Experimenting with Bulgarian dialect data

Petya Osenova

BulDialects Project
5-6 December, Tuebingen

Plan of the talk

- Background
- Infrastructure for storing and processing dialect data
- The classification task
  - The preparation stage
  - The performance stage
- The interpretation and validation stage

Background (1)

A post-doc project *Measuring Language Contact*:

- Financially supported by Dutch Foundation NWO within a Program for Bulgarian and Romanian postdocs
- Host professor: John Nerbonne (Groningen)
- Collaborator: Wilbert Heeringa
- Duration: 1 year (2004), 6 months in Sofia and 6 months in Groningen

Background (2)

To test the hypothesis whether the phonetic distances can serve as a reliable base for exploring language contact phenomena. Thus, we had to:

- Classify Bulgarian dialects using a common **set of words** and relying on **phonetic phenomena**
- Connect the results to neighboring languages

Background (3)

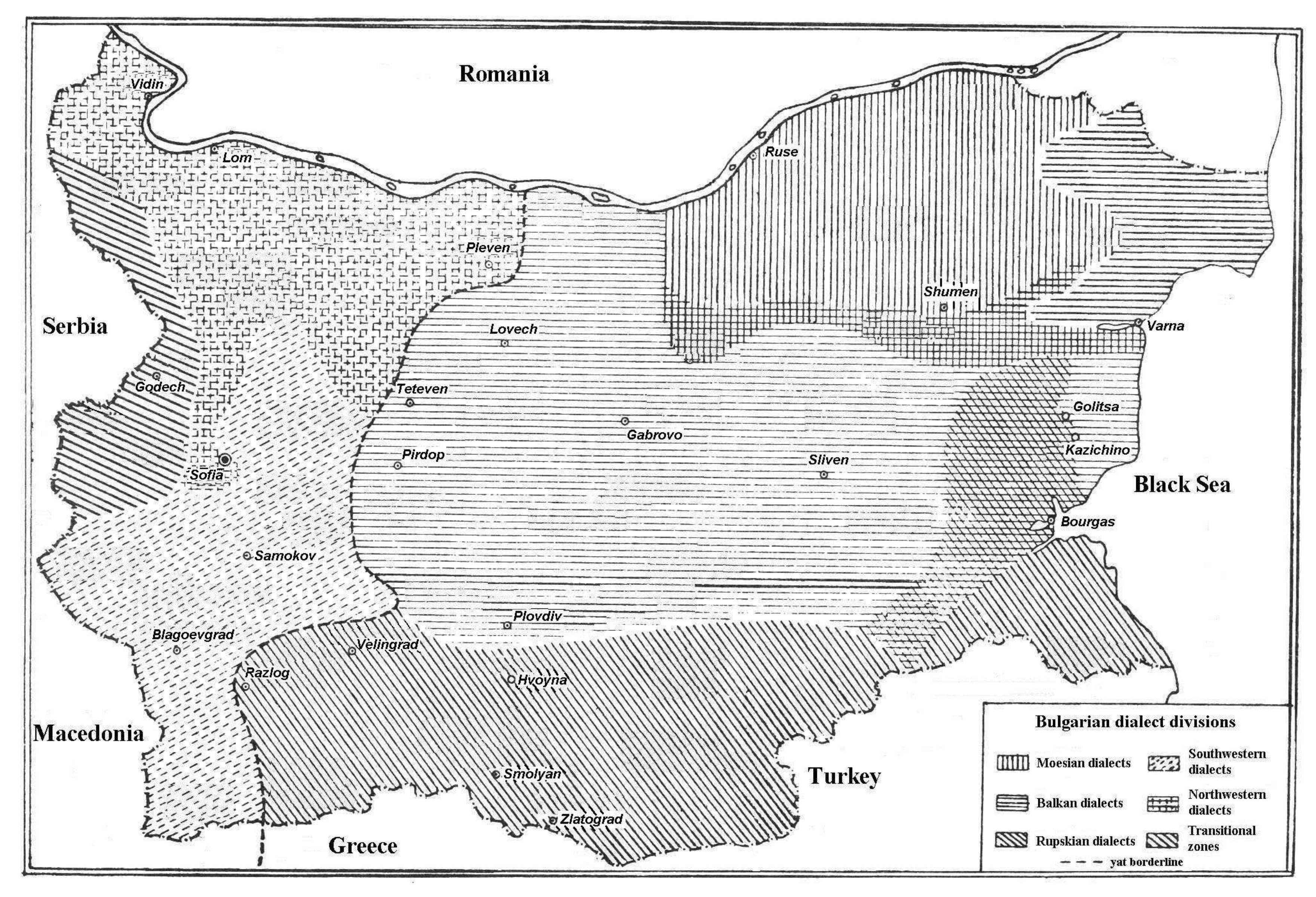
In recent years: successful phonetic measurements of Germanic and Romance dialects (Dutch, Norwegian, Sardinian, German) using Levenshtein distance as a base

Challenge to apply this technique on a Slavic language: **Bulgarian**

The classification task

To classify Bulgarian dialects using a common **set of words** and relying on **phonetic phenomena** (i.e. lexical, morphological and syntactic information was not taken into account)

To find the balance between the methods and the linguistic interpretation


Romania
Vidin
Lom
Ruse
Pleven
Shumen
Varna
Serbia
Lovech
Teteven
Godech
Golitsa
Gabrovo
Kazichino
Pirdop
Sliven
Sofia
Black Sea
Bourgas
Samokov
Plovdiv
Blagoevgrad
Velingrad
Razlog
Hvoyna
Macedonia
Smolyan
Turkey
Zlatograd
Greece
Bulgarian dialect divisions
Moesian dialects
Southwestern dialects
Balkan dialects
Northwestern dialects
Rupskian dialects
Transitional zones
yat borderline

The preparation stage

Preparation of the data

- the source (set of 36 words and 54 words)
- the digitization and conversion: **????** -> **Z@lt**

Reliability

- The correlation between distances based on the two sets is 0.97.
- The Cronbach’ s alpha is 0.84 for the 36-word set

Why then two sets?

The data

490 dialect sites within Bulgaria were chosen, and we included the Standard pronunciation (one third of the sites)

36 words, common for the 4 atlases and 54 words, common for the first two atlases

However (within the 54 set of words):

- For *North Greece*: 28 common words
- For *Serbia*: 18 common words

How to store the data?: *Linguistic adequacy*

The user decides on the level of detailness

- On one level:

  Phonetic: *palatal – semipalatal – nonpalatal (?-?')*

- On several levels:

  lemma, wordform, sense, site etc.

Mapping to some International Standard is required (for example, IPA)

How to store the data?: *software efficiency and portability*

Efficiency – we want to process the data easily. Therefore, we use simple encoding X-SAMPA which is convertible to the standard of IPA

Portability – we want to use other programs for modeling the data

Part of IPA encoding

THE INTERNATIONAL PHONETIC ALPHABET (revised to 1993)

CONSONANTS (PULMONIC)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap				ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

Some correspondences in X-SAMPA

Nasals

bilabial nasal m

labiodental nasal F

dental/alveolar nasal n

retroflex nasal n`

palatal nasal J

velar nasal N

uvular nasal N\

Efficiency

IPA (Unicode)	X-SAMPA (ASCII)
æ	{
?	E

Conversion

Via Cascaded Regular Grammars

From “ whatever working digitization” to SAMPA es

Left Regular Expression	Regular Expression	Right Regular Expression	Return Markup	Comments
	"л"		l	
	"м"		m	
	"н"		n	
	"п"		p	
	"р"		r	
	"с"		s	
	"т"		t	
	"ф"		f	
	"х"		x	
	"ц"		t_s	
	"ч"		t_S	
	"ш"		S	
	"ѕ"		d_z	
	"~"		d_Z	
	"й"		j	
	"ы"		l	
	"ь"		'	
	"я"		'a	
	"ю"		'y	
	"щ"		St	

ion

Context Check Order: ○ Left -> Right ◉ Right -> Left

iables

ame	Positive Values	Negative Values	Match
	[]	[]	Longest

Save Compile Exit

es

<table>
<tr><th>Left Regular Expression</th><th>Regular Expression</th><th>Right Regular Expression</th><th>Return Markup</th><th>Comments</th></tr>
<tr><td></td><td>"(","#",")"</td><td></td><td>\w_X</td><td></td></tr>
<tr><td></td><td>"e","-","."</td><td></td><td>8_:</td><td></td></tr>
<tr><td></td><td>"и",":"|"и",".","."</td><td></td><td>Y</td><td></td></tr>
<tr><td></td><td>".",".",".","e"|"e","/","...</td><td></td><td>3\</td><td></td></tr>
<tr><td></td><td>"e",".",".","."|"e",".","."</td><td></td><td>9</td><td></td></tr>
<tr><td></td><td>"a",".",".","."</td><td></td><td>A_0</td><td></td></tr>
<tr><td></td><td>"y",".","."</td><td></td><td>u_c</td><td></td></tr>
<tr><td></td><td>"a",".","."</td><td></td><td>A</td><td></td></tr>
<tr><td></td><td>".",".",".","л"</td><td></td><td>=</td><td></td></tr>
<tr><td></td><td>".",".",".","p"</td><td></td><td>r_=</td><td></td></tr>
<tr><td></td><td>".",".",".","l"</td><td></td><td>L_=</td><td></td></tr>
</table>

on

Context Check Order: ○ Left -> Right ◉ Right -> Left

ables

ame	Positive Values	Negative Values	Match
	[]	[]	Longest

Save **Compile** **Exit**

Portability

To structure the data in a database which is easy to explore for various purposes

XML

From content point of view the data can include the concept, the standard pronunciation, the base form, the recorded word form, the different pronunciations per site of the word form   we can remove the unnecessary information or transform, add, derive it in different formats

Targeted format

```
# zhylt
* 1
: Standard
- Z@lt
: Brezovo
- Z@lt
: Gradina
- Z@lt
: Seltsi
- Z@lt
```

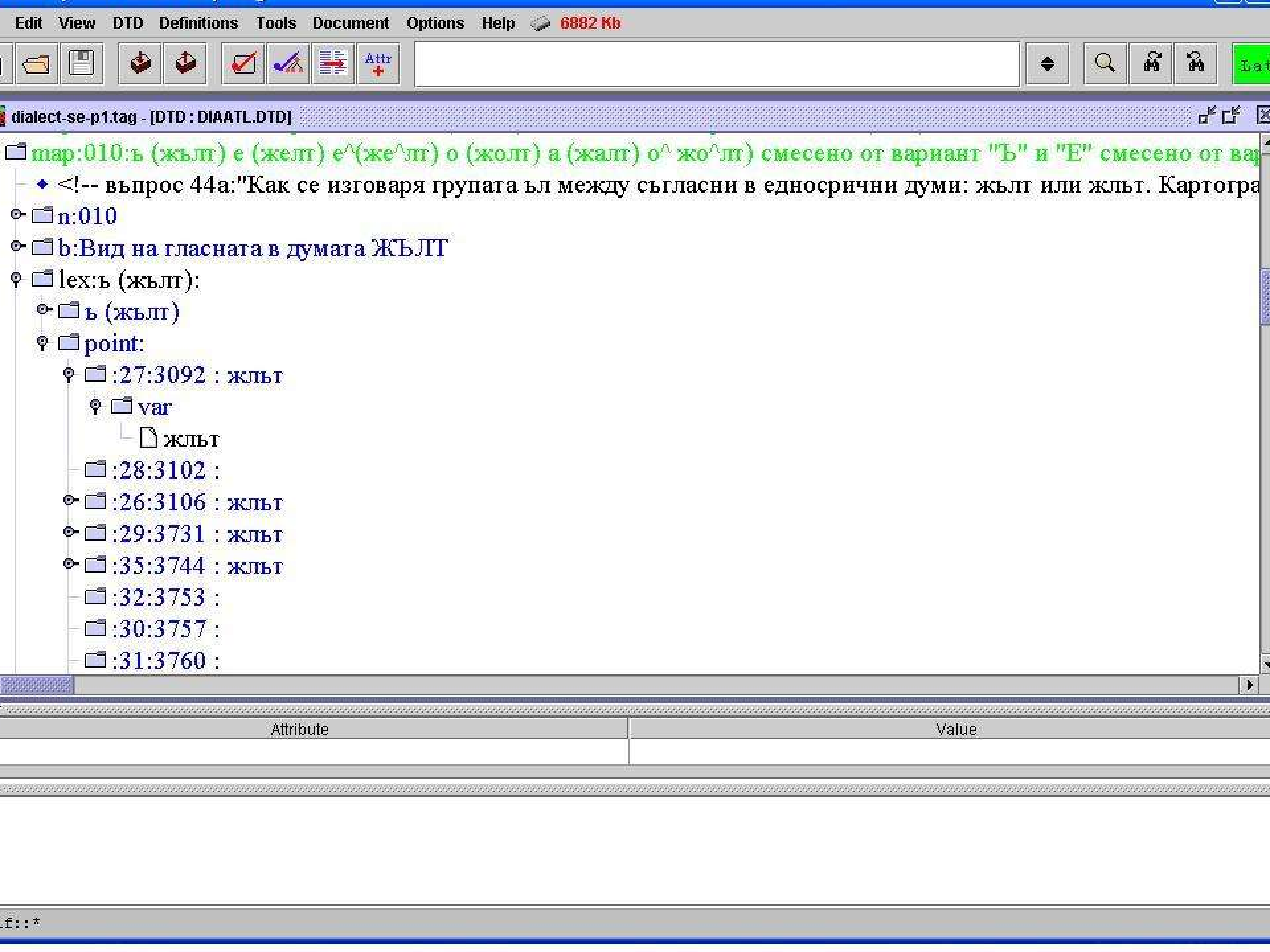
The structure of the digitized maps

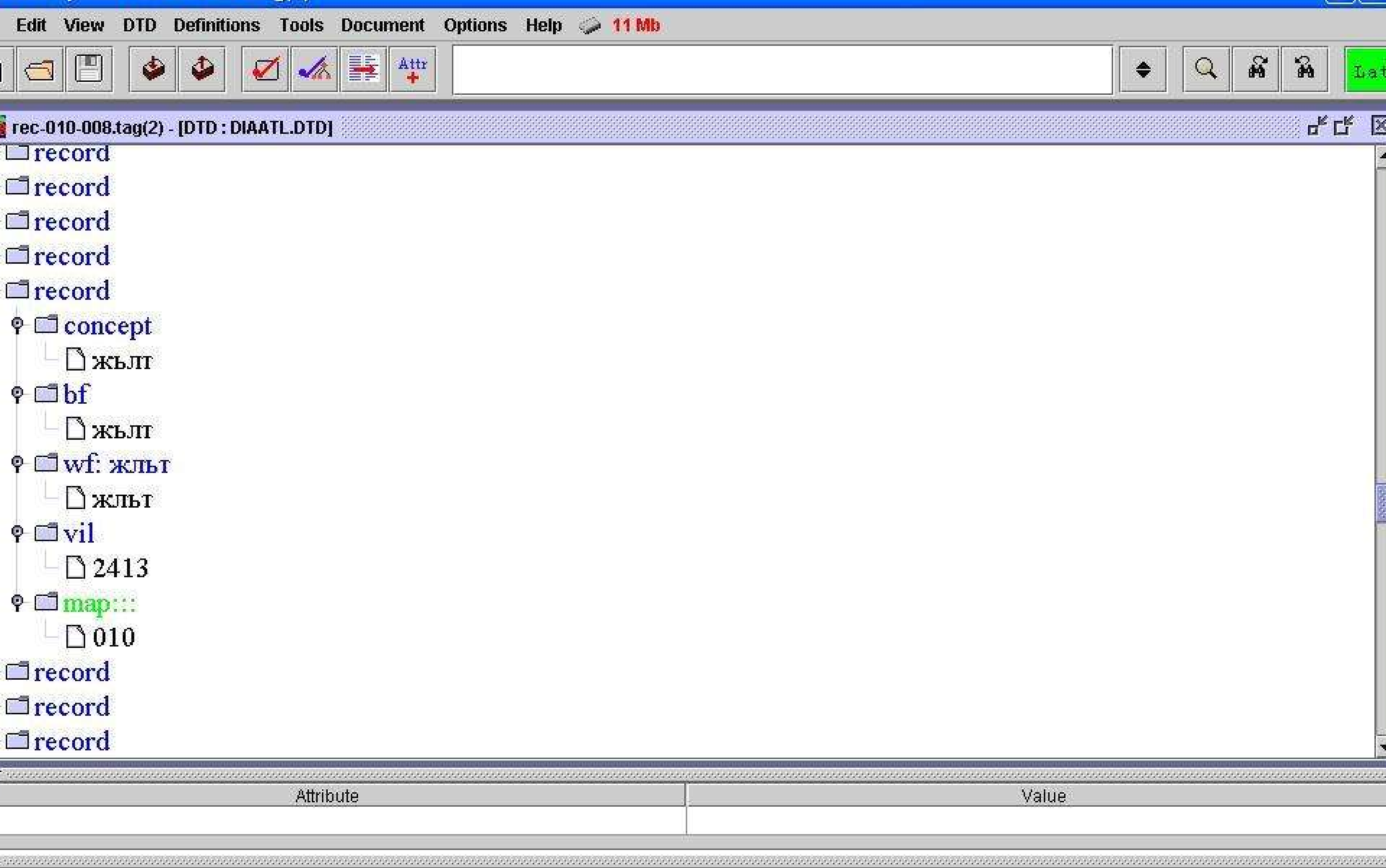
The maps include the following information:

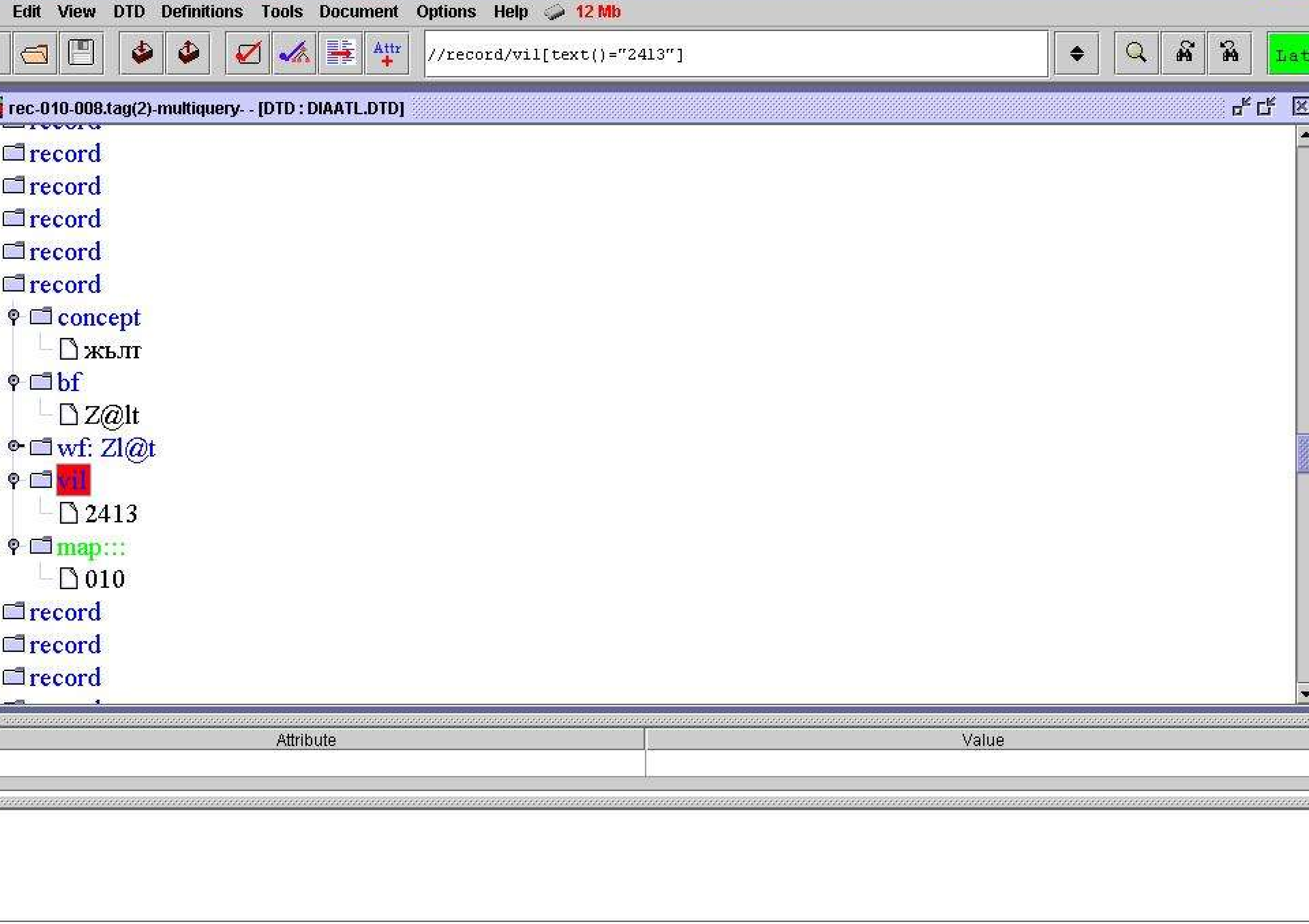
- The question
- The standard pronunciation
- All the pronunciations per site

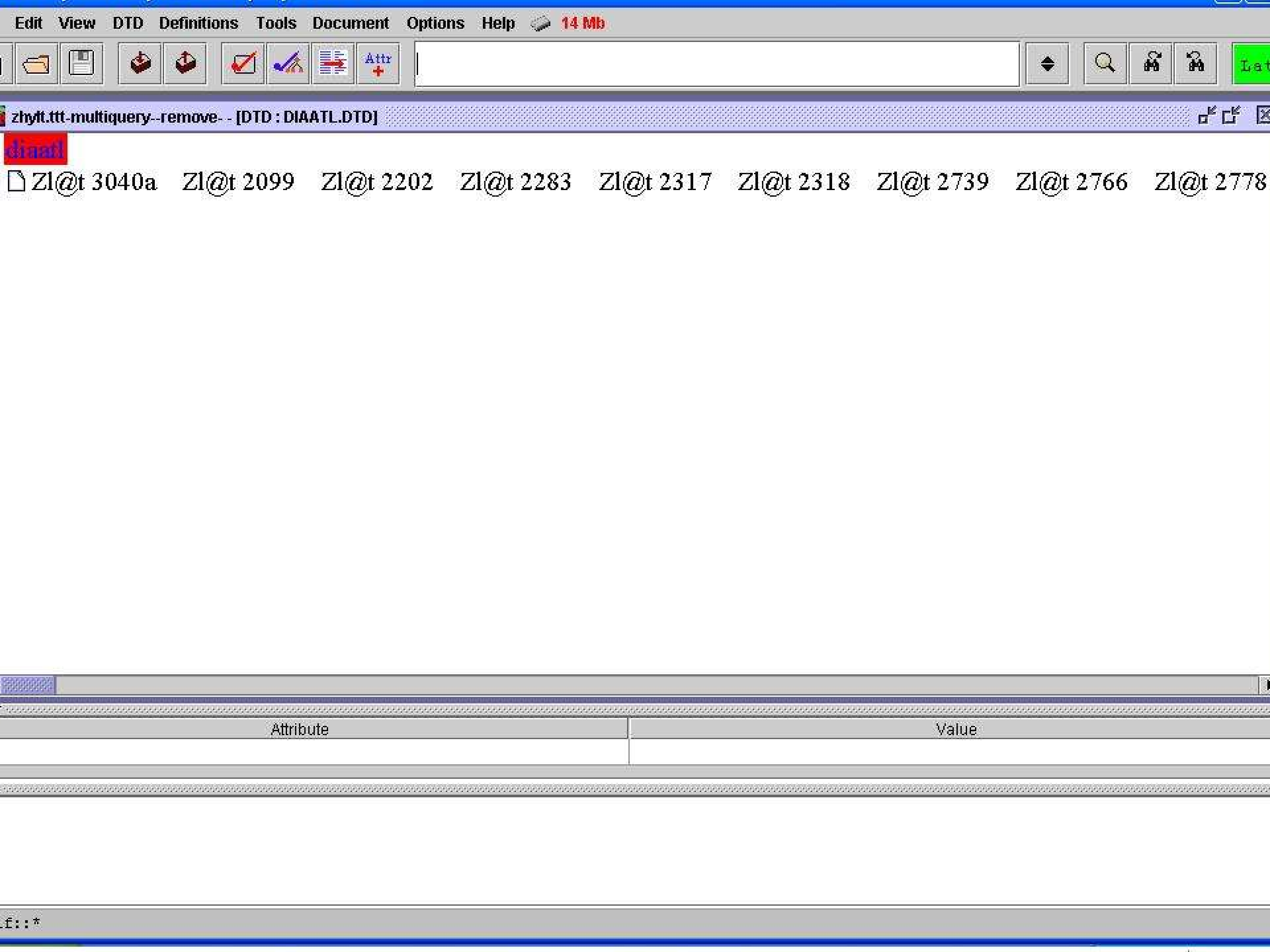
Additional information

- Concept - lexeme in the standard language
- Coordinates of the sites


Edit View DTD Definitions Tools Document Options Help 6882 Kb
Attr
dialect-se-p1.tag - [DTD : DIAATL.DTD]
map:010:ъ (жълт) е (желт) е^(же^лт) о (жолт) а (жалт) о^ жо^лт) смесено от вариант "Ъ" и "Е" смесено от ва
<!-- въпрос 44а:"Как се изговаря групата ъл между съгласни в едносрични думи: жълт или жлът. Картогра
n:010
b:Вид на гласната в думата ЖЪЛТ
lex:ъ (жълт):
ъ (жълт)
point:
:27:3092 : жлът
var
жлът
:28:3102 :
:26:3106 : жлът
:29:3731 : жлът
:35:3744 : жлът
:32:3753 :
:30:3757 :
:31:3760 :
Attribute
Value
.f::*


Edit View DTD Definitions Tools Document Options Help 11 Mb
Attr
rec-010-008.tag(2) - [DTD : DIAATL.DTD]
record
record
record
record
record
concept
жьлт
bf
жьлт
wf: жльт
жльт
vil
2413
map:::
010
record
record
record
Attribute
Value


Edit View DTD Definitions Tools Document Options Help 12 Mb
Attr
//record/vil[text()="2413"]
rec-010-008.tag(2)-multiquery- - [DTD : DIAATL.DTD]
record
record
record
record
record
concept
жълт
bf
Z@lt
wf: Zl@t
vil
2413
map:::
010
record
record
record
Attribute
Value


Edit View DTD Definitions Tools Document Options Help 14 Mb
Attr
zhyft.ttt-multiquery--remove- - [DTD : DIAATL.DTD]
diaatl
Zl@t 3040a Zl@t 2099 Zl@t 2202 Zl@t 2283 Zl@t 2317 Zl@t 2318 Zl@t 2739 Zl@t 2766 Zl@t 2778
Attribute
Value
.f::*

Carrying out the task

Measuring Levenshtein distances between each two pronunciations (line maps)

Clustering (Weighted Pair Group Method using Arithmetic averages (WPGMA) – explains 50.4 % of the Levenshtein distances) (dendograms, area maps, composite maps)

Multidimensional scaling (continuum maps)

Briefly About Levenshtein distance

In Levenshtein distance, two dialects are compared by comparing the pronunciation of words in the first dialect with the pronunciation of the same words in the second

Levenshtein distance may be approached by considering how one pronunciation may be transformed into the other by **inserting, deleting** or **substituting** sounds.

An example (1)

Tsreshn{j}a and Tcheresha

*Ts*reshna	subst. ts by tch	1
*Tch*reshna	insert e	1
Tcheresh*n*a	delete n	1

Tcheresha		3

An example (2)

Normalization: the sum of the operations is divided by the length of the longest alignment which gives the minimum cost:

Ts   0 r e S n a

Tsc   e r e S 0 a

----

1   1   1

3 (1+1+1) : 7 = 0.43

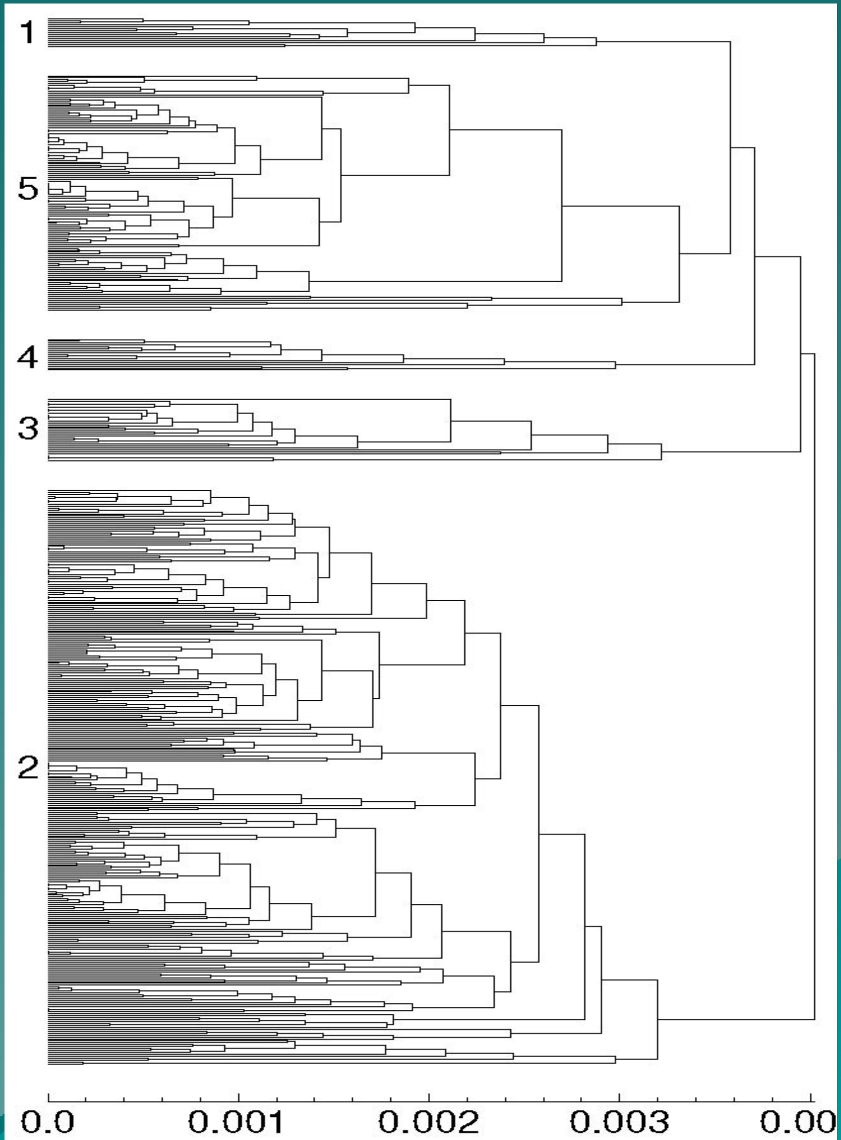
Clustering: Matrix updating algorithms

The way in which the distances between a newly formed cluster and the remaining points is calculated, is called MUA:

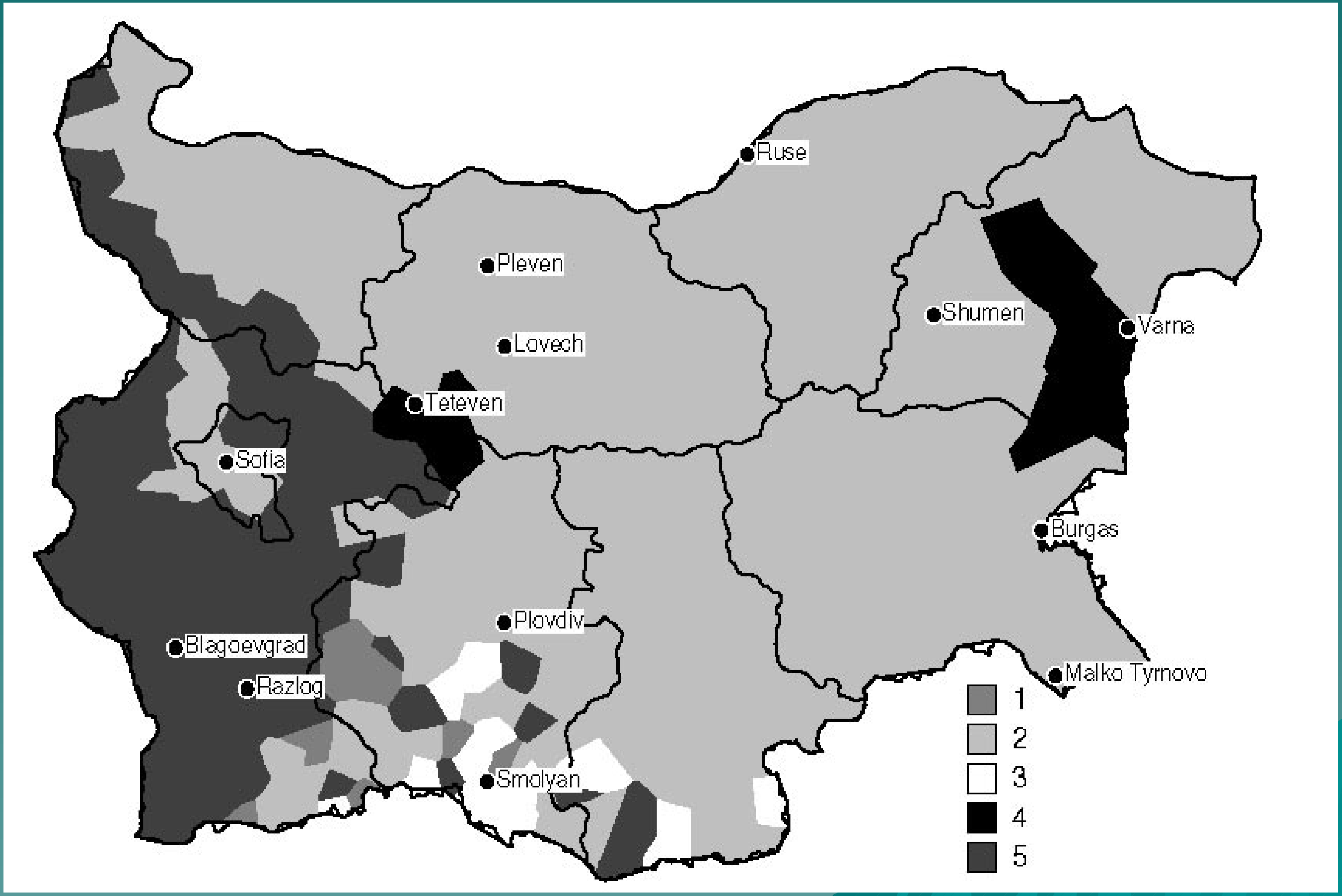
- Single link (nearest neighbor)
- Complete link (farthest neighbor)
- UPGMA (Unweighted Pair Group Method using Arithmetic averages)
- WPGMA (Weighted Pair Group Method using Arithmetic averages)
- UPMGC
- WPGMC
- Ward’ s method (minimum variance)

Clustering: dendogram

- Iteratively we select the shortest distance in the matrix of Levenshtei distances and we fuse the two data points
- Weighted Pair Group Method using Arithmetic averages (WPGMA)
- Unstable technique!!!

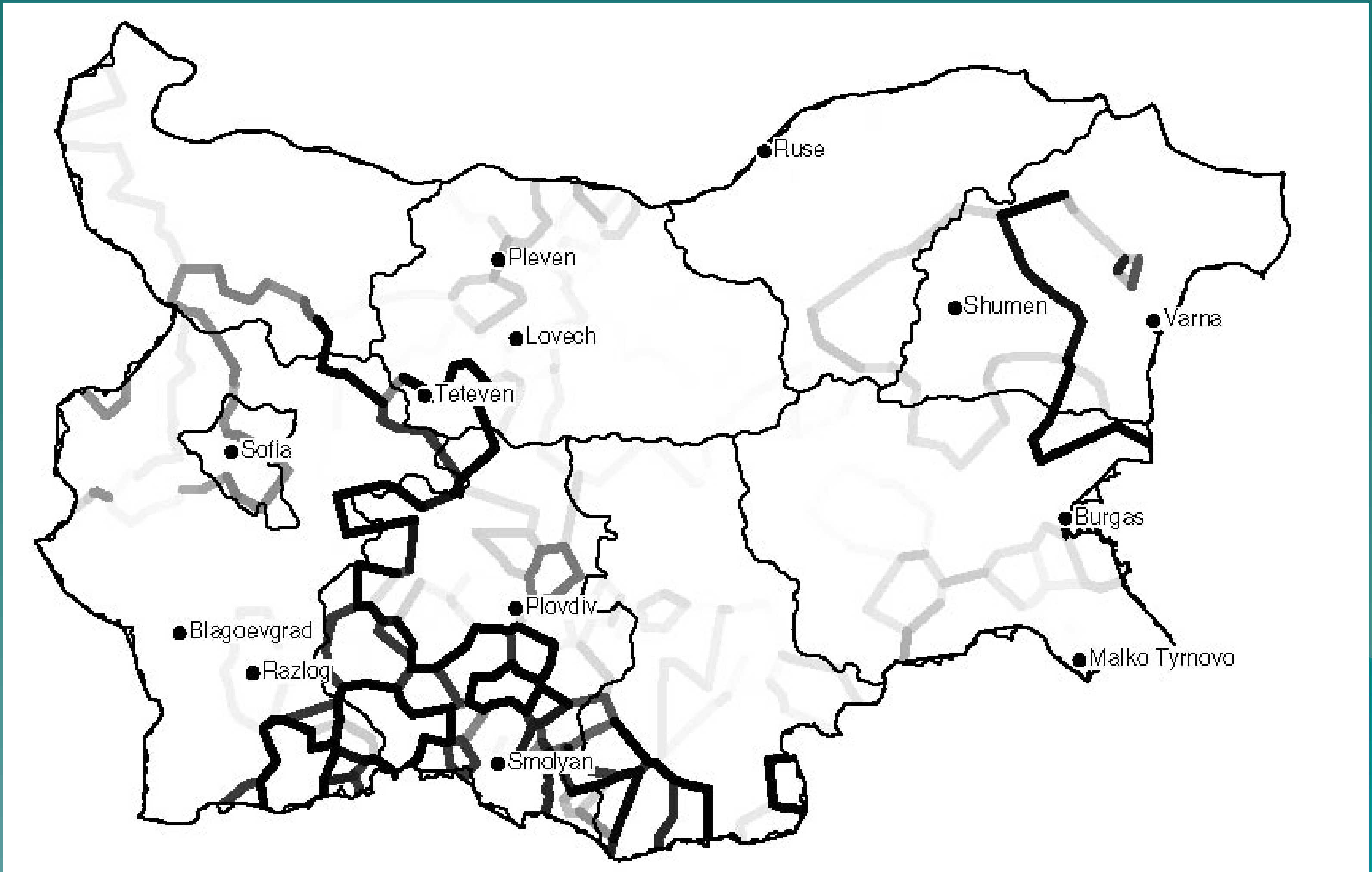


Area map



Ruse
Pleven
Shumen
Varna
Lovech
Teteven
Sofia
Burgas
Plovdiv
Blagoevgrad
Razlog
Malko Tyrnovo
Smolyan
1
2
3
4
5

Composite map



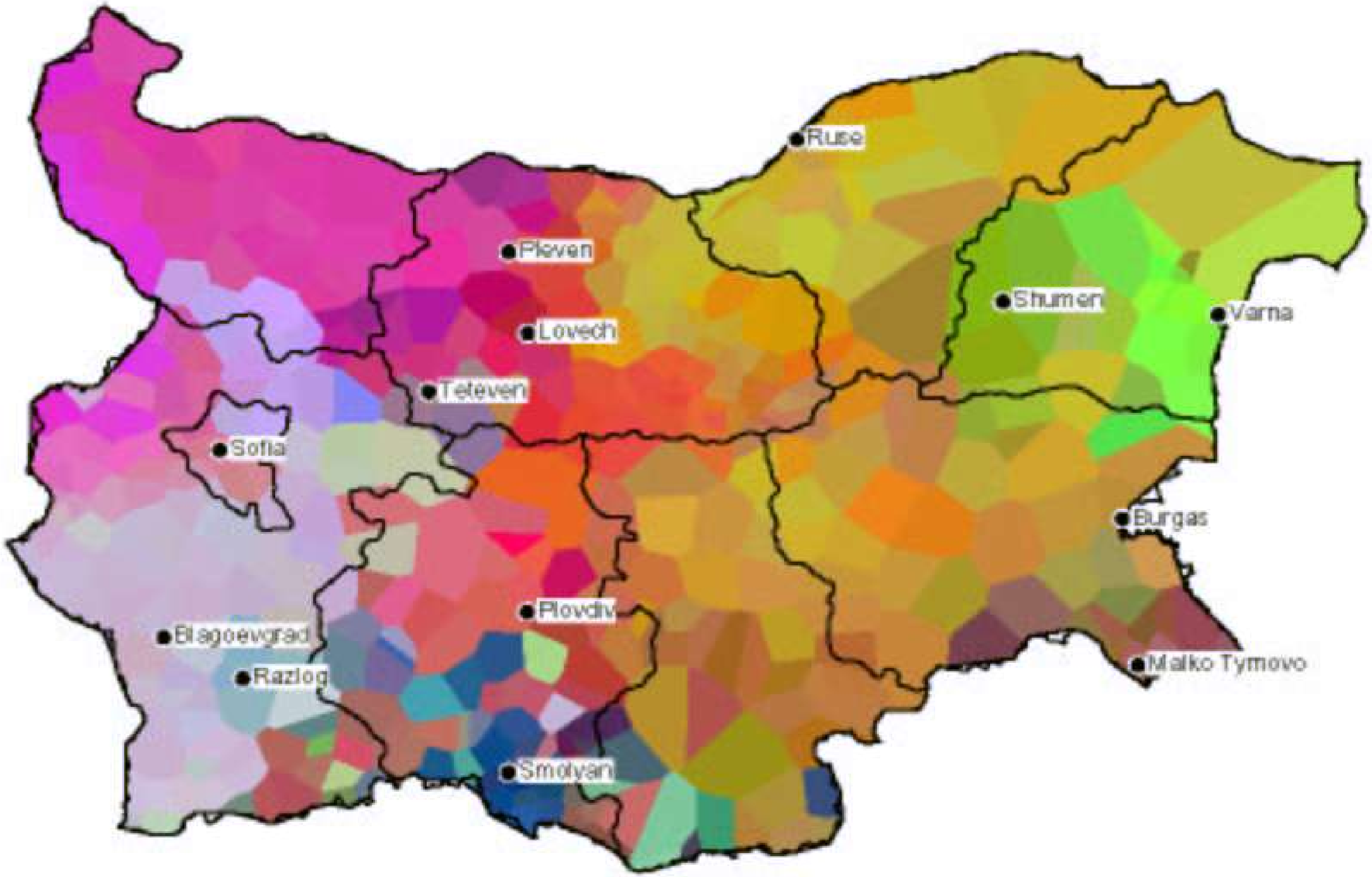
Multidimensional scaling

On the basis of mutual distances, an optimal coordinate system can be determined within which the data points may be located. The last is realized by a technique known as \`multidimensional scaling‘

On a multidimensional scaling plot, strongly related dialects are located close to each other, while very different dialects are located far away from each other (Krushkal and Wish 1984).


Duvanli (Pba)
Kaloyanovo (Pba)
Selishten dol (YuzP)
Dorkovo (Ro)
Dragoevo (M)
Breze (Ro)
Ribnovo (Zrup)
Vievo (Ro)
Kazichino (Ba)
Mugla (Ro)
Selcha (Ro)
Shiroka Lyka (Ro)
Mogilitsa (Ro)

Continuum map



The interpretation (some conclusions)

Comparison with traditional dialect maps showed that the most significant border remains the phonetic ‘ yat’ realization border

Western dialects are more coherent to each other than the Eastern ones are

The Rodopian region and the Northeastern Balkan region are distinct

The Standard is most similar to some South Balkan dialects from the Northeast

Validation

Comparison to the traditional maps and works

Estimating parameters of reliability

Experiments

Impact

- Bulgarian dialectology
- The development and testing of the technique and its implementation in a software package
- Future research intended in the area of language contact

Preliminary attempts on language contact: background

Hypothesis:

– Whether the similarity with the neighboring language is bigger near the border of the neighboring country

– Is it possible to test the hypothesis on a phonetic base?

The 36 words were translated into Standard Greek, Turkish, Romanian, Macedonian, Serbian

An example

17.pyt

: Greek

- "Dromos

: Macedonian

- pat

: Romanian

- drum

: Serbian

- put

: Turkish

- jol

Language contact: background (2)

Sparseness of data:

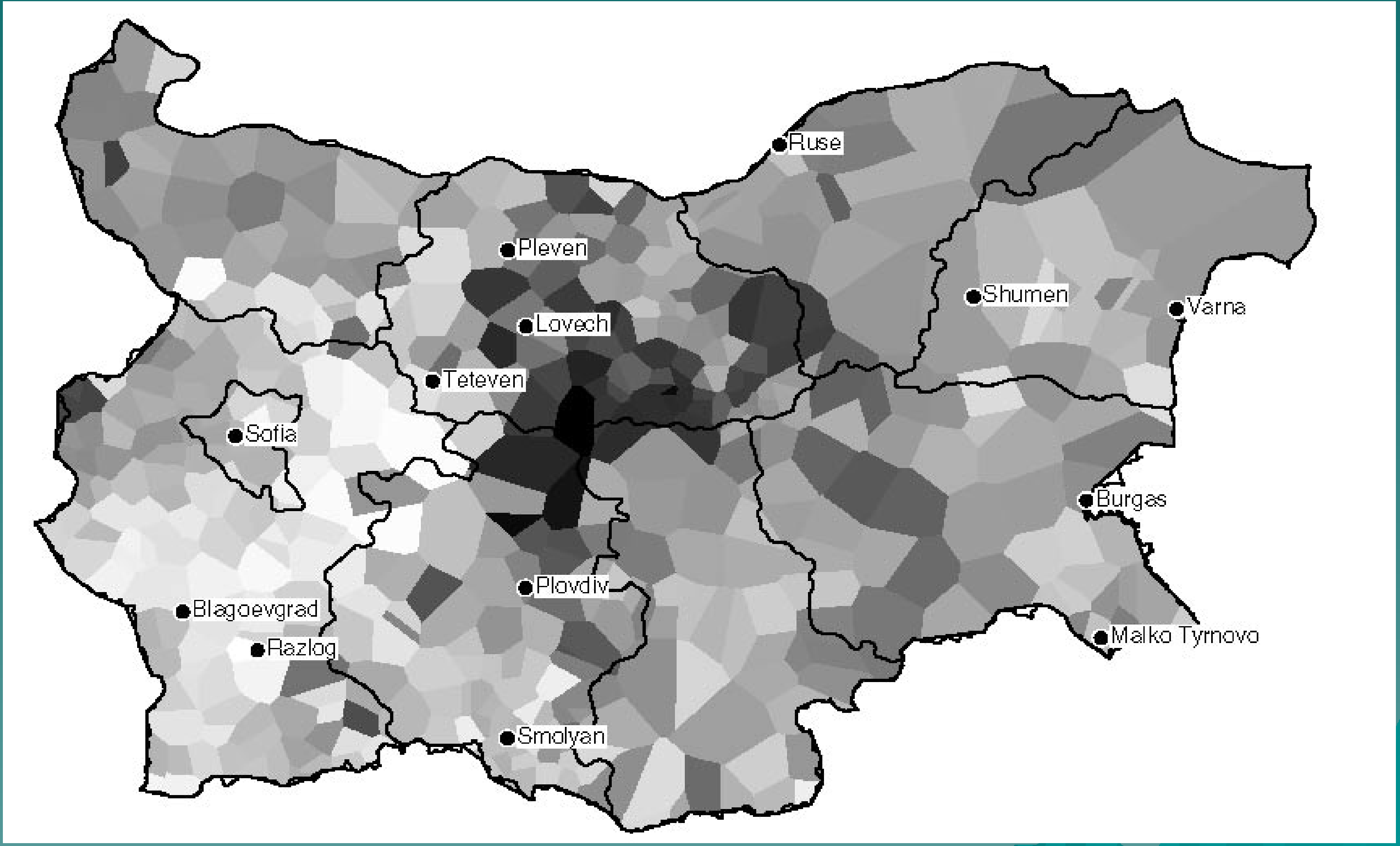
- The 36 words are basically with Slavic origin
- Only two loanwords are present: ???? (pocket) and ????????? (pot) -> both from Turkish
- Lexical variation is limited and not systematic -> only for 3 words

Corpus frequency method

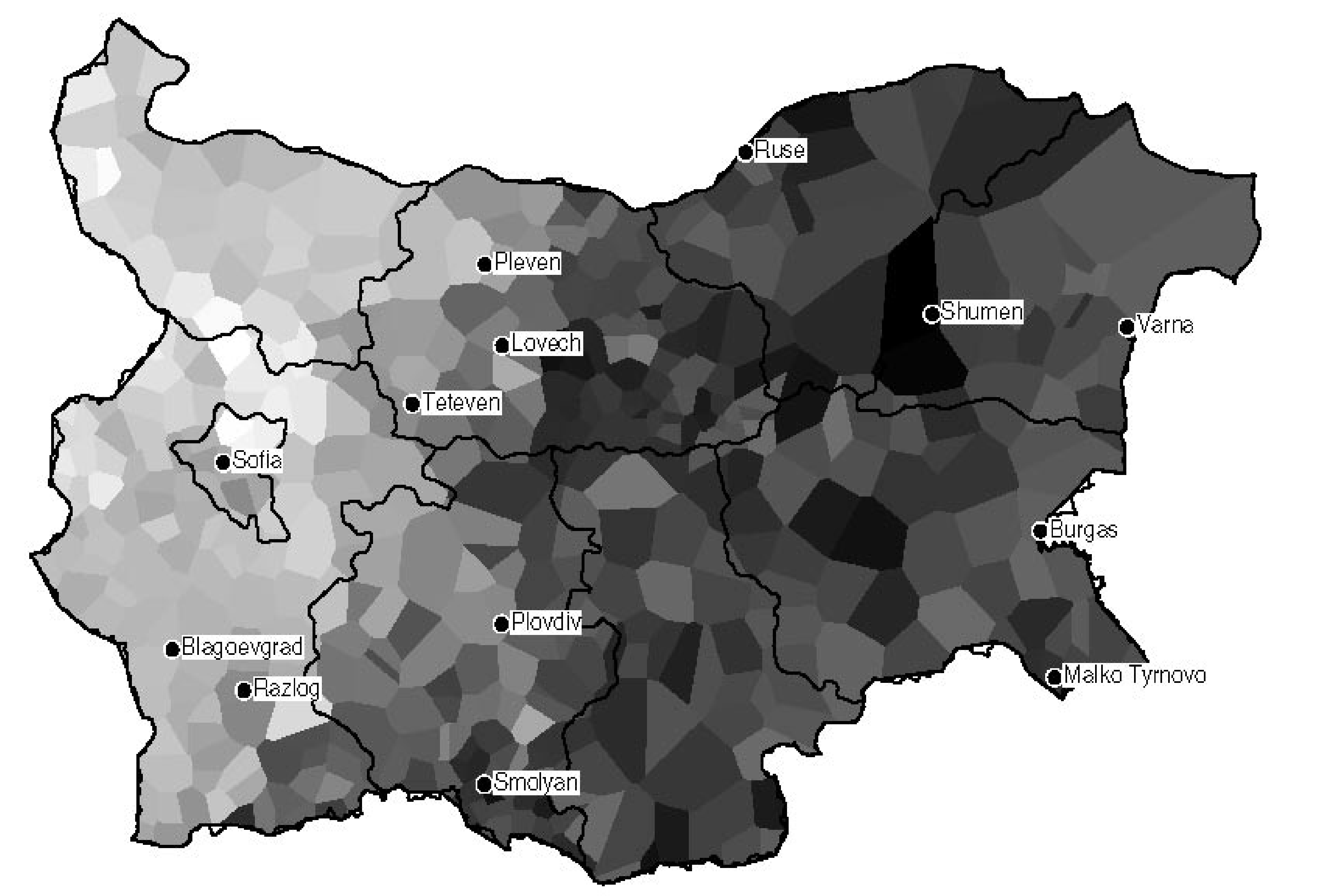
For the description of features we used Almeida and Brown system, which is based on the IPA system (Every language to Bulgarian). Two feature encodings:

- a scale-based encoding (suitable for comparing distances)
- a non-scale-based encoding (suitable for graphic presentation and interpretation)
- the correlation with the scale representation is very high - 9.2

The scale-based encoding: maps (Frequency)



The scale-based encoding: maps (Levenshtein)



Regression Analysis

The idea of the regression analysis in our case is to estimate the relation between the g*eographical* and *linguistic* distances

Our hypothesis is to check whether the greater geographical distance from the border causes greater linguistic distances

The procedure

50 Bulgarian dialect sites, evenly distributed throughout the country

The shortest distances to the five bordering countries were measured

Levenshtein tool was run on these 50 dialects plus the five standard languages

Regression analysis was performed with the help of Spss statistical package

Regression analysis

	Linear Corr	Linear sq Corr	Logarithmic sq Corr
Greek	-0.201	0.040	0.129
Romanian	0.431*	0.186	0.149
Turkish	-0.297*	0.088	0.024
Macedonian	0.651*	0.424	0.489
Serbian	0.763*	0.582	0.612

Conclusions

The data set should be expanded for better results in language contact

The lexical maps should be taken into consideration as well

To take into account that considering only Standard contact languages is very restrictive for our real purposes