Gabmap
— doing dialect analysis on the web

Therese Leinonen

university of groningen

Potsdam, December 7th 2010

Outline

- Background and theory
  - background
  - preparing dialect data for Gabmap
  - data inspection
  - calculation and mapping of linguistic distances
  - statistical analysis
- Hands-on exercises

Background

- R*u*G/L04: free software for dialectometrics and cartography
- developed by Peter Kleiweg, University of Groningen
- exists since 2001, has been freely distributed since 2004
- no graphical user interface = too complex for many potential users (dialectologists, variationist linguists)
- project 2010, financed by CLARIN-NL, for developing a web application of the R*u*G/L04 software $\rightarrow$ Gabmap

Dialectometry

- dialectometry = the measuring of dialects
- aims: defining dialect areas and describing dialect continua
- data-driven methods
- common statistical methods: multidimensional scaling, factor analysis, cluster analysis

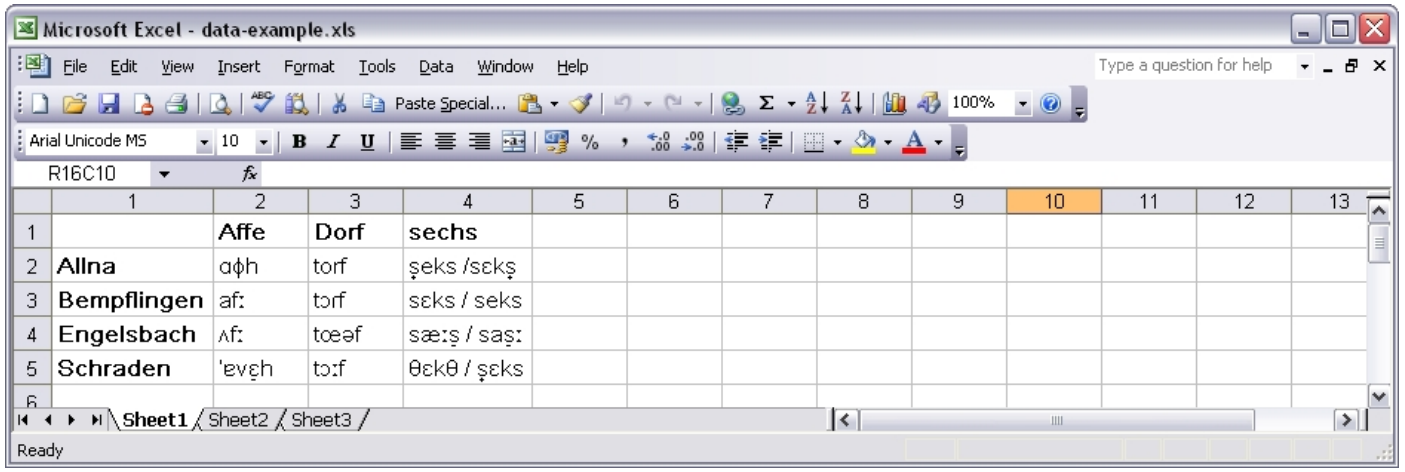
Dialect data

- data types: string data (= transcriptions), numeric data, categorical data
- input format: tab separated table (rows = sites; columns = linguistic variables)
- text file, character encoding: Unicode (UTF-8, UTF-16)
- data can be prepared for example in Microsoft Excel:
  Save as... → Unicode Text (*.txt)

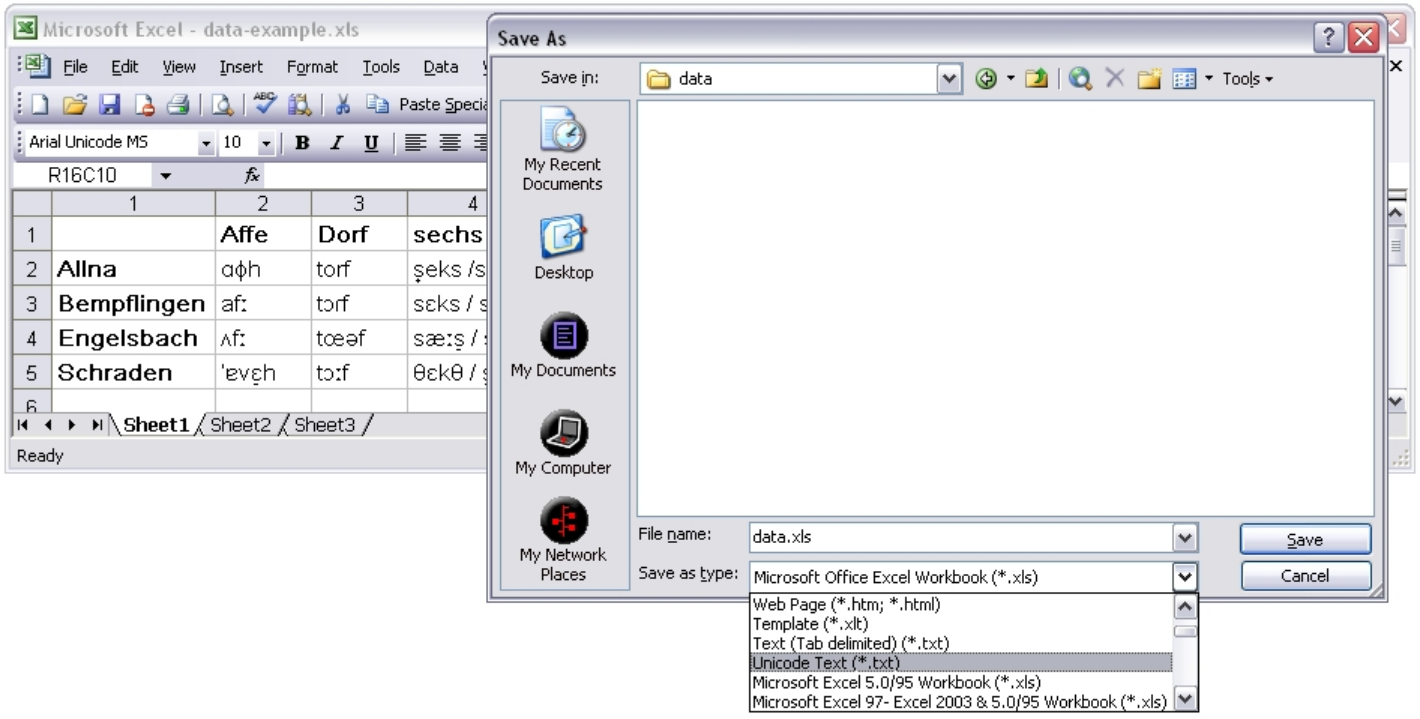
Example:

	Affe	Dorf	sechs
Allna	ɑɸh	torf	seks / sɛks
Bempflingen	afː	tɔrf	seks / sɛks
Engelsbach	ʌfː	tœəf	sæːs / sasː
Schraden	ˈɐvɛh	tɔːf	sɛks

Dialect data



	1	2	3	4
1		Affe	Dorf	sechs
2	Allna	ɑɸh	torf	ʂeks /sɛkʂ
3	Bempflingen	af:	tɔrf	sɛks / seks
4	Engelsbach	ʌf:	tœəf	sæːʂ / saʂː
5	Schraden	ˈɐvɕh	tɔːf	θɛkθ / ʂɛks
6				



Geographic data

- collect geografic data (data sites, borders) using Google Earth (http://earth.google.com/)
- save as .kml or .kmz file
- same place names in the data file as in the map file $\rightarrow$ Gabmap will automatically connect the data to the geographic space
- a number of map resources (Bantu, Bulgaria, Dutch, Germany, Pennsylvania, Norway, Swedish) available at http://www.let.rug.nl/~kleiweg/L04/Maps/

Data inspection in Gabmap

- **data overview** (number of sites, number of linguistic variables, number of characters/tokens etc.)
- **character/token list** (good way of detecting errors in the input data: infrequent character likely typos)
- **distribution maps** of items/characters/regular expressions (correspond to traditional isogloss maps)

Linguistic distances

- dialectometric analyses are applied to aggregate linguistic distances between dialects, that is, distances based on all the variables in the input table

- an appropriate distance measure (type of processing) is chosen according to the data type:

– phonetic transcriptions → string edit distance
– numeric data → Euclidean/Manhattan distance
– categorical data → binary comparison/Gewichteter Identitätswert

String edit distance (Levenshtein distance)

- calculates the smallest cost of changing one string into another
- operations: subsitutions, insertions, deletions
- cost: 1 per operation, if only a difference in diacritics 0.5

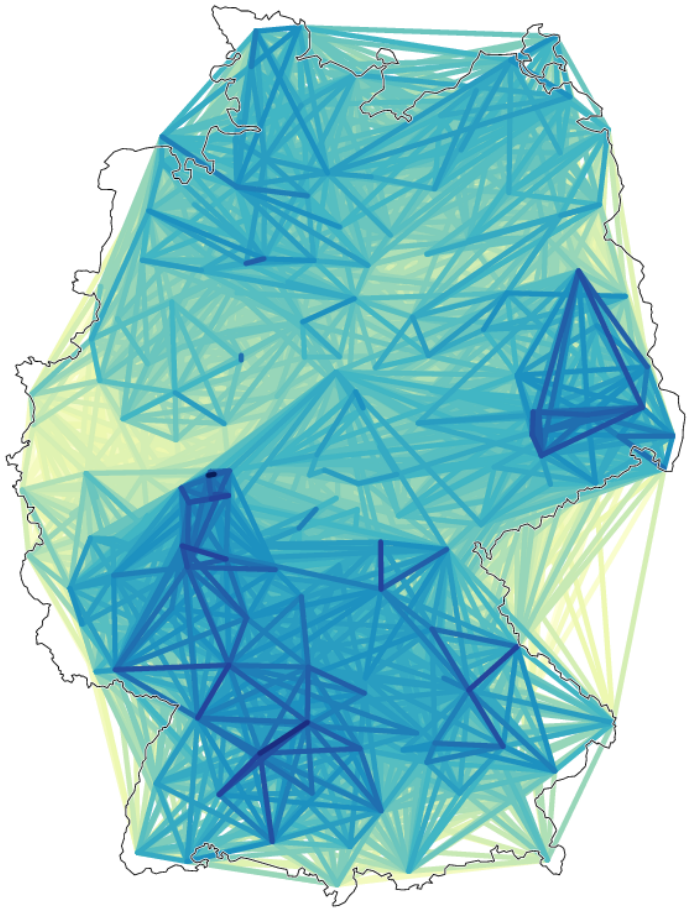
Examples:

a	fː			ʌ	fː			t	o	r	f		t	ɔ	r	f			
ʌ	fː			a	ɸ	h		t	ɔ	r	f		t	ɔː		f			
1	0	1		1	1	1	3	0	1	0	0	1	0	0.5	1	0	1.5		

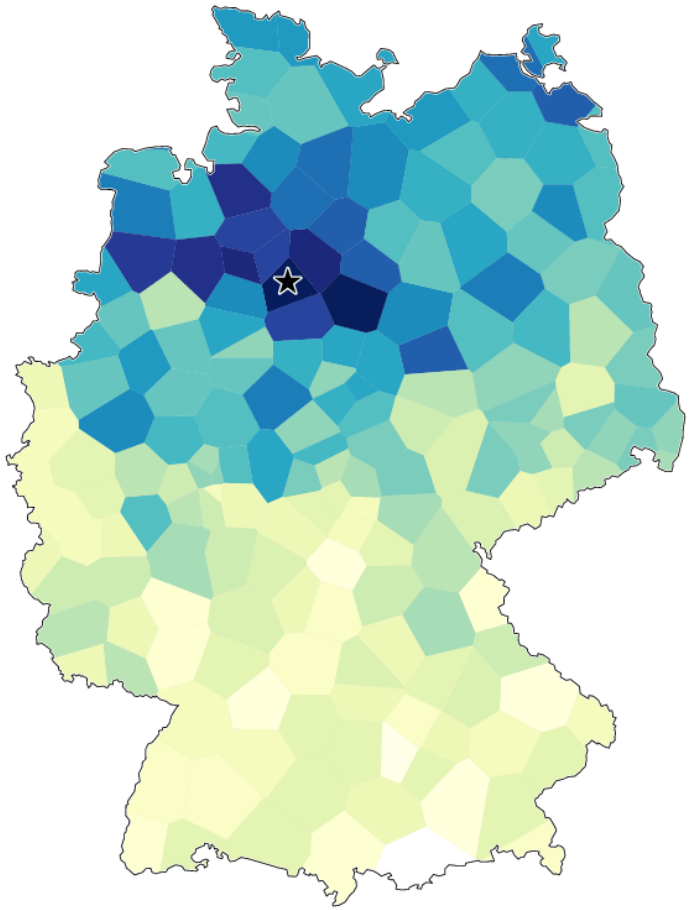
- distance computed for all words for all pairs of dialects
- distance between two dialects = average distance of all words elicited in both dialects
- all alignments can be inspected in Gabmap

Mappings of raw aggregate distances

- the darker the color the smaller the linguistic distance



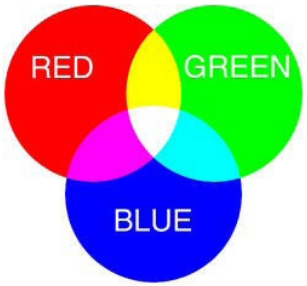
difference maps: lines drawn between locations displaying the linguistic distance

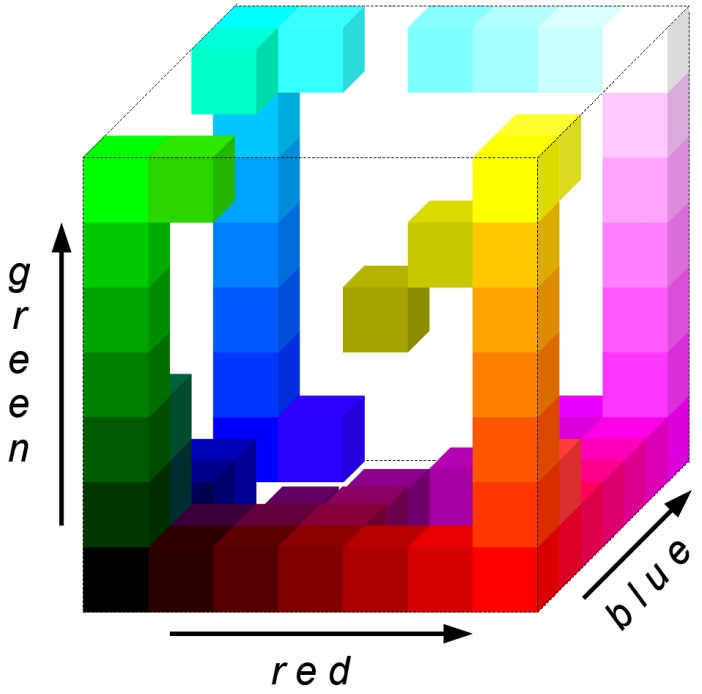


reference point maps (Goebl maps): linguistic distance from one site (star) to all other sites

Multidimensional scaling

- method for visualizing and exploring similarities/dissimilarities in data
- with given pair-wise distances positions in a low-dimensional space can be assigned to data points
- 3 dimensions visualized in red, green and blue → maps where the language varieties form a continuum



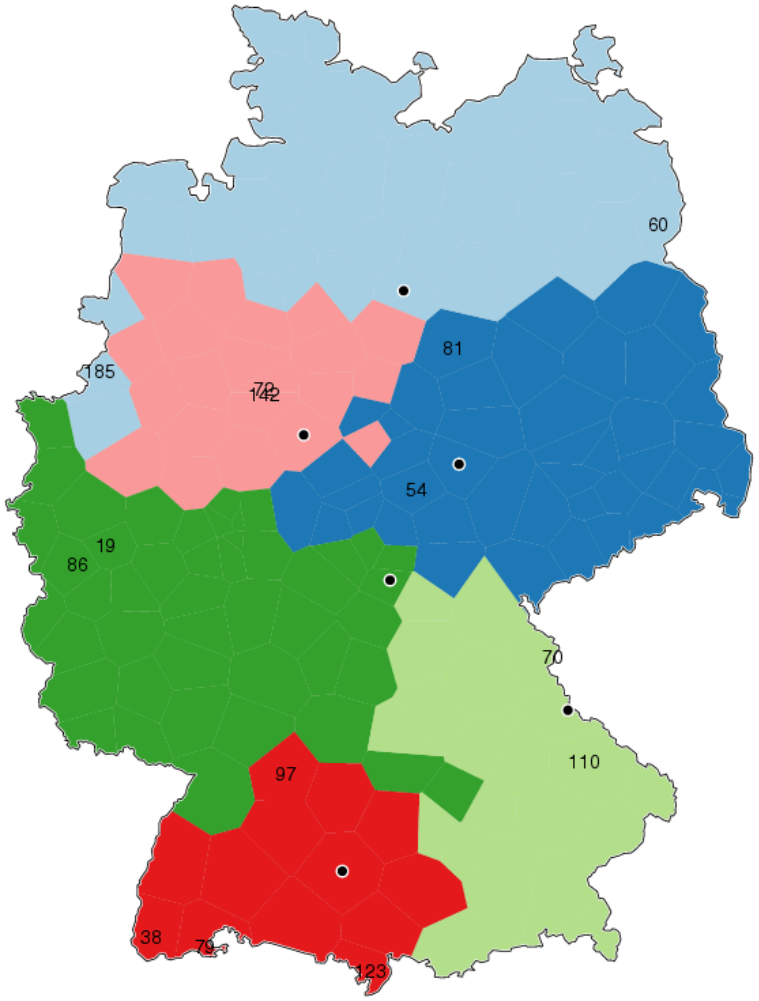


Multidimensional scaling

- MDS displays the relationships between all varieties as a continuum

Cluster analysis

- partitioning a set of objects into groups/clusters
- the most similar varieties are put in the same group → dialect classification
- less stable method than MDS: small changes in input data can lead to substantial differences in cluster division
- should be validated
- fuzzy clustering and bootstrapping can be used for obtaining more stable clusters



Gabmap

http://www.gabmap.nl/

If you have comments or questions
please mail t.leinonen@rug.nl.
We are happy to get
feedback from users!