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
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
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
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
### THE INTERNATIONAL PHONETIC ALPHABET (revised to 1993)

#### CONSONANTS (PULMONIC)

<table>
<thead>
<tr>
<th></th>
<th>Bilabial</th>
<th>Labiodental</th>
<th>Dental</th>
<th>Alveolar</th>
<th>Postalveolar</th>
<th>Retroflex</th>
<th>Palatal</th>
<th>Velar</th>
<th>Uvular</th>
<th>Pharyngeal</th>
<th>Glottal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plosive</td>
<td>p b</td>
<td></td>
<td></td>
<td>t d</td>
<td>t d c j</td>
<td>k g q g</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nasal</td>
<td>m m̃</td>
<td></td>
<td></td>
<td>n</td>
<td>η j η N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trill</td>
<td>B</td>
<td></td>
<td></td>
<td>r</td>
<td></td>
<td></td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tap or Flap</td>
<td>r</td>
<td></td>
<td></td>
<td>ɾ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fricative</td>
<td>φ θ</td>
<td>f v θ δ</td>
<td>s z</td>
<td>s z θ j</td>
<td>s z θ j x y</td>
<td>χ k h f</td>
<td>h h</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lateral fricative</td>
<td></td>
<td></td>
<td></td>
<td>l ɾ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approximant</td>
<td>v</td>
<td>ɾ j l j</td>
<td>m̃</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lateral approximant</td>
<td>l</td>
<td></td>
<td></td>
<td>l ɻ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 $N\backslash$
### Efficiency

<table>
<thead>
<tr>
<th>IPA (Unicode)</th>
<th>X-SAMPA (ASCII)</th>
</tr>
</thead>
<tbody>
<tr>
<td>æ</td>
<td>{</td>
</tr>
<tr>
<td>?</td>
<td>E</td>
</tr>
</tbody>
</table>
Conversion

Via Cascaded Regular Grammars

From “whatever working digitization” to SAMPA
<table>
<thead>
<tr>
<th>Left Regular Expression</th>
<th>Regular Expression</th>
<th>Right Regular Expression</th>
<th>Return Markup</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;й&quot;</td>
<td></td>
<td>&quot;м&quot;</td>
<td>n</td>
<td></td>
</tr>
<tr>
<td>&quot;п&quot;</td>
<td></td>
<td>&quot;д&quot;</td>
<td>r</td>
<td></td>
</tr>
<tr>
<td>&quot;р&quot;</td>
<td></td>
<td>&quot;т&quot;</td>
<td>s</td>
<td></td>
</tr>
<tr>
<td>&quot;с&quot;</td>
<td></td>
<td>&quot;с&quot;</td>
<td>t</td>
<td></td>
</tr>
<tr>
<td>&quot;ч&quot;</td>
<td></td>
<td>&quot;ч&quot;</td>
<td>f</td>
<td></td>
</tr>
<tr>
<td>&quot;х&quot;</td>
<td></td>
<td>&quot;х&quot;</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>&quot;ц&quot;</td>
<td></td>
<td>&quot;ц&quot;</td>
<td>t_s</td>
<td></td>
</tr>
<tr>
<td>&quot;ш&quot;</td>
<td></td>
<td>&quot;ш&quot;</td>
<td>t_s</td>
<td></td>
</tr>
<tr>
<td>&quot;щ&quot;</td>
<td></td>
<td>&quot;щ&quot;</td>
<td>s</td>
<td></td>
</tr>
<tr>
<td>&quot;з&quot;</td>
<td></td>
<td>&quot;з&quot;</td>
<td>d_z</td>
<td></td>
</tr>
<tr>
<td>&quot;э&quot;</td>
<td></td>
<td>&quot;э&quot;</td>
<td>d_z</td>
<td></td>
</tr>
<tr>
<td>&quot;и&quot;</td>
<td></td>
<td>&quot;й&quot;</td>
<td>j</td>
<td></td>
</tr>
<tr>
<td>&quot;ь&quot;</td>
<td></td>
<td>&quot;ь&quot;</td>
<td>l</td>
<td></td>
</tr>
<tr>
<td>&quot;а&quot;</td>
<td></td>
<td>&quot;а&quot;</td>
<td>'a'</td>
<td></td>
</tr>
<tr>
<td>&quot;ъ&quot;</td>
<td></td>
<td>&quot;ъ&quot;</td>
<td>'y'</td>
<td></td>
</tr>
<tr>
<td>&quot;ш&quot;</td>
<td></td>
<td>&quot;ш&quot;</td>
<td>St</td>
<td></td>
</tr>
<tr>
<td>Left Regular Expression</td>
<td>Regular Expression</td>
<td>Right Regular Expression</td>
<td>Return Markup</td>
<td>Comments</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-------------------</td>
<td>--------------------------</td>
<td>---------------</td>
<td>----------</td>
</tr>
<tr>
<td>&quot; ( &quot; # &quot; )&quot;</td>
<td></td>
<td></td>
<td>\w_X</td>
<td></td>
</tr>
<tr>
<td>&quot;e&quot; , &quot;-e&quot; , &quot;e&quot;</td>
<td></td>
<td></td>
<td>_ 8</td>
<td></td>
</tr>
<tr>
<td>&quot;u&quot; , &quot;</td>
<td>u&quot; , &quot;</td>
<td>u&quot; , &quot;</td>
<td>u&quot;</td>
<td></td>
</tr>
<tr>
<td>&quot;e&quot; , &quot;e&quot; , &quot;e&quot;</td>
<td></td>
<td></td>
<td>3 \</td>
<td></td>
</tr>
<tr>
<td>&quot;a&quot; , &quot;a&quot; , &quot;a&quot;</td>
<td></td>
<td></td>
<td>A 0</td>
<td></td>
</tr>
<tr>
<td>&quot;y&quot; , &quot;y&quot; , &quot;y&quot;</td>
<td></td>
<td></td>
<td>u_c</td>
<td></td>
</tr>
<tr>
<td>&quot;a&quot; , &quot;a&quot; , &quot;a&quot;</td>
<td></td>
<td></td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>&quot;r&quot; , &quot;r&quot; , &quot;r&quot;</td>
<td></td>
<td></td>
<td>r_</td>
<td></td>
</tr>
<tr>
<td>&quot;r&quot; , &quot;r&quot; , &quot;r&quot;</td>
<td></td>
<td></td>
<td>L_</td>
<td></td>
</tr>
</tbody>
</table>
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
Въпрос 44a: Как се изговаря групата ъъл между съгласни в едносрични думи: жълт или жълт. Картография на гласната в думата ЖЪЛТ.
<table>
<thead>
<tr>
<th>Attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zl@t 3040a</td>
<td>2</td>
</tr>
<tr>
<td>Zl@t 2202</td>
<td>2</td>
</tr>
<tr>
<td>Zl@t 2317</td>
<td>2</td>
</tr>
<tr>
<td>Zl@t 2739</td>
<td>2</td>
</tr>
<tr>
<td>Zl@t 2778</td>
<td>2</td>
</tr>
</tbody>
</table>
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

\textit{Tsreshna} subst. \textit{ts} by \textit{tch} 1
\textit{Tchreshna} insert \textit{e} 1
\textit{Tchereshna} delete \textit{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:

\[
\begin{array}{cccc}
T_s & 0 & r & e & S \ n \ a \\
T_{sc} & e & r & e & S \ 0 \ a \\
\end{array}
\]

-----------------------------

\[
\begin{array}{ccc}
1 & 1 & 1 \\
3 \ (1+1+1) : 7 = 0.43 \\
\end{array}
\]
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 Levenshtein distances and we fuse the two data points.

Weighted Pair Group Method using Arithmetic averages (WPGMA)

Unstable technique!!!
Area map
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).
Multidimensional scaling
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 geographical 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

<table>
<thead>
<tr>
<th>Language</th>
<th>Linear Corr</th>
<th>Linear sq Corr</th>
<th>Logarithmic sq Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greek</td>
<td>-0.201</td>
<td>0.040</td>
<td>0.129</td>
</tr>
<tr>
<td>Romanian</td>
<td>0.431*</td>
<td>0.186</td>
<td>0.149</td>
</tr>
<tr>
<td>Turkish</td>
<td>-0.297*</td>
<td>0.088</td>
<td>0.024</td>
</tr>
<tr>
<td>Macedonian</td>
<td>0.651*</td>
<td>0.424</td>
<td>0.489</td>
</tr>
<tr>
<td>Serbian</td>
<td>0.763*</td>
<td>0.582</td>
<td>0.612</td>
</tr>
</tbody>
</table>
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.