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, 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 жиелi /зиелi/ ‘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 a 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*. 
2.2. SELECTION OF WORDS AND FEATURES

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 *é (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...
CHAPTER 2. DATA

Example: [xl¹ap] vs. [xlep] vs. [xlrp] ‘bread’

Words in the data set:

- бела /bel/ ‘white - pl.’,

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

Example: [ja¹def] vs. [e¹def] ‘eat-you’

Words in the data set:


3. Presence or absence of initial prothetic [j]

Example: [agne] vs. [jagne] ‘lamb’

Words in the data set:


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

Example: [ko¹e] vs. [koºe] ‘which’

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

5. Elision or no elision of [j]

Example: [neja] vs. [neaj] ‘she - accusative’

Words in the data set:


6. Reflexes of the back nasalized vowel

Example: [kaºde] vs. [kuºde] ‘where’

Words in the data set:


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

7. Reflexes of the front nasalized vowel

Example: [zet] vs. [z\!'at] vs. [z\!'it] vs. [zent] ‘son-in-low, brother-in-low’


8. Reflexes of the back yer

Example: [ta\!'kof] vs. [ta\!'kvf] vs. [ta\!'kaf] vs. [ta\!'kef] ‘such’


9. Reflexes of the front yer

Example: [tv\!'ko] vs. [te\!'ko] vs. [\!'\!'ko] vs. [t\!'\!'ko] ‘thin - neut’


10. Choice of the vowel inserted between the two last consonants in words \!'at\!'er ‘wind’ and \!'\!'an ‘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: [v\!'at\!'r] vs. [veter] ‘wind’

Words in the data set: вятър /v\!'at\!'r/ ‘wind’, огън /\!'og\!'n/ ‘fire’

11. Vowel reduction

Example: [pe\!'pl] vs. [pe\!'pl] vs. [pe\!'pl\!\!] ‘ash’

Words in the data set: вечер /vet\!'er/ ‘evening’, пепел /pe\!'pl/ ‘ash’, понеделник /pone\!'delnik/ ‘Monday’

12. Reflexes of yer

Example: [e\!'zik] vs. [e\!'zik] ‘tongue’

Words in the data set: език /e\!'zik/ ‘tongue’, сирене /sire\!'ne/ ‘cheese’

13. Rounding of front vowels

Example: [\!'\!'f] vs. [\!'\!'f] vs. [\!'\!] ‘alive’

4In 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.
CHAPTER 2. DATA


14. Unrounding of front vowels
Example: [klʃutʃ] vs. [klɪʃ] ‘key’
Words in the data set: ключ /klʃutʃ/ ‘key’

15. Alternation /o/-/e/
Example: [dʒɔp] vs. [dʒɛp] ‘pocket’
Words in the data set: джоб /dʒɔb/ ‘pocket’, наже /naʃe/ ‘ours’, пепел /pepet/ ‘ash’

16. Presence or absence of vowel elision
Example: [neˈdelə] vs. [nɛdə] ‘Sunday’

17. Change by analogy, like [*dolu] vs. [*dole] ‘down’, presumably due to analogy with [*gore] ‘up’
Example: [*dolu] vs. [*dole] ‘down’ analogy with [*gore] ‘up’
Words in the data set: долу /dolu/ ‘down’, пека /peˈkɔ/ ‘bake - 1st sg’ (due to analogy with се/ka /seˈkɔ/ ‘chop - 1st sg’)

18. Reflexes of syllabic liquids

19. Reflexes of *tʃ, *dʃ
Example: [ʃɛfta] vs. [ʃɛʃta] vs. [ʃɛʃa] ‘lentils’

20. Variation of the original initial cluster чр + following vowel ь or ь
Example: [ʃerˈven] vs. [ʃyrˈven] ‘red’
Words in the data set: червен /ʃerˈven/ ‘red’, черен /ʃerən/ ‘black’, череша /ʃeˈrefa/ ‘cherry’

21. Epenthetic [l]
Example: [zəˈmla] vs. [zemˈla] vs. [zemˈla] ‘land’
Words in the data set: земя /zeˈmla/ ‘land’

22. Presence or absence of voiced affricates
2.2. SELECTION OF WORDS AND FEATURES

Example: [dʒɔp] vs. [ʒɔp] ‘pocket’

23. Presence or absence of palatalized consonants
Example: [ʃtʃ] vs. [ʃɪʃ] ‘wolf’

24. Results of palatalization of /st/, /zd/ in words corresponding to Standard Bulgarian ['gosti] guests, ['grozde] grapes
Example: ['gosti] vs. ['ɡosɪ] vs. ['ɡɔjse] ‘guest - pl’
Words in the data set: гости /ˈɡosti/ ‘guest - pl’, грозде /ˈɡrozde/ ‘grapes’

25. Presence or absence of simplification of the clusters стр /str/, здр /zdr/
Example: [ˈsestrax] vs. [ˈsestrə] ‘sister’
Words in the data set: здрав /ˈzdrav/ ‘healthy’, сестра /ˈsestrə/ ‘sister’, страх /ˈstrax/ ‘fear’

26. Presence or absence of epenthesis of [t] and [d] in the clusters [sr] and [zd]
Example: [strədə] vs. [strədə] ‘Wednesday’
Words in the data set: сръдна /ˈstrədə/ ‘Wednesday’

27. Presence or absence of the voiceless velar fricative
Example: [strax] vs. [stra] ‘fear’

28. Presence or absence of the voiceless labiodental fricative
Example: [ˈfurna] vs. [ˈvurna] vs. [ˈvɔɾna] vs. [ˈxurna] vs. [ˈʃurna] ‘oven’
Words in the data set: фурна /ˈfurna/ ‘oven’

29. Preservation or loss of */v/ before rounded vowels
Example: [vol] vs. [ol] ‘ox’

30. Presence or absence of prothetic [v] before rounded vowels
Example: [ˈɔɡn] vs. [ˈvɔɡn] ‘fire’
Words in the data set: огън /ˈɔɡn/ ‘fire’, опех /ˈɔpɛx/ ‘walnut’

31. Devoicing of obstruents in certain positions
Example: [ʒif] vs. [ʒɪv] ‘alive’
CHAPTER 2. DATA

32. The form of the preposition *vz and the prefix *v
Example: [vlizam] vs. [u]lizam] 'enter - 1st sg'
Words in the data set: ВЛИЗАМ /vlizam/ 'to enter - 1st sg', ВЪВ /vvv/ 'in'

33. Various assimilations and dissimilations
Example: [of şi] vs. [os şi] 'sheep'
Words in the data set: едно /e'dno/ 'one - neut', МНОГО /mnogo/ 'much, many', ОВЦА /ovţa/ 'sheep', ОВЦЕ /ovţe/ 'sheep - pl', ОВЧАР /ovţar/ 'shepherd', ОВЧАРИ /ovţari/ 'shepherd - pl', ТЪМНО /t7mno/ 'dark - neut'

34. Nonsystematic changes in individual words
Example: [bryzo] vs. [bryze] 'quickly'
Words in the data set: бързо /bryzo/ 'quickly', ВЕЧЕ /veçe/ 'already', ВЧЕРА /v7era/ 'yesterday', ЧОВЕК /ʧo'vek/ 'person'

35. Morphophonemic alternations or suffixes connected with the formation of secondary imperfective verbs
Example: [vlizam] vs. [vlazam] vs. [vlavam] 'enter - 1st sg'
Words in the data set: ВЛИЗАМ /vlizam/ 'enter - 1st sg', ВРЪЩАМ /vryšam/ 'give back - 1st sg', ПЛАЩАМ /plaštam/ 'pay - 1st sg'

36. Form of certain grammatical endings, such as that of the first person plural in all tenses
Example: [b]axme vs. [bexmo] 'were - 1st pl'
Words in the data set: бъхме /b]axme/ 'were - 1st pl'

37. Choice of the suffix in certain nouns that originally belonged to the n-stem nouns:
Example: [kamik] vs. [kamik] vs. [kamen] 'stone'
Words in the data set: ЕЧЕМИК /eʧe'mik/ 'barley', КАМЪК /kamik/ 'stone'

38. Various forms of words that are derived from a common Old Bulgarian form
Example: [vie] vs. [vi] vs. [ve] 'you'
Words in the data set: ВИЕ /vie/ 'you', И /i/ 'she - dative', ИМ /im/ 'they - dative', НИЕ /nie/ 'we', ОНЕЗИ /ondezi/ 'those', ТОБА /tø'va/ 'this - neut', ТОГАВА /tø'gava/ 'then', Я /ja/ 'she - accusative'

39. Different position of stress
Example: [vino] vs. [vi'no] '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 *ē (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. [bjal], [bjael] or [bje]. 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 [r]. For example [beli] vs. [bali], [bajeli] or [bali].

Taking into account various phonetic features, including reflexes of *ē (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 *ˇe* (yat) before non-palatal syllables is [ɬa] and before palatal syllables is [e] ([b'ål] vs. [beli]). Under the influence of the Balkan dialects [e] 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 /t/ and /u/
- 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 *ˇe* (yat) before non-palatal syllable is [ɬa] and before palatal syllable is [e] ([b'ål] 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/, /ʒʃ/, /ŋʃ/, 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:
2.3. TRADITIONAL SCHOLARSHIP

- large number of palatal consonants in various positions
- soft pronunciation of consonants /ʒ/, /ʃ/ and /ʃʃ/
- 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 *ě (yat) is always [e]
- the reflex of old back nasal vowel yus and back vowel yer is [y]
- the reflexes of old groups *Ѧ and *Ѧ are /ʃ/ and /ʒd/
- ending /e/ instead of formal Bulgarian /i/ for plural past active aorist participles
- the masculine definite article is /а/ 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 *ě (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 [o] is found.
- the reflex of Old Bulgarian back yer s and front yer ҫ is mostly [a], but in the western parts reflex [o] is found instead of [a]
- the reflexes of old groups *Ѧ and *Ѧ are /ʃ/ and /ʒd/
- change of /ɔ/ into [e] after /ʒ/, /ʃ/, /ʃʃ/ and /ʃl/
- single masculine definite article is /о/ 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 *č (yat) is always [e]
- the reflexes of old groups *p and *d are /ŋ/ and /ŋʒ/
- 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 2. DATA