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. Levenshtein 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;
CHAPTER 1. INTRODUCTION

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?
1.2. THE MAIN RESEARCH QUESTIONS

- 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 1. INTRODUCTION