Dialects

GC

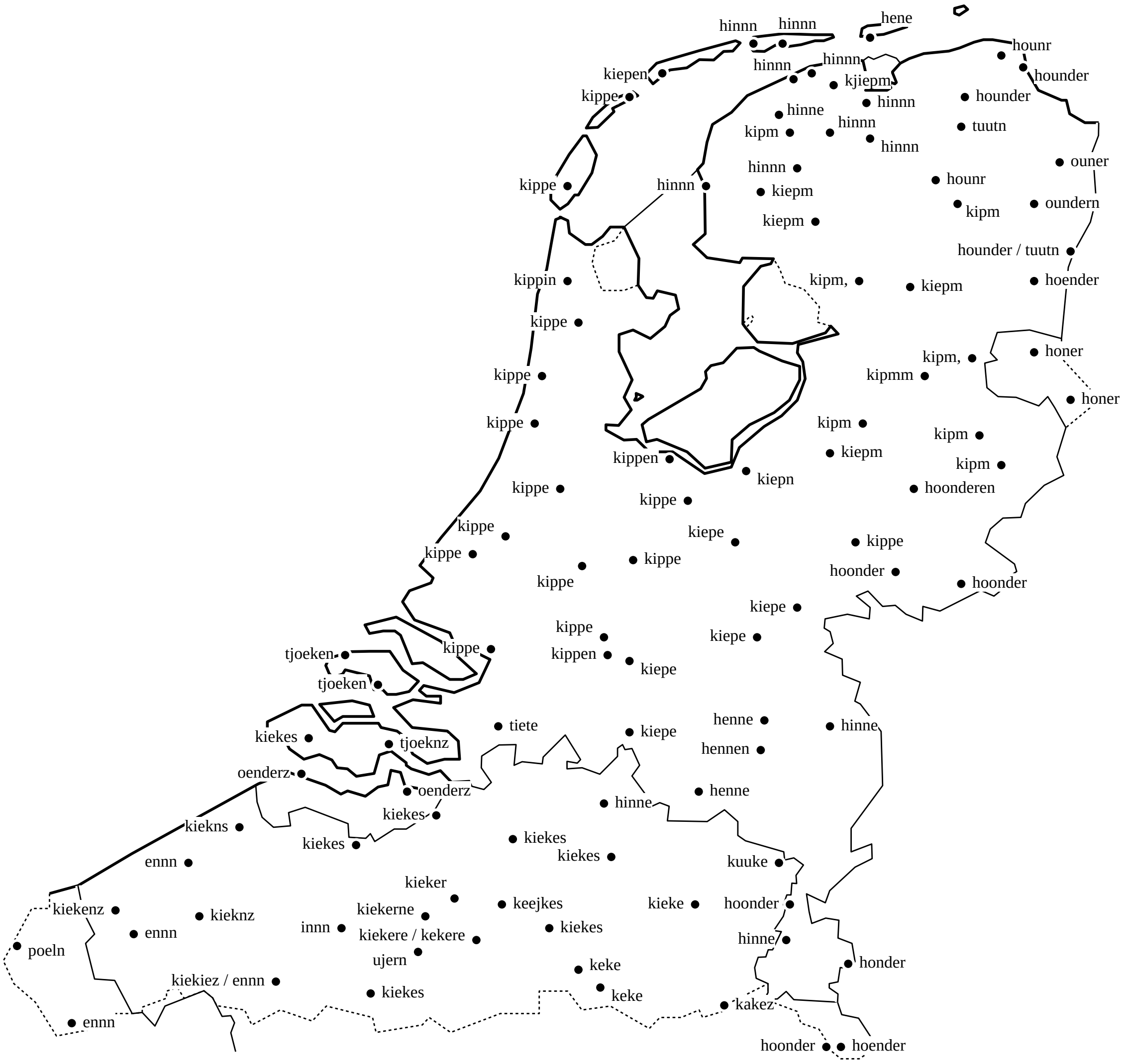
Within Humanities, maps are important in Linguistics, History and Art (and Architecture)

In analogy to *isotherm* in climate map, linguists draw lines around areas in which same or similar forms are used. The lines are ISOGLOSSES.

They interesting because they show cultural affinity which might be due to social or commercial ties, migration, or conquest.

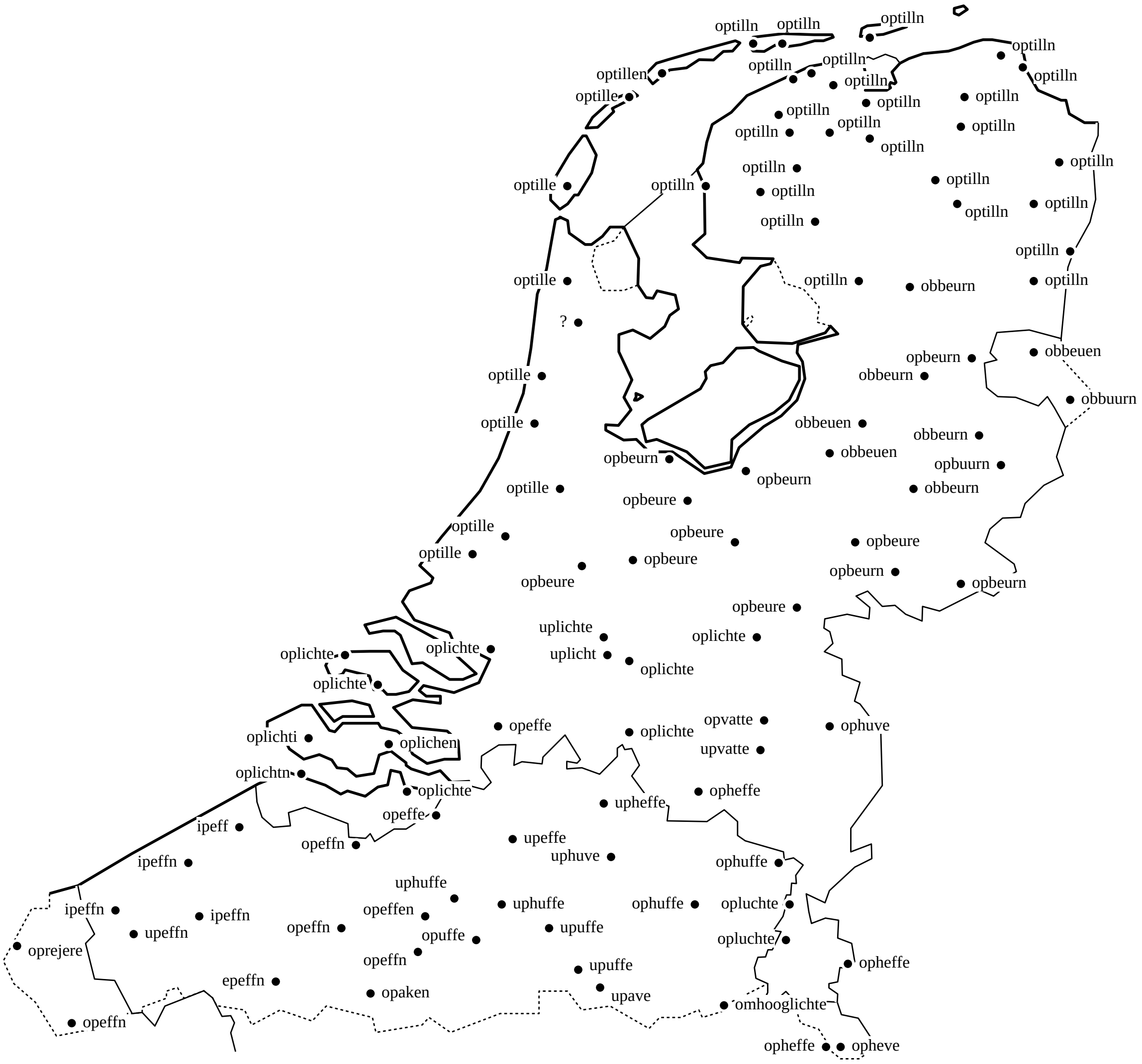
Isoglosses

GC



Isoglosses for different forms of 'kippen' (chicken) would be drawn North-South around eastern border (variants of *hounder*), and in Flanders (variants of *kieken*).

Isoglosses



Isoglosses for different forms of for different forms of 'optillen' (lift up) would run East-West.

Isoglosses

Isoglosses are an important tool, but they are insufficient for the identification of DIALECT AREAS — areas in which the same or similar varieties are spoken. Bloomfield (^{1}1916,1933) summarized this, but the problem was already well-known in his time

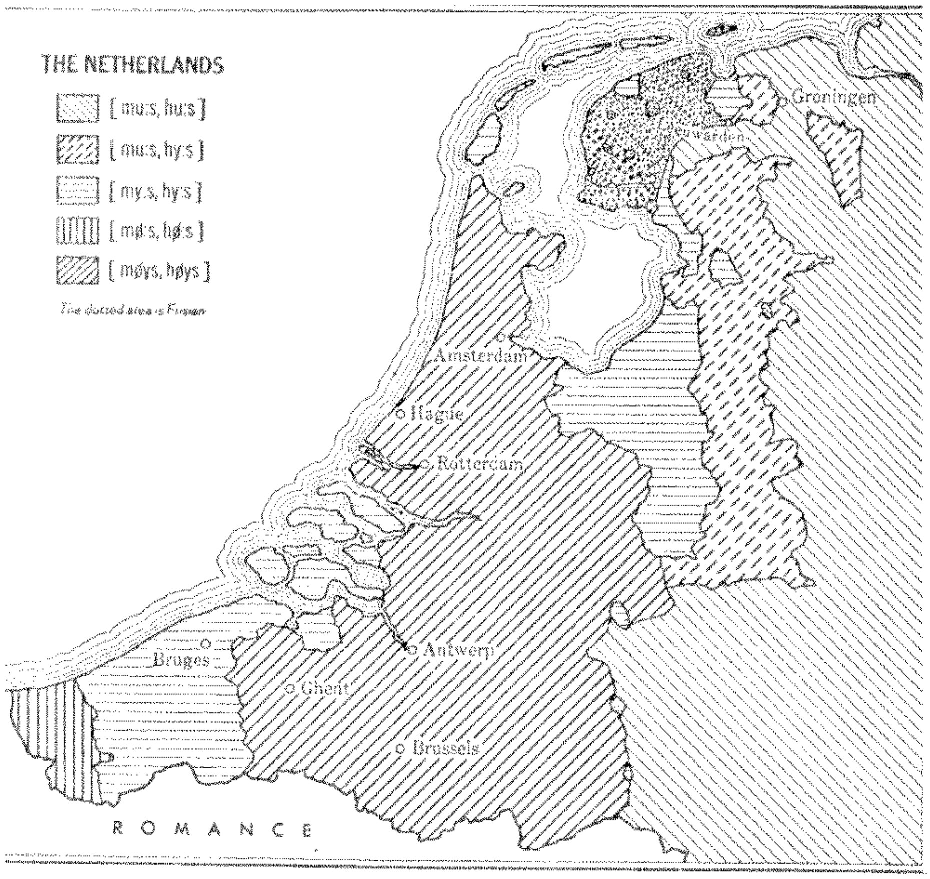


FIGURE 6. Distribution of syllabic sounds in the words *mouse* and *house* in the Netherlands. — After Kloeke.

"every word has its history"

Edit Distance

- Edit Distance (= Levensthein Distance)
  - equals the cost of (the least costly set of) operations mapping one string to another
  - basis costs are insertions (1), deletions (1), substitutions (2)
  - two strings are compared by calculating their Levenshtein distance

adresse	insert d	1
addresse	delete e	1
address		2

How do you know it's the *cheapest*?

Try *all* the sequences of operations?

Algorithm

Levenshtein distance($adresse, address$)

		a	d	d	r	e	s	s
	0	1	2	…				
a	1							
d	2							
r	⋮							
e								
s								
s								
e								

Top horizontal row is always $1, 2, \ldots$ —cost of insertions
Left vertical column is always $1, 2, \ldots$ —cost of deletions

- begin at upper left ($\Leftarrow 0$)
- to fill in a cell:

diag	above
left	min(above + delete, diag + replace, left + insert)

- lower right corner of table contains LevD

Algorithm

Levenshtein distance($adresse, address$)

		a	d	d	r	e	s	s
	0	1	2	3	4	5	6	7
a	1	0	1	2	3	4		
d	2	1	0	1	2			
r	3	2	1	2	1			
e	4	3	2			1		
s	5	4					1	
s	6							1
e	7							2

address, adresse are two Levenshtein units apart.

Alignment

Levenshtein distance($adresse, address$)

		a	d	d	r	e	s	s
	0	1	2	3	4	5	6	7
a	1	0	1	2	3	4		
d	2	1	0	1	2			
r	3	2	1	2	1			
e	4	3	2			1		
s	5	4					1	
s	6							1
e	7							2

path of lowest scores shows *alignment* of strings

a	d	d	r	e	s	s	
|	|	|	|	|	|	|	|
a	d		r	e	s	s	e

Applications

other

biologie align DNA sequences

ethology map evolution in bird songs

In language

spell checker given misspelling, find closest match in dictionary
more is needed for this!

alignment align bilingual texts
use sentence length as indicator of base similarity

language therapy identify sources of deviant pronounciation

language variation measure differences among dialects or social groups

Dialect Pronunciation

- use 100-word sample in large number of varieties
- dialect distance is equal to the sum of the word distances
- first applied for dialect comparison by Kessler (1995) for Irish dialects
- applied for Dutch dialects by Nerbonne et al. (1996), Nerbonne and Heeringa (1997), Nerbonne and Heeringa (1999, to appear).
- example

kœstə	verwijder ə	1
kœst	vervang œ door ɔ	2
kɔst	voeg toe r	1
kɔrst		
		4

Levenshtein distance

- Calculate the cost of changing one string into another
- Example: 'saw a girl' is pronounced as [sɔ:əgIrl] (Standard American) and [sɔ:rəgø:l] (Boston). Change the first pronounciation into the other.

sɔəgIrl	delete r	1
sɔəgIl	replace I/ø	2
sɔəgøl	insert r	1
sɔrəgøl		
		4

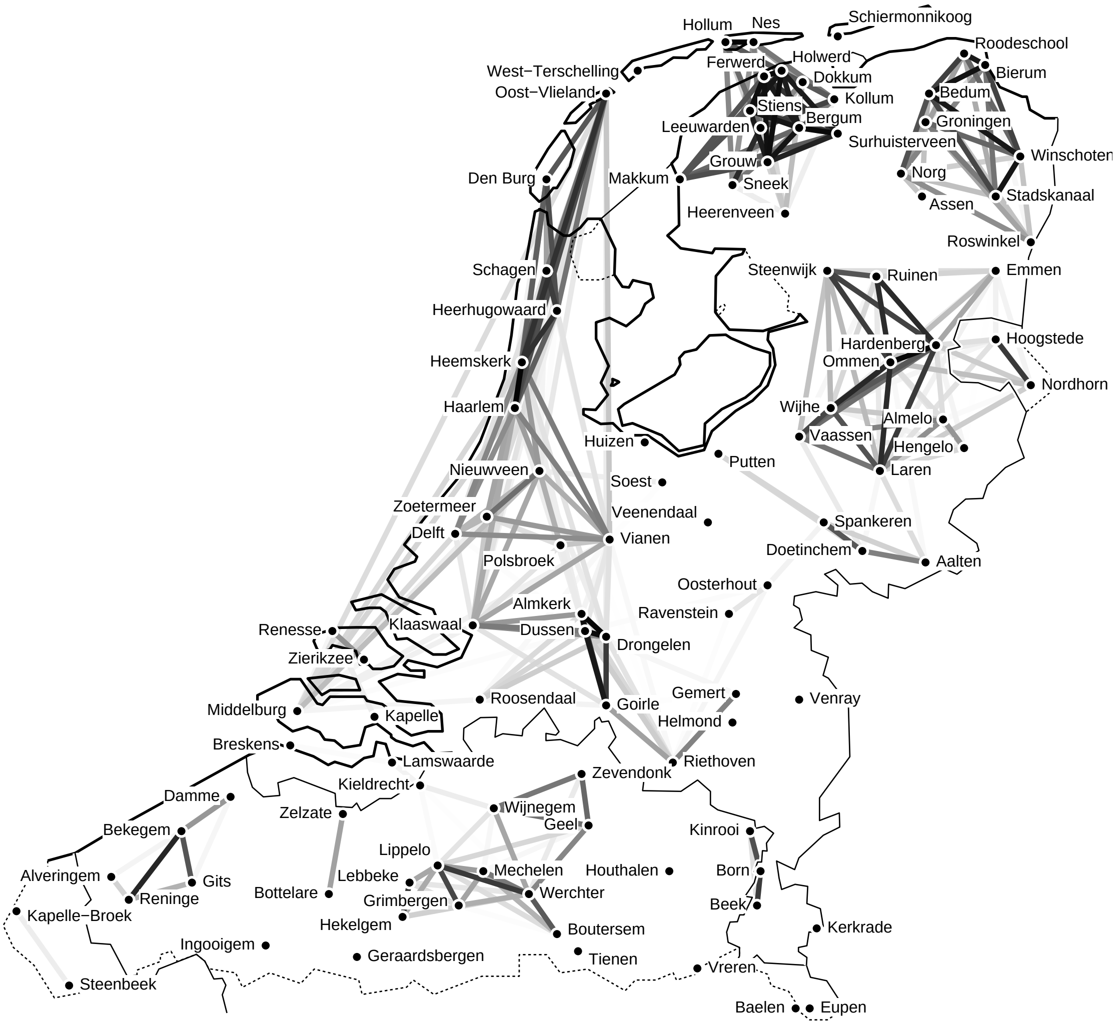
- Refinement: by looking at the features the value of a replacement varies between 0 and 2. Diacritics [ĩ,e:,ə^r] can also be taken into account.
- Example: the difference between [i] and [e] is much smaller than the difference between [i] and [u].

	i	e	u	i-e	i-u
advancement	2(front)	2(front)	6(back)	0	4
high	4(high)	3(mid high)	4(high)	1	0
long	3(short)	3(short)	3(short)	0	0
rounded	0(not rounded)	0(not rounded)	1(rounded)	0	1
				1	5

Levenshtein distance

- By looking at the discrimination of the segments for each feature a weight can be calculated (Quinlan, 1993).
- Many sequence operations map [sɔ:əgIrl] → [sɔ:rəgø:l].
  Levenshtein distance = cost of cheapest mapping.
- Using 100 words the distance between two dialects is equal to the the sum of 100 Levenstein distances.
- All distances between n dialects are arranged in a n × n matrix.

Levenshtein

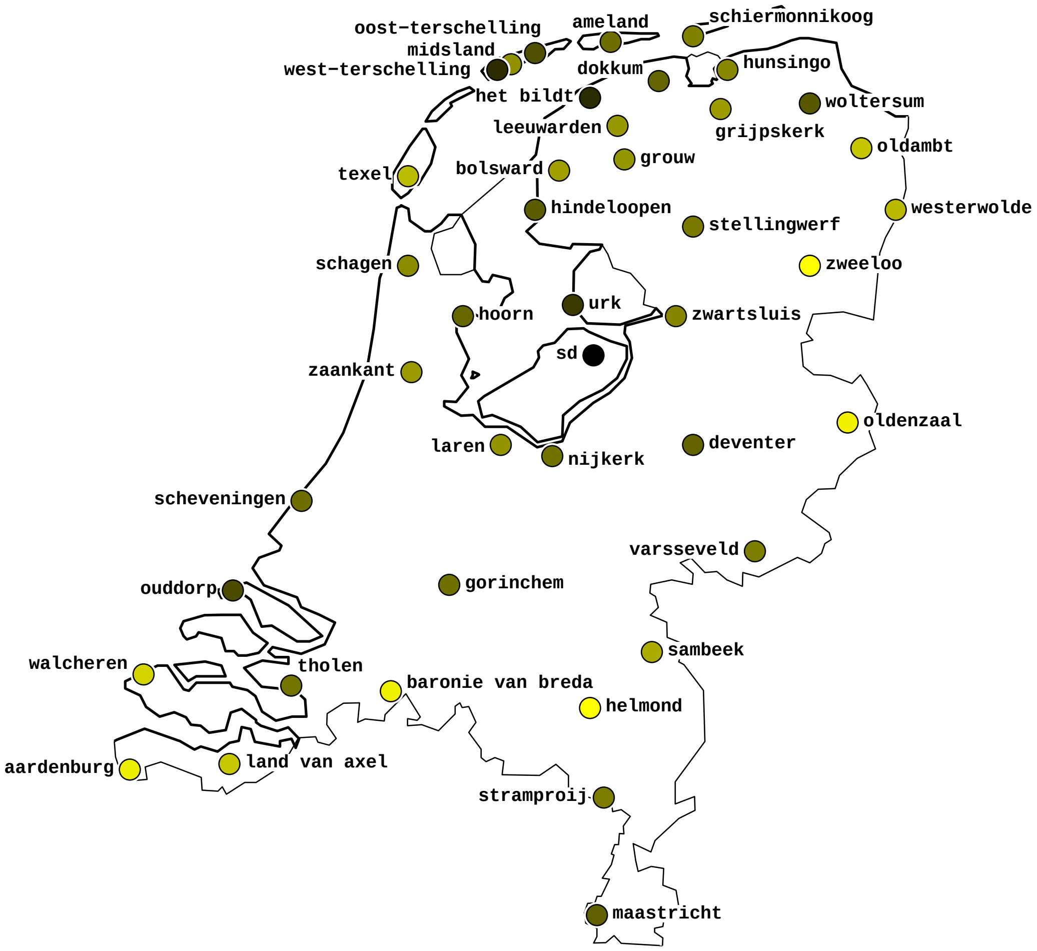


Average Levensthein distances between dialects. Darker lines connect closer points, lighter lines more remote ones. Notice that what's being mapped is (the strength of) a RELATION between two geographic points.

History

Languages change. To see how, we can compare pronunciation differences from two time periods.

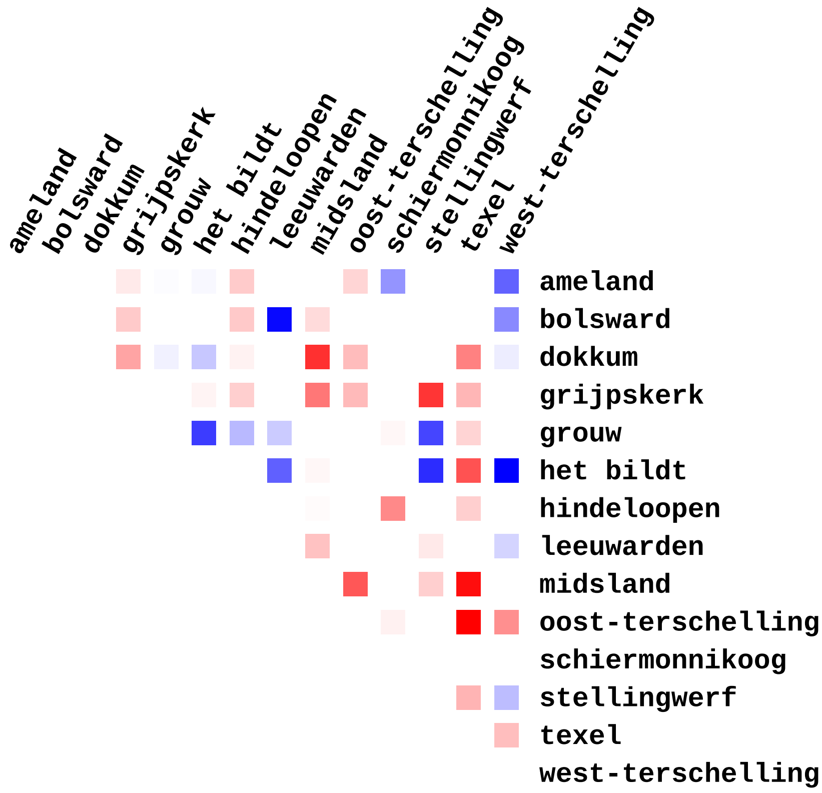
Winkler (1874) "dialect atlas" of Dutch, Flemish, Low German



yellow indicates most extreme changes

Details — Relations

We can also examine which relations changed. Which pairs of varieties became more or less alike?



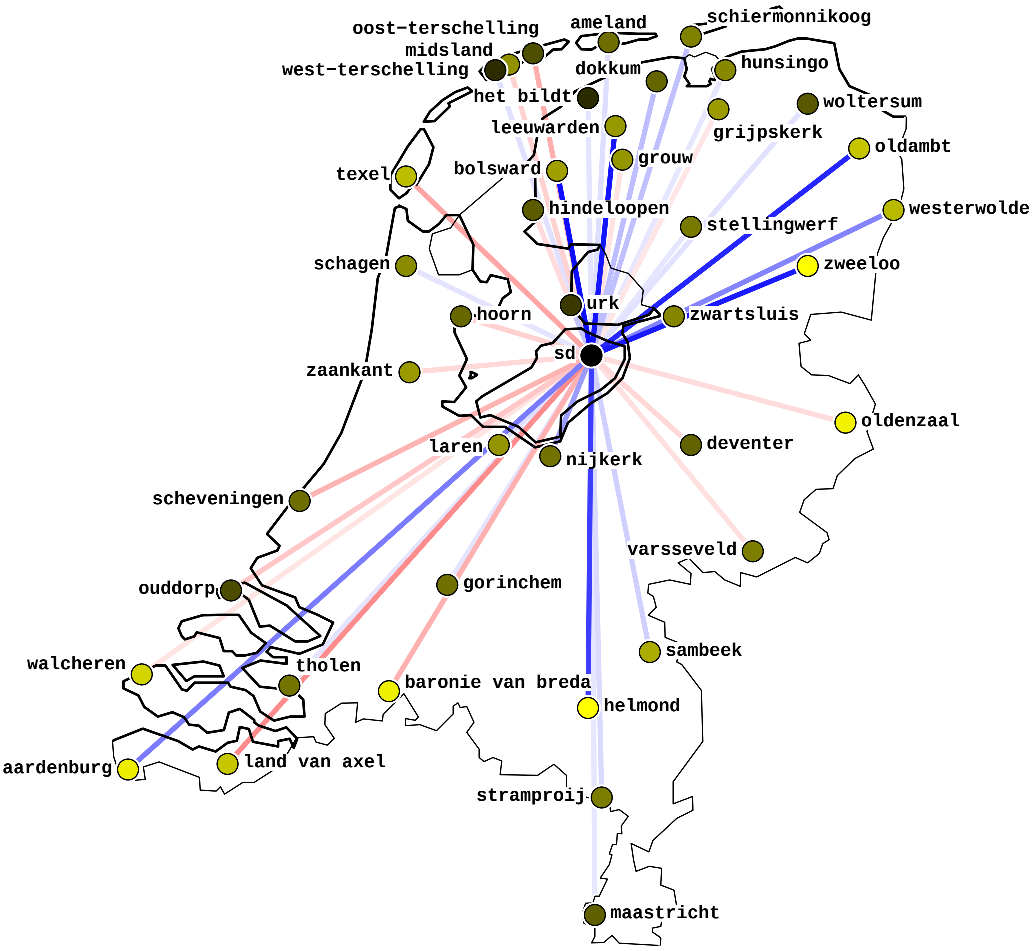
Blue convergence, red divergence.

Some rows show red and blue. Why?

GC

Combining Views

Which varieties changed (yellow of site) and how did they change vis-à-vis others? **sn** is 'Standard Netherlands'.



Effective, or cluttered?

Suppose earlier graphics had not been shown?

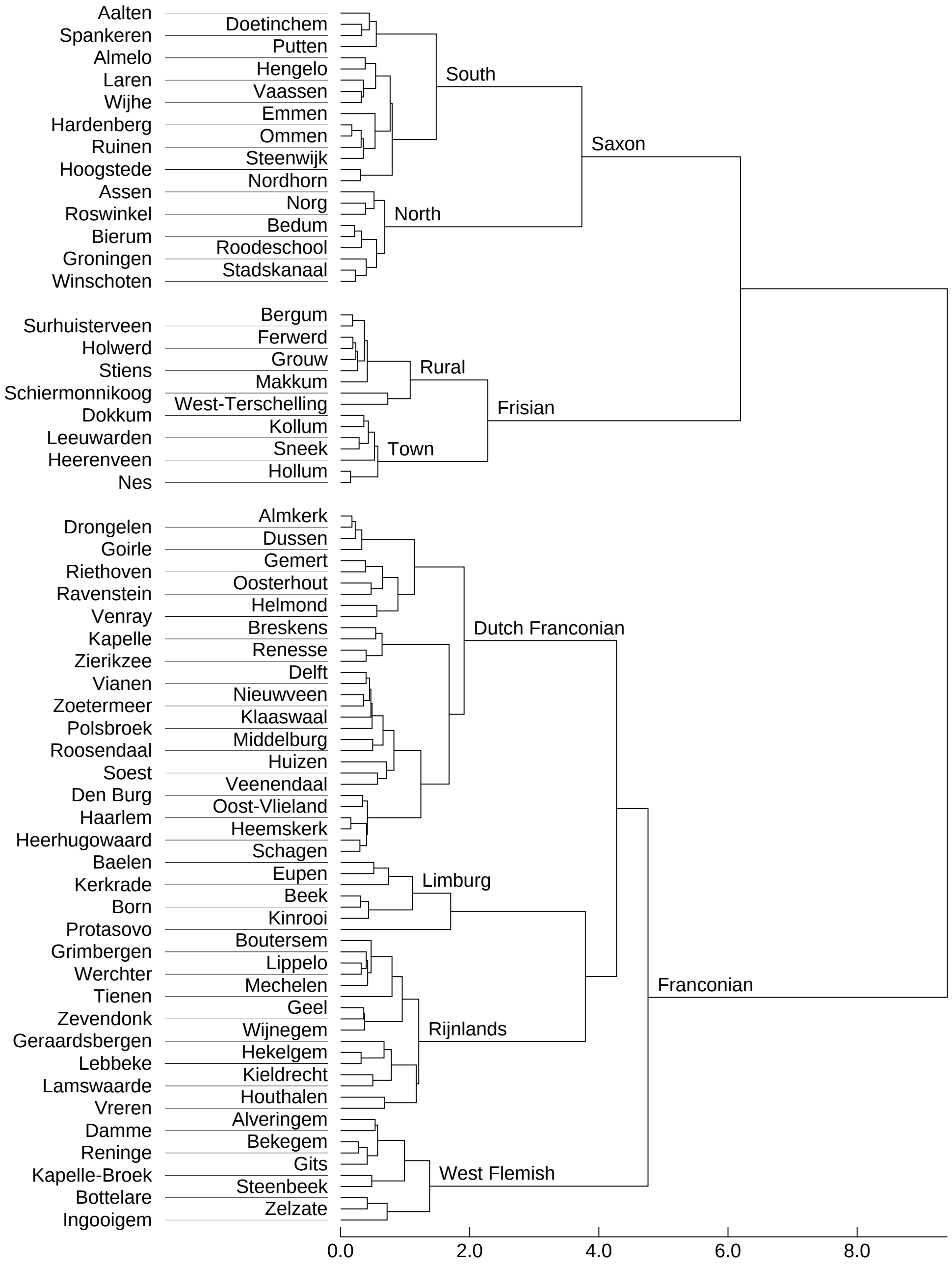
Clustering

	Assen	Delft	Kollum	Nes	Soest
Assen	0	73	64	67	79
Delft	73	0	81	74	68
Kollum	64	81	0	43	91
Nes	67	74	43	0	86
Soest	79	68	91	86	0

- Only the upper half of the matrix (blue values) is used.
- 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).

Clustering identifies groups —dialect areas?

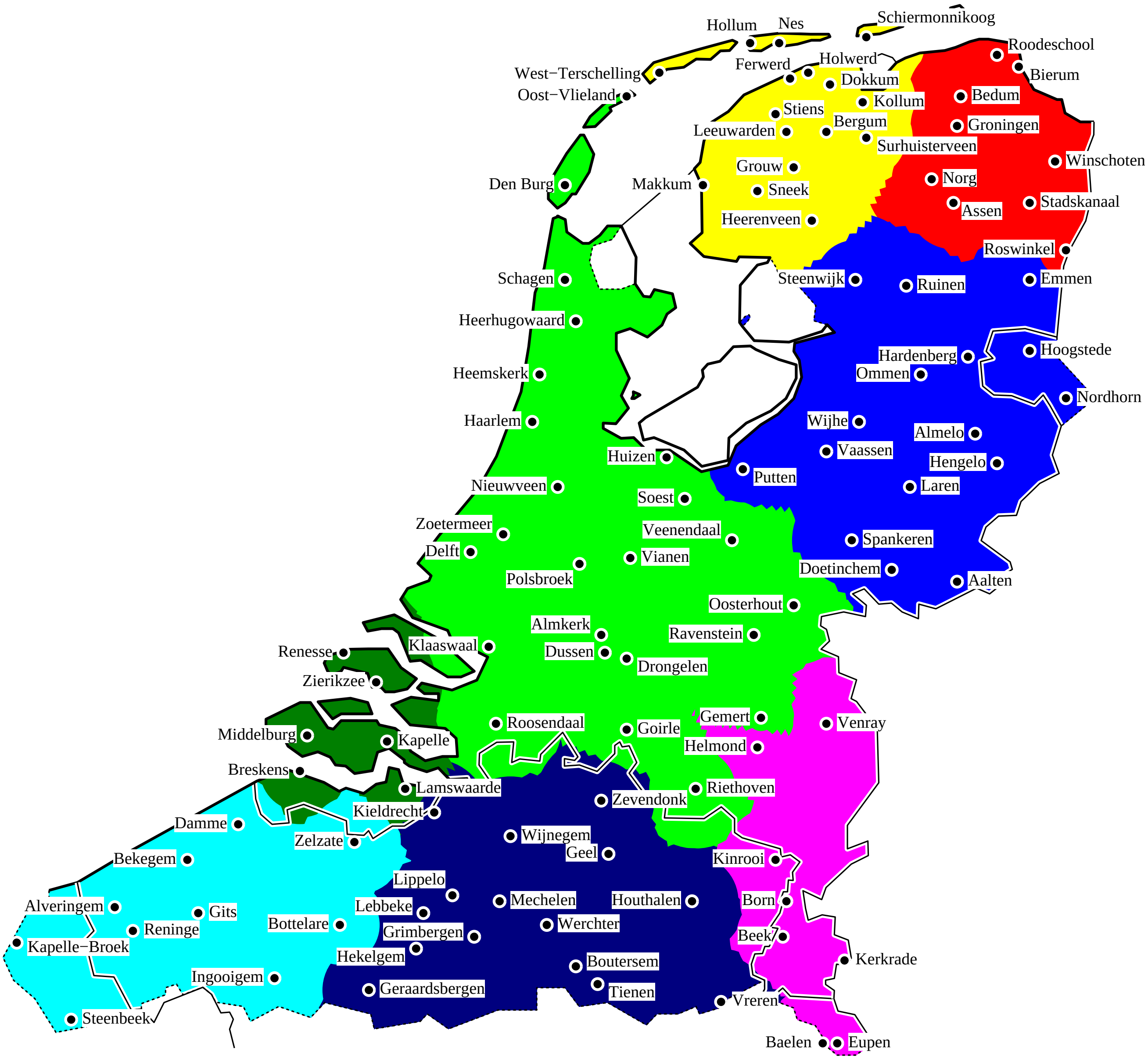
Clustering



Dendrogram derived from 104 × 104 matrix (see node-edge graphs).

Clustering

GC



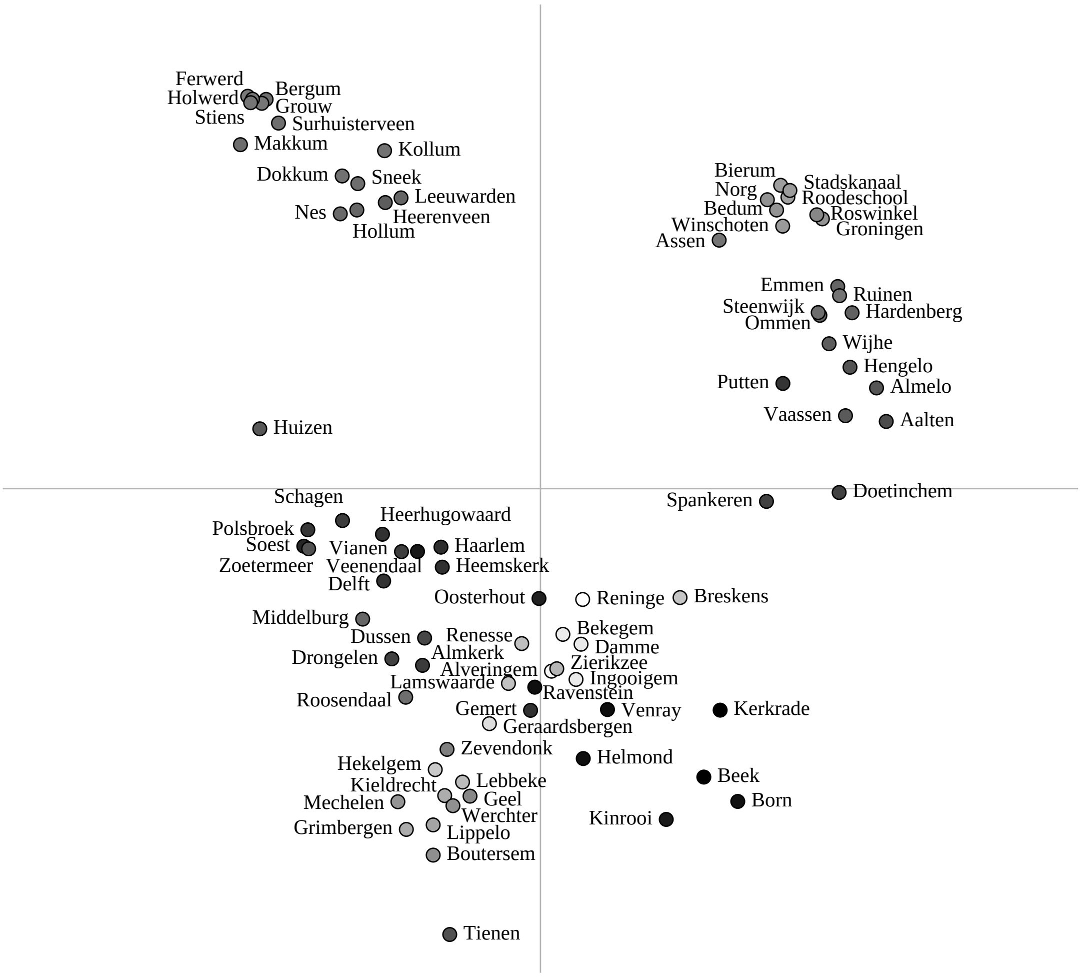
8 most significant groups in dendrogram.

RuG

Multidimensional scaling

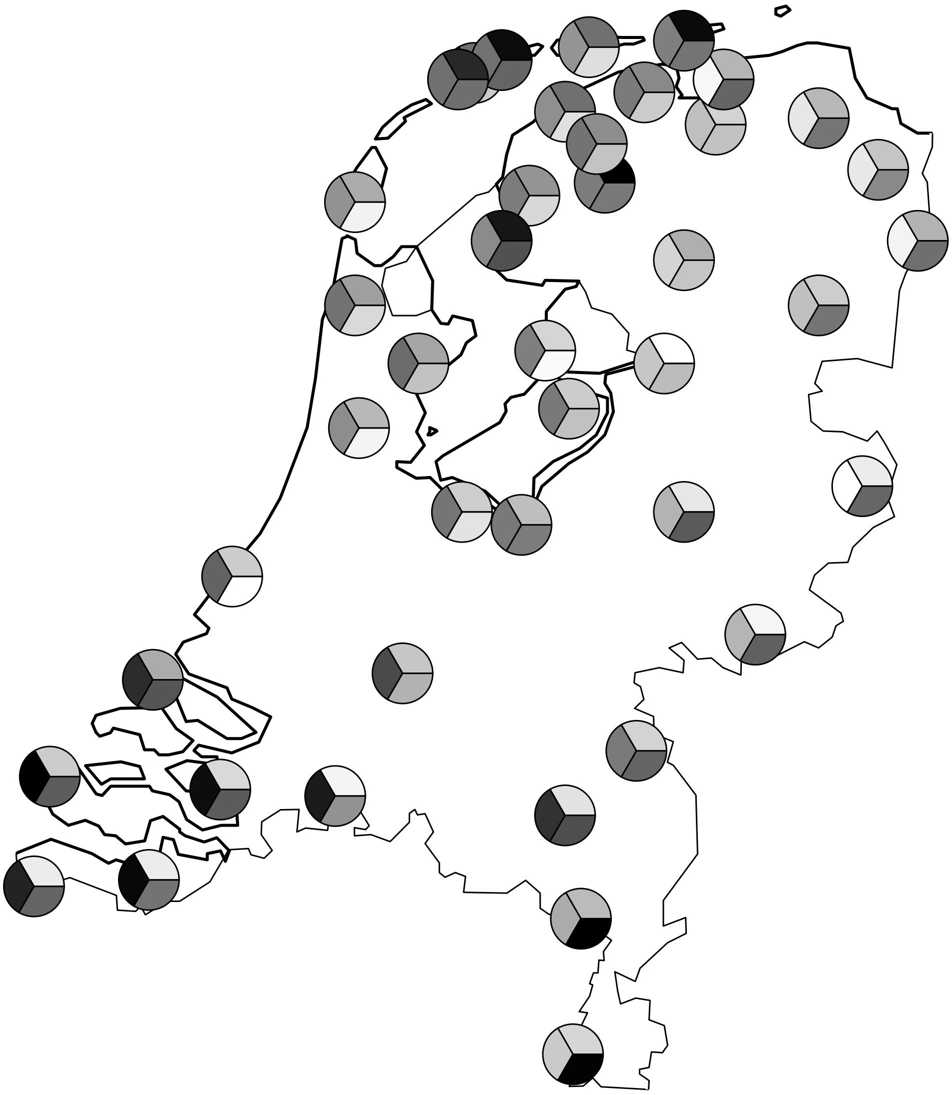
- Given a geographic map, distances between locations can be measured.
- Multidimensional scaling: given distances, locations on a map can be inferred.
- In our case: from n × n distances we infer coordinates in 2- (or 3-) dimensional space. So n dimensions are reduced to two (or three).

Multidimensional scaling



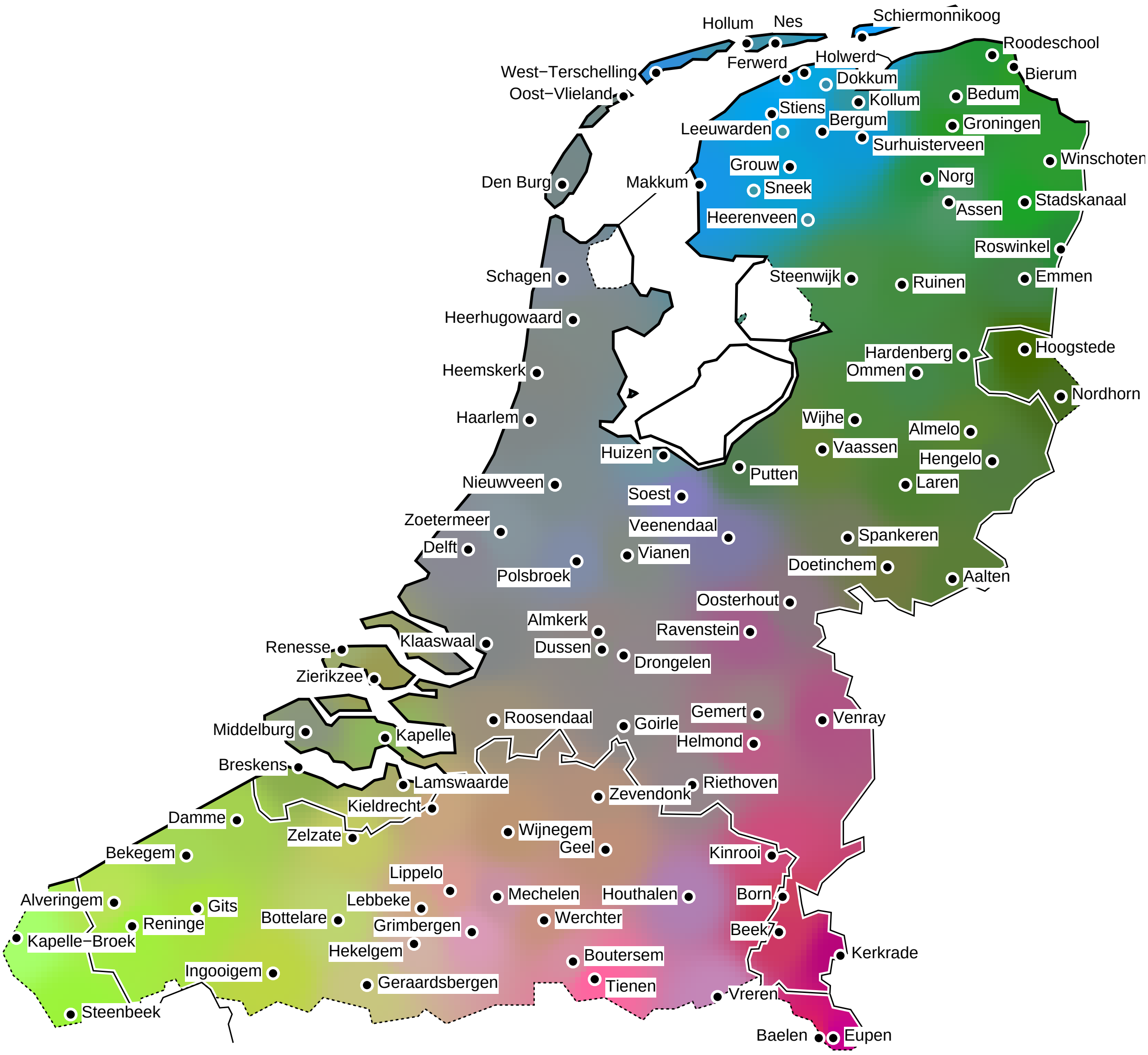
82 dimensions reduced to 3 using multidimensional scaling. x-coordinates represent the third, y-coordinates represent the first, and darkness represents the second dimension. Above left Frisian, above right the Saxon, and under Franconian dialects.

Combining Results



Show relative proportion of most significant dimensions at a range of points. Effective?

Dialect Continuum?



3 major MDS dimensions mapped to red, green and blue, and interpolated using Inverse Distance Weighting.