Analyzing Dialect/Varietal Distances

John Nerbonne
j.nerbonne@rug.nl

Center for Language and Cognition, University of Groningen

Measuring Pronunciation Diferences
LOT Winter School
University of Tilburg, 12 Jan 2012

Groningen dialectology team!

Charlotte Gooskens, Peter Houtzagers, Hermann Niebaum, Wilbert Heeringa, Jelena Prokic, Therese Leinonen, Martijn Wieling, Marco Spruit, Peter Kleiweg, Christine Siedle, Jens Moberg, ...

...

Sebastian Kürschner, Alexandra Lenz, Bob Shackleton, Renée van Bezooijen, ...

...

Bob de Jonge, Agnes de Bie, Cornelius Hasselblatt

...

Simonetta Montemagni, Franz Manni, Petja Osenova, Esteve Valls, Lucija Šimičić, Kristel Uiboaed, Boudewijn van den Berg

Overview

- Why apply a string distance measure in dialectolgy?
  - Massive variation (seen categorically)
- Why measure in an aggregate way?
  - Counterindicating signals
- Aggregating signals (dialectometry)
  - Levenshtein distance
- Analyzing aggregate measurements
  - MDS
  - Clustering
- Dialectological law enabled by aggregate view
  - Séguy's curve
- Features, "ranking isoglosses" (Chambers & Trudgill, p.97)

One old problem in dialectology

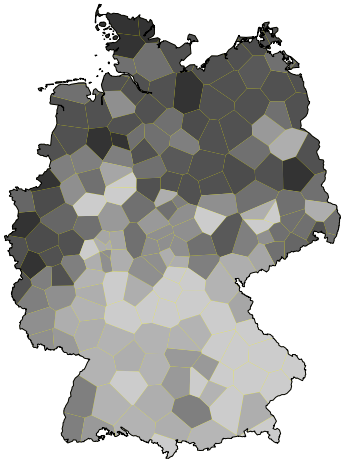
- Pronunciations are *very* variable
  — 87 different pronunciations of *ich* in the PAD

ɨ	ɐɪç	ɐɪç	ɐ̲ɪç	ʕɪ̰k	ʕɪk	əɪʃ	ə̃ig	ç	ɛɪ̲ʃ	ɛç̃k	ɛ̰g	ɛ̤ɪç̟	ɛ̤ɪ̲ʃ
ɛ̤ɪk	ɛk	ɛkʰ̃	ɪ	ɪː	ɪʔ	ɪç	ɪç̟	ɪç̲	ɪɣ	ɪɣ̟	ɪʃ	ɪʃ̟	ɪ̲ʃ
ɪ̲ç	ɪ̲ç	ɪ̲ɣ	ɪ̲g	ɪ̲k	ɪ̲k.	ɪ̲ɕ	ɪ̲ʑ	ɪ̰k	ɪ̰ç	ɪ̰g	ɪ̰g.	ɪ̰j	ɪ̰k
ɪ̤ɕ	ɪ̤x	ɪ̤	ɪ̤ç	ɪ̤ː	ɪ̤ːç	ɪ̤ç	ɪ̤χ	ɪ̤g	ɪ̤g.	ɪ̤k	ɪ̤ɕ	ɪ̤ʑ	ɪg
ɪj	ɪj̃	ɪk	ɪkʰ	ɪɕ	ɪx̟	ʏç	ʏʑ	e	e̲ɪɣ	ɐ̰ʔk	ɐ̰ç	ɐ̰g	ɐ̤ʃ̃
ɐ̤ç͡j	ɐ̤ç	ɐ̤ɣ	ɐ̤g	ɐ̤j	ɐ̤ɕ	eg	ek	ekx̃	i	iː	iːç	iːç̃	iç
ḭ	ḭːj͡ç	ḭk											

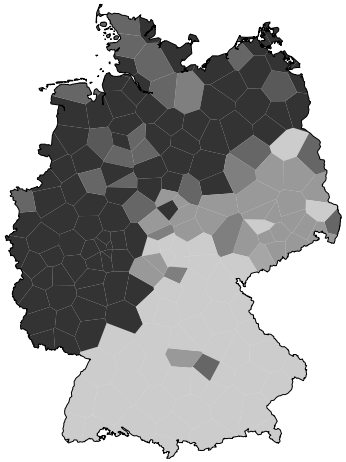
- In fact *all* analyses abstract from the recorded, observed variation.
- Relevance here: measuring sequence distance is a similar step in abstraction

A second old problem in dialectology

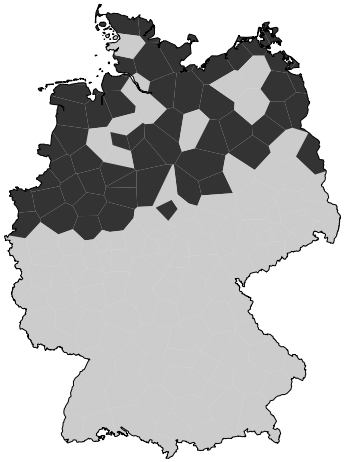
- We receive *noisy* signals of provenance.



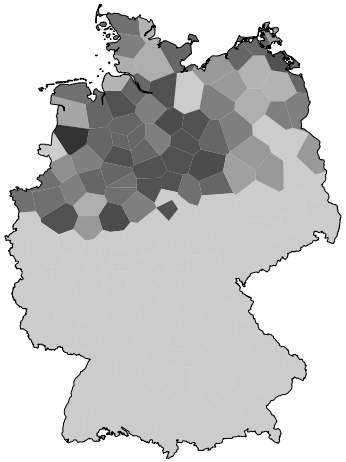
front/low V in *Haus*



[p] (dark) vs. [p͡f]



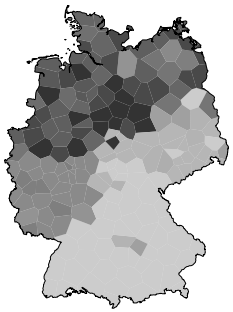
[t] vs. [t͡s]



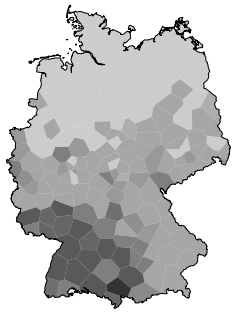
[k] vs. [x(ç)]

“non-overlapping isoglosses”

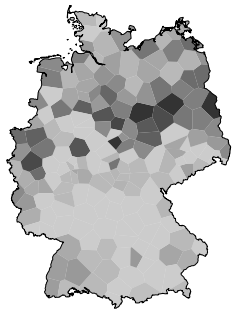
Isoglosses seldom overlap



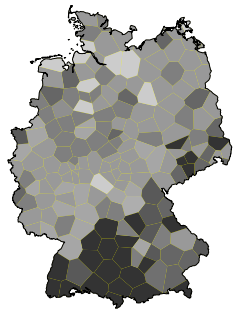
aggregate 2nd shift



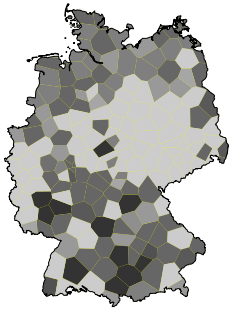
[ʃ] (dark) vs. s (non-initially)



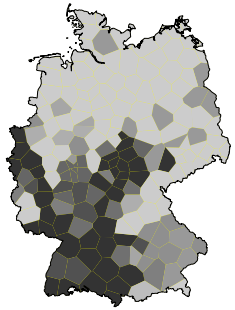
[z] (dark) vs. [s] (initially)



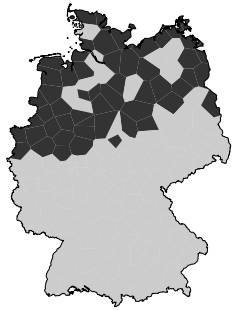
N_ d/t (dark)



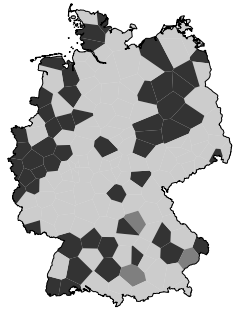
apical [r] (dark) vs. uvular [ʀ]



final [n] drop (dark) vs. retention



medial [t] vs. s



init. lenited /g/

Why dialectometry?

- Strengthen geographic signals by aggregating
- Solve problems of earlier dialectology
  - Non-overlapping distributions
  - Selection of features too arbitrary
  - “Atomism” (Coseriu), idiosyncratic words (Bloomfield)
- Introduce replicable procedures
- Following Séguy, Goebl, Schiltz, Kretzschmar, Shackleton, ...
- Seeking law-like relations in linguistic variation

Calculating dialect distances

- To determine the aggregate distance between dialects:
  - We determine the distance between each dialect pair for every single linguistic element (in sample, e.g. dialect atlas)
    - Perhaps just same (0) vs. different (1)
    - ... but we've developed more sensitive measures (below)
  - We sum these distances for every element (hundreds of them)
  - Immediate result: place $\times$ place table of dialect differences
- Séguy (1971), Goebl (1980s and on), many others

Dialectometric “feature ranking”

- Chambers & Trudgill (1998) ask for a ranking of features (and isoglosses) in order to identify dialect boundaries.
- Implicit “feature ranking” in dialectometry: a feature that’s instantiated *n* times in dialect atlas material is weighted *n* times more heavily than one that appears once.
  - Lexical items uniformly weighted
  - Phonetic segment distances weighted in proportion to their frequency in the word list
- Note that Goebl has also experimented with “inverse frequency” weighting of responses.

Aside: more sensitive pronunciation distance measure

- Levenshtein distance enables analysis of phonetic transcriptions without manual alignment
  —move from categorical to numerical analysis of data.
- One of the most successful methods to determine sequence distance (Levenshtein, 1964)
  - biological molecules, software engineering, ...
- Levenshtein distance: minimum number of insertions, deletions and substitutions to transform one string into the other
  Syllabicity constraint add: vowels never substitute for consonants

Example of the Levenshtein distance

mɔəlkə	delete ə	1
mɔlkə	subst. ɔ/ɛ	1
mɛlkə	delete ə	1
mɛlk	insert ə	1
mɛlək		
		4

m	ɔ	ə	l		k	ə
m	ɛ		l	ə	k	
	1	1		1		1

Example

- Based on Dutch pronunciation data from the Goeman-Taeldeman-Van Reenen-Project data (GTRP; Goeman and Taeldeman, 1996)
  - We use 562 words for 424 varieties in the Netherlands
- Wieling, Heeringa & Nerbonne (2007) An Aggregate Analysis of Pronunciation in the Goeman-Taeldeman-van Reenen-Project Data. In: *Taal en Tongval* 59(1), 84-116
- Calculating Levenshtein distances yields interesting sound correspondences contained in the alignments (more on that later)
  - Note that a 100-word comparison already yields about 500 sound correspondences

Distribution of sites

Analytical steps

- Obtain the distances between each of the $\approx 90,000$ pairs of varieties
  - n.b. this involves 500×5^2 segment comparisons
  - $\approx 1.1 \times 10^9$ segment comparisons in total
- Organize these in a 400×400 table
- Seek groups (dialect areas) or continuum-like relations, e.g. by applying clustering or multi-dimensional scaling, respectively

Multi-Dimensional Scaling (MDS)

- Input: site $\times$ site table of distances
- Output: optimal low-dimensional representation
  - Each site assigned coordinates in each of the dimensions
  - A measure of quality
- Important property: stability

  —small input changes (in distance table) do *not* lead to lead output changes
- Desirable property: interpretability

  —what does dimension 1 (2/3/...) represent?

MDS Quality

- Stress in SPSS, R
  - Lower stress is a better fit, 0 is perfect.
- Correlation of distances implicit in n-dimensional solution with input distances
  - If site 1 is at x_1, y_1, and site 2 is at x_2, y_2, then:
    $d(\text{site}_1, \text{site}_2) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)}$

  What about in a three-dimensional solution?
  - Correlation calculates how well the measures agree (input distances and distances in fewer dimenstions).
    1 is perfect, -1 is a perfect mismatch, 0 is unrelated.

MDS Quality

- Stress in SPSS, R
  - Lower stress is a better fit, 0 is perfect.
- Correlation of distances implicit in n-dimensional solution with input distances
  - If site 1 is at x_1, y_1, and site 2 is at x_2, y_2, then:
    $d(\text{site}_1, \text{site}_2) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)}$

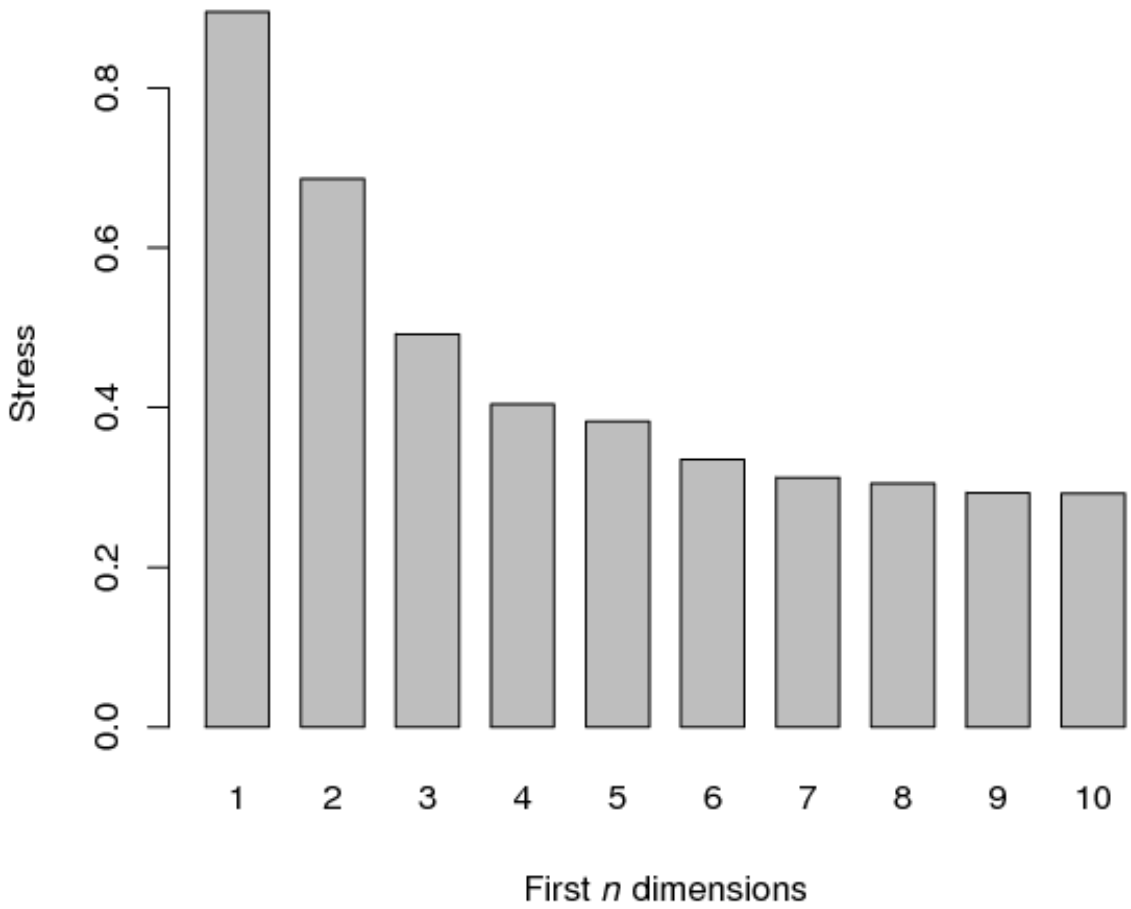
    What about in a three-dimensional solution?
  - Correlation calculates how well the measures agree (input distances and distances in fewer dimenstions).
    1 is perfect, -1 is a perfect mismatch, 0 is unrelated.

MDS Quality

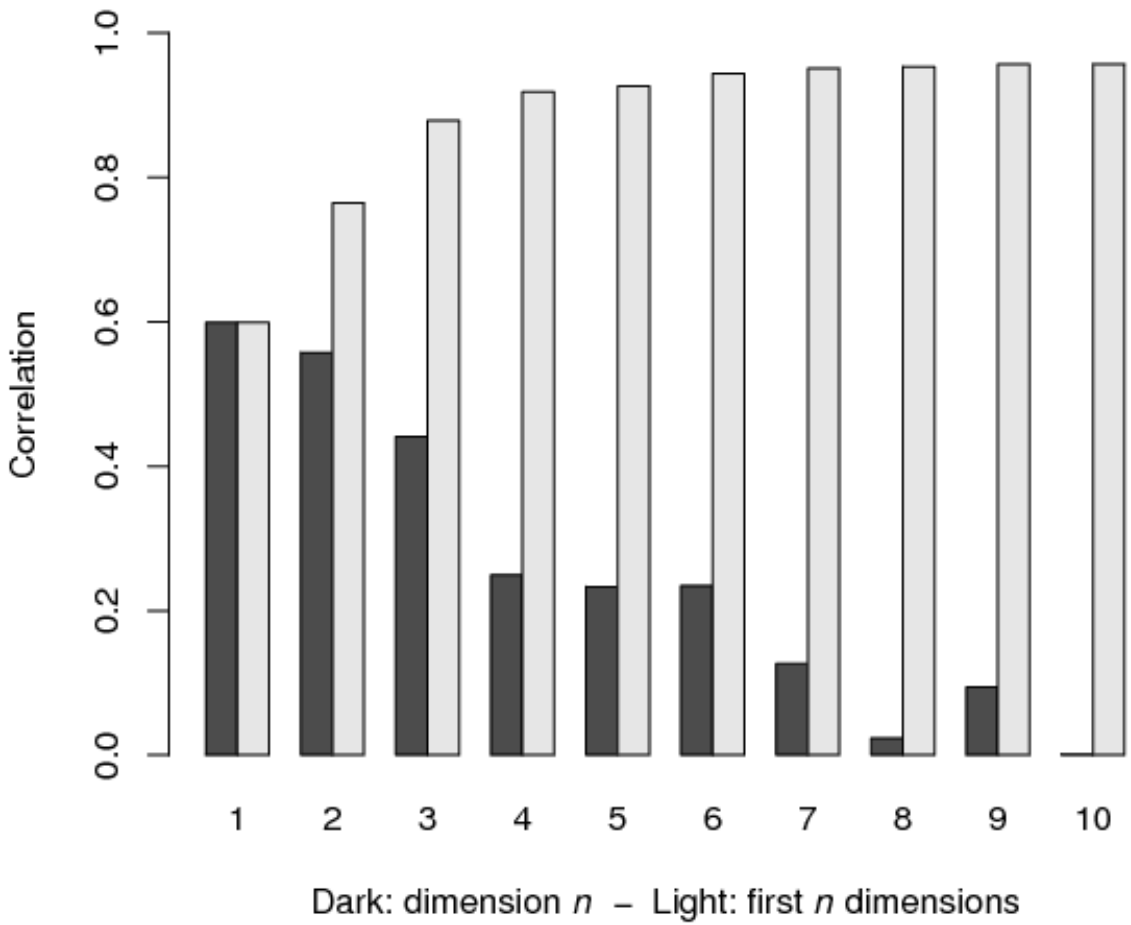
- Stress in SPSS, R
  - Lower stress is a better fit, 0 is perfect.
- Correlation of distances implicit in n-dimensional solution with input distances
  - If site 1 is at x_1, y_1, and site 2 is at x_2, y_2, then:
    $d(\text{site}_1, \text{site}_2) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)}$

    What about in a three-dimensional solution?
  - Correlation calculates how well the measures agree (input distances and distances in fewer dimenstions).
    1 is perfect, -1 is a perfect mismatch, 0 is unrelated.

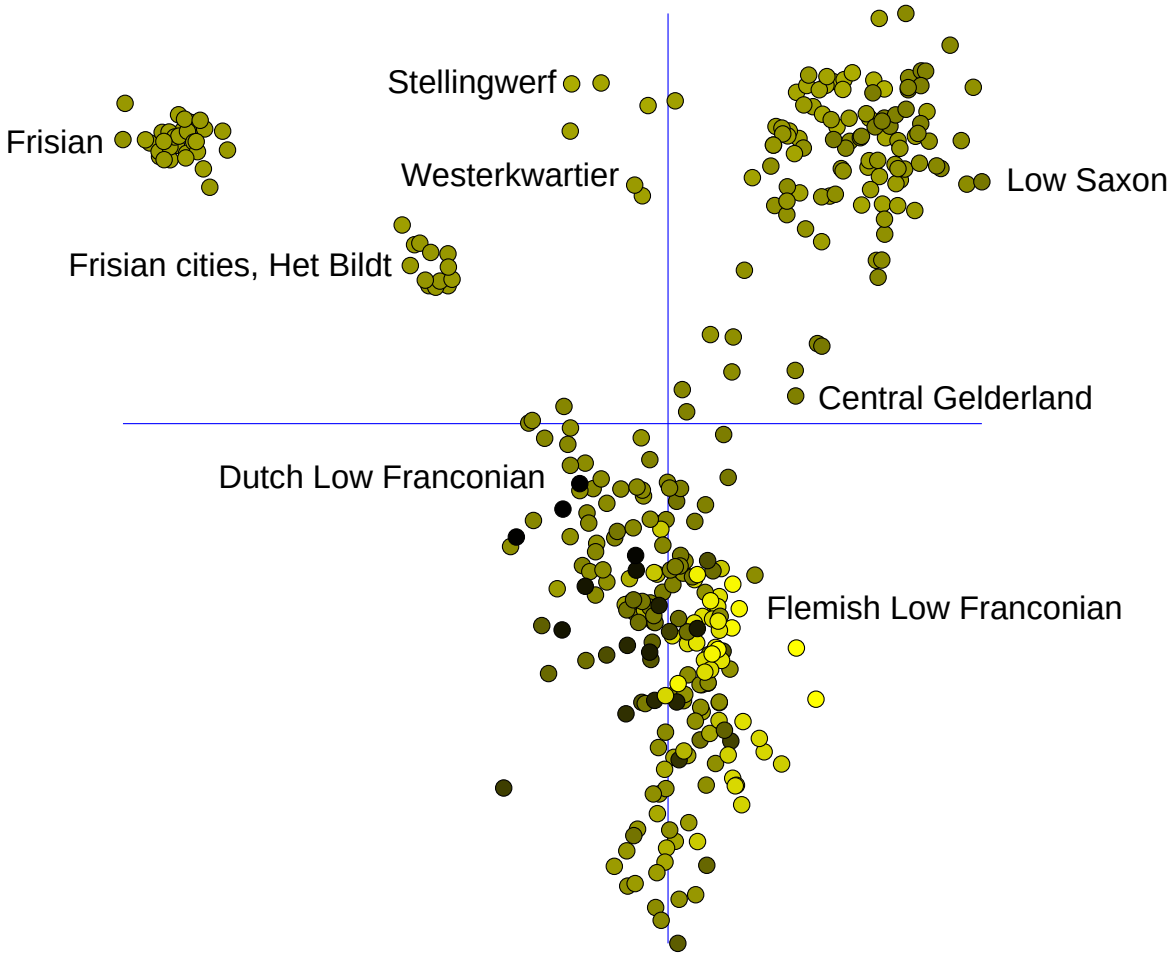
MDS Stress



MDS Stress

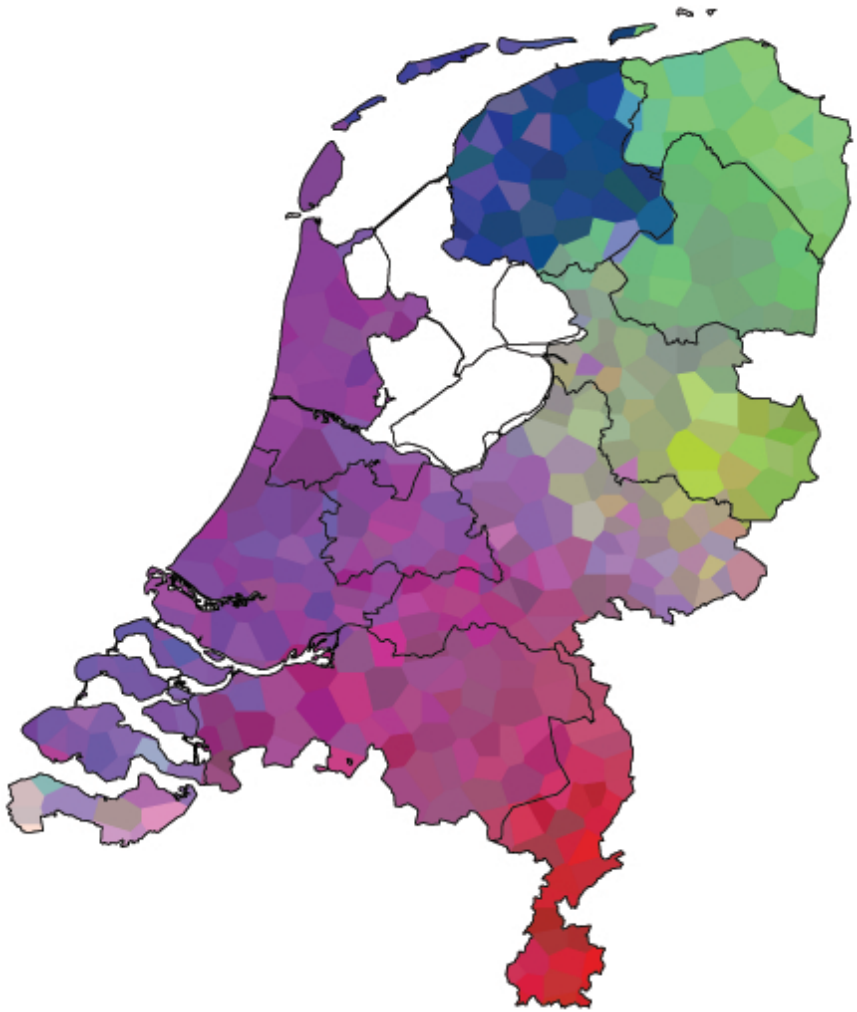

Correlation
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
6
7
8
9
10
Dark: dimension n – Light: first n dimensions

Multi-Dimensional Scaling

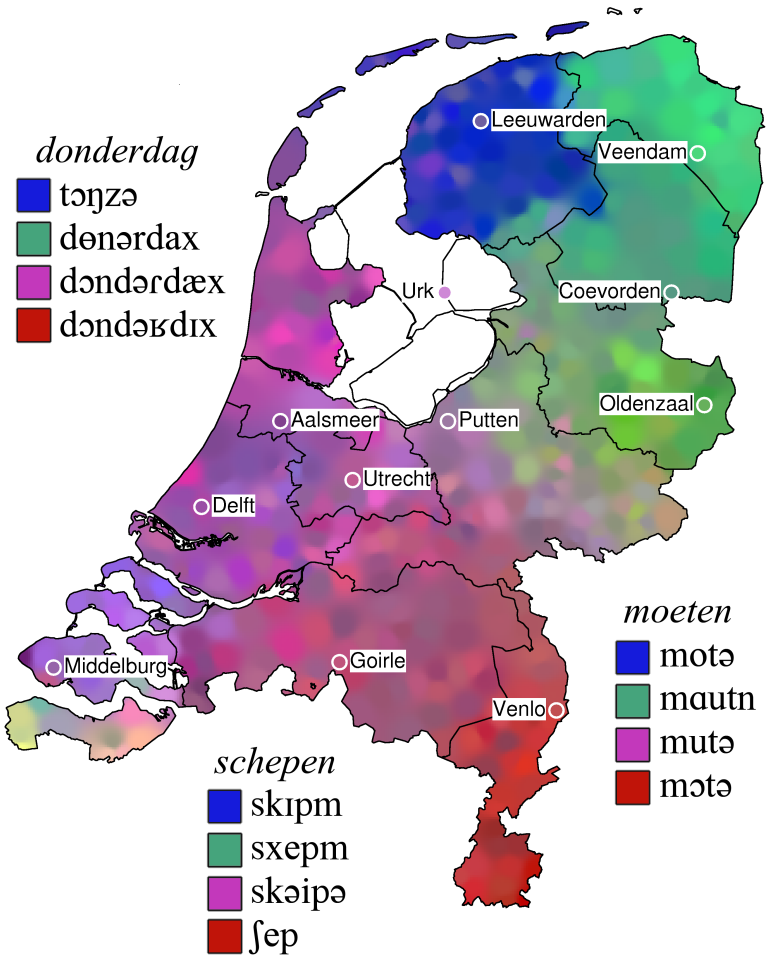


Corr. $r = 0.88$

MDS dimensions → colors, projected to map



MDS Interpretation



Clustering

- Clustering seeks “natural” groups (of most similar elements in data.
  - Because older dialectology often organizes its results via DIALECT AREAS, we wish to find groups
- **Many** clustering options, we discuss (i) simple ones; and esp. (ii) options that have proven themselves in dialectology.
- Since dialect areas are often hierarchical, we apply hierarchical (agglomerative) clustering.

Hierarchical Clustering

- Input: distance tables (same as MDS)
- Procedure: find smallest distance in table, between i and j, then fuse the two.
  - This means that the $n \times n$ table becomes a $(n-1) \times (n-1)$ table.
  - It also means that we need to update distances between all the unfused elements and the newly fused one.
  - Options: average distance, weighted average, minimal-error distance, ...
- Output: a dendrogram, a tree with sites as leaves, and internal nodes showing where two elements were fused.
- Quality: often measure via CO-PHENETIC CORRELATION, correlation between input distances and distances in dendrogram.
- Problem: because of the focus on the *smallest* element, clustering is not STABLE. Small changes in input many cause large changes in the output dendrogram.

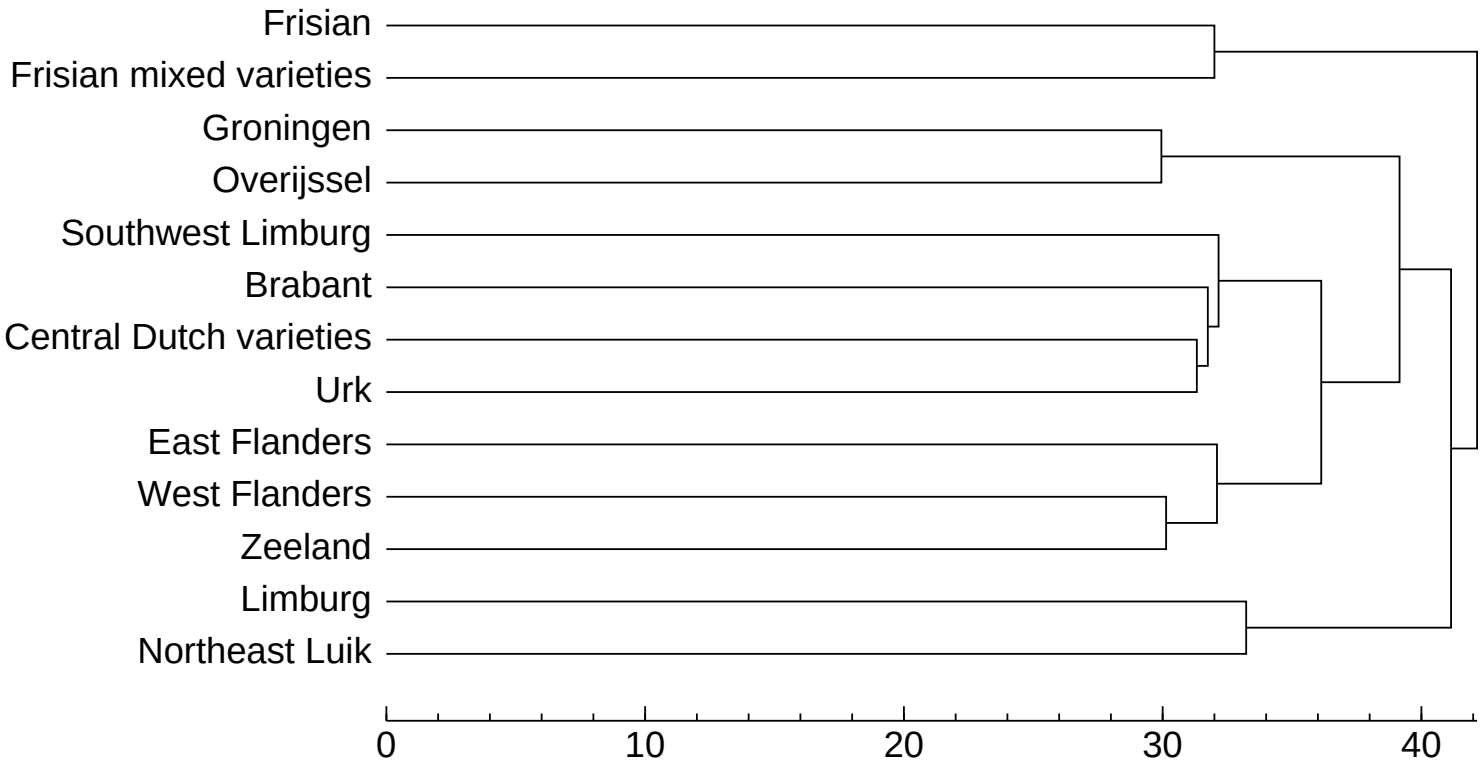
Example of Clustering

	Grouw	Haarlem	Delft	Hattem	Lochem
Grouw	0	41	44	45	46
Haarlem	41	0	16	34	36
Delft	44	16	0	37	38
Hattem	45	34	37	0	20
Lochem	46	36	38	20	0

Apply Johnson's algorithm to the upper half of the matrix (blue values):

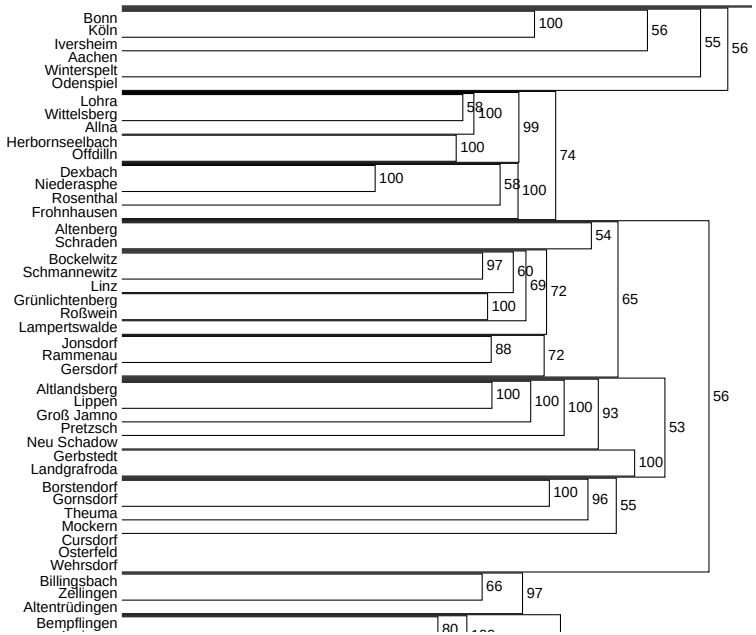
- Iteratively,
  1. select shortest distance in matrix,
  2. fuse the two datapoints involved.
- To iterate, we have to assign a distance from the newly formed cluster to all other points (several alternatives, we used UPGMA).
- Repeat until one cluster is left over.

Example Clustering Output



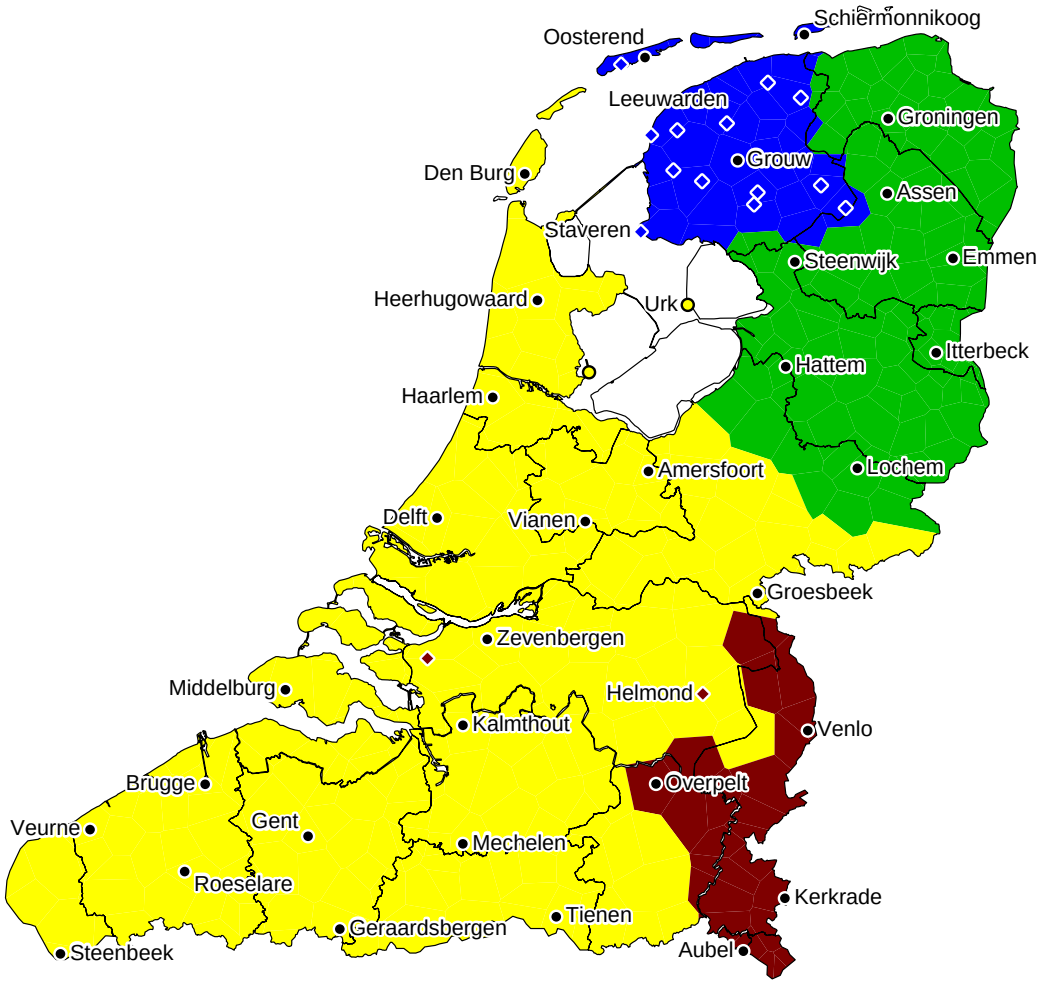
Using cluster analysis a dendrogram is derived from the 360 $\times$ 360 matrix. The scale distance shows percentages. Each of the 13 most significant groups is summed in one label.

To Improve Stability: Noisy Clustering



- Seeks groups in data, enabling comparison to older dialectology which sought *areas*
- Only bootstrap (or noisy) clustering to avoid instability

Projecting groups to geography



Large body of dialectometric work—positive aspects

- Dutch, German, American English, Norwegian, Swedish, Afrikaans, Sardinian, Tuscan, Catalan, Bulgarian, Croatian, Estonian, Sino-Tibetan, Chinese, Central Asian (Turkic & Indo-Iranian), ...
- Development of consistency measure (Cronbach's α) indicting whether data set is sufficiently large
- Novel reflection, work on validation aimed at assessing degree of detection of SIGNALS OF PROVENANCE
  - Gooskens & Heeringa (2004) Perceptive Evaluation of Levenshtein Dialect Distance Measurements using Norwegian Dialect Data. *Language Variation and Change* 16(3), 189-207.

Criticisms of dialectometry, esp. Levenshtein-based work

- Measure is too insensitive, 0/1 segment differences
- Too little attention to phonetic/phonological conditioning
- Too reliant on transcription—what about acoustics?
- Where is the sociolinguistics? Isn't variationist linguistics mostly about sociolinguistics?
- "Distance-based" methods yield too little insight into the linguistic basis of differences (concrete differences lost in the aggregate sums)

  —the hint is that it may be all smoke & mirrors
- So what? Isn't this all just confirming what we knew earlier?

... progress on all fronts, but presentation would take too long

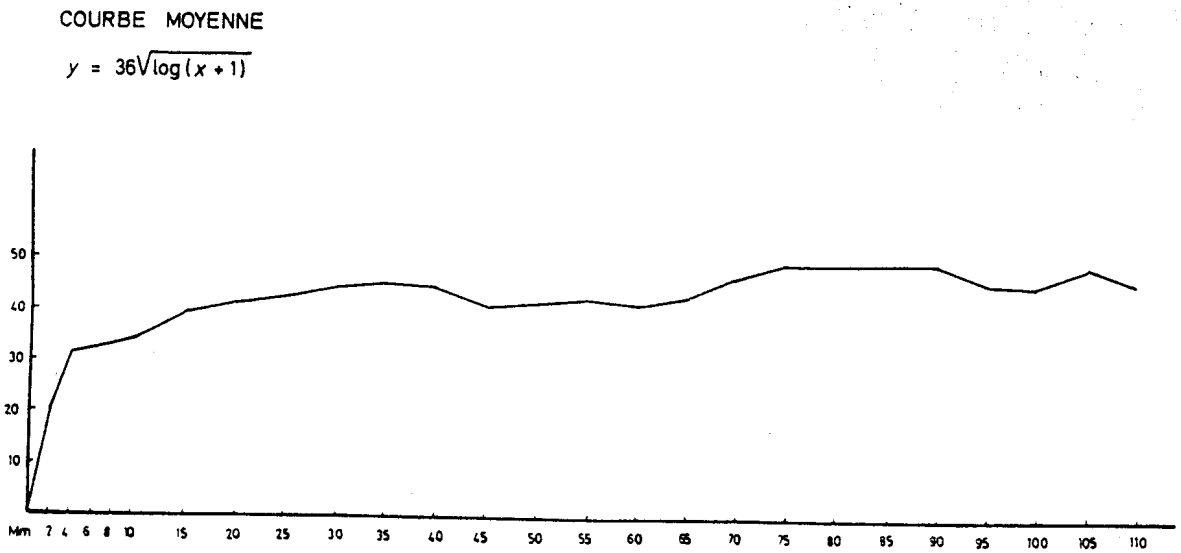
—question and discussion period for those interested

The Influence of Geography

- Regression design
- Dependent variable: varietal distance, as measured by aggregate categorical distance or Levenshtein distance
- Independent variable: geographical distance, regarded as an operationalization of the chance of social contact
- Statistical cautions:
  1. correlations involving averages are inflated
     — but we're interested in the entire varieties (dialects)
  2. distances are not independent, so significance may be inflated
     — Mantel tests

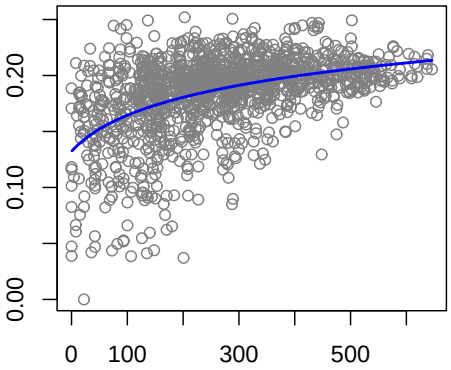
Inspiration: Jean Séguy

- Séguy (1971) La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane* 35(138), 335-357: Aggregate variation increases sublinearly with respect to geography

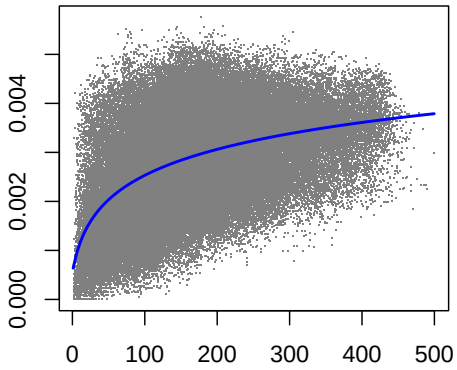


Sublinear spread is general

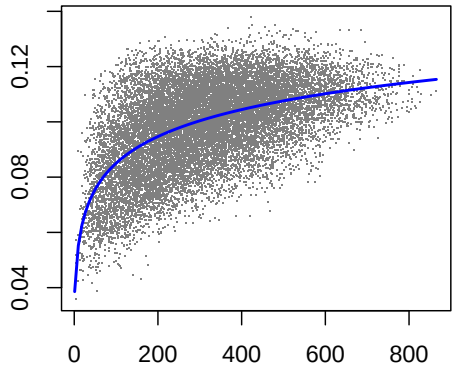
Bantu



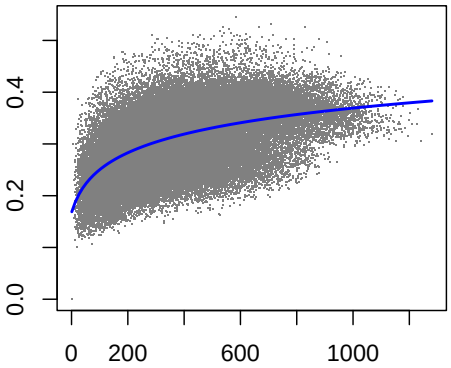
Bulgaria



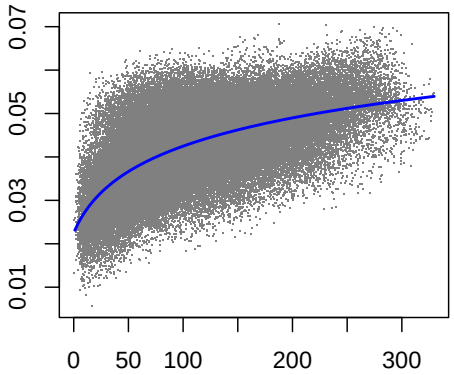
Germany



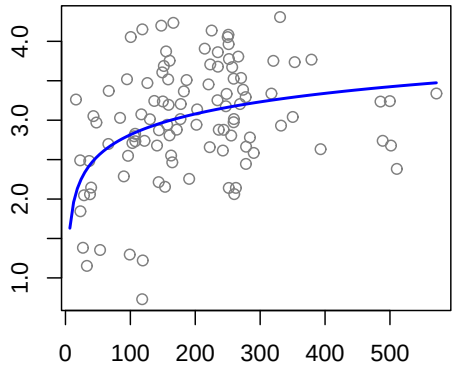
LAMSAS / Lowman



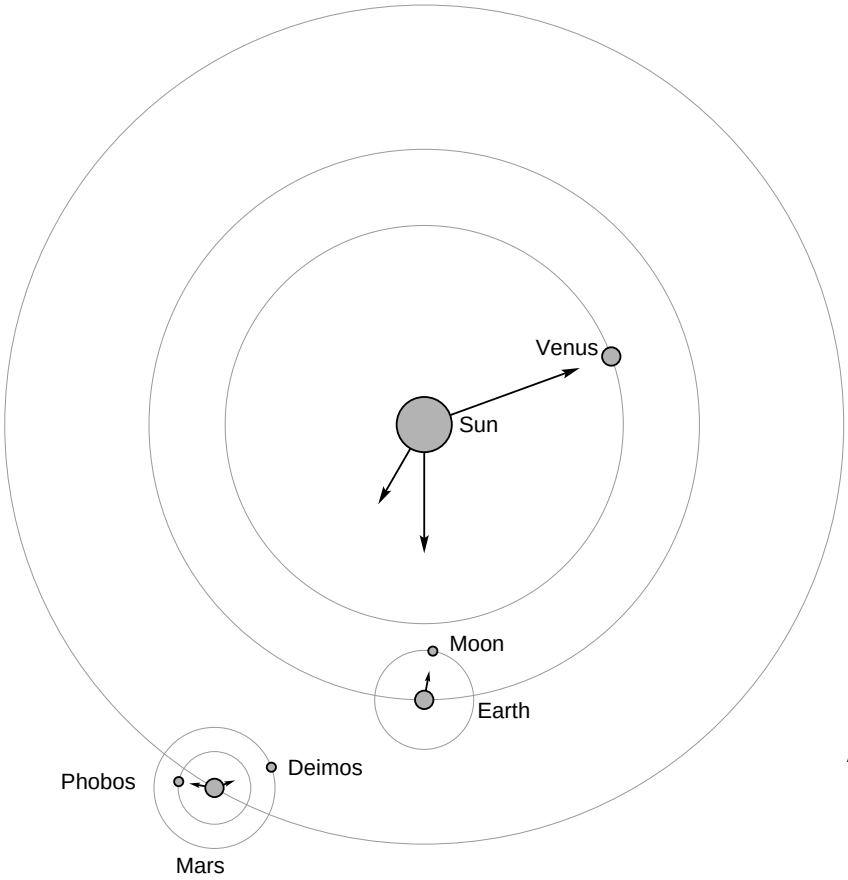
The Netherlands



Norway



Aside: Trudgill's "Gravity hypothesis"



According to Trudgill (1972) diffusion follows an inverse square law, with the consequence that linguistic distance should likewise increase with the square of the distance. Population size plays the role of mass.

Trudgill's "Gravity hypothesis"

- Sublinear aggregate relation incompatible with a quadratic influence (on individual features)

J.Nerbonne (2010) Measuring the Diffusion of Linguistic Change. *Phil. Transactions of the Royal Society B: Biological Sciences* 365.

How much does distance influence language?

Area	Corr.(l,geo)	r^2
Gabon Bantu	0.47	0.22
Bulgaria	0.49	0.24
Germany	0.57	0.32
Eastern U.S.	0.51	0.26
Netherlands	0.62	0.38
Norway	0.41	0.16

Norwegian ling. dist. correlates better w. travel time in 1900 ($r = 0.54$) Gooskens (2005) *Dialectologia et Geolinguistica* 13.

Adding areas increases explained variance 50% (forthcoming in a Freiburg volume)

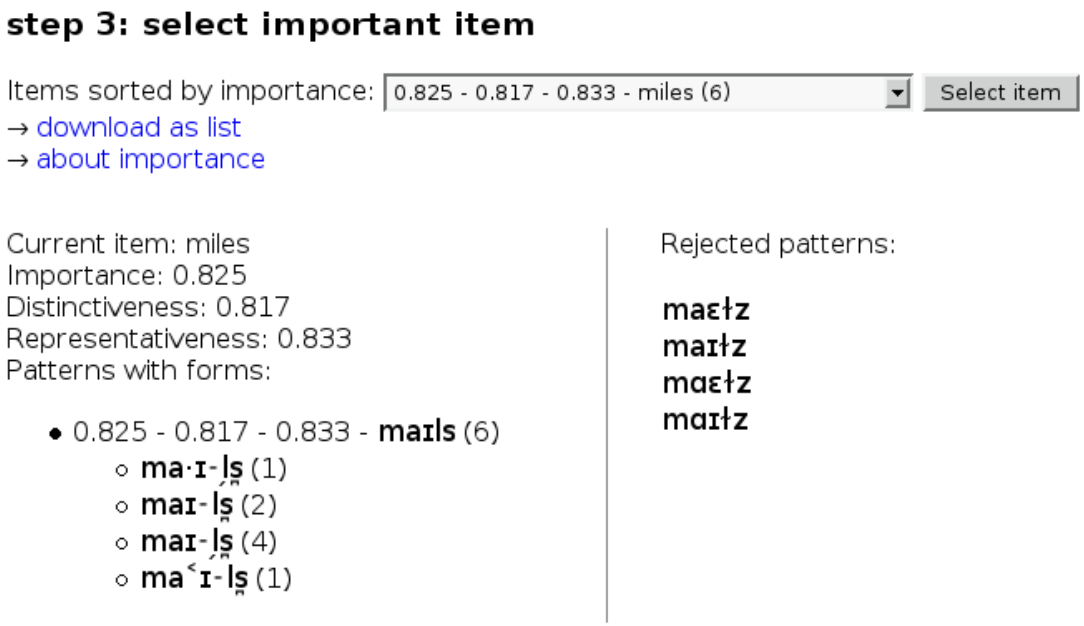
Geographic influence on language

- Geography accounts for 33 – 57% of aggregate linguistic variation.
- General — sublinear — characterization of relation between geographical distance and linguistic differences
- Like population geneticists' "isolation by distance" (Wright, 1943; Malécot, 1955)

Features? (assuming aggregate analysis)

- *Argumentum ad auctoritatem* Groningen software supports free search (with measures of "importance")
- Post-hoc "feature mining": We can look for words that correlate with significant dimensions of MDS solutions (of aggregate analyses).
- Bipartite spectral graph partitioning (like two-dimensional factor analysis).
  - Begin with matrix of varieties $\times$ features
  - Cluster varieties and features simultaneously.
- Mixed models
  - Include feature choice (words) as random-effect factor in regression model.

“Importance” of feature wrt area

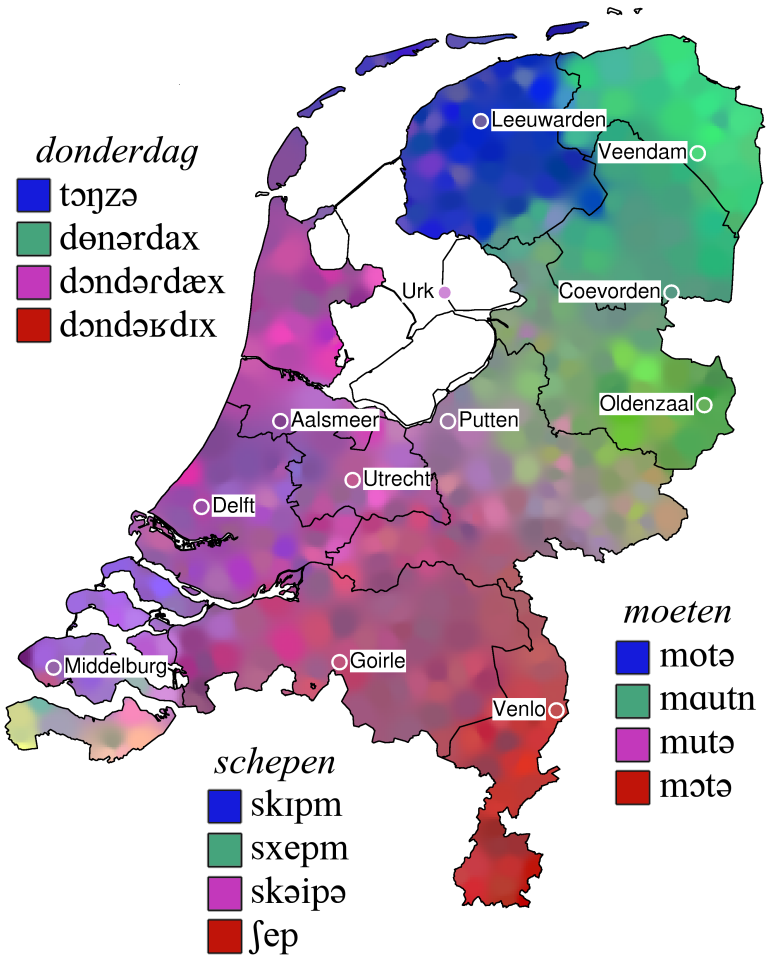


Representative(f,a) ≈ relative frequency of f among sites

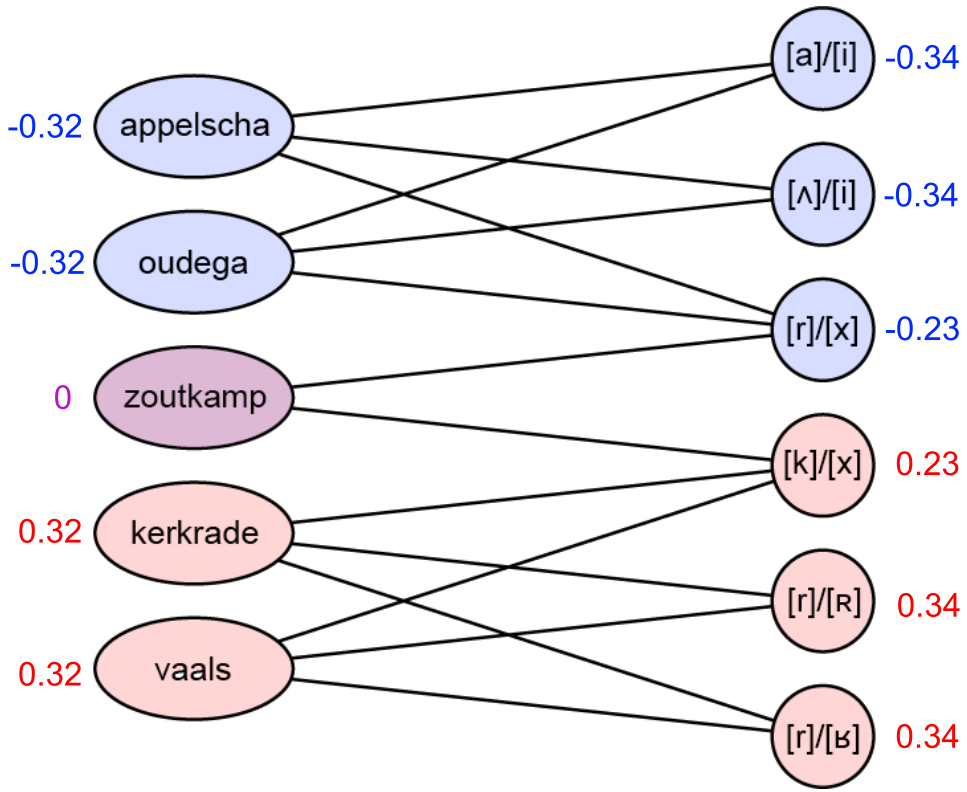
Distinctive(f,a) ≈ proportion of occurrences of f in a as opposed to outside a

Importance(f,a) is average of representativeness and distinctiveness

MDS-based feature-mining

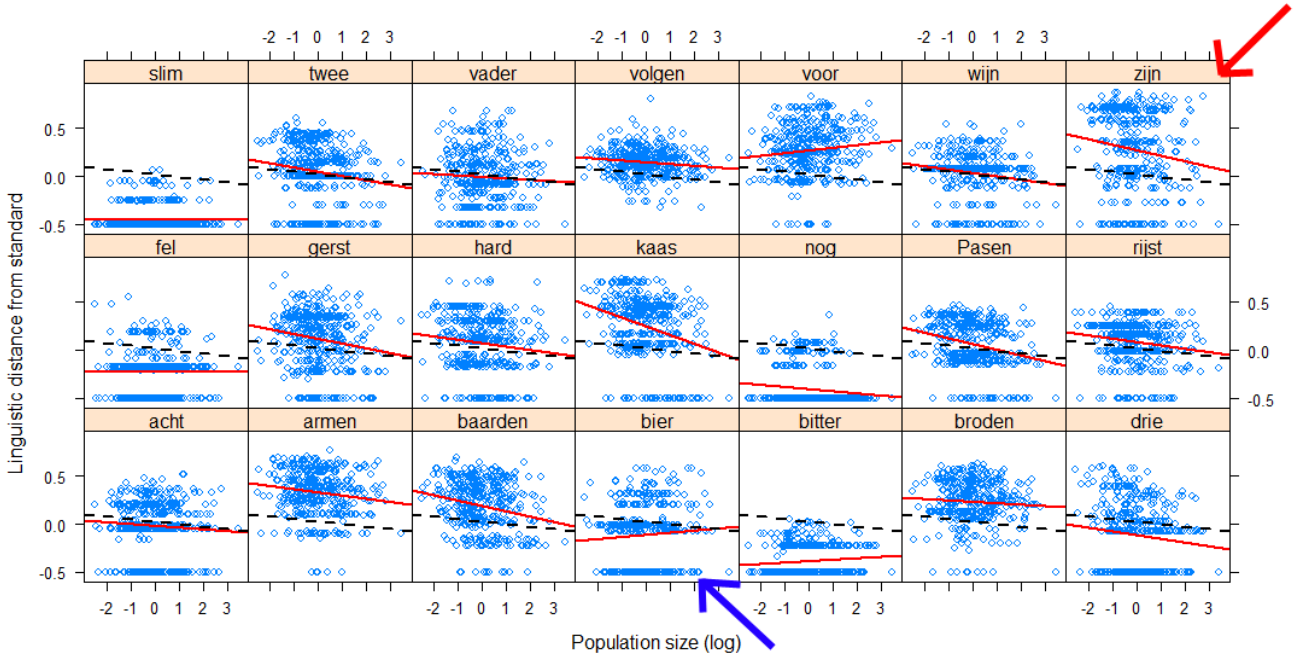


Co-clustering bi-partite spectral graph



Details during discussion if wanted.

“Mixed models”: modeling each word

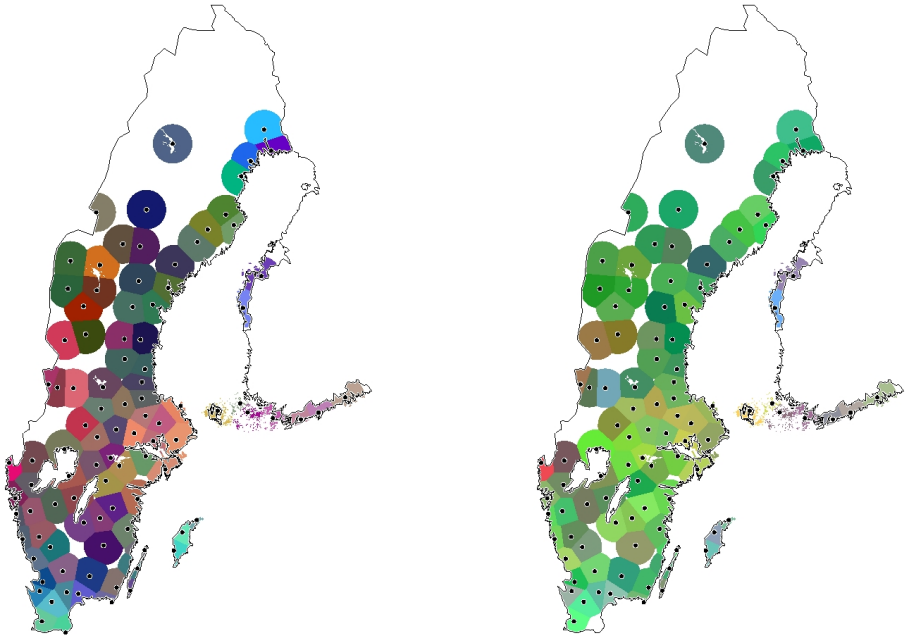


- $LD = \quad 0.00 + 0.01\,WF - 0.005\,PS + 0.004\,PA$ (general model)
- $LD = -0.01 + 0.01\,WF + 0.010\,PS + 0.004\,PA$ (word: *bier*)
- $LD = \quad 0.20 + 0.01\,WF - 0.008\,PS + 0.004\,PA$ (word: *zijn*)

Ongoing work by Martijn Wieling (submitted)

A caution: dialect continua

Old vs. young speakers in Sweden (SveDia, Therese Leinonen, 2010)



“Feature ranking” could *create* spurious dialect areas, even where scientific consensus sees continua.

Features in aggregate analysis

- Aggregate perspective enables identification & formulation of general law: distance models explain $22\% - 38\%$ of aggregate linguistic variation.
  - Areal distinctions a bit collinear, but add ($\approx 50\%$).
- Features naturally ranked in dialectometric view, either as uniform, or as reflected in item sample / lexicon
- Several means of identifying and ranking features
- Emerging questions:
  - What is the linguistic structure of the dialect differences we find?
  - Do typological constraints play a (confounding) role?
  - Can we tease apart geographical and historical explanations, and how?

Try Gabmap! www.gabmap.nl

Questions?

Thank You!