl’, средата /sreˈdata/ ‘the middle’, сряда /ˈsrʲada/ ‘Wednesday’, трева /treˈva/ ‘grass’, утре /ˈutre/ ‘tomorrow’ хляб /xlʲab/ ‘bread’, цял /ʦʲal/ ‘whole’, череша /ʧeˈreʃa/ ‘cherry’, човек /ʧoˈvek/ ‘human’

2. **Etymological** *ja*: The term etymological *ja* refers to the vowel [a] preceded by the palatal approximant [j] or a post-alveolar consonant.

Example: [jaˈdeʃ] vs. [eˈdeʃ] ‘eat-you’

Words in the data set: аз /az/ ‘I’, агне /ˈagne/ ‘lamb’, нея /ˈneja/ ‘she - accusative’, овчар /ovˈʧar/ ‘shepherd’, овчари /ovˈʧari/ ‘shepherd - pl’, чакат /ˈʧakat/ ‘wait - 3rd pl’, ябълка /ˈjabɤlka/ ‘apple’, ябълки /ˈjabɤlki/ ‘apple - pl’, яйца /jajˈʦa/ ‘egg - pl’, яйце /jajˈʦe/ ‘egg’, ям /jam/ ‘eat - 1st sg’, ядеш /jaˈdeʃ/ ‘eat - 2nd sg’

3. **Presence or absence of initial prothetic [j]**

Example: [ˈagne] vs. [ˈjagne] ‘lamb’

Words in the data set: аз /az/ ‘I’, агне /ˈagne/ ‘lamb’, един /eˈdin/ ‘one - masc’ едно /eˈdno/ ‘one - neut’, език /eˈzik/ ‘tongue’, ечемик /eʧeˈmik/ ‘barley’, утре /ˈutre/ ‘tomorrow’, ябълка /ˈjabɤlka/ ‘apple’, ябълки /ˈjabɤlki/ ‘apple - pl’, яйца /jajˈʦa/ ‘egg - pl’, яйце /jajˈʦe/ ‘egg’

4. **Presence or absence of [j] before front vowels**

Example: [koˈe] vs. [koˈje] ‘which’

Words in the data set: кое /koˈe/ ‘which’

5. **Elision or no elision of [j]**

Example: [ˈneja] vs. [ˈnea] ‘she - accusative’

Words in the data set: майка /ˈmajka/ ‘mother’, нея /ˈneja/ ‘she - accusative’, ябълка /ˈjabɤlka/ ‘apple’, ябълки /ˈjabɤlki/ ‘apple - pl’, яйца /jajˈʦa/ ‘egg - pl’, яйце /jajˈʦe/ ‘egg’, ям /jam/ ‘eat - 1st sg’, ядеш /jaˈdeʃ/ ‘eat - 2nd sg’

6. **Reflexes of the back nasalized vowel**

Example: [kaˈde] vs. [kuˈde] ‘where’

Words in the data set: берат /beˈrɤt/ ‘pick up - 3rd pl’, вътре /ˈvɤtre/ ‘inside’, дера /deˈrɤ/ ‘flay - 1st sg’, къде /kɤˈde/ ‘where’, мъж /mɤʒ/ ‘man’, мъже /mɤˈʒe/ ‘men’, мъжът /mɤˈʒɤt/ ‘the man’, носят /ˈnosʲɤt/ ‘carry - 3rd pl’, пека /peˈkɤ/ ‘bake - 1st sg’, път /pɤt/ ‘road’, ръце /rɤˈʦe/ ‘hand - pl’, седя /seˈdʲɤ/ ‘sit - 1st sg’, събота

phonetic transcription so that e.g. final devoicing, which is quite common, is ignored.

/ˈsɤbota/ ‘Saturday’, чакат /ˈt͡ʃakat/ ‘wait - 3rd pl’, чета /t͡ʃeˈtɤ/ ‘read - 1st sg’, вода /voˈda/ ‘water’,[4] глава /glaˈva/ ‘head’, жена /ʒeˈna/ ‘woman’, звезда /zvezˈda/ ‘star’, земя /zeˈmʲa/ ‘Earth’, река /reˈka/ ‘river’, ръка /rɤˈka/ ‘hand’, сестра /sesˈtra/ ‘sister’, трева /treˈva/ ‘grass’, чешма /t͡ʃeʃˈma/ ‘fountain’

7. **Reflexes of the front nasalized vowel**

Example: [zet] vs. [zʲɔt] vs. [zʲɤt] vs. [zit] vs. [zent] ‘son-in-low, brother-in-low’

Words in the data set: агне /ˈagne/ ‘lamb’, време /ˈvreme/ ‘time’, говедо /go͡ˈvedo/ ‘beef’, дете /deˈte/ ‘child’, десет /ˈdeset/ ‘ten’, език /eˈzik/ ’tongue’, ечемик /et͡ʃeˈmik/ ‘barley’, жътва /ˈʒɤtva/ ‘harvest’, зет /zet/ ‘son-in-low, brother-in-low’, име /ˈime/ ‘name’, леща /ˈleʃta/ ‘lentil - pl’, месец /ˈmeseʦ/ ‘month’, месо /meˈso/ ‘meat’, петък /ˈpetɤk/ ‘Friday’, се /se/ ‘one’s self’

8. **Reflexes of the back** *yer*

Example: [taˈkof] vs. [taˈkɤf] vs. [taˈkaf] vs. [taˈkɔf] vs. [taˈkɛf] ‘such’

Words in the data set: във /vɤv/ ‘in’, вънка /ˈvɤnka/ ‘outside’, градът /graˈdɤt/ ‘the town’, дъжд /dɤʒd/ ‘rain’, дъно /ˈdɤno/ ‘bottom’, мъжът /mɤˈʒɤt/ ‘the man’, петък /ˈpetɤk/ ‘Friday’, първият /ˈpɤrvijɤt/ ‘the first’, пясък /ˈpʲasɤk/ ‘sand’, със /sɤs/ ‘with’, такъв /taˈkɤv/ ‘such’

9. **Reflexes of the front** *yer*

Example: [ˈtɤŋko] vs. [ˈteŋko] vs. [ˈtʲɔŋko] vs. [ˈtɛŋko] ‘thin - neut’

Words in the data set: гладен /ˈgladen/ ‘hungry’, ден /den/ ‘day’, днес /dnes/ ‘today’, дошъл /doˈʃɤl/ ‘come - aor part’, един /eˈdin/ ‘one - masc’, лесно /ˈlesno/ ‘easily’, петел /peˈtel/ ‘rooster’, сега /seˈga/ ‘now’, старец /ˈstareʦ/ ‘old man’, тъмно /ˈtɤmno/ ‘dark -neut’, тънко /ˈtɤnko/ ‘thin - neut’

10. **Choice of the vowel inserted between the two last consonants in words** *вятър* **‘wind’ and** *огън* **‘fire’**: The elision of the word-final, and therefore weak, *yer* likely resulted in an inadmissible syllabic structure, more specifically, in a syllable-final combination of obstruent and sonorant, and a vowel was inserted between the two consonants. The vowel inserted is often specific for this word alone.

Example: [ˈvʲatɤr] vs. [ˈveter] ‘wind’

Words in the data set: вятър /ˈvʲatɤr/ ‘wind’, огън /ˈogɤn/ ‘fire’

11. **Vowel reduction**

Example: [ˈpepel] vs. [ˈpepil] vs. [ˈpepʲɤl] ‘ash’

Words in the data set: вечер /ˈvet͡ʃer/ ‘evening’, пепел /ˈpepel/ ‘ash’, понеделник /poneˈdelnik/ ‘Monday’

12. **Reflexes of** *yery*

Example: [eˈzik] vs. [eˈzɨk] ‘tongue’

Words in the data set: език /eˈzik/ ‘tongue’, сирене /ˈsirene/ ‘cheese’

13. **Rounding of front vowels**

Example: [ʒif] vs. [ʒyf] vs. [ʒuf] ‘alive’

[4] In those East Bulgarian dialects where the general singular form of feminine nouns derives from the accusative. Also for all the examples till the end of point 6.

Words in the data set: ечемик /et͡ʃemˈik/ ‘barley’, желязо /ʒeˈlʲazo/ ‘iron’, жив /ʒiv/ ‘alive’, живели /ʒiˈveli/ ‘live - past 3rd pl’ име /ˈime/ ‘name’, череша /t͡ʃeˈreʃa/ ‘cherry’, чешма /t͡ʃeʃˈma/ ‘fountain’

14. **Unrounding of front vowels**

Example: [klʲut͡ʃ] vs. [klit͡ʃ] ‘key’

Words in the data set: ключ /klʲut͡ʃ/ ‘key’

15. **Alternation /o/-/e/**

Example: [d͡ʒop] vs. [d͡ʒep] ‘pocket’

Words in the data set: джоб /d͡ʒob/ ‘pocket’, наше /ˈnaʃe/ ‘ours’, пепел /ˈpepel/ ‘ash’

16. **Presence or absence of vowel elision**

Example: [neˈdelʲa] vs. [ˈndelʲa] ‘Sunday’

Words in the data set: ечемик /et͡ʃeˈmik/ ‘barley’, млякото /ˈmlʲakoto/ ‘the milk’, неделя /neˈdelʲa/ ‘Sunday’, овца /ovˈt͡sa/ ‘sheep’, овце /ovˈt͡se/ ‘sheep - pl’, понеделник /poneˈdelnik/ ‘Monday’, събота /ˈsɤbota/ ‘Saturday’, това /toˈva/ ‘this - neut’

17. **Change by analogy, like [ˈdolu] vs. [ˈdole] ‘down’, presumeably due to analogy with [ˈgore] ‘up’**

Example: [ˈdolu] vs. [ˈdole] ‘down’ analogy with [ˈgore] ‘up’

Words in the data set: долу /ˈdolu/ ‘down’, пека /peˈkɤ/ ‘bake - 1st sg’ (due to analogy with сека /seˈkɤ/ ‘chop - 1st sg’)

18. **Reflexes of syllabic liquids**

Example: [vɤlk] vs. [vlɤk] vs. [vl̩k] vs. [vɤk] vs. [vuk] vs. [vɔlk] vs. [vɛlk] ‘wolf’

Words in the data set: бързо /ˈbɤrzo/ ‘quickly’, връх /vrɤx/ ‘peak’, връщам /ˈvrɤʃtam/ ‘give back - 1st sg’, вълк /vɤlk/ ‘woolf’, вълна /ˈvɤlna/ ‘wool’, дълбок /dɤlˈbok/ ‘deep’, дърво /dɤrˈvo/ ‘tree’, жълт /ʒɤlt/ ‘yellow’, кръв /krɤv/ ‘blood’, пръч /prɤt͡ʃ/ ‘he-goat’, пръвият /ˈprɤvijɤt/ ‘the first’, сърп /sɤrp/ ‘sickle’, червен /t͡ʃerˈven/ ‘red’, черен /ˈt͡ʃeren/ ‘black’, ябълка /ˈjabɤlka/ ‘apple’, ябълки /ˈjabɤlki/ ‘apple - pl’

19. **Reflexes of *tj, *dj**

Example: [ˈleʃta] vs. [ˈleʃt͡ʃa] vs. [ˈlet͡ʃa] ‘lentils’

Words in the data set: вежда /ˈveʒda/ ‘eyebrow’, връщам /ˈvrɤʃtam/ ‘give back - 1st sg’, леща /ˈleʃta/ ‘lentil - pl’, неще /neˈʃte/ ‘not want - 3rd sg’, нощ /noʃt/ ‘night’, плащам /ˈplaʃtam/ ‘pay - 1st sg’, ще /ʃte/ ‘will’

20. **Variation of the original initial cluster чр + following vowel ъ or ѣ**

Example: [t͡ʃerˈven] vs. [t͡sɤrˈven] ‘red’

Words in the data set: червен /t͡ʃerˈven/ ‘red’, черен /ˈt͡ʃeren/ ‘black’, череша /t͡ʃeˈreʃa/ ‘cherry’

21. **Epenthetic [l]**

Example: [zeˈmʲa] vs. [zemˈlʲa] vs. [zemˈnʲa] ‘land’

Words in the data set: земя /zeˈmʲa/ ‘land’

22. **Presence or absence of voiced affricates**

Example: [d͡ʒop] vs. [ʒop] ‘pocket’

Words in the data set: джоб /d͡ʒob/ ‘pocket’, желязо /ʒeˈlʲazo/ ‘iron’, звезда /zvezˈda/ ‘star’

23. **Presence or absence of palatalized consonants**

Example: [vɫk] vs. [vɤlʲk] ‘wolf’

Words in the data set: агне /ˈagne/ ‘lamb’, бране /braˈne/ ‘pick - verb. noun’, вълк /vɤlk/ ‘wolf’, кон /kon/ ‘horse’, майка /ˈmajka/ ‘mother’, носят /ˈnosʲɤt/ ‘carry - 3rd pl’, огън /ˈogɤn/ ‘fire’, понеделник /poneˈdelnik/ ‘Sunday’, път /pɤt/ ‘road’, седя /seˈdʲɤ/ ‘sit - 1st sg’, сирене /ˈsirene/ ‘cheese’, сол /sol/ ‘salt’, фурна /ˈfurna/ ‘oven’, ябълка /ˈjabɤlka/ ‘apple’, ябълки /ˈjabɤlki/ ‘apple - pl’

24. **Results of palatalization of /st/, /zd/ in words corresponding to Standard Bulgarian ['gosti] guests, ['grozde] grapes**

Example: [ˈgosti] vs. [ˈgosʲe] vs. [ˈgojse] ‘guest - pl’

Words in the data set: гости /ˈgosti/ ‘guest - pl’, грозде /ˈgrozde/ ‘grapes’

25. **Presence or absence of simplification of the clusters стр /str/, здр /zdr/**

Example: [seˈstra] vs. [seˈsra] ‘sister’

Words in the data set: здрав /zdrav/ ‘healthy’, сестра /sesˈtra/ ‘sister’, страх /strax/ ‘fear’

26. **Presence or absence of epenthesis of [t] and [d] in the clusters [sr] and [zr]**

Example: [ˈsrʲada] vs. [ˈstrʲada] ‘Wednesday’

Words in the data set: сряда /ˈsrʲada/ ‘Wednesday’

27. **Presence or absence of the voiceless velar fricative**

Example: [strax] vs. [stra] ‘fear’

Words in the data set: бяхме /ˈbʲaxme/ ‘were - 1st pl’, видях /viˈdʲax/ ‘see - aor 1st sg’, връх /vrɤx/ ‘peak’, дадоха /ˈdadoxa/ ‘give - aor 3rd pl’, орех /ˈorex/ ‘walnut’, страх /strax/ ‘fear’, сух /sux/ ‘dry’, ухо /uˈxo/ ‘ear’, хляб /xlʲab/ ‘bread’, хоро /xoˈro/ ‘chain dance’, хубав /ˈxubav/ ‘beautiful - masc’, хубаво /ˈxubavo/ ‘beautiful - neut’

28. **Presence or absence of the voiceless labiodental fricative**

Example: [ˈfurna] vs. [ˈvurna] vs. [ˈxurna] vs. [ˈhurna] vs. [ˈɸurna] ‘oven’

Words in the data set: фурна /ˈfurna/ ‘oven’

29. **Preservation or loss of */v/ before rounded vowels**

Example: [vol] vs. [ol] ‘ox’

Words in the data set: вол /vol/ ‘ox’, двор /dvor/ ‘yard’, дърво /dɤrˈvo/ ‘tree’, твой /tvoj/ ‘yours’, хубаво /ˈxubavo/ ‘pretty - neut’

30. **Presence or absence of prothetic [v] before rounded vowels**

Example: [ˈogɤn] vs. [ˈvogɤn] ‘fire’

Words in the data set: огън /ˈogɤn/ ‘fire’, орех /ˈorex/ ‘walnut’

31. **Devoicing of obstruents in certain positions**

Example: [ʒif] vs. [ʒiv] ‘alive’

Words in the data set: джоб /d͡ʒob/ ‘pocket’, дъжд /dɤʒd/ ‘rain’, жив /ʒiv/ ‘alive’, здрав /zdrav/ ‘healthy’, кръв /krɤv/ ‘blood’, мъж /mɤʒ/ ‘man’, овца /ovˈt͡sa/ ‘sheep’,

овце /ovˈt͡se/ ‘sheep - pl’, овчар /ovˈt͡ʃar/ ‘shepherd’, овчари /ovˈt͡ʃari/ ‘shepherd - pl’, такъв /taˈkɤv/ ‘such’, хляб /xlʲab/ ‘bread’, хубав /ˈxubav/ ‘pretty - masc’

32. **The form of the preposition *ъ and the prefix *ь**

Example: [ˈvlizam] vs. [uˈlizam] ‘enter - 1st sg’

Words in the data set: влизам /ˈvlizam/ ‘to enter - 1st sg’, във /vɤv/ ‘in’

33. **Various assimilations and dissimilations**

Example: [ofˈt͡sa] vs. [osˈt͡sa] ‘sheep’

Words in the data set: едно /eˈdno/ ‘one - neut’, много /ˈmnogo/ ‘much, many’, овца /ovˈt͡sa/ ‘sheep’, овце /ovˈt͡se/ ‘sheep - pl’, овчар /ovˈt͡ʃar/ ‘shepherd’, овчари /ovˈt͡ʃari/ ‘shepherd - pl’, тъмно /ˈtɤmno/ ‘dark -neut’

34. **Nonsystematic changes in individual words**

Example: [ˈbɤrzo] vs. [ˈbɤrʒe] ‘quickly’

Words in the data set: бързо /ˈbɤrzo/ ‘quickly’, вече /ˈvet͡ʃe/ ‘already’, вчера /ˈvt͡ʃera/ ‘yesterday’, човек /t͡ʃoˈvek/ ‘person’

35. **Morphophonemic alternations or suffixes connected with the formation of secondary imperfective verbs**

Example: [ˈvlizam] vs. [ˈvlazam] vs. [ˈvlʲavam] ‘enter - 1st sg’

Words in the data set: влизам /ˈvlizam/ ‘enter - 1st sg’, връщам /ˈvrɤʃtam/ ’give back - 1st sg’, плащам /ˈplaʃtam/ ’pay - 1st sg’

36. **Form of certain grammatical endings, such as that of the first person plural in all tenses**

Example: [ˈbʲaxme] vs. [ˈbexmo] ‘were - 1st pl’

Words in the data set: бяхме /ˈbʲaxme/ ‘were - 1st pl’

37. **Choice of the suffix in certain nouns that originally belonged to the n-stem nouns**:

Example: [ˈkamɤk] vs. [ˈkamik] vs. [ˈkamen] ‘stone’

Words in the data set: ечемик /et͡ʃeˈmik/ ‘barley’, камък /ˈkamɤk/ ‘stone’

38. **Various forms of words that are derived from a common Old Bulgarian form**

Example: [ˈvie] vs. [vi] vs. [ve] ‘you’

Words in the data set: вие /ˈvie/ ‘you’, и /i/ ‘she - dative’, им /im/ ‘they - dative’, ние /ˈnie/ ‘we’, онези /oˈnezi/ ‘those’, това /toˈva/ ‘this - neut’, тогава /toˈgava/ ‘then’, я /ja/ ‘she - accusative’

39. **Different position of stress**

Example: [ˈvino] vs. [viˈno] ‘wine’

2.3 Traditional scholarship

In this section we give a short overview of the main dialect areas distinguished by traditional Bulgarian dialectology.

As found in Boyadzhiev (2004), the development of modern Bulgarian dialectology started in 1848 when Russian Slavist Viktor Grigorovich published a book *Очерк путешествия по Европейской Турции* [A Sketch of a Journey in European Turkey] (Grigorovich, 1848) in which, for the first time, he proposed division of the Bulgarian dialect area into west and east, describing at the same time linguistic features responsible for this division. After the liberation of Bulgaria from the Ottoman Empire in 1878, the interest in Bulgarian dialects increased, which resulted in numerous studies of the various individual dialects. The most significant period in the development of Bulgarian dialectology came after World War II and is related to the work of Prof. Stoyko Stoykov. Prof. Stoykov, who was the head of the Bulgarian dialectology section within the Institute for Bulgarian Language and the leading expert in Bulgarian dialectology, organized field expeditions, and set the foundations for *Bulgarian dialect atlas* (Stoykov and Bernstein, 1964; Stoykov, 1966; Stoykov et al., 1974; Stoykov, Kochev, and Mladenov, 1981). Led by Prof. Stoykov, Bulgarian dialectologists compiled reference books, atlases, dictionaries, monograph descriptions of individual dialects, as well as analytic surveys on a different topics from dialectology (Alexander, 2004). Stoykov's basic assumptions were that a dialect is a self-contained linguistic system and that a satisfactory dialect description should provide a thorough account of all levels of this system, contrary to the practice of collecting and describing only exotic and rare words and features (Prokić et al., 2009). On the basis of Prof. Stoykov's work, Bulgarian dialectology continues to develop till present times.

In *Българска диалектология* [Bulgarian dialectology] (Stoykov, 2002), Stoykov described the main dialect areas in Bulgaria (Figure 2.2). This division was based on the variation of different phonetic features and no lexical or syntactic variation was taken into account. According to Stoykov, the main division of Bulgarian dialects is into western and eastern. The border between these two areas is the so-called *yat* border that reflects different pronunciations of the Old Bulgarian vowel *yat*. It goes from Nikopol in the north, near Pleven and Teteven down to Petrich in the south, represented by the bold dashed line in Figure 2.2. This is the oldest dialect border that is still very well preserved. In a nonpalatal environment, i.e. before a syllable that does not contain post-alveolar consonant, palatalized consonant or a front vowel, in the west the Old Bulgarian vowel **ě* (*yat*) is always pronounced as [e], while in the east it is pronounced either as [a] or a low variant of [e]. If the reflex of *yat* is [a] or a very low variant of [e], a preceding consonant is usually palatalized. For example [bel] vs. [b^{j}al], [b^{j}æl] or [bɛl]. This isogloss divides Bulgarian language area into west and east. According to Stoykov (2002), east of the *yat* line there is a division into northeastern and southeastern areas based on the pronunciation of the old vowel *yat* in a palatal environment, i.e. if there is a post-alveolar consonant, palatalized consonant or a front vowel in the following syllable. In the northeast *yat* is pronounced as [e], while in the southeast it is pronounced as [a], [æ] or [ɛ]. For example [beli] vs. [b^{j}ali], [b^{j}æli] or [bɛli].

Taking into account various phonetic features, including reflexes of **ě* (*yat*) as well,

Stoykov divides the Bulgarian dialect area first into two zones—eastern and western along the *yat* line. These two areas are further divided into six dialect zones, which can also be seen on the map in Figure 2.2. In the east, there are Moesian, Balkan and Rupian dialects. In the west, he distinguishes southwestern, northwestern dialects and the transitional zone at the border with Serbia.

Moesian dialects are situated in the northeastern part of Bulgaria. According to Stoykov (2002, 101-103) the most important phonetic and morphophonetic characteristics of this dialect are the following:

- In stressed syllables, the reflexes of Old Bulgarian vowel **ě* (*yat*) before non-palatal syllables is [ʲa] and before palatal syllables is [ɛ] ([bʲal] vs. [bɛli]). Under the influence of the Balkan dialects [ɛ] is almost completely replaced by [e].
- velarized realization of the Old Bulgarian back *yer* in a stressed position
- non-existence of consonants /f/ and /x/
- change of consonant /d/ into [n] before /n/ (**dn* > [nn])
- the masculine definite article is /o/ (stressed) and /u/ (unstressed) instead of formal Bulgarian /ɤt/ and /ɤ/
- ending /e/ instead of formal Bulgarian /i/ for multi-syllable masculine nouns
- ending /e/ in stressed syllables instead of formal Bulgarian /i/ for plural past active aorist participles

Balkan dialects cover the central area of present Bulgaria and represent the most extensive group of dialects of the Bulgarian language. The main characteristics of the Balkan dialects are the following (Stoykov, 2002, 107):

- the reflexes of Old Bulgarian vowel **ѣ* (*yat*) before non-palatal syllable is [ʲa] and before palatal syllable is [e] ([bʲal] vs. [beli])
- reductions of vowels /a/, /e/ and /o/, which are usually reduced to [ə], [i] and [u] respectively
- realization of /a/ is [e] after a soft consonant or /ʒ/, /ʃ/, /t͡ʃ/, /d͡ʒ/, and before a soft syllable

Rupian dialects are found in the southeastern part of Bulgaria, and include the southern part of Trakia, the region of Haskovo, the Rodopes and the most southeastern region of Bulgaria around Malko Tarnovo. Rupian dialects comprise varieties that are heterogeneous and have vastly different phonetic characteristics. However, according to Stoykov (2002, 120-122) the following characteristics are present in all Rupian dialects:

- large number of palatal consonants in various positions
- soft pronunciation of consonants /ʒ/, /ʃ/ and /t͡ʃ/
- preserved consonant /x/ in all positions
- widespread labialization of /i/ into /u/
- change of consonant /d/ into [n] before /n/ (*dn > [nn])

Northwestern dialects are situated in the area between the border with Serbia in the west and the *yat* border on the east, and between *Stara planina* mountain in the south and the river Danube in the north. The phonological characteristics of this group of dialects are the following (Stoykov, 2002, 146):

- the reflex of Old Bulgarian vowel *$\check{e}$ (*yat*) is always [e]
- the reflex of old back nasal vowel *yus* and back vowel *yer* is [ɤ]
- the reflexes of old groups *t^j and *d^j are /ʃt/ and /ʒd/
- ending /e/ instead of formal Bulgarian /i/ for plural past active aorist participles
- the masculine definite article is /ə/ in a stressed syllable and /a/ in an unstressed syllable

Southwestern dialects are situated west of the *yat* line, occupying the territory that lies between Rupian and Balkan dialects in the east, northwestern dialects in the north and transitional dialects at the border with Serbia on the west. The main characteristics of these dialects are the following (Stoykov, 2002, 149):

- the reflex of Old Bulgarian vowel *$\check{e}$ (*yat*) is always [e]
- the reflex of Old Bulgarian back nasal vowel *yus* is in most cases [a]
  The exception is Sofia area where the reflex [ə] is found.
- the reflex of Old Bulgarian back *yer* ъ and front *yer* ь is mostly [a], but in the western parts reflex [o] is found instead of [a]
- the reflexes of old groups *t^j and *d^j are /ʃt/ and /ʒd/
- change of /o/ into [e] after /ʒ/, /ʃ/, /t͡ʃ/ and /j/
- single masculine definite article is /o/ or /a/

Transitional dialects lie at both sides of today's Bulgarian-Serbian border. In this thesis we are interested only in the varieties that are within the Bulgarian administrative border. At the Bulgarian side, these dialects occupy very small area near the border and represent a transition between Serbian and Bulgarian language varieties. They are characterized by the following features (Stoykov, 2002, 164-165):

- the reflex of Old Bulgarian vowel **ě* (*yat*) is always [e]
- the reflexes of old groups **tʲ* and **dʲ* are /t͡ʃ/ and /d͡ʒ/
- the reflex of Old Bulgarian back nasal vowel *yus* is [u]
- the reflex of Old Bulgarian back and front *yer* is always [ə]
- articulation of voiced consonants at the end of the word (as in Serbian)
- softer [l] than in other Bulgarian dialects, but not palatalized
- complete loss of consonant /f/ in all positions—in new words it is replaced with /v/
- complete loss of consonant /x/ in all positions
- frequent usage of palatalized /n/ and /l/ in word final position and before front vowels /e/ and /i/

In the following chapters of this thesis we apply various quantitative methods on the dialect pronunciation data from the *Buldialect* project in order to automatically detect main dialect groups and calculate the distances between them. In Chapter 4 we compare in detail the results of the computational analysis to the traditional divisions of Bulgarian dialects. The aim of the comparison is to evaluate our computational methods but also to check the distribution of the phonetic features responsible for traditional divisions within the *Buldialect* data set. The results will show that the features responsible for the traditional dialect divisions, according to Stoykov (2002), are well represented in our data set.

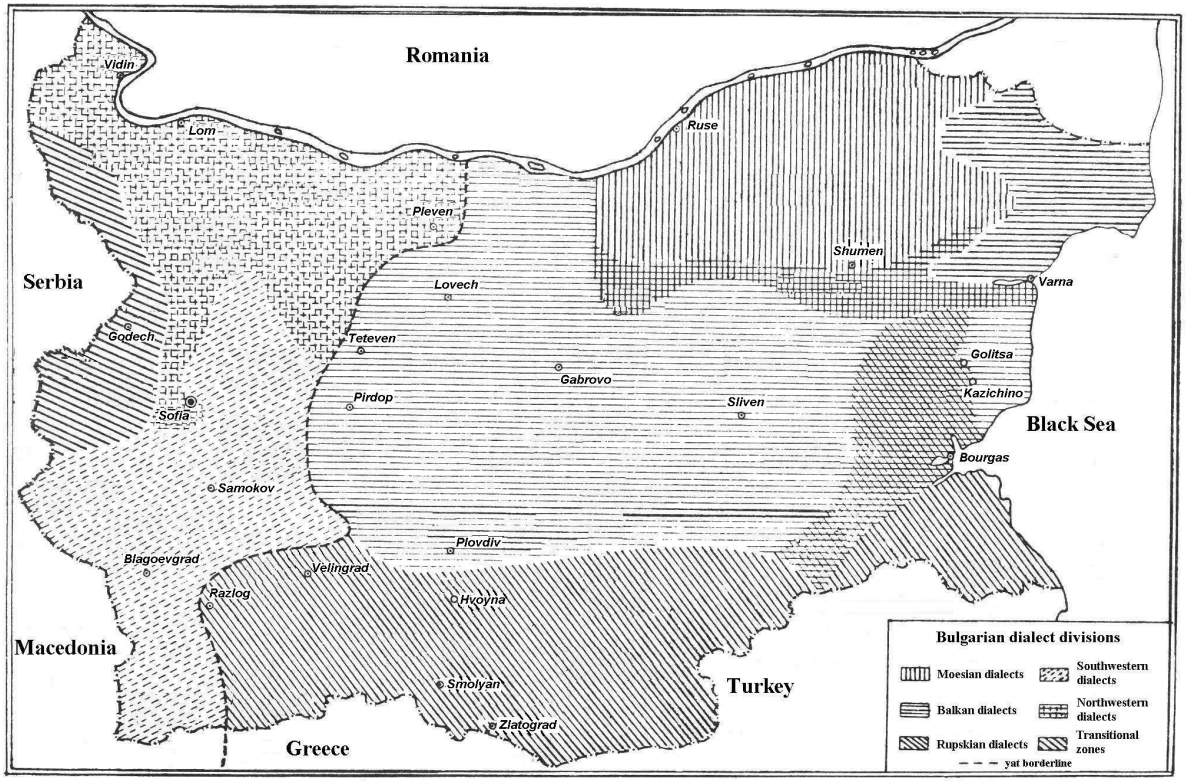

Figure 2.2: Traditional dialect division according to Stoykov (2002). The black dashed line is the *yat* line that divides the country into east and west. Each of these areas is further divided into three dialect zones: northwestern, southwestern dialects and transitional zone at the border with Serbia west of the *yat* line, and Balkan, Rupian and Moesian dialects east of the *yat* line.

Chapter 3

Distance-based methods

In this chapter we present a group of methods that all proceed from a distance matrix that stores information on the distances between each two sites in the data set and try to group those sites based on different criteria depending on a method. The distances between the sites are calculated using the Levenshtein algorithm that at the same time pairwise aligns all corresponding word pronunciations in the data set. This algorithm is described in more detail in Section 3.1. The so-called distance-based methods used to analyze the distance matrix include multidimensional scaling, seven hierarchical clustering algorithms, k-means clustering algorithm, neighbor-joining and neighbor-net, described in Sections 3.2, 3.3 and 3.4. We also present the results of applying these techniques in Section 3.5. In Section 3.6 we propose various evaluation techniques and use them to evaluate the performance of the mentioned classification algorithms. We conclude this chapter with the discussion and conclusions in Section 3.7. Work presented in this chapter was published as Prokić and Nerbonne (2008).

3.1 Levenshtein distance

The Levenshtein, or string edit distance, algorithm (Levenshtein, 1966) is a dynamic programming algorithm used to measure the differences between two strings. The distance between two strings is the smallest number of insertions, deletions, and substitutions needed to transform one string to the other. For example, in order to transform one word transcription in Figure 3.1 into the other we would need 3 operations: [bʲ] has to be replaced by [b], [ə] by [e] and [i] by [e]. In this chapter all three operations were assigned the same value, namely 1. This means that the distance between two strings in Figure 3.1 is 3. Every sequence, i.e. word transcription is represented as a sequence of phones which are not further defined. As a consequence, pair [bʲ]-[b] counts as different

to the same degree as pair [i]-[e].[1]

bʲ	ə	r	ˈɑ	n	i
b	e	r	ˈɑ	n	e
1	1				1

Figure 3.1: Levenshtein distance between these two strings is 3.

In his thesis Heeringa (2004) has shown that in the aggregate analysis of dialect differences, more detailed feature representation of segments does not improve the results obtained by using simple phone representation. Another motivation for using the simple phone representation is to keep the analysis as robust as possible, without going into the language specific details of feature representation. In Chapter 5 we present a method, called pointwise mutual information, that can be used to automatically acquire the distances between the segments in the transcriptions. We incorporate this method into the Levenshtein procedure and obtain alignments that are of a better quality than those obtained by the simple Levenshtein algorithm. This results in a slight improvement of the results in the aggregate analysis of dialect differences.

The Levenshtein algorithm is also directly used to align two sequences, as can be seen in Figure 3.1. The transcriptions used for experiments in this thesis were aligned based on the following principles: a) a vowel can be aligned only with a vowel b) a consonant can be aligned with a consonant, a sonorant or a semivowel such as [j] and [w]. After aligning all word transcriptions, which also results in a calculation of the distances between each two strings, we calculate the distances between the sites. The distance between two sites is the mean of all word distances calculated for those two sites. The final result is a distance matrix that contains the distances between each two sites in the data set. We note that using the mean Levenshtein distance over a large sample of pronunciations effectively aggregates over a large number of individual segment differences, the basis of most isoglosses. Brett Kessler (Kessler, 1995) was the first to use Levenshtein distance in order to calculate the linguistic distance between the dialects. Later it was successfully applied to many other languages. An overview of the application of the Levenshtein algorithm in dialectology can be found in Nerbonne (2009).

3.2 Multidimensional scaling

Multidimensional scaling is a dimension-reducing method used in exploratory data analysis and a data visualization method, often used to look for separation of data clusters (Legendre and Legendre, 1998). The goal of the analysis is to detect meaningful underlying dimensions that allow the researcher to explain observed similarities or dissim-

[1] For technical reasons, the sign for primary stress is moved to the first vowel in stressed syllable.

ilarities between the investigated objects. It displays the structure of distance-like data as a geometrical picture by attempting to arrange 'objects' in a space within a certain small number of dimensions, which, however, accord with the observed distances. As a result, dissimilar objects are plotted far apart from each other, while similar objects are close to one another. This enables us to 'explain' the distances in terms of underlying dimensions. It has been used in linguistics and dialectology since Black (1973). In this thesis Kruskal's non-metric MDS is being used.

3.3 Clustering

Cluster analysis is the process of partitioning a set of objects into groups or clusters (Manning and Schütze, 1999). The goal of clustering is to find the structures in the data by finding objects that are similar enough to be put in the same group and by identifying distinctions between the groups. The data in each subset share some common trait—often proximity according to some defined distance measure. Clustering methods can be classified into several types, based on different criteria. One classification is into *soft* or *hard*, where in soft clustering objects can belong to a cluster to a certain degree. In hard clustering objects can be assigned only to one cluster. Another division of clustering algorithms is into *hierarchical* and *partitional* clustering. Partitional clustering algorithms produce mutually exclusive partitions of the data where each instance can belong only to one cluster. Unlike in hierarchical clustering, all clusters are determined in one step. Hierarchical clustering is usually hard, while partitioning clustering can be both hard and soft. Hierarchical clustering algorithms produce a set of nested partitions of the data by finding successive clusters using previously established clusters. This kind of hierarchy is represented with a dendrogram—a tree in which more similar elements are grouped together. Hierarchical clustering algorithms can be further divided into agglomerative (bottom-up) and divisive (top-down) clustering. In agglomerative clustering, the procedure begins by putting each object in a separate cluster, and later successively grouping them into larger and larger clusters until a single cluster is obtained. In divisive clustering, the procedure goes in the opposite direction: at the beginning of the procedure all objects are put into one cluster and later successively divided into smaller and smaller subclusters.

In this thesis one partitional and seven hierarchical agglomerative clustering algorithms will be examined in more detail. Since traditional scholarship agrees that, in general, dialect areas are organized hierarchically, and since, in particular, Bulgarian dialect areas are claimed to be hierarchically organized (see Chapter 2), we are particularly interested in the hierarchical techniques. We examine the partitioning techniques to see whether they might aid in detecting groups at any level of hierarchy.

3.3.1 Hierarchical agglomerative clustering

In this section we present seven hierarchical clustering algorithms whose performance on the dialect pronunciation data is examined in this thesis. Hierarchical clustering algorithms can be described by the following scheme formalized by Johnson (1967):

- estimate pairwise distances
- put information on distances into matrix
- find the shortest distance in the matrix
- fuse two closest points
- calculate the distance between the newly formed node and the rest of the nodes (matrix updating algorithms)
- repeat until there are no more nodes to be fused

Based on the way in which the distances between a newly formed node and the rest of the nodes are calculated, there are seven different algorithms (Jain and Dubes, 1988) and they will be described in more details.

Single link method, also known as nearest neighbor, is one of the oldest methods in cluster analysis. The similarity between two clusters is computed as the distance between the two most similar objects in the two clusters.

$$d_{k[ij]} = minimum(d_{ki}, d_{kj}) \tag{3.1}$$

In this formula, as well as in other formulae in this subsection, i and j are two closest points that are fused into one cluster$[i, j]$, and k represents all the remaining points (clusters). In single link clustering the similarity function is locally defined, resulting in clusters of good local coherence, but bad global quality (Manning and Schütze, 1999). As noted in Jain and Dubes (1988), single link clusters easily chain together, yielding a so-called *chaining effect*, and produce elongated clusters. The presence of only one intermediate object between two compact clusters is enough to turn them into a single cluster. For that reason, this method is sensitive to noise, and as we shall see later, not suitable for dialectometric analysis.

Complete link method, also called furthest neighbor, uses the most distant pair of objects while fusing two clusters. The algorithm first compares all existing clusters searching for the most distant pairs of objects belonging to two different clusters. In the second step it merges two clusters that have the smallest value for the most distant objects found in the first step. In that way, an object joins a cluster only when it is linked to all the objects already members of the cluster (Legendre and Legendre, 1998).

$$d_{k[ij]} = maximum(d_{ki}, d_{kj}) \tag{3.2}$$

Unlike the single link method, this method produces sphere-like clusters that have good global quality. In ecology, complete link clustering is often used in order to delineate clusters with clear discontinuities (Legendre and Legendre, 1998).

Unweighted Pair Group Method using Arithmetic averages (UPGMA) belongs to a group of average clustering methods, together with three methods that will be described below. In UPGMA, the distance between any two clusters is the average of distances between all members of the two clusters being compared. All objects, i.e. single elements, receive the same weight in the computation regardless of the number of objects in the cluster.

$$d_{k[ij]} = (n_i/(n_i+n_j)) \times d_{ki} + (n_j/(n_i+n_j)) \times d_{kj} \tag{3.3}$$

As a consequence, the clusters themselves are weighted according to the number of elements that belong to them, i.e. clusters with the smaller number of elements will be weighted less and the other way around.

Weighted Pair Group Method using Arithmetic averages (WPGMA), the same as UPGMA, calculates the distance between the two clusters as the average of distances between all members of two clusters. But in WPGMA, the clusters that fuse receive equal weight regardless of the number of members in each cluster.

$$d_{k[ij]} = (\frac{1}{2} \times d_{ki}) + (\frac{1}{2} \times d_{kj}) \tag{3.4}$$

Because all clusters receive equal weight, objects in smaller clusters are more heavily weighted than those in the big clusters. This modification of the UPGMA algorithm was proposed by Sokal and Michener (1958) since sometimes UPGMA results can be distorted during the fusion of a large group of objects with the small group of objects.

Unweighted Pair Group Method using Centroids (UPGMC) In this method, the members of a cluster are represented by their mean point, called centroid. This centroid represents the cluster while calculating the distance between the clusters to be fused.

$$d_{k[ij]} = (n_i/(n_i+n_j)) \times d_{ki} + (n_j/(n_i+n_j)) \times d_{kj} - ((n_i \times n_j)/(n_i+n_j)^2) \times d_{ij} \tag{3.5}$$

In the unweighted version of centroid clustering the clusters are weighted based on the number of elements that belong to that cluster. This means that bigger clusters receive higher weight, and sometimes centroids can be biased towards bigger clusters.

Centroid clustering methods can occasionally produce reversals–partitions where the

distance between two clusters is smaller than the distance between the subclusters in one of the two clusters (Legendre and Legendre, 1998). These dendrograms are hard to draw and interpret and for that reason often not used by researchers.

Weighted Pair Group Method using Centroids (WPGMC) Somewhat as in WPGMA, in WPGMC all clusters are assigned the same weight regardless of the number of objects in each cluster. In that way the centroids are not biased towards well-sampled clusters.

$$d_{k[ij]} = (\frac{1}{2} \times d_{ki}) + (\frac{1}{2} \times d_{kj}) - (\frac{1}{4} \times d_{ij}) \tag{3.6}$$

Ward's method This method is also known as the minimal variance method. At each stage in the analysis clusters that merge are those that result in the smallest increase in the sum of the squared distances of each individual from the mean of its cluster.

$$\begin{aligned} d_{k[ij]} &= ((n_k + n_i)/(n_k + n_i + n_j)) \times d_{ki} \\ &\quad +((n_k + n_j)/(n_k + n_i + n_j)) \times d_{kj} \\ &\quad -(n_k/(n_k + n_i + n_j)) \times d_{ij} \end{aligned} \tag{3.7}$$

This method uses an analysis of variance approach to calculate the distances between clusters. One of the main drawbacks of this method is that it tends to create clusters of the same size (Legendre and Legendre, 1998).

3.3.2 K-means

The k-means algorithm belongs to the non-hierarchical algorithms which are often referred to as *partitional* clustering methods (Jain and Dubes, 1988). Unlike hierarchical clustering algorithms, partitional clustering methods generate a single partition of the data. A partition implies a division of the data in such a way that each instance can belong only to one cluster. The number of groups in which the data should be partitioned is usually determined by the user.

The k-means is the most commonly used partitional algorithm, which despite its simplicity, works sufficiently well in many applications (Manning and Schütze, 1999). The main idea of k-clustering is to find the partition of n objects into k clusters such that each object is assigned to the cluster with the nearest mean. In other words, given a set of objects $X = \{x_1, x_2, ..., x_n\}$, *k-means* tries to put n objects into k groups such that the total error sum of squares is minimized:

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x_j \in S_i} ||x_j - \mu_i||^2 \tag{3.8}$$

where μ_i is the mean of S_i. In this chapter we use squared Euclidean distance to compute centroid clusters: each centroid is the mean of the points in that cluster.

In the simplest version, the algorithm consists of the following steps:

1. pick at random initial cluster centers
2. assign objects to the cluster whose mean is closest
3. recompute the means of clusters
4. reassign every object to the cluster whose mean is closest
5. repeat steps 3 and 4 until there are no changes in the cluster membership of any object

Eventually, the algorithm converges finding the best solution given the starting centroids. This means that the final solution depends on the initial position of the centroids and that the algorithm is guaranteed to find only the 'local minimum', but not necessarily the 'overall minimum' as well. This is considered one of the main drawbacks of the *k-means* algorithm. In order to overcome this problem, different solutions have been suggested throughout vast literature on partitional clustering. Here we list some of the possibilities:

- In the rare cases where it is possible, start with the centroids placed in positions already close to the final solution.
- Repeat the whole procedure several times starting every time from a different random configuration. Take the solution that minimizes the most sum of square errors.
- Use the output of some of the hierarchical clustering algorithms as the starting point.

Another drawback of *k-means* algorithm is that number of groups k has to be defined in advance. More information on the *k-means* algorithm can be found in some of the classical references to k-means: Hartigan (1975), Everitt (1980) and Jain and Dubes (1988).

3.4 Neighbor-joining and neighbor-net

Apart from the k-means and seven hierarchical clustering algorithms, we also investigate the performance of the neighbor-joining and neighbor-net algorithms. These two algorithms were developed for making phylogenetic trees and networks in biology. In the past decade, there has been an increasing interest in the application of computational phylogenetic methods from biology to the study of language variation and language change, including these two techniques (Nakhleh et al., 2005; Hamed, 2005; Bryant,

Filimon, and Gray, 2005; Wichmann and Saunders, 2007). In this research we are particularly interested in seeing the performance of these algorithms on the dialect pronunciation data, since most of the previous studies were conducted on data from different languages where the boundaries between the various varieties are much sharper than in the case of our data.

3.4.1 Neighbor-joining

Neighbor-joining is a method for reconstructing phylogenetic trees that was first introduced by Saitou and Nei (1987). The main principle of this method is to find pairs of taxonomic units that minimize the total branch length at each stage of clustering. The distances between each pair of instances (in our case data collection sites) are calculated and put into the $n \times n$ matrix, where n represents the number of instances. The matrices are symmetrical since distances are symmetrical, i.e. distance (a,b) is always the same as distance (b,a). Based on the input distances, the algorithm finds a tree that fits the observed distances as closely as possible. While choosing the two nodes to fuse, the algorithm always takes into account the distance from every node to all other nodes in order to find the smallest tree that would explain the data. Once found, two optimal nodes are fused and replaced by a new node. The distance between the new node and all other nodes is recalculated, and the whole process is repeated until there are no more nodes left to be paired. The algorithm was modified by Studier and Kepler (Studier and Kepler, 1988) and the complexity was reduced to $O(n^3)$. The steps of the algorithm are as follows (taken from Felsenstein (2004)):

- For each node compute u_i which is the sum of the distances from that node to all other nodes

$$u_i = \sum_{j:j\neq i}^{n} \frac{D_{ij}}{(n-2)} \tag{3.9}$$

- Choose i and j for which $D_{ij} - u_i - u_j$ is smallest
- Join i and j. Compute the length from i and j to the newly formed node v using the equations below. Note that the distances from the new node to its children (leaves) need not be identical. This possibility does not exist in hierarchical clustering.

$$v_i = \frac{1}{2}D_{ij} + \frac{1}{2}(u_i - u_j) \tag{3.10}$$

$$v_j = \frac{1}{2}D_{ij} + \frac{1}{2}(u_j - u_i) \tag{3.11}$$

- Compute the distance between the new node and all of the remaining nodes

$$D_{(ij),k} = \frac{(D_{ik} + D_{jk} - D_{ij})}{2} \tag{3.12}$$

- Delete nodes i and j and replace them by the new node

This algorithm produces a unique unrooted tree under the principal of minimal evolution (Saitou and Nei, 1987). Trees can generally either be rooted or unrooted. A rooted tree has a *root* node from which all other nodes descend. The closer a node is to the root of the tree, the older is in time. Unrooted trees do not have a root node, and they do not allow us to define ancestor-descendant relationship between nodes (Page and Holmes, 2006).

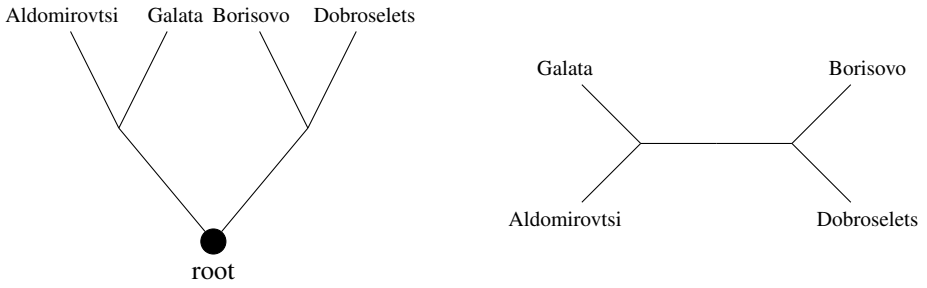

Figure 3.2: Rooted tree on the left hand side and unrooted tree on the right hand side.

In biology, the neighbor-joining algorithm has become a very popular and widely used method for reconstructing trees from distance data. It is fast and can be easily applied to a large amount of data. Unlike most hierarchical clustering algorithms, it will recover the true tree even if there is not a constant rate of change among the taxa (Felsenstein, 2004).

3.4.2 Neighbor-net

Neighbor-net is a network construction and data-representation tool and is, just as the neighbor-joining algorithm, agglomerative: taxa are combined into progressively larger and larger units (Bryant and Moulton, 2004). Unlike the neighbor-joining method, it reconstructs networks rather than trees. In each iteration it selects a pair of taxa to be grouped together, but it does not agglomerate those pairs immediately. That is done at a later stage when the second neighbor of one of the previously paired nodes is found. At that point, three nodes are replaced by two and the distance between newly formed nodes and the rest of the nodes is calculated.

In the first step a pair of clusters C_i and C_j is found to minimize the standard NJ formula:

$$Q(C_i,C_j) = (m-2)d(C_i,C_j) - \sum_{k=1,k\neq i}^{m} d(C_i,C_k) - \sum_{k=1,k\neq j}^{m} d(C_j,C_k) \tag{3.13}$$

where distance $d(C_i,C_j)$ between two clusters is the average of the distances between elements in each cluster.

$$d(C_i,C_j) = \frac{\sum_{x\in C_i}\sum_{y\in C_j} d(x,y)}{|C_i||C_j|} \tag{3.14}$$

C_i and C_j can contain one or two neighboring nodes. At the very beginning every node forms a separate cluster. Later on, some pairs of nodes will have been identified as neighbors. These pairs of neighbors are taken into account when selecting nodes to agglomerate (Bryant and Moulton, 2004).

In the second step, we find $x \in C_i$ and $y \in C_j$ that minimize

$$\hat{Q}(x_i,x_j) = (\hat{m}-2)d(x_i,x_j) - \sum_{k=1,k\neq i}^{\hat{m}} d(x_i,C_k) - \sum_{k=1,k\neq j}^{\hat{m}} d(x_j,C_k) \tag{3.15}$$

In the agglomeration step three closest nodes (x, y, z) are replaced by two new nodes (u, v). The distance from the two newly formed nodes to the rest of the nodes is calculated using the following formulae:

$$d(u,a) = (\alpha+\beta)d(x,a) + \gamma d(y,a) \tag{3.16}$$

$$d(v,a) = \alpha d(y,a) + (\beta+\gamma)d(z,a) \tag{3.17}$$

$$d(u,v) = \alpha d(x,y) + \beta d(x,z) + \gamma d(y,z) \tag{3.18}$$

where α, β, and γ are positive real numbers with $\alpha+\beta+\gamma = 1$.

In the graph generated by neighbor-net, splits of the taxa are represented by classes of parallel edges. Conflicting signals appear as boxes. Unlike in the neighbor-joining algorithm, the edge length estimation is done at the end and not during the agglomeration stage.

In Figure 3.3 we can see pronunciations of word агне /ˈagne/ ‘lamb’ collected at four different sites. These four pronunciations differ in positions 1 and 5. Since we do not have any model of phonetic evolution, we will use a very simple model and calculate the divisions of the four sites based on the number of positions in which they differ. Position 1, where we have initial prothetic /j/, gave the following division (Aldomirovtsi, Dobroselets) — (Borisovo, Galata). In position number 5, representing reflexes of the front nasalized vowel in word final position, the division of the sites is (Aldomirovtsi, Galata) — (Borisovo, Dobroselets). Splits at positions 1 and 5 are incompatible, since

	1	2	3	4	5
Aldomirovtsi:	j	ˈɑ	g	n	e
Borisovo:	-	ˈɑ	g	n	i
Dobroselets:	j	ˈɑ	g	n	i
Galata:	-	ˈɑ	g	n	e

Figure 3.3: Four pronunciations of word 'lamb'.

neither of the splits is the refinement of the other. Two different splits $S = X|Y$ and $S' = X'|Y'$ are compatible if one of the four conditions holds:

$$X \subset X', X \subset Y', Y \subset X', \text{or } Y \subset Y'$$

With the tree representation it would not be possible to represent this ambiguity with a single tree, but rather with two trees (Figure 3.4).

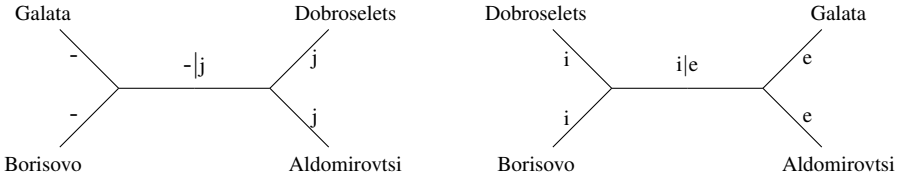

Figure 3.4: Two trees representing two incompatible splits in the data.

Unlike trees, network representation enables us to represent these conflicting signals within one graph (Figure 3.5). The incompatible splits are represented with reticulation that reflects the fact that in position 1 we have (Aldomirovtsi, Dobroselets) — (Borisovo, Galata) split and in position 5 (Aldomirovtsi, Galata) — (Borisovo, Dobroselets) split. We present these two splits in Figure 3.6.

In the final stage each branch in the network is assigned a length that represents the number of changes between each two nodes. In our network in Fugure 3.5 each branch has length 1, since there is only one change between the each two connected nodes.

$$\begin{aligned} d(Aldomirovtsi, Galata) = d(Galata, Borisovo) &= \\ d(Borisovo, Dobroselets) = d(Dobroselets, Aldomirovtsi) &= 1 \end{aligned} \quad (3.19)$$

For example, pronunciations for the villages Aldomirovtsi and Galata differ only at position 1, where we have insertion of the sound [j] at the beginning of the word collected at the site Galata. The distance between Aldomirovtsi and Borisovo is two since

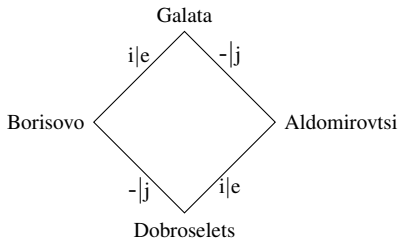

Figure 3.5: Neighbor-net representing two incompatible splits in one graph.

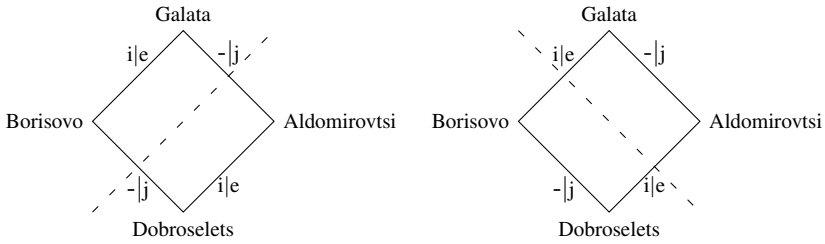

Figure 3.6: There are two different ways in which we can split this reticulation.

the pronunciations from these two sites differ in two positions, 1 and 5. The same holds for the distance between Dobroselets and Galata. The more changes between two nodes, the longer the branch in the network. This allows us to visualize the distances between the nodes and determine which nodes group together. The division of the nodes into a groups is done manually by visually inspecting the length of the branches.

One important property of the neighbor-net algorithm is that if the input distance is circular it will return the collection of circular splits. If the input distance is additive, it will return the corresponding tree (Bryant and Moulton, 2004). This property of neighbor-net enables us to see if the data is tree-like or non-tree-like. This can be very useful in the investigations of language change, since throughout the history of linguistics two models of language change have been competing—family tree model (Schleicher, 1853) and wave model (Schmidt, 1872). The main advantages of the network representation is that it allows us a) to check if the data in question is tree-like or network-like and b) to represent both models of language change at the same time.

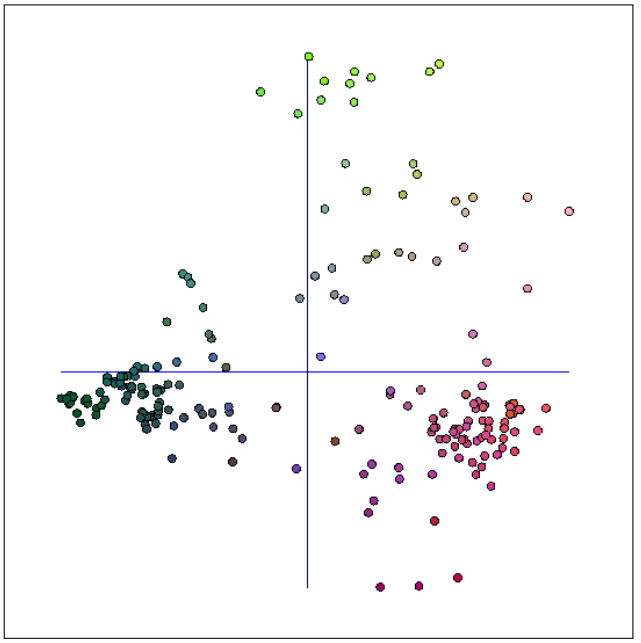

Figure 3.7: First two dimensions extracted by MDS are plotted against x- and y-axes. We additionally mark the first three dimensions using red, green and blue color of the dots.

3.5 Visual inspection

In this section we visually inspect the results of the classifications produced by various distance-based methods.

3.5.1 MDS

The results of performing MDS analysis can be seen in Figure 3.7 where the first two extracted dimensions are plotted against x- and y-axes. We additionally represent the first three dimensions using different proportions of red, green and blue color, the so-called RGB color model. This is done by translating every position in three-dimensional MDS space into a distinct color. The amount of red, green and blue represents the first three MDS dimensions respectively. A very good explanation on how to display the results of MDS using the full RGB color spectrum can be found in Leinonen (2010, 208-211).

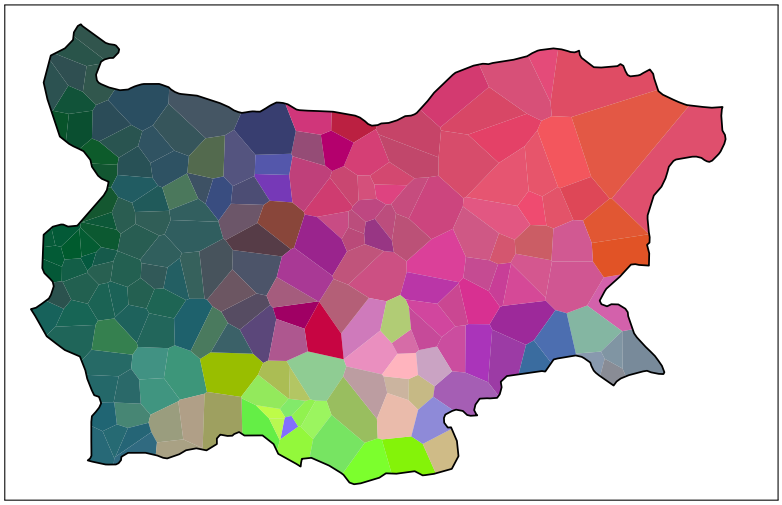

Figure 3.8: MDS map. First three extracted dimensions are represented with different amount of red, green and blue.

MDS plot in Figure 3.7 shows two relatively clearly separated groups of dialects along the x-axis. Along the y-axis the third group of varieties is visible, although it is not clearly separated from any of the two previously identified groups. We can also identify these three groups of dots based on their different colors in the MDS plot.

In Figure 3.8 we color the area around each site on the map of Bulgaria using the color assigned by MDS, which allows us to see if there is a geographical cohesion of the extracted groups. The MDS map in Figure 3.8 shows that the two groups identified on the MDS plot in Figure 3.7 correspond to the western, colored green on the map, and eastern group, colored red, of varieties. The separation of western and eastern varieties approximately follows the so-called *yat* border described in Chapter 2. The third group of varieties are the sites located in the south of the country in the area of Rodopi mountains, colored with various shades of green. These findings correspond well with the traditional scholarship described in Chapter 2. According to traditional scholarship, these three dialect areas are three out of six main dialect groups identified by Stoykov (2002). Three extracted dimensions explain 95.45 per cent of the variation found in the distance matrix.

In Figure 3.9 we display the values of each of the MDS dimensions separately. The first dimension itself explains 64.84 per cent of the variation. The map on the top in Figure 3.9 reveals that the variation captured by the first MDS dimension follows approximately the *yat* line and divides the country into the west and east. The second

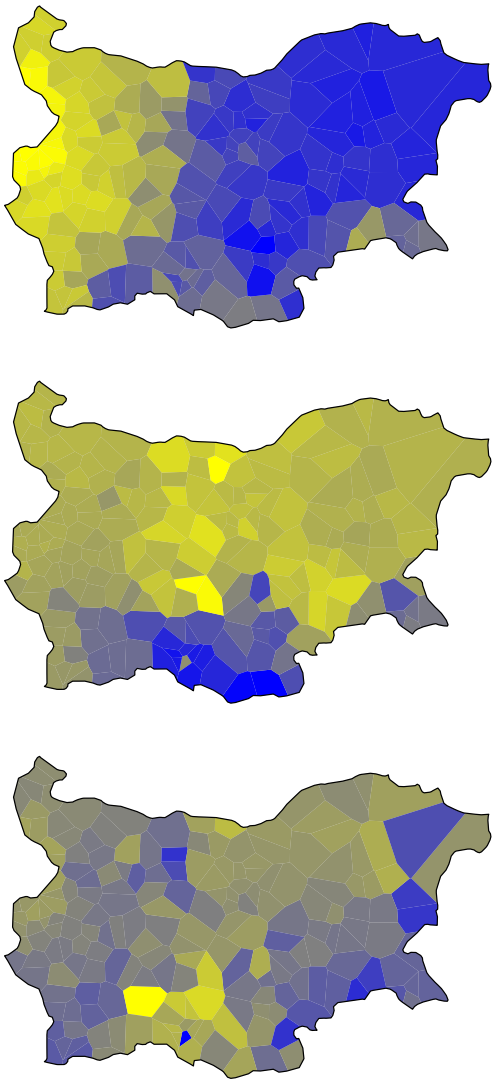

Figure 3.9: MDS values for each of the dimensions separately projected on the map of Bulgaria. Top: first dimension explains 64.84 per cent of the total variation. Middle: second dimension explains 27.56 of the variation. Bottom: third dimension explains 3.05 per cent of the variation.

dimension explains 27.56 per cent of the total variation found in the data. We display the values for this dimension on the middle map in Figure 3.9. This map shows clear separation of the southern varieties from the rest of the country. The first two dimensions, responsible for the division of the varieties into eastern, western and southern group, already account for the 92.40 per cent of the variation. The third MDS dimension explains only 3.05 per cent of the variation. On the bottom map in Figure 3.9, where we represent the values of the third dimension, we do not see any pattern in the geographic distribution of the colors.

3.5.2 Clustering

In cluster analysis, the number of groups that will be retrieved by a certain algorithm has to be specified in advance. For all clustering algorithms we performed analyses for the number of groups ranging from 2 to 10. In this thesis we present only one part of the maps important for the further analyses.

Visual inspection has revealed that three hierarchical clustering algorithms fail to identify any structure in the data, namely single link (Figure 3.10) and two centroid algorithms, UPGMC and WPGMC (Figure 3.14 and Figure 3.15). Closer inspection of the single link dendrogram shows the presence of the *chain effect* (left Figure D.1 in Appendix D), while dendrograms drawn using two centroid methods reveal a large number of reversals (left hand side dendrograms in Figure D.5 and Figure D.6 in Appendix D). In Appendix D we present dendrograms for all seven hierarchical clustering algorithms, plain on the left hand side and with the noise (see Section 3.6.2) on the right hand side.

Three hierarchical clustering algorithms, UPGMA, WPGMA and Ward's method show exactly the same two-way split into eastern and western group that approximately corresponds to the *yat* border. This split is also visible on the map drawn using k-means algorithm. A similar split, which also includes several sites from the southern area, is found using complete link algorithm.

On the level of three dialect groups, except for the three algorithms that do not find any structure in the data, the remaining four hierarchical clustering algorithms, as well as k-means algorithm, distinguish eastern, western and southern group of dialects. This finding correspond well both with the MDS analysis and with the traditional scholarship as well. The three-way split is also found using the neighbor-joining algorithm, although the southern group is larger when compared to the results obtained using other algorithms (Figure 3.18).

Since traditional scholarship distinguishes six main dialect areas, we wanted to see if any of the algorithms would give the same analysis of the sites. At this level of hierarchy different algorithms found different groups and none of them corresponds completely with the traditional division. UPGMA distinguishes only three dialect areas, eastern, western and southern which is further divided into smaller groups. Apart from these three areas, WPGMA also distinguishes the group of dialects at the border with Ser-

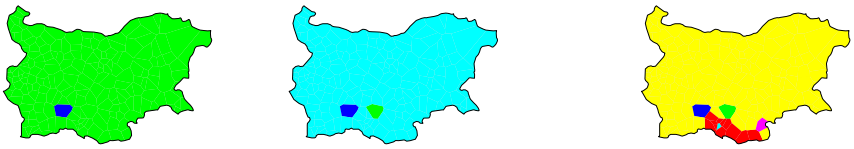

Figure 3.10: Single link: 2-way, 3-way and 6-way splits.

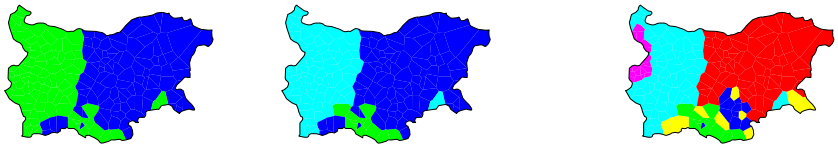

Figure 3.11: Complete link: 2-way, 3-way and 6-way splits.

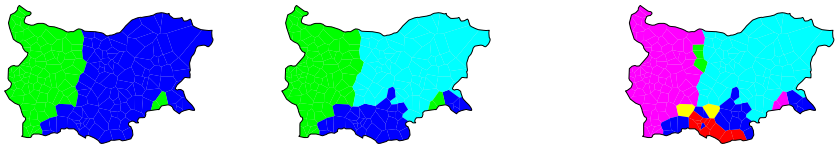

Figure 3.12: UPGMA: 2-way, 3-way and 6-way splits.

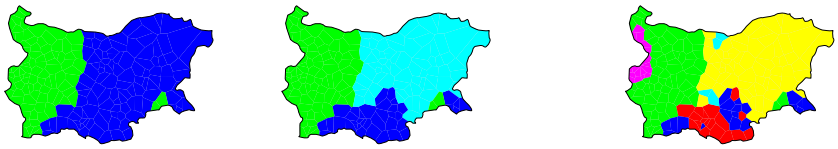

Figure 3.13: WPGMA: 2-way, 3-way and 6-way splits.

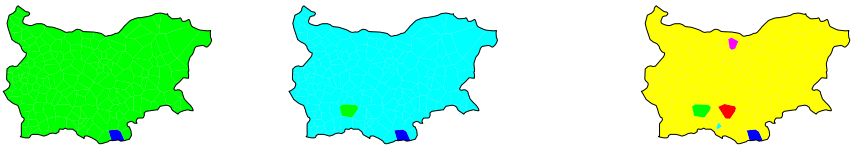

Figure 3.14: UPGMC: 2-way, 3-way and 6-way splits.

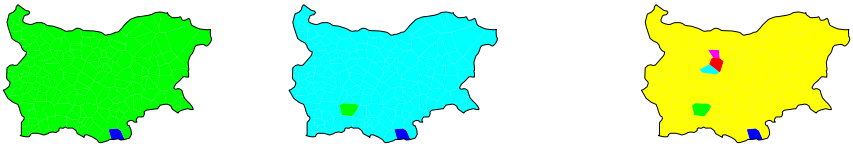

Figure 3.15: WPGMC: 2-way, 3-way and 6-way splits.

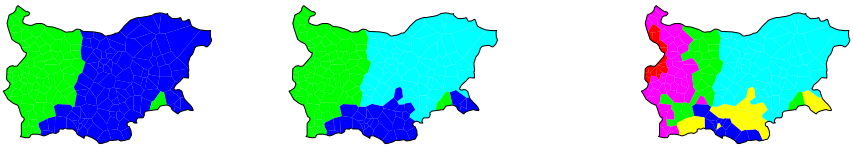

Figure 3.16: Ward's method: 2-way, 3-way and 6-way splits.

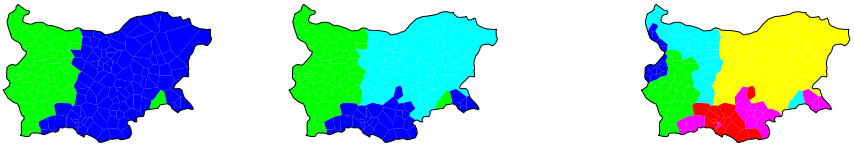

Figure 3.17: K-means: 2-way, 3-way and 6-way splits.

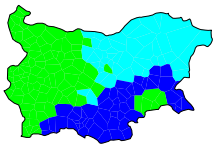

Figure 3.18: Neighbor-joining: 3-way split.

bia that corresponds well with the transitional zone in Stoykov's classification. These four traditional groups are also found by Ward's method, complete link and k-means algorithm. The k-means algorithm also distinguishes two groups that resemble northwest-southwest split described by Stoykov. On the other hand, Ward's method further divides the western area into eastern and western groups, which is not found in any traditional atlases.

Visual inspection of the maps shows that different clustering algorithms give different analysis of the distances obtained using Levenshtein algorithm. They differ among each other, and also from the traditional scholarship.

3.5.3 Neighbor-joining and neighbor-net

In Figure 3.19 we can see the unrooted tree produced by neighbor-joining algorithm.[2] In the tree, there is a three-way split of the varieties that corresponds to the east, west, and south division (Figure 3.18). The split produced by neighbor-joining is geographically coherent and corresponds to some extent to the three-way divisions produced by previously described clustering techniques. However, while four clustering algorithms, namely UPGMA, WPGMA, Ward's method and k-means algorithm, almost perfectly agree on the three-way split of the varieties, neighbor-joining produces a much larger southern group of varieties which includes the transitional zone between Balkan and Rupian dialects in the southeast. This transitional zone can be seen on the map in Figure 2.2.

The network produced by the neighbor-net algorithm can be seen in Figure 3.20. Since neighbor-net is using the same selection criteria and formulae for computing the distances between the nodes that are to be fused, the grouping of the sites matches quite well the one done by neighbor-joining. In the network in Figure 3.20 we can distinguish eastern, western and southern groups. The detection of the groups is done by visually inspecting the network and looking for the longest branches since they signal us which groups of varieties are the most distant ones. In Figure 3.20 the branches connecting western varieties from the rest of the sites are the longest in the network. This means that this group of varieties is the most distant from the rest of the sites. We mark this split with the yellow dashed line that cuts the network in two. Accordingly, the data can be first split into two groups, which put on the map, roughly corresponds with the east-west split along the *yat* border. Eastern varieties can be further divided into two groups, southern and northern. Branches connecting southern varieties to each other are longer when compared to the branches within other groups, which suggests that the language varieties found in this region are more heterogeneous than in other areas. All these findings correspond well with the traditional scholarship described in Chapter 2.

[2]Neighbor-joining tree and neighbor-net were produced using SplitsTree software that can be freely downloaded at http://www.splitstree.org.

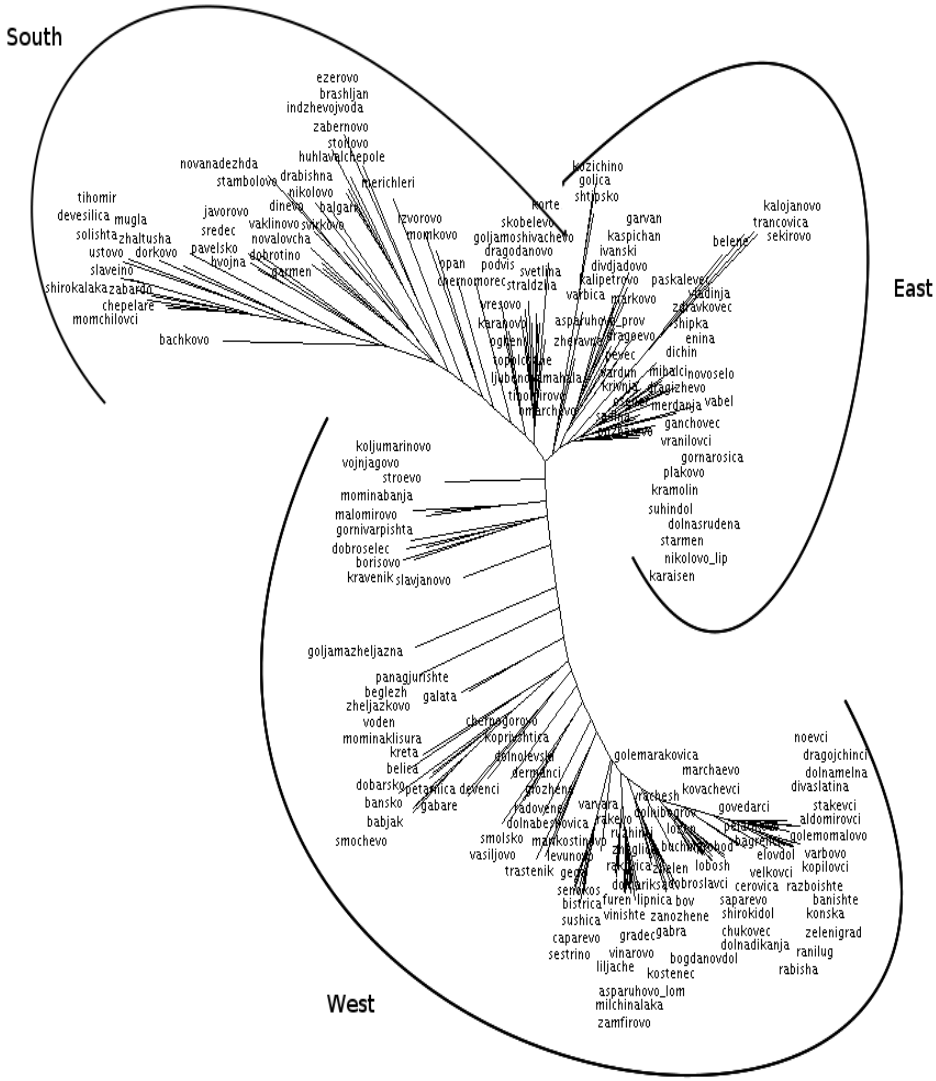

Figure 3.19: The neighbor-joining tree shows a three-way division of Bulgarian dialects into western, eastern and southern groups.

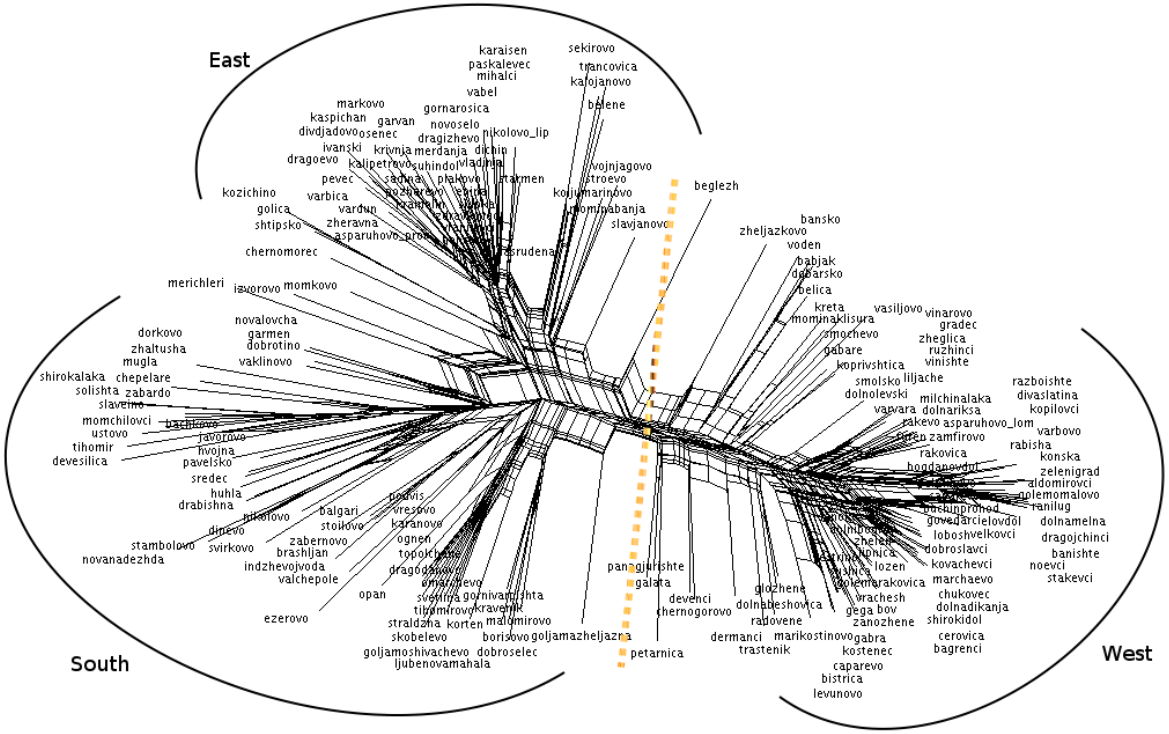

Figure 3.20: Neighbor-net shows many conflicting signals in the data. The biggest split is between the western sites on one side and the eastern and southern on the other.

However, the detection of groups in the network is done visually which makes the divisions to some extent arbitrary. Network representation allows us to see that there are many conflicting signals represented as reticulations, which makes the data look more network-like than the tree-like. These conflicting signals show that there are more ways in which the sites could be grouped. For example, on one side there are features shared between southern and western varieties and we can group these varieties together, while on the other hand there are features shared between southern and western varieties, and we can see splits that would classify these two groups together. At this moment, it is still not possible to automatically detect which specific features are responsible for which divisions.

3.6 Evaluation of the results of distance-based methods

Although instable, clustering techniques are still the most commonly used tool in dialectometry for group detection within a certain dialect area. In this section we propose several evaluation techniques that should be used in order to deal with the instability of the clustering algorithms. Since there is no direct way to evaluate the performance of clustering algorithms, we propose a combination of different techniques that can help us determine if the results of the applied clustering technique are artifacts of the algorithm or the detection of real groups in the data. The proposed evaluation methods can be divided into external and internal. External validation of the clustering results include the modified Rand index, purity and entropy. External validation involves comparison of the structure obtained by different algorithms to a *gold standard*. In our study we used the manual classification of all the sites produced by an expert on Bulgarian dialects as a *gold standard*. Internal validation included examining the cophenetic correlation coefficient, noisy clustering and a consensus tree, which do not require comparison to any *a priori* structure, but rather try to determine if the structure obtained by algorithms is intrinsically appropriate for the data.

3.6.1 External validation

The **modified Rand index** (Hubert and Arabie, 1985) is used for comparing two different partitions of a finite set of objects. It is a modified form of the Rand index (Rand, 1971), one of the most popular measures for comparing partitions. Given a set of n elements $S = o_1, ...o_n$ and two partitions of S, $U = u_1, ...u_R$ and $V = v_1, ...v_C$ we define

a the number of pairs of elements in S that are in the same set in U and in the same set in V

b the number of pairs of elements in S that are in different sets in U and in different sets in V

c the number of pairs of elements in S that are in the same set in U and in different sets in V

d the number of pairs of elements in S that are in different sets in U and in the same set in V

The Rand index R is

$$R = \frac{a+b}{a+b+c+d} \approx \frac{|\text{agreeing pairs}|}{|\text{all pairs}|}$$

In this formula a and b are the number of pairs of elements in which two classifications agree, while c and d are the number of pairs of elements in which they disagree. The value of the Rand index is between 0 and 1, with 0 indicating that the two data clusters do not agree on any pair of points and 1 indicating that the data clusters are exactly the same. In dialectometry, this index was used by Heeringa et al. (2002) to validate dialect comparison methods. A problem with the Rand index is that it does not return a constant value (zero) if two partitions are picked at random. Hubert and Arabie (1985) suggested a modification of Rand index that corrects this property. It can be expressed in the general form as:

$$\frac{RandIndex - ExpectedIndex}{MaximumIndex - ExpectedIndex}$$

The expected index is the expected number of pairs which would be placed in the same set in U and in the same set in V by chance. The maximum index represents the maximum number of objects that can be put in the same set in U and in the same set in V. The Modified Rand Index (MRI) value ranges between -1 and 1, where 1 represents an upper bound (perfect overlap) and 0 indicates that the index equals its expected value. For a more detailed explanation of the modified Rand index, please refer to Hubert and Arabie (1985).

Entropy and **purity** are two measures used to evaluate the quality of clustering by looking at the reference class labels of the elements assigned to each cluster (Zhao and Karypis, 2001). Entropy measures how different classes of elements are distributed within each cluster. The entropy of a single cluster is calculated using the following formula:

$$E(S_r) = -\frac{1}{\log q} \sum_{i=1}^{q} \frac{n_r^i}{n_r} \log \frac{n_r^i}{n_r}$$

where S_r is a particular cluster of size n_r, q is the number of classes in the reference data set, and n_r^i is the number of the elements of the ith class that were assigned to the rth cluster. The overall entropy is the sum of all cluster entropies weighted by the size of the cluster:

$$E = \sum_{r=1}^{k} \frac{n_r}{n} E(S_r)$$

The **purity** measure is used to determine to which extent a cluster contains objects from primarily one class. The purity of a cluster is calculated as:

$$P(S_r) = \frac{1}{n_r} max(n_r^i)$$

while the overall purity is the weighted sum of the individual cluster purities:

$$P = \sum_{r=1}^{k} \frac{n_r}{n} P(S_r)$$

3.6.2 Internal validation

The **cophenetic correlation coefficient** (Sokal and Rohlf, 1962) is Pearson's correlation coefficient computed between the cophenetic distances produced by clustering and those in the original distance matrix. The cophenetic distance between two objects is the similarity level at which those two objects become members of the same cluster during the course of clustering (Jain and Dubes, 1988) and is represented as branch length in dendrogram. It measures the extent to which the clustering results correspond to the original distances by comparing the distances between each two objects calculated from the dendrogram to the original distances. When the clustering functions perfectly, the value of the cophenetic correlation coefficient is 1.

Noisy clustering, also called composite clustering, is a procedure in which small amounts of random noise are added to matrices during repeated clustering. The main purpose of this procedure is to reduce the influence of outliers on the regular clusters and to identify stable clusters. As shown in Nerbonne et al. (2008) it gives results that nearly perfectly correlate with the results obtained by bootstrapping—a statistical method for measuring the support of a given edge in a tree (Felsenstein, 2004). The advantage of noisy clustering, compared to bootstrapping, is that it can be applied on a single distance matrix—the same one used as input for the classification algorithms. In this thesis noisy clustering analysis was done using L04 software. The amount of noise c was set to one-half standard deviation of distances in the matrix. To each cell in the distance matrix we add different random amounts of noise which ranges between 0 and c. This process is repeated 100 times, resulting in the same number of distance matrices. We apply clustering to each of the matrices and finally calculate composite dendrogram which contains groups of sites that are clustered together in more than 50 per cent of the iterations.

A **consensus dendrogram**, or consensus tree, is a tree that summarizes the agreement between a set of trees (Felsenstein, 2004). A consensus tree that contains a large number of internal nodes shows high agreement between the input trees. On the other hand, if a consensus tree contains few internal nodes, it is a sign that input trees classify the data in conflicting ways. The majority rule consensus tree, used in this study, is a tree that consists of the groups, i.e. clusters, which are present in the majority of the trees under study. Clusters that appear in the consensus tree are those supported by the majority of algorithms and can be taken with greater confidence to be true clusters. In this research a consensus dendrogram was created with the L04 software from four dendrograms produced by four different hierarchical clustering methods, namely complete link, UPGMA, WPGMA and Ward's method.

3.6.3 Results

In this section we present the results obtained by the above described methods.

External validation

In order to compare divisions done by clustering algorithms with the division of sites done by experts we calculated the modified Rand index, entropy and purity for the 2-fold, 3-fold, and 6-fold divisions done by algorithms on the one hand, and those divisions according to the experts on the other. The results can be seen in Table 3.1. The neighbor-joining algorithm produced an unrooted tree (Figure 3.19), where only 3-fold division of the sites can be identified. This classification of the sites is represented on the map in Figure 3.18. Hence, all the indices were calculated only for the 3-fold division made by neighbor-joining. Since even the detection of the main groups in the neighbor-net is pretty arbitrary, we do not evaluate the divisions done by neighbor-net using any of the proposed evaluation techniques.

Table 3.1: Results of external validation: the modified Rand index (MRI), entropy (E) and purity (P). Results for the 2, 3 and 6-fold divisions are reported.

Algorithm	MRI(2)	MRI(3)	MRI(6)	E(2)	E(3)	E(6)	P(2)	P(3)	P(6)
single link	-0.004	0.007	-0.001	0.958	0.967	0.881	0.614	0.396	0.360
complete link	0.495	0.520	0.350	0.510	0.542	0.467	0.848	0.766	0.645
UPGMA	**0.700**	**0.627**	0.273	0.368	**0.445**	0.583	0.914	**0.853**	0.568
WPGMA	**0.700**	0.626	0.381	0.368	0.445	0.448	0.914	**0.853**	0.665
UPGMC	-0.004	0.007	-0.006	0.959	0.967	0.926	0.614	0.396	0.310
WPGMC	-0.004	0.007	-0.005	0.958	0.967	0.925	0.614	0.396	0.305
Ward's method	**0.700**	**0.627**	0.398	0.368	**0.445**	0.441	0.914	**0.853**	0.675
k-means	**0.700**	0.625	**0.471**	**0.354**	0.451	**0.355**	**0.919**	0.756	**0.772**
neighbor-joining	-	0.461	-	-	0.550	-	-	0.777	-

In Table 3.1 we can see that the values of the modified Rand index for single link and two centroid methods are very close to 0, which is the value we would get if the partitions were picked at random. UPGMA, WPGMA, Ward's method and k-means, which gave nearly the same 2-fold division of the sites, show the highest correspondences with the divisions done by experts. For 3-fold and 6-fold divisions the values for the modified Rand index went down for all algorithms, which was expected since the number of groups increased. The two algorithms with the highest values of the index are Ward's method and UPGMA for 3-fold, and k-means for the 6-fold division. Just as in the case of the 2-fold division, the single link, UPGMC, and WPGMC algorithms have values of the modified Rand index close to 0. Neighbor-joining produced a relatively low correspondence with expert opinion for the 3-fold division—0.461. Similar results for all algorithms and all divisions were obtained using entropy and purity measures. We

conclude from this that the modified Rand index is a good measure of the agreement of one partition with another, and that entropy and purity impose on it only in providing measures per cluster.

Internal validation

In the next step internal validation methods were used to check the performance of the algorithms: the cophenetic correlation coefficient, noisy clustering and consensus tree. Since k-means does not produce a dendrogram, it was not possible to calculate the cophenetic correlation coefficient. The values of the cophenetic correlation coefficient for the remaining eight algorithms can be seen in Table 3.2. We can see that clustering results of the UPGMA have the highest correspondence to the original distances of all algorithms—90.26 per cent. They are followed by the results obtained by using complete link and neighbor-joining algorithm.

Table 3.2: Cophenetic correlation coefficient.

Algorithm	CCC	p
single link	0.7804	0.0001
complete link	0.8661	0.0001
UPGMA	**0.9026**	0.0001
WPGMA	0.8563	0.0001
UPGMC	0.8034	0.0001
WPGMC	0.6306	0.0001
Ward's method	0.7811	0.0001
neighbor-joining	0.8587	0.0001

All correlations are highly significant with $p < 0.0001$ calculated using a Mantel test. Given the poor performance of the centroid and single link methods in detecting the dialect divisions scholars agree on, we note that cophenetic correlation coefficients are not successful in distinguishing the better techniques from the weaker ones. We conjecture that the reason for this lies in the fact that the cophenetic correlation coefficient so dependent is on the lengths of the branches in the dendrogram, while our primary purpose is the classification. Although we had expected neighbor-joining to benefit from the fact that it assigns different branch lengths in its fusion step, Table 3.2 shows that it was not able to convert this additional freedom to an improved cophenetic correlation.

Results of noisy clustering can be seen in Appendix D, where dendrograms on the right hand side are created by applying noisy clustering to the original distance matrix. Noisy clustering, that was applied with the seven hierarchical algorithms, has confirmed that there are only two relatively stable groups in the data: eastern and western. Dendrograms obtained by applying noisy clustering to the whole data set show low confidence

for the two-way split of the data, between 52 and 60 per cent. After removing the southern villages from the data set, we obtained dendrograms that confirm two-way split of the data along the *yat* border with much higher confidence, ranging around 70 per cent. These values are also not very high. In order to check the reason of the influence of the southern varieties on the noisy clustering we examine an MDS plot (Figure 3.21) in two dimensions with cluster groups marked by symbols.

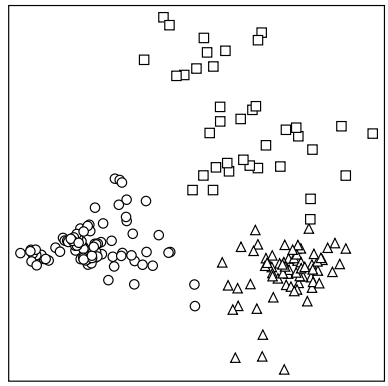

Figure 3.21: MDS plot: different symbols present grouping done by WPGMA algorithm, while the distances between the objects present the MDS analysis. Circles are western, triangles eastern and squares Rupian dialects.

In Figure 3.21 we can see first two dimensions extracted by MDS plotted against x- and y-axes. These two dimensions represent together 92.40 per cent of the variation in the data. The distance between the points in the plot accords with the linguistic difference between them: similar varieties are located close to each other while more different varieties are placed further apart from each other. In that way we can see if there are groups of varieties put together. The orientation of the x- and y-axes is not relevant for the analysis. We are interested in detecting clouds of symbols and their distance from the other symbols of groups of symbols. Additionally, grouping produced by the WPGMA algorithm is represented by different symbols: western varieties (approximately west of the *yat* line) are marked with the circles, Rupian dialects are marked with square symbols, while the rest of the varieties (central and northeastern) are marked with triangles. The MDS plot reveals two homogeneous groups and a third, more diffuse, group that lies at a remove from them. The third group of the sites represents the southern group of varieties, marked with square symbols, and is much more heterogeneous than the rest of the data. Closer inspection of the MDS plot in Figure 3.21 also shows that this group of dialects has a particularly unclear border separating it from the eastern dialects, which could explain the results of the noisy clustering applied to the whole data set.

Since different algorithms gave different divisions of sites, we used a consensus dendrogram in order to detect the clusters on which most algorithms agree. Since single link, UPGMC and WPGMC have turned to be inappropriate for the analysis of our data, they were not included in the consensus dendrogram. The consensus dendrogram drawn using complete link, UPGMA, WPGMA and Ward's method can be seen in Figure 3.22. The names of the sites are colored based on the experts' opinion, i.e. the same as on the map placed next to the consensus dendrogram in Figure 3.22. The dendrogram shows strong support for the east-west division of sites, but no agreement on the division of sites within the eastern and western areas.

At this level of hierarchy, i.e. 2-way division, there are several sites classified differently by algorithms and by experts. They are colored black on the map in Figure 3.23. This map clearly shows that these sites follow the *yat* border and represent the border cases. The only two exceptions are villages in the southeast, namely Voden and Zhelyazkovo. However, according to many traditional dialectologists these villages should be classified as western dialects due to many features that they share with the dialects in the west (personal communication with Prof. Vladimir Zhobov). The four algorithms show agreement only at the very low level where several sites are grouped together and again on the highest level. It is not possible to extract any hierarchical structure that would be present in the majority of four analyses.

3.7 Discussion and conclusions

We were unusually fortunate in obtaining very low-dimensional MDS solutions which represent over 90 per cent of the variation in the data. For this reason, we relied on MDS not only for a map of Bulgarian dialect variation (Figure 3.7), but also as a diagnostic to understand the less reliable clustering techniques (Figure 3.21). We tentatively infer that the clustering results are less stable due to the fact that the dialect groups, which are completely obvious in the MDS plots, are not so distinct that borderline cases are impossible. There is no wide swath of clear space between the different groups in Figure 3.21.

Different clustering validation methods have shown that three algorithms are not suitable at all for the data we are working with, namely single link, UPGMC and WPGMC. The remaining four hierarchical clustering algorithms gave different results depending on the level of hierarchy, but all four algorithms had fairly high agreement on the detection of two main dialect areas within the dialect space. At the lower level of hierarchy, i.e where there are more clusters, the performance of the algorithms is poorer, both with respect to the expert opinion and with respect to the mutual agreement as well. As shown by noisy clustering, the 2-fold division of the Bulgarian language area is the only partition of sites that can be asserted with high confidence.

The division of sites done by the k-means algorithm corresponded well with the expert divisions. Two and three-way divisions also correspond well with the divisions of four hierarchical clustering algorithms. What we find more important is the fact that in

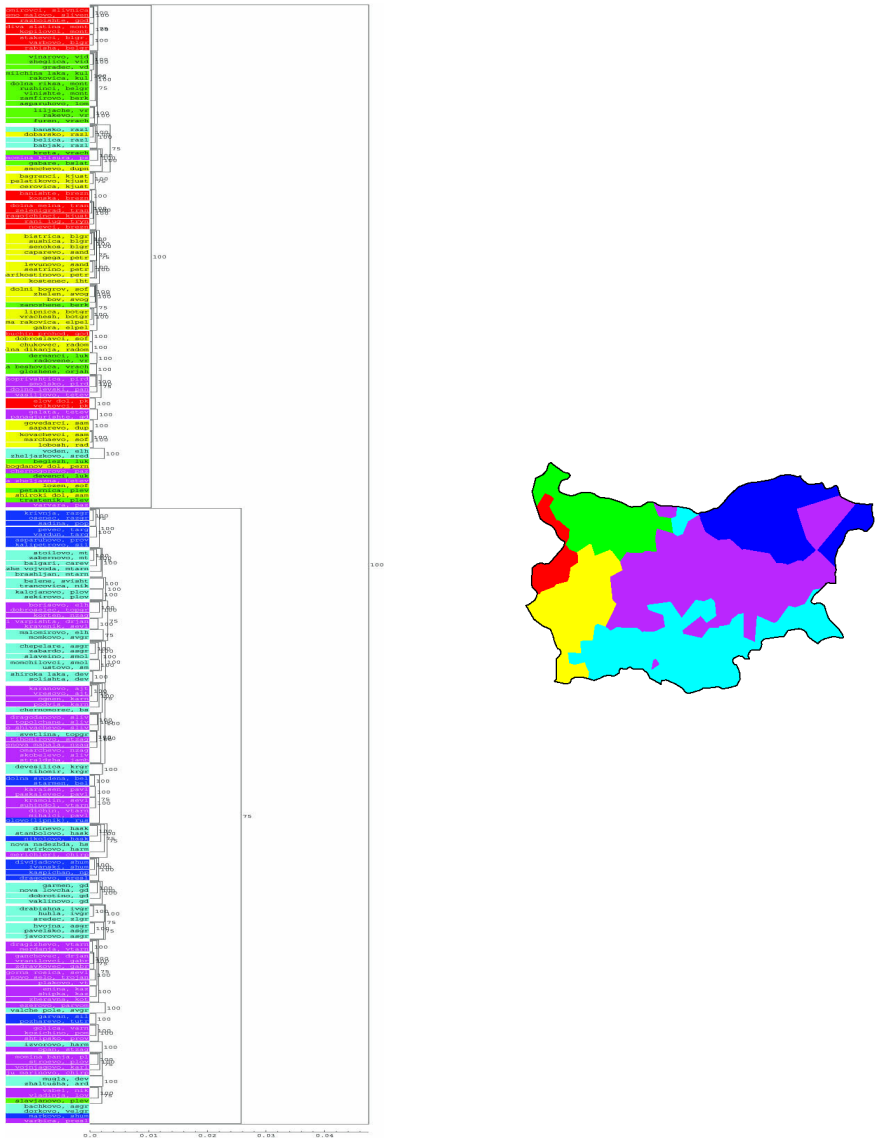

Figure 3.22: Consensus dendrogram on the left-hand side drawn using complete link, UPGMA, WPGMA and Ward's method. Division of the sites in the data set done by an expert is on the right-hand side.

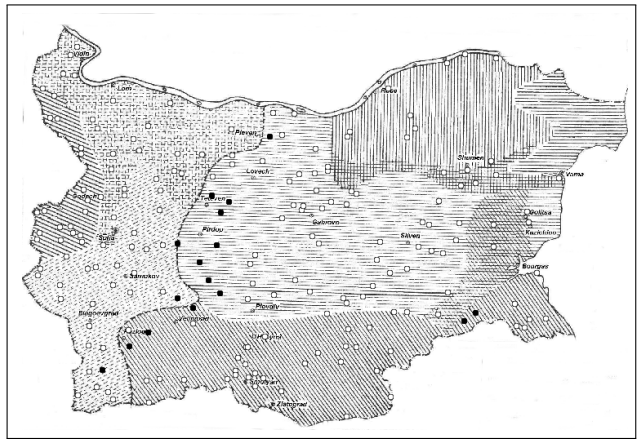

Figure 3.23: Sites differently classified by four hierarchical clustering algorithms and experts are colored black. Except for 2 villages, they all follow the *yat* line.

the divisions obtained by the k-means algorithm into 2, 3, 4, 5 and 6 groups the two-way division into the eastern and western groups is the only stable division that appears in all partitions.

The results of the neighbor-joining algorithm were a bit less satisfactory. The reason for this could be in the fact that our data is not tree-like, but rather contains a lot of borrowings due to contact between different dialects. Unlike biological species and languages, where neighbor-joining was earlier successfully applied, dialect varieties form a continuum, rather than well defined groups. A recent study of Chinese dialects (Hamed and Wang, 2006) has shown that their development is not tree-like and that in such cases usage of tree-reconstruction methods can be misleading.

Neighbor-net has confirmed that there are many conflicting signals in the data, represented as reticulations. In the neighbor-net there are three distinguishable groups, although for many sites it is not clear to which group they belong to. Detection of the groups in networks is pretty arbitrary, but we find neighbor-net to be a very useful representation tool since it is possible to see to which extent the data is tree-like.

This research shows that clustering algorithms should be applied with caution as classifiers of language dialect varieties. Where possible, several internal and external validation methods should be used together with the clustering algorithms in order to validate their results and make sure that the classifications obtained are not mere artifacts of algorithms but natural groups present in the data set. Since performance of clustering algorithms depends on the sort of data used, evaluation of algorithms is a necessary step in order to obtain results that can be asserted with high confidence.

The fact that there are two distinct groups in our data set that can be asserted with high confidence, and that the third one that was found with less confidence, even though six are found in the traditional atlases, could possibly be due to the simplified representation of the data. It is also possible that some of the features responsible for the traditional 6-way division are not present in our data set. The quality of the data set and detail comparison of the automatically produced and traditional maps is described in Chapter 4. Regardless of the quality of the input data set, clustering algorithms will partition data into any given number of groups even if there is no natural separation of the data. For this reason it is essential to use different evaluation techniques along with the clustering algorithms.

We are fully aware of the fact that in this kind of research the so-called *gold standard* is not something that should be taken for granted. Classification of language varieties done by experts suffers itself from certain flaws. These classifications can often be subjective and based on the non-linguistic factors. Even when they are linguistically motivated, very often the classification is done using a single feature or a small number of features. However, traditional scholarship is valuable source of information in dialectometry. It helps us evaluate different quantitative methods that are being developed or adapted from other disciplines, and better understand how different varieties are perceived and classified by humans. At the same time quantitative methods enable us to reevaluate traditional divisions by using more objective techniques based on large amount of features which is usually hard or impossible to do using traditional approach.

Chapter 4

Comparison to the traditional maps

In Chapter 3 we have presented the results of analyzing the dialect pronunciation data using various classification methods which proceed from a matrix that stores information on the distances between each two sites in the data set. These distances represent linguistic distances and are calculated using the Levenshtein method. The resulting dialect divisions agree to different extents among each other and with the traditional scholarship. The differences between computational and traditional methods could be due to: a) the Levenshtein method used to calculate the linguistic distances between the sites; b) problems with the quantitative classification methods; c) the possible absence of some of the features responsible for traditional divisions from our data set; d) the fact that sometimes traditional divisions are based on criteria other than linguistic ones, or e) linguistic criteria that are not sound enough. In this chapter we investigate the differences between computational and traditional classifications in more depth in order to get better insight into these issues. This task is very difficult since on one hand we are trying to develop new methods that are tested against the traditional divisions and on the other we apply quantitative methods hoping to improve traditional classifications and get new insights into dialect divisions and dialect change. By comparing the two classification on a level of a very fine detail, we hope to find out more about both the effectiveness of our method and the representativeness of our data set. This chapter is based on the work presented in Houtzagers, Nerbonne, and Prokić (2010).

4.1 East-west division

According to various clustering techniques, the east-west division of Bulgarian dialect area is the most important division found by most of the algorithms (see Chapter 3). This division was also found using multidimensional scaling. It corresponds well with the *yat* boundary described in traditional literature (Stoykov, 2002, 83-87) as the main dialect border in Bulgaria.[1] In Figure 4.1 we can see two classifications projected on the same map. The division resulting from weighted pair group method using arithmetic averages (WPGMA) clustering algorithm is marked by different shades, while the traditional boundary as found in Stoykov (2002) is marked with black line. It is evident that there is high correspondences between two divisions, except that the computational one is further east.

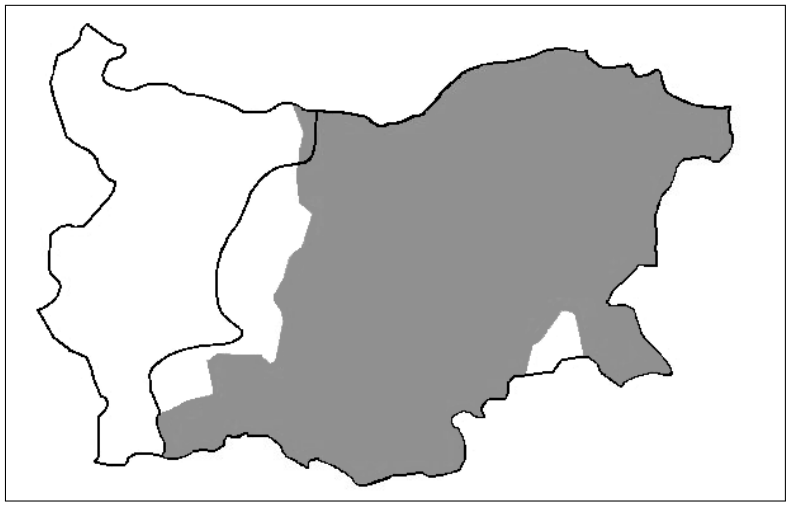

Figure 4.1: Two-way classification done by WPGMA algorithm and traditional two-way division of sites.

As found in Houtzagers et al. (2010) Stoykov's *yat* border is based on the bundle of 48 isoglosses which is the number of corresponding maps in OT.[2] These isoglosses reflect various phonetic phenomena which are present in 101 words in the Buldialect data set:

- reflexes of *yat* in specific positions

[1] Detail description of different reflexes of *yat* in Bulgarian is given is Chapter 2.

[2] OT: *Български диалектен атлас, обобщаващ том I-III. Фонетика, акцентология, лексика* [Atlas of Bulgarian Dialects: Phonetics. Intonation. Lexicology, Vol. I - III], (Kochev et al., 2001)

- presence vs. absence of mixture of the reflexes of the two *yers* and the two nasal vowels
- vowel reduction phenomena
- presence vs. absence of epenthetic /l/
- change of $*d^j$, $*t^j$ into [g^j], [k^j]
- reflexes of $*l_b$, $*l_o$ and syllabic $*l$
- presence vs. absence of the changes $*a >$ [e] in certain positions
- presence vs. absence of the change $*dn >$ [nn]

Close inspection of these words has revealed that 68 of them show the east-west division of the sites. Very few of them perfectly match the traditional division, and most of the isoglosses run east of the *yat* line. The large number of words where the east-west division is present explains the stability of the *yat* line in most of the computational analyses. This division is absent only from the analyses done by three clustering algorithms that have proven to be unsuitable for the analysis of our data (see Chapter 3, 38). Since most of the isoglosses run east of the *yat* line, this is also reflected in the aggregate analysis: the east-west division on the quantitative maps represents the average of all isoglosses in this bundle. We note that generally speaking the two classifications correspond to a high degree and that features responsible for the east-west division are well represented in our data set.

In the rest of this chapter we look more closely into divisions of the areas west and east of the *yat* line. While the east-west division corresponds well on the quantitative and traditional maps, further classifications into smaller dialect zones show much bigger differences.

4.2 Western dialects

On the map shown in Stoykov (2002) (Figure 2.2), there are three dialect areas west of the *yat* line: transitional zone with Serbia (TZS), northwestern (NW) and southwestern (SW) dialects. While Stoykov names a number of features that distinguish these three dialect areas, none of them is recognized constantly on the computational maps. The transitional zone at the border with Serbia is present in most of the cluster analyses, but it is not recognized by UPGMA which is one of the most widely used hierarchical clustering algorithms. The northwest-southwest split shows even less stability and something that resembles this division is present only on the map drawn using k-means algorithm. We first look into the divergence between the computational and traditional maps with respect to the NW-SW division. After that we examine the issue of the instability of the transitional zone in some of the computational analyses.

4.2.1 Northwest-southwest split

As found in Houtzagers et al. (2010), in OT we find the following phonetic characteristics responsible for the NW-SW split:

- the reflexes of back *yer* in specific phonetic environments or in specific words
- the reflexes of front *yer*
- the reflexes of the back nasal
- presence or absence of mixture of reflexes of back and front nasal
- reflex of *yat* in цял /t͡sʲal/ ‘whole - masc sg’ and цели /ˈt͡seli/ ‘whole - pl’
- final [o] or [e] in such words as наше /ˈnaʃe/ ‘ours’
- presence or absence of the second [j] in яйце /jajˈt͡se/ ‘egg’

These features are present on 21 maps in OT and in 21 words in our data set. On the map in Figure 4.2 we present isoglosses based on the relevant segments from the 21 words from our data set.

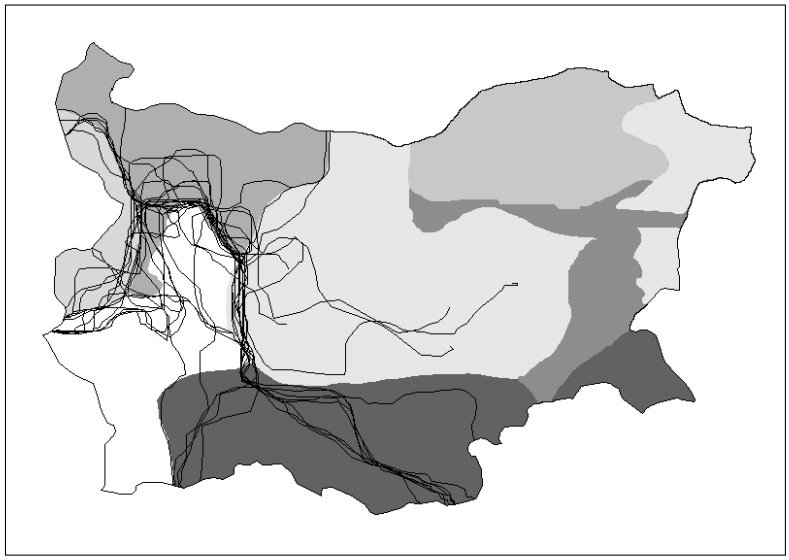

Figure 4.2: Isoglosses of the segments from 21 words that show NW-SW split.

The bundle of isoglosses separates northwest and the southwest areas, and additionally TZS. The same features also delineate western and eastern parts of the country along the *yat* line. In the north, the *yat* line is strengthened only by two of the features. This can be seen on the the map in Figure 4.2. The majority of the 21 features clearly delineates the southwest from the rest of the country, while the northwestern part shares many characteristics with the eastern part across the *yat* line. We find this type of distribution, for example, in the reflexes of the back nasal in words мъж /mɤʒ/ 'man', път /pɤt/ 'road' and събота /'sɤbota/ 'Saturday' where in the TZS the reflex is [u], in the SW it is [ɑ], and in the NW and in most of the parts east of the *yat* line [ɤ]. However, there are numerous features presented in the Section 4.1 that strengthen the *yat* line and make the west-east split undisputed on all our computational maps.

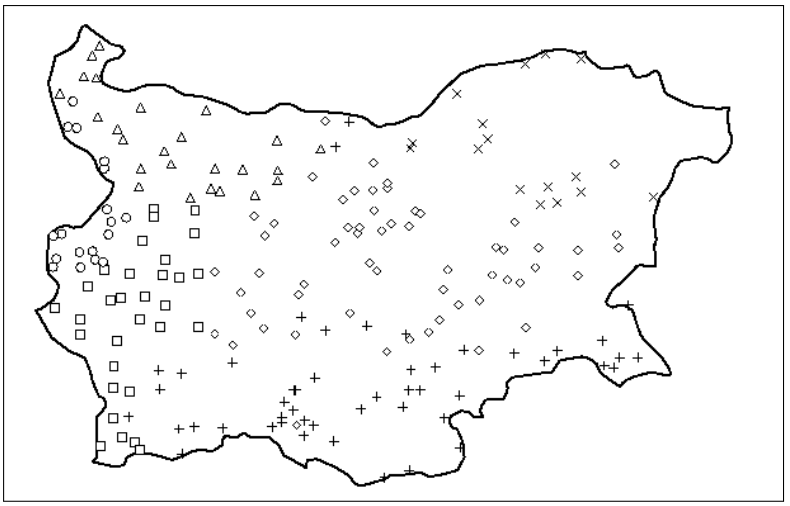

Figure 4.3: Stoykov's 6-way classification represented with different symbols.

Since there is a number of words that support the NW-SW split, we analyzed various MDS plots in order to try to explain the instability of this division on the quantitative maps. In MDS plots, we use different symbols to distinguish six traditional groups according to Stoykov (2002), while the linguistic distances obtained using the Levenshtein method are represented by the distance of symbols in a Cartesian coordinate system. In Figure 4.3, as well as in all MDS plots, we present Stoykov's six dialect areas using the following symbols:'○' for TZS, '△' for northwestern dialects, '□' for southwestern dialects, '◇' for Balkan dialects, '×' for Moesian, and '+' for Rupian dialects.

On the left in Figure 4.4 we present the MDS plot of the whole data set, 156 words and 197 sites, with all the sites placed into six groups according to Stoykov (2002). In the

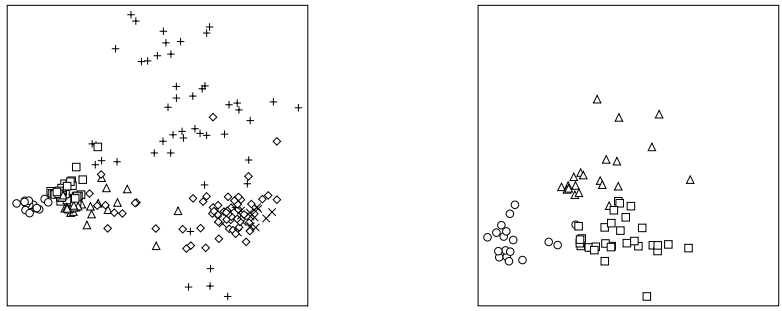

Figure 4.4: Left: MDS plot of 197 sites based on 156 words. Right: MDS plot of the 70 sites west of the *yat* line based on 156 words.

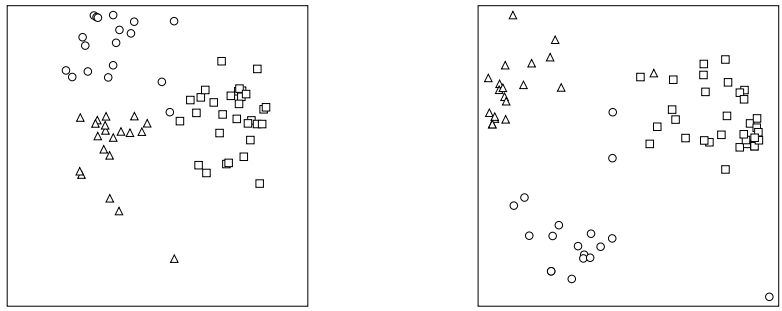

Figure 4.5: Left: MDS plot of the 70 western sites based on the 21 words selected. Right: MDS plot of the 70 western sites based on the specific segments from 21 words.

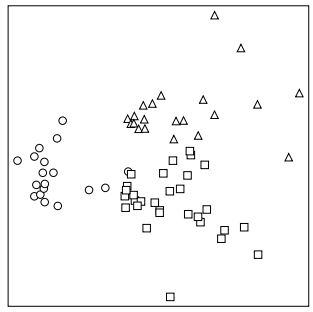

Figure 4.6: MDS plot of the 70 western sites based on the 135 words.

plot it is not possible to distinguish the two groups of symbols that represent the NW and SW varieties, since they form one compact group. We also note a number of Stoykov's Rupian and Balkan varieties that are put into the same group with NW and SW dialects in the computational analysis. The reason for this is that the computational division of the sites into eastern and western runs more to the east and includes parts of dialects that Stoykov classifies as Rupian and Balkan (Section 4.1). In order to investigate the division of the western varieties in more detail, we have removed all villages east of the *yat* line and repeated the analysis (MDS plot on the right in Figure 4.4). The MDS plot shows minor changes when compared to the previous one: it still remains very hard to distinguish the NW and SW varieties. These two groups are, indeed, more separate than on the previous plot, but the region between them is not empty. This means that while it is possible to distinguish north and south, the decision where to separate them would be arbitrary if we based our estimations of difference on aggregate Levenshtein distance.

In the two MDS plots in Figure 4.5 we examine the aggregate distances based on just 21 words in which the features relevant for NW vs. SW division appear, and also the aggregate distances based on just the single segments themselves. The left MDS plot in Figure 4.5 shows the Levenshtein distance based on 21 words without focusing on the relevant segments. Even if we base our analysis only on the words chosen, the two dialect varieties (NW and SW) are not clearly separated. NW and SW varieties are more distinct than in previous MDS plots, but there is still no clear separation between two clouds of symbols. However we note that varieties from the TZS do form a distinct group on this MDS plot, although our analysis is based on the features that are in traditional atlases specified as responsible primarily for the NW-SW split. We have also checked the distances between the western varieties based on the whole data set excluding the chosen 21 words. This analysis was performed in order to check whether the rest of the words would contain any conflicting signals with respect to the NW vs. SW division. As can be seen in Figure 4.6, the distances on the MDS plot are fully in accordance with the analysis based just on the 21 words chosen: TZS is a separate cluster, while there is no clear separation between NW and SW. When we base our analysis only on the specific sounds that Stoykov uses for distinguishing NW and SW dialects (right MDS plot in Figure 4.5) all three western varieties are clearly distinct. However, even in this focused view there are borderline cases shown by the single triangle within the group of squares (Kreta, Vrach), and the two circles which are closer to the squares than to the other circles (Buchin prohod, Sofia province, and Velkovtsi, Pernik province). If we use whole words instead of relevant segments, NW and SW varieties are not distinct since other segments in the words cloud the information provided by relevant segments. This also makes classification much more difficult since the separation between the varieties is less clear.

MDS plots have shown that in the aggregate analysis NW and SW varieties are not distinct even if the linguistic distances are based just on 21 selected words that contain features that do distinguish these two dialects. They become clearly separated only when

the analysis is based on the specific segments. We conclude that the features responsible for the traditional NW vs. SW division are present in the Buldialect data set, but in the aggregate analysis we do not find evidence that there is a categorical division between these two varieties.

4.2.2 Transitional zone

The transitional zone at the border with Serbia is recognized on most of the computational maps, but some clustering techniques, like UPGMA, fail to identify it as a separate zone. In Houtzagers et al. (2010) we find that the following characteristics present in OT maps distinguish TZS:

- the reflexes of back and front *yer* in specific phonetic environments or in specific words
- reduction or not of front *yer* in the suffix of such words as жаден /ˈʒaden/ 'thirsty'
- the reflexes of the back nasal in specific words
- reflexes of Old Bulgarian **tj*, **ktj* and **dj* in general and in specific words
- palatalized or nonpalatalized /l/ in such words as болна /ˈbolna/ 'ill - fem sg'
- labialization or not of /e/ in certain phonetic environments

These characteristics are found on 16 maps in OT. In our data set, we find 22 words in which these features are present. On the map in Figure 4.7 we draw isoglosses using relevant segments from each of those words. Most of the isoglosses drawn match almost perfectly Stoykov's TZS forming a bundle that delineates clearly this area from the surrounding varieties. We also note that the isoglosses drawn using 21 words that delineate NW and SW also clearly distinguish TZS as a separate area. To check the instability of this area on some of the computational maps we reexamine the two MDS plots in Figure 4.4. The left plot shows the distances among all the sites in the data based on 156 words, while on the right hand side we show a plot of the distances among the 70 sites west of the *yat* line. On both MDS plots group of circles that represents Stoykov's TZS forms a separate group with some intermediate varieties between the TZS and SW. The villages Buchin Prohod, Elov Dol and Velkovtsi, all classified as TZS by Stoykov, are closer to the SW varieties than to the rest of the TZS in our quantitative analysis.

With respect to our question concerning the reason for differences between the quantitative and the traditional maps we conclude that the Levenshtein method separates the TZS varieties from the other western varieties, but some clustering techniques fail to recognize this. Some pairs of sites, one from each area, remain very close in aggregate Levenshtein distance. In the data set there are 22 words that show features described by Stoykov (2002) as characteristic for this dialect. Additionally, 21 words which contain

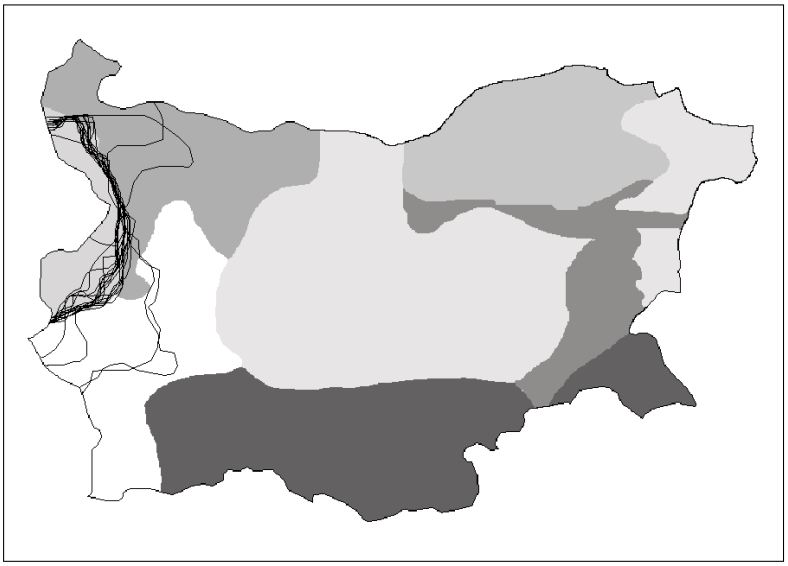

Figure 4.7: Isoglosses based on relevant segments from 22 words that delineate TZS.

features responsible for the NW vs. SW division also distinguish TZS from the rest of the western varieties. Despite the substantial number of relevant words in the data set, in the aggregate analysis the distance between TZS and the rest of the western varieties is not large and also contains intermediate varieties. This poses problems for some clustering techniques, like UPGMA, that fail to recognize this area as a separate dialect zone. Another reason for the poor performance of UPGMA in this case could be the fact that the results of this clustering technique could be distorted during the fusion of the large group of objects with the small group of objects (see Chapter 3), since there is a significant difference in the number of objects belonging to TZS and the rest of the western varieties.

4.3 Eastern dialects

In the east, i.e. east of the *yat* line, all computational maps distinguish the area in the south that corresponds well with Rupian dialects and the large area in the north that comprises the Balkan and Moesian dialects as defined on the traditional maps. In this subsection we address both of these issues.

4.3.1 Rupian dialects

Rupian dialects are detected on all the quantitative maps presented in Chapter 3, except for the three algorithms that did not identify any groups in the data. MDS analysis has shown that this is one of the three main dialect areas that can be asserted with some confidence. Moreover, it has been shown that this is the most heterogeneous area, not only in the east, but with respect to all other dialect zones in Bulgaria identified by computational methods. As found in Houtzgers et al. (2010) the same picture can be found on maps in OT: there are many maps on which this area is distinct from the surrounding varieties, but there is also a substantial number of maps where this applies only to part of Rupian dialects. Many characteristics are shared between parts of the Rupian area and areas outside this territory, especially in the northeast. For example, on maps OT F 40-46, which show reflexes of ě *yat* in word две /dve/ 'two', and in certain verbal endings there is a geographically variable central area within Rupian that differs from its immediate surroundings but shows similarities with varying subareas mostly in the east and northeast.[3] There are also maps on which a larger part of the southeastern area is distinguished from the northeast. Following Houtzagers et al. (2010) we give two examples:

1. OT F 9: presence of epenthetic [ə] in such *l*-participles as Standard Bulgarian пекла /ˈpekla/ 'bake - fem 1st sg' ([ˈpekla] vs [ˈpekəla). This characteristic is shared by most (but not all) of the southeast and two noncontingent areas in the northeast.

2. OT F 19: absence of a vowel in the verbal root **mъk-* (Old Bulgarian) 'weave'. The whole southeast is opposed to the northeast here, but it shares its characteristic with the entire west.

In the *Buldialect* data set, we also find numerous words which contain the features that are shared between Rupian and eastern (Figure 4.1), or Rupian and western varieties (Figure 4.2). As a result, in the aggregate analysis this area is more diffuse than eastern or western varieties and lies at a remove from them. This can be clearly seen on the right MDS plot in Figure 4.4 (see the higher part of the plot). In our data set we find 31 words which contain features characteristic for the language varieties in the area of the Rodopi mountains. Isoglosses based on the specific segments, drawn using white lines, can be seen in Figure 4.8.

Regarding Rupian dialects we find fairly high correspondences between computational and traditional maps. This area is identified both on MDS plots and in various maps obtained using clustering techniques. The substantial number of words which delineate this area from the surrounding territories is reflected in relatively clear separation of this area in MDS analyses.

[3] OT F is used to refer to the maps in OT that regard phonetics.

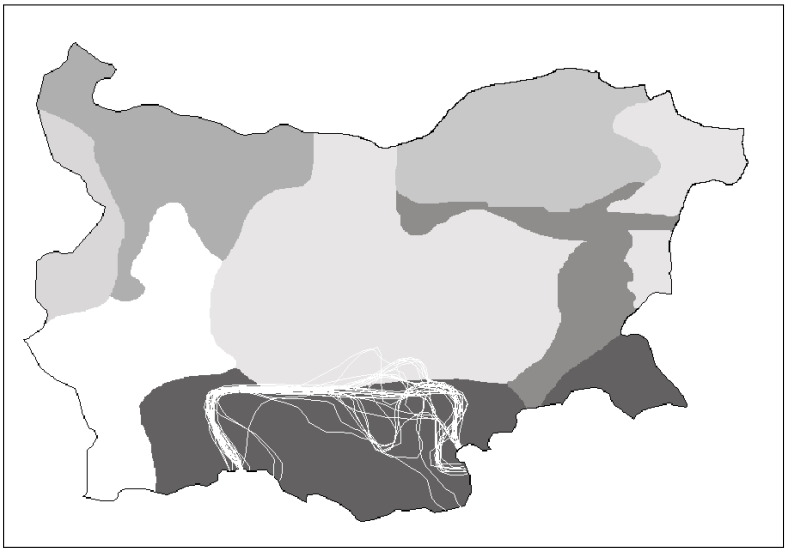

Figure 4.8: Isoglosses based on the relevant segments from 31 words that distinguish Rupian varieties.

4.3.2 Moesian dialects

Unlike Rupian dialects, the Moesian area as defined in Stoykov (2002) does not appear on any of the computational maps. Stoykov mentions four phonetic characteristics of this area:[4]

- velarized realization of the Old Bulgarian back *yer* in a stressed position
- In stressed syllables, the reflexes of Old Bulgarian vowel *ě (*yat*) before hard syllable is [ʲa] and before soft syllable is [ɛ] ([bʲal] vs. [bɛli]). Under the influence of the Balkan dialects [ɛ] is almost completely replaced by [e].
- change of consonant /d/ into [n] before /n/ (*$dn >$ [nn])
- non-existence of consonants /f/ and /x/

Three of these distinguishing characteristics are not supported by (his own) OT and BDA maps.[5] Velarized pronunciation of the back *yer* is found neither in OT nor in BDA.

[4]Repeated from the Section 2.3 for the convenience of the reader.

[5]BDA: *Български диалектен атлас.* [Atlas of Bulgarian dialects] (Stoykov and Bernstein, 1964; Stoykov, 1966; Stoykov et al., 1974; Stoykov, Kochev, and Mladenov, 1981)

It is also not present in our data set. Regarding the reflexes of the *yat* and the $*dn > $ [nn] change, the maps in OT (OT F 35 and OT 166 respectively) show that these characteristics are not typical only for the Moesian area, since they spread far outside the area labeled as Moesian by Stoykov. Characteristics mentioned are common to almost the whole area east of the *yat* boundary. With respect to the fourth characteristic, nonexistence of /f/ and /x/, on some maps in OT (135-141) it is possible to distinguish an area that corresponds to Stoykov's Moesian dialects. However, the relevant characteristic is often shared with the areas to the east, west, or south. In the data set there are 23 words that contain this feature, but only 15 of them show an isogloss that runs more or less along the boundary of Stoykov's Moesian area. In Figure 4.9 we present isoglosses drawn using only relevant segments from those 15 words.

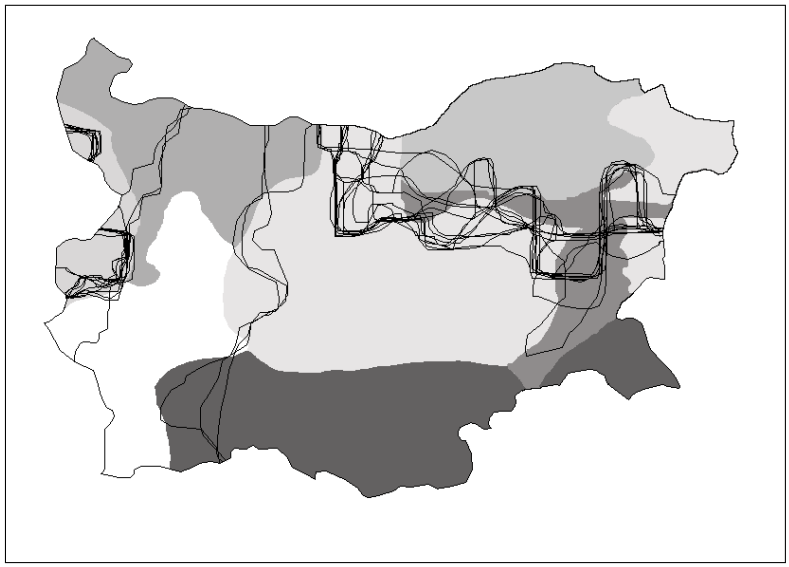

Figure 4.9: Isoglosses drawn using segments from 15 words in the data set where features that distinguish Moesian zone are present.

Even if we focus on the relevant segments, the isoglosses do not delineate only Stoykov's Moesian area, but also other parts of Bulgaria as well. The MDS plot in Figure 4.10 confirms that in the aggregate analysis based on the chosen 15 words, Stoykov's Moesian area is not distinguishable. In this MDS plot sites that belong to the Stoykov's Moesian area ('$\times$' sign) are concentrated in the right low corner of the plot, together with the Balkan varieties ('$\diamondsuit$' sign). It is not possible to detect a separate cloud representing Moeasian varieties since two groups of signs are mixed. On the right MDS plot

in Figure 4.10 we show aggregate analysis based on relevant segments from 15 chosen words. Moesian and Balkan varieties, concentrated in the left upper corner are to some extent more separated from the rest of the varieties than on the previous plot. However, the two groups of symbols are mixed and cannot be separated from each other. It is clear that 15 chosen segments are not distinctive for Stoykov's Moesian area, but are shared with a considerable number of sites from the Balkan dialects.

We conclude that as far as phonetics is concerned there is not enough evidence that Moesian area should be treated as a separate dialect. Most of the phonetic characteristics that traditional literature considers typical for this region is actually shared with the neighboring Balkan dialects. Using only relevant segments from 15 words that show nonexistence of /f/ and /x/ we manage to detect a very weak signal that distinguishes northeastern area but broader than suggested by Stoykov. The strength of this signal is lost when the data as a whole is taken into account.

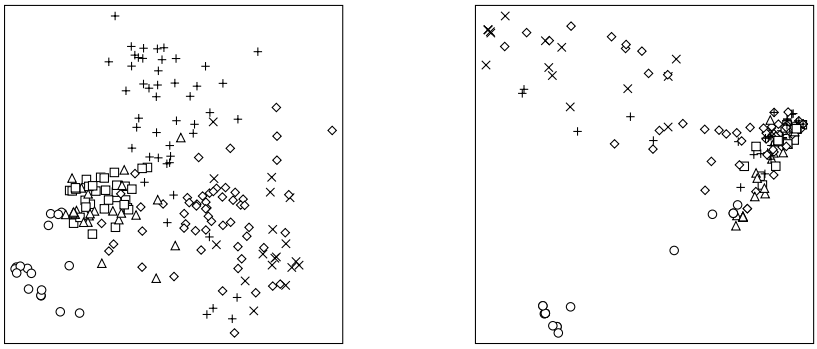

Figure 4.10: Left: MDS plot made using the chosen 15 words. Right: MDS plot made using only relevant segments from the chosen 15 words. The Moesian dialects, symbolized by '$\times$', do not emerge coherently.

4.4 Discussion

Our goal in this chapter was to compare traditional and quantitative classifications of Bulgarian dialects. We drew on Stoykov's authoritative work for our views on traditional classification, and we used a simple version of Levenshtein distance to provide a base for a quantitative view. The general lines of the two views of the Bulgarian dialect landscape are similar. Both see the language area dominated by an east-west division, i.e. Stoykov's *yat* line, and both identify the Rupian south as a third most significant area. The quantitative work located the *yat* line slightly to the east of where Stoykov had drawn it, and it failed to identify anything like his Moesian area. In both of these cases

we find for the quantitative work, and conclude that it improves on Stoykov's. Assuming that Levenshtein distance is yielding a probative measure of aggregate pronunciation differences, we relied on multidimensional scaling (MDS) to visualize the more than 19,000 distances between the pairs in our 197-site sample, encouraged by the fact that over 92 per cent of the variation is captured in the first two dimensions. This allowed us to see that the Rupian area is much more diverse than either the east or the west in the north.

Regarding the situation in the west, the MDS plot demonstrates that the transitional zone at the border with Serbia, the northern and southern parts of the west, all of which Stoykov postulated, may indeed be distinguished when using aggregate pronunciation distance, but the borders are not linguistically prominent. It is not surprising that clustering fails to distinguish these areas reliably.

We noted above that most of the work presented here proceeds from the assumption that Levenshtein distance is a valid measure of the pronunciation differences found in dialects. Naturally this assumption may be questioned: for example, the built-in sensitivity to segment frequency in Levenshtein distance may be inappropriate. For example, for the most prominent division into the east and west, we find 68 relevant words, for the TZS we find 41 words, while for the Moesian area we find only 15 words that contain relevant features. It is evident that the clearer the separation of an area is, the bigger the number of relevant words in the data set. While traditional dialectologists often use their own intuition in giving certain features more weight, our aggregate method treats all features as equal and tries to infer dialect divisions based on all features in the data set. In our data set we are not able to determine if the distribution of chosen features corresponds well with their distribution in Bulgarian language. However, as described in Chapter 2, the data was collected in a such way that there is a balance between various phonetic features, which ensures that the data set is not biased towards certain phonetic phenomena and as a consequence certain dialect divisions.

Computational measures of pronunciation differences may be modified in many ways. While in the research addressed in the current chapter we have applied the simple version of the Levenshtein algorithm and represented every segment as a distinct unit that is not further defined, in Chapter 5 we automatically infer the distances between the segments in the data set and use that information to get more accurate alignments and consequently more accurate distances between the sites.

Detailed comparison between computational and traditional maps has shown that the features responsible for traditional divisions of Bulgarian dialect varieties are well represented in our data set. The simple version of Levenshtein algorithm was successful in identifying three main dialect groups and in showing that Moesian area cannot be identified as a separate dialect purely based on phonetic evidence. We see a three-way division in the west of Bulgaria reflected in MDS plots, but not distinctly enough to be detected reliably by clustering. In the next chapter we show how Levenshtein approach can further be improved by introducing segment distances in the alignment procedure.

The instability of clustering techniques poses a problem in dialect data classification and we argue that MDS is more reliable in the analysis of dialect varieties.

Chapter 5

Segment distances

In this chapter we apply pointwise mutual information (PMI) in order to automatically acquire segment distances from the phonetic transcriptions. Information on the distances between the phones can help us estimate more precisely the distances between two strings and consequently the distances between two language varieties. Instead of using only same vs. different as a comparison between the phones, we can use information on the phone distances together with Levenshtein algorithm in order to get better distances and better alignments (Chapter 3). There are alternatives to our empirically deriving segment distances from dialect atlas samples. The distances between the phones can also be calculated using a linguistically more informed approach by representing each phone as a bundle of features where every feature is a certain phonetic property (Heeringa, 2004). The distances can also be measured acoustically, which is less arbitrary than using feature representation of phones since it is based on physical measures (Heeringa, 2004). However, both of these approaches have their disadvantages. The former relies on a language-dependent feature system, while the letter requires acoustic data to be available. Since very often neither of the two is available, we propose a technique to acquire the distances automatically. Similar research was presented in Wieling et al. (2007) where the distances between the phones were automatically acquired using pair hidden Markov models (PHMM). As reported in Wieling, Prokić, and Nerbonne (2009), where both PMI and PHMM techniques were applied on the same data set used in this thesis, they produce pairwise alignments of a very similar quality. However, PMI is much faster, and the alignment errors made by this algorithm are *a priori* predictable and much easier to comprehend than errors induced by PHMM algorithm. Part of the work presented in this chapter was published in Wieling, Prokić, and Nerbonne (2009).

5.1 Pointwise mutual information

Pointwise mutual information (PMI) is a measure of association between two events x and y. It measures the amount of information one event tells us about the other. It was first introduced by Fano (1961). Given a pair of outcomes x and y, the pointwise mutual information I is measured as:

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)} \tag{5.1}$$

The numerator $P(x,y)$ tells us how often we have observed the two events together, while the denominator tells us how often we would expect these two events to occur together assuming that they each occurred independently. The ratio between these two shows us if two events co-occur together more often than just by chance. Positive values of I show that there is a genuine association between x and y. This measure was first used in computational linguistics by Church and Hanks (1989) for calculating associations between words.

In this research, PMI is used to automatically learn the distances between the phones in aligned word transcriptions and also to improve the automatically generated alignments. Equation 5.1 is used to calculate PMI values for each pair of segments in alignments, and later these values are transformed into segment distances (see below). Applied to aligned transcriptions, $P(x,y)$ represents the relative frequency of two segments being aligned together and is calculated by dividing the number of times two phones were aligned together by the total number of aligned segments in the data set. $P(x)$ and $P(y)$ are relative frequencies of segments x and y—the number of times segments x and y occur in the data set divided by the total number of segments.

The procedure of calculating the segment distances and improving the alignments is iterative and consists of the following steps:

1. Align all word transcriptions using Levenshtein algorithm where only the vowel-vowel consonant-consonant constraint is given. No detailed information on the distances between the segments is provided in this step.

2. From the obtained alignments, for all pairs of segments calculate PMI values $I(x,y)$ using formula 5.1.

3. Transform PMI values into distances by substracting each value from 0 and normalize the distances to insure that the smallest distance is 0.

$$dist(x,y) = \frac{(0 - I(x,y)) - min}{max - min} \tag{5.2}$$

where *min* and *max* are the minimal and maximal values obtained after substracting PMI values for all pairs of segments from 0, and $I(x,y)$ is given in 5.1.

4. For all pairs of segments that never align in the data, set the distance to an arbitrary large value.[1]

5. Align all word transcriptions once more using Levenshtein algorithm, but based on the segment distances generated in the previous step.

6. Repeat steps 2 and 3 until there are no changes in segment distances and alignments.

The final result are distances between each two segments as well as the alignments that show improvement when compared to the alignments obtained by using the Levenshtein algorithm with only the vowel-vowel consonant-consonant constraint (Levenshtein VV-CC). To illustrate how the algorithm works, we will examine two pronunciations of the word *днес* /dnes/ 'today': [deˈneskɑ] and [ˈnʲeskə]. In the first step we align these two pronunciation using the simple Levenshtein algorithm. Using only the vowel-vowel consonant-consonant constraint, there are two different alignments of these two strings that have the same minimal cost:[2]

d	e	n	ˈe	s	k	ɑ		d	e	n	ˈe	s	k	ɑ
-	-	nʲ	ˈe	s	k	ə		nʲ	-	-	ˈe	s	k	ə

Figure 5.1: Two alignments produced by the Levenshtein VV-CC that have the same cost of 4.

In the next step we use these alignments to calculate the PMI values for all pairs of aligned segments.

Table 5.1: Values for the [n], [nʲ] and [d] segments calculated from the alignments produced by the Levenshtein VV-CC.

x	y	$f(x)$	$f(y)$	$f(x,y)$	$P(x)$	$P(y)$	$P(x,y)$	$I(x,y)$	$dist(x,y)$
n	nʲ	1250463	268725	135477	0.038178	0.008205	0.008273	4.723003	0.199076
d	nʲ	1228146	268725	223	0.037497	0.008205	0.000013	-4.497805	0.601548

The total number of segment pairs in the data set was 16376419 and the total number of segments 32752838. In Table 5.1 we present the frequencies (f) and relative frequencies (P) for two pairs of sound correspondences: [n]-[nʲ] and [d]-[nʲ]. By multiplying relative frequencies of two segments $P(x)$ and $P(y)$ we can see the probability of seeing two segments aligned by chance.

[1] In our experiment it was set to 1000.

[2] The sign for primary stress is moved to the first vowel in stressed syllable in all examples presented in this chapter.

$$P(n) \times P(nʲ) = 0.038178 \times 0.008205 = 3.13 \times 10^{-4}$$

$$P(d) \times P(nʲ) = 0.037497 \times 0.008205 = 3.07 \times 10^{-4}$$

For the first pair, [n]-[nʲ] this value is 0.000313, and for the second 0.000307. If we compare these values to the relative frequencies of each of the two segments being aligned together in our data set $P(x,y)$, we see that the segments [n]-[nʲ] were aligned 28 times more than we would expect by chance, since $P(n,nʲ)$ is 0.008273. For the pair [d]-[nʲ] the situation is opposite: they align 22 times less than we would expect by chance ($P(d,nʲ) = 0.000013$).

For all pairs of segments, we use information on the relative frequencies to obtain pointwise mutual information (formula 5.1), which shows negative association for the [d]-[nʲ] pair. PMI values are transformed into the distances using formula 5.2. We find that the distance between [n]-[nʲ] is 0.199076 and between [d]-[nʲ] 0.601548.

In the next step we align two strings using Levenshtein algorithm based on the calculated distances between the segments and recalculate the distances between the segments based on the new alignments. These steps are repeated until there are no more changes in the distances between the segments. The final values for the segments [n]-[nʲ] and [d]-[nʲ] can be seen in Table 5.2.

Table 5.2: Values for the [n], [nʲ] and [d] segments calculated from the alignments produced by the Levenshtein PMI.

x	y	$f(x)$	$f(y)$	$f(x,y)$	$P(x)$	$P(y)$	$P(x,y)$	$I(x,y)$	$dist(x,y)$
n	nʲ	1193681	261135	134527	0.037449	0.008192	0.008441	4.782038	0.196556
d	nʲ	12281456	261135	0	0.038530	0.008192	0	0	1000

Based on the calculated distances, where the distance between [n]-[nʲ] is much smaller than between [d]-[nʲ], the outcome of the Levenshtein algorithm is only one alignment of the strings [denˈeskɑ] and [nʲˈeskə] (Figure 5.2). The final distance between [n]-[nʲ] was reduced to 0.196556, while the distance between [d]-[nʲ] was set to 1000, i.e. an arbitrary large value, since in the improved alignments these two segments are never aligned.

d	ɤ	n	ˈɤ	s	k	ɑ
-	-	nʲ	ˈe	s	k	ə

Figure 5.2: Alignment of the strings produced by Levenshtein PMI.

The distances among vowels and consonants are all set to an arbitrary large value since they never align in our alignment procedure. In the first step of the procedure we use the Levenshtein algorithm with the constraint that the vowels and consonants cannot align. Without this constraint, the Levenshtein algorithm produces several alignments for many pairs of transcriptions. Since only one of them is correct, this means that in the first step of our PMI procedure we would get a large number of erroneous alignments. Segment distances induced from such a large number of erroneous alignments are themselves erroneous and they cannot improve the quality of the alignments if used within the Levenshtein algorithm.

Using the PMI procedure we have managed to automatically infer the distances between the segments, but also to improve the quality of the alignments as we shall show in the next section. In Section 5.2 we present the results of evaluating pairwise aligned strings obtained using the Levenshtein algorithm with and without the PMI procedure on the segment level. In the Section 5.3 we analyze the automatically acquired distances between the tokens using multidimensional scaling in order to check if they correspond well with our linguistic knowledge on the distances between the phones. We also investigate the influence of the automatically acquired segment distances on the aggregate analysis of dialect divisions and report on our findings in Section 5.4. A short discussion on the merits of PMI in dialectometrical research is presented in Section 5.5.

5.2 Evaluation of the pairwise alignments

In this section we describe a method for quantitatively evaluating pairwise aligned strings and report on the quality of the alignments obtained using the Levenshtein algorithm with and without segment distances induced using PMI. The comparison of the two techniques was done by comparing each of them, on the segment level, to the gold standard pairwise alignment. We also report on the qualitative analyses of the alignments produced.

The gold standard alignment was generated from the gold standard multiple alignments described in Section 6. The gold standard multiple alignments were automatically generated using some heuristics and later manually corrected (for the details see Chapter 6). They consist of all pronunciations for a single word aligned simultaneously, instead of aligning it pair-by-pair which is done using pairwise-aligning algorithms like Levenshtein. Using multiple aligned strings it was possible to manually go through the whole data set, since this technique gives us 156 files with approximately 200 aligned strings. In pairwise approach each of 156 files contains around 12090 pairwise alignments which would be very time consuming to correct manually. Since for 4 out of 156 words experts could not agree on what a correct alignment is, those 4 entries were removed from the data set. An example is word ɵʊɵ/vɤv/ ‘in’. Some of the dialect variants of this word contain only one segment [v] which could be aligned with two segments in other transcriptions, but the experts could not agree which of the two is more likely to

be the correct one (Figure 5.3). Since it was very difficult for humans to make a decision which alignment is the correct one, this word, as well as three others that posed similar problems to the human experts, was left out of the evaluation procedure.

v	ˈɤ	v		v	ˈɤ	v
v	-	-		-	-	v

Figure 5.3: Two possible alignments of the word ‘in’ on which experts could not agree.

All further analyses were done on 152 words for which the gold standard alignments were available. Out of the manually corrected multiple string alignments we have extracted all pairwise alignments and used them as a gold standard to evaluate the results of Levenshtein PMI.

d	ɤ	n	ˈɤ	s	k	ɑ		d	ɤ	n	ˈɤ	s	k	ɑ
-	-	nʲ	ˈe	s	k	ə		nʲ	-	-	ˈe	s	k	ə

Figure 5.4: The gold standard alignment on the left and the alignment produced by Levenshtein VV-CC on the right.

The evaluation procedure consists of the following steps:

1. For each pair of aligned strings, take every pair of aligned segments and convert the pair into a single token. For example, the first two aligned strings in Figure 5.4 would give the following tokens: d/-, ɤ/-, n/nʲ, ˈɤ/ˈe, s/s, k/k, ɑ/ə.

2. Concatenate all tokens obtained into a single string. Segments generated in Step 1 would give the following string for the first alignment: d/- ɤ/- n/nʲ ˈɤ/ˈe s/s k/k ɑ/ə.

3. Use the Levenshtein algorithm without any restrictions on segment distances and align corresponding strings, i.e. transformed strings generated by Levenshtein VV-CC and Levenshtein PMI against the gold standard alignments. Since there are no restrictions on segment distances, two segments match only if their both parts match. For example, the distance between two generated strings from Figure 5.5 would be 2:

d/-	ɤ/-	n/nʲ	ˈɤ/ˈe	s/s	k/k	ɑ/ə
d/nʲ	ɤ/-	n/-	ˈɤ/ˈe	s/s	k/k	ɑ/ə
1		1				

Figure 5.5: Levenshtein distance between these two strings is 2.

4. The distances for all alignments are automatically calculated using Levenshtein algorithm and summed up giving the total distance between alignments produced by two versions of Levenshtein algorithm and the gold standard.

5.2.1 Results

The quantitative results of the evaluation can be seen in Table 5.3. We report the error rate at the segment level and the percentage of missaligned strings. The error rate in second column represents the number of incorrectly-aligned segments divided by the total number of aligned segments in the gold standard. In the third column we report the percentage of the strings that are not aligned in the same way as found in the gold standard alignments. In both cases Levenshtein PMI outperforms Levenshtein VV-CC algorithm. On the segment level, error rate drops from 0.040 to 0.032, while at the word alignment level error of 7.614 per cent in the basic algorithm improved to 6.263 per cent when including the PMI-derived distance. The difference is statistically significant with $p < 0.001$ by the exact binomial test.

Table 5.3: Comparison of the alignments generated by Levenshtein VV-CC and Levenshtein PMI algorithms to the gold standard alignments.

Algorithm	Error rate for segments	Incorrect alignments(%)
Levenshtein VV-CC	0.040	7.614
Levenshtein PMI	**0.032**	**6.263**

The qualitative error analysis has shown that most of the errors arising using the simple Levenshtein algorithm come from the constraint that vowels and consonants cannot be aligned. Although this holds in most of the cases, there are, however, exceptions where vowels should be aligned with consonants. An example of these types or error can be seen in the alignments where metathesis is present. Metathesis is a change where sounds switch their places within a word (for example [vɤrx] vs. [vrɤx]). Metathesis of liquid consonants is an important historical change in Slavic languages and is present in 18 words from our data set (11.84 per cent of the data). More on the metathesis in Slavic languages can be found for example in Sussex and Cubberley (2006). Due to a VV-CC constraint, this poses a problem for the Levenshtein algorithm. Instead of aligning a vowel with a consonant, additional gaps are introduced by Levenshtein algorithm (Figure 5.6).

Since the PMI alignment procedure proceeds from the Levenshtein VV-CC algorithm, a vowel can also not be aligned with a consonant. For that reason, this type of error is also present in the alignments produced with Levenshtein PMI algorithm.

Another type of error detected in the alignments produced by Levenshtein VV-CC

v	ˈɤ	r	x
v	r	ˈɤ	x

v	ˈɤ	r	-	x
v	-	r	ˈɤ	x

Figure 5.6: The gold standard alignment on the left and the erroneous alignment produced by Levenshtein on the right.

arises in cases when one vowel (consonant) has to be aligned with one of the two adjacent vowels (consonants). Since the distance between all vowels on one hand, and all consonant on the other is the same, the algorithm often yields erroneous alignments. For Levenshtein VV-CC algorithm both alignments in Figure 5.7 are correct since the distance between two strings is 3.

v	ˈɤ	n	-	-
v	ˈɤ	ŋ	k	ə
		1	1	1

v	ˈɤ	-	n	-
v	ˈɤ	ŋ	k	ə
		1	1	1

Figure 5.7: The alignments produced by Levenshtein VV-CC algorithm: the correct one on the left and the erroneous one on the right.

Unlike Levenshtein VV-CC, Levenshtein PMI algorithm generates only the correct alignment since it 'learns' that the distance between [n] and [ŋ] is smaller than the distance between [n] and [k]. Correction of these types of errors is where the PMI procedure improves the performance of simple Levenshtein VV-CC algorithm and generates more correct alignments.

5.3 Analysis of segment distances

Comparison of the alignments produced using the Levenshtein VV-CC and the Levenshtein PMI to the gold standard alignments has shown that the PMI procedure can improve the quality of the obtained alignments. We were also interested in the nature of the automatically obtained segment distances. In order to check whether they reflect any of the 'traditional' phonetic features of language sounds, we have performed MDS analysis (Chapter 3) of the phone distances calculated using PMI.

In Figure 5.8 we can see two-dimensional plot of all the sounds in the data set. The first extracted dimension, plotted against the x-axis explains 11.69 per cent of the variation, and the second dimension, plotted against the y-axis, explains 4.06 per cent of the variation. Along the x-axis there is a clear separation between vowels and consonants. This was expected since in our PMI procedure vowels and consonants cannot be aligned and the distances between them were set to an arbitrary large value. More interesting is the variation along the y-axis. Along the y-axis we can see that the distances

between vowels are much smaller, (in fact there is almost no variation), than the distances between the consonants. It means that in our data set vowel changes are much more frequent than consonant changes, since the more often two tokens correspond the smaller the PMI distance between them.

In order to analyze the distances more accurately, we performed MDS analyses separately for vowels and consonants. In Figure 5.9 we present the plot of all vowels in the data set, with all the diacritics preserved. With all the consonants removed, it is possible to analyze the relationship between the vowels in more depth. The first two extracted dimensions explain 16.06 per cent of the variation, with the first one explaining 10.39 per cent. With a very few exceptions, along the x-axis there is a separation between stressed and unstressed vowels. The distance between stressed vowels is larger than that between the unstressed, meaning that unstressed vowels are more similar than stressed vowels. In the upper right corner of the MDS plot we have a group of front vowels, while in the opposite, low left, corner there is a cluster of back vowels. We note that the separation between front and back vowels does not go along x- or y-axis, since first two dimensions extracted by MDS do not correspond to any of the two most prominent oppositions based on articulatory features of vowels—back/front or open/close opposition. However, it is still possible to distinguish front/back vowels contrast. To check this we have extracted all vowel correspondences from the aligned transcriptions. In Table 5.4 we present the 10 most frequent. We can see that among the most frequent correspondences we indeed do have neutralization of the contrast of vowel height, [e]-[i], [ɑ]-[ə], [o]-[u]. Since these correspondences occur more frequently than others, the distances between these phones calculated using PMI are small and in MDS analysis they are not separated by any of the first two dimensions. These findings conform with the traditional Bulgarian phonology scholarship according to which the elimination of the contrast of vowel height in unstressed vowels is the most common vowel reduction phenomenon in Bulgarian (Wood and Pettersson, 1988; Barnes, 2006).

In Figure 5.10 a MDS plot of consonant distances is presented. The first two extracted dimensions explain only 6.68 per cent of the variation, the first dimension explains 3.53 per cent and the second 3.15 per cent of the variation. The main division goes along the y-axis where in the upper part we have mostly plosives and sonorants and their palatalized counterparts. The distances between them are smaller than between the segments in the lower part, mostly fricatives, indicating that palatalization of consonants is the most frequent consonant variation in our data set. It can be seen in Table 5.5 where we present the 30 most frequent consonant correspondences in the data set. Unlike vowels, consonants show much less variation and in the 10 most frequent correspondences there are no consonant changes. In the 30 most frequent correspondences extracted, the most frequent consonant change is the insertion/deletion of [j], followed by the palatalization of [n], [r] and [l].

The analyses of vowel and consonant distances obtained using Levenshtein PMI have shown that these distances correspond to a certain extent to the vowel and consonant

Table 5.4: Ten most frequent vowel correspondences in the data set. Note that there are only three pairs of non-identical vowels: [e]-[i], [ɑ]-[ə] and [o]-[u].

Number of occurrences	Vowel pair
592274	[ˈe]-[ˈe]
497495	[ˈɑ]-[ˈɑ]
371146	[ˈo]-[ˈo]
287243	[e]-[i]
273473	[e]-[e]
257192	[ɑ]-[ə]
225142	[ə]-[ə]
214763	[o]-[u]
211673	[u]-[u]
204639	[i]-[i]

characteristics we know from phonetic and phonological theory. Unfortunately it was not possible to obtain data that would contain acoustic distances between the segments for Bulgarian and compare it directly to the automatically induced distances.

5.4 PMI and the aggregate analysis of dialects

In Section 5.2.1 we have shown that PMI can improve the quality of the alignments produced using Levenshtein VV-CC algorithm. In this section we examine if this improvement will show in the analysis of dialect divisions at the aggregate level. We analyze distances between the sites obtained using Levenshtein PMI with MDS and compare the results to the divisions obtained using Levenshtein VV-CC.

All analysis for Levenshtein PMI and Levenshtein VV-CC were done on 152 words, used also to evaluate the alignments on the segment level. We calculated Pearson's correlation coefficient between distance matrices obtained using these two versions of Levenshtein algorithm and found that they correlate to a high extent, namely the coefficient is $r = 0.98$.

In Figure 5.11 we present MDS maps based on the Levenshtein VV-CC and the Levenshtein PMI distance matrices next to each other. Although the two distance matrices correspond highly, there are differences in two MDS maps in Figure 5.11, most notably in the western part of the country. On the left map derived from Levenshtein VV-CC, this part of the country forms a homogeneous area, with no distinct groups. The map produced using Levenshtein PMI distances distinguishes the transitional zone at the border with Serbia as a separate group. At the same time, there is some distinction between northwestern and southwestern areas, that cannot be detected on the other map.

Table 5.5: Thirty most frequent consonant correspondences in the data set. Note that there are 10 pairs of non-identical consonants.

Number of occurrences	Vowel pair
626676	[r]-[r]
595761	[t]-[t]
524857	[d]-[d]
517440	[s]-[s]
503480	[n]-[n]
471794	[k]-[k]
423509	[v]-[v]
374155	[m]-[m]
326280	[l]-[l]
261019	[b]-[b]
237165	[g]-[g]
233534	[ʃ]-[ʃ]
222479	[p]-[p]
212851	[j]-[-]
202390	[ʧ]-[ʧ]
188601	[j]-[j]
150620	[z]-[z]
136781	[f]-[f]
134527	[n]-[nʲ]
133659	[r]-[rʲ]
127755	[ʒ]-[ʒ]
126525	[ʦ]-[ʦ]
114261	[d]-[dʲ]
102062	[l]-[lʲ]
93188	[v]-[vʲ]
79601	[v]-[-]
68514	[lʲ]-[lʲ]
65616	[ˈr̩]-[r]
64385	[t͡ɕ]-[ʧ]
61771	[ʃ]-[ɕ]

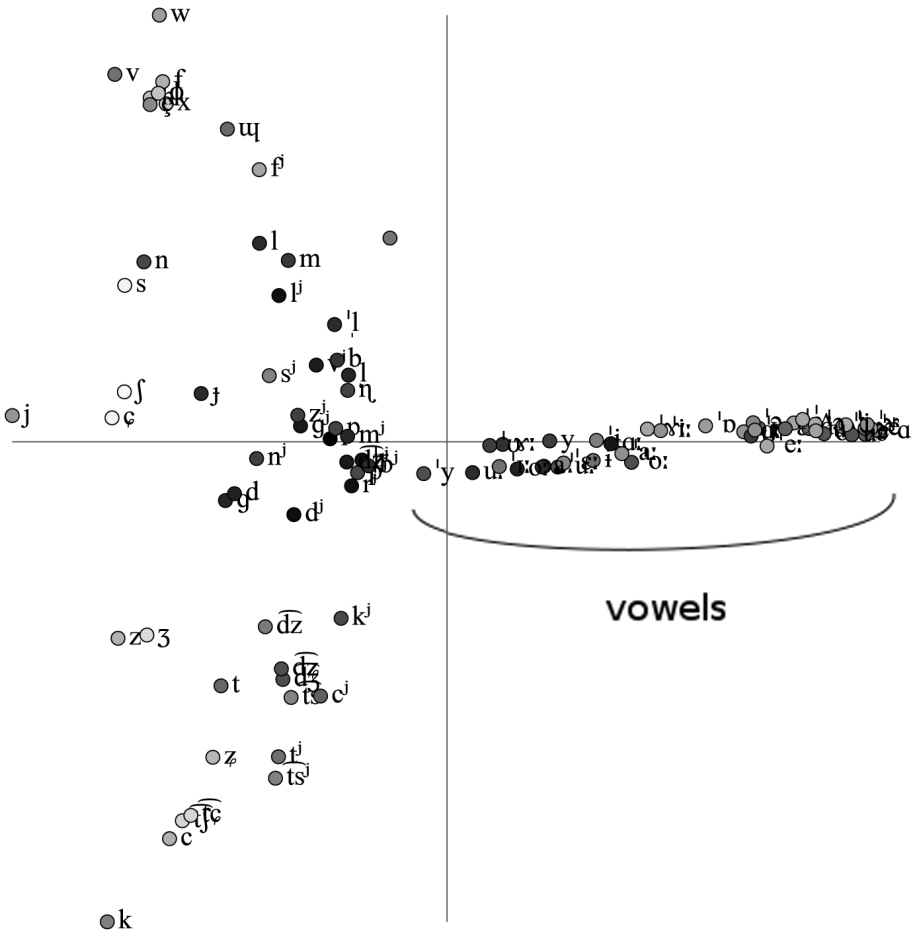

Figure 5.8: MDS plot of all phones in the data set. The distances between the vowels and the consonants are set to an arbitrary large value, which resulted in the clear separation between them along the x-axis. Note much larger distances between the consonants than between the vowels along the y-axis.

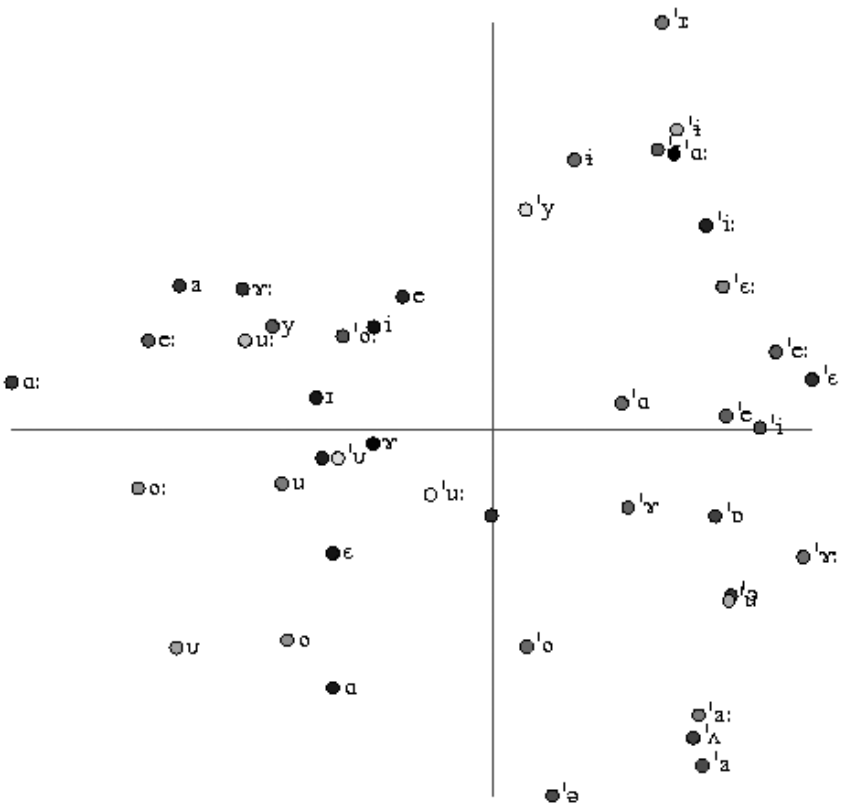

Figure 5.9: MDS plot of all vowels in the data set. Along the x-axis there is a separation between the stressed and unstressed vowels.

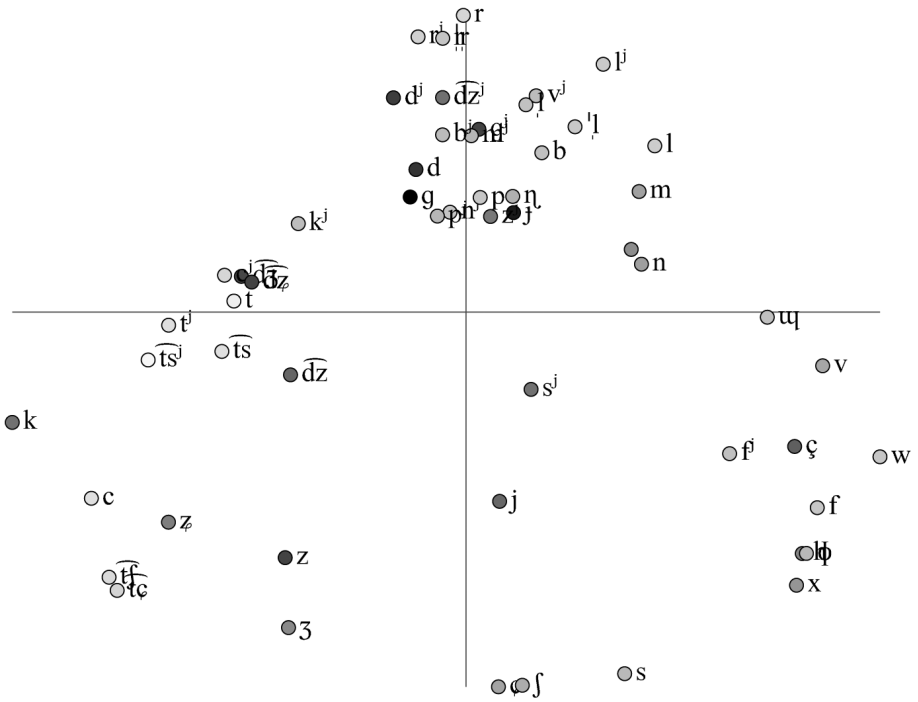

Figure 5.10: MDS plot of all consonants in the data set. In the upper part of the y-axis we note small distances between the plosives and sonorants and their palatalized counterparts.

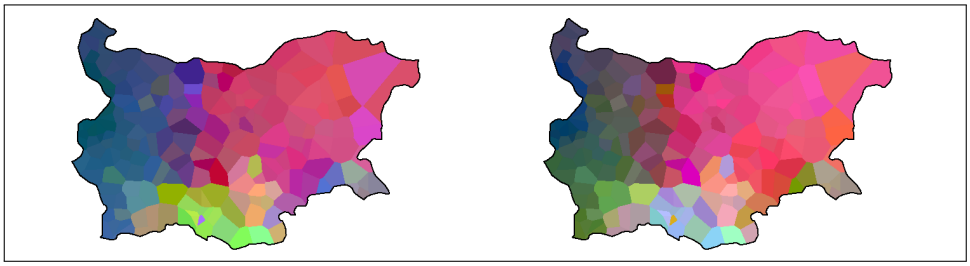

Figure 5.11: Left: MDS map of the distances produced using Levenshtein VV-CC. Right: MDS map of the distances produced using Levenshtein PMI.

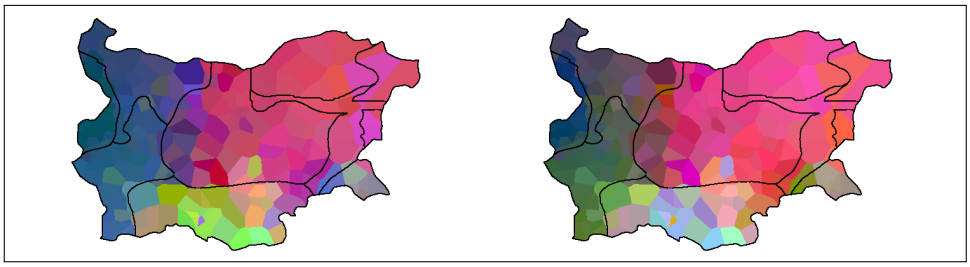

Figure 5.12: Left: Traditional borders projected on the MDS map of the distances produced using Levenshtein VV-CC. Right: Traditional borders projected on the MDS map of the distances produced using Levenshtein PMI.

In the maps in Figure 5.12 we project Stoykov's traditional dialect borders on two MDS maps in order to compare both Levenshtein algorithms to the traditional scholarship. We note that the divisions of the western part of the country visible on the Levenshtein PMI map correspond better with the traditional scholarship. On both maps there is no sign of a Moesian area, which was expected considering the findings in Chapter 4 that there is no phonetic evidence that this area is a separate dialect zone.

We also examine PMI induced distances using MDS plots presented in Figure 5.13. On both plots sites that belong to the transitional zone at the border with Serbia are located in the low left corner. We note that, on the right plot made using Levenshtein PMI produced distances, this area is more clearly separated from the rest of the western varieties than on the left plot. The distance between northwestern and southwestern varieties on the right-hand plot is also much bigger when compared to the corresponding area on the plot to the left. However northwester and southwestern varieties do not form two separate groups but rather a continuum.

Dots in the right upper corner on two plots represent sites from the southern part of the country. We also note differences in the distances between the dots in this part of the

two plots. In the left-hand plot produced using Levenshtein VV-CC, there is much bigger separation of the dots representing Rupian and western varieties than on the right-hand plot. In the right-hand plot there is no clear separation between Rupian and southwestern varieties. We also note smaller separation between Balkan and northwestern varieties on this plot, which conforms well with the findings reported in Chapter 4, where we have found number of features in the data set that are shared between northwestern area and the eastern varieties, including parts of the Rupian area.

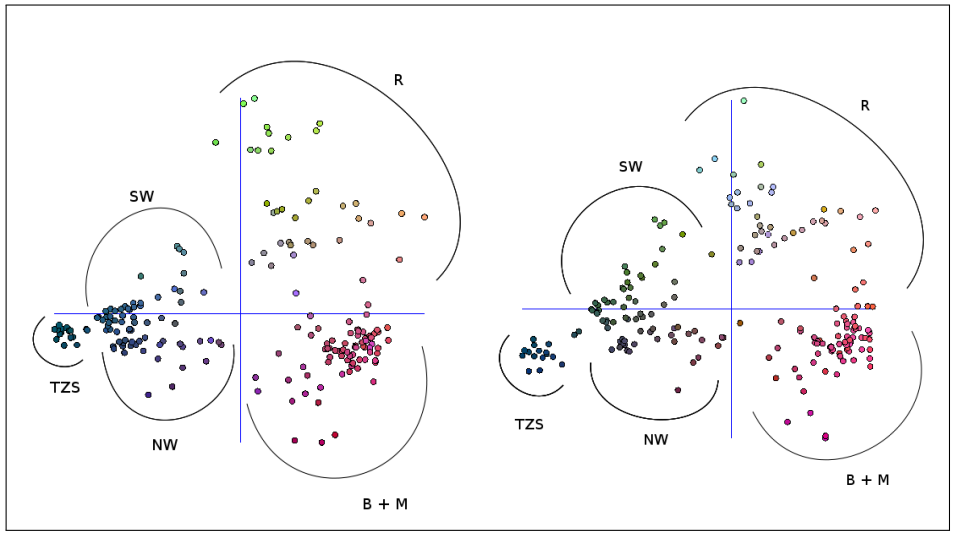

Figure 5.13: Left: MDS plot of the distances produced using Levenshtein VV-CC. Right: MDS plot of the distances produced using Levenshtein PMI.

We note that despite high correlation between two distance matrices there are differences in the MDS analyses performed on those two matrices. While distances obtained using Levenshtein VV-CC show three main groups in the data, Levenshtein PMI distances additionally separate TZS from the neighboring western varieties more clearly. At the same time, distances between northwestern and southwestern dialects are bigger, although there is no clear distinction between the two. The distances produced using Levenshtein PMI correspond to a higher extent to the traditional dialect divisions as described in Chapter 2 and Chapter 4. Levenshtein PMI distances also correctly reflect the fact that there are a number of features shared between northwestern and Balkan dialects in the east. In future we would also like to check the effect of using PMI induced distances on the clustering results.

5.5 Discussion

Measured at the segment level, pointwise mutual information has shown to be successful in improving the quality of the pairwise alignments obtained using Levenshtein algorithm. The PMI algorithm automatically learns the distances between each two phones in the data set. These automatically extracted distances are useful information for making the transcription alignments more accurate. To some extent these distances reflect the phonetic and phonological features of Bulgarian described in the traditional literature. Techniques that exploit the aligned segments extracted from the transcriptions, like those described in Prokić (2007) and Prokić and Van de Cruys (2010) can benefit from more accurate alignments.

At the aggregate level, distances between the sites calculated using Levenshtein VV-CC and Levenshtein PMI show very high correspondences ($r = 0.98$). Despite the high correlation, MDS maps produced using these two matrices show some slight differences in the analysis of the western varieties, with the Levenshtein PMI-produced map corresponding a bit better with the traditional divisions of this dialect zone.

The main limitation of the Levenshtein PMI is the constraint that vowels can align only with vowels and consonant only with consonants. Due to that restriction we cannot get information on the distances between vowels and consonants. The real merit of PMI used in string aligning would be to use this procedure without VV-CC constraint, in which case some previously described errors would be avoided and the distances between vowels and consonants would also be retrieved. However, if at the beginning of the PMI procedure no VV-CC constraint is given to the Levenshtein algorithm, the starting alignments are erroneous and as a consequence the segment distances and alignments produced by the PMI algorithm are also erroneous. This problem still remains to be solved.

Chapter 6

Multiple string alignments in linguistics

In this chapter we present and evaluate an algorithm used to produce multiple sequence alignments in linguistics ALPHAMALIG (Alonso et al., 2004). Originally used for text alignment, we adapted it slightly and applied it to our dialect pronunciation data. The alignments produced are evaluated by comparing them to the manually corrected alignments, the so-called *gold standard*. The results of evaluating the two alignments show that automatically induced alignments are of a good quality, highly corresponding with the manually produced alignments. This chapter is structured as follows. In Section 6.1 we introduce multiple sequence alignments and give our motivation for using this type of alignment. Section 6.2 gives a description of the ALPHAMALIG algorithm. We then present our gold-standard, but also simple and advanced baseline alignments in Section 6.3. Both advanced and baseline alignments are used to evaluate the quality of the automatically induced alignments. The evaluation of the alignments was done using two novel methods that we present in more detail in Section 6.4. A short discussion and some pointers for future work are given in Section 6.5. Work presented in this chapter was published as Prokić, Wieling, and Nerbonne (2009).

6.1 Multiple sequence aligning

In bioinformatics, sequence alignment is a way of arranging DNA, RNA or protein sequences in order to identify regions of similarity and determine evolutionary, functional or structural similarity between the sequences. There are two main types of string alignment: pairwise and multiple string alignment. Pairwise string alignment methods compare two strings at a time and cannot directly be used to obtain multiple string alignment

methods (Gusfield, 1997, 343-344). In multiple string alignment all strings are aligned and compared at the same time, making it a good technique for discovering patterns, especially those that are weakly preserved and cannot be detected easily from sets of pairwise alignments. Multiple string comparison is considered to be *the holy grail* of molecular biology (Gusfield, 1997, 332):

> It is the most critical cutting-edge tool for *extracting and representing* biologically important, yet faint or widely dispersed, commonalities from a set of strings.

Multiple string comparison is not new in linguistic research. In the late 19th century the Neogrammarians proposed the hypothesis of the regularity of sound change. According to *the Neogrammarian hypothesis* sound change occurs regularly and uniformly whenever the appropriate phonetic environment is encountered (Campbell, 2004). Ever since then the understanding of sound change has played a major role in the comparative method that is itself based on the simultaneous comparison of different languages, i.e. lists of cognate terms from the related languages. The correct analysis of sound changes requires the simultaneous examination of corresponding sounds in order to compare hypotheses about their evolution. Alignment identifies which sounds correspond. Historical linguists align the sequences manually, while we seek to automate this process.

In recent years there has been a strong focus in historical linguistics on the introduction of quantitative methods in order to develop tools for the comparison and classification of languages. For example, in his PhD thesis, Kondrak (2002) presents algorithms for the reconstruction of proto-languages from cognates. Warnow et al. (2006) applied methods taken from phylogenetics to Indo-European phonetic data in order to model language evolution. Heeringa and Joseph (2007) applied the Levenshtein algorithm to the Dutch pronunciation data taken from *Reeks Nederlandse Dialectatlassen* and tried to reconstruct a 'proto-language' of Dutch dialects using the pairwise alignments.

Studies in historical linguistics and dialectometry where string comparison is used as a basis for calculating the distances between language varieties will profit from tools to multi-align strings automatically and to calculate the distances between them. Good multiple alignment is of benefit to all those methods in diachronic linguistics such as the comparative reconstruction method or the so-called character-based methods taken from phylogenetics, which have also been successfully applied in linguistics (Gray and Jordan, 2000; Gray and Atkinson, 2003; Atkinson et al., 2005; Warnow et al., 2006). The multi-alignment systems can help historical linguistics by reducing the human labor needed to detect the regular sound correspondences and cognate pairs of words. They also systematize the linguistic knowledge in intuitive alignments and provide a basis for the application of the quantitative methods that lead to a better understanding of language variation and language change.

In this study we apply an iterative pairwise alignment program for linguistics, ALPHAMALIG, to the phonetic transcriptions of words used in dialectological research. We automatically multi-align all transcriptions and compare these generated alignments

with manually aligned gold standard alignments. At the same time we propose two methods for the evaluation of the multiple sequence alignments (MSA).

6.1.1 Example of multiple sequence alignment

In this section we will give an example of the automatically multi-aligned strings from our data set and point out some important features of the simultaneous comparison of more than two strings.

Aldomirovtsi:	j	ˈɑ	-	-	-	-
Beglezh:	-	ˈɑ	s	-	-	-
Belene:	-	ˈɑ	s	-	-	-
Chukovets:	j	ˈɑ	z	e	k	a
Dinevo:	j	ˈɑ	-	-	-	-
Dobroselets:	-	ˈɑ	s	-	-	-

Figure 6.1: Example of multiple string alignment for six villages. Sign for primary stress is moved to the first vowel in the stressed syllable.

In Figure 6.1 we have multi-aligned pronunciations of the word аз /az/ 'I' automatically generated by ALPHAMALIG. The advantages of this kind of alignment over pairwise alignment are twofold:

- First, it is easier to detect and process corresponding phones in words and their alternations (like [s] and [z] in the third column in Figure 6.1).
- Second, the distances/similarities between strings can be different in pairwise comparison as opposed to multiple comparison. This is so because multi-aligned strings, unlike pairwise aligned strings, contain information on the positions where phones were inserted or deleted in both strings. For example, in Figure 6.1 the pairwise alignment of the pronunciations from the villages Aldomirovtsi and Beglezh would be:

Aldomirovtsi:	j	ˈɑ	-
Beglezh:	-	ˈɑ	s

  These two alignments have one matching element out of three in total, which means that the similarity between them is $1/3 = 0.33$. At the same time the similarity between these two strings calculated based on the multi-aligned strings in Figure 6.1 would be $4/6 = 0.66$:

Aldomirovtsi:	j	ˈɑ	-	-	-	-
Beglezh:	-	ˈɑ	s	-	-	-

The measurement based on multi-alignment takes the common missing material into account as well. For example, the last three positions are not present in the pairwise alignments, which is, in some cases, an important information loss.

6.2 Iterative pairwise alignment

Multiple alignment algorithms iteratively merge two multiple alignments of two subsets of aligned strings into a single multiple alignment that is union of those subsets (Gusfield, 1997). The simplest approach is to align the two strings that have the minimum distance over all pairs of strings and iteratively align strings having the smallest distance to the already aligned strings in order to generate a new multiple alignment. Other algorithms use different initializations and different criteria in selecting the new alignments to merge. Some begin with the longest (low cost) alignment instead of the pair with the least cost absolutely. A string with the smallest edit distance to any of the already merged strings is chosen to be added to the strings in the multiple alignment. In choosing the pair with the minimal distance, all algorithms are greedy, and risk missing optimal alignments.

ALPHAMALIG is an iterative pairwise alignment program for bilingual text alignment. It uses the strategy of merging multiple alignments of subsets of strings, instead of adding just one string at the time to the already aligned strings.[1] It was originally developed to align corresponding words in bilingual texts, i.e. to work with textual data, but it functions with any data that can be represented as a sequence of symbols of a finite alphabet. In addition to the input sequences, the program needs to know the alphabet and the distances between each token pair and each pair consisting of a token and a gap.

In order to perform multiple sequence alignments of X-SAMPA word transcriptions we modified ALPHAMALIG slightly so it could work with the tokens that consist of more than one symbol, such as ["e], ["e:] and [t_S], i.e. IPA [ˈe], [ˈeː] and [t͡ʃ] respectively. The distances between the tokens were specified in such a way that vowels can be aligned only with vowels and consonants only with consonants. The same tokens are treated as identical and the distance between them is set to 0. The distance between any token in the data set to a gap symbol has the same cost as replacing a vowel with a vowel or a consonant with a consonant. Except for this very general linguistic knowledge, no other data-specific information was given to the program. In this chapter we do not use any phonetic features in order to define the segments more precisely or to calculate the distances between them in a more sensitive way other than making a binary 'match/does-not-match-distinction', since we want to keep the system language independent and robust to the highest possible degree.

To illustrate the algorithm we will look at the six pronunciations of the word aɜ /az/ 'I' presented in Figure 6.1.

[1] http://alggen.lsi.upc.es/recerca/align/alphamalig/intro-alphamalig.html

Aldomirovtsi:	j	ˈɑ				
Beglezh:	ˈɑ	s				
Belene:	ˈɑ	s				
Chukovets:	j	ˈɑ	z	e	k	a
Dinevo:	j	ˈɑ				
Dobroselets:	ˈɑ	s				

Figure 6.2: Pronunciations of word 'I' collected at six places.

In the first step the algorithm forms 2 groups: pronunciations for villages Aldomirovtsi and Dinevo [jˈɑ] are put in one and pronunciations for villages Beglezh, Belene and Dobroselets [ˈɑs] in the other. The distance between the strings within these two groups is 0, i.e. they have the same pronunciation of the word in question. Pronunciation for village Chukovets would still be non-aligned.

Aldomirovtsi:	j	ˈɑ		Beglezh:	ˈɑ	s
Dinevo:	j	ˈɑ		Belene:	ˈɑ	s
				Dobroselets:	ˈɑ	s

Chukovets: j ˈɑ z e k a

Figure 6.3: In the first step strings that have distance 0 are grouped together.

In the next step, pronunciations for villages Aldomirovtsi and Dinevo are aligned with the pronunciations for villages Beglezh, Belene and Dobroselets since the distance between these pronunciations is smaller than the distance between any of the two groups of strings to the pronunciation from village Chukovec. The distance between [jˈɑ] and [ˈɑs] is 2, between [jˈɑ] and [jˈɑzeka] is 4, while the distance between [ˈɑs] and [jˈɑzeka] is 5.

Aldomirovtsi:	j	ˈɑ	-
Dinevo:	j	ˈɑ	-
Beglezh:	-	ˈɑ	s
Belene:	-	ˈɑ	s
Dobroselets:	-	ˈɑ	s

Chukovets: j ˈɑ z e k a

Figure 6.4: In the second step strings that have distance 2 are grouped together.

In the last step already aligned strings for 5 villages are aligned against the pronunciation for village Chukovets. The distance between them is 5 since they only match in the second position [ˈɑ].

Aldomirovtsi:	j	ˈɑ	-	-	-	-
Dinevo:	j	ˈɑ	-	-	-	-
Beglezh:	-	ˈɑ	s	-	-	-
Belene:	-	ˈɑ	s	-	-	-
Dobroselets:	-	ˈɑ	s	-	-	-
Chukovets:	j	ˈɑ	z	e	k	a

Figure 6.5: In the last step all six strings are aligned.

6.3 Gold standard and baseline

In order to evaluate the performance of ALPHAMALIG, we compared the alignments obtained using this program to manually aligned strings, our gold standard, and to the alignments obtained using two very simple techniques that are described next: simple baseline and advanced baseline.

6.3.1 Simple baseline

The simplest way of aligning two strings would be to align the first element from one string with the first element from the other string, the second element with the second and so on. If two strings are not of equal length, the remaining unaligned tokens are aligned with the gap symbol which represents an insertion or a deletion. This is the alignment implicit in Hamming distance, which ignores insertions and deletions.

By applying this simple method, we obtained multiple sequence alignments for all words in our data set. An example of such a multiple sequence alignment is shown in Figure 6.6. These alignments were used to check how difficult the multiple sequence alignment task is for our data and how much improvement is obtained using more advanced techniques to multi-align strings.

Aldomirovtsi:	j	ˈɑ	-	-	-	-
Chukovets:	j	ˈɑ	z	e	k	a
Dobroselets:	ˈɑ	s	-	-	-	-

Figure 6.6: Simple baseline produced by aligning the first element from one string with the first element from the other string, the second element with the second and so on.

6.3.2 Advanced baseline

Our second baseline is more advanced than the first and was created using the following procedure:

1. for each word the longest string among all pronunciations is located
2. All strings are pairwise aligned against the longest string using the Levenshtein algorithm. We refer to the two sequences in a pairwise alignment as 'aligned sequences'. Note that aligned sequences include hyphens indicating the places of insertions and deletions.
3. the aligned sequences—all of equal length—are extracted
4. all extracted aligned sequences are placed below each other to form the multiple alignment

An example of combining pairwise alignments against the longest string (in this case [jˈɑzeka]) is shown in Figure 6.7.

Chukovets:	j	ˈɑ	z	e	k	a		Chukovets:	j	ˈɑ	z	e	k	a
Aldomirovtsi:	j	ˈɑ	-	-	-	-		Dobroselets:	-	ˈɑ	s	-	-	-

Aldomirovtsi:	j	ˈɑ	-	-	-	-
Chukovets:	j	ˈɑ	z	e	k	a
Dobroselets:	-	ˈɑ	s	-	-	-

Figure 6.7: Advanced baseline. The top two alignments each contain two aligned sequences, and the bottom one contains three.

6.3.3 Gold standard

Our gold standard was created by manually correcting the advanced baseline alignments described in the previous section. The gold standard results and both baseline results consist of 152 files with multi-aligned strings, one for each word. The pronunciations are ordered alphabetically according to the village they come from. If there are more pronunciations per site, they are all present, one under the other.

6.4 Evaluation

Although multiple sequence alignments are broadly used in molecular biology, there is still no widely accepted objective function for evaluating the goodness of the multiply

aligned strings (Gusfield, 1997). The quality of the existing methods used to produce multiple sequence alignments is judged by the 'biological meaning of the alignments they produce'. Since strings in linguistics cannot be judged by the biological criteria used in string evaluation in biology, we are forced to propose evaluation methods that are suitable for the strings in question. One of the advantages we have is the existence of the gold standard alignments, which makes our task easier and more straightforward—in order to determine the quality of the multi-aligned strings, we compare outputs of the different algorithms to the gold standard. Since there is no off-the-shelf method that can be used for comparison of multi-aligned strings to a gold standard, we propose two novel methods—one sensitive to the order of positions in two alignments and another that takes into account only the content of each position.

6.4.1 Order dependent method

The first method we develop compares the contents of the position in two alignments and also takes the position sequence into account. A position is a certain vertical position in the multiple alignments and can be best illustrated on the multiply aligned strings in Figure 6.8 (repeated from Figure 6.1) where we mark the first position:

Aldomirovtsi:	j	ˈɑ	-	-	-	-
Beglezh:	-	ˈɑ	s	-	-	-
Belene:	-	ˈɑ	s	-	-	-
Chukovets:	j	ˈɑ	z	e	k	a
Dinevo:	j	ˈɑ	-	-	-	-
Dobroselets:	-	ˈɑ	s	-	-	-

Figure 6.8: This multiple alignment contains 6 positions. We mark the first position.

The order dependent evaluation (ODE) procedure is as follows:[2]

- Each gold standard column is compared to the most similar column out of two neighboring columns of a candidate multiple alignment. The two neighboring columns depend on the previous matched column j and have indices $j+1$ and $j+2$ (at the start $j=0$). It is possible that there are columns in the candidate multiple alignment which remain unmatched, as well as columns at the end of the gold standard which remain unmatched.

- The similarity of a candidate column to a gold standard column is calculated by dividing the number of correctly placed elements in every candidate column by the

[2] In Prokić, Wieling, and Nerbonne (2009) we use the name 'column dependent method' for the same method.

total number of elements in the column. A score of 1 indicates perfect overlap, while a score of 0 indicates the columns have no elements in common. This calculation is performed for each column.

- The similarity score of the whole multiple alignment (for a single word) is calculated by summing the similarity score of each candidate column and dividing the resulting sum by the total number of matched columns plus the total number of unmatched columns in both multiple alignments.
- The final similarity score between the set of gold standard alignments with the set of candidate multiple alignments is calculated by averaging the multiple alignment similarity scores for all strings.

As an example consider the multiple alignments in Figure 6.9, with the gold standard alignment (GS) on the left and the generated alignment (GA) on the right.

w	rʲ	ˈɛ	m	e
v	r	ˈe	m	i
u	rʲ	ˈe	m	i
v	rʲ	ˈe	m	i

w	-	rʲ	ˈɛ	m	e
v	-	r	ˈe	m	i
-	u	rʲ	ˈe	m	i
v	-	rʲ	ˈe	m	i

Figure 6.9: GS and ALPHAMALIG multiple string alignments, the gold standard alignment left, the ALPHAMALIG output right.

The evaluation starts by comparing the first column of the GS with the first and second column of the GA. The first column of the GA is the best match, since the similarity score between the first columns is 0.75 (3 out of 4 elements match). In similar fashion, the second column of the GS is compared with the second and the third column of the GA and matched with the third column of GA with a similarity score of 1 (all elements match). The third GS column is matched with the fourth GA column, the fourth GS column with the fifth GA column and the fifth GS column with the sixth GA column (all three having a similarity score of 1). As a consequence, the second column of the GA remains unmatched. In total, five columns are matched and one column remains unmatched. The total score of the GA equals:

$$\frac{(0.75+1+1+1+1)}{(5+1)} = 0.792$$

It is clear that this method punishes unmatched columns by increasing the value of the denominator in the similarity score calculation. As a consequence, swapped columns are punished severely, which is illustrated in Figure 6.10. In the alignments in Figure 6.10, the first three columns of GS would be matched with the first three columns of GA with a score of 1, the fourth would be matched with the fifth, and two columns would be left unmatched: the fifth GS column and the fourth GA column yielding a

total similarity score of $4/6 = 0.66$. Especially in this case this is undesirable, as both sequences of these columns represent equally reasonable multiple alignment and should have a total similarity score of 1. We therefore need a less strict evaluation method which does not insist on the exact ordering. An alternative method is introduced and discussed in the following section.

ˈo	rʲ	ə	j	-		ˈo	rʲ	ə	-	j
ˈo	rʲ	ə	-	u		ˈo	rʲ	ə	u	-
ˈo	rʲ	ə	f	-		ˈo	rʲ	ə	-	f

Figure 6.10: Two alignments with swapped columns.

6.4.2 Modified Rand index

In developing an alternative evaluation we proceeded from the insight that the columns of a multiple alignment are a sort of partition of the elements of the alignment strings, i.e., they constitute a set of disjoint multi-sets whose union is the entire multi-set of segments in the multiple alignment. Each column effectively assigns its segments to a partition, which clearly cannot overlap with the elements of another column (partition). Since every segment must fall within some column, the assignment is also exhaustive.

Our second evaluation method is therefore based on the modified Rand index (Hubert and Arabie, 1985) described in Chapter 3. The modified Rand index is used in classification for comparing two different partitions of a finite set of objects. In Chapter 3 we have used it to compare the classification of sites done by various clustering algorithms to the traditional division of the sites. In this chapter we use it to assess the quality of each column from the automatically induced multiple sequence alignments.

We would like to emphasize that it is clear that the set of columns of a multi-alignment have more structure than a partition *sec*, in particular because the columns (subpartitions) are ordered, unlike the subpartitions in a partition. But we shall compensate for this difference by explicitly marking order.

ˈo [1]	rʲ [2]	ə [3]	j [4]	-
ˈo [5]	rʲ [6]	ə [7]	-	u [8]
ˈo [9]	rʲ [10]	ə [11]	f [12]	-

Figure 6.11: Annotated multiple sequence alignment.

In our study, each segment token in each transcription is treated as a different object (see Figure 6.11), and every column is taken to be a sub-partition to which segment

tokens are assigned. Both alignments in Figure 6.10 have 12 phones that are put into 5 groups. We 'tag' each token sequentially in order to distinguish the different tokens of a single segment from each other, but note that the way we do this also introduces an order sensitivity in the measure. Since columns 4 and 5 in two of the multiple sequence alignments in Figure 6.10 are swapped, we obtain the following two partitions:

GS1 = {1,5,9}	GA1 = {1,5,9}
GS2 = {2,6,10}	GA2 = {2,6,10}
GS3 = {3,7,11}	GA3 = {3,7,11}
GS4 = {4,12}	GA4 = {8}
GS5 = {8}	GA5 = {4,12}

Using the modified Rand index the quality of each column is checked, regardless of whether the columns are in order. The MRI for the alignments in Figure 6.10 will be 1, because both alignments group segment tokens in the same way. Even though columns four and five are swapped, in both classifications phones [j] and [f] are grouped together, while sound [u] forms a separate group.

The MRI itself only takes into account the quality of each column separately since it simply checks whether the same elements are together in the candidate alignment as in the gold-standard alignment. It is therefore insensitive to the ordering of columns. While it may have seemed counterintuitive linguistically to proceed from an order-insensitive measure, the comparison of 'tagged tokens' described above effectively reintroduces order sensitivity.

In the next section we describe the results of applying both evaluation methods on the automatically generated multiple alignments.

6.4.3 Results

After comparing all files of the baseline algorithms and ALPHAMALIG against the gold standard files according to the order dependent evaluation method and the modified Rand index, the average score is calculated by summing up all scores and dividing them by the number of word files (152).

The results are given in Table 6.1 and also include the number of words with perfect multi-alignments (i.e. identical to the gold standard). Using ODE, ALPHAMALIG scored 0.932 out of 1.0 with 103 perfectly aligned files. The result for the simple baseline was 0.710 with 44 perfectly aligned files. As expected, the result for the advanced baseline was in between these two results—0.869 with 72 files that were completely identical to the GS files. Using MRI to evaluate the alignments generated we obtained generally higher scores for all three algorithms, but with the same ordering. ALPHAMALIG scored 0.982, with 104 perfectly aligned files. The advanced baseline had a lower score of 0.937 and 74 perfect alignments. The simple baseline performed worse, scoring 0.848 and having 44 perfectly aligned files.

Table 6.1: Results of evaluating outputs of the different algorithms against the GS.

	ODE	ODE perfect columns	MRI	MRI perfect columns
Simple baseline	0.710	44	0.848	44
Advanced baseline	0.869	72	0.937	74
ALPHAMALIG	**0.932**	**103**	**0.982**	**104**

The scores of the ODE evaluation method are lower than the MRI scores, which is due to the first method's problematic sensitivity to column ordering in the alignments. It is clear that in both evaluation methods ALPHAMALIG outperforms both baseline alignments by a wide margin.

It is important to notice that the scores for the simple baseline are reasonably high, which can be explained by the structure of our data set. The variation of word pronunciations is relatively small, making string alignment easier. However, ALPHAMALIG obtained much higher scores using both evaluation methods.

Additional qualitative error analysis reveals that the errors of ALPHAMALIG are mostly caused by the vowel-vowel consonant-consonant alignment restriction. In the data set there are 21 words that contain metathesis, i.e. switched sounds within the words. More on metathesis can be found in Section 5.2.1. Since vowel-consonant alignments were not allowed in ALPHAMALIG, alignments produced by this algorithm were different from the gold standard, as illustrated in Figure 6.12. The vowel-consonant re-

v	ˈɣ	r	x		v	ˈɣ	r	-	x
v	r	ˈɣ	x		v	-	r	ˈɣ	x

Figure 6.12: Two alignments with metathesis. The gold standard on the left hand side, and the erroneous produced by ALPHAMALIG on the right hand side.

striction is also responsible for wrong alignments in some words where metathesis is not present, but where the vowel-consonant alignment is still preferred over aligning vowels and/or consonants with a gap (see for example Figure 6.9).

The other type of error present in the ALPHAMALIG alignments is caused by the fact that all vowel-vowel and consonant-consonant distances receive the same weight. In Figure 6.13 the alignment of word бяхме /ˈbʲaxme/ 'were - 1st pl' produced by ALPHAMALIG is wrong because instead of aligning [mʲ] with [m] and [m] it is wrongly aligned with two tokens of [x], while a third token of [x] is aligned with [ʃ] instead of aligning it with [x] and [x]. This is the sort of error which segment-weighted alignments

such as the one presented in Chapter 5 might be expected to prevent, at least to some extent.

b	ˈɛ	**ʃ**	u	**x**	-	m	e	-
bʲ	ˈɑ	**-**	-	**x**	-	m	i	-
b	ˈe	**x**	-	**mʲ**	-	-	ɤ	-

Figure 6.13: Alignment error produced by ALPHAMALIG.

6.5 Discussion

In this chapter we have presented a technique to automatically multi-align phonetic transcriptions. We have also introduced two novel techniques that can be used to evaluate the quality of the multi-aligned strings. Both evaluation methods are based on comparing the automatically induced alignments to gold-standard alignments. The results have shown that the automatically produced multi-alignments are of a good quality with less than 2 per cent error on the segment level. However, in order to apply either of these two methods it is necessary to have a gold standard alignment against which the automatically induced alignments are compared to. We are aware that for many data sets this is neither available nor easily obtainable. But in cases where it exists, we find these techniques a useful evaluation tool.

The comparison between our simple baseline to the gold standard alignments has shown that our data set contains strings with relatively simple structure. Pronunciation variation in our dialects is relatively small, especially if compared to cross-linguistic data. The structure of syllables is also relatively simple, very often showing only CV structure. It would be very important to apply the ALPHAMALIG algorithm to the data from some other languages in order to obtain further insight into the quality of the alignments produced.

In the alignment procedure, we have used vowel-vowel consonant-consonant restriction as the only 'linguistic' knowledge given as an input to the ALPHAMALIG algorithm. This, on one side, makes the alignment robust and language independent, but, on the other, introduces some errors in the alignments like those presented in Figure 6.13. We believe that the quality of the generated alignments could be further improved if some kind of segment weighting were introduced into the alignment procedure, such as the one presented in Chapter 5. Weighting of the segments could also enable the vowel-vowel consonant-consonant constraint to be completely eliminated from the aligning. This could lead to better alignments in the cases where vowels and consonants need to be aligned (see for example Figure 6.12).

The automatic multi-aligning could be improved in many other ways. There are

also various algorithms used to multi-align sequences in biology. Some of them could potentially be adopted to work with the strings in linguistics. We hope that our first experiments with the multiple aligned phonetic transcriptions have shown the usefulness of this type of approach to string comparison in linguistics and that in future further experiments in this direction will be conducted.

Chapter 7

Bayesian phylogenetic inference

In this chapter we use automatically multi-aligned phonetic transcriptions to infer the historic relationships between the language varieties, but also to explore the relationship between the various phones in the data set. Multi-aligned transcriptions are analyzed using Bayesian Monte Carlo Markov Chain inference (MCMC), in recent years one of the most popular and the most powerful methods in molecular phylogeny for inferring the relationships between species. Bayesian MCMC inference belongs to the so-called character based methods, together with some other popular methods like maximum parsimony and maximum likelihood methods for phylogenetic inference. In the next section we briefly introduce molecular phylogenetics based on Page and Holmes (2006), followed by an introduction to character-based methods in Section 7.2. Section 7.3 gives an overview of the application of the methods taken from phylogenetics in linguistics. We then give introduction to Bayesian phylogenetic inference in Section 7.4. In Section 7.5 we present our experiment, followed by the results that we report in Section 7.6. We conclude the chapter with the discussion presented in Section 7.7.

7.1 Phylogenetic inference

Phylogenetics is a branch of biology that studies the evolutionary relatedness among various species. The relatedness can be inferred at the molecular level by examining the differences between DNA or protein sequences of the organisms. DNA sequences are composed of four nucleotides (A, C, T, and G), while protein sequences comprise 20 different amino acids. Closely related organisms have similar structure of DNA (protein) sequences, i.e. similar order of the nucleotides (amino acids) in their DNA (protein) sequences. More distantly related organisms show more dissimilarity if we compare their DNA (protein) sequences. Another approach to phylogenetic inference is to compare various morphological characteristics of the organisms. In this chapter we focus on mo-

lecular phylogenetics and try to use some of the models developed for the evolution of DNA and protein sequences on language data.

One of the most important events in the development of molecular evolution was the discovery of the molecular structure of DNA in 1953 by James Watson and Francis Crick (Page and Holmes, 2006, 4). The first comparison of amino acid sequences came in 1955, when Fred Sanger and his colleagues used it to compare protein insuline from cattle, pigs and sheep.

In order to recover evolutionary information from DNA and protein sequences, it is necessary to formalize the process of sequence change over time. In 1960s different models of molecular evolution started being developed. The comparison of sequences proceeds from their alignment— either pair-wise or multiple sequence alignment depending on the approach. The distances between the aligned sequences can be, in the simplest case, expressed as the number of the segments in which two sequences differ. Since there are usually multiple changes at each position within a sequence, distances inferred in this way are actually underestimating the amount of evolutionary change. To correct for this, different evolutionary models based on the frequency of the nucleotides and the probability of a nucleotide substitution have been developed. The most simple model, the so-called Jukes-Cantor model, assumes that the four nucleotides have equal frequencies and that all substitutions are equally likely. The most general model, general reversible model, allows each possible nucleotide substitution to have its own probability.

Information from the aligned and compared sequences is turned into an evolutionary tree that is used to represent genetic relatedness among the species. A tree consists of nodes connected by branches. There are three types of nodes: terminal nodes that represent organisms (sequences), internal nodes that represent hypothetical ancestors and a root node that is the ancestor of all organisms (Figure 7.1).

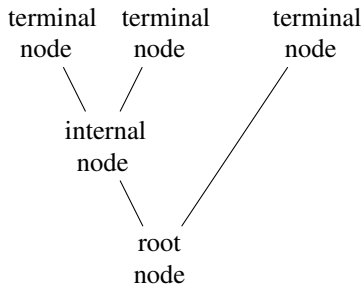

Figure 7.1: An example of a phylogenetic tree.

There are numerous methods that are used to convert aligned sequences into trees. Based

on how the data is treated they can be divided into distance-based methods and character-based methods. In distance-based methods, the distance between each two strings is represented as a single number and stored into a matrix. This matrix is used by various tree building methods to construct an evolutionary tree. This kind of approach is used in Chapter 3 of this thesis to infer the relatedness among the dialect varieties in Bulgaria. Character-based methods analyze each position in the aligned sequences separately. In the next section we describe character-based methods in more detail.

7.2 Character-based methods

Character-based methods (CBM) comprise various methods used in phylogenetics to study the evolutionary relatedness among species. In CBM each species is described in the terms of the states of certain characters. The term 'character' is used to refer to a different position in a DNA or a protein sequence. For every species, a character is in one of the states inherent for that character. Some states only vary between present/absent states, while the so-called multi-state characters can have multiple states. CBM proceed from the simultaneously aligned sequences of various species and perform the analysis based on each of the characters separately. A scheme of the aligned sequences for 3 species and 5 characters would look like this:

	character1	character2	character3	character4	character5
species1	state1	state1	state1	state1	state1
species2	state2	state1	state1	state2	state1
species3	state2	state2	state1	state3	state2

Figure 7.2: A scheme of the aligned sequences for 3 species and 5 characters.

Unlike the distance methods described in Chapter 3 that aggregate all the differences between each two strings into a single distance, character-based methods infer relatedness between the species separately on each character and later combine those analyses into a single tree. In that way the information provided by each character is retained and information loss that results from converting the sequence data into distance scores is avoided (Penny, 1982). Two well known CBM are maximum parsimony and maximum likelihood.

Parsimony methods were among the first methods to be used to infer phylogenies. They are based on the idea of the 'minimum evolution'. While reconstructing the phylogenetic tree, the algorithm seeks the tree with the smallest number of events, i.e. for the smallest tree that would explain the data. This method has often been criticized because it does not rely on any model of evolution, but seeks for the most simple explanation of the data. The other problem related to this approach is the so-called *long branch attrac-*

tion phenomenon where species that evolve rapidly are grouped together in the phylogenetic analysis regardless of their true genetic relationship (Page and Holmes, 2006).

Unlike parsimony, probabilistic methods for phylogenetic inference, like maximum likelihood and Bayesian inference, are based on a specific model of evolution. The maximum likelihood method, as the name suggests, is based on the concept of likelihood, the probability of observing the data given a particular model or hypothesis. Given some data D and a hypothesis H the likelihood L can be expressed as:

$$L = P(D|H) \tag{7.1}$$

In phylogenetics, D is a set of aligned sequences, and H is a phylogenetic tree. The tree that makes our data the most probable is the maximum likelihood tree. Detailed explanation on the parsimony and likelihood methods in phylogenetics can be found in Felsenstein (2004). In our experiments we have used Bayesian inference, in recent years one of the most popular character-based methods for inferring phylogenies. In Section 7.4 we present it in more detail. Before that, we will look into the usage of phylogenetic methods in linguistics.

7.3 Phylogenetic inference in linguistics

In the last decade there has been an increasing interest in the application of the methods taken from phylogenetics to the language data. This line of research starts from the premise that there is a genuine similarity between the evolution of species and the evolution of languages. Although there are some important differences in their evolution, the mechanisms of the change of species and languages are the same: they split into new species/languages, mutate, borrow material from neighboring species/languages, and innovations in both languages and species appear independently in unrelated elements. They both document evolutionary history, species in molecules and various morphological characteristics, languages in phonetics/phonology, morphology, syntax. Evolution and relatedness of both the species and languages can be described using family trees.

Methods taken from computational phylogenetics have been applied to lexical (Gray and Jordan, 2000; Gray and Atkinson, 2003) and phonetic data (Warnow, 1997; Nakhleh, Ringe, and Warnow, 2005) to study evolutionary relationships between languages or dialects (Hamed, 2005; Hamed and Wang, 2006; McMahon et al., 2007). They have been used to address the problems of the origins of Indo-European (Gray and Jordan, 2000) and Bantu languages (Holden, 2002; Holden and Gray, 2006). They were also applied to the problems of the subgrouping of Indo-European (Ringe, Warnow, and Taylor, 2002; Nakhleh, Ringe, and Warnow, 2005), as well as to test various hypotheses about human prehistory (Dunn et al., 2005; Greenhill and Gray, 2005; Gray, Drummond, and Greenhill, 2009). As pointed out in Greenhill and Gray (2009), computational phylogenetic methods are seen as ‘a powerful supplement to the comparative method used

in historical linguistics'. They are not a replacement of the traditional well-established methods in linguistics, but help in resolving some rather old questions on the history of languages. Although developed to work with different types of data, the use of the new techniques developed for phylogenetic inference opens new perspectives in the field of historical linguistics and potentially in dialectology. However, it does not come without its problems and concerns. Although they share the same mechanisms of change, species and languages differ in many ways. Languages change much faster than species. Borrowing between neighboring languages, regardless of their genetic relatedness, is much more common than between species. The two most important preconditions for analyzing languages using methods from phylogenetics are the adequate linguistic data coding and the choice of an appropriate model of language change. If the data employed in the analyses is not well analyzed and coded, it will lead to wrong results. The same holds for the wrong choice of the evolutionary models. All models implemented in the computational phylogenetic software are naturally designed to cover various aspects of the evolution of species. Most of them cannot be applied to the linguistic data since the assumptions behind those models violate the known facts of the linguistic change. But those that fit linguistic data well are a good start for the exploration of the possibilities of using the phylogenetic methods on the language data, as they enable researchers to analyze larger bodies of data while systematically controlling many aspects of analysis.

In this chapter we apply Bayesian methods used to infer phylogenies to the Bulgarian phonetic data. It is, to our knowledge, the first time that methods borrowed from phylogenetics are directly applied to phonetic transcriptions of words. We first present our coding of the dialect pronunciation data. On one hand, the data had to be simplified so that we would be able to use software developed for biological data. This simplification led to an information loss, since we were not able to use all the phones in our data to infer relatedness between the dialect varieties. On the other hand, the coding was linguistically informed so that we could preserve enough relevant information to allow us to address certain issues related to language change. We try to find a good compromise between the two. We then present several models of evolution that we find appropriate for our dialect data and apply them to the previously coded data. As a result we obtain dialect divisions based on an approach that is historically motivated and compare them to the divisions obtained using methods that focus on geographic organization of Bulgarian dialects. We also test various hypotheses of vowel changes.

In the next section we give an introduction to Bayesian inference of phylogeny, and after that focus on the experiment.

7.4 Bayesian inference of phylogeny

In probability theory, Bayes theorem dates back to the 18th century. It gives a mathematical representation of how a conditional probability of event A given event B is related to the conditional probability of B given A:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{7.2}$$

where

$P(A)$ is the prior probability of A
$P(B)$ is the prior probability of B
$P(A|B)$ is the conditional probability of A given B, also called posterior probability
$P(B|A)$ is the conditional probability of B given A, also called likelihood

Bayesian inference of phylogeny, based on Bayes theorem, was independently proposed by several authors in 1996 (Rannala and Yang, 1996; Mau, 1996; Li, 1996). Just as the maximum likelihood method, it is based on the likelihood function, i.e. the probability of observing the data given a tree. In addition to the maximum likelihood method, it includes the prior probability of a phylogeny, i.e. tree, in the testing of a hypothesis (Huelsenbeck et al., 2002). In phylogenetic inference, Bayes theorem (Equation 7.2) can be expressed as:

$$P(\tau|D) = \frac{P(D|\tau)P(\tau)}{P(D)} \tag{7.3}$$

where

$P(\tau)$ is the prior probability of a tree
$P(\tau|D)$ is the posterior probability of a tree
$P(D|\tau)$ is the likelihood of a tree
$P(D)$ is the probability of data, which is an aligned sequence of characters

Unlike maximum likelihood that searches for the most likely tree, Bayesian inference of phylogeny is based upon finding a large number of trees with a high posterior probability. The number of all possible trees $B(s)$ for s species depends on the number of species (s). For rooted bifurcating trees

$$B(s) = \frac{(2s-3)!}{2^{s-2}(s-2)!} \tag{7.4}$$

B(2) = 1
B(3) = 3
B(4) = 15
B(5) = 105
B(6) = 945
B(7) = 10395

B(8) = 135135
B(9) = 2027025
B(10) = 34459420
B(20) = 8.200795×10^{21}
B(50) = 2.752921×10^{76}

while for unrooted bifurcating trees

$$B(s) = \frac{(2s-5)!}{2^{s-3}(s-3)!} \tag{7.5}$$

B(2) = 1
B(3) = 1
B(4) = 3
B(5) = 15
B(6) = 105
B(7) = 945
B(8) = 10395
B(9) = 135135
B(10) = 2027025
B(20) = 2.22×10^{20}
B(50) = 2.84×10^{74}

It is clear that for both rooted and unrooted trees the number of trees grows very fast as the number of species increases. It should be noted that term 'tree' refers to the way in which terminal nodes are grouped, regardless of the assignments to the internal nodes. For example, for three species (or in our case villages) we can have three different rooted trees (tree topologies). In Figure 7.3 we present three possible trees for villages Lobosh, Mihaltsi and Slaveino.

As in other character-based methods, all calculations in Bayesian inference are based on each of the sites, i.e. positions in the aligned sequences separately.[1] To calculate the posterior probability of a tree (tree topology), we need a prior probability of a tree (tree topology) and a likelihood of a tree which is based on the observed data in each of the positions in the alignments separately. The posterior probability of a phylogenetic tree τ_i for the ith position can be calculated using the following formula:

$$P(\tau_i|D) = \frac{P(\tau_i)P(D|\tau_i)}{\sum_{j=1}^{B(s)} P(D|\tau_j)P(\tau_j)} \tag{7.6}$$

[1] In molecular biology term 'site' is used to refer to a position in a DNA or protein sequence. In dialectometry, and through out this thesis, we use term 'site' to refer to a location where the data comes from. In order to avoid misunderstanding, in this chapter we will refer to a specific position in a sequence as a 'position' or simply try to give a descriptive explanation.

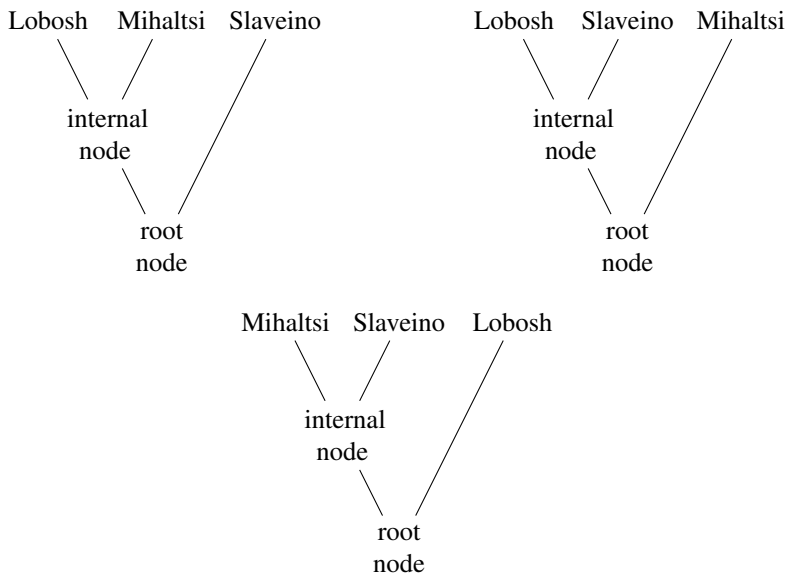

Figure 7.3: The 3 possible trees for 3 villages.

$P(\tau_i)$ is a prior probability of the ith tree. The use of prior probability sets Bayesian inference apart from the maximum likelihood method. This is considered the strongest and at the same time the weakest point of the Bayesian inference. If we have reliable information on the priors, it can help us get better posterior estimates, and it can be very powerful tool. But, in reality it is very hard to find realistic estimates for the priors. In the case of phylogenetic inference, usually all trees are considered equally probable and they are assigned the so-called flat priors where $P(\tau_i) = \frac{1}{|B(s)|}$. In this case, Bayes inference and maximum likelihood do not differ in the trees they prefer. However, final result in a maximum likelihood approach is a single tree, while Bayesian approach provides the whole distribution of trees. This enables us to sample a large number of high probability trees from the posterior.

$P(D|\tau_i)$ is the likelihood of the ith tree, i.e. the probability of observing the data at the ith site. To be able to calculate the likelihoods we need the phylogenetic model that consists of a tree τ_i, branch lengths on the tree v_i and the substitution model θ. To illustrate how the likelihoods are calculated we use aligned transcriptions of the word бели /'beli/ 'white - pl.' for three villages as our observed data (Figure 7.4). In this example the states of the characters are the phones themselves. In our example we will focus on the second position in our aligned data (character2). In Figure 7.5 we present one of the trees for the second position in the alignment given in Figure 7.4.

	character1	character2	character3	character4
Lobosh:	b	ˈe	l	i
Mihaltsi:	b^j	ˈe	l	i
Slaveino:	b	ˈɛ	l	i

Figure 7.4: A scheme of the aligned transcriptions for word 'white' for 3 villages.

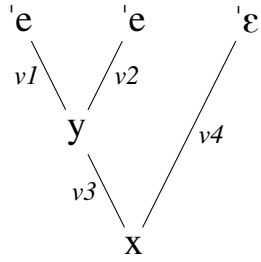

Figure 7.5: An example of a phylogenetic tree for 3 species for the second character.

There are three terminal nodes ˈe, ˈe, ˈɛ, one internal (y) and a root node (x). The branches are labeled from $v1$ to $v4$. An internal node and a root node can have any state inherent for the second character. In our case, it could be any of the 43 tokens that we use for various vowels in our data set, since vowels can align only with other vowels and consonants only with the consonants. For the two nodes we get $43 \times 43 = 1849$ possible combinations for state assignments. In Figure 7.6 we present one of the possible assignments of the states for the nodes x and y. We note that there is only one change of states on the tree in Figure 7.6: ˈe $\rightarrow$ ˈɛ. We mark it with a dashed horizontal line on branch $v4$. Branch lengths in a tree represent the number of changes that have occurred in a certain branch. For example, in Figure 7.6 there is one change on branch $v4$, meaning that this branch has length 1.

To be able to calculate the likelihood of the ith tree $P(D|\tau_i)$, apart from a tree τ_i with branch lengths v_i, we need a substitution model θ. The substitution model θ is a model of how one state changes into the other, i.e. a model that specifies the probability of one state changing into the other. θ operates both on the leaf nodes such as [ˈe] and [ˈɛ], but also on internal nodes such as x and y. In our example, we would need to know the probability of one phone changing into any other phone present in the aligned sequences. In the simplest model, a character can go from any state into any other state. The probability of going from one into the other state is equal for all pairs of states. This is neither very realistic for most of the data in biology, nor for the language data. We know that phones are not equally likely to change into all other phones, but prefer some

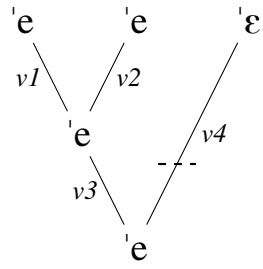

Figure 7.6: An example of a possible state assignments for the internal and a root node.

changes. More complex phylogenetic models allow different rates of change between the states, which suits our data better. In more complex substitution models it is also possible to specify the directionality of a change. The substitutions may have different values for 'e $\rightarrow$ 'ε and for 'ε $\rightarrow$ 'e. Another parameter that we can add to the phylogenetic model is the 'site heterogeneity rate'. It allows us to specify if different characters, i.e. positions, evolve at the same or different rate. In our linguistic example, it is more likely that some characters evolve faster since changes are more frequent at the beginning and at the end of words, than in the middle. In phylogenetic models this is set by having a distribution of character rates instead of a uniform rate. It is usually done by estimating a so-called gamma distribution of rate changes from the data (Yang, 1994).

The likelihood of a tree $P(D|\tau_i)$ is calculated by integrating over all possible combinations of branch lengths (v_i) and substitution model parameters (Huelsenbeck et al., 2002):

$$P(D|\tau_i) = \int_{v_i,\theta} P(D|\tau_i, v_i, \theta)P(v_i, \theta)dv_i d\theta \tag{7.7}$$

where $P(v_i, \theta)$ is the prior probability density of the branch lengths and substitution model parameters, and $d\theta$ is an infinitesimal interval. The likelihood $P(D|\tau_i, v_i, \theta)$ is normally calculated under a Markov model of character evolution—the probability of every node is dependent only on the preceding node and the branch length between these two nodes. This assumes that all positions and all lineages (villages in our case) evolve independently. The likelihood of the tree in Figure 7.6 is the product of the probabilities of every node in the tree:

$$L = P('\text{e})P('\text{e} \rightarrow '\text{e}|v3)P('\text{e} \rightarrow '\text{e}|v1)P('\text{e} \rightarrow '\text{e}|v2)P('\text{e} \rightarrow '\varepsilon|v4)$$

Probability of one state changing into the other, 'e $\rightarrow$ 'e or 'e $\rightarrow$ 'ε in our example, given a certain branch length ($v1$-$v4$), is defined by the substitution model θ. The likelihood of a tree τ_i for the position i is the product of all possible ancestral states combinations

for that position (combinations of all possible assignments for the internal node y a root node x given a certain branch length v_i).

We now go back to Formula 7.6 (we repeat it for convenience) used to calculate the posterior probability of a phylogenetic tree τ_i :

$$P(\tau_i|D) = \frac{P(\tau_i)P(D|\tau_i)}{\sum_{j=1}^{B(s)} P(D|\tau_j)P(\tau_j)}$$

where $P(\tau_i)$ is a prior probability of the ith tree, and $P(D|\tau_i)$ the likelihood of the ith tree. The remaining element is a denominator $\sum_{j=1}^{B(s)} P(D|\tau_j)P(\tau_j)$ used as a normalizing constant. It denotes marginal probability of the data, obtained by summing the probability of the data under the assumption of all the different trees. $B(s)$ is a number of all possible trees for s species. For both rooted and unrooted trees the number of trees grows very fast as the number of species increases. It is computationally extremely expensive to calculate the denominator in Equation 7.4 (repeated in 7.8) and in the general case not feasible at all.

$$\sum_{j=1}^{B(s)} P(D|\tau_j)P(\tau_j) \tag{7.8}$$

We need to do calculations for all possible trees and for each tree to integrate over all possible combinations of branch lengths and parameter values of the substitution model. In order to sample from a posterior probability distribution on trees, Bayesian inference in phylogeny uses Markov Chain Monte Carlo (MCMC) modeling. MCMC involves three steps: a) pick a tree randomly or one that is a good description of the data; b) propose a new tree by stochastically perturbing the current tree; and c) accept or reject new tree with a probability described by Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970). The number of generations that the MCMC algorithm will execute is set by the user. It depends on the size of the data set and the complexity of the model. The chain length should be run enough to obtain a good approximation of the posterior probabilities of trees and the parameters. As a result of Bayesian inference we do not get a single tree, as in other character-based methods, but a sample of trees chosen according to their posterior probability. Information from sample trees can be summarized in a single tree using different methods, such as the 'maximum clade credibility tree', 'majority rule consensus tree' or simply a single tree that seems most probable. A tree where the information is summarized, contains the information on the posterior probabilities of the nodes and particular clades, i.e. branches in a phylogenetic tree.

In the past decade Bayesian MCMC inference has become a very popular method in molecular phylogenetics. The possibility of including priors in the analysis makes it a potentially very powerful technique that sets it apart from similar statistical methods. Thanks to Monte Carlo sampling, it is also faster than the maximum likelihood method, which requires heavy computation, an issue with both of these methods (Archibald,

Mort, and Crawford, 2003). Recently there have been several attempts to apply this method to language data (Gray and Atkinson, 2003; Pagel, Atkinson, and Meade, 2007; Greenhill and Gray, 2009). However, they were used either on cognate sets or on lexical data from various languages. The present chapter is, to our knowledge, the first attempt to apply it directly on dialect phonetic data.

In this section we have tried to give a general overview of how Bayesian phylogenetic method works. For more technical and detailed explanation on Bayesian inference we refer an interested reader to Huelsenbeck et al. (2001) and Huelsenbeck et al. (2002). A very good, less technical, description of Bayesian inference can be found in Greenhill and Gray (2009).

7.5 Experiment

In the research described in this section, we apply Bayesian inference to the dialect phonetic data in order to discover the relationships between various sites, but also between the phones found in our data set.[2] All calculations related to the Bayesian MCMC inference were done using the BEAST software (Drummond and Rambaut, 2007). The experiment was set as follows.

- We proceed by automatically multi-aligning 152 word transcriptions in the data set. We use the ALPHAMALIG algorithm described in Section 6.2. The algorithm is given a constraint that vowels can be aligned only with vowels and consonants only with consonants. The evaluation of the multiple sequence alignments produced by the ALPHAMALIG algorithm, when this constraint is used, has shown that they correspond well with the gold standard alignments and can be used in our experiment for further analyses. For example, in Figure 7.7 we present multi-aligned pronunciations for words вечер /ˈvet͡ʃer/ ‘evening’, дъно /ˈdɤno/ ‘bottom’ and лесно /ˈlesno/ ‘easily’ for five villages.

- If there are multiple pronunciation of a certain word in some villages, we randomly chose only one pronunciation per site in order to conform to the format that can be handled by the software used for Bayesian inference.

- After multi-aligning transcriptions for every word separately, we merge all aligned transcriptions into a single set of multi-aligned strings, where each string contains transcriptions of all 152 pronunciations collected at a certain village. Bayesian MCMC inference infers the relationships between language varieties by processing multiple alignments position by position. This allows us to merge the transcriptions of all words into a single set of multi-aligned strings, since our calculations

[2] This experiment was conducted during the research visit to the University of Auckland. We would like to thank Prof. Russell Gray and Prof. Alexei Drummond for their help with setting this experiment and using the BEAST software.

Aldomirovtsi:	v	ˈe	t͡ʃ	e	r	Aldomirovtsi:	d	-	n	ˈo
Asparuhovo-Lom:	v	ˈe	t͡ʃ	e	r	Asparuhovo-Lom:	d	-	n	ˈo
Asparuhovo-Prov :	vʲ	ˈe	t͡ʃ	ə	r	Asparuhovo-Prov:	d	ˈɤ	n	u
Babyak:	v	ˈe	t͡ʃ	e	r	Babyak:	d	-	n	ˈo
Bachkovo:	v	ˈe	t͡s	e	r	Bachkovo:	d	ˈɑ	n	u

Aldomirovtsi:	l	ˈɤ	s	n	o
Asparuhovo-Lom:	l	ˈe	s	n	o
Asparuhovo-Prov:	lʲ	ˈe	s	n	u
Babyak:	?	?	?	?	?
Bachkovo:	lʲ	ˈe	s	n	u

Figure 7.7: Multiple alignments for three words and five villages.

do not take into account any information related to the word level (e.g. lexical identity, lexical semantics, specific context in which certain phone occurs). In Figure 7.8 all pronunciations of the three words presented in step 1 are merged into a single set of multi-aligned strings.

Aldomirovtsi:	v	ˈe	t͡ʃ	e	r	d	-	n	ˈo	l	ˈɤ	s	n	o
Asparuhovo-Lom:	v	ˈe	t͡ʃ	e	r	d	-	n	ˈo	l	ˈe	s	n	o
Asparuhovo-Prov:	vʲ	ˈe	t͡ʃ	ə	r	d	ˈɤ	n	u	lʲ	ˈe	s	n	u
Babyak:	v	ˈe	t͡ʃ	e	r	d	-	n	ˈo	?	?	?	?	?
Bachkovo:	v	ˈe	t͡s	e	r	d	ˈɑ	n	u	lʲ	ˈe	s	n	u

Figure 7.8: Pronunciation of different words merged into a single string.

We do not use any information on where one words begins or ends. Merging all multi-aligned transcriptions in our data set resulted in 620 columns that contain either consonants or vowels. For the missing words in our data set we use symbol '?' to mark each of the positions where the corresponding phones would have been placed if the pronunciation for that village had been available. For the phones that were deleted in a certain pronunciation, we use symbol '-' in order to keep these two types of missing tokens separate.

- It is evident that these multi-aligned sequences are very different from the sequences used in biology. Our linguistic alignment contains a large number of sites, 197, and relatively short strings comprising 620 positions in total. At the same time alignments in biology would normally contain longer sequences for a much smaller number of species. The other difference is in the number of unique

tokens: for protein sequences there are 20 different proteins, while in our linguistic alignments the number of unique phonetic segments was 97: 55 for consonants and 43 for vowels. Having a large number of different symbols in some columns on one hand, and such a small number of columns on the other, makes it impossible for the algorithm to reach convergence and obtain the desired analyses correctly. For that reason the data set was reduced to only the columns that contain vowels. As we have seen in Chapter 5 vowel changes are more frequent and more diverse. Consonant changes occur much less frequently and in most of the cases involve palatalization. We argue that on dialect level, most of the information on the language change and variation can be inferred from the processes related to vowel changes. Since, for technical reasons, we are forced to reduce the number of analyzed phones, we chose to base our analyses on the vowel changes only. From the merged alignments we removed all columns that contain consonants, making the total number of columns 303. After removing all the consonants, our example presented in Figure 7.8 would look like this:

Aldomirovtsi:	ˈe	e	-	ˈo	ˈɤ	o
Asparuhovo-Lom:	ˈe	e	-	ˈo	ˈe	o
Asparuhovo-Prov:	ˈe	ə	ˈɤ	u	ˈe	u
Babyak:	ˈe	e	-	ˈo	?	?
Bachkovo:	ˈe	e	ˈɑ	u	ˈe	u

Figure 7.9: Only columns with vowels are kept in the merged multiple string alignment.

- After reducing our data set only to vowels, there were still 43 different phonetic segment symbols, including various diacritics and suprasegmentals. It is still a much larger number of segments that any software made to process biological data is able to handle. In order to get smaller number of symbols, we have removed all diacritics and suprasegmentals and reduced our set of symbols to 16. In Table 7.1 we list the reduced set of symbols on the left hand side, and the full, unreduced, set on the right hand side.

- Taking into consideration the short length of strings (303 positions), 16 different symbols was still too large a number to be processed successfully. In the final reduction step, all 16 symbols were put into one of the 8 groups based on their position in the vowel chart (Figure 7.10). Finally, the data set is transformed into the format shown in the example in Figure 7.11.

Table 7.1: Vowel inventory after removing all diacritics and suprasegmentals.

reduced set	full set			
ɑ	ɑ	ˈɑ	ɑː	ˈɑː
e	e	ˈe	eː	ˈeː
ɛ	ɛ	ˈɛ	ˈɛː	
ɤ	ɤ	ˈɤ	ɤː	ˈɤː
ɒ	ˈɒ			
ɪ	ɪ	ˈɪ	ˈɪː	
o	o	ˈo	oː	ˈoː
u	u	ˈu	uː	ˈuː
ʊ	ʊ	ˈʊ		
ə	ə	ˈə		
a	a	ˈa	ˈaː	
i	i	ˈi	ˈiː	
ɔ	ˈɔ			
ʌ	ˈʌ			
ɨ	ɨ	ˈɨ		
y	y	ˈy		

Representation of the pronunciation dialect data with only 8 symbols leads to information loss. We have completely discarded consonant changes, and, additionally, we have merged all 43 vowels into only 8 groups. However, we believe that this type of data representation still contains enough information for the exploration of dialect variation and change. As mentioned earlier, consonant changes in our data set are less frequent and less various if compared to vowel changes. For that reason, we choose to focus on vowels. We group all vowels in our data set based on their articulatory features, so that each of them can be defined based on the front/back and close/open opposition. For example, we can describe group 6 as a group comprising close front vowels. By grouping vowels in such a way, we hope to be able to discover some of the general principles of substantial vowel changes within the vowel chart.

After putting our data into the format described, our next step was to choose suitable models of sound changes. We tested three models of evolution on our data set and they will be explained in more detail in the next subsection.

7.5.1 Different models of sound change

All models implemented in the BEAST software that we have used in our experiment for Bayesian inference were originally developed to analyze molecular sequences. Among

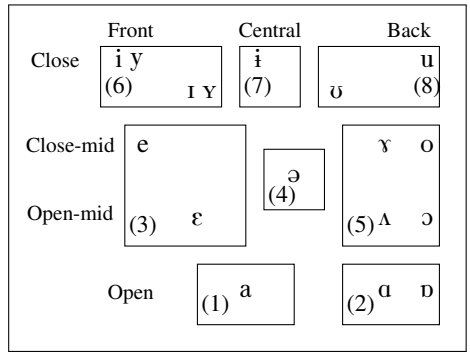

Figure 7.10: All vowels in the data set were placed into one of the 8 groups.

Aldomirovtsi::	3	3	-	5	5	5
Asparuhovo-Lom:	3	3	-	5	3	5
Asparuhovo-Prov:	3	4	5	8	3	8
Babyak:	3	3	-	5	?	?
Bachkovo:	3	3	2	8	3	8

Figure 7.11: Final format of our data set used for Bayesian inference analysis.

various possibilities, we have chosen to test three settings that can be applied to our phonetic data. In each of the settings we specify the following categories: a) a substitution model (s); b) a position heterogeneity model (h) and c) a molecular clock (m).

Substitution models for biological data describe the process of one nucleotide or amino acid being replaced by another. In our case, they describe the process of one vowel, or more precisely one of our 8 groups, being substituted for another. In Figure 7.12, we repeat the alignment presented in Figure 7.11 but mark it with 's', 'h' and 'm' to show which model applies to which part of the alignments. In our example substitution model, marked with 's', calculates the probability of group 5 being substituted for group 8, or the other way around. In this model we were not able to specify the directionality of the change. As a result we get only one probability of change for each pair of phones.

The site (position) heterogeneity model allows us to specify whether the rate of variation in different position, marked with 'h1', 'h2', ..., 'h6' in our example in Figure 7.12, is the same or whether it varies from column to column. For our data it would mean that we can specify whether vowel changes occur more frequently in some positions in words than in others. We do not specify in which positions the substitutions are more or less frequent, but some settings allow different columns to vary at different rates.

		h(1)	h(2)	h(3)	h(4)	h(5)	h(6)
m(1)	Aldomirovtsi:	3 [ˈe]	3 [e]	-	5 [ˈo]	5 [ˈɤ]	5 [o]
m(2)	Asparuhovo-Lom:	3 [ˈe]	3 [e]	-	5 [ˈo]	3 [ˈe]	5 [o]
m(3)	Asparuhovo-Prov:	3 [ˈe]	4 [e]	5 [ˈɤ]	8 [u]	3 [ˈe]	8 [u]
m(4)	Babyak:	3 [ˈe]	3 [e]	-	5 [ˈo] ↕**s**	?	?
m(5)	Bachkovo:	3 [ˈe]	3 [e]	2 [ˈɑ]	8 [u]	3 [ˈe]	8 [u]

Figure 7.12: Three models of evolution apply to the parts of the alignments marked with ‘s’ (substitution model), ‘h’ (rate heterogeneity model), and ‘m’ (molecular clock model).

For all three settings we set the molecular clock option to the strict molecular clock. This setting specifies that different branches in a tree have the same rate of variation, i.e. that different species, in our case language varieties marked with ‘m1’, ‘m2’, ..., ‘m5’, change constantly over time. This is the basic, and the simplest molecular clock model implemented in BEAST. Since in this experiment our data is rather limited, we tried to build simple models and get reliable estimates of our parameters. In the future, we would certainly like to test the relaxed molecular clock options that assume independent rates on different branches.

Our **Setting 1** is the simplest one, with the following values for the two models:

- Substitution model: any state, i.e. phone, is equally likely to change into any other state. For example, vowel [a] (group 1) can change into a vowel from any other group and the probability of, for example, [a] changing into [ə] is the same as [a] changing into [u].
- Site (position) heterogeneity model was set to ‘None’, meaning that all sounds in all positions in words evolve at the same rate.

In **Setting 2** we have the following options:

- Substitution model: General Time Reversible (GTR) model. Under a GTR model any state, i.e. phone, can change into any other, but the probability of change differs depending on the phones involved. The rate of change is not set in advance, but calculated from the data. In this setting the probability of, for example, [a] changing into [ə] is not the same as the probability of [a] changing into [u]. This allows us to calculate which phone changes are more likely than some others.
- Site (position) heterogeneity model was set to ‘None’. The same as in the Setting 1, i.e. all sounds in all positions in words are assumed to evolve at the same rate.

Setting 3 comprises the following options:

- Substitution model: General Time Reversible (GTR) model. The same as in the Setting 2, any phone can change into any other. The probability of one phone changing into the other may vary depending on the phones involved. The directionality of the change is not specified.
- Site heterogeneity model was set to Gamma. This setting allows various substitution rates between different positions, i.e. it allows for the phones in different positions within the words to evolve differently. Unlike in the previous two settings, we assume that, for example, position h(1) might evolve slower or faster than position h(6).

The length of the chain, i.e. the number of generations that the MCMC algorithm ran for, was 4×10^7 for all three settings. The trees were sampled after every 8000 generations, which gave us a final sample of 5000 trees. This number of generations was sufficient in all three runs to get a representative sample of trees.

Some assumptions made by the various models might seem more or less plausible depending on one's linguistic intuition. By using rigorous quantitative methods, we want to test the validity of different hypotheses and try to answer some questions about language evolution and change in a more exact manner. In the next section we present the results for each of the settings tested.

7.6 Results

We use the TreeAnnotator program from the BEAST package to summarize the information from the sample trees produced by BEAST into a single tree. We select the option 'maximum clade probability tree' in order to get a tree where the node height[3] and rate statistics are summarized on the tree in the posterior sample that has the maximum sum of posterior probabilities on its $n-2$ internal nodes.

In Figure 7.13 we present the dendrogram where the trees produced using Setting 1 are summarized. On all dendrograms in this section we present the posterior probabilities of nodes. Due to the large number of sites in our data set, node labels were not readable. We have removed them from all dendrograms. In the dendrogram in Figure 7.13 we can see that on the highest level the split at the root node has maximum posterior probability 1. We mark two-way split with red and blue, where red represents eastern varieties and blue western and southern. In order to see the geographical distribution of the two groups of sites, we present this two-way split of Bulgarian dialect varieties on a map (see Figure 7.15). Two groups of sites are marked with red dots (eastern varieties) and blue dots (western and southern varieties). The two-way division of

[3]The height of a node is the length of the longest downward path from that node to a leaf.

sites is geographically coherent and divides the Bulgarian language area in a such way that eastern varieties, in traditional literature referred to as Balkan and Moesian dialects, and on our map marked with red symbol are put in one group, while western and Rupian dialects, marked with blue, are put in an other group. Unlike in the aggregate analyses presented in Chapter 3, Rupian dialects are grouped together with the western, rather than with the eastern varieties. One step lower in the dendrogram, there is a split with posterior probability of 0.898. According to the analysis performed, we can assume this split with a high confidence. It divides southern varieties from the western. On dendrogram in Figure 7.14 we mark the southern varieties with green and western with blue. Classification of the western varieties into a single group is supported with maximum posterior probability, while the grouping of southern varieties is much less certain since the node that is on the top of this group has posterior probability of 0.531. Although according to the posterior probability it is not highly certain that these sites form a group, they largely occupy a geographically coherent area in the south of the country. Some of the varieties placed in this group are found along the *yat* border. We present the three-way classification produced using Setting 1 in Figure 7.16. Based on the branch lengths in the dendrogram, groups presented on this map form three distinct varieties. Since in Setting 1 the probability of any state, i.e. any phone changing into any other state was set to be equal we could not get any interesting information on vowel changes from this setting.

In Figure 7.17 we present the tree that summarizes the trees resulting from the Bayesian inference performed once we adopted the General Time Reversible (GTR) model. The two-way split at the root node that has maximum posterior probability, shows a split of the sites into western and eastern. The southern group of varieties is classified with the eastern dialects (Figure 7.17). Just as with the previous dendrogram, we show this split on the map of Bulgaria (Figure 7.19). This division corresponds well with the division of the sites based on the aggregate analysis (Chapter 3) since the split follows approximately the *yat* line and groups all the sites into eastern and western. Unlike in the Setting 1, varieties in the south are grouped with the eastern dialects (see map in Figure 7.20). However, the support for this grouping is relatively low (0.505) and cannot be taken with any great confidence. Groupings of both southern and eastern varieties have low posterior probabilities, namely 0.134 and 0.526. The former has little basis in the model. Unlike the eastern division of the sites, the western varieties are grouped under the node with the high posterior probability and can be taken with great confidence to form a coherent group. Apart from reconstructing phylogenies, i.e. grouping of the varieties, Setting 2 also allows us to investigate how probable certain sound changes are. In Setting 2, we used a General Time Reversible Model to model sound changes. As a reminder, we recall that any group of sounds was allowed to change into any other group but the changes did not receive equal probability as in the Setting 1. One of the outputs of the Bayesian inference analysis were the probabilities of change between each two groups of sounds calculated from the data.

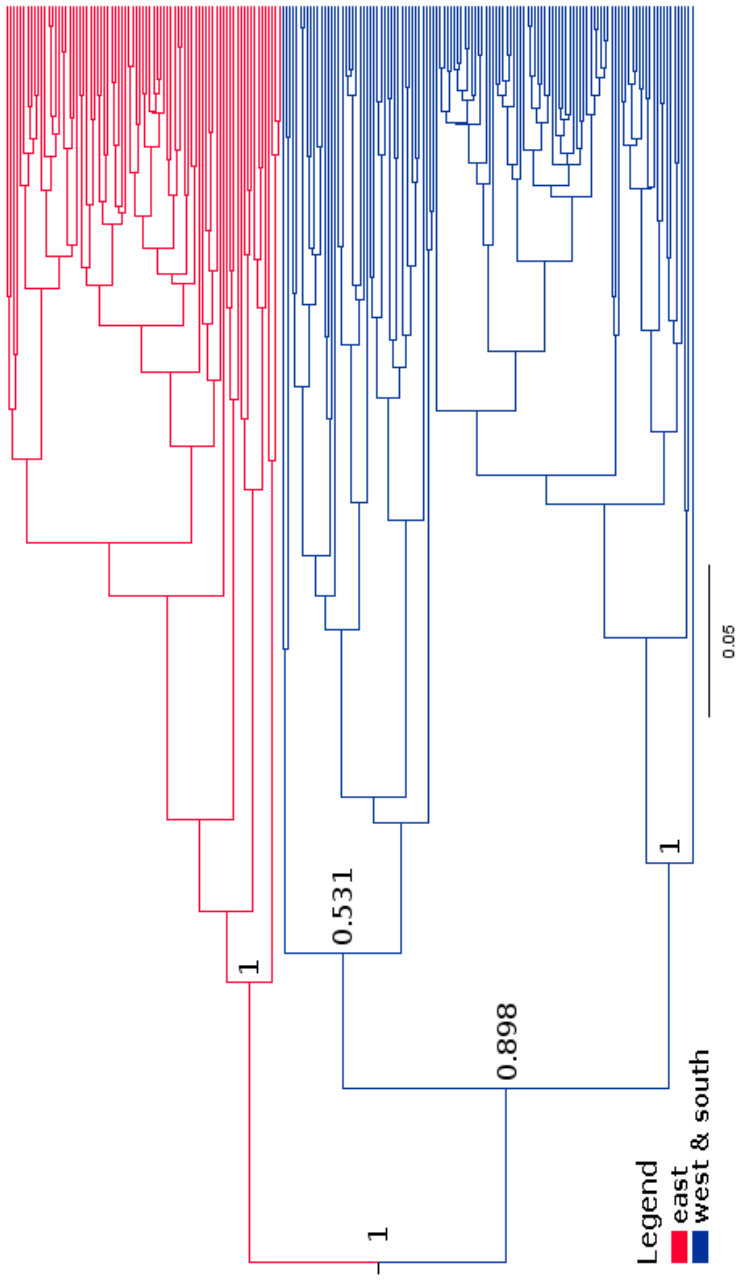

Figure 7.13: Dendrogram that summarizes the trees produced using Setting 1: free substitution model and no positional heterogeneity. Two-way division of the sites.

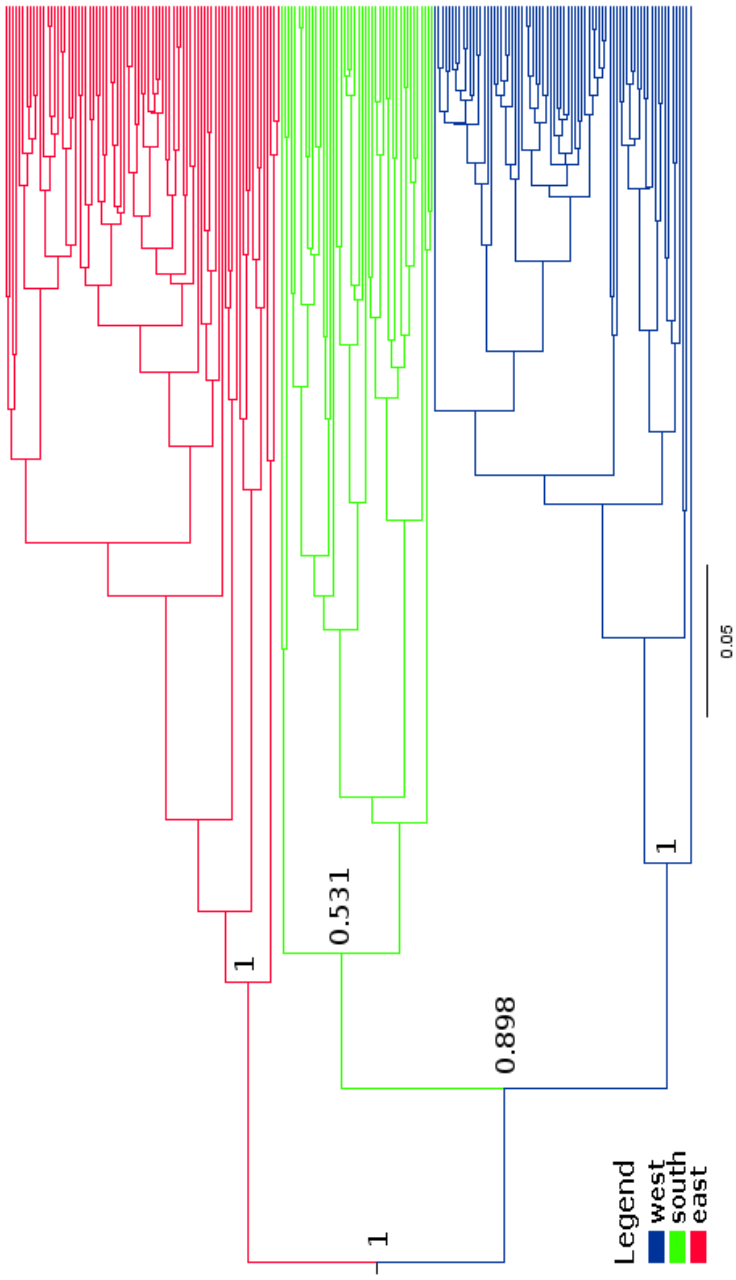

Figure 7.14: Dendrogram that summarizes the trees produced using Setting 1: free substitution model and no positional heterogeneity. Three-way division of the sites. This is a more detailed view of the dendrogram in Figure 7.13

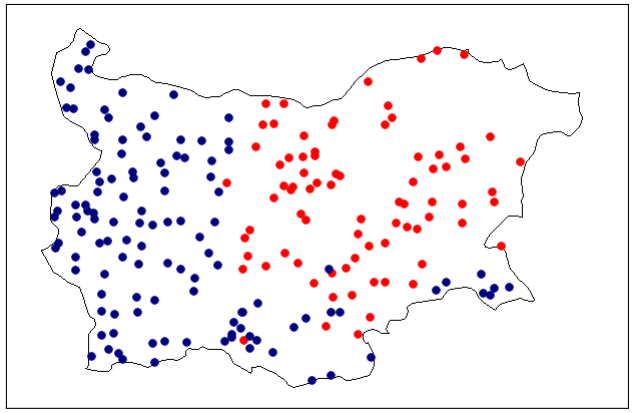

Figure 7.15: Distribution of the two group of sites using a free substitution model and no positional heterogeneity model (Setting 1).

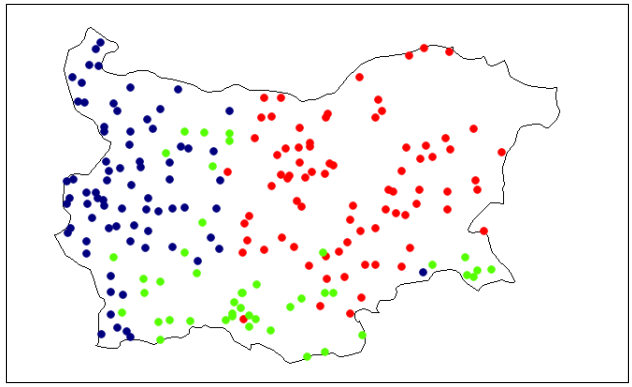

Figure 7.16: Distribution of the three group of sites (Setting 1).

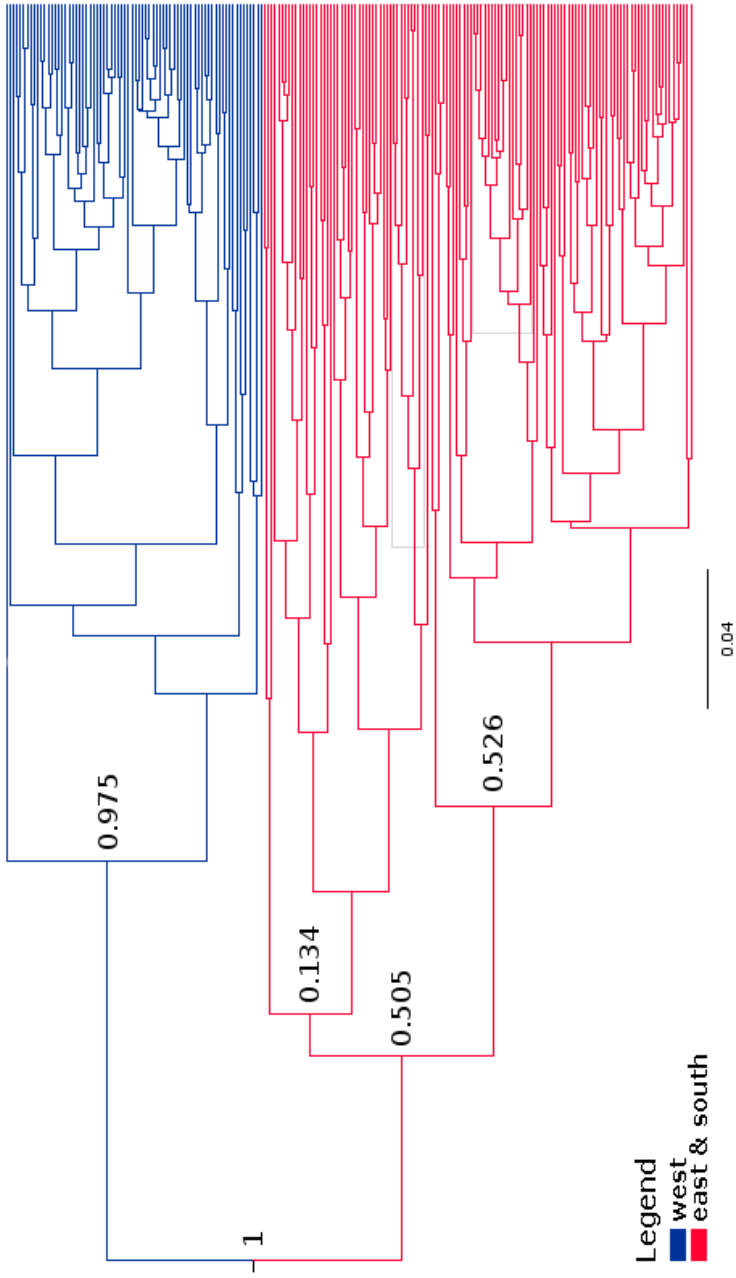

Figure 7.17: Dendrogram that summarizes the trees produced using Setting 2: GTR model with no positional heterogeneity. Two-way division of the sites.

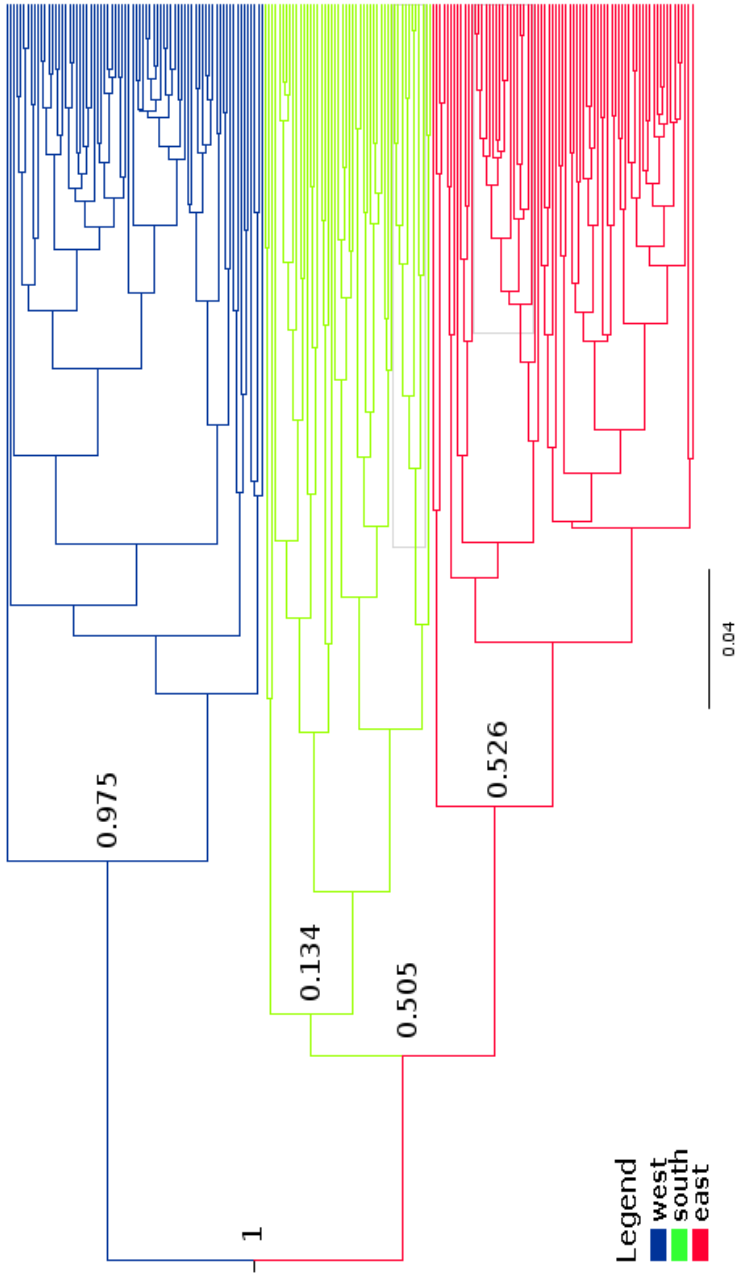

Figure 7.18: Dendrogram that summarizes the trees produced using Setting 2: GTR model and no positional heterogeneity. Three-way division of the sites.

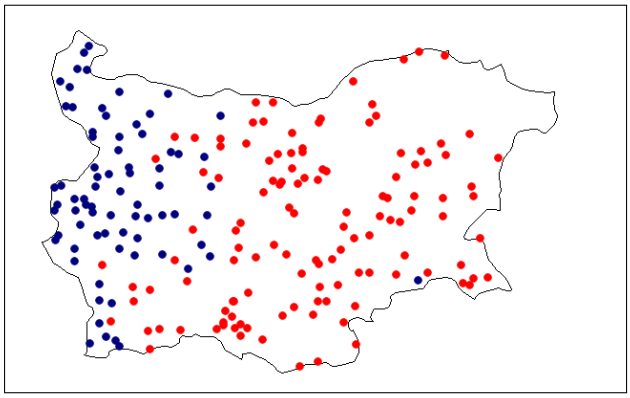

Figure 7.19: Distribution of the two group of sites using a GTR substitution model and no positional heterogeneity (Setting 2).

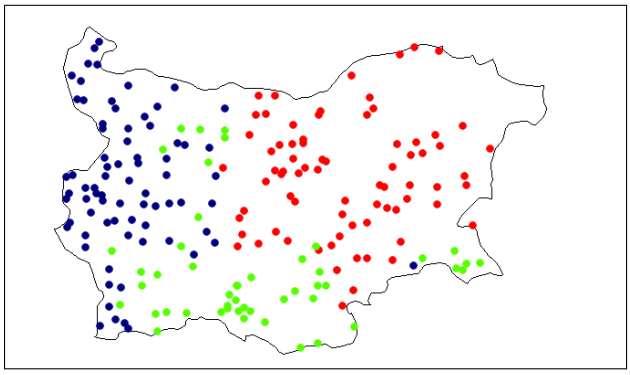

Figure 7.20: Distribution of the three group of sites (Setting 2).

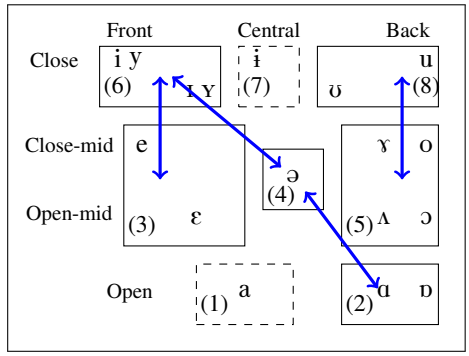

Figure 7.21: The most probable vowel transitions, marked with blue, under the GTR model no positional heterogeneity. Groups (1) and (7) are put in dashed boxes to indicate that our estimations concerning these groups are unreliable.

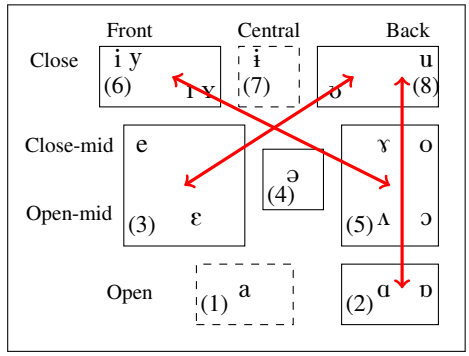

Figure 7.22: The least probable vowel transitions, marked with red, under the GTR model with no positional heterogeneity.

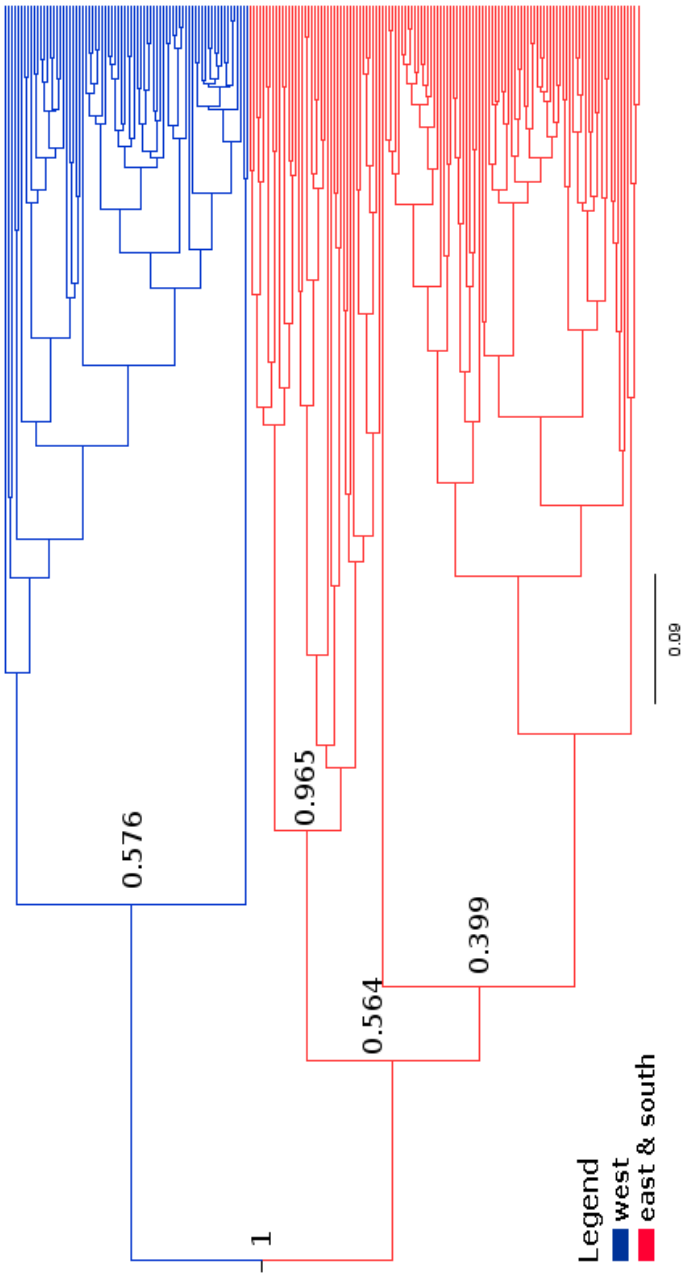

Figure 7.23: Dendrogram that summarizes the trees produced using Setting 3: GTR model with gamma positional heterogeneity. Two-way division of the sites.

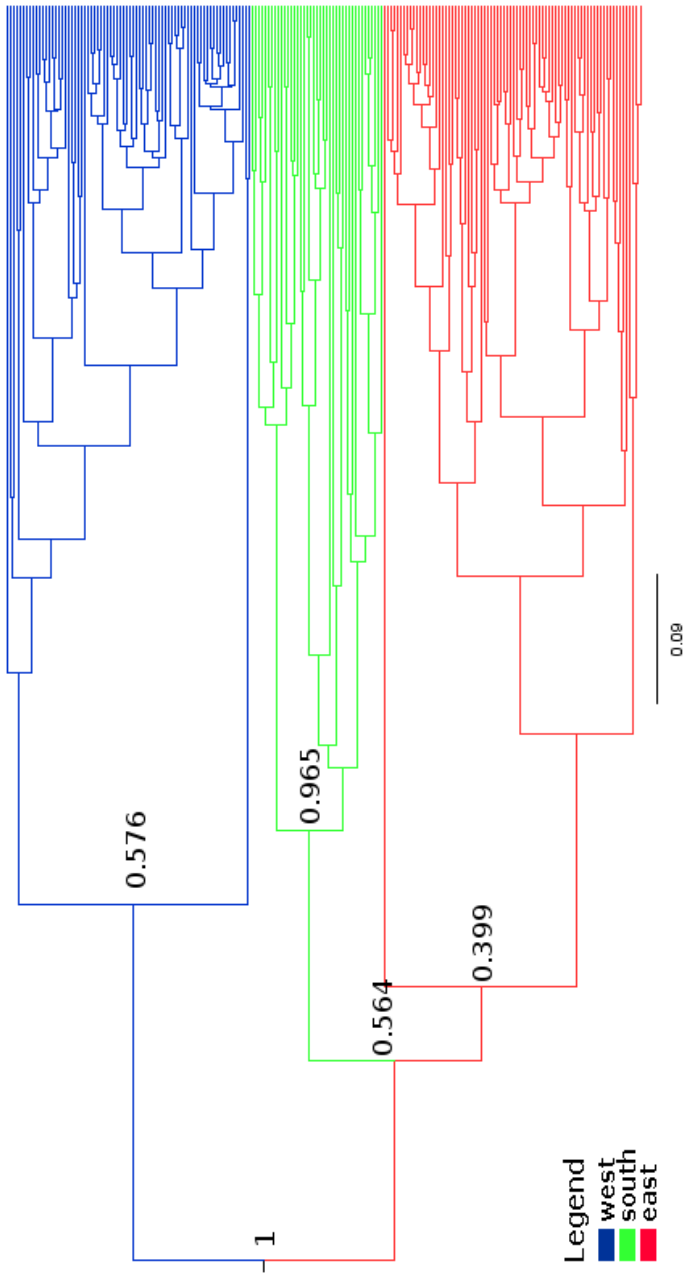

Figure 7.24: Dendrogram that summarizes the trees produced using Setting 3: GTR model and gamma positional heterogeneity. Three-way division of the sites.

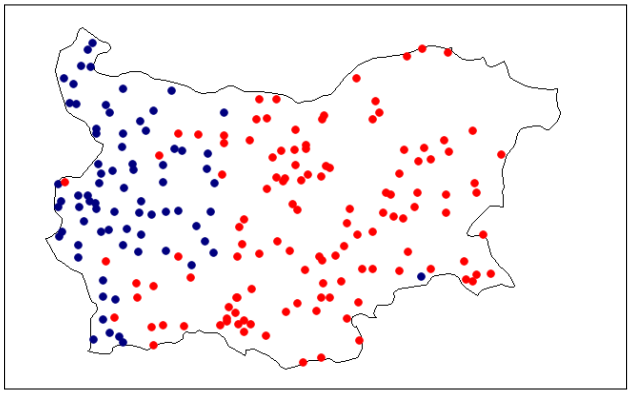

Figure 7.25: Distribution of the two group of sites using a GTR model with gamma positional heterogeneity (Setting 3).

The results can be seen in Figure 7.21, where we present sound changes with the highest probabilities (connected with blue lines) and in Figure 7.22 where we show changes that have the lowest probability (connected with red lines). For clarity, we put both numbers and sounds in the charts. Since all our sounds in the data are put into one of the eight groups, we can naturally talk only about how probable the change of a vowel in one group into a vowel in another is. In Figures 7.21, 7.22, 7.27 and 7.28 groups 1 and 7, which stand for [a] and [ɨ] sounds are put in dashed squares since we could not get any reliable estimations for them. The reason for this is their very low frequency in the data set. The sound [a] appears only 147 times in our multiple alignment, while the sound [ɨ] is present only 40 times. Vowels from the third group [ɛ, e], which is the most frequent group in the data set, appear 14663 times. As marked with the blue lines in the vowel chart in Figure 7.21, changes that received the highest probability are between the following groups: 5 [ʌ, ɔ, ɤ, o] and 8 [ʊ, u], 3 [ɛ, e] and 6 [ɪ, y, i], 4 [ə] and 6 [ɪ, y, i], and 2 [ɑ, ɒ] and 4 [ə]. We can see in the chart that those changes involve moving only one step within the vowel chart. Unfortunately it was not possible to infer the directions of the changes and see whether, for example, it is more probable that vowels from group 3 would change into vowels from group 6 ($3 \rightarrow 6$) or the other way around ($6 \rightarrow 3$). However, our findings correspond well with the findings reported in the literature on the traditional analyses of the vowel reduction in Bulgarian (Wood and Pettersson, 1988; Barnes, 2006). According to them the most common vowel change in Bulgarian dialects is rise of unstressed midvowels [e] and [o] to neutralize with the high vowels [i] and [u]. The low unstressed vowel [a] rises to neutralize with [ə].

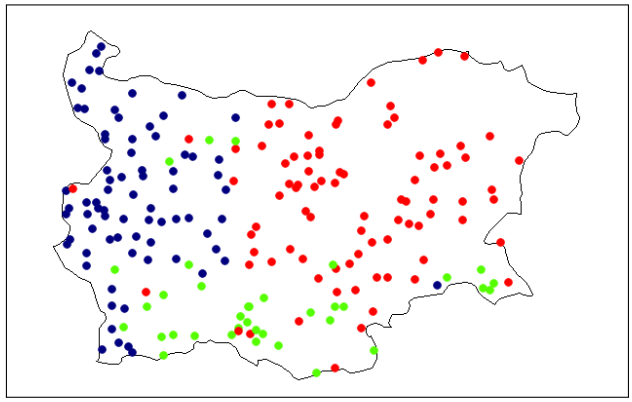

Figure 7.26: Distribution of the three group of sites (Setting 3).

In the chart in Figure 7.22 we mark the changes between the groups with the lowest probabilities using red lines: 2 [ɑ, ɒ] and 8 [ʊ, u], 3 [ɛ, e] and 8 [ʊ, u], and 5 [ʌ, ɔ, ɤ, o] and 6 [ɪ, y, i]. In contrast to the alternations with the highest probabilities, they do not involve changes between the adjacent groups but rather between the groups separated by at least one group within the vowel chart.

In the Setting 3 under the General Time Reversible model, just as in the Setting 2, every state was allowed to change into any other state with the transition probabilities being inferred from the data. It was again not possible to calculate the directionality of the changes. The difference between the two settings is that in the Setting 3 the positions in the alignments were allowed to vary at different rates. From the dendrogram in Figure 7.23 we also extracted the two-way division of the sites and represented it on the map in Figure 7.25. In Figure 7.24 we mark three groups extracted and show that division in Figure 7.26. Both the two-way and the three-way divisions of the sites are almost identical to the divisions for Setting 2: the first one goes along the *yat* line, while the second additionally distinguishes the southern area as separate. Division into western and eastern dialects gets the highest posterior probability, while other major splits were supported with much smaller posterior probabilities.

In Figure 7.27 and Figure 7.28 we present vowel charts with the changes that are the most and the least probable. The sound changes with the highest probabilities are those between the groups 5 [ʌ, ɔ, ɤ, o] and 8 [ʊ, u], 3 [ɛ, e] and 6 [ɪ, y, i], and 4 [ə] and 6 [ɪ, y, i]. Just as in the previous analysis, sound correspondences that involve two adjacent groups within the vowel chart are the most probable. The least probable sound correspondences include alternations between the sounds that are more than one step apart within the vowel chart.

In Table 7.6 we give the values of the modified Rand index (MRI) presented in Sec-

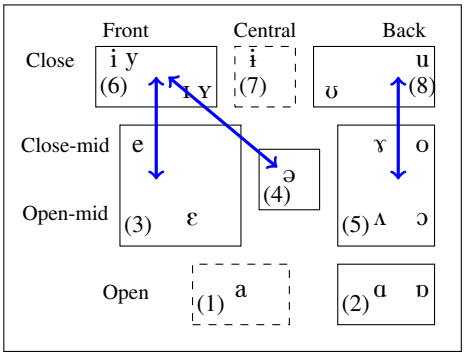

Figure 7.27: The most probable vowel transitions, marked with blue, using GTR and gamma site heterogeneity model.

tion 3.6.1 for the pairwise comparison of the classifications produced in all three settings, classification done by weighted pair group method using arithmetic averages (WPGMA) clustering algorithm and the traditional division of the sites according to Stoykov (2002). We note very high agreement between the 2-way divisions produced using Setting 2 and Setting 3: 0.939. There is also very high agreement between the 3-way divisions produced by Setting 1 and Setting 2: 0.945. Agreement on the 2-way and 3-way divisions produced by Setting 2 and Setting 3 in Bayesian inference experiment and WPGMA clustering algorithm is lower, but still high, ranging from 0.686 to 0.763. The 2-way division produced by the Setting 1 has lower values for MRI since, unlike WPGMA, it groups southern varieties with the western and not with the eastern dialects. Comparison of the divisions resulting from Setting 2 and Setting 3 to the traditional divisions as suggested by Stoykov (2002), shows that they give similar values of MRI that we get by comparing the divisions produced by WPGMA and traditional classification.

Settings 2 and 3 gave very similar results, both with the respect to the classification of villages and to the vowel transition probabilities. Although the results were similar, the two settings contain two different hypotheses about sound changes. In Setting 2 we assume that in all positions in words sounds change at the same rate. In Setting 3 we allowed that at some positions in words some sound changes are more likely than in some others. In order to check which of the two hypotheses is more probable, we calculated Bayes factor (K) for the two settings, which is a Bayesian alternative to a classical hypothesis testing in statistics. The Bayes factor was calculated using the following formula which examines the ratio of the marginal likelihoods of the two models:

$$K = \frac{P(D|H1)}{P(D|H2)}$$

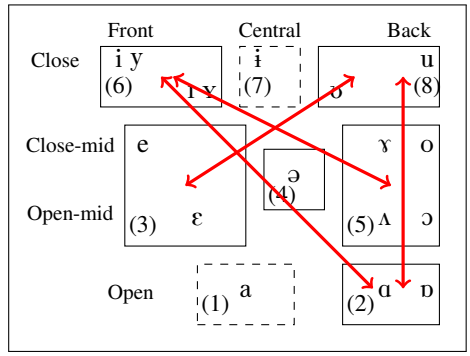

Figure 7.28: The least probable vowel transitions, marked with red, using GTR and gamma site heterogeneity model.

where $P(D|H)$ expresses the marginal likelihood of a hypothesis H. For a more detail explanation see Kass and Raftery (1995) or MacKay (2003). For our two settings we calculated the Bayes factor using the Tracer software.[4] In Table 7.3 we present the values of the Bayes Factor in log 10 scale obtained after pairwise comparing all three settings.

All values of K $>$ 2 for the log 10 scale indicate strong support for a favored model. All values for comparing our three settings are much bigger. It shows that there is a very strong evidence in favor of Setting 3. Setting 2 is much more strongly supported than the Setting 1, while the Setting 3 is much more strongly supported than the Setting 1 and 2. Explanation of the scale for K can be found in Jeffreys (1961) and Kass and Raftery (1995). These results show that there is a strong evidence in our data that different vowel changes are not equally probable. Some changes are much more likely to occur than others. The data also strongly supports the hypothesis that vowel changes occur at different rates in various positions in words.

7.7 Discussion

In recent years there has been an increasing number of studies that apply methods taken from phylogenetics to the research of language change and evolution. However, only very few of them apply those techniques on the phonetic or phonological data (Nakhleh, Ringe, and Warnow, 2005; Warnow et al., 2006; McMahon et al., 2007). In previous work phonological and phonetic data was very carefully manually selected and coded, based on the substantial linguistic knowledge. We do not argue against the linguistic

[4] http://tree.bio.ed.ac.uk/software/tracer

Table 7.2: The modified Rand index (MRI) for the 2-fold and 3-fold divisions of sites produced by three various Bayesian inference settings ('s1', 's2', 's3'), WPGMA ('WA') and traditional scholarship ('trad.').

	s1 (2)	s2 (2)	s3 (2)	WA (2)	trad. (2)	s1 (3)	s2 (3)	s3 (3)	WA (3)	trad. (3)
s1 (2)	-	0.301	0.312	0.467	0.290	-	-	-	-	-
s2 (2)	-	-	0.939	0.717	0.716	-	-	-	-	-
s3 (2)	-	-	-	0.734	0.665	-	-	-	-	-
wa (2)	-	-	-	-	0.700	-	-	-	-	-
trad. (2)	-	-	-	-	-	-	-	-	-	-
s1 (3)	-	-	-	-	-	-	0.945	0.854	0.727	0.601
s2 (3)	-	-	-	-	-	-	-	0.829	0.763	0.597
s3 (3)	-	-	-	-	-	-	-	-	0.686	0.543
wa (3)	-	-	-	-	-	-	-	-	-	0.626
trad. (3)	-	-	-	-	-	-	-	-	-	-

Table 7.3: Values of the Bayes factor in log 10 scale. There is a strong support for Setting3 when compared to both Setting1 and Setting2.

	Setting 1	Setting 2	Setting 3
Setting 1	-	-573.369	-938.271
Setting 2	573.369	-	-364.902
Setting 3	938.271	364.90	-

data coding, but we do try to apply a more robust and language-independent approach. In this research we have tried to automate the process of character selection by automatically multi-aligning phonetic transcriptions and using them as input to software for phylogenetic inference. However, we were not able to explore sound correspondences in all their varieties, since the number of phones in our data set was too large to be successfully processed by any software developed for the computational phylogenetics. We have restricted ourselves to the investigation of the vowel changes, since the analysis of the sounds presented in Chapter 5 has shown that most of the variation in our dialect data is between the vowels. In comparison to the consonants, they are more likely to contain sufficient information on dialect change. By putting all the vowels into eight groups we have tried to keep in our analyses at least the main articulatory characteristics (open/close and front/back opposition) of the vowels. This multi-state character encoding enabled us to test the probability of sound changes within the vowel chart. The coding of the characters can naturally be done differently, but we leave this to future research. We hope that in future it will become computationally feasible to process the data using a larger set of states.

The application of Bayesian inference allows us to test various models of evolution

and to investigate how related certain species are. By applying this method to the phonetic data, we were able to test various hypothesis about the mechanisms of sound change. Each model of evolution contains its own explicit assumptions. Relying on the models of evolution created for biological data, we were forced to draw parallels between the evolution of species and the evolution of languages. But very often models developed for the evolution of species contain assumptions that are not very realistic for the language data. For example, all character-based methods, including the Bayesian inference of phylogeny, assume that each position in the alignments evolves independently. For our phonetic data, it would mean that the changes of the phones are not influenced by the changes of the proceeding or the following sounds. Although this is not true for the mechanism of a sound change, it is one of the simplifications that we had to introduce in our analyses. In future we hope to implement a model that would relax the assumption of independence, at the cost of substantial complexity.

One of the models that is being heavily debated in linguistics is the lexical clock. While some of the authors used this assumption in their attempts to date Proto Indo-European (Forster and Toth, 2003), others heavily criticize the usage of a uniform lexical clock (Eska and Ringe, 2004). A strict molecular clock model assumes that all lineages (language varieties) evolve at a constant rate. We have used this assumption in our experiments since it is the basic molecular clock model in the software for phylogenetic inference, and it makes the estimation of the other parameters easier, especially with such a small data set as ours. All the trees produced in our experiments have shown an expected topology (structure), which suggests that the assumption of a constant molecular clock is not extreme a simplification in the models examined here. These were, however, initial experiments and in the future, we would like to apply other molecular clock models, and statistically test wether other molecular clock hypotheses fit our data better.

By initially choosing simple models of evolution to be tested on our language data, we have tried to justify more complicated assumptions step by step. None of the models developed for the biological data can cover all aspects of language evolution and change. The possibility to test various hypotheses separately makes Bayesian inference a potentially very useful technique in exploration of languages. But its true potential in linguistics can be achieved only if models are developed specifically for language data.

The results of applying Bayesian phylogenetic inference to Bulgarian dialect data have shown that three dialect areas appear as the most prominent under various models of evolution: western, eastern, and southern. This three-way division also conforms to the traditional scholarship on Bulgarian dialectology (Stoykov, 2002). We have obtained the same division of Bulgarian dialect area using the Levenshtein method that is based on the similarity between the pronunciation strings without any assumptions on the genetic relatedness of the compared varieties. Two alternative approaches gave very similar picture of the Bulgarian dialect variation. However, these two approaches are very hard to separate in the case of dialect data where we *a priori* test varieties that are genetically

very closely related.

We have also shown that for the Bulgarian language the most probable vowel changes are those that involve neighboring vowels within the vowel chart. Most of the highly probable changes involve vowel height. The probability of vowels changing into vowels that are far apart in the vowel chart is very low. We were not able to include the directionality of vowel changes into our analysis, and see if, for example, [e] is more likely to change into [i] or the other way around. We hope to achieve this in future. Testing of different models of evolution has also shown that vowels change faster in some positions within the words. In future we would like to investigate changes of various positions in multi-aligned sequences in more detail and try to discover patterns of variation and how regular certain sound changes are.

Chapter 8

Conclusions and discussion

The aim of this thesis was to develop and apply a quantitative analysis of the Bulgarian dialect pronunciation data. The data set used in this thesis was gathered and put into a machine-readable format as part of the *Buldialect* project. It consists of 157 transcribed words collected at 197 sites distributed over most of Bulgaria. The main source of the data was the large dialect archive at the University of Sofia. The words in the data set contain in total 39 various phonetic features that are commented on in the traditional scholarship on Bulgarian dialects and which have been used as a basis for determining dialect divisions. The most widely known and the most authoritative study of Bulgarian dialects is one published by Stoyko Stoykov (Stoykov, 2002). Throughout this thesis we use his classification of Bulgarian dialects against which we compare our computational methods. Main dialect divisions suggested by Stoykov are presented in Chapter 2. The data for *Buldialect* project was collected in a such way that there is a balance between various phonetic features that Stoykov (2002) uses as a basis for classification of Bulgarian dialects. In Chapter 2 we give a list of the phonetic features present in our data set, and in Appendix A we list the words from the data set and additionally mark which phonetic features are present in which word.

In the first experiments on the Bulgarian pronunciation data, we have used the Levenshtein algorithm to measure the differences between Bulgarian dialect varieties. We used the simple version of the Levenshtein algorithm, where weights were set to make it impossible for vowels and consonants to align. The distances between each two sites in the data set were analyzed using multidimensional scaling (MDS) and numerous clustering techniques, neighbor-joining and neighbor-net. MDS is a dimension-reduction technique, used to look if there are any distinct clusters in the data. The analysis has shown that there are two clearly separated groups of dialects and the third one that is at a remove from them. Multidimensional scaling proved to be quite reliable in the exploration of continuous data, like ours, since it can detect if there are any distinct groups in

the data. The results of applying different classifying techniques were compared to each other and to the traditional scholarship. We proposed several methods that were used to compare the outputs of the classification algorithms. Some of them, like the modified Rand index, entropy and purity, require the existence of a gold-standard classification provided by the experts in the field. The other evaluation methods, such as the cophenetic correlation coefficient, noisy clustering and consensus dendrograms, can be used in a more realistic scenario when the classification provided by the experts is not available. They do not rely on a comparison to any *a priori* structure, but try to determine if the structure obtained by the classification algorithm is appropriate for the data. From the methodological side, the results have shown that clustering algorithms should be used with great caution in dialectometry since there are often no sharp borders between the dialect varieties. There is no one single algorithm that we can use to obtain reliable classifications. We can only look for the most probable dialect divisions by applying some of the techniques presented in Chapter 3. Our results have shown that three hierarchical clustering techniques, namely single link, unweighted pair group method using centroids (UPGMC) and weighted pair group method using centroids (WPGMC), failed to identify any structure in the data. The rest of the clustering techniques tested gave different results depending on the level of hierarchy. All algorithms had high agreement on the detection of the two main dialect areas within the dialect space, the western and the eastern varieties along the *yat* line. Though less consistently, we could also identify the Rodopi area in the south of the country. No other dialect groups were identified in a consistent manner. These results correspond well with the division suggested by Stoykov, but are of course less elaborate. The results of the neighbor-joining algorithm were less satisfactory, most probably due to the continuous structure of our dialect data. Neighbor-net has proven to be a nice representation tool, since it can tell us if the data is a tree- or a net-like. Using neighbor-net we have detected many conflicting signals and showed that Bulgarian dialect data is to a high extent network-like.

In Chapter 4 we compare traditional dialect divisions suggested by Stoykov to the divisions that we obtain using various clustering techniques. We focus mainly on the differences between traditional and computational methods and try to explain them by comparing two classifications on the level of a very fine detail. We look into the features responsible for each of the six main traditional divisions of the Bulgarian dialect area, check their distribution in our data set and how they are reflected in the aggregate analysis done using the Levenshtein algorithm. We applied the Levenshtein algorithm to the word segments that reflect specific traditional divisions and also to the words that contain those segments in order to check how and whether the traditional division in question would be reflected in our computational analyses. The distances obtained using the Levenshtein algorithm were analyzed using MDS plots. The results have shown that, with some differences in frequencies, all the examined features are present in our data and that our data set is a reliable basis on which to compare quantitative and traditional classifications. The results also suggest that the differences between computational and

traditional approaches cannot be attributed to a single factor. Regarding the most prominent division into the western and eastern dialects, along the *yat* line, the border between the two areas on the computational maps is further east. On the quantitative maps this border represents the average of all isoglosses in the bundle of 68 which we have detected in our data. The traditional *yat* border matches few of the isoglosses found perfectly. The difference between the computational and traditional border can be attributed to the different criteria used to define the line of separation between two dialect areas. Unlike the west-east division, which showed up in all computational analyses, the Moesian area could not be detected since none of the features mentioned in the traditional literature were characteristic only for this area. As far as the phonetics is concerned, we did not find enough evidence that Moesia is a separate area. There were probably some non-phonetic factors that the traditional linguists took into account while defining this area as one of the six most important dialect areas in Bulgaria (although Stoykov emphasized that his divisions were based on pronunciation). The area around the border with Serbia, the so-called transitional zone, appears as a separate zone on all MDS plots, both based on the relevant segments and the whole words as well. We attribute the fact that some of the clustering techniques, like UPGMA, fail to recognize it to a shortcoming of the clustering technique itself. The northwest-southwest split is detected on MDS plot only if we base our Levenshtein analysis on specific segments (features), while there is no clear distinction between these two areas if we repeat the same analysis using whole words that contain the relevant segments. The comparison of the analyses done on the segment and on the word level has shown that if we perform analysis only on the relevant segments we can see the divisions clearly, while the signal gets weaker, or even lost, if we take whole words into account. The additional segments add noise to the signal of separation.

We conclude that while some of the differences between the traditional and computational divisions can be attributed to the way we calculated the distances using the Levenshtein method, the others are the result of how the dialect borders are defined in the traditional and the computational approach. While computational techniques rely only on the data that is analyzed using exact methods, the divisions done in traditional scholarship are very often more subjective and maybe led by some extra-linguistic factors. Some differences can be attributed to the biases of certain clustering techniques, which is why we argue that MDS is more suitable technique for the continuous data such as dialect data.

In Chapter 5 we have applied pointwise mutual information (PMI) technique to a table summing the frequency with which one segment aligns with another in order to automatically induce the distances between the phones in the data set. PMI was combined with the Levenshtein algorithm, which enabled us to obtain the distances between each two vowels and each two consonants. Since the Levenshtein algorithm was used with the vowel-vowel consonant-consonant constraint we never obtained non-zero frequencies with which vowels and consonants aligned. The idea behind the PMI pro-

cedure is that segments that tend to correspond more frequently in the alignments are closer to each other than the segments that rarely or never align. We analyzed the PMI distances using MDS plots in order to discover which phones tend to correspond more frequently. We were especially interested in whether there are any patterns in frequently co-occurring sounds. The MDS analyses have shown that vowels tend to vary more frequently than the consonants which resulted in much smaller PMI distances between the vowels than between the consonants. The analysis of the vowel PMI distances has shown that the changes between the unstressed vowels are much more frequent than the changes between the stressed vowels or between the stressed and unstressed vowels. The MDS plot of the vowel distances also revealed that the separation between the front and back vowels is bigger than the separation between the high and low vowels. The reason for this are smaller PMI distances between high and low vowels caused by their frequent co-occurrence in the alignments. The analysis of the distances between the consonants has shown that the consonants change less frequently than the vowels. The only pattern of change that we could discover using the MDS plot is that consonants most frequently correspond with their palatalized counterparts. No other pattern of the corresponding consonants was discovered.

We have also shown that by using these PMI induced distances in the Levenshtein alignment procedure we can get more accurate alignments compared to the alignments produced with only the vowel-vowel consonant-consonant constraint. The percentage of the incorrect alignments was reduced from 7.614 per cent to 6.236 per cent. This improvement was also reflected in the better estimation of the distances between the language varieties at the aggregate level. The main drawback of the procedure in which we have combined PMI and the Levenshtein algorithm is that we could not calculate the distances between the vowels and the consonants. We had to introduce the restriction that the consonant and the vowels cannot be aligned, since without this constraint Levenshtein algorithm produces alignments of a low quality that cannot be used to accurately estimate the distances between the phones.

In Section 6 we have presented an adapted version of the ALPHAMALIG algorithm, that can be used to multi-align strings in linguistics. Multiple alignments of strings is used, for example, in comparative method to detect sound correspondences. Here we tried to automate the process, which is a necessary step for working with larger data sets. This format of data, when compared to the pairwise-aligned strings produced by the Levenshtein algorithm, allows us to detect the patterns of phone correspondences much easier and much more accurately. It also gives us a better estimation of the distances between the strings. We have applied the ALPHAMALIG algorithm to our phonetic data and evaluated the alignments produced using two novel techniques. Both evaluation techniques are based on comparing the automatically aligned strings to the so-called gold standard alignments produced by the experts in Bulgarian phonetics/phonology. They compare the contents of the columns, i.e. positions in word transcriptions, in the two multiple alignments compared. While one of the evaluation techniques takes into

account the order in which columns appear, the other is focused solely on the content of the positions examined. Application of the two evaluation techniques has proven that the automatically multi-aligned strings are of a good quality when compared to the manually multi-aligned data. Using the first method, ALPHAMALIG scored 0.932 out of 1.0, while according to the second method ALPHAMALIG scored 0.982 out of 1.0. Although the alignments produced were of a good quality, the error analysis has shown that some of the errors are caused by the constraint that vowels cannot be aligned with the consonants. As an input the algorithm needs to know the alphabet, i.e. the segments that need to be aligned and the distances between each two segments. In our experiment we have set the weights between the segments so that vowels and consonants cannot be aligned. In the future we would like to introduce some kind of feature weighting into the alignment procedure in order to correct some of the errors present in the current alignments.

In order to get better insight into the quality of the alignments produced by ALPHAMALIG, we have also created simple and advanced baseline alignments and compared them to the gold standard alignments. The results have shown that ALHPAMALIG produces alignments of a better quality than any of the baseline techniques proposed. However, the comparison between the simple baseline and the gold standard alignments has revealed that our data set contains strings with a relatively simple CV syllable structure. The variation in the pronunciation is also relatively small if compared to cross-linguistic data. For that reason it would be necessary to validate the performance of the ALPHAMALIG against some other language data.

By multi-aligning phone transcriptions from our data set, we were able to analyze them using a Bayesian inference method designed to analyze DNA or protein sequences in molecular biology. We use Bayesian phylogenetic inference method in order to reexamine the relatedness of Bulgarian dialect varieties from a historically motivated perspective. It is an alternative to the Levenshtein approach (used in Chapters 3 and 4) which is focused on the similarity of Bulgarian dialects.

First we had to code our data in a way that would on one hand be acceptable to the software and on the other linguistically motivated. The biggest problem was large number of different phones in the data, 98 in total including all diacritics and suprasegmentals. Software designed for DNA or protein sequences can normally process up 21 different symbols used for protein data. Another issue with our linguistic data were short strings, 620 phones per each string, compared to much longer sequences in biology. Even if we reduce our set of symbols to 21, we would not be able to get reliable estimates of our parameters using such a small segment inventory. Considering the large number of tokens and relatively short strings, we reduced our data only to vowels and placed all vowels in the eight groups based on their position in the vowel chart. This enabled us to test some general principles of vowel changes within the vowel space. One of the main issues with applying computational phylogenetic methods on the linguistic data in general, is the amount of the data in linguistics that we can gather. This problem

looms particularly large in historical linguistics where we do not have large data bases and where the collection of new data cannot be done automatically since it needs to be carefully prepared.

In our experiment we have tested two hypotheses about vowel changes. We were interested to see a) if vowels change more frequently in some positions in words and b) which vowel changes are the most likely. The results have shown that there is strong support for the hypothesis that in some positions vowels change must faster. There was also very strong support for the hypothesis that vowels are not likely to change into just any other vowel, but change into vowels that are very close in the vowel chart. For our data set the most probable changes were those that involve change of vowel height. Unfortunately, it was not possible to calculate which direction of the changes are more probable. We hope to achieve this goal in the future.

Regarding dialect divisions in Bulgaria, the results of applying Bayesian MCMC inference to the Bulgarian pronunciation dialect data correspond well to the findings obtained using the Levenshtein method. The most prominent dialect division follows the *yat* line and divides the Bulgarian dialect area into western and eastern. The third area that appears as the most important under the various models of evolution is the Rodopi area in the south.

In our experiment with the Bayesian inference, we included a strict molecular clock hypothesis in our calculations. For all our settings, the resulting trees have shown the expected topology, i.e. structure of the dendrogram, with no major differences when compared both to the traditional dialectology and the computational methods applied earlier. Although we could not test the constant molecular clock assumption, we note that our finding good classifications suggests that it is a reasonable simplification. Given the fact that we were not using Bayesian MCMC to infer any dates related to the history of Bulgarian language, we find that molecular clock assumption can be used as a starting point for the experiments. However, in future we would like to repeat our experiments without the molecular clock assumption and compare the results of these two experiments.

Despite some significant differences in the evolution of species and languages, the general mechanism of evolution that they share allows us to try to take the advantage of the very powerful computational techniques developed in biology to address some problems in linguistics. The models that we have tested in this thesis are relatively simple models of the evolution of species that can be applied to linguistic data. Only if models specifically designed for linguistics are developed will we be able to have complex models that cover more aspects of the evolution of language. In the meantime we have to try to find the appropriate models, although probably not perfect.

In the future we hope to be able to use a larger set of segments in the analyses which would enable us to code the data differently and reduce the information loss introduced by putting all vowels in our data set into 8 groups. We would also like to try to introduce directionality of the phone changes in our analyses and examine in more detail the

patterns of sound changes. In this research we have restricted ourselves to the vowel changes, but it would also be interesting to reexamine our findings by exploring variation of the consonants. One of the results of our experiment has proven that sounds vary at a different rate in different positions in words. Further research might investigate which positions in words shows similar patterns of variation and how regular sound changes are. A number of possible future studies using the same experimental set up are apparent.

List of abbreviations

aor	aorist
BDA	*Български диалектен атлас* – Stoykov (1966) and Stoykov et al. (1964; 1974; 1981)
CBM	character-based methods
CV	consonant vowel
D	data
DNA	deoxyribonucleic acid
E	entropy
fem	feminine
GA	generated alignment
GS	gold standard alignment
GTR	general time reversible
H	hypothesis
IPA	international phonetic alphabet
K	Bayes factor
masc	masculine
MCMC	Monte Carlo Markov chain
MDS	multidimensional scaling
MRI	modified Rand index
MSA	multiple sequence alignments
neut	neuter
NJ	neighbor-joining
NW	northwest
ODE	order dependent evaluation
OT	*Български диалектен атлас, обобщаващ том I-III. Фонетика, акцентология, лексика* – Kochev at al. (2001)
P	purity
par	participle
PHMM	pair hidden Markov models
pl	plural

PMI	pointwise mutual information
RGB	red, green, and blu (color model)
RNA	ribonucleic acid
sg	singular
SW	southwest
TZS	transitional zone at the border with Serbia
UPGMA	unweighted pair group method using arithmetic averages
UPGMC	unweighted pair group method using centroids
WA	weighted pair group method using arithmetic averages
WPGMA	weighted pair group method using arithmetic averages
WPGMC	weighted pair group method using centroids

Bibliography

Alexander, Ronelle. 2004. The vitality, and the revitalizing, of Bulgarian dialectology. In Ronelle Alexander and Vladimir Zhobov, editors, *Revitalizing Bulgarian Dialectology*, volume 1. University of Californian Press.

Alonso, Laura, Irene Castellon, Jordi Escribano, Xavier Messeguer, and Lluis Padro. 2004. Multiple Sequence Alignment for characterizing the linear structure of revision. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*.

Archibald, Jenny K., Mark E. Mort, and Daniel J. Crawford. 2003. Bayesian inference of phylogeny: a non-technical primer. *Taxon*, 52:187–191.

Atkinson, Quentin, Geoff Nicholls, David Welch, and Russell Gray. 2005. From words to dates: water into wine, mathemagic or phylogenetic inference. *Transcriptions of the Philological Society*, 103:193–219.

Barnes, Jonathan. 2006. *Strength and Weakness at the Interface: Positional Neutralization in Phonetics and Phonology*. Walter de Gruyter GmbH, Berlin.

Black, Paul. 1973. Multidimensional scaling applied to linguistic relationships. In *Cahiers de l'Institut de Linguistique Louvain*, volume 3. Expanded version of a paper presented at the Conference on Lexicostatistics. Montreal. University of Montreal.

Bolognesi, Roberto and Wilbert Heeringa. 2002. De invloed van dominante talen op het lexicon en de fonologie van Sardische dialecten. *Gramma/TTT: tijdschrift voor taalwetenschap*, 9(1):45–84.

Boyadzhiev, Todor. 2004. The achievements and tasks of Bulgarian dialectology. In Ronelle Alexander and Vladimir Zhobov, editors, *Revitalizing Bulgarian Dialectology*, volume 1. University of Californian Press.

Bryant, David, Flavia Filimon, and Russell D. Gray. 2005. Untangling our past: Languages, trees, splits and networks. In Ruth Mace, Clare J. Holden, and Stephen Shennan, editors, *The Evolution of Cultural Diversity: Phylogenetic Approaches*. UCL Press, pages 67–84.

Bryant, David and Vincent Moulton. 2004. Neighbornet: an agglomerative algorithm for the construction of plenar phylogenetic networks. *Molecular Biology and Evolution*, 21:255–265.

Campbell, Lyle. 2004. *Historical Linguistics: An Introduction*. Edinburgh University Press, second edition.

Chambers, Jack and Peter Trudgill. 2007. *Dialectology*. 13th edition.

Church, Kenneth Ward and Patrick Hanks. 1989. Word association norms, mutual information, and lexicography. In *Proceedings of the 27th annual meeting on Association for Computational Linguistics*, pages 76–83, Morristown, NJ, USA. Association for Computational Linguistics.

Crystal, David. 1987. *The Cambridge Encyclopedia of Language*. Cambridge University Press, Cambridge.

Drummond, Alexei J. and Andrew Rambaut. 2007. Beast: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7(214).

Dunn, Michael, Angela Terrill, Ger Reesnik, Robert A. Foley, and Stephen C. Levinson. 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science*, (309).

Eska, Joseph F. and Donald Ringe. 2004. Recent work in computational linguistic phylogeny. *Language*, (80).

Everitt, Brian S. 1980. *Cluster Analysis*. Halsted Press, New York.

Fano, Robert Mario. 1961. *Transmission and Information: A Statistical Theory of Communications*. MIT Press.

Felsenstein, Joseph. 2004. *Inferring Phylogenies*. Sinauer Associates, Inc.

Forster, Peter and Alfred Toth. 2003. Toward a phylogenetic chronology of ancient Gaulish, Celtic and Indo-European. In *Proceedings of the National Academy of Sciences*, number 100, pages 9079–9084.

Goebl, Hans. 1982. *Dialektometrie: Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie in Bereich der Dialektgeographie*. Wien: Österreichischen Akademie der Wissenschaften.

Goebl, Hans. 1984. *Dialektometrische Studien: Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF*, volume 3. Tübingen: Max Mayer.

Gooskens, Charlotte and Wilbert Heeringa. 2004. Perceptive evaluation of Levensthein dialect distance measurments using Norwegian dialect data. *Language Variation and Change*, 16:189–207.

Gray, Russel D. and Fiona M. Jordan. 2000. Language trees support the express-train sequence of Austronesian expansion. *Nature*, 405:1052–1055.

Gray, Russell, Alexei Drummond, and Simon J. Greenhill. 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science*, (323).

Gray, Russell D. and Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426:435–438.

Greenhill, Simon J. and Russell D. Gray. 2005. Testing population dispersal hypothesis: Pacific settlement, phylogenetic trees and Austronesian languages. In R. Mace, C.J. Holden, and S. Shennan, editors, *The evolution of cultural diversity: phylogenetic approaches*. UCL Press, pages 31–52.

Greenhill, Simon J. and Russell D. Gray. 2009. Austronesian language phylogenies: myths and misconceptions about Bayesian computational methods. *Austronesian historical linguistics and culture history: a festschrift for Robert Blust*.

Grigorovich, Viktor. 1848. *Очерк путешествия по Европейской Турции.* [A Sketch of a Journey in European Turkey]. Kazan.

Gusfield, Dan. 1997. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge University Press.

Hamed, Mahé Ben. 2005. Neighbour-nets portray the Chinese dialect continuum and the linguistic legacy of China's demic history. *Proceedings of the Royal Society B*, 272(1567):1015–1022.

Hamed, Mahe Ben and Feng Wang. 2006. Stuck in the forest: Trees, networks and Chinese dialects. *Diachronica*, 23:29–60(32).

Hartigan, John A. 1975. *Cluster Algorithms*. John Wiley & Sons, New York.

Hastings, W. Keith. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Heeringa, Wilbert. 2004. *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. Ph.D. thesis, University of Groningen.

Heeringa, Wilbert and Brian D. Joseph. 2007. The relative divergence of Dutch dialect pronunciations from their common source: An exploratory study. In John Nerbonne, T. Mark Ellison, and Grzegorz Kondrak, editors, *Proceedings of the Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*.

Heeringa, Wilbert, John Nerbonne, and Peter Kleiweg. 2002. Validating dialect comparison methods. In Wolfgang Gaul and Gunter Ritter, editors, *Classification, Automation, and New Media. Proceedings of the 24th Annual Conference of the Gesellschaft für Klassifikation e. V., University of Passau, March 15-17, 2000*, pages 445–452. Springer, Berlin, Heidelberg and New York.

Holden, Clare J. 2002. Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum parsimony approach. In *Proceedings of the Royal Society of London, B Sciences*, volume 269, pages 793–799.

Holden, Clare J. and Russell D. Gray. 2006. Rapid radiation, borrowing and dialect continua in the Bantu languages. In P. Foster and C. Renfrew, editors, *Phylogenetic methods and the prehistory of languages*. McDonald Institute for Archeological Research, pages 19–31.

Houtzagers, Peter, John Nerbonne, and Jelena Prokić. 2010. Quantitative and traditional classifications of Bulgarian dialects compared. *Scando Slavica*, 56(2):29–54.

Hubert, Lawrence and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification*, 2:193–218.

Huelsenbeck, John P., Bret Larget, Richard E. Miller, and Fredrik Ronquist. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. *Systematic Biology*, 51(5):673–688.

Huelsenbeck, John P., Fredrik Ronquist, Rasmus Nielsen, and Jonathan P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, 294:2310–2314.

Jain, Anil K. and Richard C. Dubes. 1988. *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, New Yersey.

Jeffreys, Harold. 1961. *The Theory of Probability*. Oxford University Press, 3rd edition.

Johnson, Stephen C. 1967. Hierarchical clustering algorithms. *Psychometrika*, 32(3):241–254.

Joseph, Brian D. 2004. Rescuing traditional (historical) linguistics from grammaticalization 'theory'. In Olga Fischer, Muriel Norde, and Harry Perridon, editors, *Up and Down the Cline - The Nature of Grammaticalization*. Amsterdam: John Benjamins Publishing Co.

Kass, Robert E. and Adrian E. Raftery. 1995. Bayes factors. *Journal of the American Statistical Association*, 90:773–795.

Kessler, Brett. 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.

Kochev, Ivan, Donka Vakarelska-Chobanska, Tinka Kostova, Elena Kiaeva, and Margarita Tetovska-Troeva. 2001. *Български диалектен атлас, обобщаващ том I-III. Фонетика, акцентология, лексика.* [Atlas of Bulgarian Dialects: Phonetics. Intonation. Lexicology, Vol. I - III]. Sofia: Trud.

Kondrak, Grzegorz. 2002. *Algorithms for Language Reconstruction*. PhD Thesis, University of Toronto.

Kroeber, Alfred L. and Charles D. Chrétien. 1939. The statistical technique and Hittite. *Language*, 15(2).

Legendre, Pierre and Louis Legendre. 1998. *Numerical Ecology*. Elsevier, Amsterdam, second edition.

Leinonen, Therese. 2010. *An Acoustic Analysis of Vowel Pronunciation in Swedish Dialects*. Ph.D. thesis, University of Groningen.

Levenshtein, Vladimir. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:707–710.

Li, Shuying. 1996. *Phylogenetic Tree Construction Using Markov Chain Monte Carlo*. Ph.D. thesis, Ohio State University, Columbus.

MacKay, David J.C. 2003. *Information Theory, Inference and Learning Algorithms*. CUP.

Manning, Chris and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press. Cambridge, MA.

Mau, Bob. 1996. *Bayesian Phylogenetic Inference via Markov Chain Monte Carlo Methods*. Ph.D. thesis, University of Wisconsin, Madison.

McMahon, April, Paul Heggarty, Robert McMahon, and Warren Maguire. 2007. The sound patterns of Englishes: representing phonetic similarity. *English Language and Linguistics*, 11.1:113–142.

Metropolis, Nicholas, Arianna W. Rosenbluth, Marshall N. Rosenbluth, and Augusta H. Teller. 1953. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Nakhleh, Luay, Don Ringe, and Tandy Warnow. 2005. Perfect phylogenetic networks: a new methodology for reconstructing the evolutionary history of natural languages. *Language*, 81:382–420.

Nakhleh, Luay, Tandy Warnow, Don Ringe, and Steven N. Evans. 2005. A comparison of phylogenetic reconstruction methods on an Indo-European dataset. *Transactions of the Phylological Society*, 103(2):171–192.

Nerbonne, John. 2005. Various variation aggregates in the LAMSAS South. In Catherine Davis and Michael Picone, editors, *Language Variety in the South III*. University of Alabama Press, Tuscaloosa.

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

Nerbonne, John, Wilbert Heeringa, Eric van den Hout, Peter van de Kooi, Simone Otten, and Willem van de Vis. 1996. Phonetic distance between Dutch dialects. In G. Durieux, W. Daelemans, and S. Gills, editors, *CLIN VI, Papers from the sixth CLIN meeting*. University of Antwerpen, Antwerpen, pages 185–202.

Nerbonne, John, Peter Kleiweg, Wilbert Heeringa, and Franz Manni. 2008. Projecting dialect differences to geography: Bootstrap clustering vs. noisy clustering. In Christine Preisach, Lars Schmidt-Thieme, Hans Burkhardt, and Reinhold Decker, editors, *Data Analysis, Machine Learning, and Applications. Proceedings of the 31st Annual Meeting of the German Classification Society*, pages 647–654, Berlin.

Nerbonne, John and Christine Siedle. 2005. Dialektklassifikation auf der Grundlage Aggregierter Ausspracheunterschiede. *Zeitschrift für Dialektologie und Linguistic*, 72(2):129–147.

Osenova, Petya, Wilbert Heeringa, and John Nerbonne. 2009. A quantitative analysis of Bulgarian dialect pronunciation. *Zeitschrift für Slavische Philologie*, 66(2):425–458.

Page, Roderic D. M. and Edward C. Holmes. 2006. *Molecular Evolution: A Phylogenetic Approach*. Blackwell, Oxford.

Pagel, Mark, Quentin D. Atkinson, and Andrew Meade. 2007. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature*, 449:717–720.

Penny, David. 1982. Towards a basis for classification: the incompletness of distance measures, incompatibility analysis and phenetic classification. *Journal of Theoretical Biology*, 96(2):129–142.

Prokić, Jelena. 2007. Identifying linguistic structure in a quantitative analysis of dialect pronunciation. In *Proceedings of the ACL 2007 Student Research Workshop*, pages 61–66, Prague, Czech Republic, June. Association for Computational Linguistics.

Prokić, Jelena and Tim Van de Cruys. 2010. Exploring dialect phonetic variation using PARAFAC. In *Proceedings of the 11th Meeting of ACL Special Interest Group in Computational Morphology and Phonology (SIGMORPHON)*, Uppsala, July. Association for Computational Linguistics.

Prokić, Jelena and John Nerbonne. 2008. Recognizing groups among dialects. *International Journal of Humanities and Arts Computing. Special Issue on Language Variation edited by John Nerbonne, Charlotte Gooskens, Sebastian Kürschner, and Renée van Bezooijen*, 2:153–172.

Prokić, Jelena, John Nerbonne, Vladimir Zhobov, Petya Osenova, Krili Simov, Thomas Zastrow, and Erhard Hinrichs. 2009. The computational analysis of Bulgarian dialect pronunciation. *Serdica Journal of Computing*, 3:269–298.

Prokić, Jelena, Martijn Wieling, and John Nerbonne. 2009. Multiple string alignments in linguistics. In Lars Borin and Piroska Landvai, editors, *Proceedings of Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH - SHELT&R 2009) EACL Workshop*.

Rand, William M. 1971. Objective criteria for the evaluation of clustering methods. *Journal of American Statistical Association*, 66(336):846–850, December.

Rannala, Bruce and Ziheng Yang. 1996. Probability distribution of molecular evolutionary trees. *Journal of Molecular Evolution*, (43):304–311.

Ringe, Donald, Tandy Warnow, and Ann Taylor. 2002. Indo-European and computational cladistics. *Transactions of the Philological Society*.

Saitou, Naruya and Masatoshi Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4:406–425.

Schleicher, August. 1853. Die ersten Spaltungen des indogermanischen Urvolkes. *Allgemeine Zeitung für Wissenschaft und Literatur*.

Schmidt, Johannes. 1872. *Die Verwandtschaftsverhältnisse der indogermanischen Sprachen*. Weimar:Böhlau.

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

Sokal, Robert R. and F. James Rohlf. 1962. The comparison of dendrograms by objective methods. *Taxon*, 11:33–40.

Stoykov, Stoyko. 1954. *Кратък осведомителен въпросник за проучване на Българските говори* [A Short Guide for Studying Bulgarian Dialects]. Sofia.

Stoykov, Stoyko. 1966. *Български диалектен атлас. Североизточна България.* [Atlas of Bulgarian Dialects. Northeastern Bulgaria]. Publishing House of Bulgarian Academy of Science, volume II, Sofia, Bulgaria.

Stoykov, Stoyko. 2002. *Българска диалектология.* [Bulgarian Dialectology]. Sofia, 4th ed.

Stoykov, Stoyko and Samuil B. Bernstein. 1964. *Български диалектен атлас. Югоизточна България.* [Atlas of Bulgarian Dialects. Southeastern Bulgaria]. Publishing House of Bulgarian Academy of Science, volume I, Sofia, Bulgaria.

Stoykov, Stoyko, Ivan Kochev, and Maksim Mladenov. 1981. *Български диалектен атлас. Северозападна България.* [Atlas of Bulgarian Dialects. Northwestern Bulgaria]. Publishing House of Bulgarian Academy of Science, volume IV, Sofia, Bulgaria.

Stoykov, Stoyko, Kiril Mirchev, Ivan Kochev, and Maksim Mladenov. 1974. *Български диалектен атлас. Югозападна България.* [Atlas of Bulgarian Dialects. Southwestern Bulgaria]. Publishing House of Bulgarian Academy of Science, volume III, Sofia, Bulgaria.

Studier, James A. and Karl J. Kepler. 1988. A note on the neighbor-joining algorithm of Saitou and Nei. *Molecular Biology and Evolution*, 5:729–731.

Sussex, Roland and Paul Cubberley. 2006. *The Slavic Languages*. CUP.

Warnow, Tandy. 1997. Mathematical approaches to comparative linguistics. In *Proceedings of the National Academy of Science of the USA*, volume 94, pages 6585–6590.

Warnow, Tandy, Steven N. Evans, Donald Ringe, and Luay Nakhleh. 2006. A stochastic model of language evolution that incorporates homoplasy and borrowing. In Peter Forster and Colin Renfrew, editors, *Phylogenetic Methods and the Prehistory of Languages*. MacDonald Institute for Archaeological Research, Cambridge.

Wichmann, Søren and Arpiar Saunders. 2007. How to use typological databases in historical linguistic research. *Diachronica*, 4(2):373–404.

Wieling, Martijn, Therese Leinonen, and John Nerbonne. 2007. Inducing sound segment differences using Pair Hidden Markov Models. In John Nerbonne, Mark Ellison, and Greg Kondrak, editors, *Computing and Historical Phonology: 9th Meeting of ACL Special Interest Group for Computational Morphology and Phonology*, pages 48–56.

Wieling, Martijn, Jelena Prokić, and John Nerbonne. 2009. Evaluating the pairwise string alignments of pronunciations. In Lars Borin and Prioska Lendvai, editors,

Proceedings of Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH - SHELT&R 2009) EACL Workshop, pages 26–34.

Wood, Sidney A. J. and Thore Pettersson. 1988. Vowel reduction in Bulgarian: the phonetic data and model experiments. *Folia Linguistica*, 22(3-4):239–262.

Yang, Ziheng. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution*, 39:306–314.

Zhao, Ying and George Karypis. 2001. Criterion functions for document clustering: Experiments and analysis. TR 01-40, Department of Computer Science, University of Minnesita, Minneapolis, MN.

Short summary

Dialectometry is a multidisciplinary field that uses quantitative methods in the analysis of dialect data. From the very beginning, most of the research in dialectometry has been focused on including large amounts of data in analyses and offering alternative views to researchers. Later it was used for the identification of dialect groups and development of methods that would tell us how similar (or different) one variety is when compared to the neighboring varieties. In this book we present advances in several techniques that allow the researcher to automatically measure the differences between language varieties. We test all methods on Bulgarian dialect pronunciation data.

Part of the research presented relies on the Levenshtein algorithm to aggregate over the numerous features found in the data and infer the similarities/distances among the groups of dialects. We investigate the application of clustering techniques in the detection of dialect groups, and propose several evaluation techniques that can be used to estimate the quality of the automatically obtained groups. In order to automatically infer the distances between the phones in the data set we combine the Levenshtein algorithm with the technique called pointwise mutual information. Information on the distances between the phones helps us get better estimates on the distances between the strings, and consequently on the distances between language varieties.

In this thesis we also test an alternative approach to dialect variation that is more historically motivated. We employ a method taken from phylogenetics, namely Bayesian inference of phylogeny, which focuses on systematic shared innovations as a signal of common ancestry, and reexamine the relatedness among the Bulgarian dialect varieties. This method is applied to the automatically multiply aligned strings, which we produce and evaluate using two novel methods.

The results of applying different quantitative techniques to the Bulgarian dialect data show that some of the traditional divisions of this area have to be questioned if only pronunciation data is taken into account. The comparison of the divisions resulting from the geographic and historical approaches has shown that these two different perspectives gave very similar picture of the Bulgarian dialect variation. None of the methods developed are language specific, nor are they applicable only to the dialect data.

Samenvatting

Dialectometrie is een multidisciplinair onderzoeksgebied dat kwantitatieve methoden inzet voor de analyse van dialectgegevens. Aanvankelijk was onderzoek binnen dialectometrie vooral gericht op het gebruik van grote hoeveelheden gegevens voor analyses en het bieden van nieuwe inzichten voor onderzoekers. Later werd dialectometrie ingezet voor de identificatie van dialectgroepen en de ontwikkeling van methoden die blootleggen hoe gelijk (of ongelijk) één variëteit is ten opzichte van naburige taalvariëteiten. In dit boek wordt de vooruitgang van verschillende technieken beschreven die de onderzoeker in staat stelt om geautomatiseerd verschillen te meten tussen taalvariëteiten. Alle methoden worden getest op Bulgaarse dialect uitspraakgegevens.

Een deel van het onderzoek hier beschreven is gebaseerd op het Levenshtein algoritme, dat wordt gebruikt voor het aggregeren van de vele kenmerken van de dialectgegevens om daarmee de overeenkomsten/afstanden tussen de dialectgroepen af te leiden. We onderzoeken de toepassing van clustertechnieken voor het determineren van dialectgroepen en dragen verschillende evaluatietechnieken aan die gebruikt kunnen worden voor het schatten van de kwaliteit van de geautomatiseerd verkregen groepen. Voor het geautomatiseerd afleiden van de afstanden tussen de fonemen in de gegevensverzameling, combineren we het Levenshtein algoritme met een techniek uit de informatietheorie, pointwise mutual information. We gebruiken de (empirische) frequentie van foneemcorrespondenties in aligneringen om de afstanden tussen fonemen beter in te schatten. Informatie over de afstanden tussen de fonemen helpt ons om betere schattingen te maken van de afstanden tussen de karakterreeksen en daaropvolgend de afstanden tussen taalvariëteiten.

In dit proefschrift wordt ook een alternatieve benadering van dialect variatie getest, een benadering die vooral historische affiniteiten tracht op te zoeken. We passen een methode toe die gebruikt wordt binnen de phylogenetica, namelijk Bayesiaanse inferentie van phylogenetica, die systematisch op gemeenschappelijke innovaties als teken van een gedeelde afkomst focust, en beoordelen opnieuw de gerelateerdheid tussen de Bulgaarse dialectvariëteiten. Deze methode wordt toegepast op de meervoudig opgelijnde ('aligned') karakterreeksen, die geautomatiseerd werden verkregen.

De resultaten van het toepassen van verschillende kwantitatieve methoden op de Bul-

gaarse dialectgegevens, laten zien dat er bij sommige traditionele indelingen van dit gebied vraagtekens gezet kunnen worden in het bijzonder als we slechts uitspraakgegevens in beschouwing nemen. De vergelijking van de indelingen voortkomend uit de geografische - en historische benadering, laat zien dat deze twee verschillende perspectieven eenzelfde beeld laten zien van de Bulgaarse dialectvariatie. Geen van de ontwikkelde methoden is taalspecifiek, noch slechts toepasbaar op dialectgegevens.

Appendix A

List of words

Table A.1: List of all words from *Buldialect* phonetic data set. The numbers in the right column refer to the features described in Section 2.2 which are present in a given word.

word	features
аз /az/ ‘I’	2; 3
агне /ˈagne/ ‘lamb’	2; 3; 7; 23
бели /ˈbeli/ ‘white - pl.’	1; 39
берат /beˈrɤt/ ‘pick up - 3rd pl’	6; 39
беше /ˈbeʃe/ ‘be - past 2nd sg, 3rd sg’	1
бране /braˈne/ ‘pick - verb. noun’	23; 39
брашно /braʃˈno/ ‘flour’	39
бързо /ˈbɤrzo/ ‘quickly’	18; 34
бяхме /ˈbʲaxme/ ‘be - past 1st pl’	1; 27; 36
вежда /ˈveʒda/ ‘eyebrow’	1; 19; 39
вече /ˈvet͡ʃe/ ‘already’	34
вечер /ˈvet͡ʃer/ ‘evening’	11
видях /viˈdʲax/ ‘see - aorist 1st sg.’	1; 27; 39
вие /ˈvie/ ‘you’	38
вино /ˈvino/ ‘wine’	39
влизам /ˈvlizam/ ‘enter - 1st sg’	32; 35
вода /voˈda/ ‘water’	6; 39
вол /vol/ ‘ox’	29
време /ˈvreme/ ‘time’	1; 7
връх /vrɤx/ ‘peak’	18; 27
връщам /ˈvrɤʃtam/ ‘give back - 1st sg’	18; 19; 35

word	features
вчера /ˈvt͡ʃera/ ‘yesterday’	34
във /vɤv/ ‘in’	8; 32
вълк /vɤlk/ ‘woolf’	18; 23
вълна /ˈvɤlna/ ‘wool’	18
вънка /ˈvɤnka/ ‘outside’	8
вътре /ˈvɤtre/ ‘inside’	1; 6
вятър /ˈvʲatɤr/ ‘wind’	1; 10
глава /glaˈva/ ‘head’	6; 39
гладен /ˈgladen/ ‘hungry’	9
говедо /goˈvedo/ ‘beef’	7
горе /ˈgore/ ‘up’	1
гости /ˈgosti/ ‘guest - pl’	24
градът /graˈdɤt/ ‘the town’	8; 39
грозде /ˈgrozde/ ‘grapes’	24
дадоха /ˈdadoxa/ ‘to give - aor 3rd pl’	27
две /dve/ ‘two’	1
двор /dvor/ ‘yard’	29
ден /den/ ‘day’	9; 23
дера /deˈrɤ/ ‘flay - 1st sg’	6; 39
десет /ˈdeset/ ‘ten’	7; 23
дете /deˈte/ ‘child’	1; 7; 39
джоб /d͡ʒob/ ‘pocket’	15; 22; 31
днес /dnes/ ‘today’	9
добре /doˈbre/ ‘well’	1
долу /ˈdolu/ ‘down’	17
дошъл /doˈʃɤl/ ‘come - aor part’	9
дъжд /dɤʒd/ ‘rain’	8; 31
дълбок /dɤlˈbok/ ‘deep’	18
дъно /ˈdɤno/ ‘bottom’	8
дърво /dɤrˈvo/ ‘tree’	18; 29; 39
един /eˈdin/ ‘one - masc’	3; 9
едно /edˈno/ ‘one - neut’	3; 33
език /eˈzik/ ‘tongue’	3; 7; 12; 23
ечемик /et͡ʃeˈmik/ ‘barley’	3; 7; 13; 16; 23; 37
желязо /ʒeˈlʲazo/ ‘iron’	1; 13; 22; 39
жена /ʒeˈna/ ‘woman’	6; 39
жив /ʒiv/ ‘alive’	13; 31
живели /ʒiˈveli/ ‘live - past pl’	1; 13
жълт /ʒɤlt/ ‘yellow’	18
жътва /ˈʒɤtva/ ‘harvest’	7
звезда /zvezˈda/ ‘star’	1; 6; 22; 39
здрав /zdrav/ ‘healthy’	25; 31
земя /zeˈmʲa/ ‘land’	6; 21; 39

word	features
зет /zet/ ‘brother-in-low’	7; 23
и /i/ ‘she - dative’	38
им /im/ ‘they - dative’	38
име /ˈime/ ‘name’	7; 13
камък /ˈkamɤk/ ‘stone’	37
ключ /klʲut͡ʃ/ ‘key’	14
кое /koˈe/ ‘which’	4
кон /kon/ ‘horse’	23
кръв /krɤv/ ‘blood’	18; 31
къде /kɤˈde/ ‘where’	1; 6
лесно /ˈlesno/ ‘easily’	9
леща /ˈleʃta/ ‘lentil - pl	7; 19
майка /ˈmajka/ ‘mother’	5; 23
месец /ˈmeset͡s/ ’month’	1; 7
месо /meˈso/ ’meat’	7; 39
млякото /ˈmlʲakoto/ ‘the milk’	1; 16; 39
много /ˈmnogo/ ‘much, many’	33
мъж /mɤʒ/ ‘man’	6; 31
мъже /mɤˈʒe/ ‘men’	6; 39
мъжът /mɤˈʒɤt/ ‘the man’	6; 8; 39
наше /ˈnaʃe/ ‘our - neut’	15
неделя /neˈdelʲa/ ‘Sunday’	1; 16; 39
неще /neˈʃte/ ‘not want - 3rd sg’	19; 39
нещо /ˈneʃto/ ‘something’	1
нея /ˈneja/ ‘she - accusative’	2; 5
ние /ˈnie/ ‘we’	38
носят /ˈnosʲɤt/ ‘carry - 3rd pl’	6; 23
нощ /noʃt/ ‘night’	19
няма /ˈnʲama/ ‘there is no’	1
овца /ovˈt͡sa/ ‘sheep’	16; 31; 39
овце /ovˈt͡se/ ‘sheep - pl’	16; 31; 39
овчар /ovˈt͡ʃar/ ‘shepherd’	2; 31
овчари /ovˈt͡ʃari/ ‘shepherd - pl’	2; 31
огън /ˈogɤn/ ‘fire’	10; 23; 30; 39
онези /oˈnezi/ ‘those’	1; 38
орех /ˈorex/ ‘walnut’	1; 27; 30
пека /peˈkɤ/ ‘bake - 1st sg’	6; 17; 39
пепел /ˈpepel/ ‘ash’	1; 15; 23
петел /peˈtel/ ‘rooster’	1; 9
петък /ˈpetɤk/ ‘Friday’	1; 7; 8
плащам /ˈplaʃtam/ ‘pay - 1st sg’	19; 35
понеделник /poneˈdelnik/ ‘Monday’	1; 11; 16; 23

word	features
пръч /prɤt͡ʃ/ ‘he-goat’	18
пръвият /ˈprɤvijɤt/ ‘the first’	8; 18
път /pɤt/ ‘road’	6; 23
пясък /ˈpʲasɤk/ ‘sand’	1; 8
река /reˈka/ ‘river’	1; 6; 39
ръка /rɤˈka/ ‘hand’	6; 39
ръце /rɤˈt͡se/ ‘hands’	1; 6; 39
се /se/ ’one’s self’	7
сега /seˈga/ ‘now’	9
седя /seˈdʲɤ/ ‘sit - 1st sg’	6; 23; 39
сестра /sesˈtra/ ‘sister’	6; 25; 39
сирене /ˈsirene/ ‘cheese’	12; 23
сол /sol/ ‘salt’	23
средата /sreˈdata/ ‘the middle’	1
сряда /ˈsrʲada/ ‘Wednesday’	1; 26
старец /ˈstare͡ts/ ‘old man’	9
страх /strax/ ‘fear’	25; 27
сух /sux/ ‘dry’	27
събота /ˈsɤbota/ ‘Saturday’	6; 16
сърп /sɤrp/ ‘sickle’	18
със /sɤs/ ‘with’	8
такъв /taˈkɤv/ ‘such’	8; 31
твой /tvoj/ ‘yours’	29
това /toˈva/ ‘this - neut’	16; 38
тогава /toˈgava/ ‘then’	38
тъмно /ˈtɤmno/ ‘dark -neut’	9; 33
тънко /ˈtɤnko/ ‘thin - neut’	9
трева /treˈva/ ‘grass’	1; 6; 39
утре /ˈutre/ ‘tomorrow’	1; 3
ухо /uˈxo/ ‘ear’	29; 39
фурна /ˈfurna/ ‘oven’	23; 28
хляб /xlʲab/ ‘bread’	1; 27; 31
хоро /xoˈro/ ‘chain dance’	29; 39
хубав /ˈxubav/ ‘beautiful - masc’	27; 31
хубаво /ˈxubavo/ ‘beautiful - neut’	27; 29
цял /t͡sʲal/ ‘whole’	1
чакат /ˈt͡ʃakat/ ‘wait - 3rd pl’	2; 6
червен /t͡ʃerˈven/ ‘red’	18; 20
черен /ˈt͡ʃeren/ ‘black’	18; 20
череша /t͡ʃeˈreʃa/ ‘cherry’	1; 13; 20
чета /t͡ʃeˈtɤ/ ‘read - 1st sg’	6; 39
чешма /t͡ʃeʃˈma/ ‘fountain’	6; 13; 39

word	features
човек /t͡ʃoˈvek/ ‘human’	1; 34
ще /ʃte/ ‘will’	19
я /ja/ ‘she - accusative’	38
ябълка /ˈjabɤlka/ ‘apple’	2; 5; 18; 23
ябълки /ˈjabɤlki/ ‘apple - pl’	2; 5; 18; 23
яйца /jajˈt͡sa/ ‘egg - pl’	2; 5; 39
яйце /jajˈt͡se/ ‘egg’	2; 5; 39
ям /jam/ ‘eat - 1st sg’	2; 5
ядеш /jaˈdeʃ/ ‘eat - 2nd sg’	2; 5

Appendix B

List of phones

ɑ	ˈɑ	ɑː	ˈɑː	ˈɒ	a	ˈa	ˈaː	e	ˈe
eː	ˈeː	ɛ	ˈɛ	ˈɛː	ɪ	ˈɪ	ˈɪː	i	ˈi
ˈiː	ɨ	ˈɨ	y	ˈy	o	ˈo	oː	ˈoː	ˈɔ
ʊ	ˈʊ	ˈʌ	u	ˈu	uː	ˈuː	ɤ	ˈɤ	ɤː
ˈɤː	ə	ˈə	b	bʲ	c	cʲ	ç	d	dʲ
d͡z	d͡ʑ	d͡ʒ	d͡ʒʲ	f	fʲ	g	gʲ	h	j
ɟ	k	kʲ	l	lʲ	l̩	ˈl̩	m	mʲ	ɯ̟
n	nʲ	ɳ	p	pʲ	ɸ	r	rʲ	r̩	ˈr̩
s	sʲ	ʃ	ɕ	t	tʲ	t͡ɕ	t͡s	t͡sʲ	t͡ʃ
v	vʲ	w	x	z	zʲ	ʑ	ʒ		

Appendix C

List of sites

Алдомировци (Aldomirovtsi)	Аспарухово, Лом (Asparuhovo, Lom)
Аспарухово, Пров (Asparuhovo, Prov)	Бабяк (Babyak)
Багренци (Bagrentsi)	Банище (Banishte)
Банско (Bansko)	Бачково (Bachkovo)
Беглеж (Beglezh)	Белене (Belene)
Белица (Belitsa)	Бистрица (Bistritsa)
Бов (Bov)	Богданов дол (Bogdanov dol)
Борисово (Borisovo)	Бръшлян (Brashlyan)
Бучин проход (Buchin prohod)	Българи (Balgari)
Варвара (Varvara)	Вардун (Vardun)
Васильово (Vasilyovo)	Велковци (Velkovtsi)
Винарово (Vinarovo)	Винище (Vinishte)
Владиня (Vladinya)	Воден (Voden)
Войнягово (Voynyagovo)	Враниловци (Vranilovtsi)
Врачеш (Vrachesh)	Вресово (Vresovo)
Въбел (Vabel)	Въклиново (Vaklinovo)
Вълче поле (Valche pole)	Върбица (Varbitsa)
Върбово (Varbovo)	Габаре (Gabare)
Габра (Gabra)	Галата (Galata)
Ганчовец (Ganchovets)	Гарван (Garvan)
Гега (Gega)	Гложене (Glozhene)
Голема Раковица (Golema Rakovitsa)	Големо Малово (Golemo Malovo)
Голица (Golitsa)	Голяма Желязна (Golyama Zhelyazna)
Голямо Шивачево (Golyamo Shivachevo)	Горна Росица (Gorna Rositsa)
Горни Върпища (Gorni Varpishta)	Говедарци (Govedartsi)

Градец (Gradets)	Гърмен (Garmen)
Девенци (Deventsi)	Девесилица (Devesilitsa)
Дерманци (Dermantsi)	Дива Слатина (Diva Slatina)
Дивдядово (Divdyadovo)	Динево (Dinevo)
Дичин (Dichin)	Доброселец (Dobroselets)
Доброславци (Dobroslavtsi)	Добротино (Dobrotino)
Добърско (Dobarsko)	Долна Бешовица (Dolna Beshovitsa)
Долна Диканя (Dolna Dikanya)	Долна Мелна (Dolna Melna)
Долна Рикса (Dolna Riksa)	Долна Студена (Dolna Studena)
Долни Богров (Dolni Bogrov)	Долно Левски (Dolno Levski)
Дорково (Dorkovo)	Драбишна (Drabishna)
Драгижево (Dragizhevo)	Драгоданово (Dragodanovo)
Драгоево (Dragoevo)	Драгойчинци (Dragoychintsi)
Езерово (Ezerovo)	Елов дол (Elov dol)
Енина (Enina)	Жалтуша (Zhaltusha)
Жеглица (Zheglitsa)	Желен (Zhelen)
Желязково (Zhelyazkovo)	Жеравна (Zheravna)
Заберново (Zabernovo)	Забърдо (Zabardo)
Замфирово (Zamfirovo)	Заножене (Zanozhene)
Здравковец (Zdravkovets)	Зеленига̀д (Zelenigrad)
Ивански (Ivanski)	Изворово (Izvorovo)
Индже войвода (Indzhe voyvoda)	Калипетрово (Kalipetrovo)
Калояново (Kaloyanovo)	Караисен (Karaisen)
Караново (Karanovo)	Каспичан (Kaspichan)
Ковачевци (Kovachevtsi)	Козичино (Kozichino)
Колю Мариново (Kolyu Marinovo)	Конска (Konska)
Копиловци (Kopilovtsi)	Копривщица (Koprivshtitsa)
Кортен (Korten)	Костенец (Kostenets)
Кравеник (Kravenik)	Крамолин (Kramolin)
Крета (Kreta)	Кривня (Krivnya)
Левуново (Levunovo)	Лиляче (Lilyache)
Липница (Lipnitsa)	Лобош (Lobosh)
Лозен (Lozen)	Любенова махала (Lyubenova mahala)
Маломирово (Malomirovo)	Марикостиново (Marikostinovo)
Марково (Markovo)	Марчаево (Marchaevo)
Мерданя (Merdanya)	Меричлери (Merichleri)
Милчина Лъка (Milchina Laka)	Михалци (Mihaltsi)
Момина баня (Momina banya)	Момина клисура (Momina klisura)
Момково (Momkovo)	Момчиловци (Momchilovtsi)
Мугла (Mugla)	Николово (Липник) (Nikolovo (Lipnik))

Николово (Nikolovo)	Нова Ловча (Nova Lovcha)
Нова Надежда (Nova Nadezhda)	Ново село (Novo selo)
Ноевци (Noevtsi)	Огнен (Ognen)
Омарчево (Omarchevo)	Опан (Opan)
Осенец (Osenets)	Павелско (Pavelsko)
Панагюрище (Panagyurishte)	Паскалевец (Paskalevets)
Певец (Pevets)	Пелатиково (Pelatikovo)
Петърница (Petarnitsa)	Плаково (Plakovo)
Подвис (Podvis)	Пожарево (Pozharevo)
Рабиша (Rabisha)	Радовене (Radovene)
Разбоище (Razboishte)	Ракево (Rakevo)
Раковица (Rakovitsa)	Рани луг (Rani lug)
Ружинци (Ruzhintsi)	Садина (Sadina)
Сапарево (Saparevo)	Светлина (Svetlina)
Свирково (Svirkovo)	Секирово (Sekirovo)
Сенокос (Senokos)	Сестрино (Sestrino)
Скобелево (Skobelevo)	Славейно (Slaveyno)
Славяново (Slavyanovo)	Смолско (Smolsko)
Смочево (Smochevo)	Солища (Solishta)
Средец (Sredets)	Стакевци (Stakevtsi)
Стамболово (Stambolovo)	Стоилово (Stoilovo)
Стралджа (Straldzha)	Строево (Stroevo)
Стърмен (Starmen)	Сухиндол (Suhindol)
Сушица (Sushitsa)	Тихомир (Tihomir)
Тихомирово (Tihomirovo)	Тополчане (Topolchane)
Трънчовица (Tranchovitsa)	Тръстеник (Trastenik)
Устово (Ustovo)	Фурен (Furen)
Хвойна (Hvoyna)	Хухла (Huhla)
Цапарево (Tsaparevo)	Церовица (Tserovitsa)
Чепеларе (Chepelare)	Черногорово (Chernogorovo)
Черноморец (Chernomorets)	Чуковец (Chukovets)
Шипка (Shipka)	Широка лъка (Shiroka laka)
Широки дол (Shiroki dol)	Щипско (Shtipsko)
Яворово (Yavorovo)	

Appendix D

Clustering results

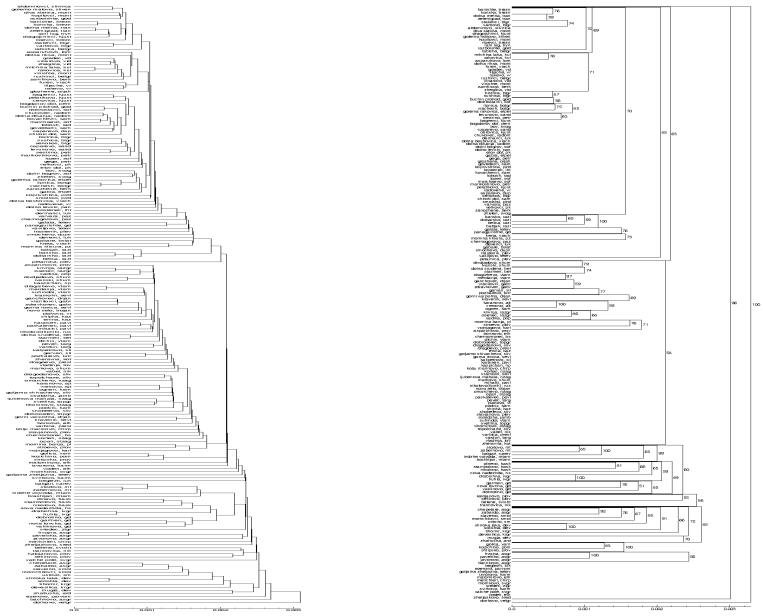

Figure D.1: Single link dendrograms, plain and with the noise.

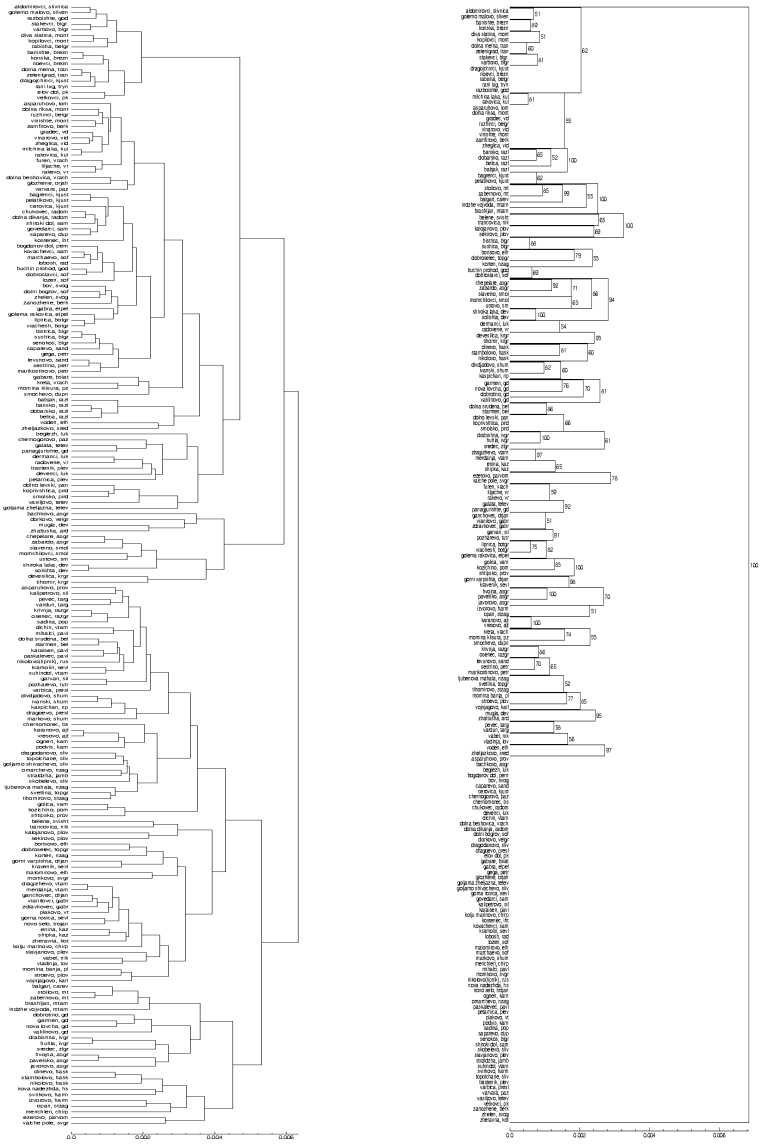

Figure D.2: Complete link dendrograms, plain and with the noise.

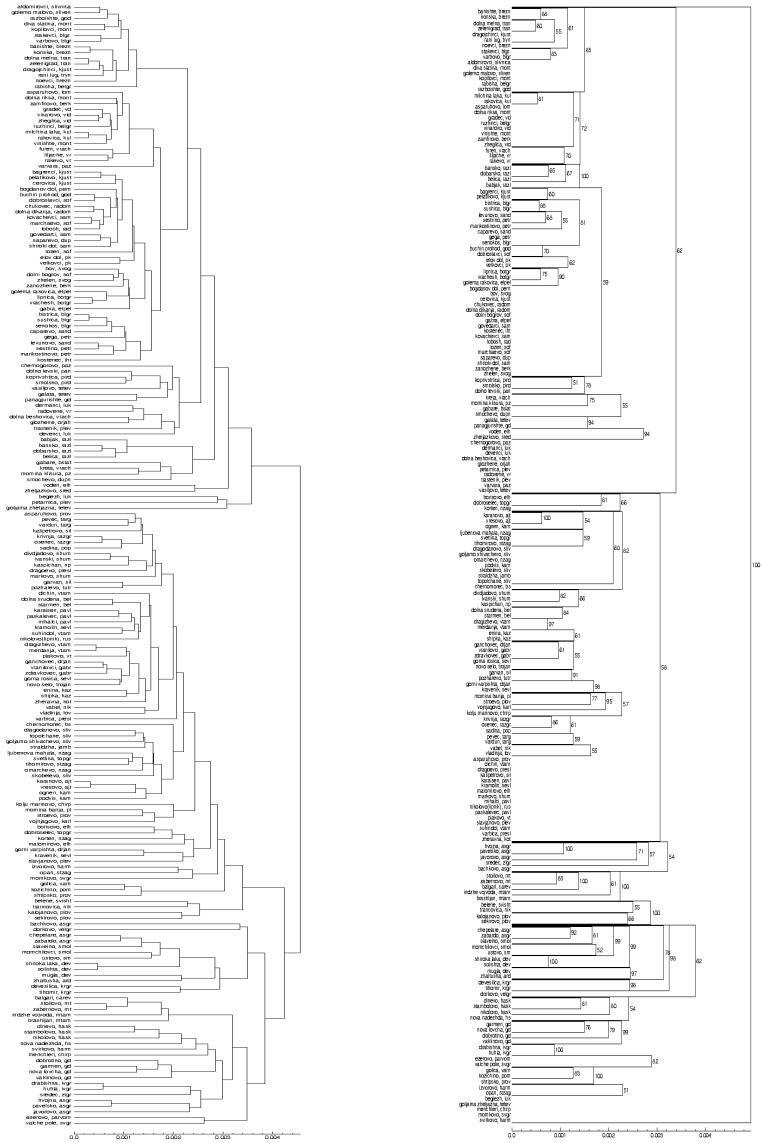

Figure D.3: UPGMA dendrograms, plain and with the noise.

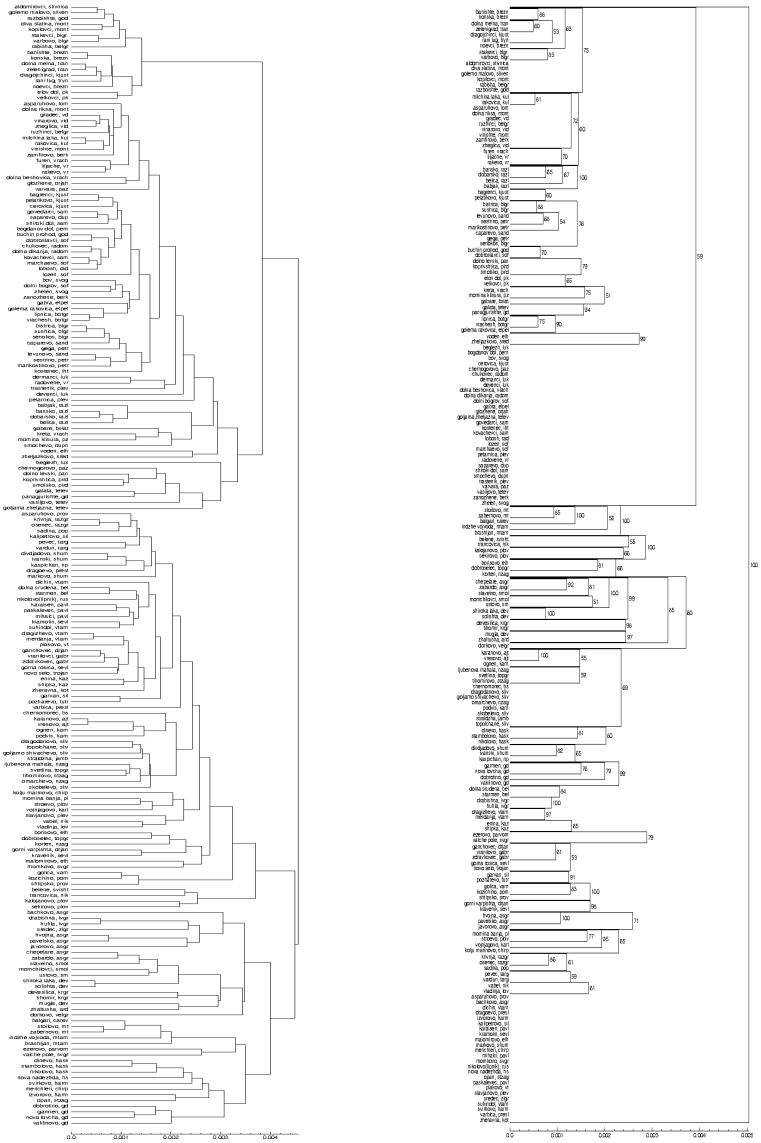

Figure D.4: WPGMA dendrograms, plain and with the noise.

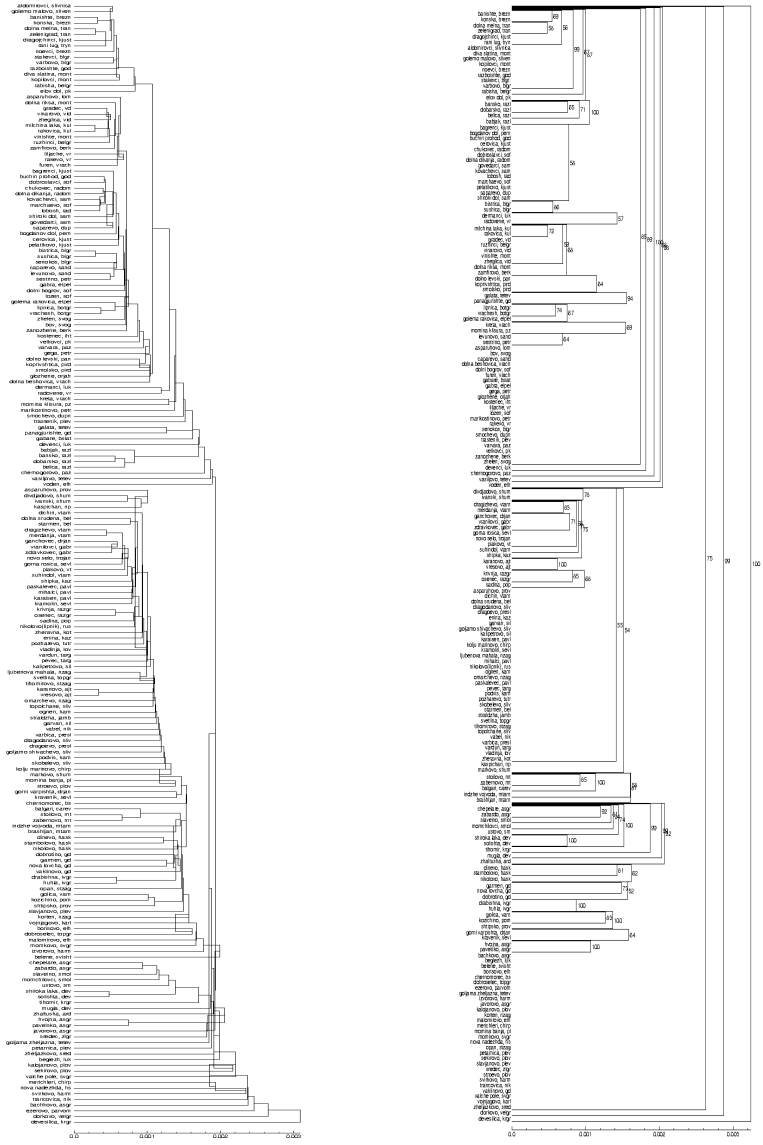

Figure D.5: UPGMC dendrograms, plain and with the noise.

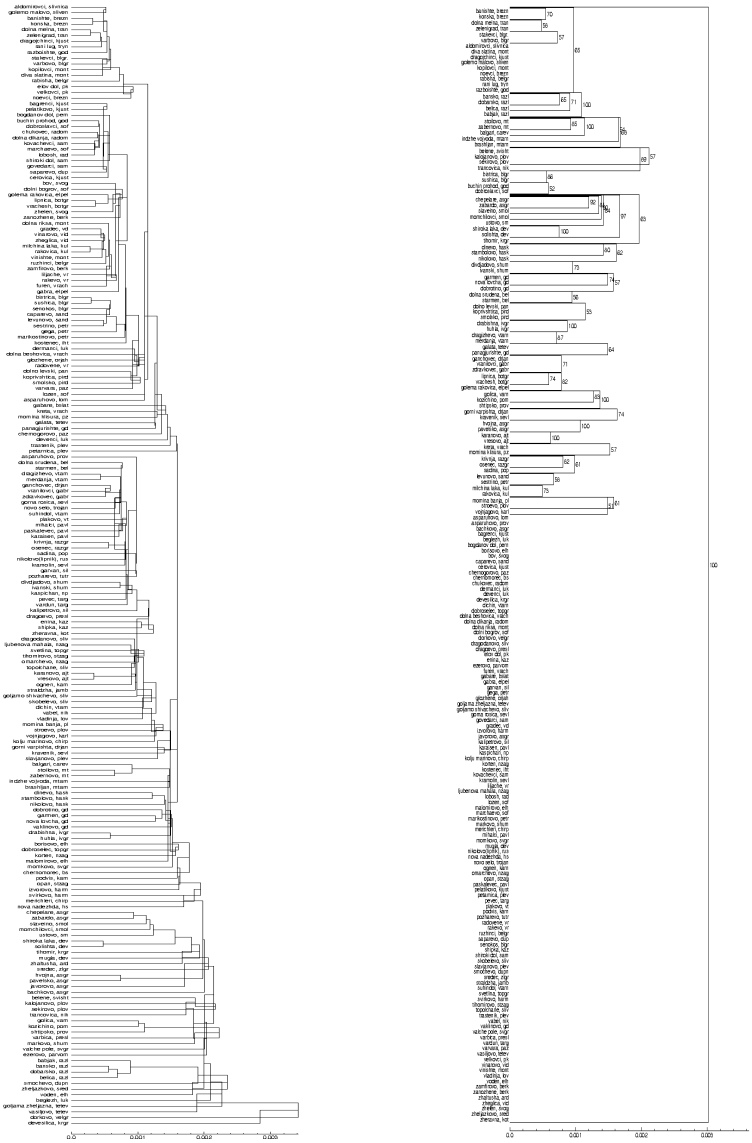

Figure D.6: WPGMC dendrograms, plain and with the noise.

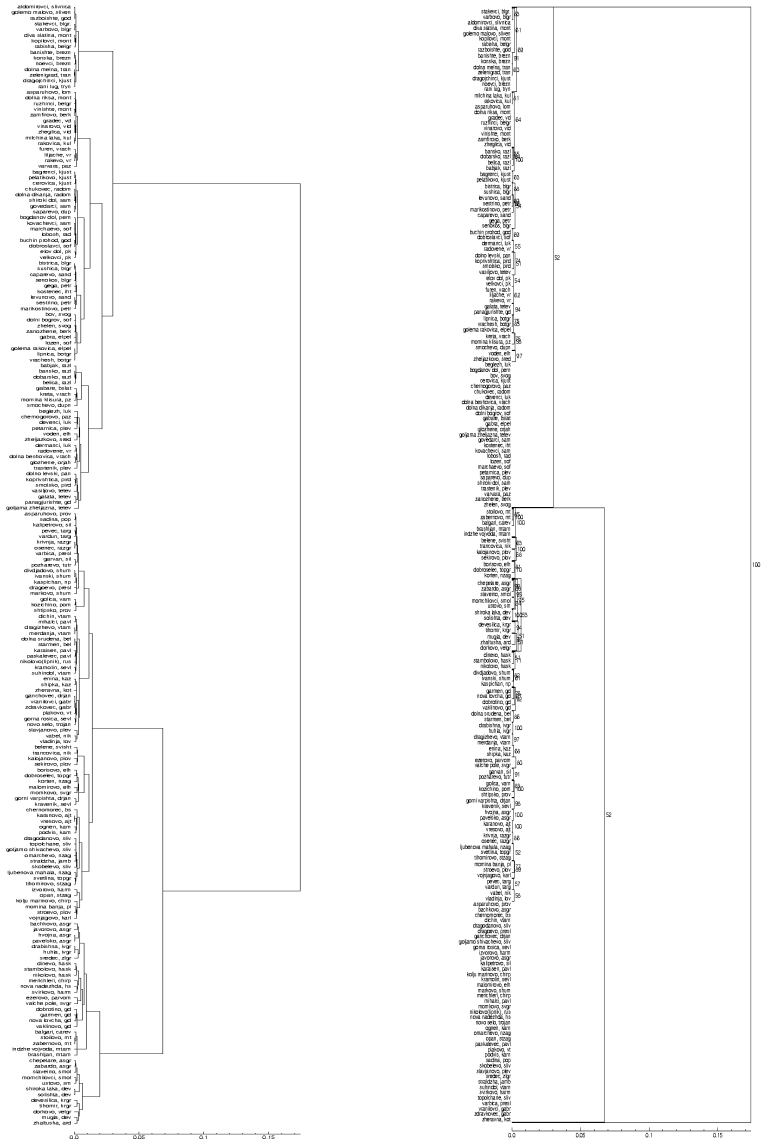

Figure D.7: Ward's method dendrograms, plain and with the noise.

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 I. Blok (1993). *The Interpretation of Focus.*
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 P. 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.*

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. Mnika 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 (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 Student.*

87. Xiaoyan Xu (2010). *English language attrition and retention in Chinese and Dutch university students.*

88. Jelena Prokić (2010). *Families and Resemblances.*

GRODIL
Secretary of the Department of General Linguistics
Postbus 716
9700 AS Groningen
The Netherlands