Identifying Linguistic Structure in a Quantitative Analysis of Bulgarian Dialect Pronunciation

Jelena Prokic
j.prokic@rug.nl

03.11.2006. Sofia

Outline

- The goal of the thesis
  - Aggregate analysis
  - Identification of linguistic structure in the aggregate analysis
- Previous work
- Aggregate analysis
  - New data set
  - L04
- Regular sound correspondences
  - Extraction
  - Quantification
  - Results

The Goal of the Thesis

- To do an aggregate analysis of the Bulgarian dialects using
  - new data set
  - L04

- To identify the underlying linguistic structure in the aggregate analysis
  - regular sound correspondences were extracted from the aligned pairs of words
  - for the 10 most frequent sound correspondences a separate analysis of each site was made

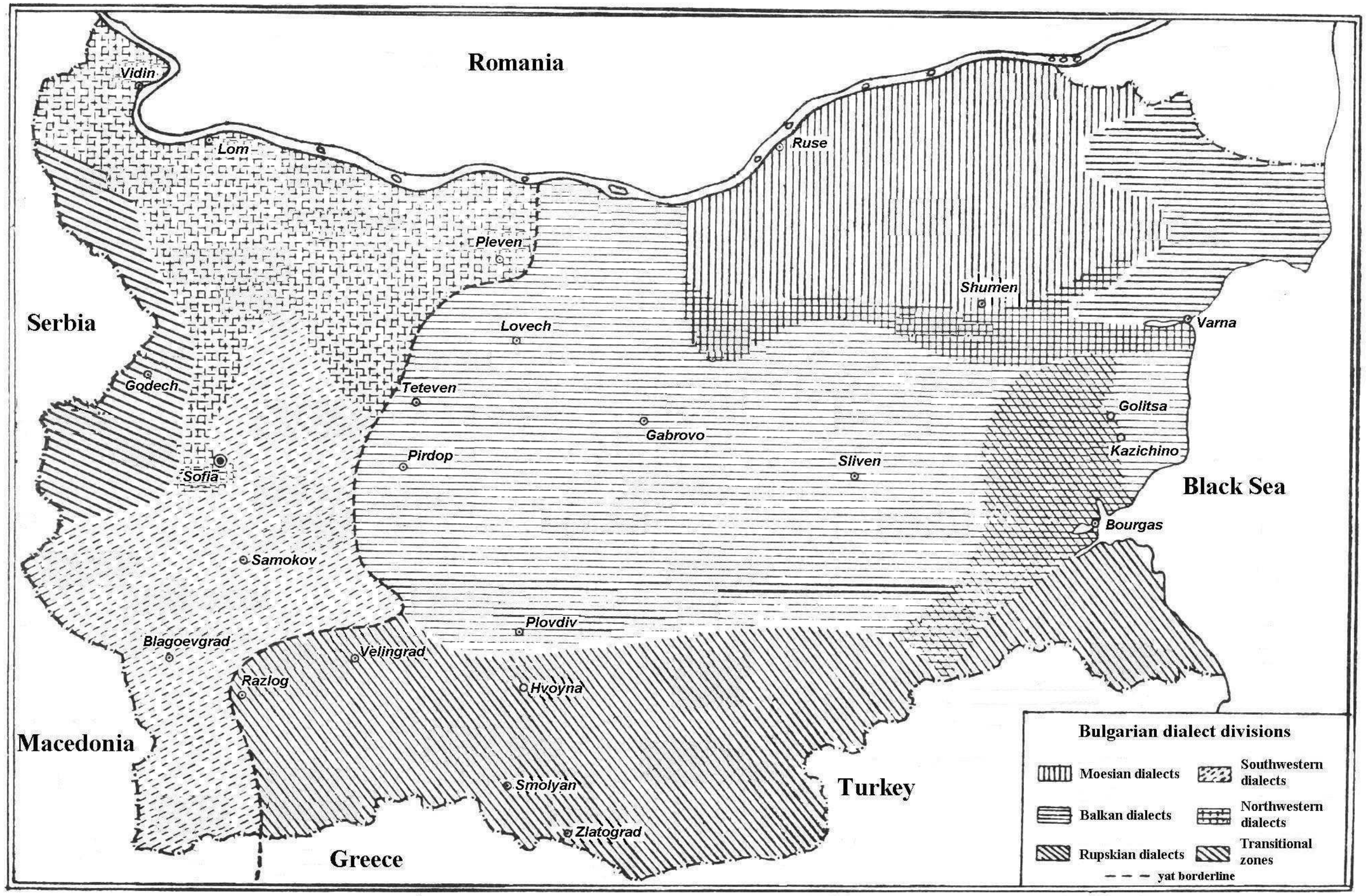
Previous Work

- Aggregate analysis of dialect divisions
  - successfully applied to various languages
  - on Bulgarian applied by Osenova et all. (2006)
- Identification of linguistic structure in the aggregate analysis
  - aggregating over a subset of data (Nerbonne, 2005)
  - factor analysis (Nerbonne, 2006)
- Extraction of sound correspondences
  - Kondrak (Kondrak, 2002) applied it in the task of cognate identification

Osenova et al. 2006

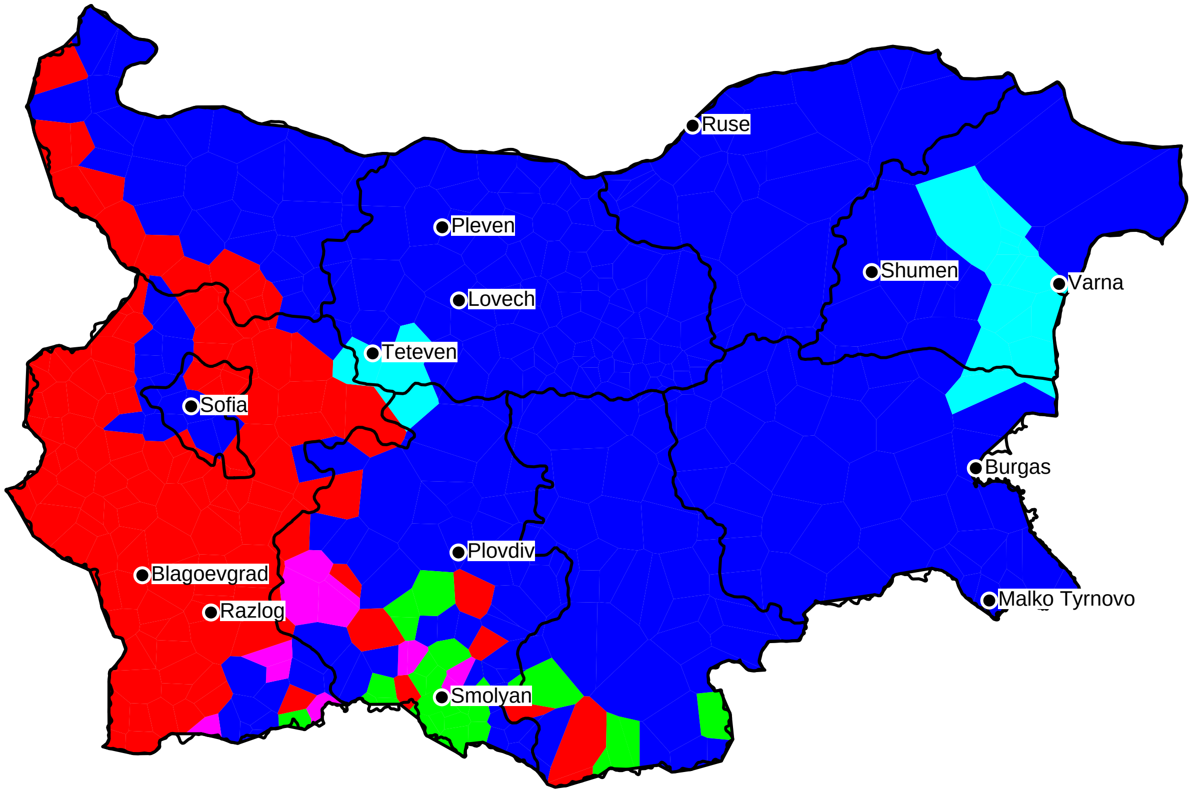
- Aggregate analysis of dialect divisions in Bulgaria
  - data set: 36 words collected from 490 sites
  - suprasegmentals and diacritics were removed
  - L04 toolkit
- Cluster analysis
- Multidimensional scaling

Osenova et al. 2006 Cont.



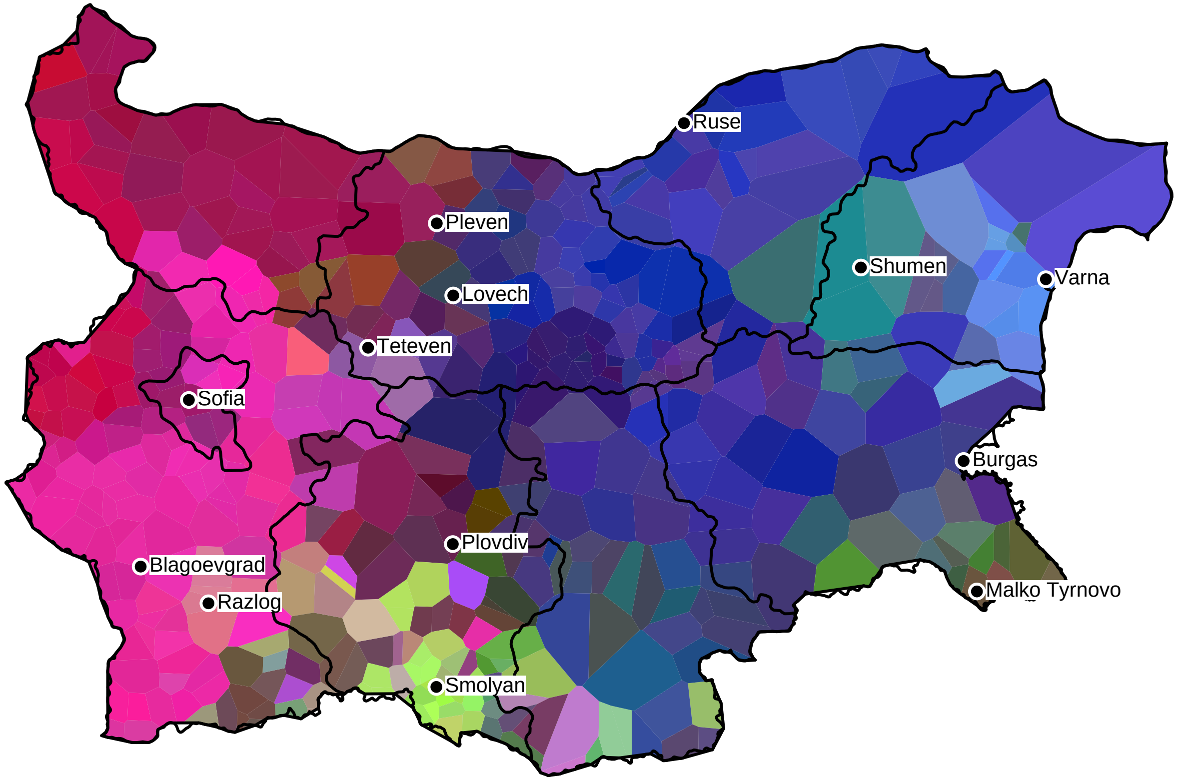
Map of Bulgarian dialect divisions taken from Stoykov (2002)

Osenova et al. 2006 Cont.



Classification map from Osenova et al. (2006)

Osenova et al. 2006 Cont.



Continuum map from Osenova et al. (2006)

Osenova et al. 2006 Cont.

- Both maps give a reliable picture of the dialect divisions
  - the most important division is between East and West
  - Rodopi area is the most incoherent
  - area around Varna and Schumen is distinct from the neighbouring areas
  - area around Teteven is also distinct

- Dialectometrical methods were successfully applied to a Slavic language for the first time

Extraction of Linguistic Structure

- Nerbonne (2005)
  - aggregates over a subset of the data, namely vowels
  - the differences between the sites are calculated using both complete phonetic transcriptions and also using only vowels
  - results: vowels are probably responsible for a great deal of aggregate differences (r = 0.936)
- Nerbonne (2006)
  - applies factor analysis to the results of the dialectometrical analysis
  - only vowels are investigated
  - results: 3 factors are most important, explaining 35% of the total amount of variance

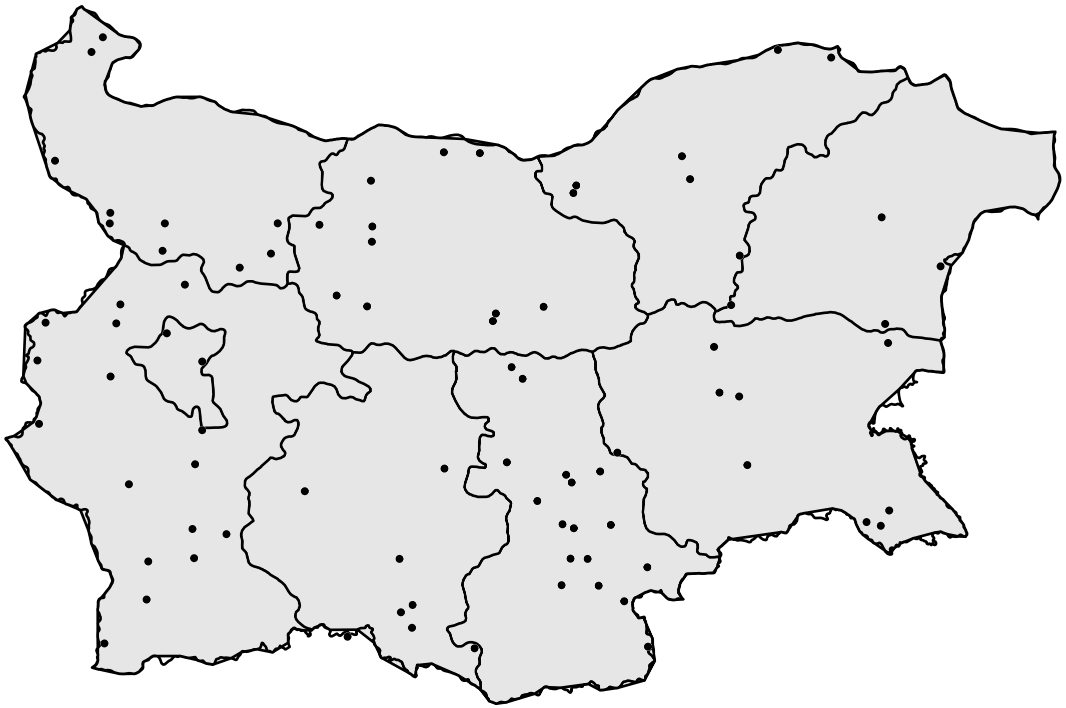
Sound Correspondences

- Kondrak (2002) extracts regular sound correspondences and uses them to identify cognates in a bilingual word list
- Melamed's parameter estimation models were adopted and used to determine sound correspondences
- The more regular sound correspondences two words contain the more likely it is that they are cognates and not borrowings
- This method has outperformed other methods for cognate identification

New Data Set

- Data from the project Buldialect – Measuring linguistic unity and diversity in Europe
- 117 words collected from 84 sites
- Words include nouns, verbs, pronouns, and prepositions in different word forms
- All phonetic transcriptions were in X-SAMPA format

Distribution of 84 Sites



Distribution of 84 sites from the new data set

Part I: Aggregate Analysis

- L04 toolkit
  - alignment of word transcriptions
  - Levensthein algorithm
  - cluster analysis
  - multidimensional scaling

- Preprocessing of the data
  - suprasegmentals and diacritics were removed
    - s’ s\ “s *s *”s “s\ all represented as s
  - palatalized/non-palatalized opposition preserved

Aggregate Analysis Cont.

- Alignments were based on the following principles:
  - vowel can match only with the vowel
  - consonant can match only with the consonant
  - [i] and [u] can match both with vowels and sonorants
  - [j] can match both with vowels and consonants

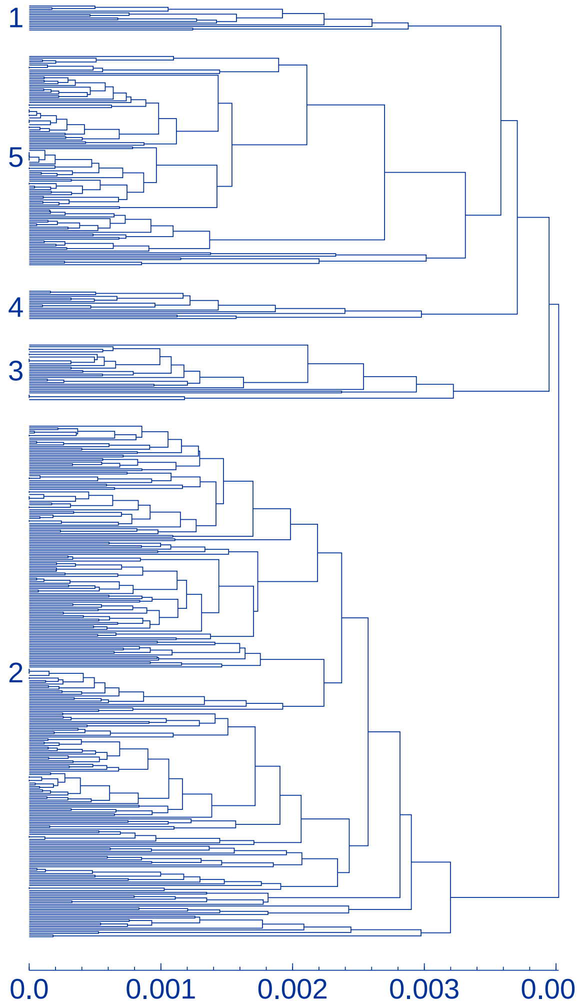
Example 1:

```
[  4] zelenigrad
[ 24] merichleri
      b    e    l    i
    b_j    a    l    i
--------------------------
      1    1
```

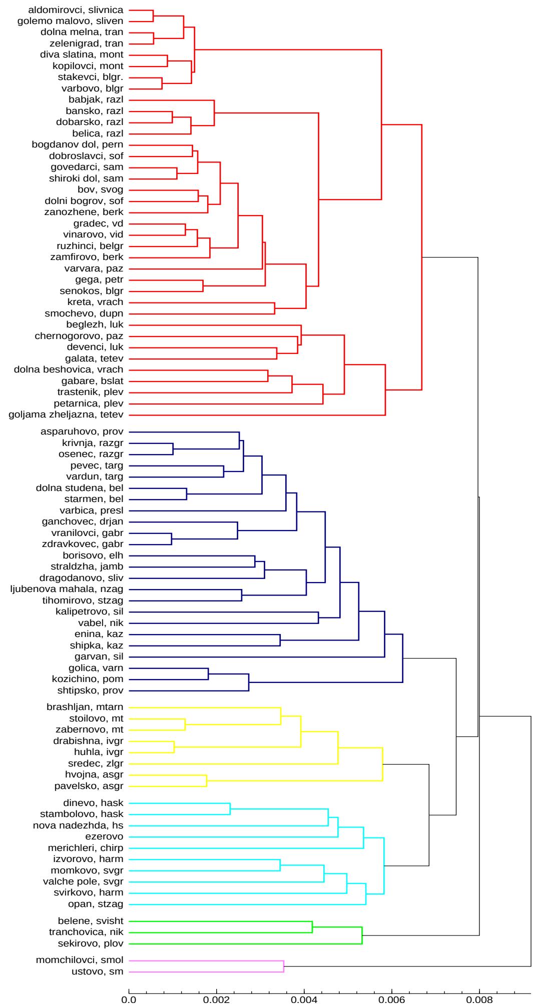
Aggregate Analysis Cont.

- Insertions, deletions, and substitutions have the same cost – 1
- The distance between two strings was normalized by the length of the longest alignment that gives the minimal cost
- The distance between two aligned strings in Example 1 would be 0.5
- Distances between the aligned pairs of transcriptions are used to calculate the distance between each pair of sites
- The results were analyzed using cluster and multidimensional scaling (MDS) analyses

Dendograms

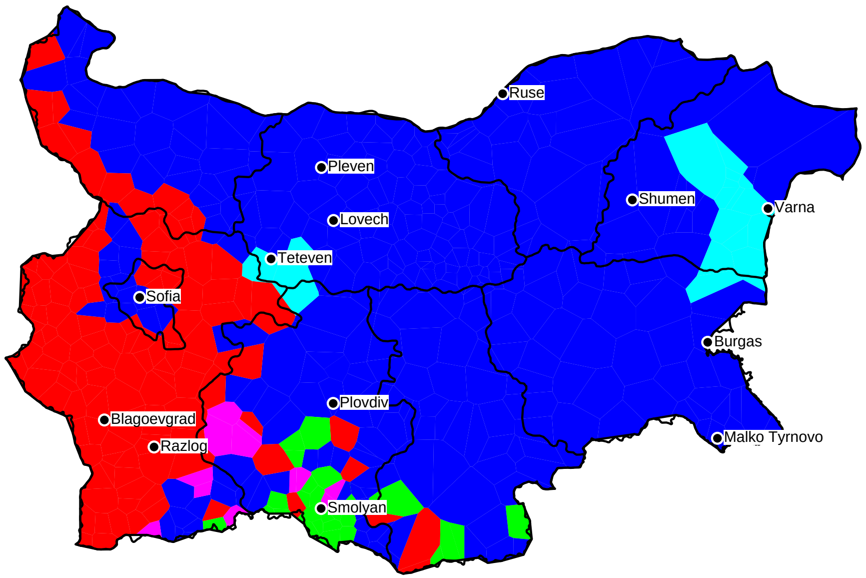


Old data set (Osenova et al., 2006)

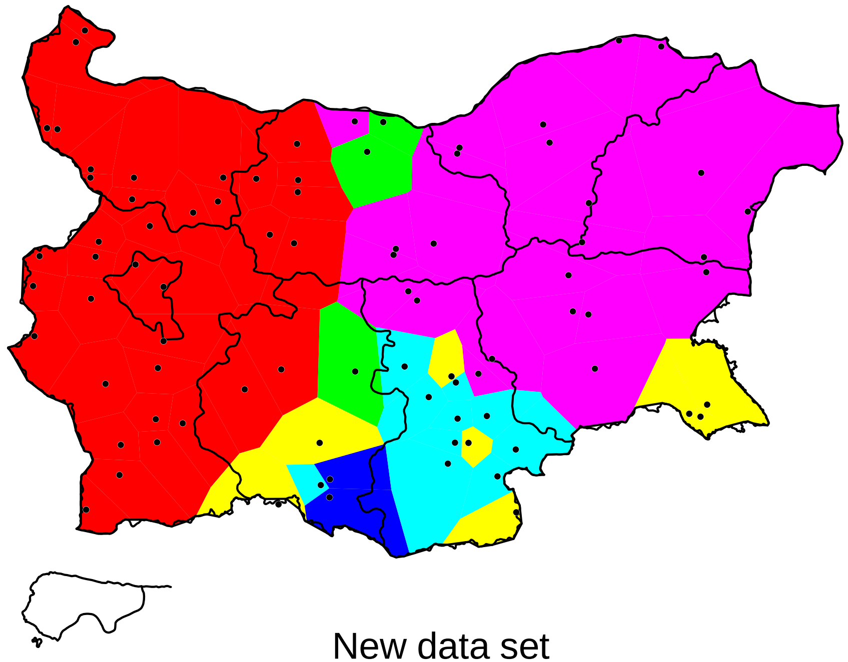


New data set

Cluster Maps

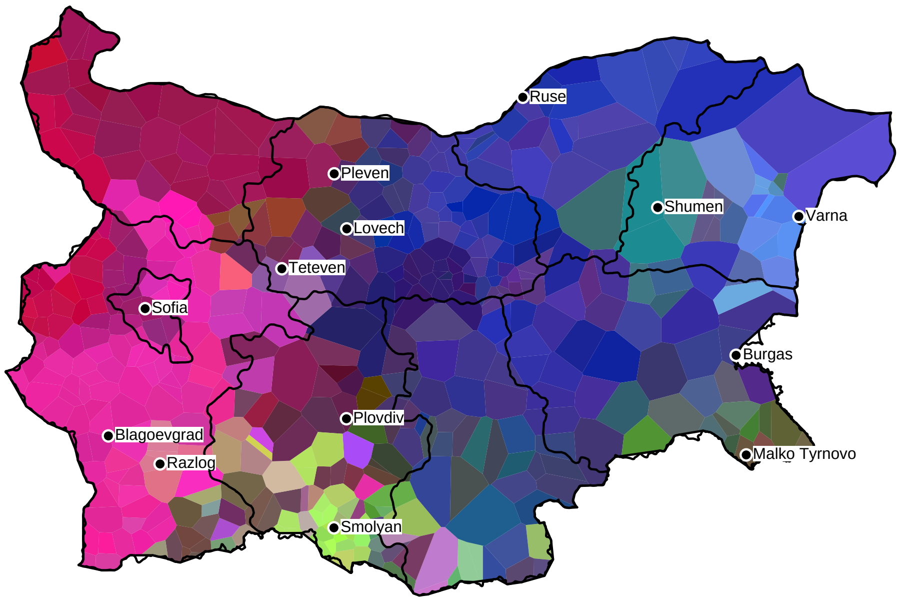


Old data set

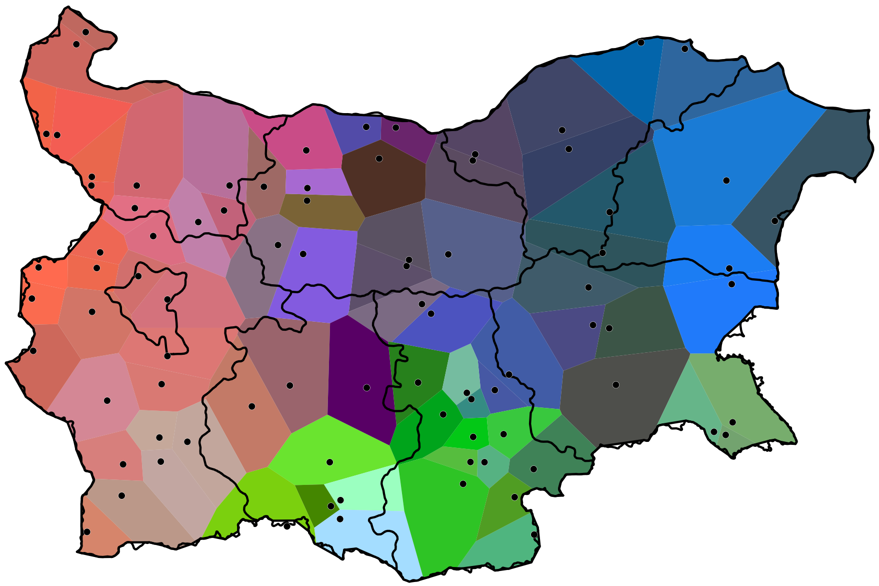


New data set

MDS Maps



Old data set



New data set

Results

- Clear division between East and West (‘yat’ realization border)
- Rodopi area is the most incoherent
- Both cluster and MDS map conforms with the maps presented in Osenova et al. (2006) and the map presented in Stoykov (2002)
- New data set gave a faithful picture of the dialect divisons in Bulgaria

Part II: Regular Sound Correspondences

- Problem: How to extract linguistic structure from aggregate comparison?

- Suprasegmentals and diacritcs were removed

- Word pronunciation transcriptions were aligned using L04

- For each pair of sites one best alignment for every word is taken into account (1.18 alignments per word pronunciation pair)

Example 2:

```
f   n   u   t   r   e              f   n   u   t   r   e
        v   ɣ   t   r   e              v       ɣ   t   r   e
------------------------------     -----------------------------
1   1   1                          1   1   1
```

Regular Sound Correspondences Cont.

- Phonetic distance between 2 segments is not taken into account, they are either identical or not
- Segments that do not match were extracted from all aligned pairs and sorted according to their frequency

Regular Sound Correspondences Cont.

Example 3:

```
Babjak     j   a                Beglezh          a    s
    Golica         ɑ    s            S. Dol     j    a
       -------------------         ------------------------------
          1    1    1                           1         1
```

phon1	j	a	
phon2		ɑ	s
No.	2	1	2

Table 1: Sound correspondences extracted from the alignments in Example 3

Regular Correspondences Cont.

- For each pair of sites and every word correspondences were summed

- Results:

e	o		ɑ	ɑ	ə	e	ɑ	v	j
i	u	ɣ	e	ɣ	ɣ	ɣ	ə		
52246	40981	39414	33391	33184	32753	32177	28976	22462	21475

Table 2: 10 most frequent correspondences from the whole data set

- Eight out of ten most frequent correspondences involve substitution or insertion/deletion of vowels

Correspondence Index

- Correspondence index is obtained by comparing every site to all other sites with respect to the first ten correspondences

- Goal:
  - to see if the site belongs to the group where 1 or the other sound is present
  - to see if there is a geographical cohesion in the sites that use 1 or the other sound in the correspondence

- Method:
  - only one best alignment for each word pronunciation pair was taken into account
  - all sound correspondences were extracted, both matching and non-matching

r	ɑ	e	o	e	s	k	d	l	v
r	ɑ	i	u	e	s	k	d	l	v
35	35	29	27	27	26	25	24	24	24

Table 3: 10 most frequent correspondences for the pair Aldomirovci-Borisovo

Correspondence Index Cont.

- For each pair of the most frequent correspondences (Table 2) a correspondence index is calculated for each site using the following formula:

$$\frac{1}{n-1}\sum_{j=1, j\neq i}^{n} S_i \to S_j, i = 1,..,n$$

n – number of sites

$S_i \to S_j$ - comparison of each 2 sites with respect to certain sound correspondence

Correspondence Index Cont.

$S_i \to S_j$ is calculated applying the following formula:

$$\frac{|s,s'|}{|s,s'| + |s,s|}$$

$|s,s'|$ - the number of times sound s seen in the word pronunciation collected at site1, was aligned with s' in the word pronunciation collected at site2

$|s,s|$ - the number of times sound s seen in the word pronunciation collected at site1 stayed unchanged

Correspondence Index Cont.

Correspondence index for the pair [e]-[i] for Aldomirovci and Borisovo:

s	e	i	e
s'	i	e	e
No.	29	0	27

Table 4: Number of times [e] correspondes to [e] and [i] for the site pair Aldomirovci-Borisovo

$$\frac{|e,i|}{|e,i|+|e,e|} = \frac{29}{29+27} = 0.517$$

Index for site1 (Aldomirovci)

$$\frac{|e,i|}{|e,i|+|e,e|} = \frac{0}{0+27} = 0.0$$

Index for site2 (Borisovo)

Correspondence Index Cont.

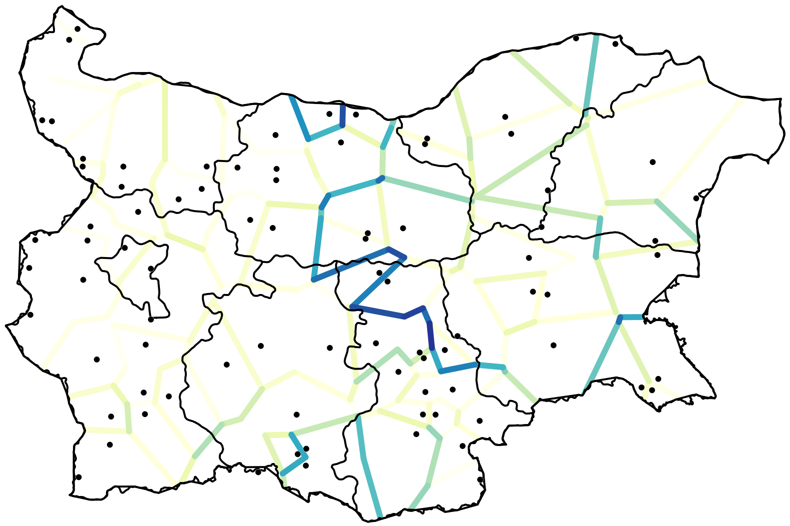
- Every site was compared to all other sites resulting in 83 indexes per site
- The general correspondence index for each site represents the mean of all 83 indexes
  - Aldomirovci 0.2328
  - Borisovo 0.1538
- Sites with the higher values of the general index represent the sites where sound [e] tends to be present
- Sites with the lower values of the general index represent the sites where sound [i] tends to be present

Correspondence Index Cont.

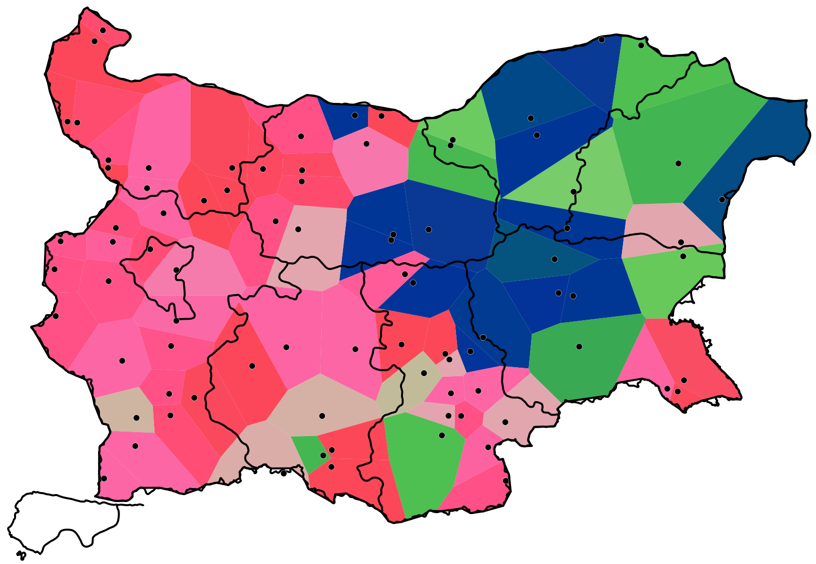
- General correspondence index was calculated for every site with the respect to the 10 most frequent correspondences found in the data set

- General indexes were analyzed using composite clustering and MDS-cophenetic method resulting in 2 types of maps:
  - composite cluster maps
  - MDS-cophenetic maps

[e]-[i] correspondence

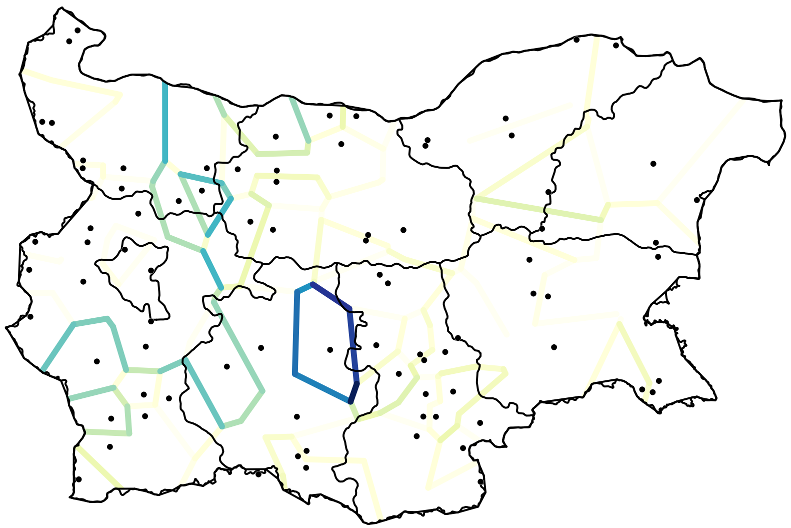


Composite cluster map

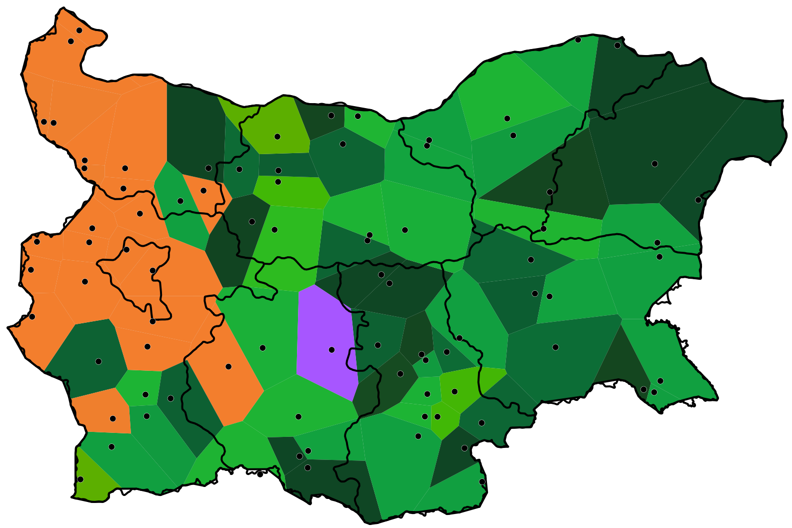


MDS-cophenetic map

[o]-[u] correspondence

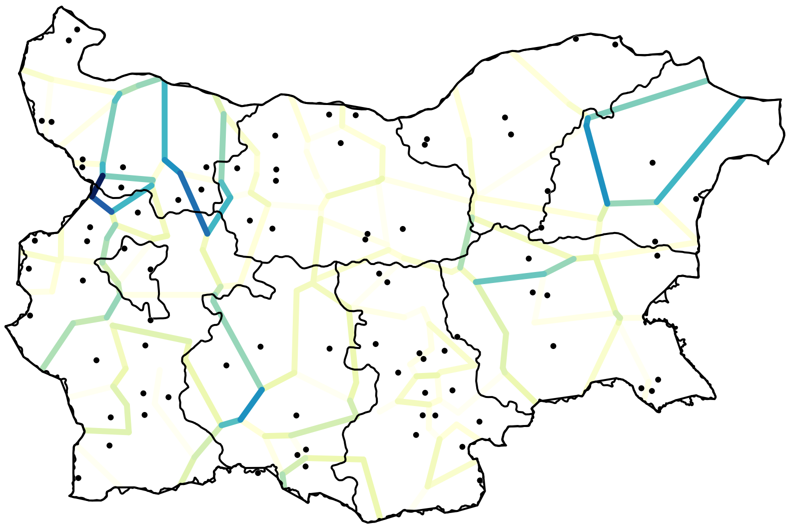


Composite cluster map

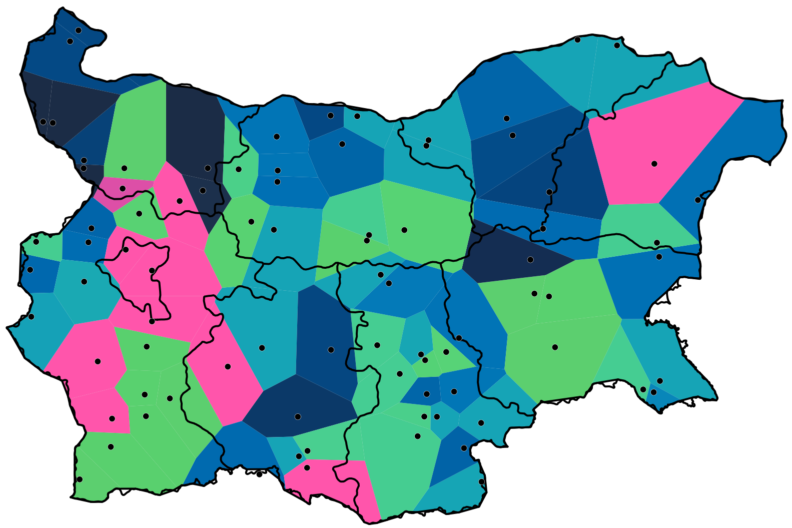


MDS-cophenetic map

[ɣ]-[ø] correspondence

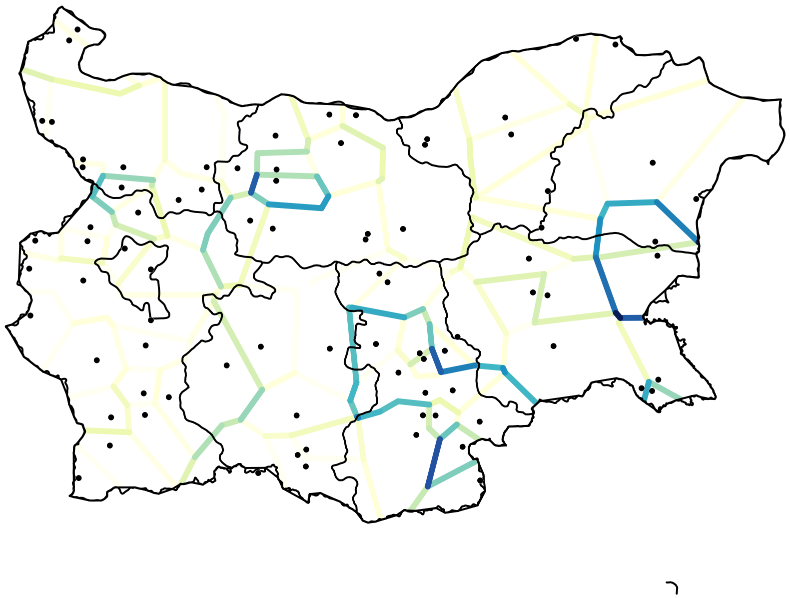


Composite cluster map

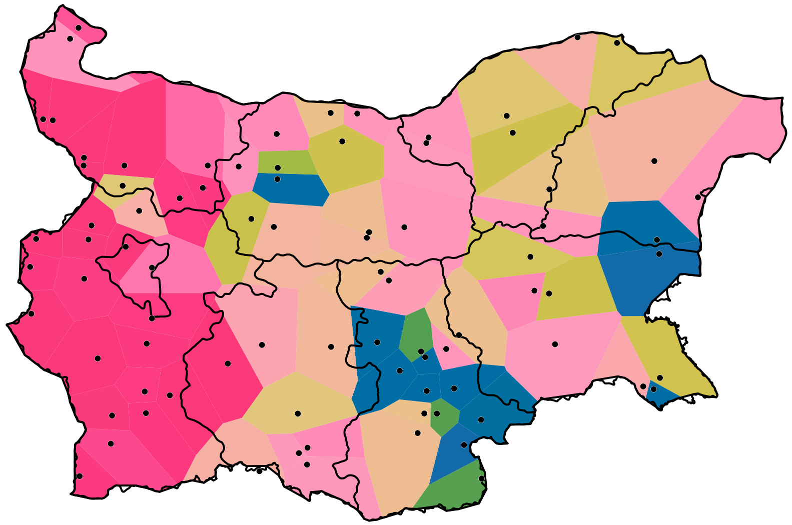


MDS-cophenetic map

[ɑ]-[e] correspondence

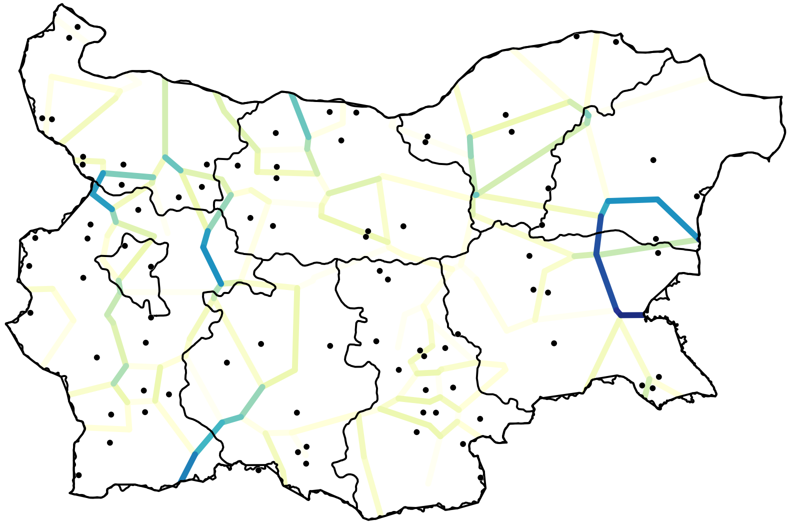


Composite cluster map

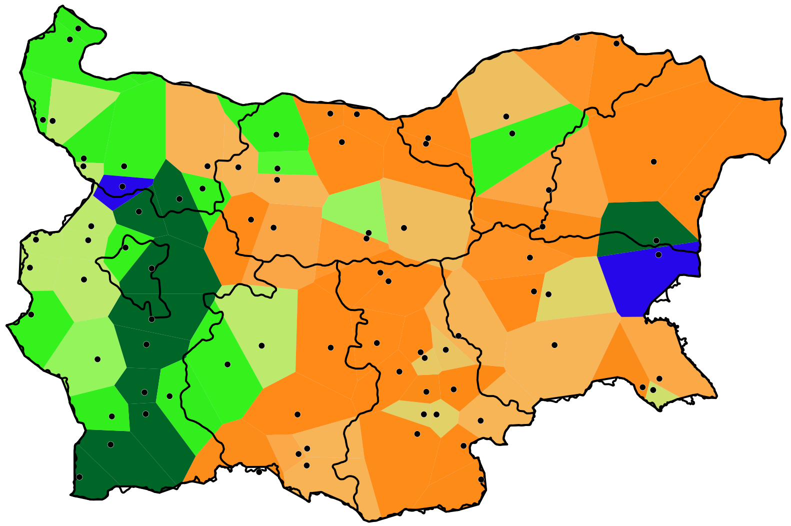


MDS-cophenetic map

[ɑ]-[ɣ] correspondence

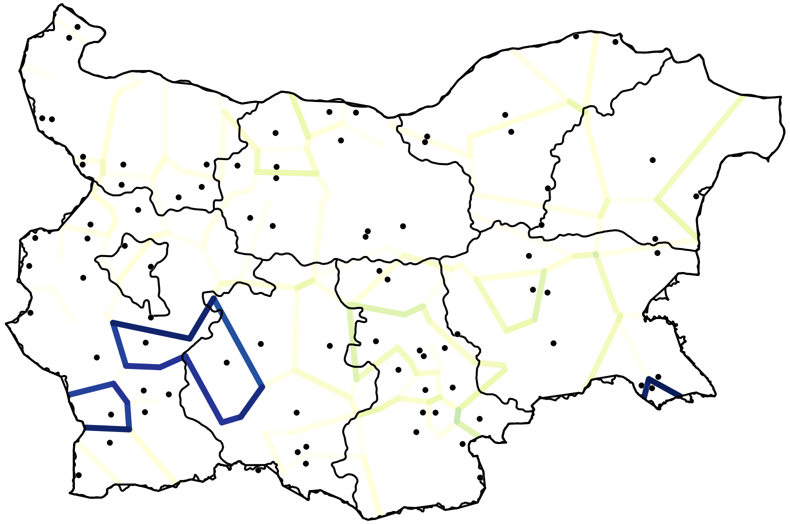


Composite cluster map

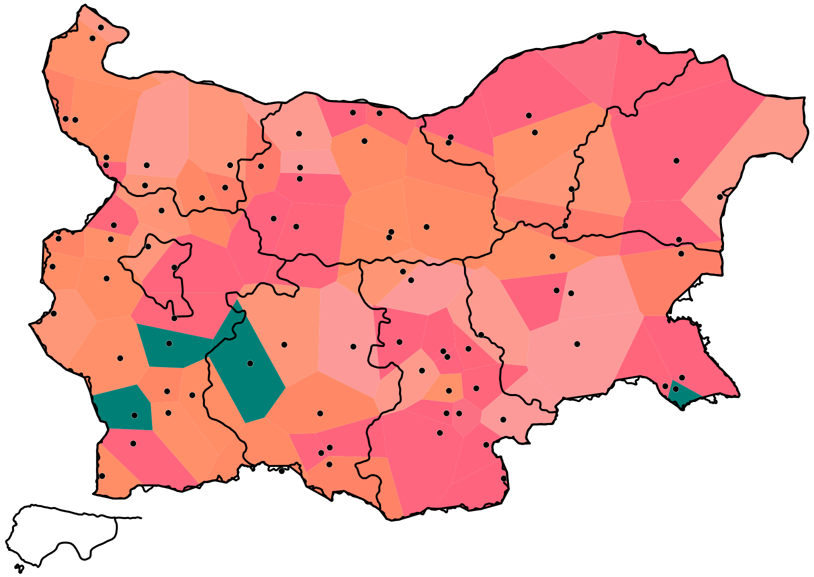


MDS-cophenetic map

[ə]-[ɣ] correspondence

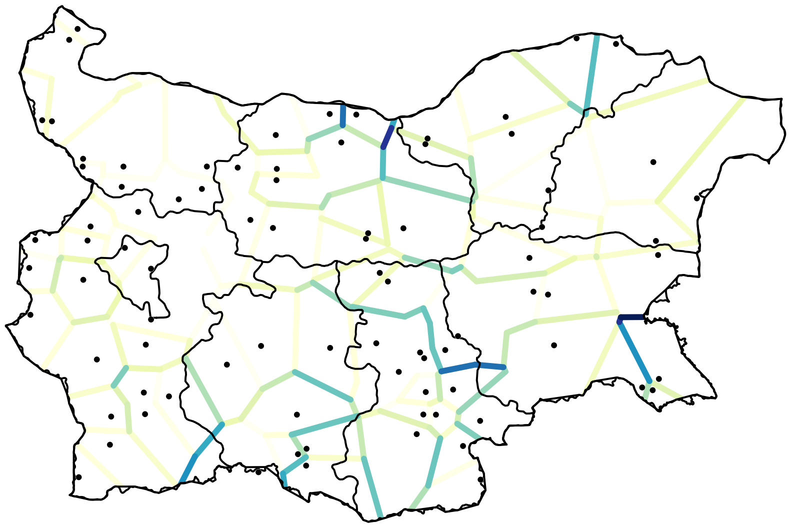


Composite cluster map

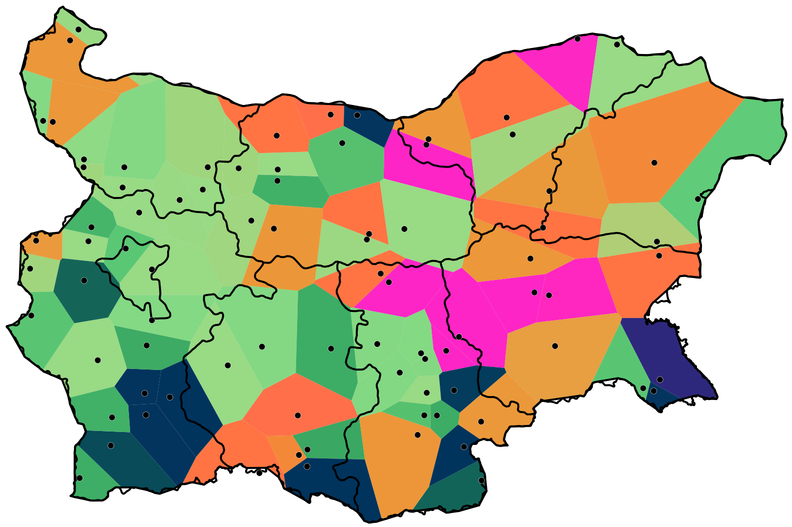


MDS-cophenetic map

[e]-[ɤ] correspondence

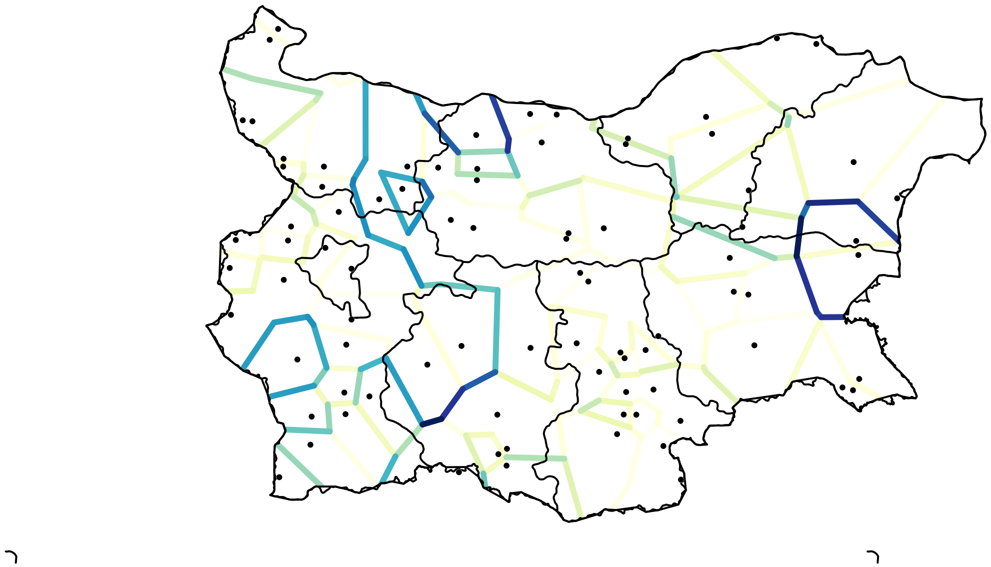


Composite cluster map

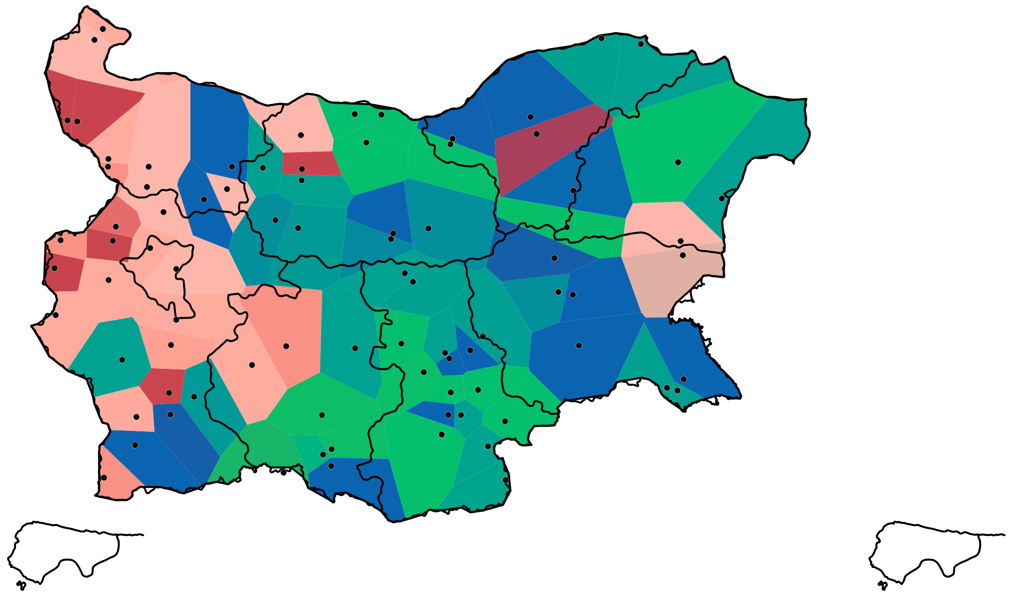


MDS-cophenetic map

[ɑ]-[ə] correspondence

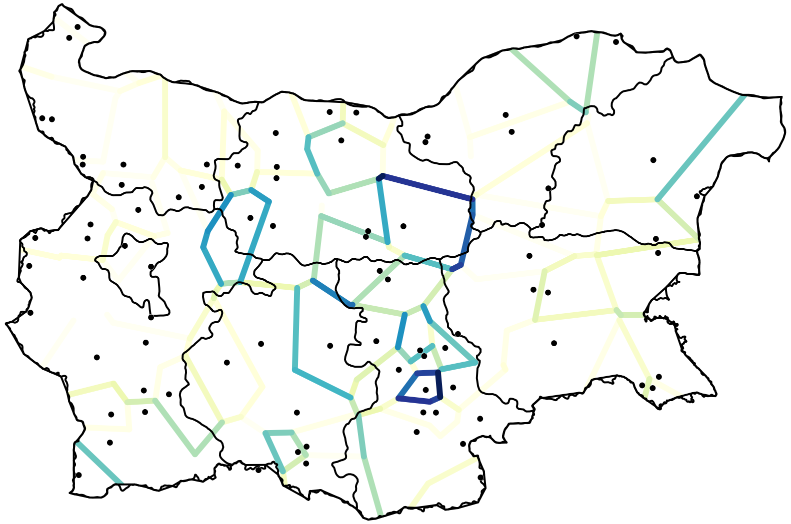


Composite cluster map

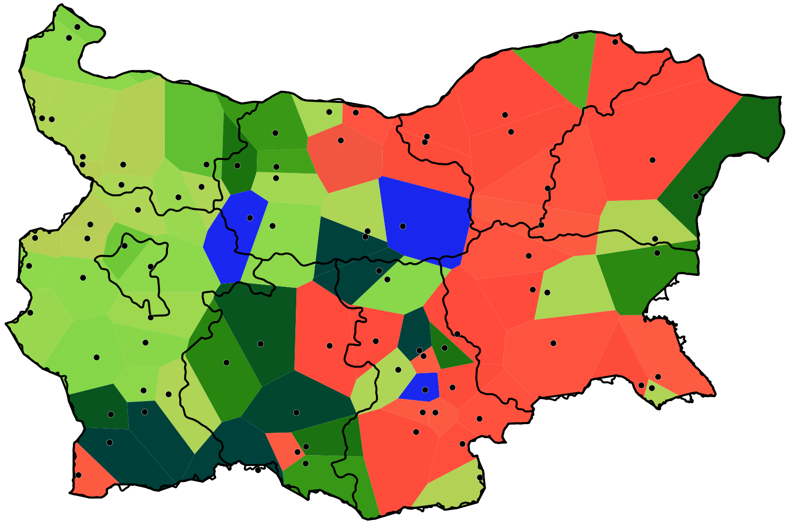


MDS-cophenetic map

[v]-[ø] correspondence

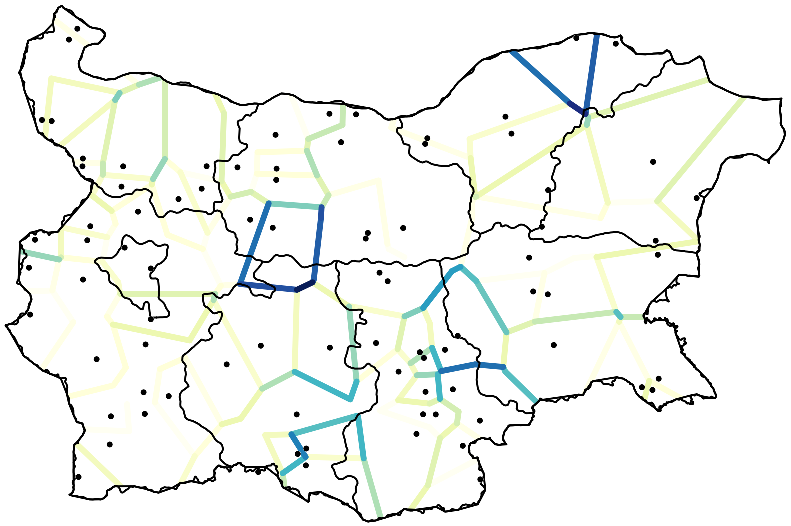


Composite cluster map

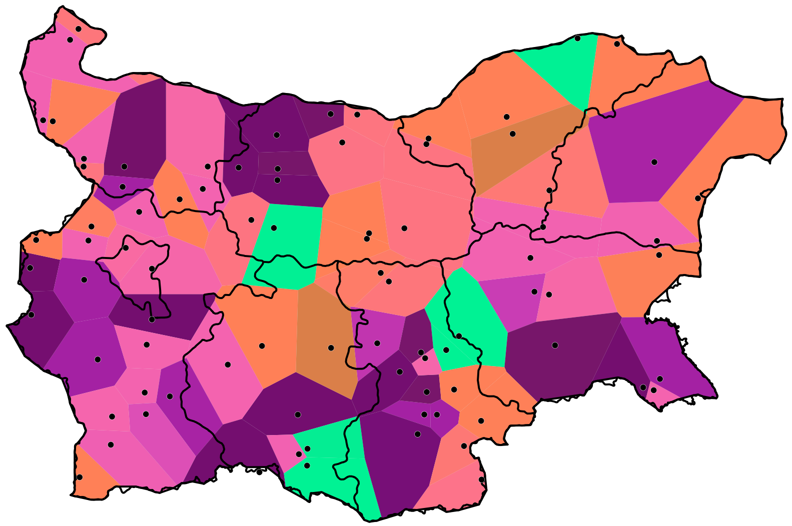


MDS-cophenetic map

[j]-[ø] correspondence



Composite cluster map



MDS-cophenetic map

Results

- Maps show that there is a geographical cohesion in the distribution of sites

- Maps show similarity with the traditional maps

- West-East division is based on the following correspondences:
  - [e]-[i] [o]-[u] [ɑ]-[e] [ɑ]-[ɤ] [e]-[ɤ] [ɑ]-[ə] [v]-[ø]

- Area around Kozichino and Golica is characterized by the presence of [e], [ɑ], and [v] sounds

Drawbacks of the Method

- Analyzes only one sound alternation at a time
- In the analysis of the sound alternations no context is taken into account

Future Work

- More sites should be included
- Instead of a simple phone representation of segments, feature representation of segments should be used
- Stress should be included
- MDS-cophenetic maps should include scale

References

- [Kondrak 2002] G. Kondrak. Algorithms for Language Reconstruction. PhD Thesis, University of Toronto.

- [Nerbonne 2005] John Nerbonne. Various Variation Aggregates in the LAMSAS South. Accepted to appear in (10/2005) Catherine Davis and Michael Picone (eds.) *Language Variety in the South III.* Tuscaloosa: University of Alabama Press.

- [Nerbonne 2006] John Nerbonne. Identifying Linguistic Structure in Aggregate Comparison. Accepted (5/2006) to appear in *Literary and Linguistic Computing* 21(4), 2006. (J.Nerbonne & W.Kretzschmar, Jr. (eds.) *Progress in Dialectometry: Toward Explanation*)

- [Osenova et al. 2006] Petya Osenova, Wilbert Heeringa and John Nerbonne A Quantitative Analysis of Bulgarian Dialect Pronunciation. To appear.

- [Stoykov 2002] S. Stoykov. Bulgarska dialektologiya. Sofia, 4th ed.