Automatic detection of patterns of sound changes

Jelena Prokić Michael Cysouw

Ludwig-Maximilians-Universität München

20th International Conference on Historical Linguistics
Osaka, July 25-30, 2011

Overview

Levenshtein distance

- One of the most successful methods to determine sequence distance (Levenshtein, 1964)
  - biological molecules, software engineering, ...
- Levenshtein distance: minimum number of insertions, deletions and substitutions to transform one string into the other
  Syllabicity constraint add: vowels never substitute for consonants

m	ɔ	ə	l		k	ə
m	ɛ		l	ə	k	
	1	1		1		1

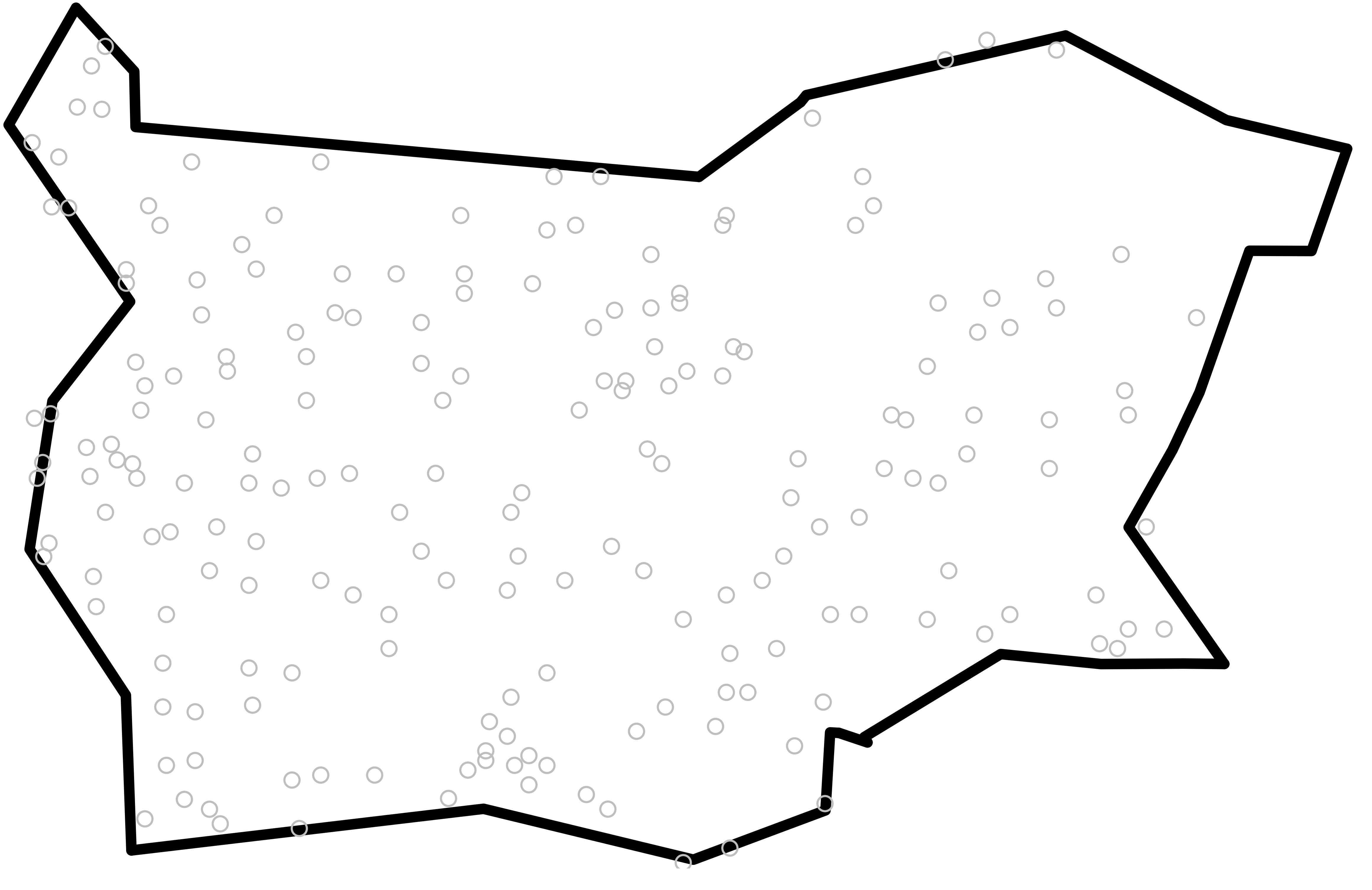
- One of the most commonly used methods in quantitative language comparison, including automatic approaches to historical linguistics

Positive aspects

- Allows automatic alignment of the strings
- Very fast and easy to implement
- Gives good general picture of the relatedness between language varieties

Negative aspects

- Usually based on 0/1 segment differences
- Yields too little insight into the linguistic basis of differences
- There is no model of linguistic change

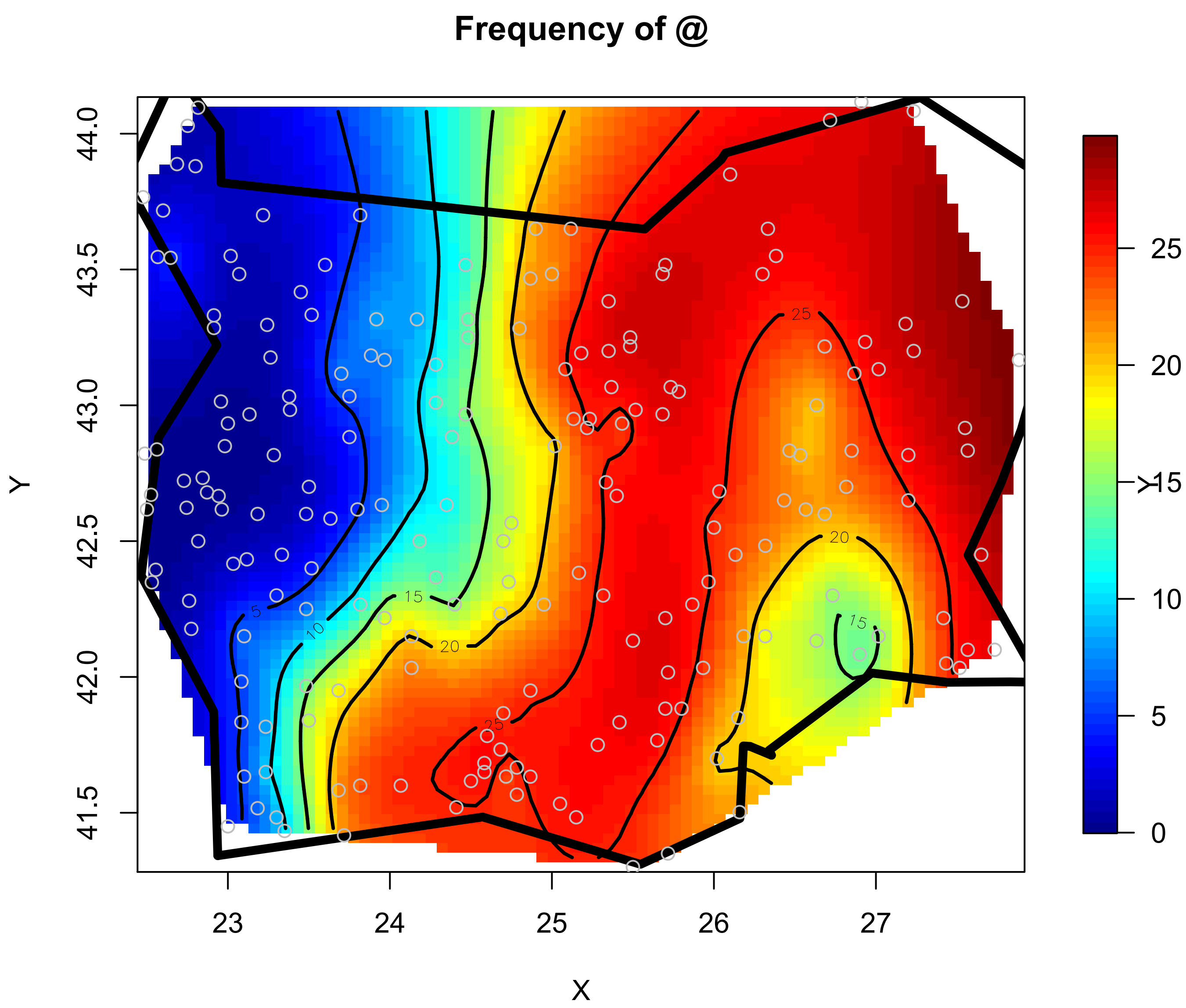


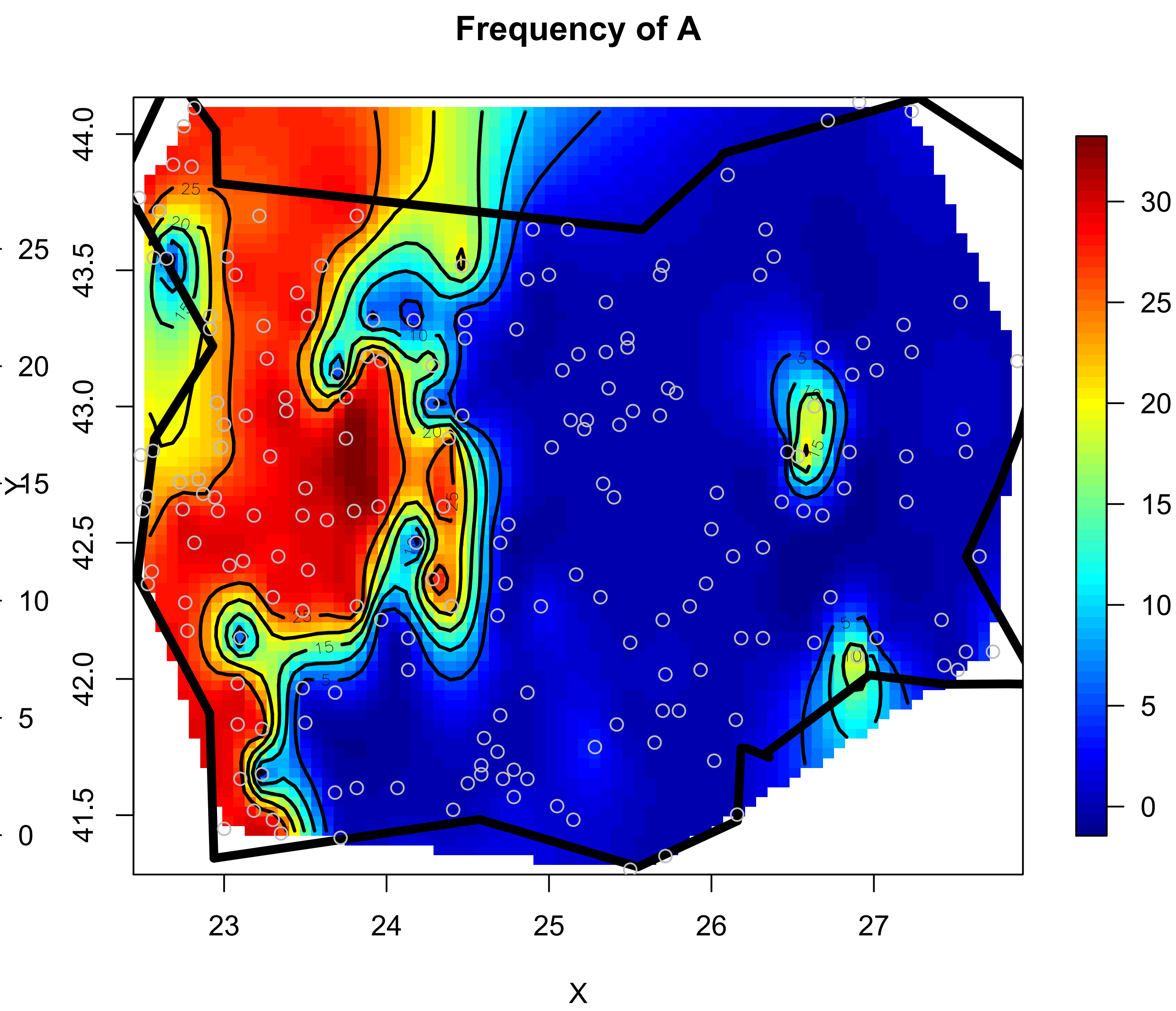
Multisequence Alignment

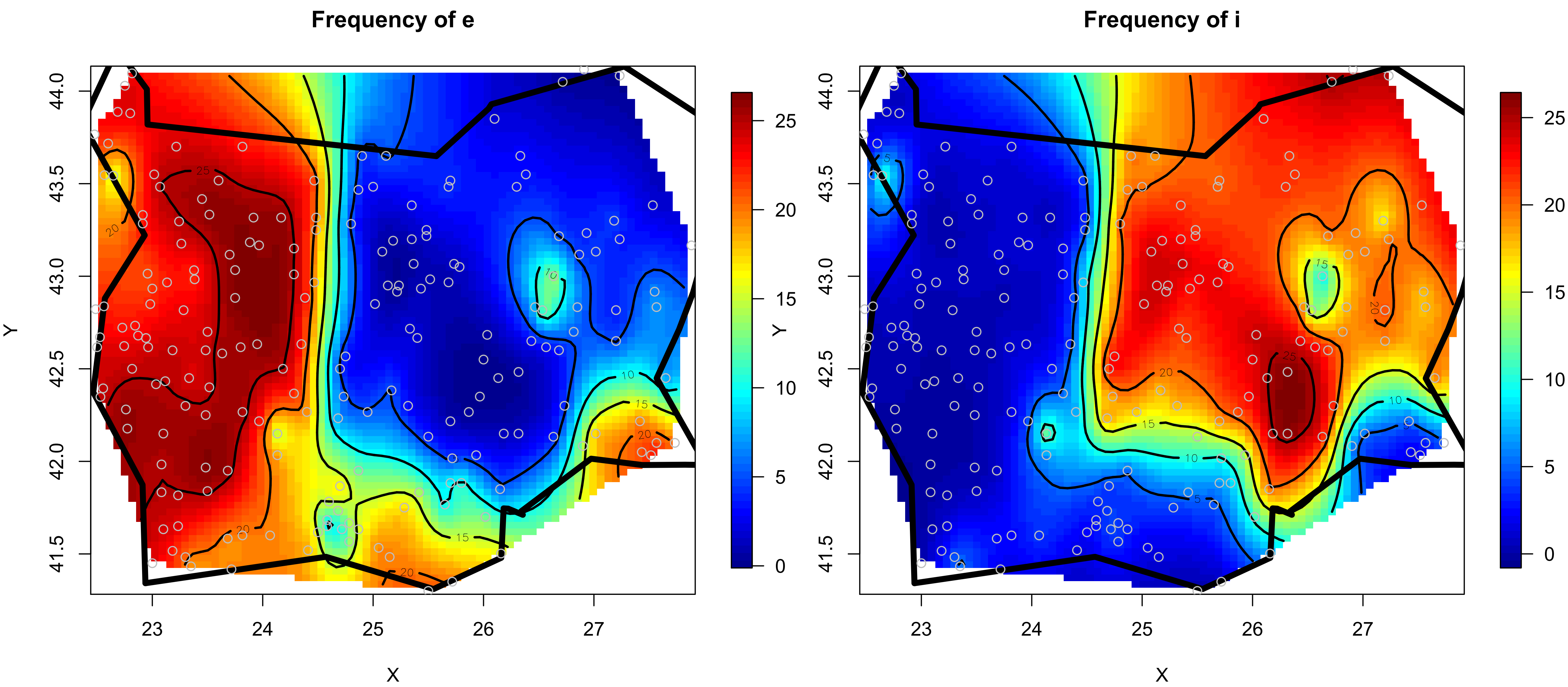
- We first need to align not only pairs of strings, but large sets.
  - Gusfield "holy grail of string algorithms"
    —no perfect solution
- Softwares for automatic multisequence alignment in linguistics:
  Prokić et al. (2009)
  Johann-Mattis (2010)
  Steiner et al. (2011)

Example of multi-aligned strings

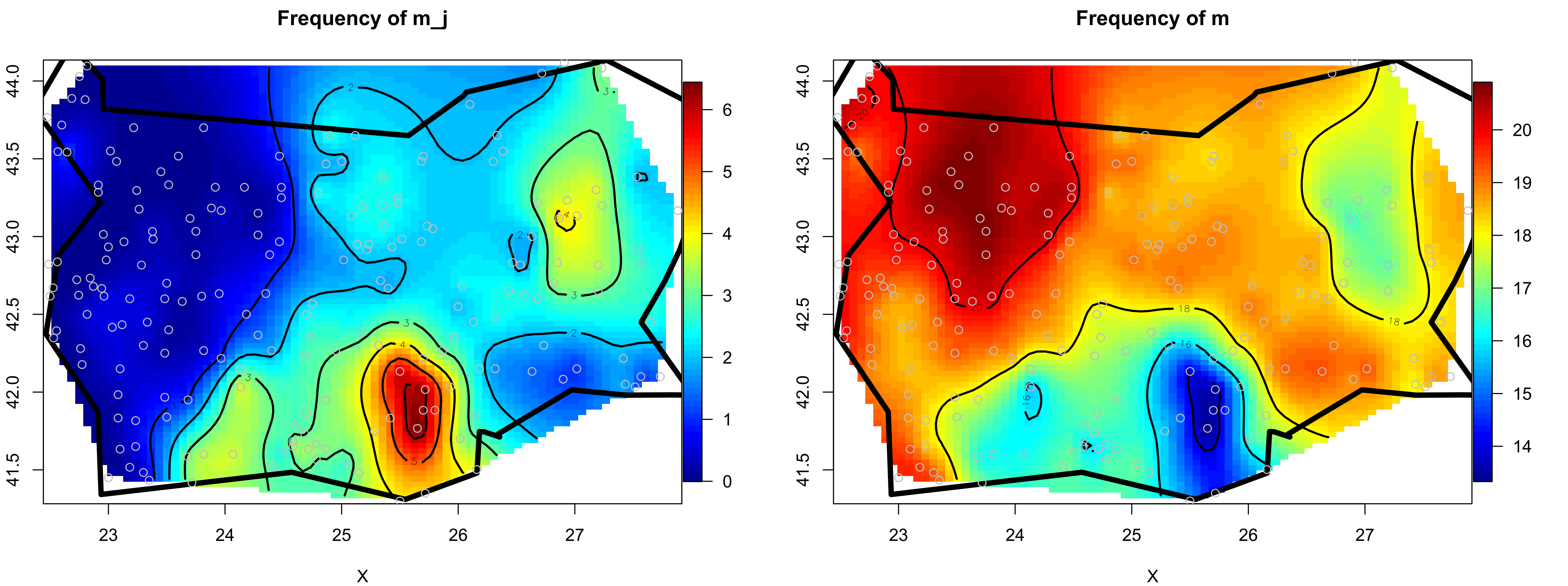
Aldomirovtsi:	v	'e	t͡ʃ	e	r	d	-	n	'o	l	'ɤ	s	n	o
Asparuhovo-Lom:	v	'e	t͡ʃ	e	r	d	-	n	'o	l	'e	s	n	o
Asparuhovo-Prov:	vʲ	'e	t͡ʃ	ə	r	d	'ɤ	n	u	lʲ	'e	s	n	u
Babyak:	v	'e	t͡ʃ	e	r	d	-	n	'o	?	?	?	?	?
Bachkovo:	v	'e	t͡s	e	r	d	'ɑ	n	u	lʲ	'e	s	n	u


Frequency of @
Y
X


Frequency of A
Y
X



Frequency of e
Frequency of i
X
Y
23 24 25 26 27
41.5 42.0 42.5 43.0 43.5 44.0
0 5 10 15 20 25


Frequency of m_j
Frequency of m
X
X
23
24
25
26
27
41.5
42.0
42.5
43.0
43.5
44.0
0
1
2
3
4
5
6
14
15
16
17
18
19
20

Example of multi-aligned strings

Aldomirovtsi:	v	'e	t͡ʃ	e	r	d	-	n	'o	l	'ɤ	s	n	o
Asparuhovo-Lom:	v	'e	t͡ʃ	e	r	d	-	n	'o	l	'e	s	n	o
Asparuhovo-Prov:	vʲ	'e	t͡ʃ	ə	r	d	'ɤ	n	u	lʲ	'e	s	n	u
Babyak:	v	'e	t͡ʃ	e	r	d	-	n	'o	?	?	?	?	?
Bachkovo:	v	'e	t͡s	e	r	d	'ɑ	n	u	lʲ	'e	s	n	u

Do various word positions vary at different rate?

- Yes (not surprising)
- Can we measure that?
- Yes: Shannon index

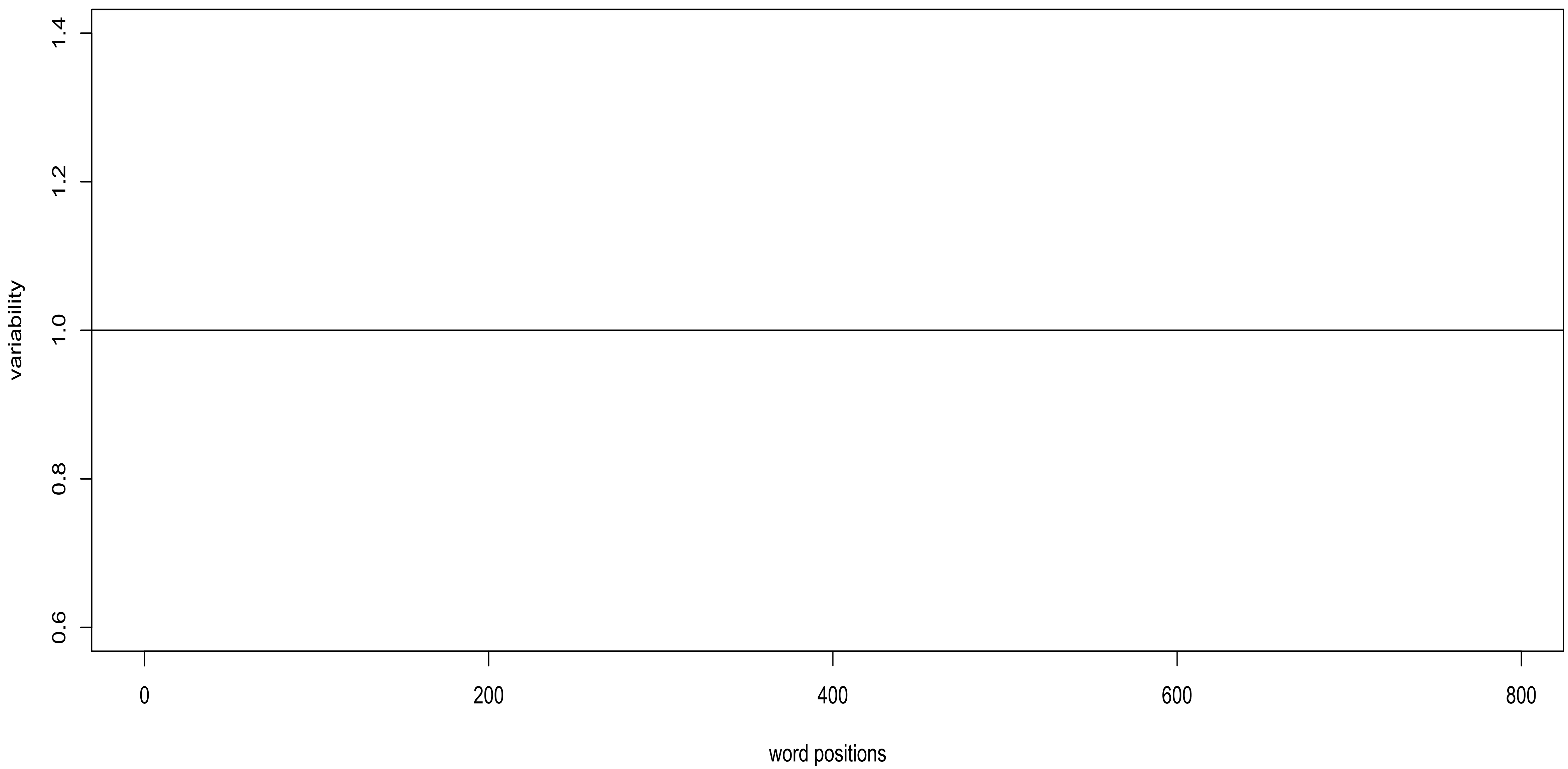
Shannon index

- Information entropy of the distribution

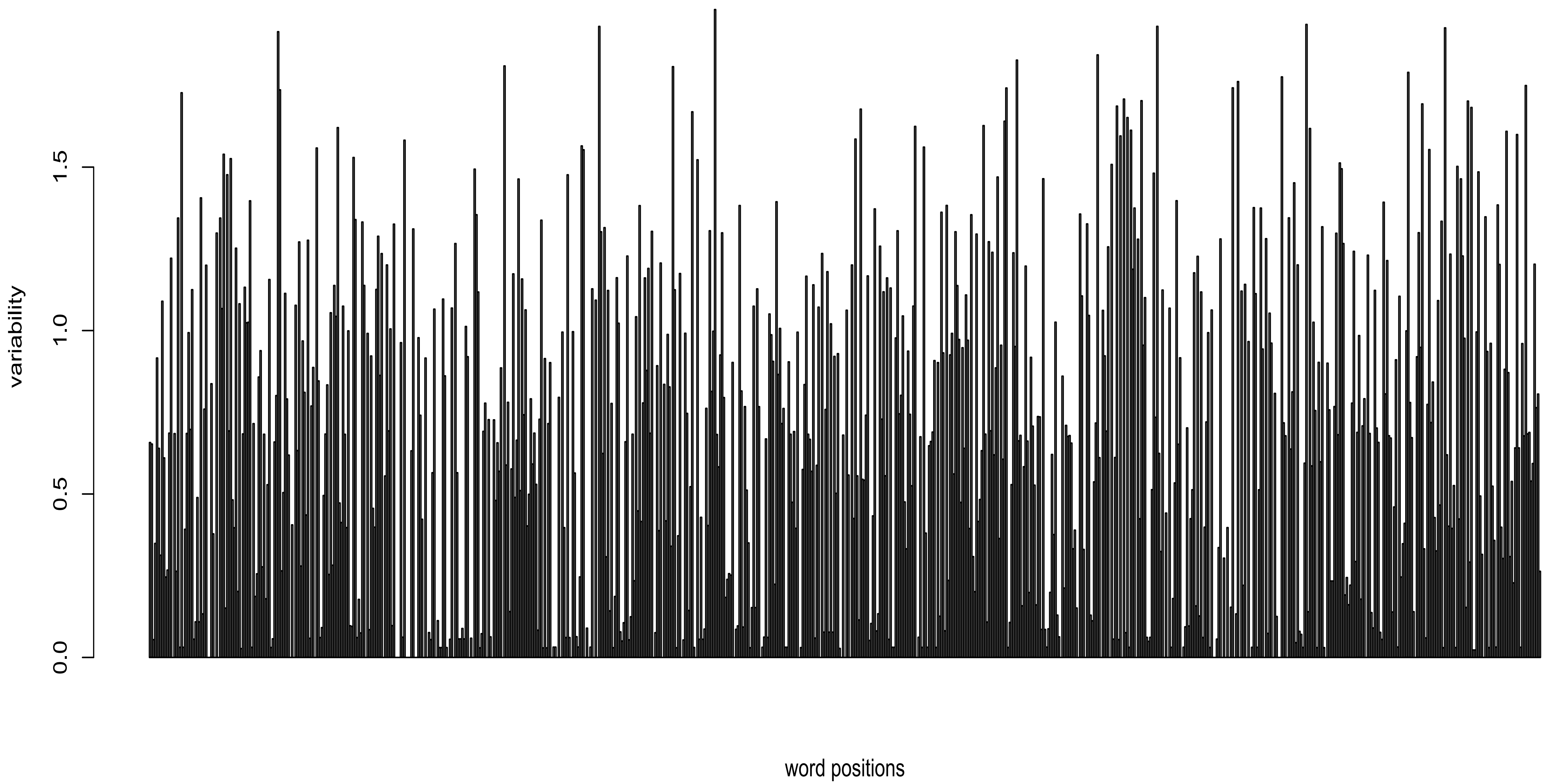
$$H = -\sum_{S}^{i=1} p_i \ln p_i$$

S - number of phones, p_i - relative abundance of phone i.

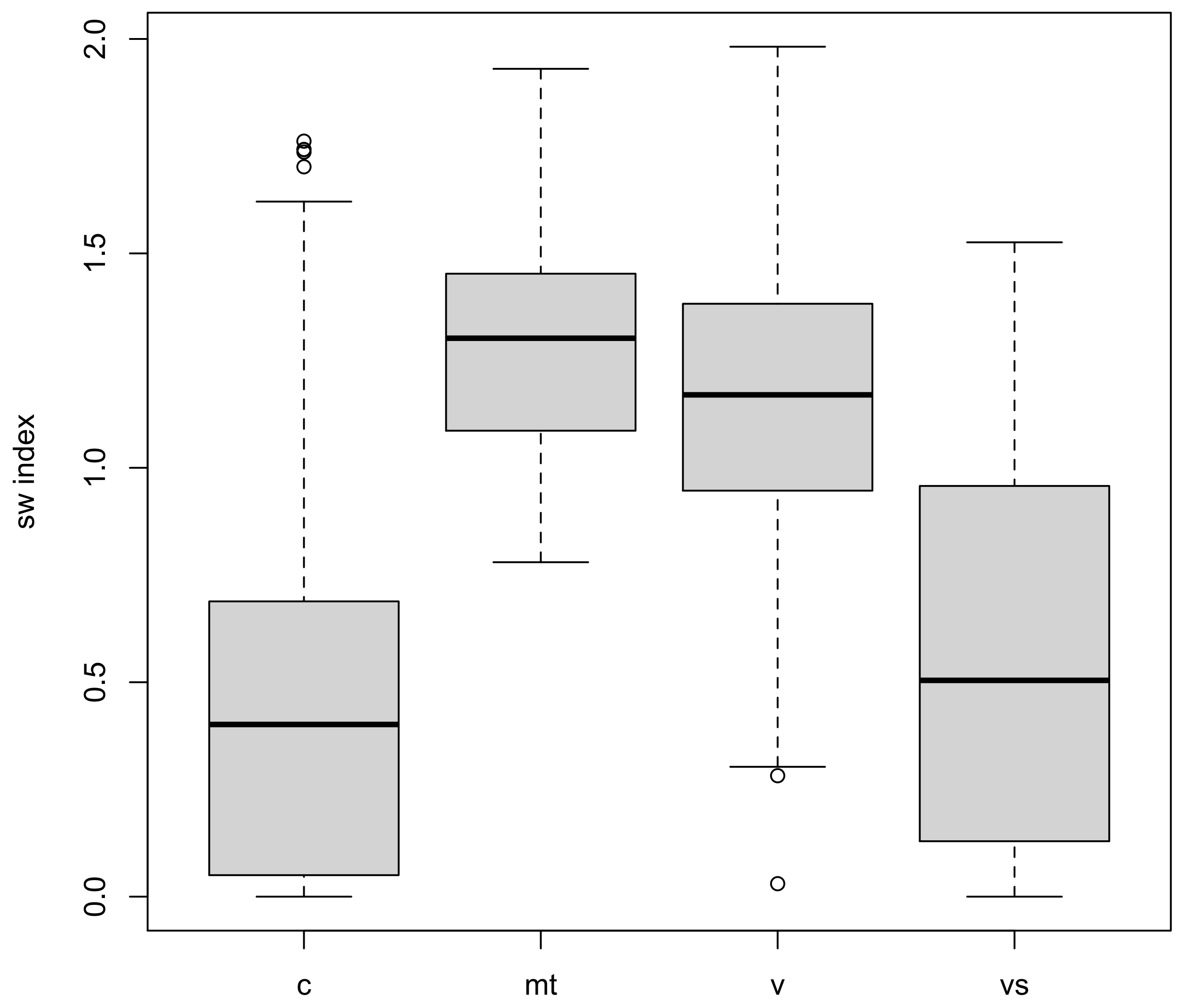
Variability under the "Levenshtein" model



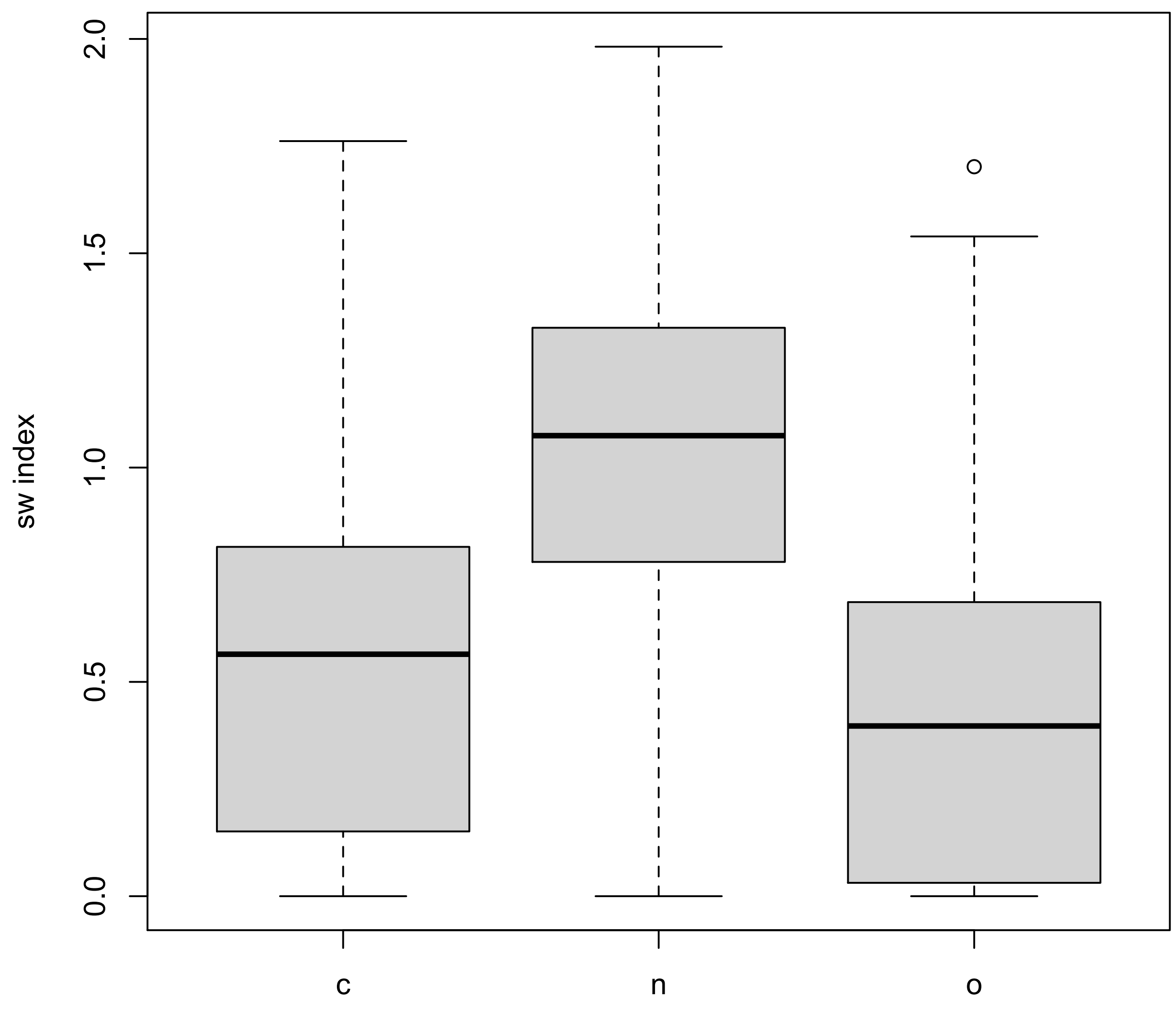
Actual variability



Variability of vowels and consonants



Variability: onset vs. nucleus vs. coda



Computing highly correlated positions

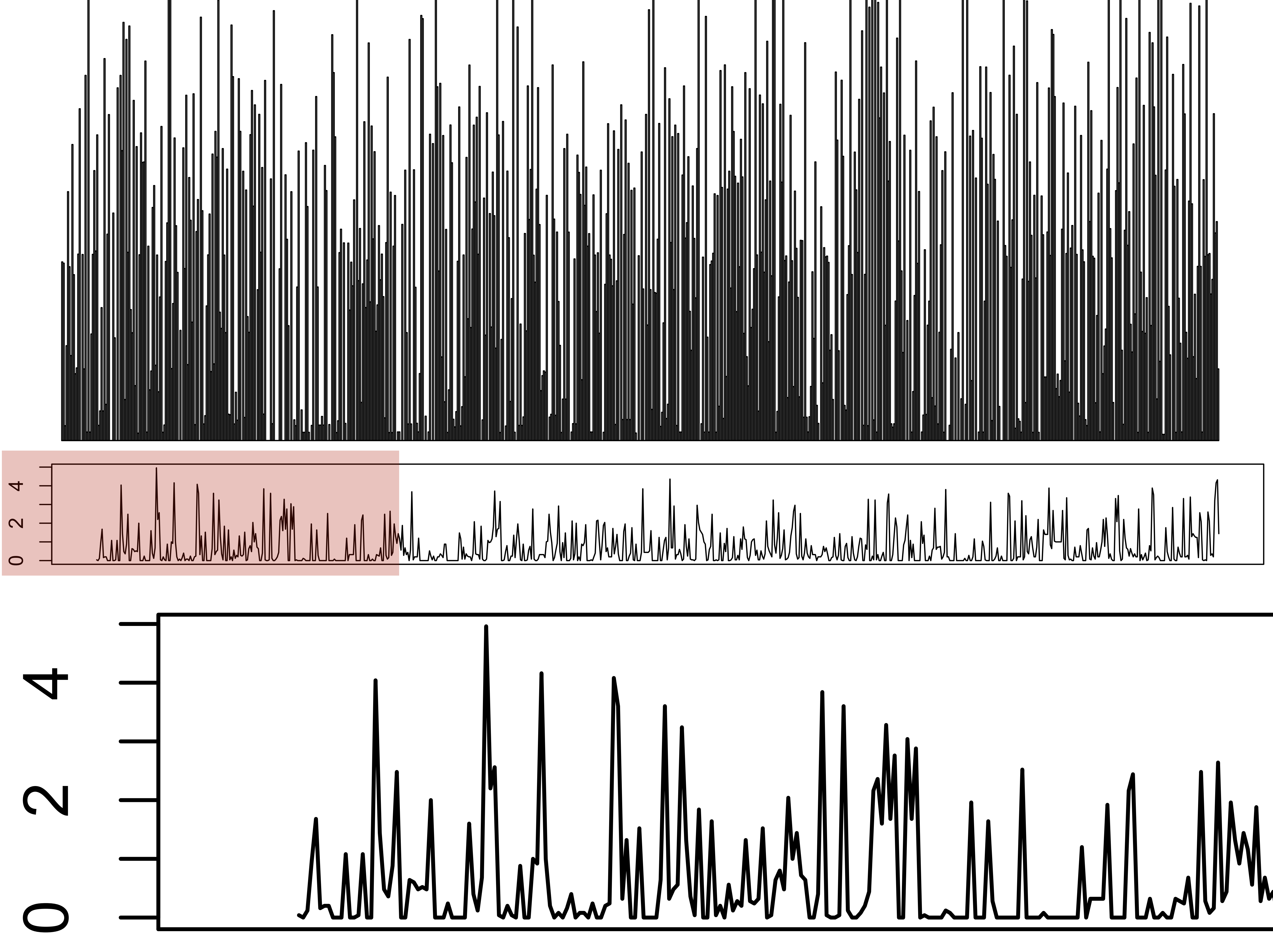
- How regular are sound changes?
- Completely identical positions were not found in the data set
- Word positions that show high correlation can be identified using Mutual Information (MI)

Mutual Information

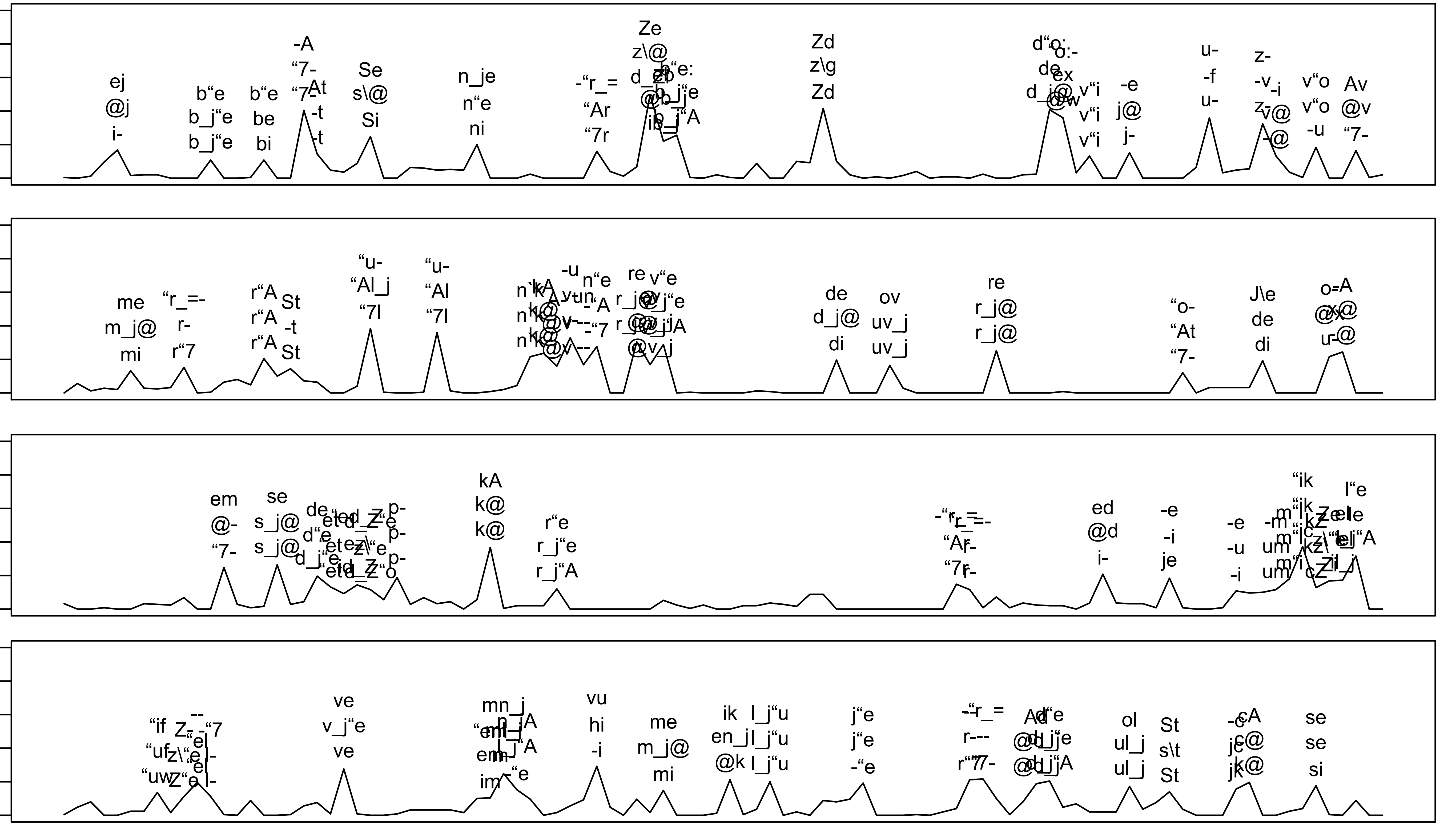
- The mutual information of two random variables is a quantity that measures the mutual dependence of the two variables.

$$I = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$$

p(x,y) is the joint probability distribution function of X and Y, and p(x) and p(y) are the marginal probability distribution functions of X and Y respectively.



0
2
4
0
2
4

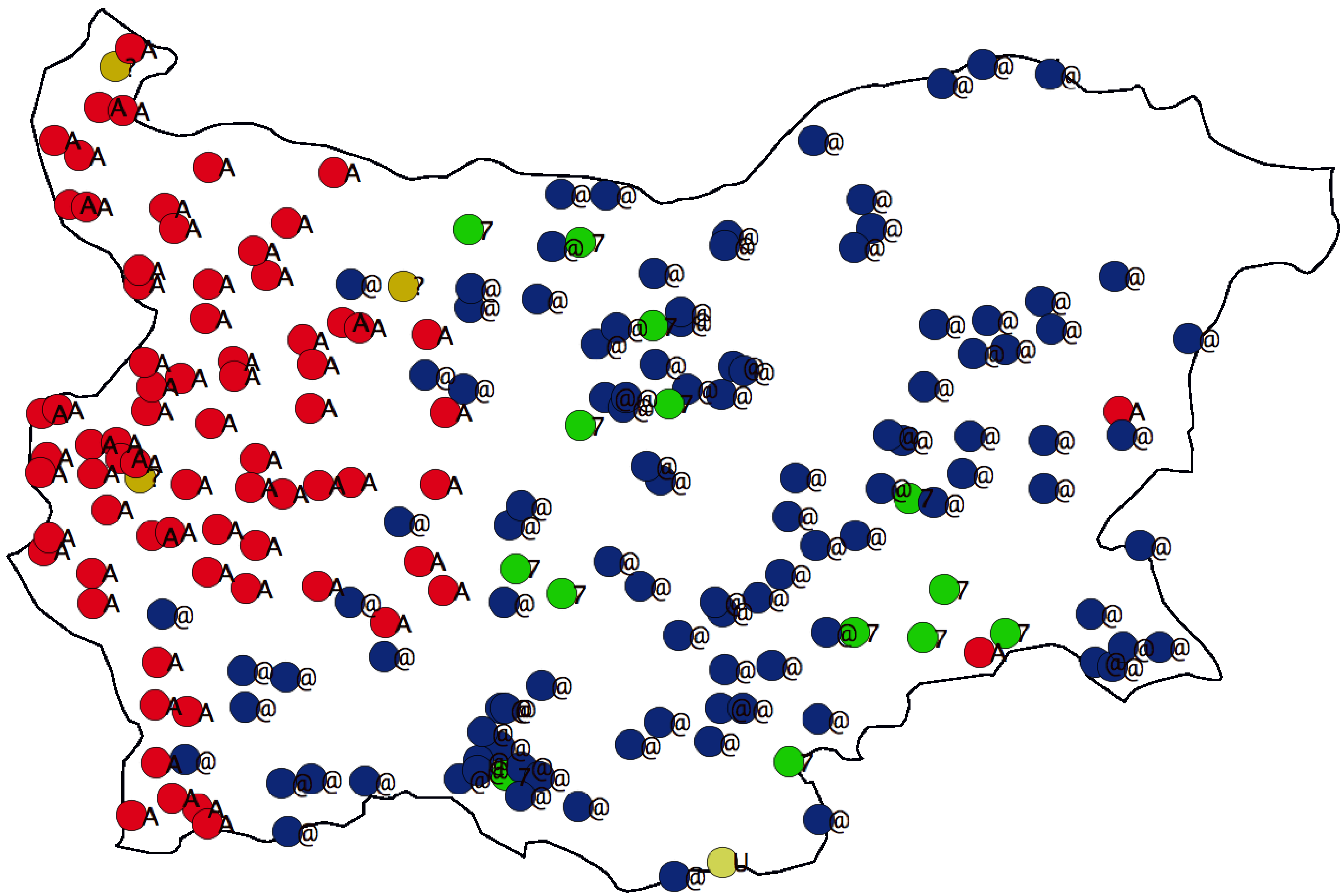


ej
@j
i-
b“e
b_j“e
b_j“e
b“e
be
bi
Se
s\@
Si
n_je
n“e
ni
-“r_=
“Ar
“7r
Zd
z\g
Zd
u-
-f
u-
v“o
v“o
-u
Av
@v
“7-
me
m_j@
mi
“u-
“Al_j
“7l
“u-
“Al
“7l
de
d_j@
di
ov
uv_j
uv_j
re
r_j@
r_j@
“o-
“At
“7-
J\e
de
di
em
@-
“7-
se
s_j@
s_j@
kA
k@
k@
r“e
r_j“e
r_j“A
ed
@d
i-
-e
-i
je
ve
v_j“e
ve
vu
hi
-i
me
m_j@
mi
ol
ul_j
ul_j
St
s\t
St
se
se
si

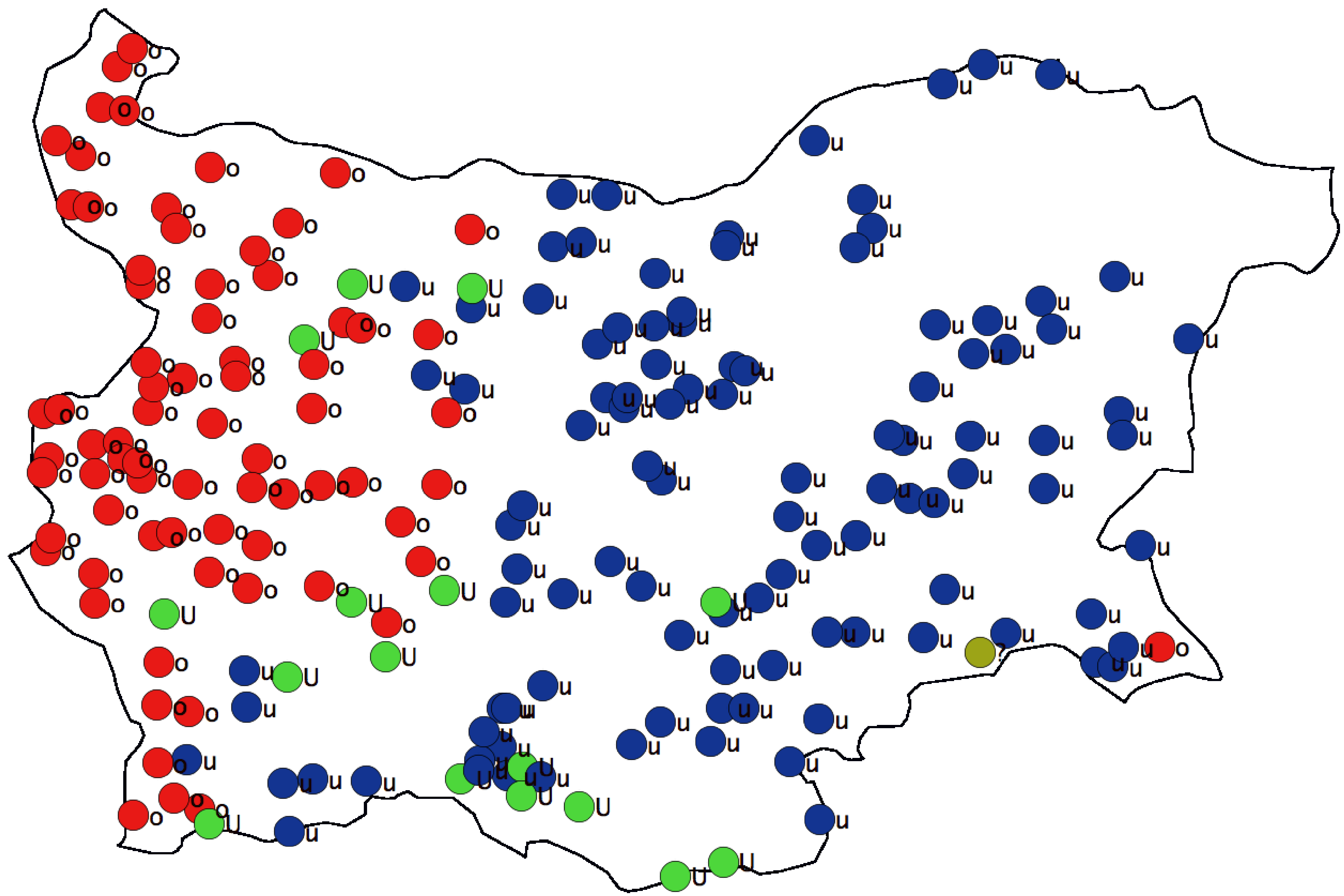
Results

- Each position was compared to all others
- Highly correlated positions were linked into groups
- Four groups of highly correlated positions were detected

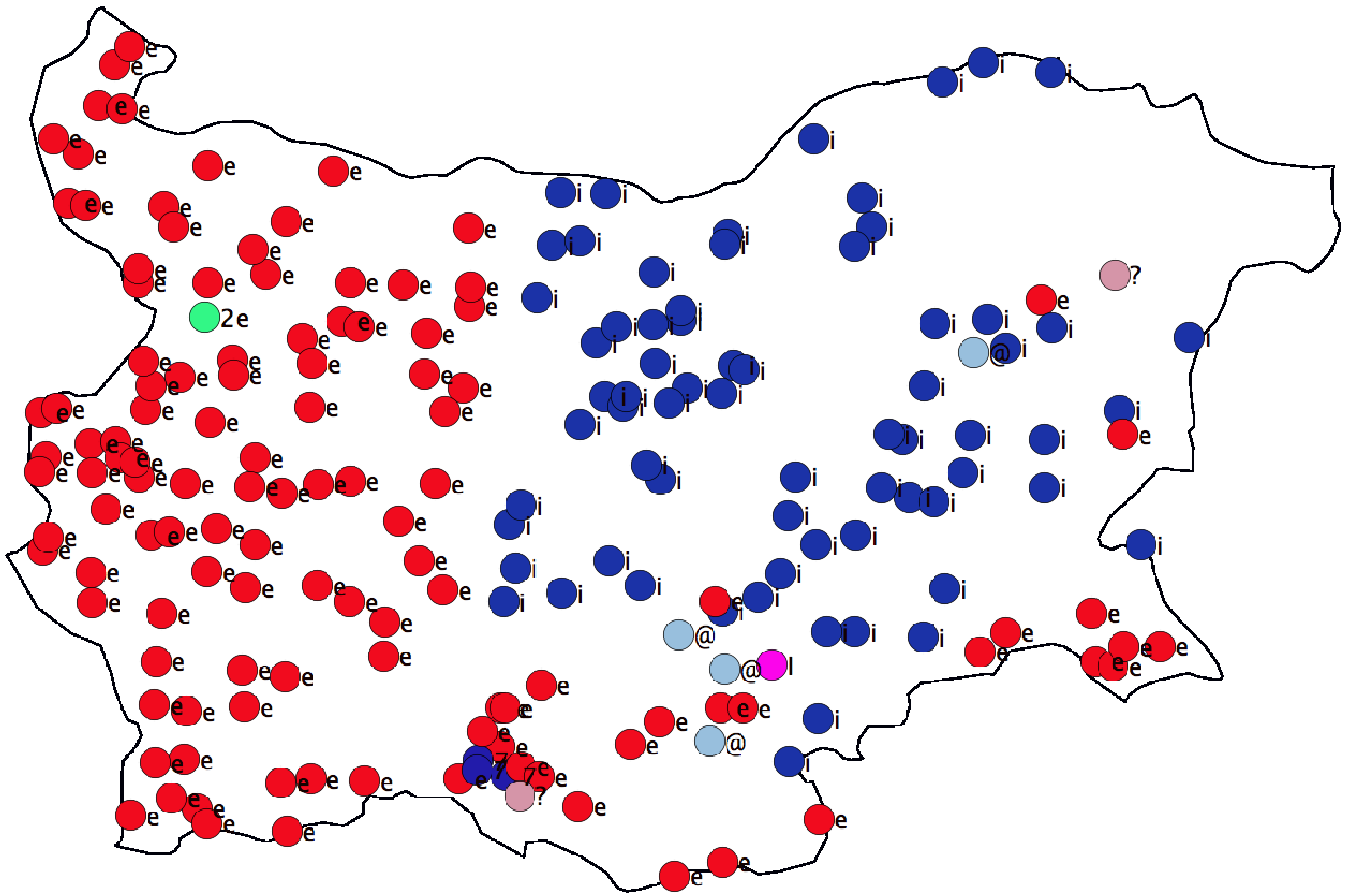
Cluster 1: back vowels



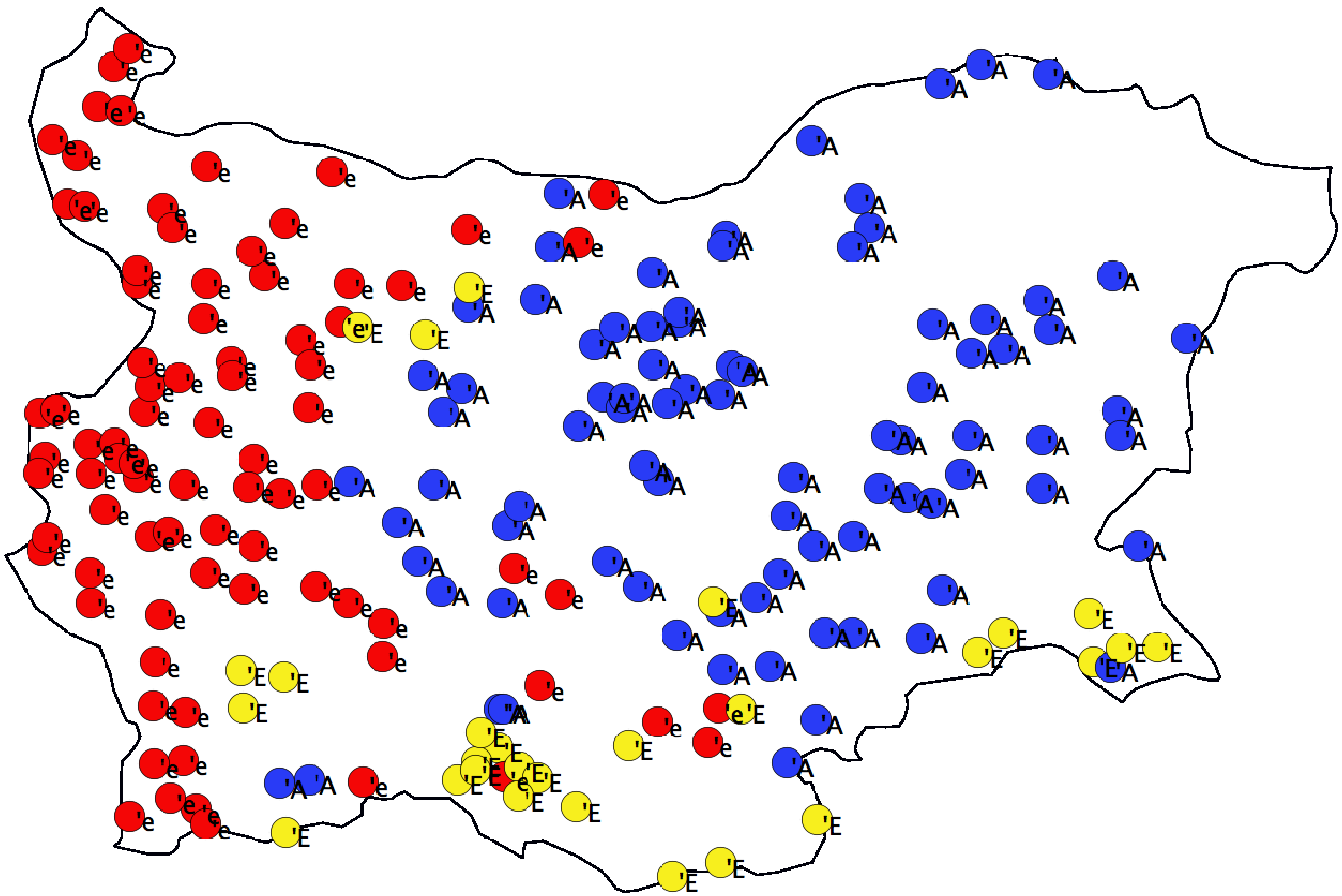
Cluster 1: back vowels



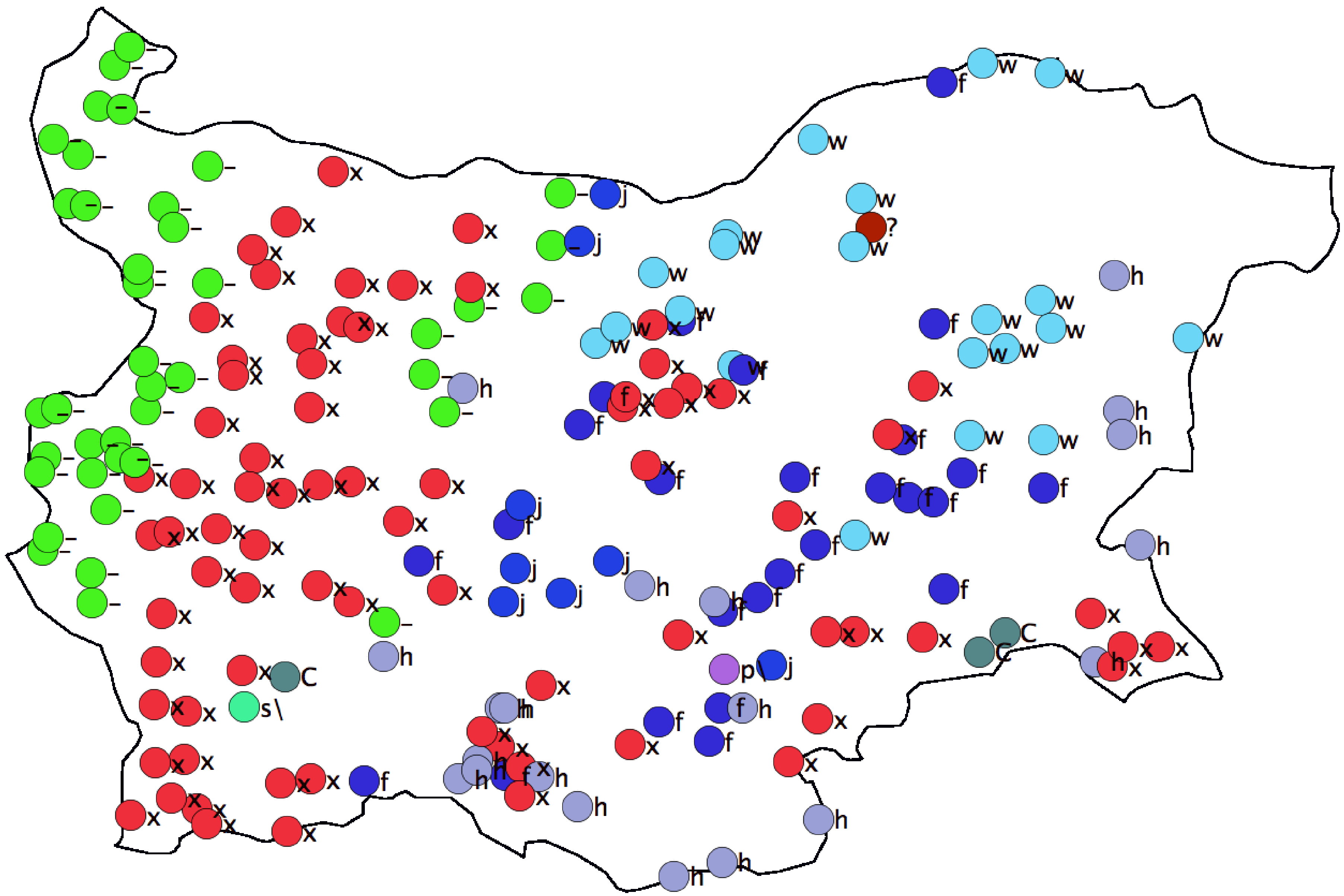
Cluster 2: front vowels



Cluster 3: the ‘yat’ vowel



Cluster 4: presence of velar fricative



Conclusions

- Occurrence of geographical patterns from the data
- MI is not content dependent, but highly correlated columns have similar phonetic content
- This approach enables us to automatically identify layers of sound changes