Dialektgeografiska kartor i Gabmap

Workshop vid Svenska litteratursällskapet i Finland r.f.

Helsingfors, den 23 och 25 november 2011

Therese Leinonen

Kursinnehåll

- introduktion och bakgrund
- dataformat och datagranskning i Gabmap
- utbredningskartor
- tillverkning av baskarta i Google Earth
- dialektometriska metoder

Bakgrund

- R*u*G/L04: gratis programvara för dialektometri och kartografi
- utvecklad av Peter Kleiweg, universitetet i Groningen
- har funnits sedan 2001, fritt tillgängligt sedan 2004
- i den tidiga unixbaserade versionen av programvaran finns inget grafiskt användargränssnitt = komplicerat att använda för många potentiella användare
- 2010: CLARIN-NL financierat projekt för att utveckla en webbapplikation av R*u*G/L04 programvaran $\rightarrow$ Gabmap

Vad behövs för att tillverka dialektkartor?

- Gabmap är en webbapplikation = behöver inte installeras på datorn utan används online i en webbläsare

- lösenordsskyddat användarkonto där upp till 20 olika projekt kan lagras

- olika typer av material kan analyseras i Gabmap:
  - transkriberat dialektmaterial
  - kategoriska variabler (t.ex. morfologiska eller syntaktiska variabler)
  - numeriska data (t.ex. akustiska mätningar)

- materialet laddas upp i Gabmap och analyseras online

- två filer måste laddas upp:
  1) fil med dialektmaterialet
  2) baskarta

- de färdiga dialektkartorna laddas ner till den egna datorn i form av bildfiler (.eps, .png, .pdf)

Dataformat för dialektmaterialet

- tab-separerad tabell (rader = orter; kolumner = lingvistiska variabler)
- textfil kodad i Unicode (UTF-8, UTF-16), fältavgränsare: tabulator
- vissa formateringsfel hittar Gabmap automatiskt när man laddar upp materialet (t.ex. tabellrader av olika längd)
- materialet kan sammanställas t.ex. i Microsoft Excel:
  (Spara som... → Unicode Text (*.txt))

Exempel:

	baka	lamm	kväll
Bergö	baːk	lamb	kveːld / kvieːld
Björkö	baːk	laːmb	kveːld / kvieːld
Borgå	bak(a)	laːmb	kveld
Bromarv	bak(a)	lamm	kveːld / kvell

Baskarta

- fil med geografiska koordinater (orter, gränser) kan skapas i Google Earth (http://earth.google.com/)
- spara som .kml- eller .kmz-fil
- dialektmaterialet kopplas till de geografiska koordinaterna genom att samma ortnamn används i båda filerna

Datagranskning i Gabmap

- *index*: orter, antal belägg per variabel
- *data overview*: antal orter, antal variabler, antal tecken i materialfilen etc.
- *character/token list*: antal belägg per tecken (kan användas för att hitta fel i materialfilen: t.ex. infrekventa tecken är ofta tryckfel)
- distributionskartor över enskilda fonetiska tecken


SLS

Utbredningskartor

- utbredningskartorna (*distribution maps*) i Gabmap visar **frekvensen** av en specifik variant

- ju mörkare färg på kartan, desto mer frekvent förekommer varianten på orten i fråga

- Gabmap är i första hand utvecklat för dialektometrisk analys
  = funktionerna för utbredningskartor inte fullt utbyggda

- klassiska utbredningskartor, som visar förekomsten av flera olika varianter på en och samma karta (t.ex. med olika symboler), finns inte direkt inbyggda i Gabmap – sådana kartor kan åstadkommas genom att lägga upp ett projekt med en enda språklig variabel och använda funktionen för klusterkartor (*cluster maps*)


SLS

Dialektometri

- dialektometri = mätning av dialekter
- mål: beskriva dialektkontinuum och klassificera dialekter
- i traditionell dialektologi har isoglosser använts för att identifiera dialektgränser och dialektområden
- inom dialektometrin vill man dels skapa en helhetsbild (dvs. inte beskriva enskilda språkdrag), dels använda mer objektiva metoder
- datadrivet tillvägagångssätt
- dialektometrisk analys omfattar ofta två steg:
  1) mätning av språkligt avstånd mellan dialekterna
  2) statistisk analys av de uppmätta avstånden

Levenshteinavstånd (string edit distance)

- algoritm för att räkna ut det lingvistiska avståndet mellan dialekter utifrån transkriberade ord
- räknar ut det minsta antalet operationer (= raderingar, infogningar och substitutioner) som krävs för att transformera en teckensträng till en annan
- kostnader: 1 per operation, 0.5 om den enda skillnaden är i diakriser

Exempel:

b	e	i	n	
b	ä	i	n	
0	1	0	0	1

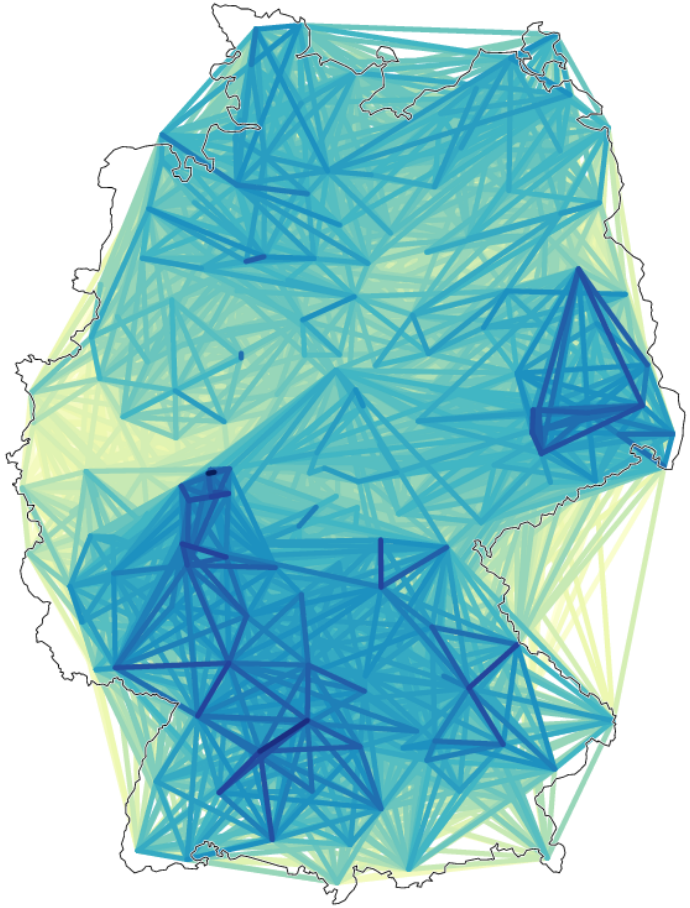
a	r	g			
a	r	g	ä	r	
0	0	0	1	1	2

k	v	e	l	l	
k	v	ä	l’	l’	
0	0	1	0.5	0.5	2

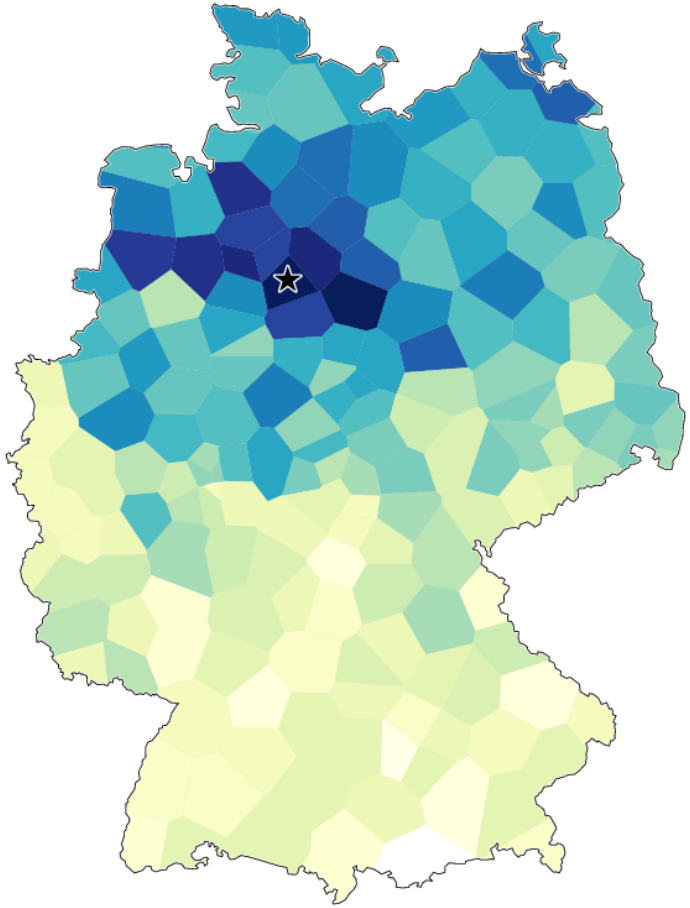
- avståndet räknas ut för alla ord och alla dialektpar
- det språkliga avståndet mellan två dialekter är medeltalet av avståndet för alla ord som finns belagda i båda dialekterna
- länkningarna kan inspekteras Gabmap

Avståndskartor

- ju mörkare färg, desto mindre språkligt avstånd



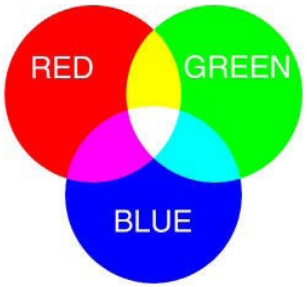
difference maps: linjer dragna mellan orterna parvis anger det språkliga avståndet

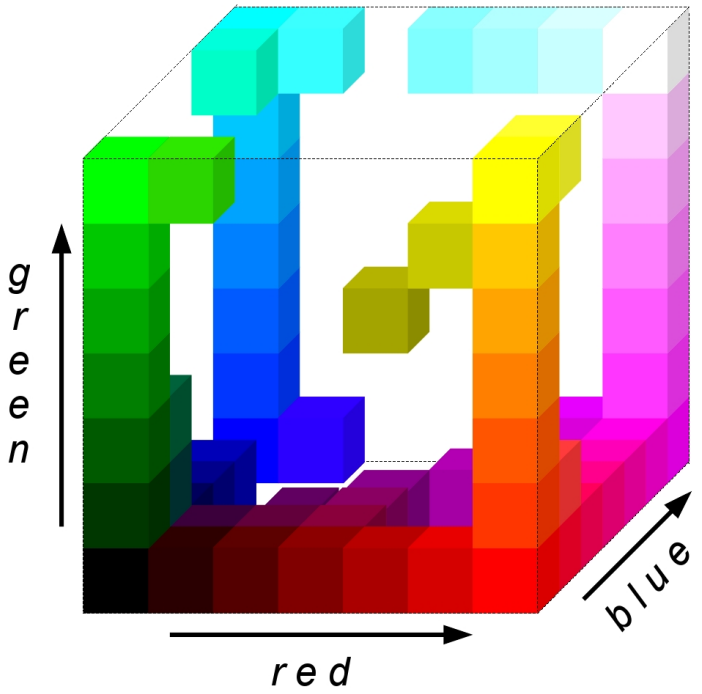


reference point maps: språkligt avstånd från en utvald ort (stjärna) till alla andra orter

Dialektkontinuum

- multidimensionell skalering (MDS) används för att reducera komplexa avståndsmatriser till ett mindre antal dimensioner
- metod för att visualisera likheter/olikheter i data som ett kontinuum
- utifrån parvis uppmätta avstånd mellan dialekter tilldelas varje dialekt en position i ett lågdimensionellt rum
- resultaten från MDS kan överföras till kartor med hjälp av RGB färgmodellen



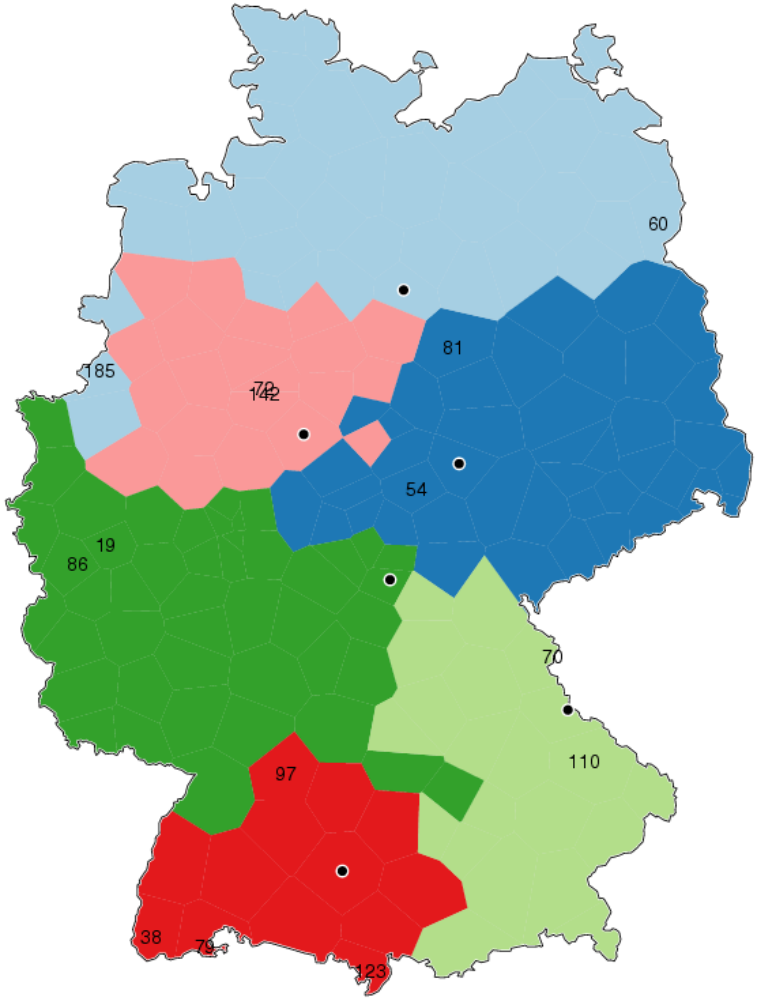


Multidimensionell skalering

- MDS visar den språkliga variationen som ett kontinuum

Klusteranalys

- delar in dataobjekt i grupper/kluster
- de dialekter som liknar varandra mest hamnar i samma grupp $\rightarrow$ dialektindelning
- denna metod är inte lika stabil som MDS: små förändringar i materialet som analyseras kan leda till stora skillnader i dialektindelningen
- kräver validering (t.ex. jämförelse med MDS, jämförelse mellan olika klustermetoder, *fuzzy clustering*)



Litteratur

Heeringa, Wilbert (2004): Measuring dialect pronunciation differences using Levenshtein distance. Doktorsavhandling, University of Groningen.

Leinonen, Therese (u.u.): Indelning av finlandssvenska dialekter med särskild hänsyn till Åboland. I: *Folkmålsstudier* 50.

Nerbonne, John (2009). Data-driven dialectology. *Language and Linguistics Compass* 3(1), 175−198.

Nerbonne, John (2010). Mapping aggregate variation. I: A. Lameli, R. Kehrein och S. Rabanus (red.), *Language and Space Vol. 2.1. Language Mapping*. Berlin: De Gruyter, 476−495.

Nerbonne, John och Wilbert Heeringa (2010): Measuring dialect differences. I: J. E. Schmidt och P. Auer (red.), *Language and Space Vol. 1. Theories and Methods*. Berlin: De Gruyter, 550−567.

Nerbonne, John, Rinke Colen, Charlotte Gooskens, Peter Kleiweg och Therese Leinonen (2011): Gabmap − A Web Application for Dialectology. *Dialectologia*, Special Issue II, 65−89.

Mera övningar finns tillgängliga online:
http://www.gabmap.nl/~app/doc/tutorial/