Gabmap
– doing dialect analysis on the web

Therese Leinonen

university of groningen

Tagung des Forums Sprachvariation

Erlangen, 14.-15. Oktober 2010

Outline

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

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 → 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

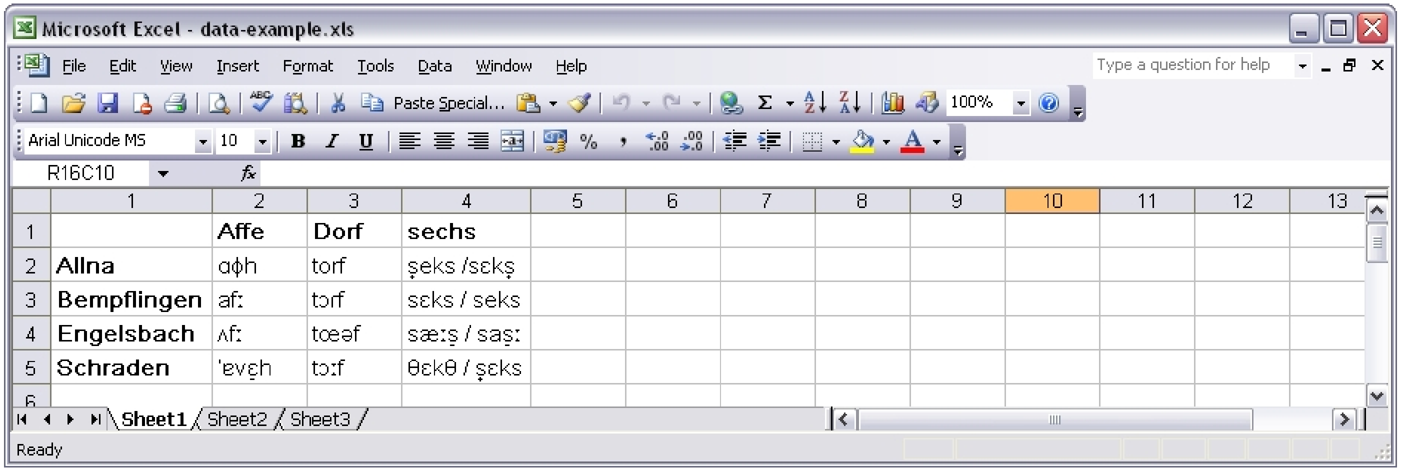
Dialect data

- data for example from dialect atlasses: phonetically transcribed lexical items
- input format: tab separated table (rows = sites; columns = words)
- text file, character encoding: Unicode (UTF-8, UTF-16)
- data can be prepared for example in Microsoft Excel:
  Save as... → Unicode Text (*.txt)

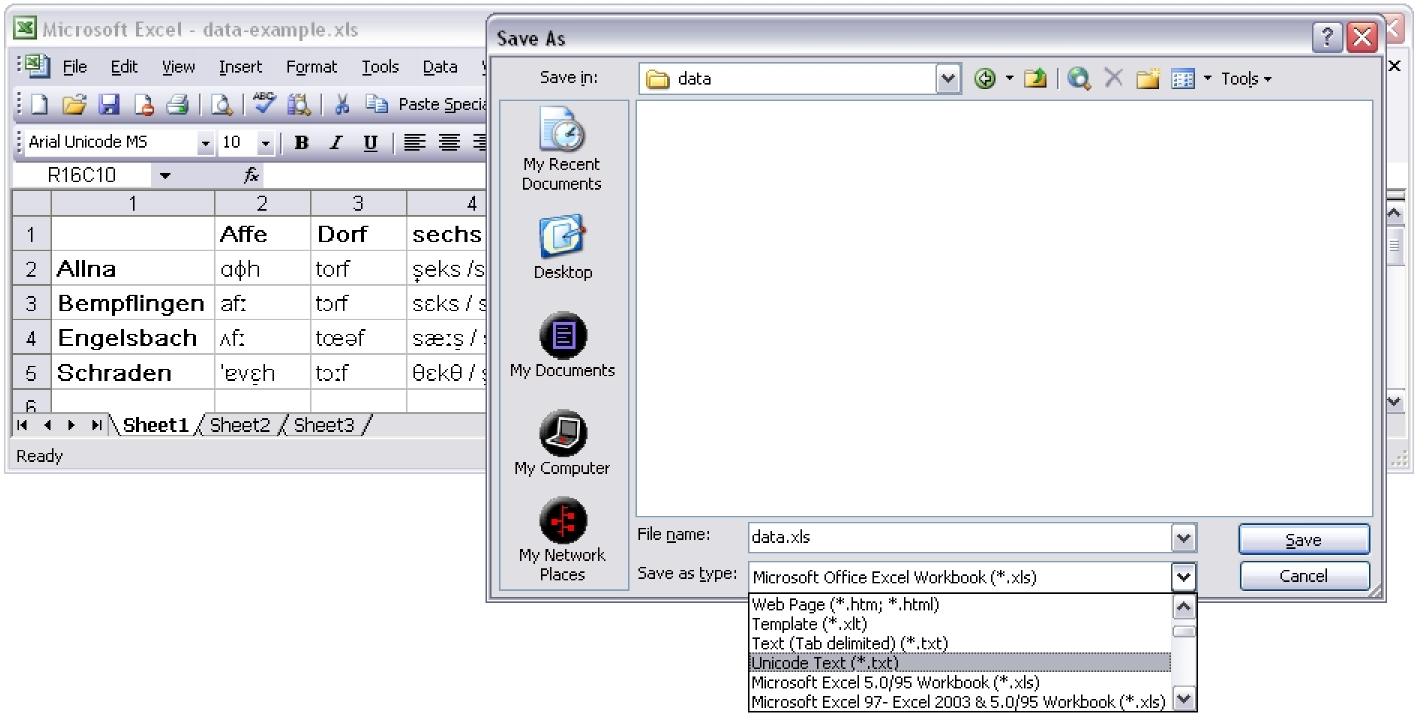
Example:

	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	s̟eks /sɛks̟
3	**Bempflingen**	af:	tɔrf	sɛks / seks
4	**Engelsbach**	ʌf:	tœəf	sæːs̠ / sas̠ː
5	**Schraden**	ˈɐvçh	tɔːf	θɛkθ / s̟ɛks
6				



Save in: data

File name: data.xls

Save as type: Microsoft Office Excel Workbook (*.xls)

- Web Page (*.htm; *.html)
- Template (*.xlt)
- Text (Tab delimited) (*.txt)
- Unicode Text (*.txt)
- Microsoft Excel 5.0/95 Workbook (*.xls)
- Microsoft Excel 97- Excel 2003 & 5.0/95 Workbook (*.xls)

Geographic data

- collect geografic data (data sites, borders) using Google Earth (http://earth.google.com/)

- save as .kml or .kmz file

- a number of map resources (Bantu, Bulgaria, Dutch, Germany, Pennsylvania, Norway, Swedish) available at http://www.let.rug.nl/~kleiweg/L04/Maps/

Data inspection

- **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)

String edit distance (Levenshtein distance)

- calculates the smallest cost of changing one string into another
- operations: subsitutions, insertions, deletions
- cost: 1 per operation

Examples:

a	fː
ʌ	fː
1	1

ʌ	fː	
a	ɸ	h
1	2	3

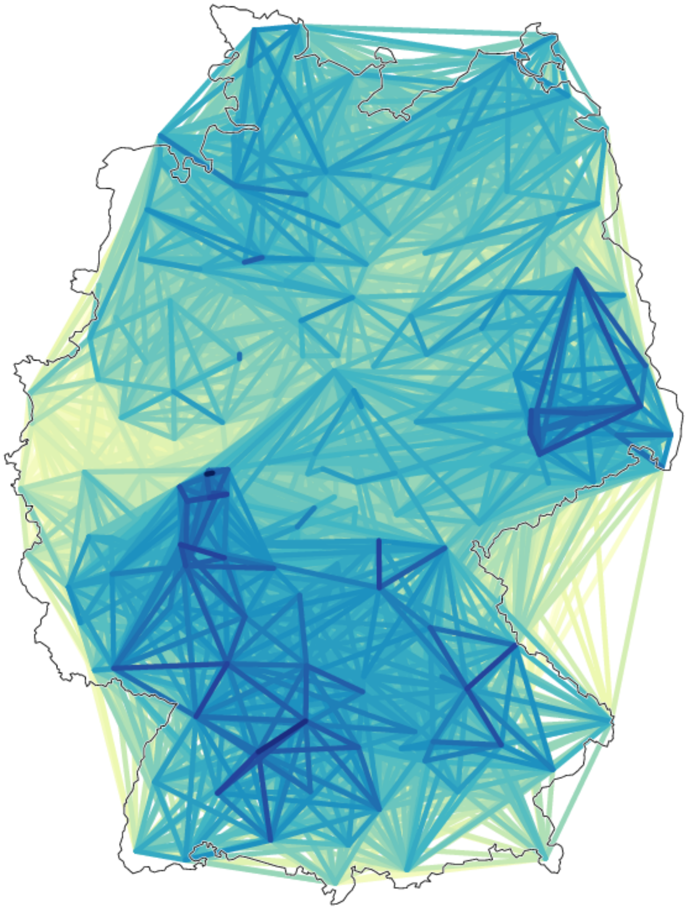
t	o	r	f
t	ɔ	r	f
0	1	1	1

t	o	r	f
t	ɔː		f
0	1	2	2

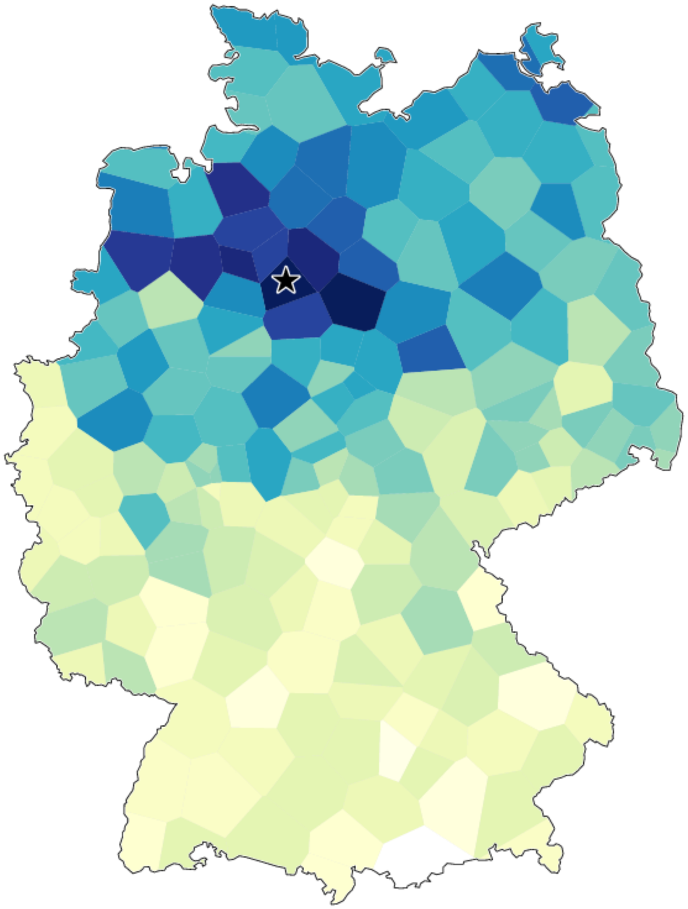
- distance computed for all words for all pairs of dialects
- distance between two dialects = average distance of all words elicited in both dialects
- all alignments can be inspected in Gabmap

Mappings of raw aggregate distances

- the darker the color the smaller the linguistic distance



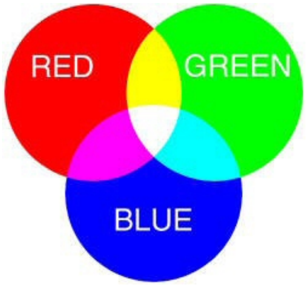
difference maps: lines drawn between locations displaying the linguistic distance

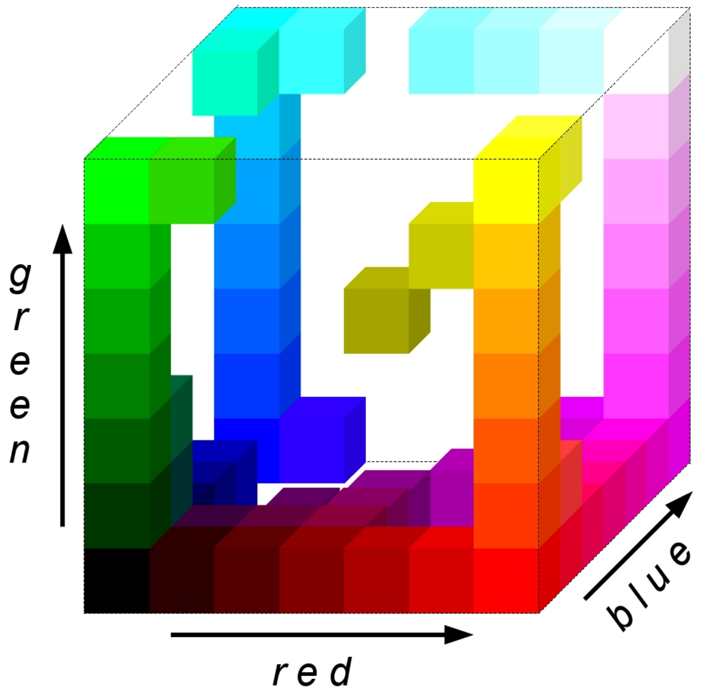


reference point maps (Goebl maps): linguistic distance from one site (star) to all other sites

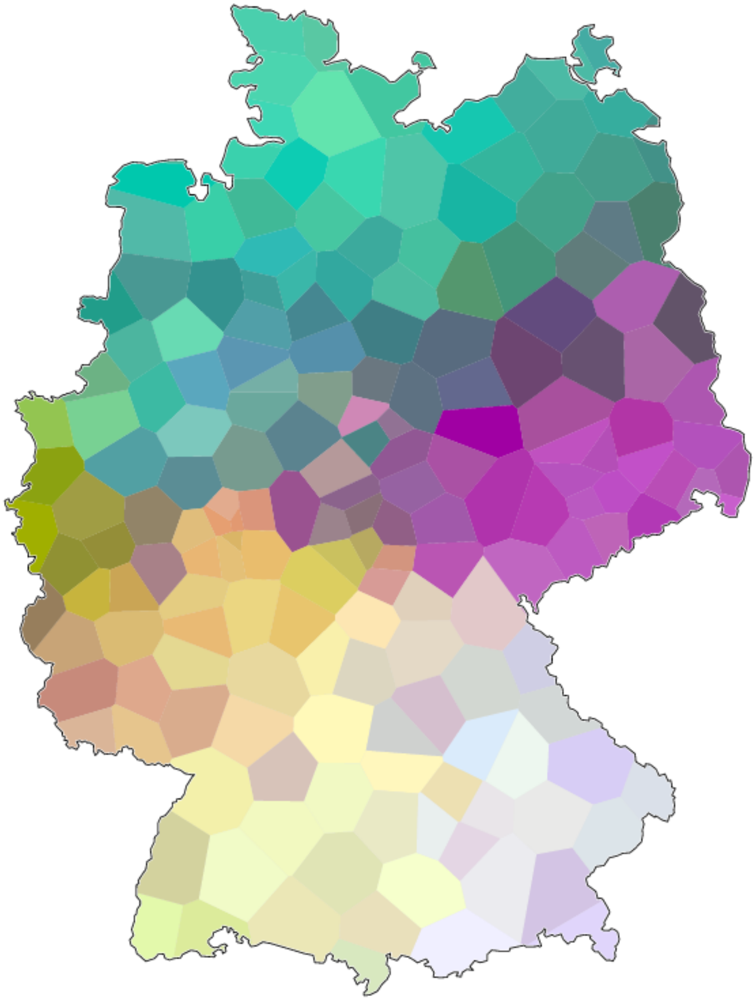
Multidimensional scaling

- 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

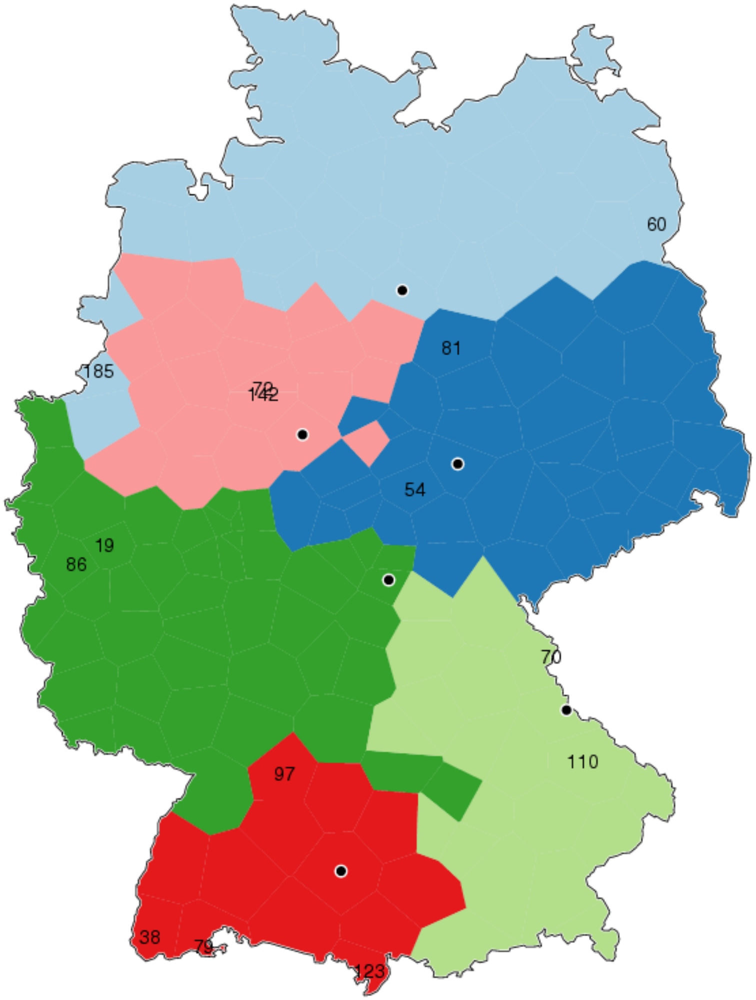
Cluster analysis

- 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

- noisy clustering and bootstrapping can be used for obtaining more stable clusters



Gabmap

http://tadept01.meertens.knaw.nl/

Under Construction!

Changes and additions will still be made, but the application is already now available to users.

If you have feedback please mail it to t.leinonen@rug.nl. We are happy to get any comments or suggestions!

Exercises

1. Some of the variants of the word *Georgia* occur only once, while others are frequent. Look at some of the most frequent variants of this variable. Can you identify any geographic areas? Which linguistic features are distinguishing for the different areas?

2. Look at some of the alignments of the words *first* and *hearth*. How does the realization of *r* influence the measured linguistic distances?

3. Compare the two difference maps of Pennsylvania. In which areas are there large dialect distances? In which areas are the dialects very similar to each other? Are there abrupt dialect borders?

4. Look at the MDS map of Pennsylvania. Are there any abrupt dialect borders? Are there transitional borders? Can clear dialect groups be identified?

5. Compare clustering using Ward's Method and Group Average. How are the maps different? What is similar? Which cluster map corresponds better to the mds map?