Application of Phylogenetic Methods on the Dialect Pronunciation Data

Jelena Prokić
University of Groningen
The Netherlands

Tübingen, 23 June 2007

Overview

- Introduction
- Phylogenetic methods
- Linguistic data
- Results
- Conclusions

Introduction

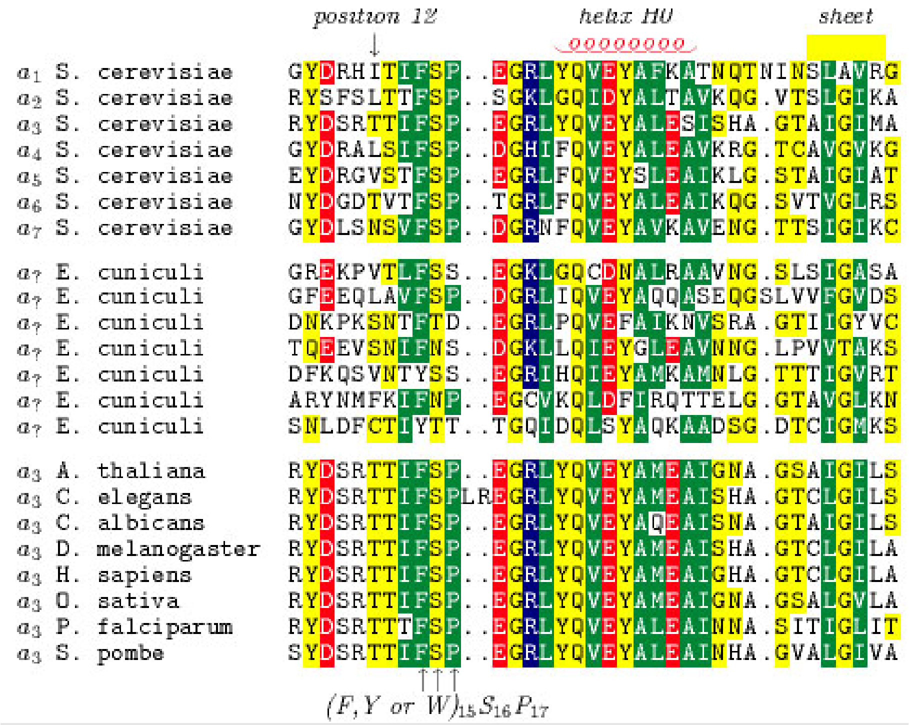
- Can methods from phylogenetics be applied to linguistic data?
    - What is the similarity between biological and linguistic data?
- What can phylogenetic analyses tell us about language change?

Phylogenetics

- Study of evolutionary relatedness among various groups of organisms

- Molecular phylogeny
  - the use of the structure of molecules to gain information
  - analysis of aligned nucleotide or amino acid sequences
  - closely related organisms have similar molecular structure

A Sequence Alignment



a_1	S. cerevisiae	GYDRHITIFSP..EGRLYQVEYAFKATNQTNINSLAVRG
a_2	S. cerevisiae	RYSFSLTTFSP..SGKLGQIDYALTAVKQG.VTSLGIKA
a_3	S. cerevisiae	RYDSRTTIFSP..EGRLYQVEYALESISHA.GTAIGIMA
a_4	S. cerevisiae	GYDRALSIFSP..DGHIFQVEYALEAVKRG.TCAVGVKG
a_5	S. cerevisiae	EYDRGVSTFSP..EGRLFQVEYSLEAIKLG.STAIGIAT
a_6	S. cerevisiae	NYDGDTVTFSP..TGRLFQVEYALEAIKQG.SVTVGLRS
a_7	S. cerevisiae	GYDLSNSVFSP..DGRNFQVEYAVKAVENG.TTSIGIKC
$a_?$	E. cuniculi	GREKPVTLFSS..EGKLGQCDNALRAAVNG.SLSIGASA
$a_?$	E. cuniculi	GFEEQLAVFSP..DGRLIQVEYAQQASEQGSLVVFGVDS
$a_?$	E. cuniculi	DNKPKSNTFTD..EGRLPQVEFAIKNVSRA.GTIIGYVC
$a_?$	E. cuniculi	TQEEVSNIFNS..DGKLLQIEYGLEAVNNG.LPVVTAKS
$a_?$	E. cuniculi	DFKQSVNTYSS..EGRIHQIEYAMKAMNLG.TTTIGVRT
$a_?$	E. cuniculi	ARYNMFKIFNP..EGCVKQLDFIRQTTELG.GTAVGLKN
$a_?$	E. cuniculi	SNLDFCTIYTT..TGQIDQLSYAQKAADSG.DTCIGMKS
a_3	A. thaliana	RYDSRTTIFSP..EGRLYQVEYAMEAIGNA.GSAIGILS
a_3	C. elegans	RYDSRTTIFSPLREGRLYQVEYAMEAISHA.GTCLGILS
a_3	C. albicans	RYDSRTTIFSP..EGRLYQVEYAQEAISNA.GTAIGILS
a_3	D. melanogaster	RYDSRTTIFSP..EGRLYQVEYAMEAISHA.GTCLGILA
a_3	H. sapiens	RYDSRTTIFSP..EGRLYQVEYAMEAIGHA.GTCLGILA
a_3	O. sativa	RYDSRTTIFSP..EGRLYQVEYAMEAIGNA.GSALGVLA
a_3	P. falciparum	RYDSRTTTFSP..EGRLYQVEYALEAINNA.SITIGLIT
a_3	S. pombe	SYDSRTTIFSP..EGRLYQVEYALEAINHA.GVALGIVA

$(F, Y \text{ or } W)_{15} S_{16} P_{17}$

Figure 1: Multiple sequence alignment

Methods

- Distance-based
  - UPGMA (Unweighted Pair Group Method with Arithmetic mean)
  - Neighbor Joining
  - Neighbor Net
- Character-based
  - Maximum Parsimony
  - Maximum Likelihood
  - Bayesian Inference

Distance-based Methods

- Step 1: estimate pairwise distances

A	ATTTGCGGTA
B	ATCTGCGATA
C	ATTGCCGTTT
D	TTCGCTGTTT

	A	B	C	D
A	0.0	0.2	0.4	0.7
B		0.0	0.5	0.6
C			0.0	0.3
D				0.0

- Step 2: apply one of the distance-based methods

Example 1

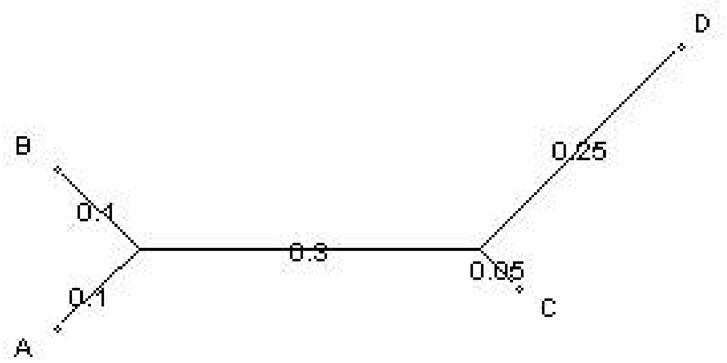


Figure 2: Neighbor-joining tree

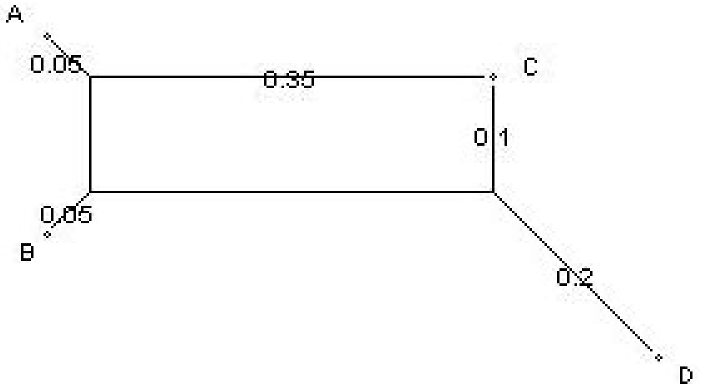


Figure 3: Neighbor net

Character-based Methods

- Carry out calculations on each of the individual residues of the sequences
- Individual variations between sequences not reduced to a single value
- Infer the phylogeny based on all the individual characters

Example 2

A	ATTTGCGGTA
B	ATCTGCGATA
C	ATTGCCGTTT
D	TTCGCTGTTT

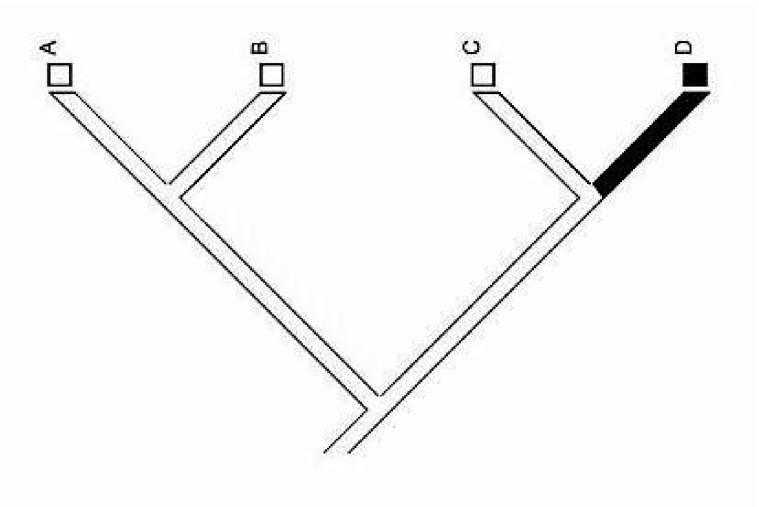


Figure 4: Maximum parsimony tree

Example 3

A	ATTTGCGGTA
B	ATCTGCGATA
C	ATTGCCGTTT
D	TTCGCTGTTT

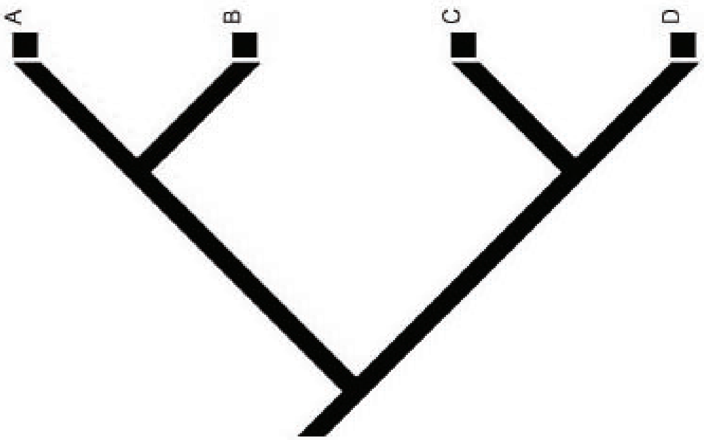


Figure 5: Maximum parsimony tree

Example 4

A	ATTTGCGGTA
B	ATCTGCGATA
C	ATTGCCGTTT
D	TTCGCTGTTT

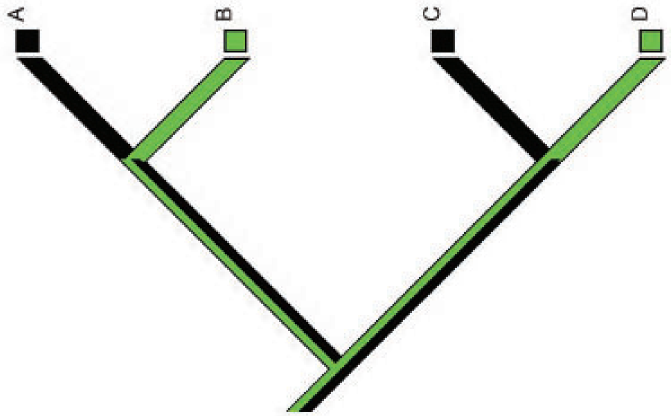


Figure 6: Maximum parsimony tree

Similarity with Linguistic Data

- Languages evolve in similar ways to biological species
  - split into new languages, mutate
- Languages, like molecules, document evolutionary history
- Language family trees
  - from lexical, morphological and phonological data

Multiple Sequence Alignments

- Pairwise sequence alignments using Levensthein

b	e	l	i
b	ɛ	l	i

b	e	l	i
b^j	ɑ	l	i

- Extracted multiple sequence alignments

b	e	l	i
b	ɛ	l	i
b^j	ɑ	l	i

Distance-based Methods

- Transcriptions of 113 different words merged into 1 string

- Distance between 2 sites - number of different positions divided by the length of the string

j	ɑ	g	n	e	j	ɑ	-	b	e	l	i
j	ɑ	g	n	e	j	ɑ	-	b	ɛ	l	i
-	ɑ	g	n	e	-	ɑ	s	b^j	ɑ	l	i

Neighbor Joining

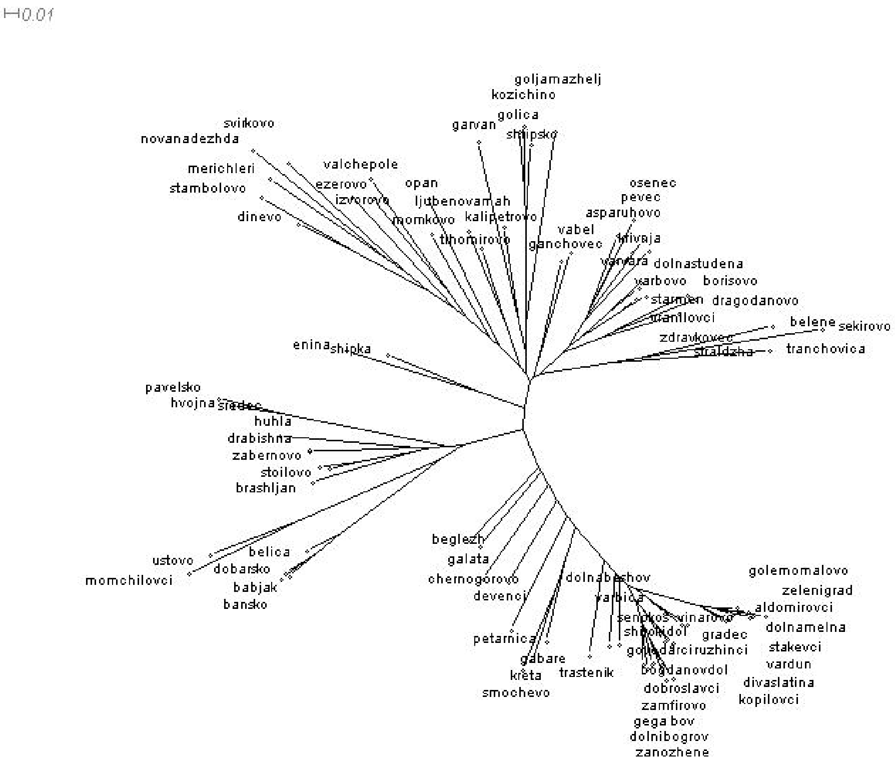


Figure 7: Neighbor Joining Tree

NJ - Dialect Divisions

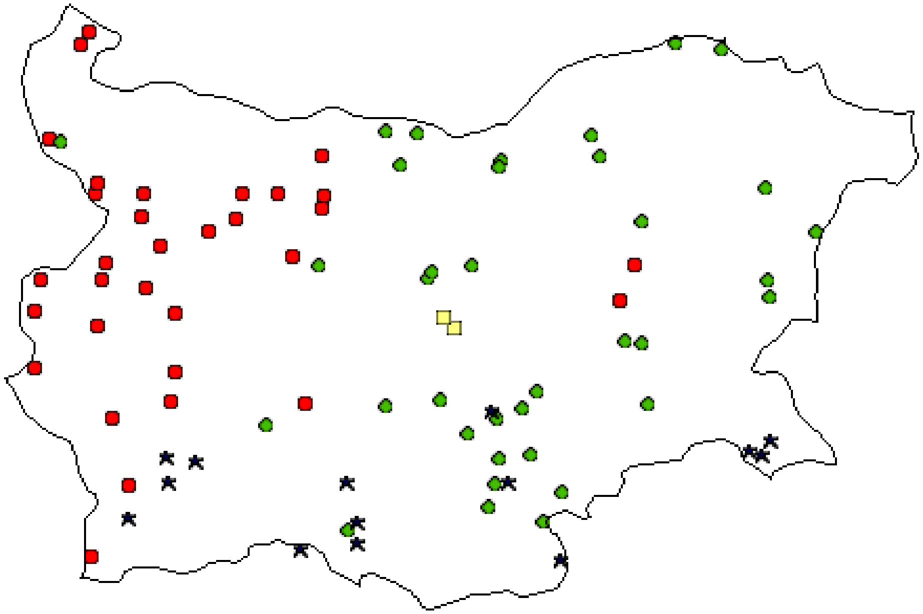


Figure 8: Four dialect areas

Neighbor Net

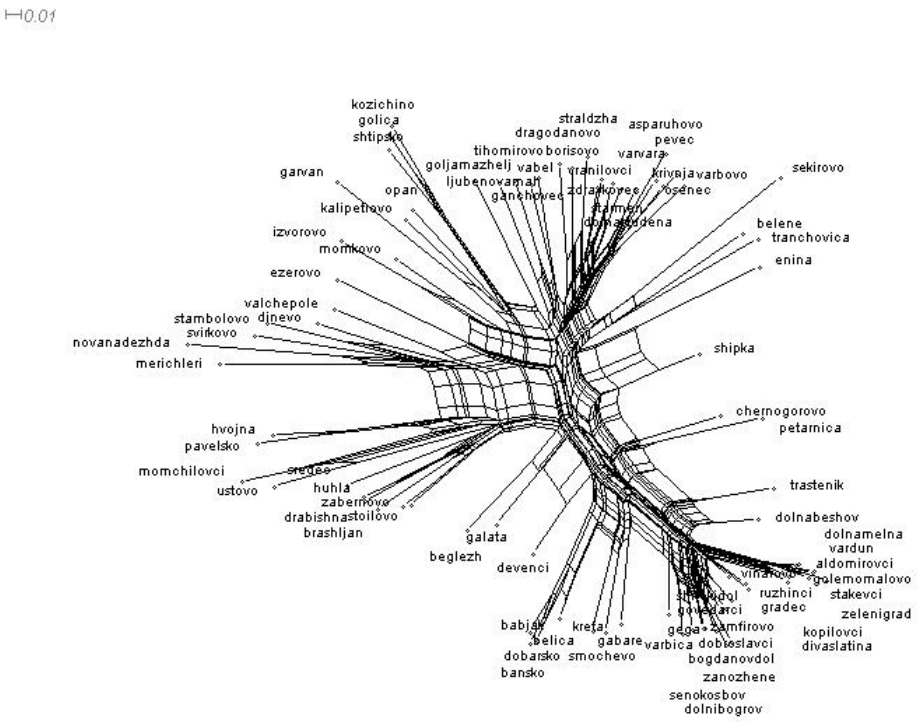


Figure 9: Neighbor Net Tree

NN - Dialect Divisions

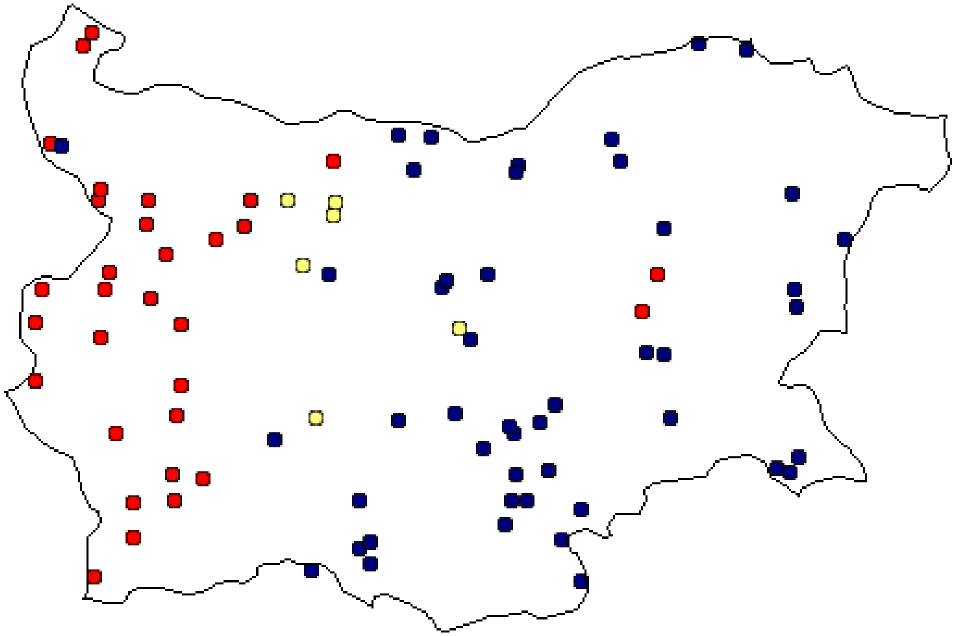


Figure 10: Three dialect areas

Bootstrapping

- Statistical method for estimating the sampling distribution of an estimator by sampling with replacement from the original sample

- Creates a new data set by sampling N characters randomly with replacement

- Resulting data set has the same size as the original

- Statistically bootstrapped data sets contain variation that you would get from collecting new data sets

- Bootstrap values are a measure of support of a given edge

NJ - Bootstrapped Tree

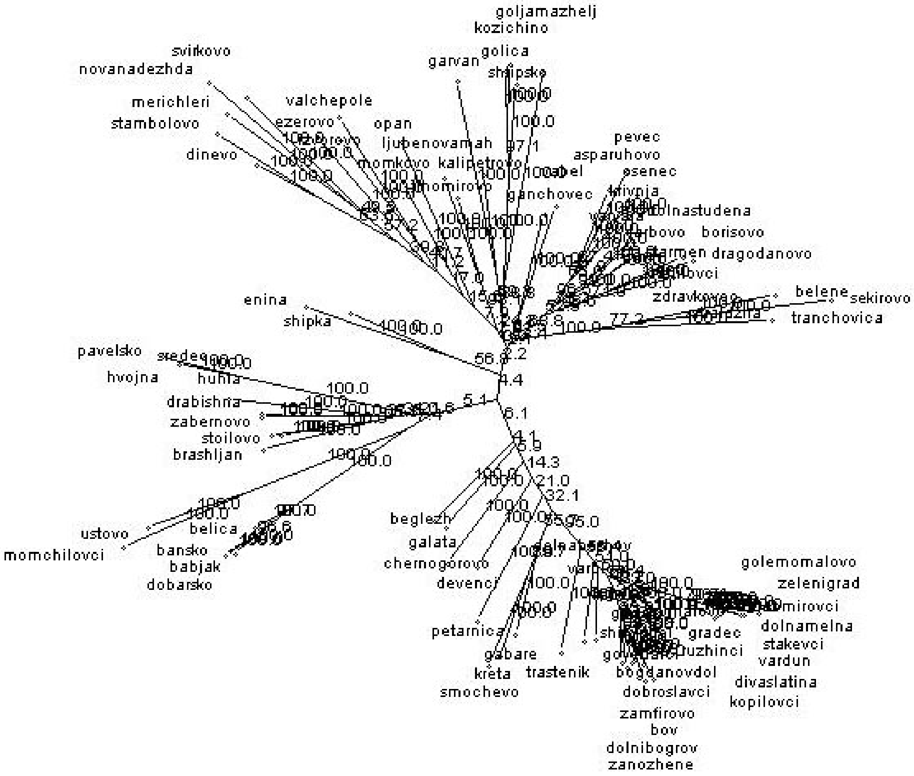


Figure 11: Bootstrapped tree

Character-based Method

- Transcriptions of 20 different words merged into 1 string
- Maximum parsimony method - method for phylogenetic inference
- Infers a phylogenetic tree by minimizing the total number of evolutionary steps required to explain a given set of data
- Cladistic method - studies the pathways of evolution
- Each character serves as an independent hypothesis of evolution

Example 5

j	ɑ	g	n	e	b	e	l	i	v	e	t	ɣ	r	g	o	v	e	d	o
-	ɑ	g	n	i	b	e	l	i	v^j	ɑ	t	ɣ	r	g	u	v	e	d	u
j	ɑ	g	n	e	b	ɛ	l	i	v	ɛ	t	e	r	g	ʊ	v	e	d	u
j	ɑ	g	n	e	b	ɛ	l	i	v	ɛ	t	e	r	g	u	v	e	d	u
-	ɑ	g	n	e	b^j	ɑ	l	i	v^j	ɑ	t	ɣ	r	g	u	v	e	d	u
j	ɑ	g	n	e	b	e	l	i	v	e	t	ɣ	r	g	u	v	i	d	u
j	ɑ	g	n	e	b	e	l	i	v	e	t	e	r	g	u	v	e	d	u

Figure 12: Each colored position reflects unique linguistic phenomenon

Maximum Parsimony I

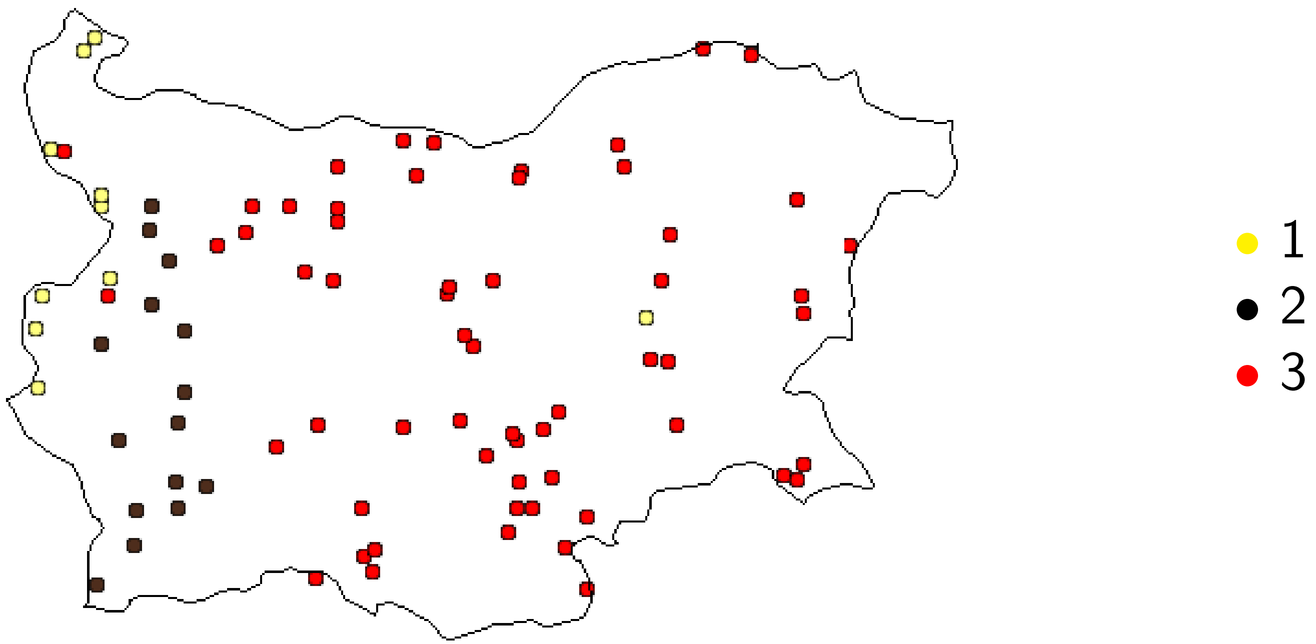


Figure 13: Three dialect areas

Maximum Parsimony II

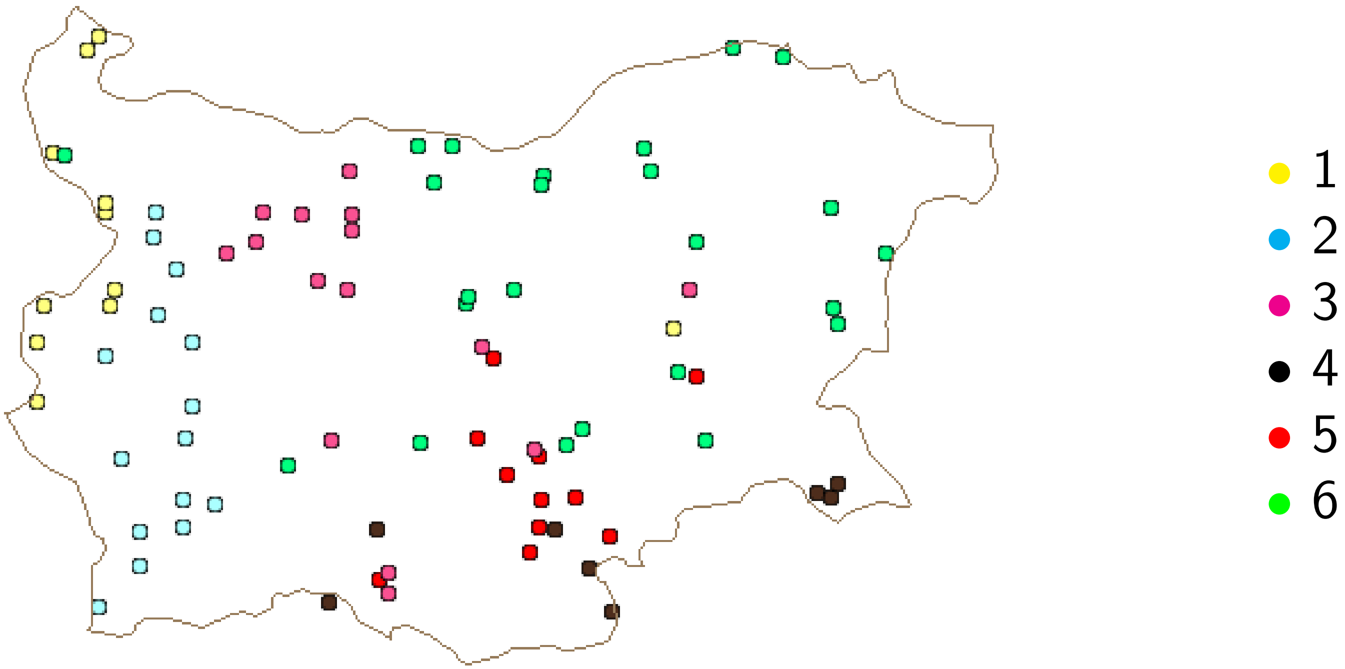


Figure 14: Six dialect areas

v - v^j opposition

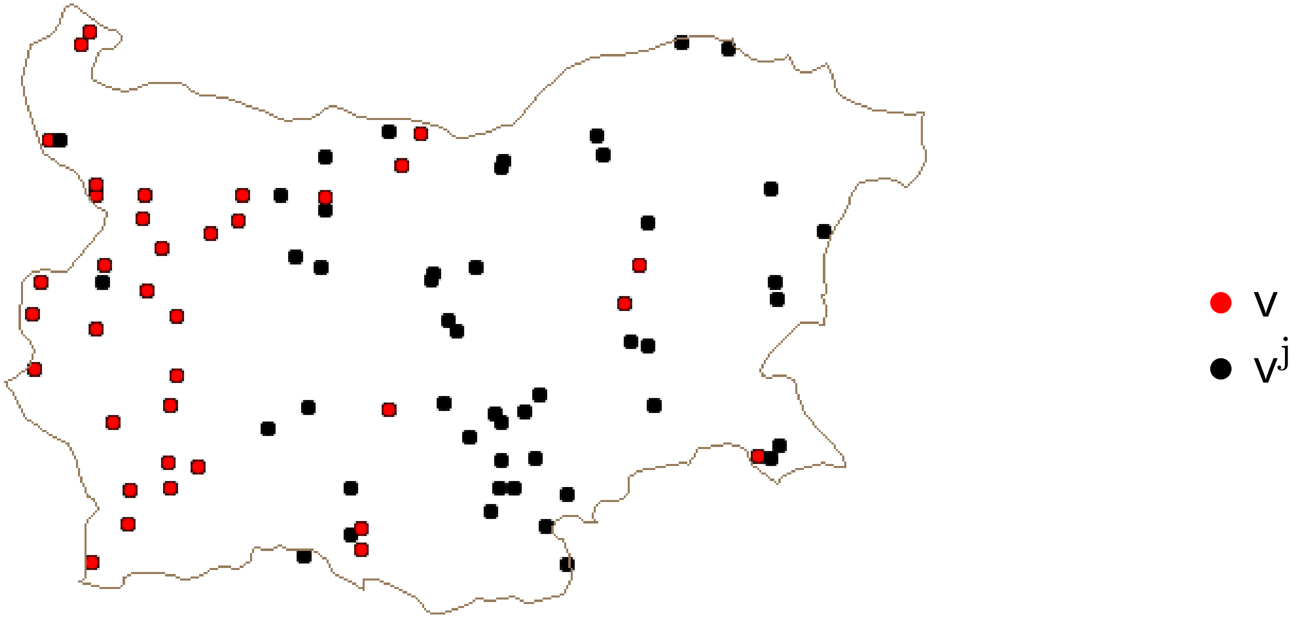


Figure 15: Palatalization of front consonant in word ”vjatyr”

Reflexes of 'jat' in Non-Palatal Environment

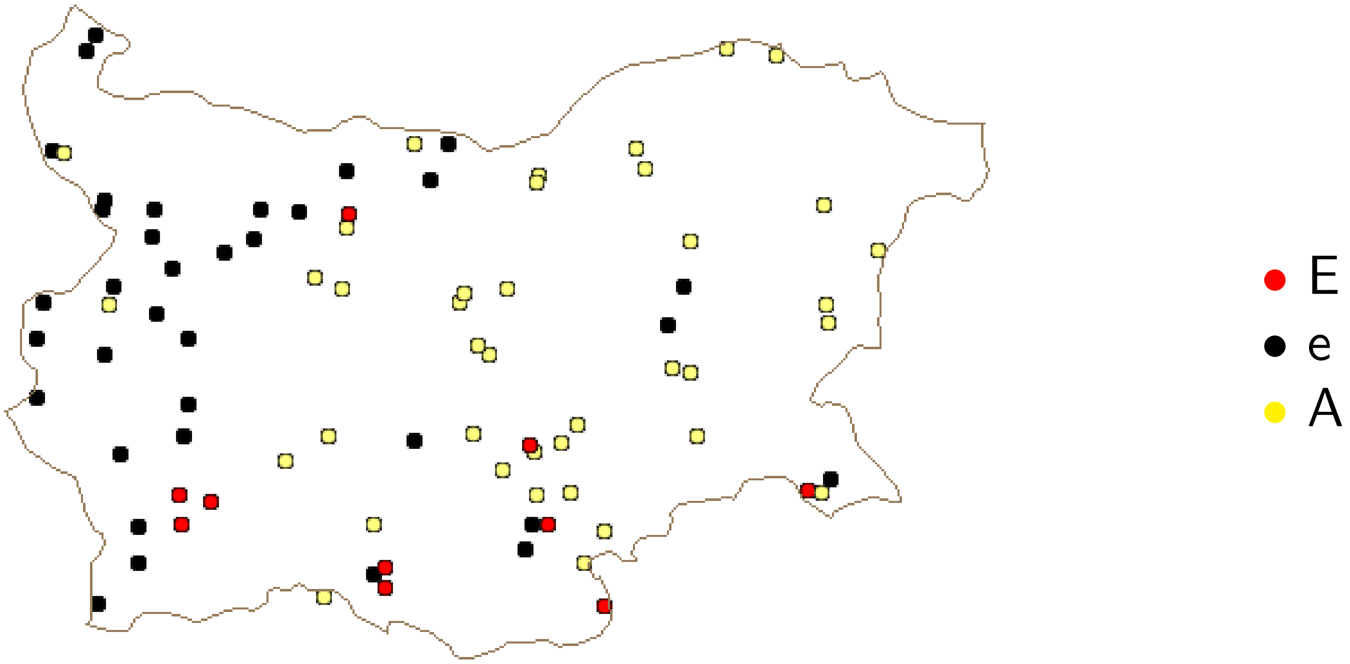


Figure 16: Reflexes of 'jat' in word 'vjatyr'

Reflexes of the Front Nasalized Vowel

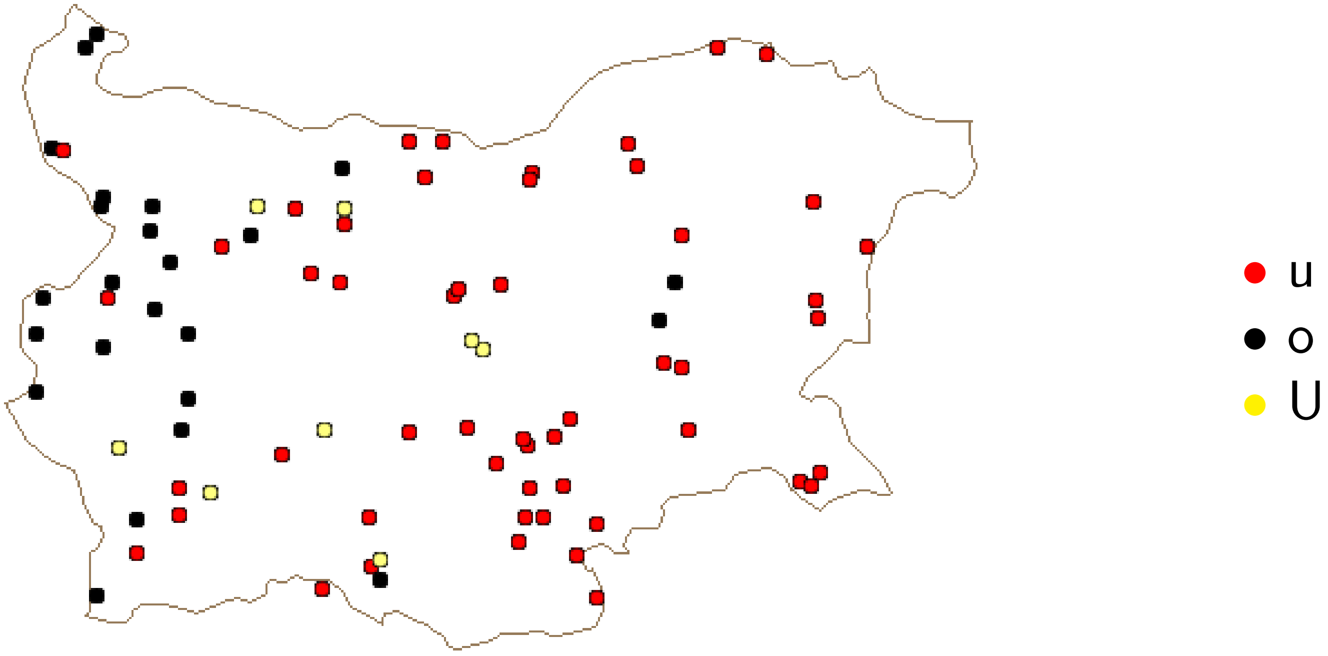


Figure 17: Reflexes of old nasalized vowel in word 'govedo'

Conclusions

- Distance-based methods give nice representation of data
- Networks are more suitable for dialect representation than trees
- Character-based methods give deeper insigths into dialect change and variation

Future Work

- Exploration of character-based methods
- Other character-based methods should be included
- Automatic detection of phonetic changes that represent the same phenomenon