Thomas Zastrow

Exploring Quantitative Methods in Dialectometry

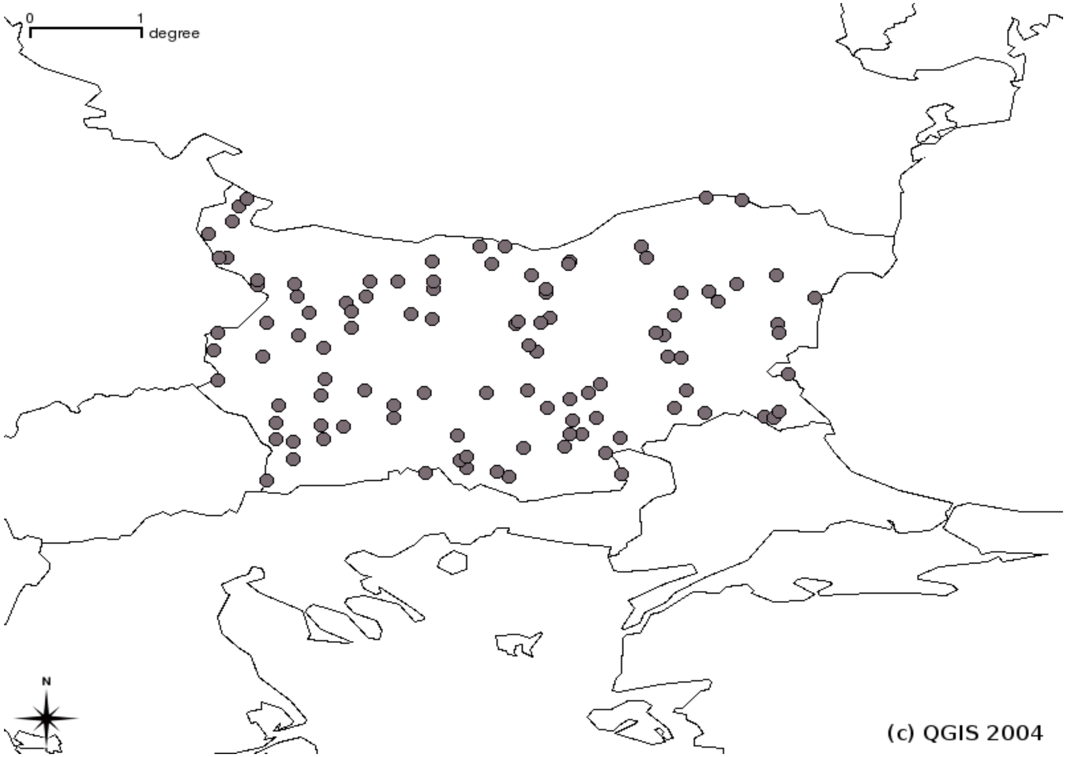
Seminar für Sprachwissenschaft
Eberhard Karls Universität Tübingen

Outline

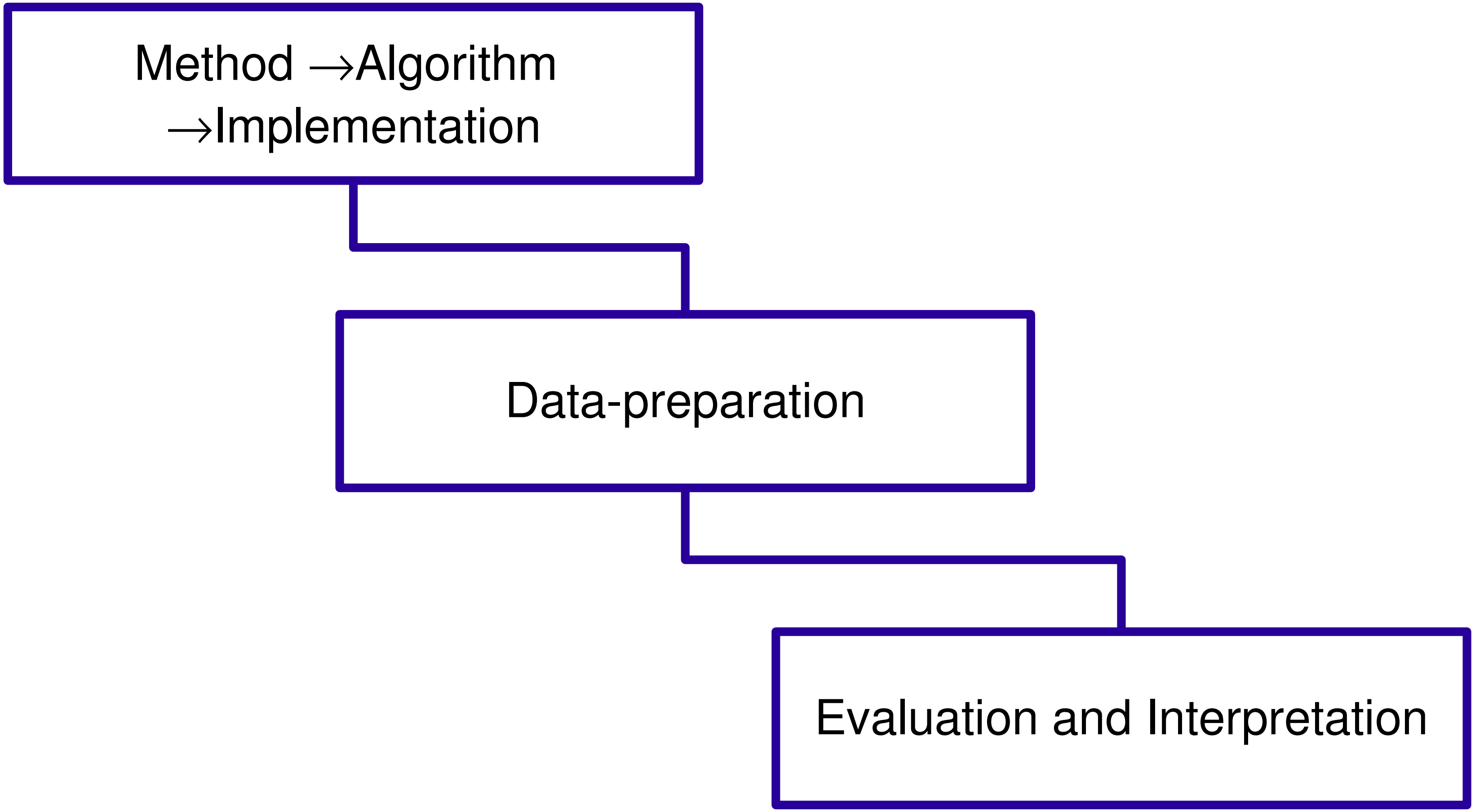
- Pronunciation Data
- General procedure for applying quantitative methods to dialect-data
- Methods: Entropy
- Methods: Vector-Analysis
- Further methods
- Technical aspects
- ToDo & Dates

1 Pronunciation data

- 106 sites
- 143 word for every site (some sites has more)



2 General procedure for applying quantitative methods to dialect-data



Method →Algorithm →Implementation

Data-preparation

Evaluation and Interpretation

2.1 Method – Algorithm - Implementation

Implemented	**ToDo**
Relative Entropy	Alingment-algorithms
Partial Information	Compression
Conditional Entropy	Bayes-filter/-nets
Vector-Analysis	

2.2 Data-preparation

- Bi-, Tri-, N-grams
- Word-list vs. site-list
- Site-list compared to the standard

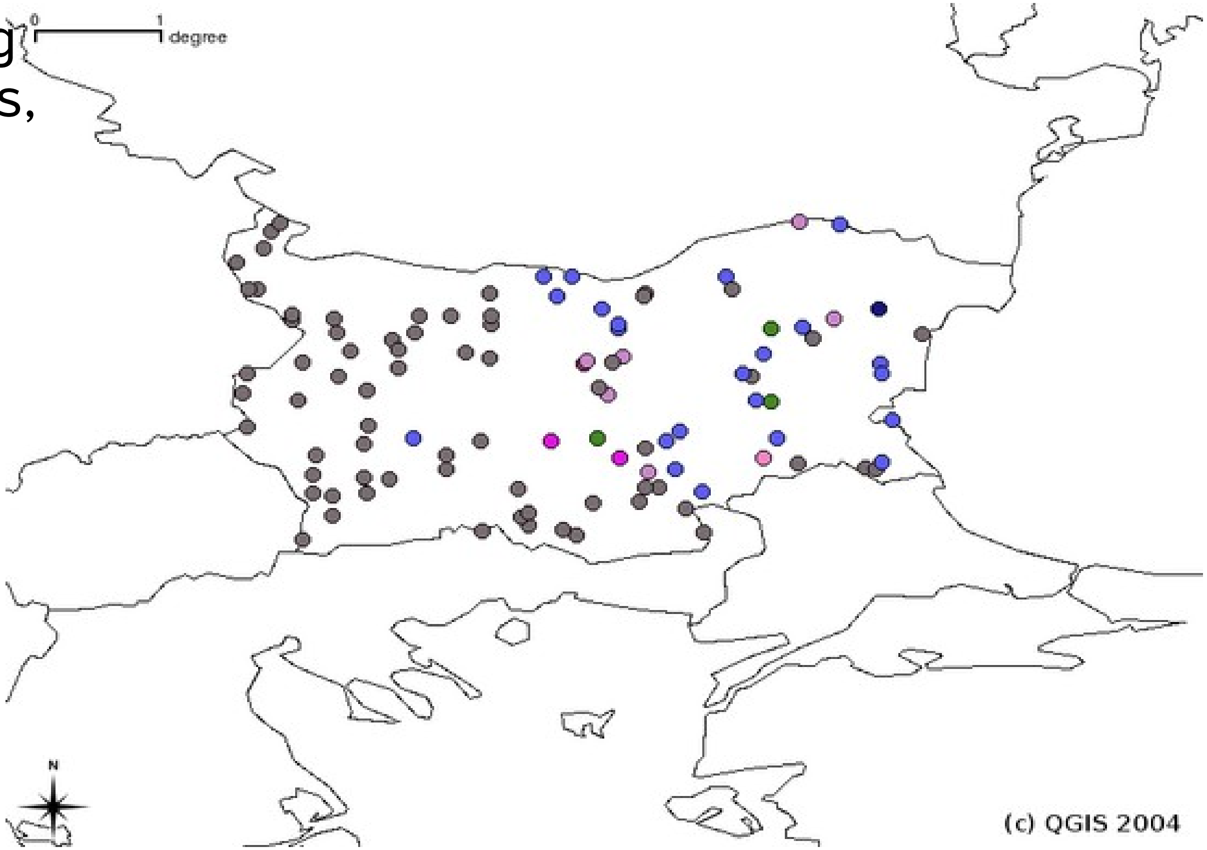
2.3 Evaluation and interpretation

- Clustering
- Geographical distribution, visualized with maps
- Mathematical visualisation with graphs, diagrams, ...
- More complex data-mining

Methods: Relative Entropy

- Formula: $$H_{rel} = \frac{H_{proper}}{H_{max}}$$

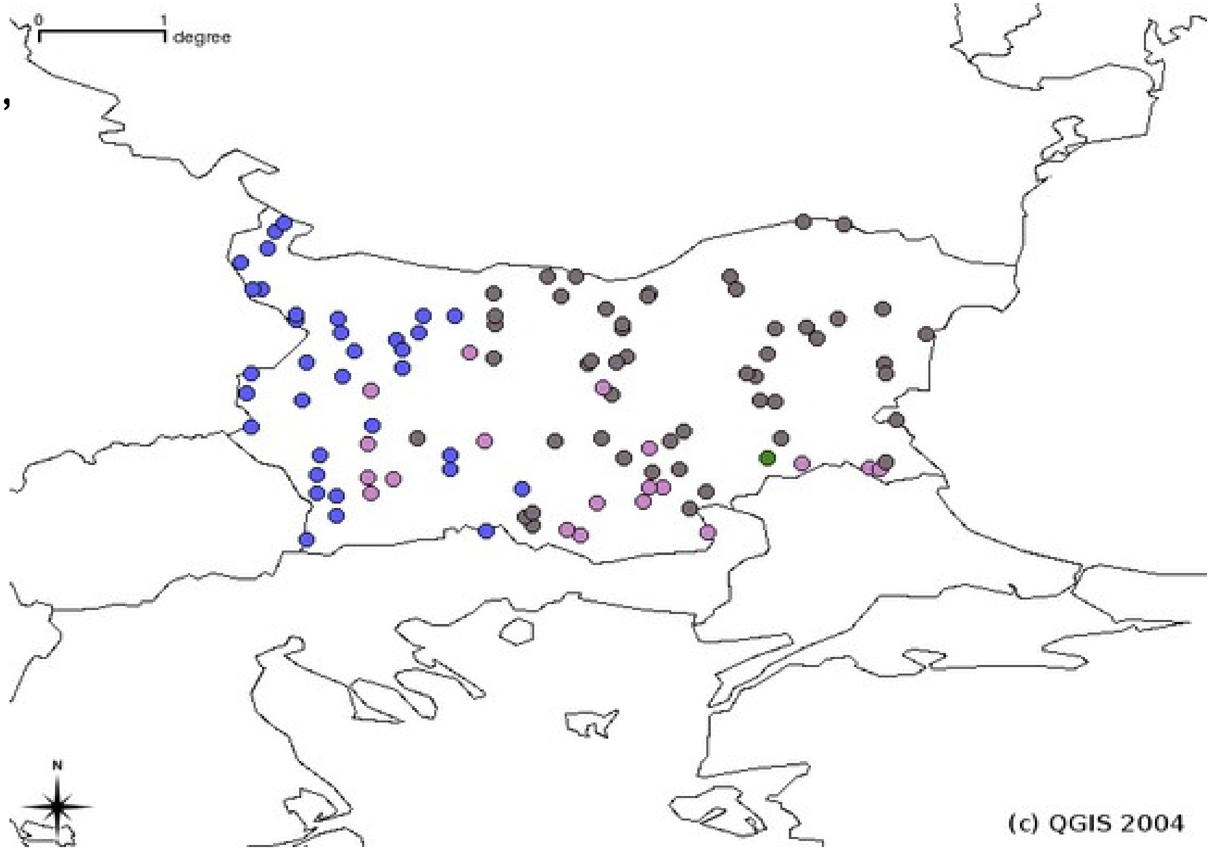
- Vowel E, clustering
- with adapted colors,
- treshhold 2:



Methods: Partial information

- Formula: $I_p(A,B) = I\frac{(A)}{I}(B) \text{ where } A \subset B$

- Vowel E, clustering
- with adapted colors,
- treshhold 6:



Methods: Conditional Entropy

- The conditional entropy compares two datasets: What is the relation between the entropies of a given, well-known (X) and an unknown dataset (Y)?[1]

- Formula: $H(Y|X) = H(X,Y) - H(X)$

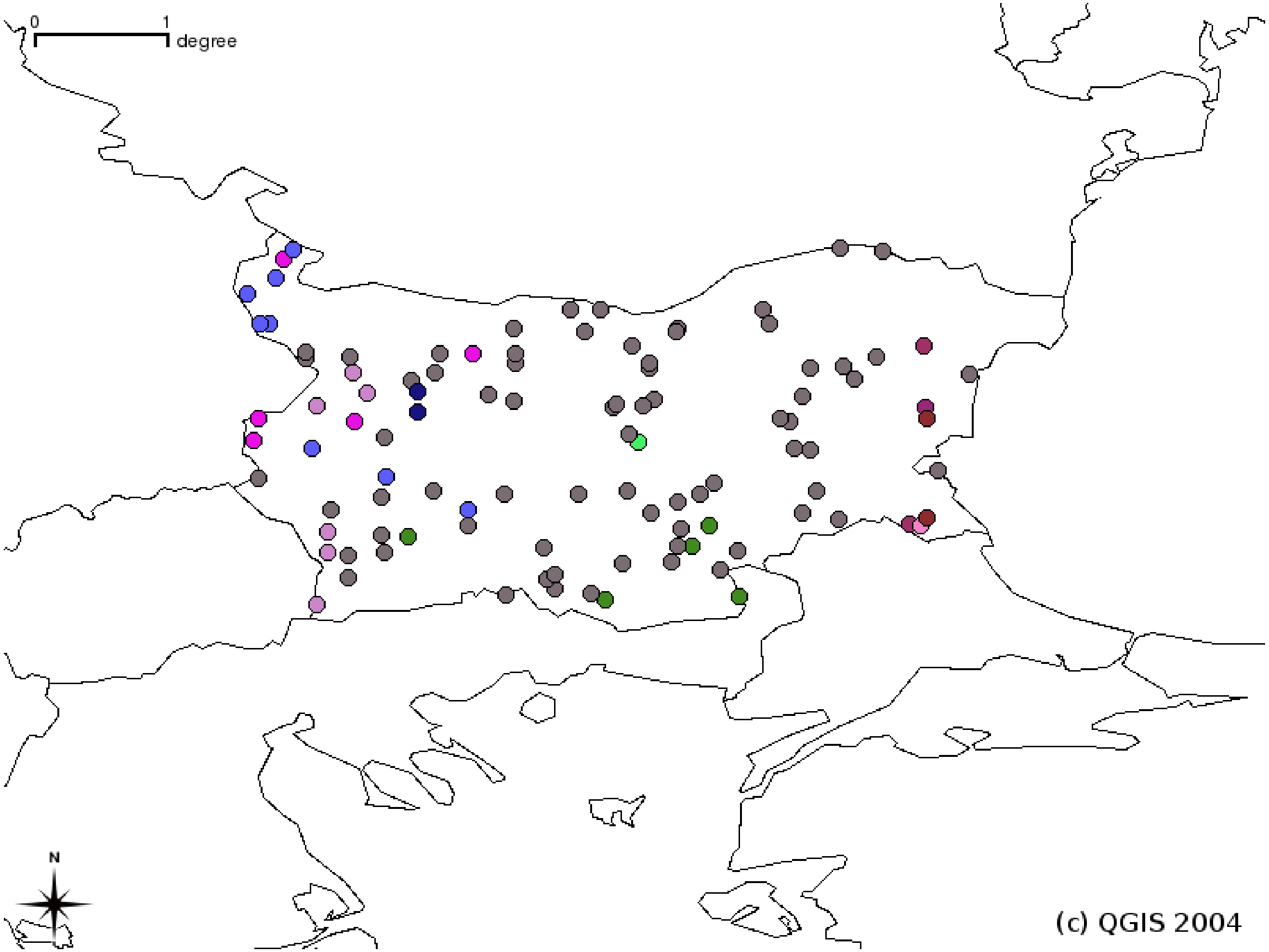
  - where H(X,Y) is the Joint Entropy:

$$H(X,Y) = -\sum_{1}^{x,y} p_{x,y} \log_2(p_{x,y})$$

- [1] See also http://en.wikipedia.org/wiki/Conditional_entropy

Methods: Conditional Entropy

- Conditional Entropy of Phone A, compared to "standard", treshhold 2:



Methods: Vector analysis

- A vector is an arrow which has two properties: Direction and length.
- In a geometrical way, an origin can be set for every site (Fig. 1) Starting from here, a chain of vectors can follow an observed phone through the site (Fig. 2.). From this vector-chain, a new vector can be compiled (Fig. 3, blue vector):

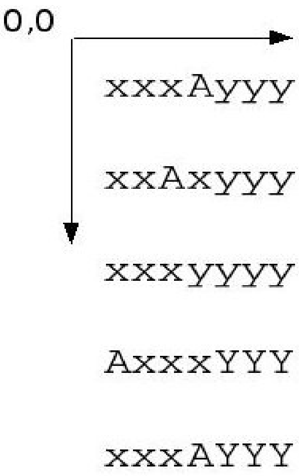


Fig. 1

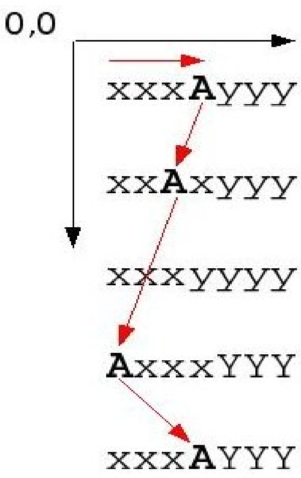


Fig. 2

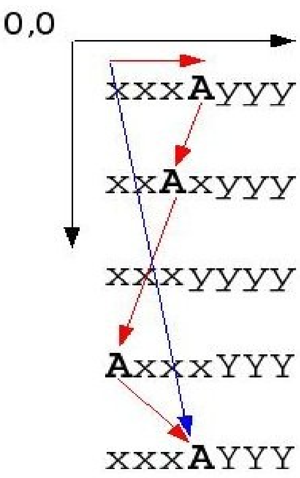
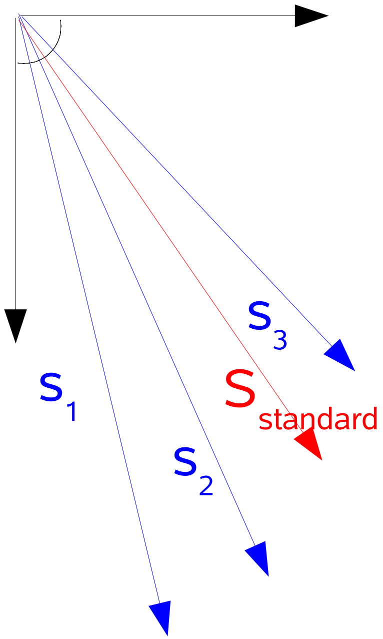


Fig. 3

Methods: Vector analysis

- When doing this for all sites, the angle between the compiled vectors should be site-specific (Fig. 4):



$$\vec{v}_1=\begin{pmatrix}3\\0\end{pmatrix}$$

$$\vec{v}_2=\begin{pmatrix}-1\\1\end{pmatrix}$$

$$\vec{v}_3=\begin{pmatrix}-2\\2\end{pmatrix}$$

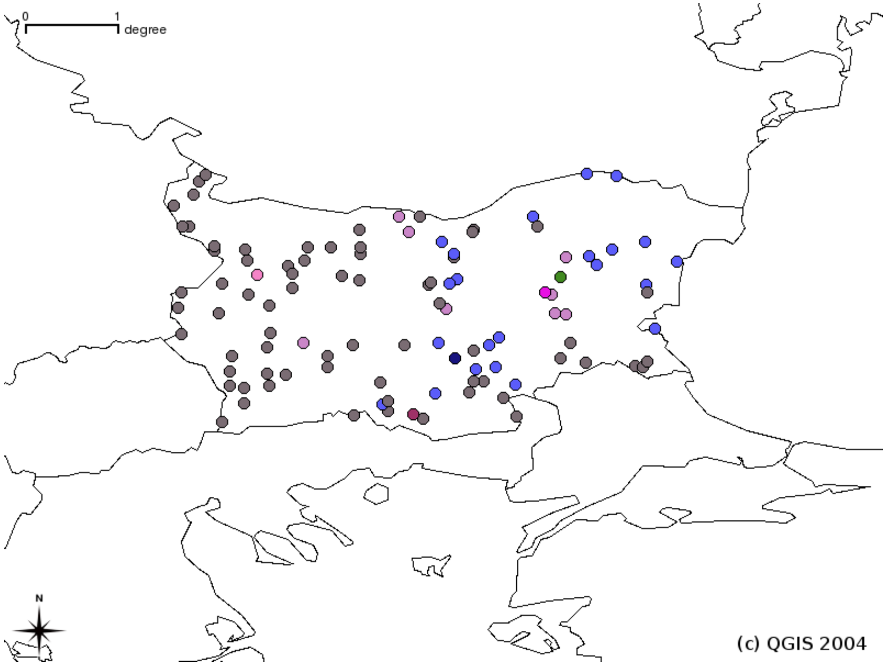
$$\vec{v}_4=\begin{pmatrix}3\\1\end{pmatrix}$$

$$\vec{v_{ges}}=\vec{v}_1+\vec{v}_2+\vec{v}_3+\vec{v}_4=\begin{pmatrix}3\\0\end{pmatrix}+\begin{pmatrix}-1\\1\end{pmatrix}+\begin{pmatrix}-2\\2\end{pmatrix}+\begin{pmatrix}3\\1\end{pmatrix}=\begin{pmatrix}3\\4\end{pmatrix}$$

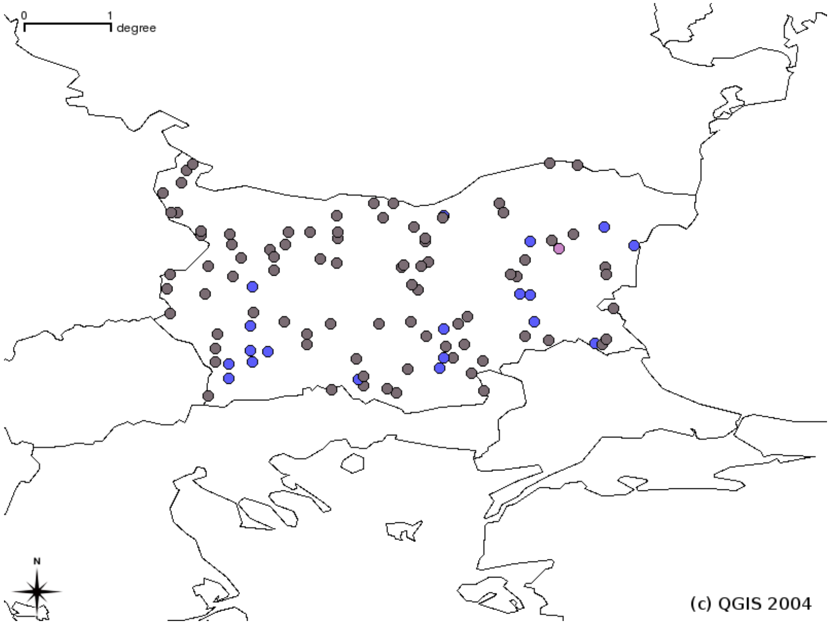
- Angle between two vectors: $\cos(\alpha)=\frac{\vec{a}*\vec{b}}{|\vec{a}|*|\vec{b}|}$

Methods: Vector analysis

- Vowel e, treshhold 4:



- Vowel u, treshhold 6:



Conclusions

- The actual implemented methods are all suitable for dialectometry
- In the current form, the methods are showing a clear distinction between west and east

To get better results:

- More data
- Better clustering

Further methods

- Alignment
  Covington, Kondrak
- Compression
  ZIP, /-Zip, RAR, ...
- Bayes-filter / -nets

ToDo & Dates

- CLIN 01/12/07
- Presentation with the geographers in Tübingen

- More geographical information (Google Maps, CIA Factbook etc.)
- Demographical data

Technical aspects - problem

The database

Until now, the OpenSource-XML-database eXist is used in the Buldialect-project.

This database has some insufficiencies:
It's slow.
Problem with big amounts of Xqueries (> 10.000)

Technical aspects - solution

The database

Switching to another, more powerful database: DB2 from IBM.

DB2 is not OpenSource, but usable without any costs.
Should be much more faster than eXist.
And much more robustly.