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.](image)

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.

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 циал /tsiːl/ ‘whole - masc sg’ and цели /tsiːli/ ‘whole - pl’
- final [o] or [e] in such words as наше /nasa/ ‘ours’
- presence or absence of the second [j] in яйце /jaʃce/ ‘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 *my’ /m7Z/ ‘man’, *πv’t /pyt/ ‘road’ and *σbota /sybota/ ‘Saturday’ where in the TZS the reflex is [u], in the SW it is [u], and in the NW and in most of the parts east of the yat line [v]. 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.](image)

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
CHAPTER 4. COMPARISON TO THE TRADITIONAL MAPS

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.
4.2. WESTERN DIALECTS

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 /5a den/ ‘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 /bolna/ ‘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
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 č yat in word đve /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. 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 /pekla/ ‘bake - fem 1st sg’ ([pekla] vs [pekəla]). This characteristic is shared by most (but not all) of the southeast and two noncontiguous areas in the northeast.

2. OT F 19: absence of a vowel in the verbal root *мак- (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.
4.3. EASTERN DIALECTS

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:

- 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 [ła] and before soft syllable is [e] ([b‘al] vs. [b‘eli]). Under the influence of the Balkan dialects [e] is almost completely replaced by [e].
- change of consonant /d/ into [n] before /n/ (*dn > [nn])
- non-existence of consonants /ʃ/ and /χ/

Three of these distinguishing characteristics are not supported by (his own) OT and BDA maps.† 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 дialeктен 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 /fl/ and /xl/, 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 (‘×’ sign) are concentrated in the right low corner of the plot, together with the Balkan varieties (‘◦’ sign). It is not possible to detect a separate cloud representing Moesian 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 '×', 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.
4.4. DISCUSSION

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.