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 yar 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-
procedure 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 ALPHAMALIG 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
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.