Information Theoretic Approaches in Computational Dialectometry

4. March 2009

Groningen, Netherlands

Outline

- Information
- Entropy
- Applying Information and Entropy to Dialectometry
- Advantages and Disadvantages
- The Bulgarian Data Set
- Map(s): Information
- Map(s): Entropy
- Conclusion

Information

- The information I of an element z in a data set:

$$I(z) = -\log_2 p(z)$$

- p(z) is the probability of z
- *Example:* using a fair 6 side dice, every throw contains the information

$$I(T) = -\log_2(1/6) = 2.5849 \text{ Bit}$$

Entropy

- Entropy measures the *amount of surprise* in a data set: entropy is the relation between the information and the noise of a data set:

$$H(X) = -\sum_{i=1}^{n} p(z_i) \log_2 p(z_i)$$

- n is the number of elements in the alphabet
- $p(z_i)$ is the propability of the actual element

Applying Information and Entropy to Dialectometry

- Information theoretic approaches can only work with phonetic data: differences between dialects are expressed as pronunciation differences
- Pronunciation differences are resulting in qualitative and quantitative different uses of elements:

Dialect 1	Dialect 2
Oder	Oder
Geht	Gäht
Andere	Andere

Applying Information and Entropy to Dialectometry

- The information is calculated on the basis of the whole data set, then the percentage of every site on this amount of information is visualized
- The entropy is calculated as individual value for every site seperately and then visualized
- Unigrams are used

Advantages and Disadvantages

- + Information theoretic approaches are taking into account the whole data set at once: *aggregate method*
- + Analysis is possible on the basis of arbitrary n-grams

- - The entropy is a pure quantitative unit ➔ the order of elements doesn't play a role
- - Word borders are ignored
- - Elements with the same number of appearance are treated identical

The Bulgarian Data Set

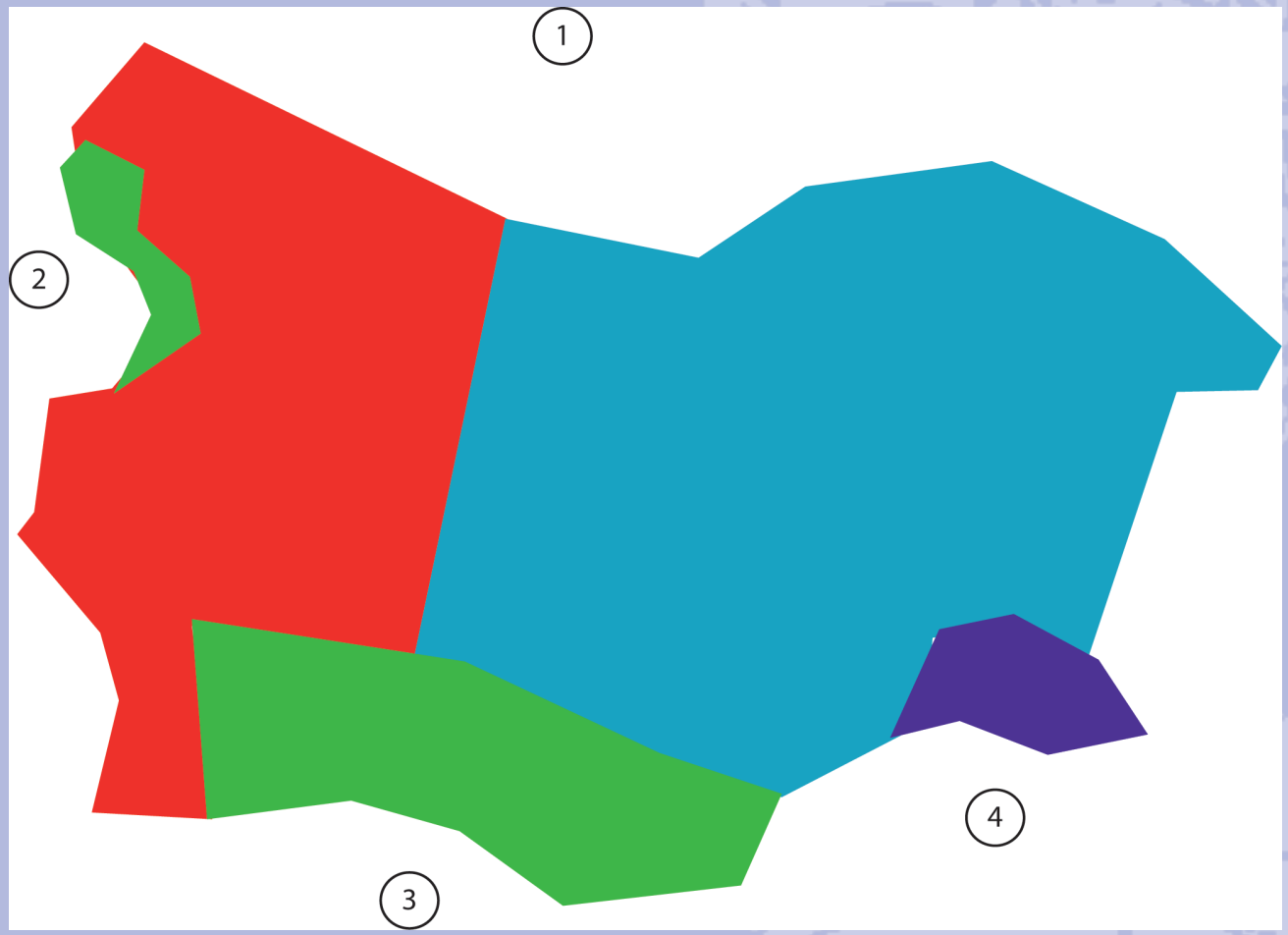
- In cooperation with the Bulgarian Academy of Science and the University of Sofia, two data sets of Bulgarian dialects have been compiled:
  - A set of phonetic data, containing 156 distinct words in 197 geographical locations (sites)
  - A set of 112 lexical lemmas, collected in the same 197 sites
  - ➔ This presentations relies on the phonetic data set, using 118 out of 156 words in 197 sites

The Bulgarian Data Set - Phonetic

- The data is encoded in XSampa, which is an electronic form of the IPA
- XML is used as container:

```
<entry id="4082-4"><key>агне</key><english>lamb</english><cform
ana="Ncnsi">lагне</cform>
<nform>`agne</nform>
<variant ana="Ncnsi">`jagne</variant>
<sampa>
<nform>"agne</nform>
<variant ana="Ncnsi">"jAgne</variant></sampa></entry>
```

Bulgarian Dialects



Analysis and Visualization with VDM

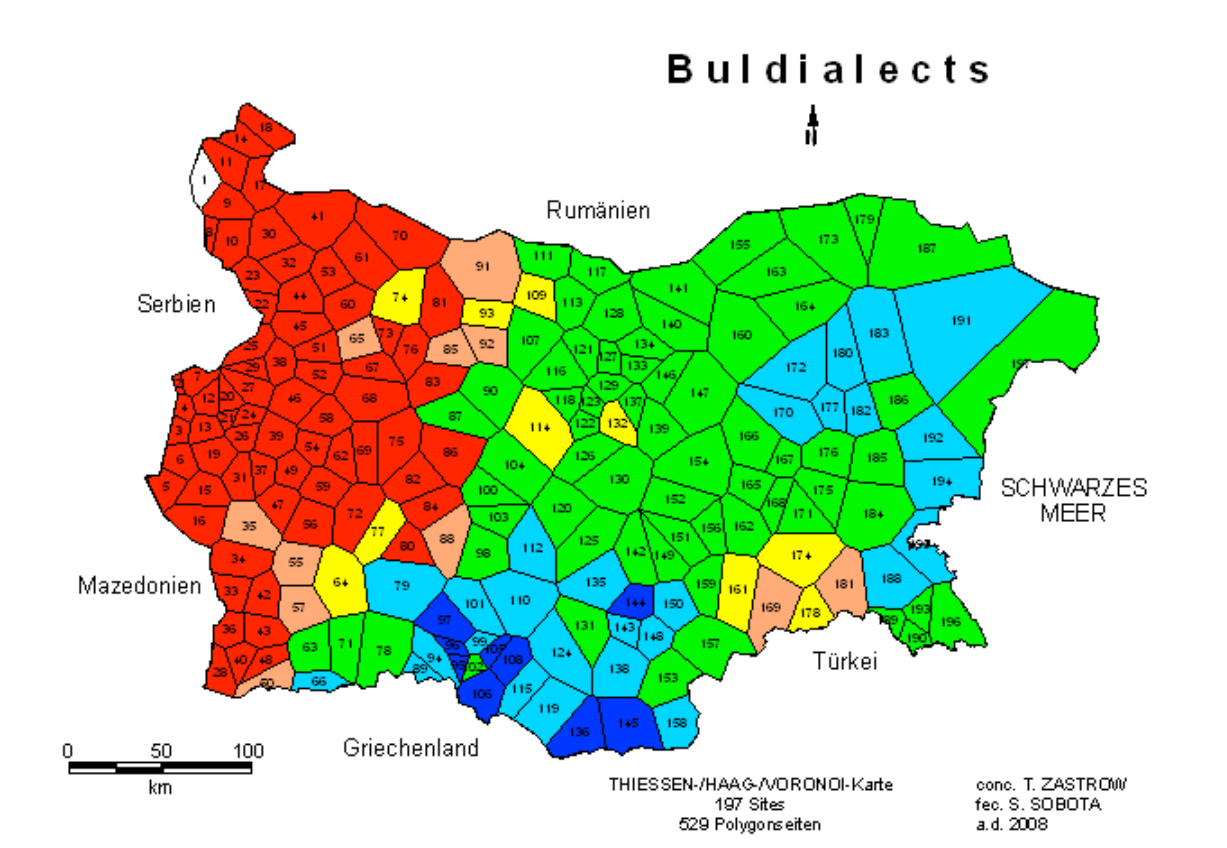
- "Visual Dialectometry", developed by Edgar Haimerl at the University Salzburg
- Initially, developed to analyse and visualize dialectometrical investigations on the basis of "Relativer Identitätswert" by Prof. Dr. Hans Goebl
- ➔ VDM takes a similarity matrix as input, so *also other data on a geographical basis can be analysed and visualized*

Analysis and Visualization with VDM

- Possible steps of analysis:
  - Classification on the basis of interval algorithms (Synopsis Map)
  - Hierarchical Clustering with several methods of distance measurement (dendrograms and maps)
  - Isogloss Map
  - Ray Map

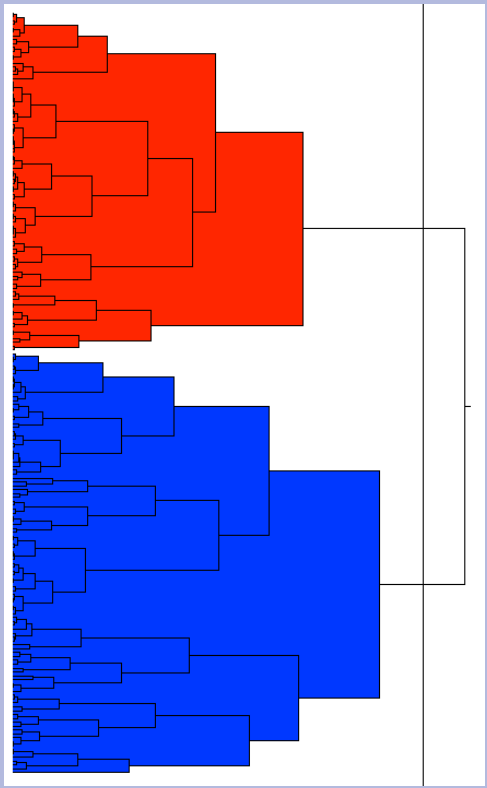
Synopsis Map

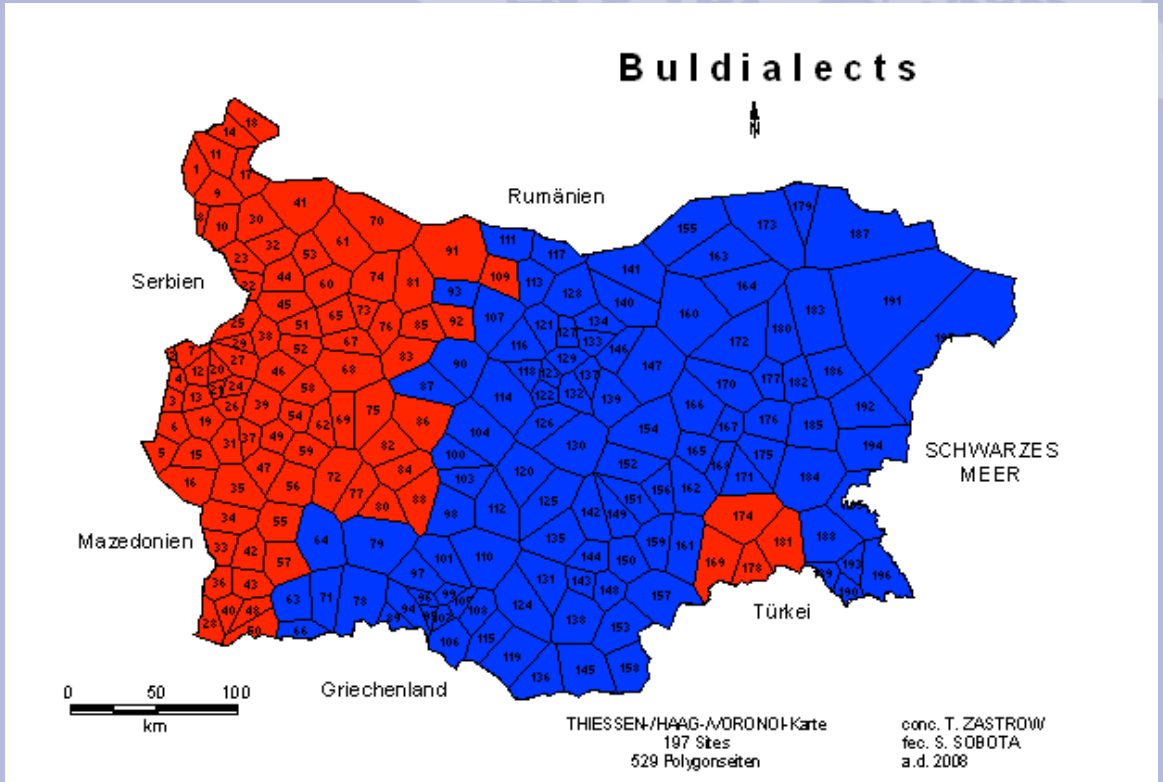
6 classes, algorithm MinMwMax, site 1 as reference point



Cluster Analysis

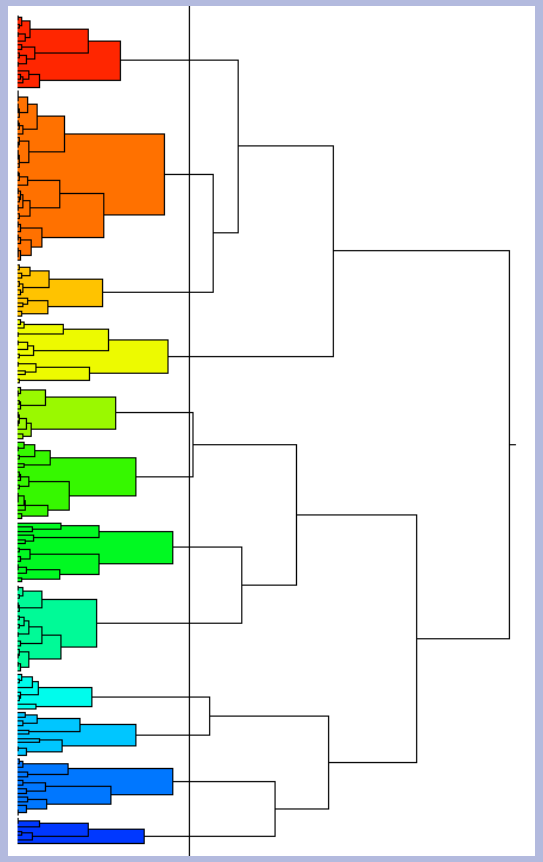
Ward method, 2 clusters

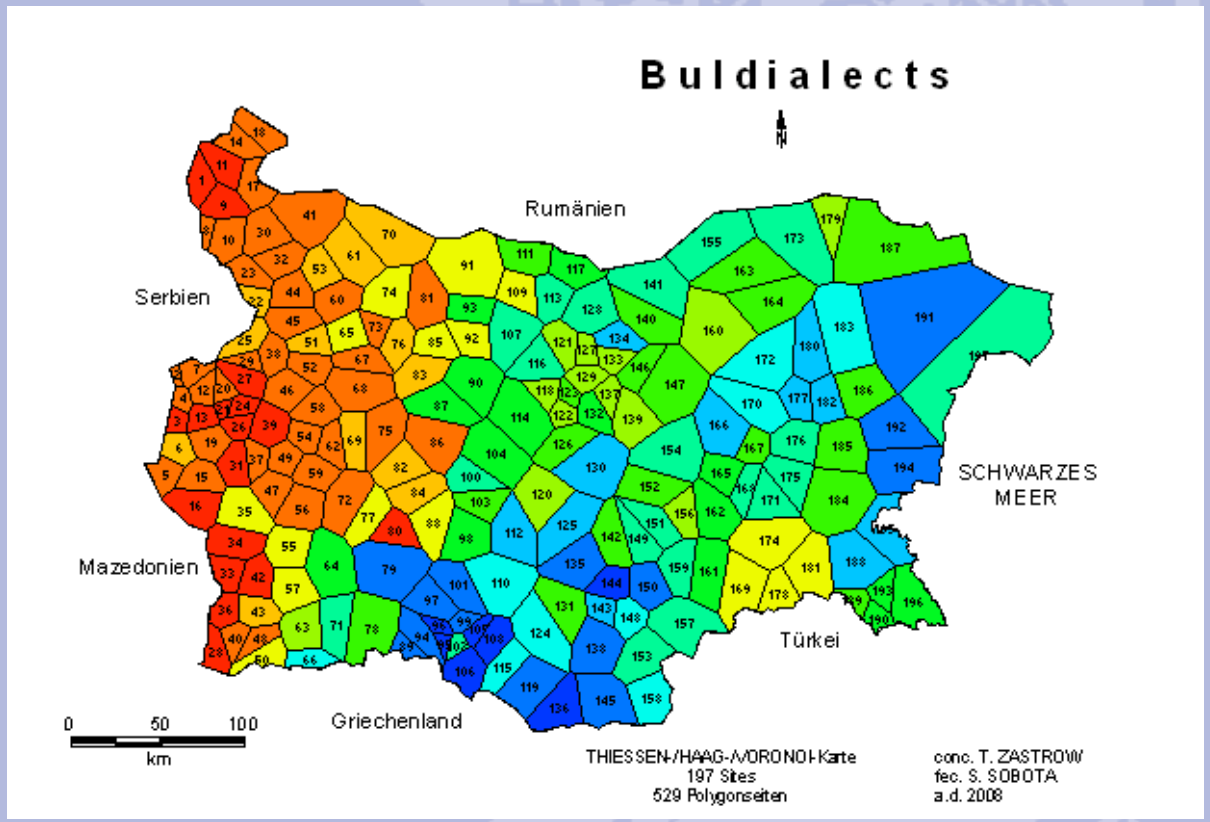




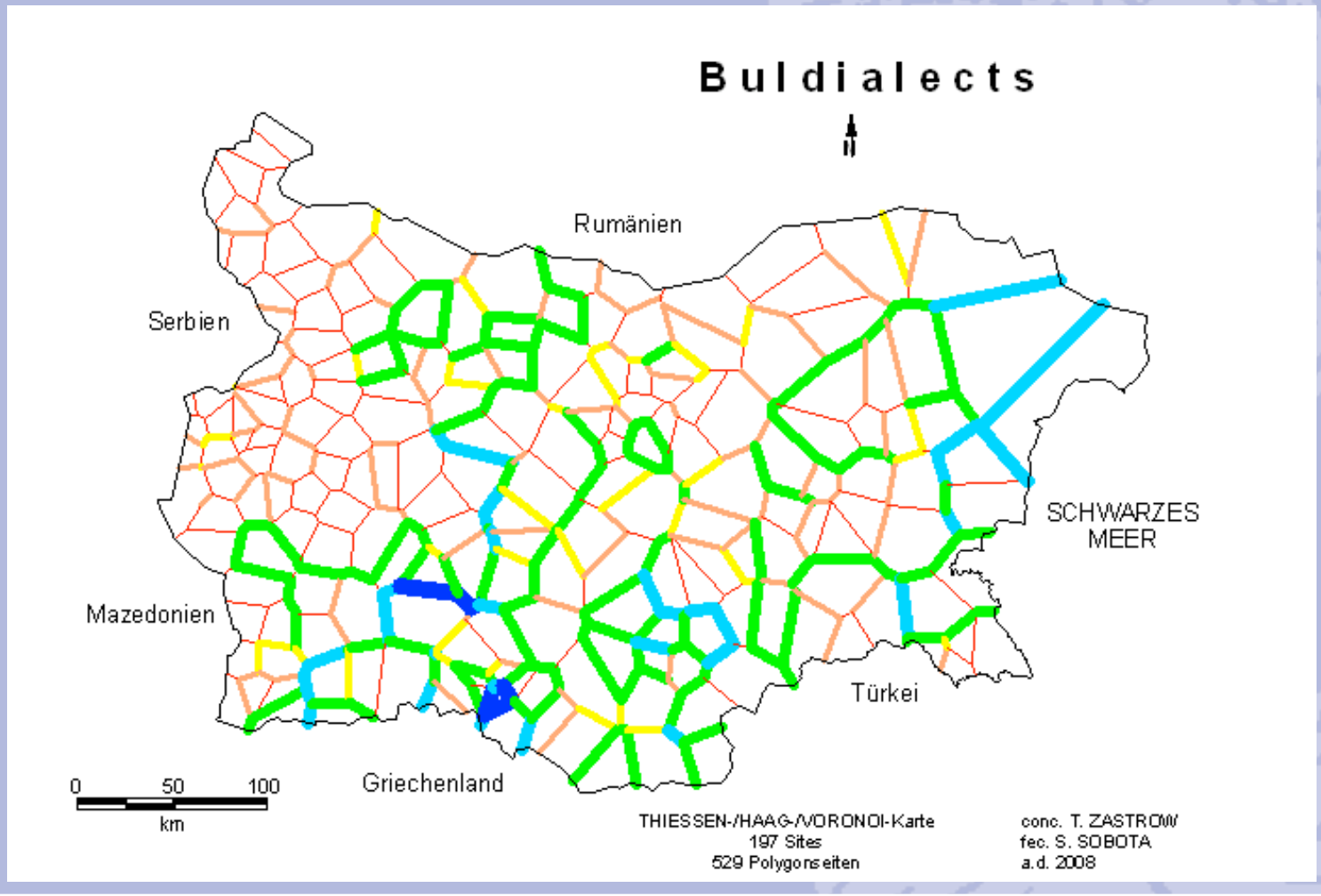
Cluster Analysis

Ward method, 12 clusters

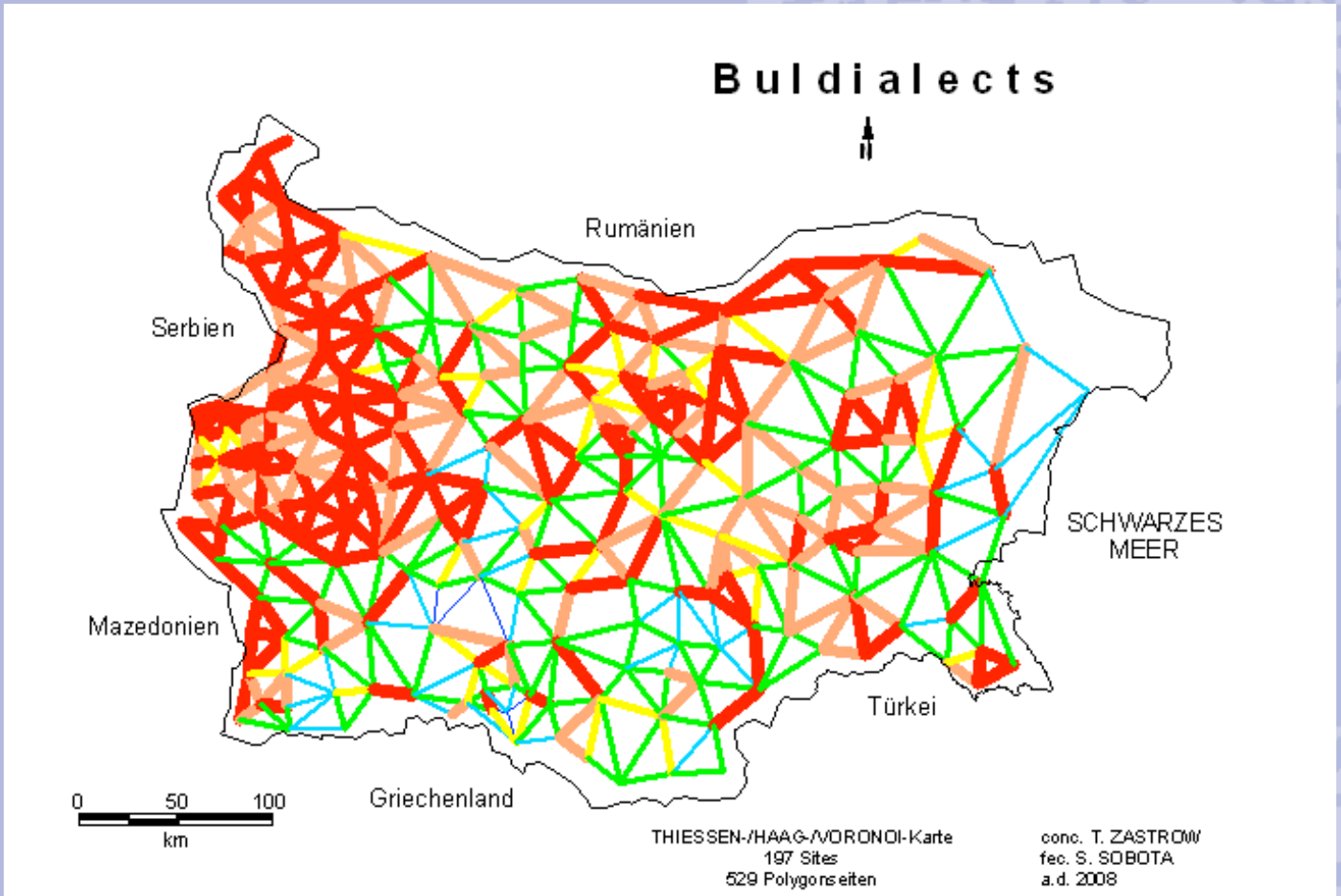




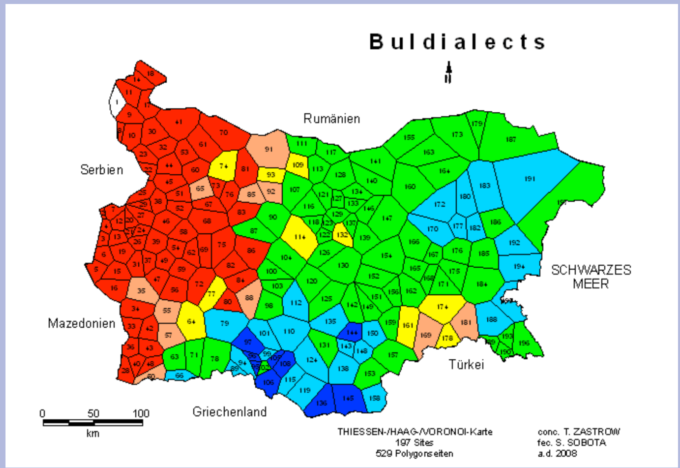
Isogloss Map

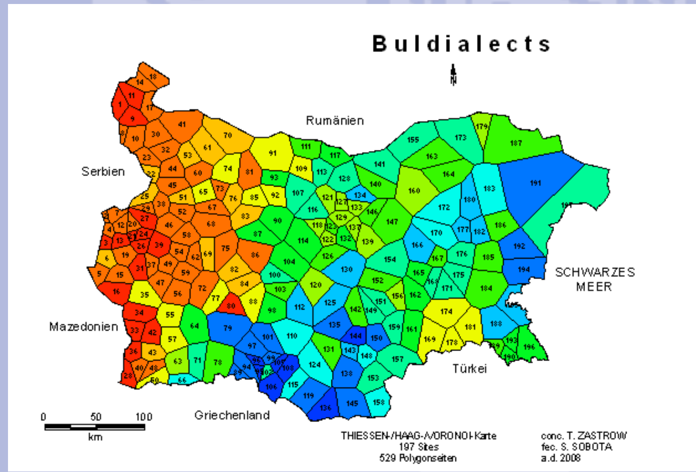


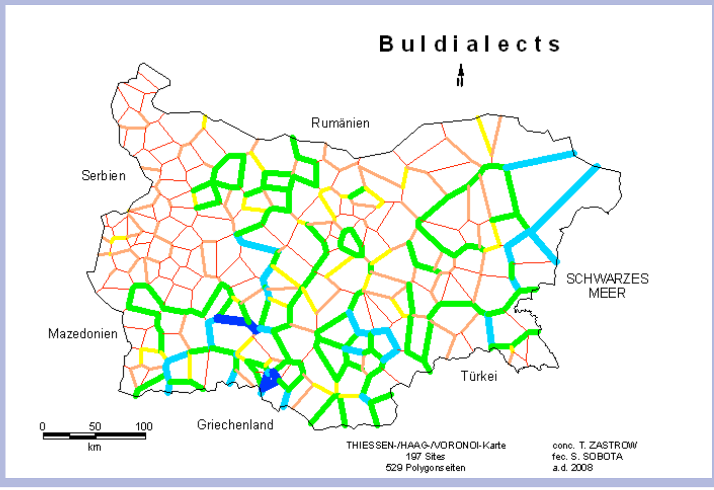
Ray Map

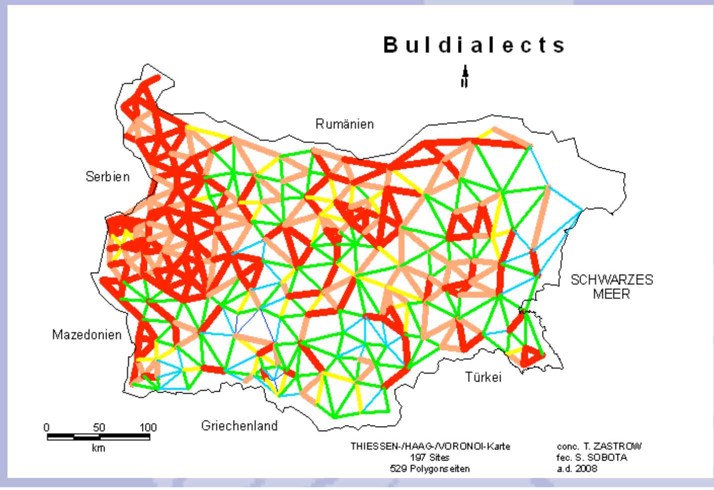

Buldialects
Rumänien
Serbien
SCHWARZES MEER
Mazedonien
Türkei
Griechenland
0 50 100
km
THIESSEN-/HAAG-/VORONOI-Karte
197 Sites
529 Polygonseiten
conc. T. ZASTROW
fec. S. SOBOTA
a.d. 2008

Overview - Information

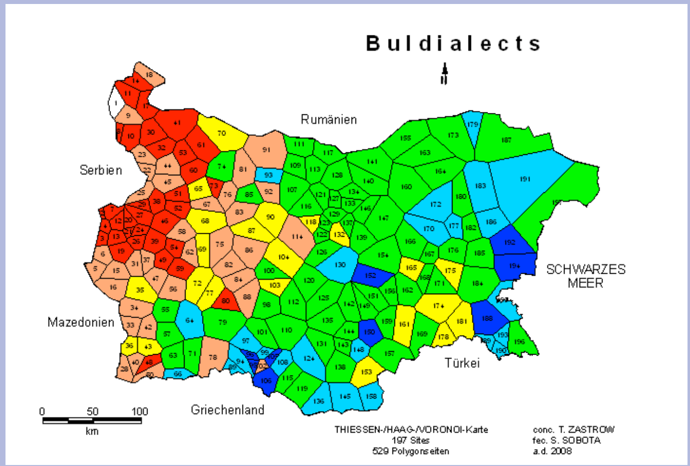


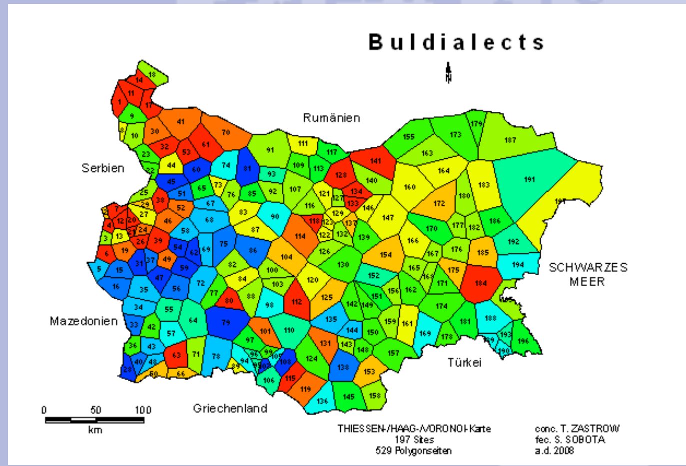


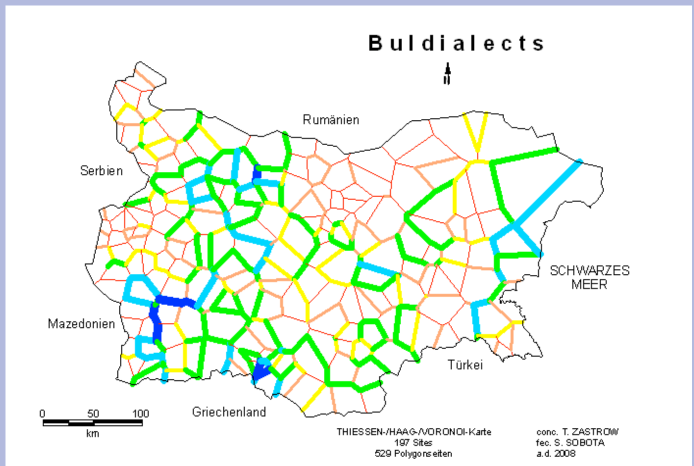


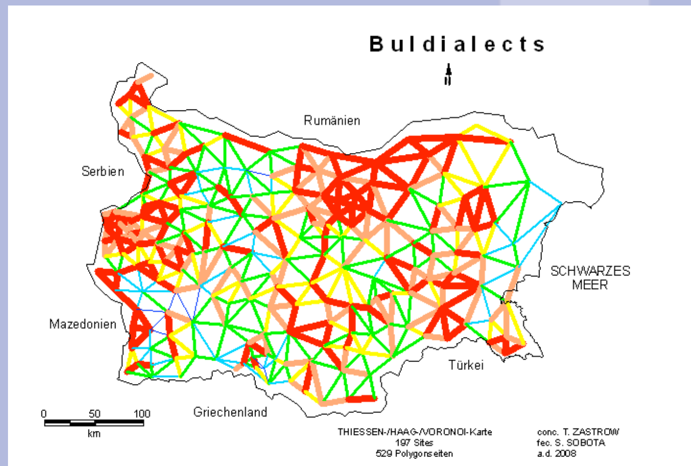


Overview - Entropy

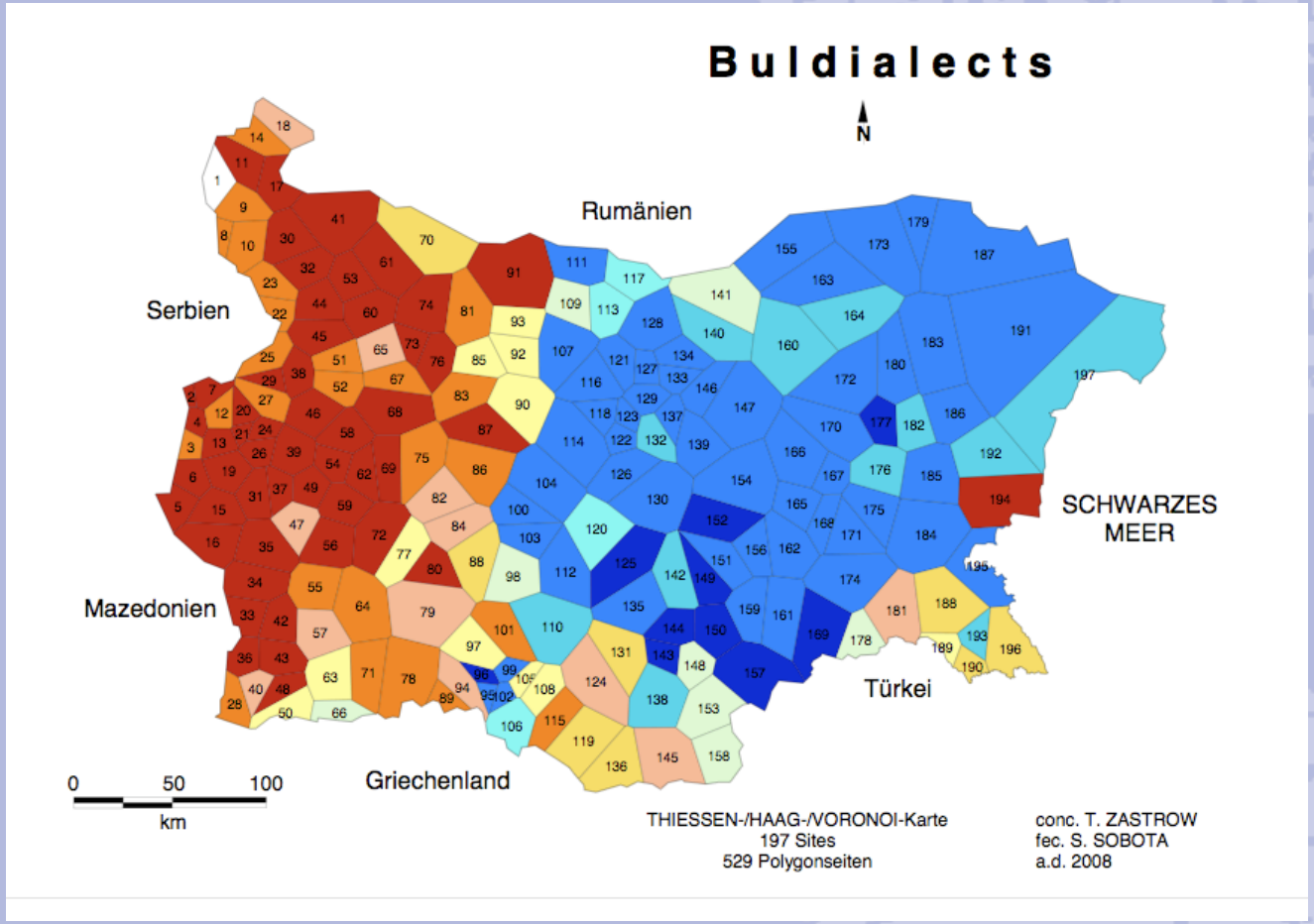








Comparison to other methods – vector analysis



Conclusions

- Despite its disadvantages, information theory based methods are able to identify the main linguistic structures between dialects
- Because of its aggregating nature, the entropy and similar methods can be used in addition to well-known methods in computational dialectometry, for example edit distance based approaches

Further Work

- Extend the analysis to n-gram models (already done for bigrams)
- Use more enhanced methods of information theory, for example conditional entropy etc. (claculating entropy values against a gold standard)
- Extensive statistical analysis of the results

Links

- http://www.thomas-zastrow.de
- http://www.sfs.uni-tuebingen.de/~dialectometry