Strings

Languages, Texts all use STRINGS.

Lots of multimedia techniques generalize from techniques for strings.

Ergo, α-**i** looks at strings.

Some useful string algorithms.

- SOUNDEX (ca. 1918!) —variant spellings
- DICE — string similarity
- Levenshtein — string similarity, alignment

Others (Boyer-Moore string search) in more advanced courses.

SOUNDEX

problem: names are often misspelled (e.g., phone reservations)

idea: code names so that likely spelling variants are coded alike

D.Knuth *The Art of Computer Programming*, Vol.3, pp.394-95

Algorithm

Nerbonne → Nerbonn

1.	Keep first letter, drop later vowels, h,w, and y	Strings
2.	Map second and further letters to numbers using this table: p,b,f,v 1 c,k,q,g,s,z,j,x 2 d,t 3 l 4 m,n 5 r 6	N6155
3.	Simplify adjacent identical codes except where vowels originally intervened.	N615
4.	Express as four symbols either by trancating (if five or more symbols long) or by padding with zeros (if three or fewer symbols)	N615

SOUNDEX

Gauss →

1. Keep first letter, drop later vowels, h,w, and y
2. Map second and further letters to numbers using this table:

p,b,f,v	1
c,k,q,g,s,z,j,x	2
d,t	3
l	4
m,n	5
r	6

3. Simplify adjacent identical codes except where vowels originally intervened.
4. Express as four symbols either by truncating (if five or more symbols long) or by padding with zeros (if three or fewer symbols)

try some more

Gauss, Ghosh	E460
Euler, Ellery	G200
Hilbert, Heilbronn	H416
Knuth, Kant	K530
Lloyd, Ladd	L300
Lukasiewicz, Lissajous	L222

Applications

SOUNDEX used frequently

- US Census bureau 1918 $\rightarrow$?
- airline reservation systems
- Welling in digitizing records of imports *"Paalgeld"*, 1771-1817 e.g. captains' and ships' names have variant spellings

  "nominal record linkage"—linking nonnumeric fields in databases

  G.Welling *The Prize of Neutrality: Trade Relations between Amsterdam and North America 1771-1817*, Groningen, 1998.

limitations

- English spelling, pronunciation
- one take on more general problem: when are strings SIMILAR?
  - spell checker: which dictionary word is most similar (to unrecognized word)?

DICE

a **measure** of string similarity

in INFORMATION RETRIEVAL (IR), we keep track of index terms throughout a document base (maybe keywords, maybe all nouns, ...)

a document is associated with a TERM VECTOR

	computer	hospital	education	...	Tot.Refs
Doc_1	0	2	1	...	5
Doc_2	2	0	1	...	5
Doc_3	1	0	0	...	1

To measure similarity, we look at the sum of products divided by the sum of the total refs.

$$\text{Sim}(\text{Doc}_2, \text{Doc}_3) = (2 \cdot 1 + 0 \cdot 0 + 1 \cdot 0)/(5 + 1) = 1/3$$

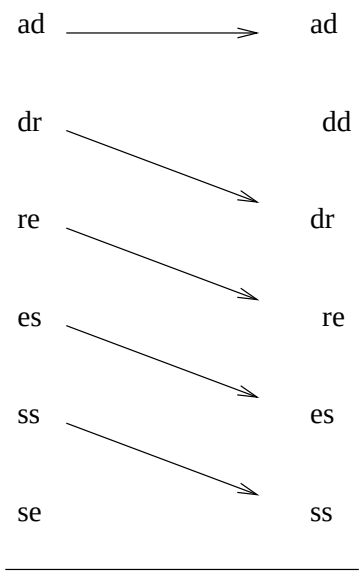
$$\text{Sim}(\text{Doc}_1, \text{Doc}_3) = (0 \cdot 1 + 2 \cdot 0 + 1 \cdot 0)/(5 + 1) = 0$$

$$\text{Sim}(\text{Doc}_1, \text{Doc}_2) = ?$$

Gerald Salton and Michael McGill *Introduction to Modern Information Retrieval* New York: McGraw Hill, 1983. pp.201-203.

DICE

to apply this to strings, look at bigrams



Totals	6	5	6
	b	a	c

$$\text{Sim}(\text{adresse}, \text{address}) = (2 \cdot 5)/(6 + 6) = 0.83$$

$$\text{Sim}(s_1, s_2) = (2 \cdot a)/(b + c)$$

Which is better?

When might you want to use SOUNDEX, when DICE?

- searching web
  how could they be used?
- spell-checking? how could they be used?

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)

- each lower right corner of table contains LevD (for strings transformed from its indices)

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

Levenshtein distance may be applied to dialect pronunciations

kœstə	delete ə	1
kœst	replace œ by ɔ	2
kɔst	insert r	1
kɔrst		
		4

Could this be interesting?

Some unsolved problems in dialectology

- What is the analytical basis of 'dialect areas'?
  Coastal New England, U.S. Southern Coastal, Saxon (Dutch)
- Can we say more precisely in what sense dialectal differences are "cumulative" (Chalmers and Trudgill)?
- How do we reconcile the notions 'dialect area' and "dialect continuum"?

Dialect Geography

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

Originally pursued (late 19th cent.) in order to see whether local linguistic change might be more phonetically regular than global change (it isn't).

Dialectology

Isoglosses are important, but insufficient for identifying DIALECT AREAS — areas with similar varieties. Bloomfield ([1]1916,1933) summarized this, but the problem was already well-known:

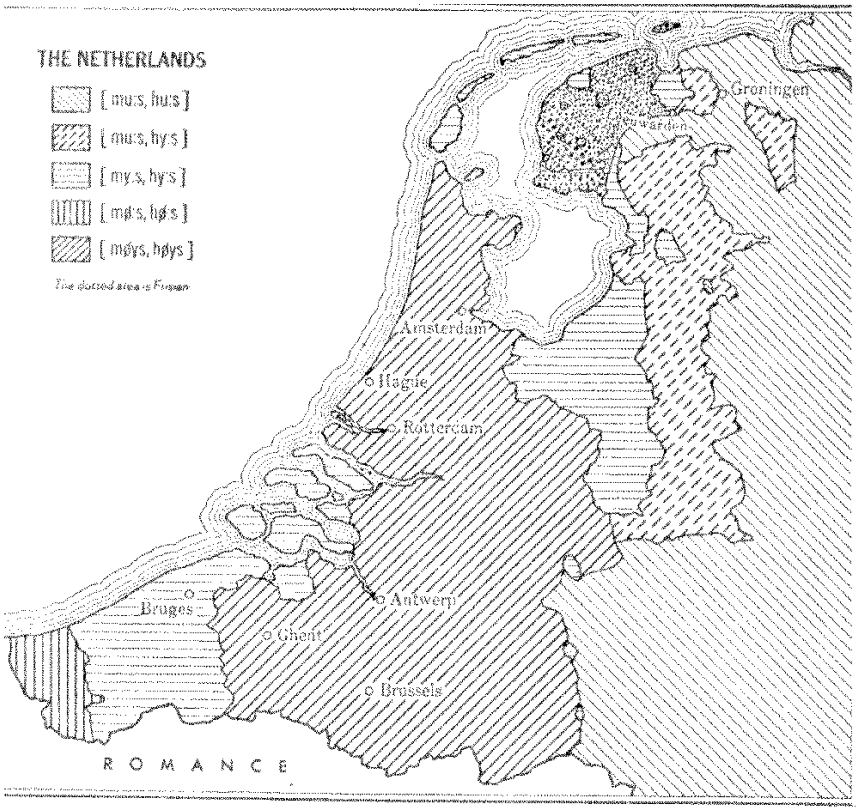


FIGURE 6. Distribution of syllabic sounds in the words *mouse* and *house* in the Netherlands. — After Klocke.

Bloomfield: "every word has its history"

Coseriu (1956): "danger of *atomistic* view"

Computational Perspective

need a way to AGGREGATE individual differences — a numerical view

- 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

Levenshtein

Average Levensthein distances between dialects. Feature system used; diphthongs represented as two phones, and Euclidean distance between feature bundles is calculated. Darker lines connect closer points, lighter lines more remote ones. Pearson's r with geographic distances is 0.6792 (significant).

Cumulativity

Chambers and Trudgill (1980) § 1.3, §§ 8.1-8.6 speculate that, although geographic distribution is irregular, it is nonetheless CUMULATIVE — geographic distance goes hand in hand with linguistic distance.

Using Levenshtein distance, we can measure the degree to which this holds:

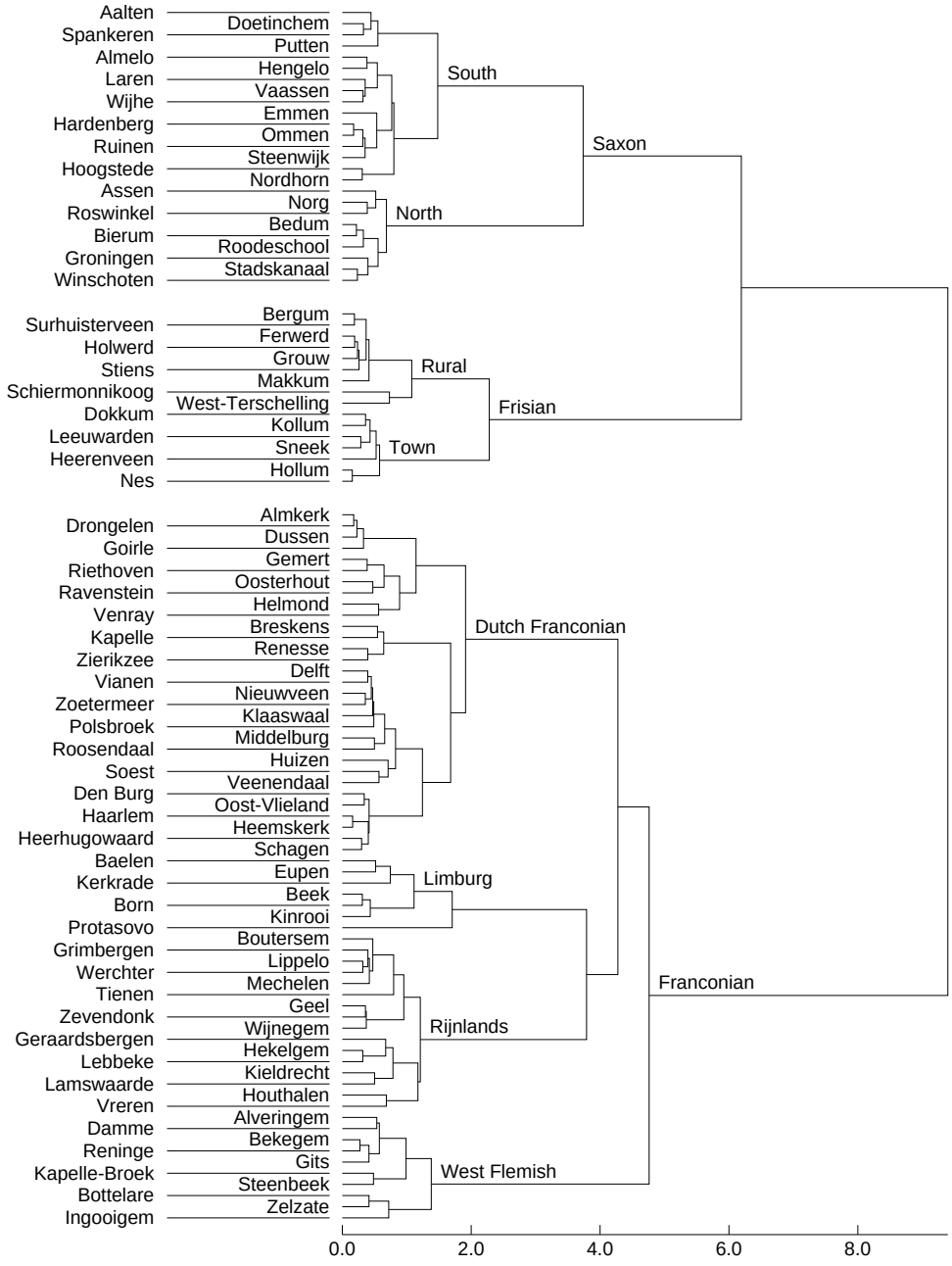
Dutch dialect distance correlates highly with geographic distance $r = 0.68$, accounting for 45% of linguistic variance (the height of parents correlates with the height of children much less $r = 0.5$)

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).

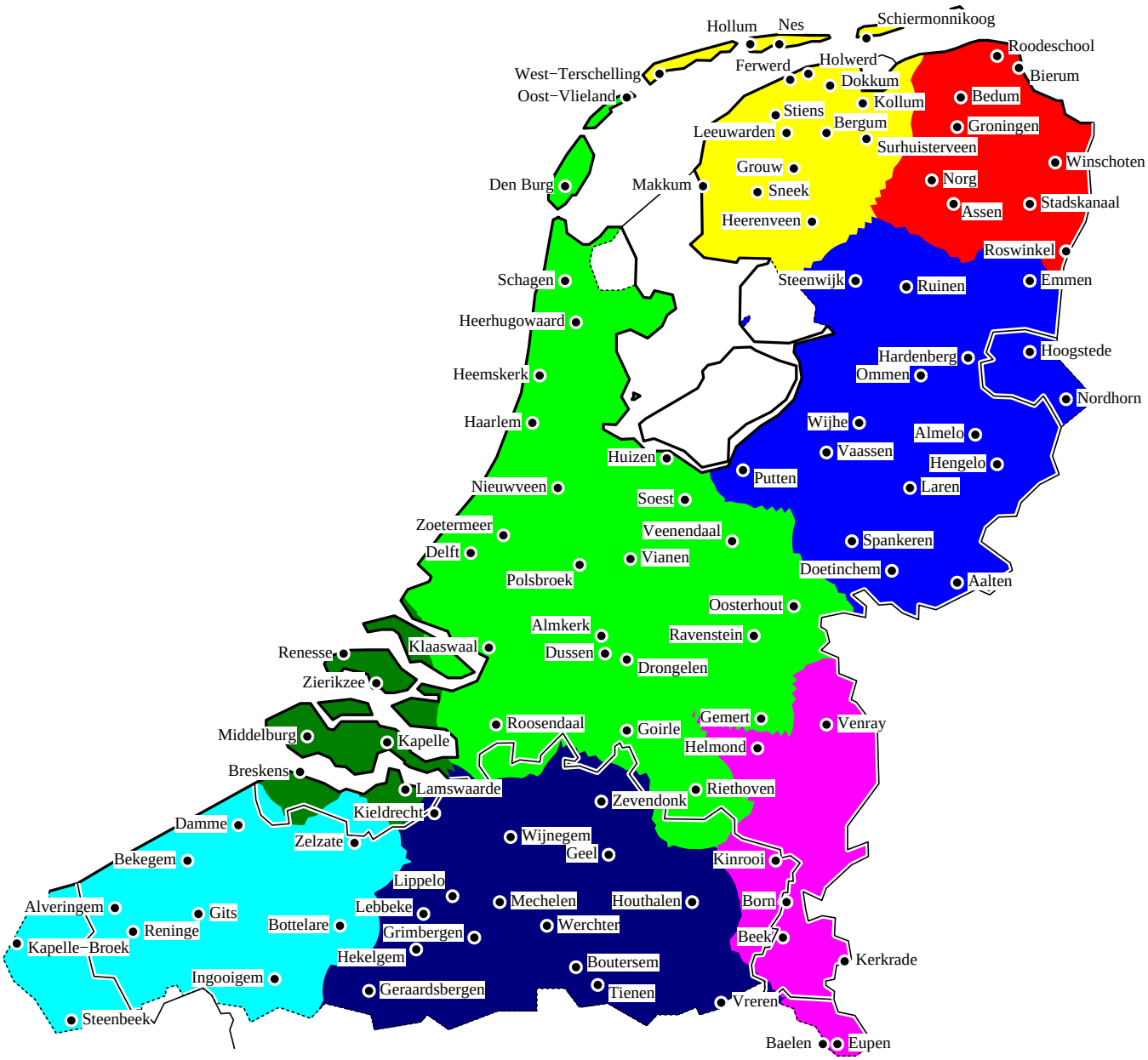
Clustering



A dendrogram derived from the 104 $\times$ 104 matrix.

Clustering

Strings



RuG

8 most significant groups in dendrogram show **areas**.

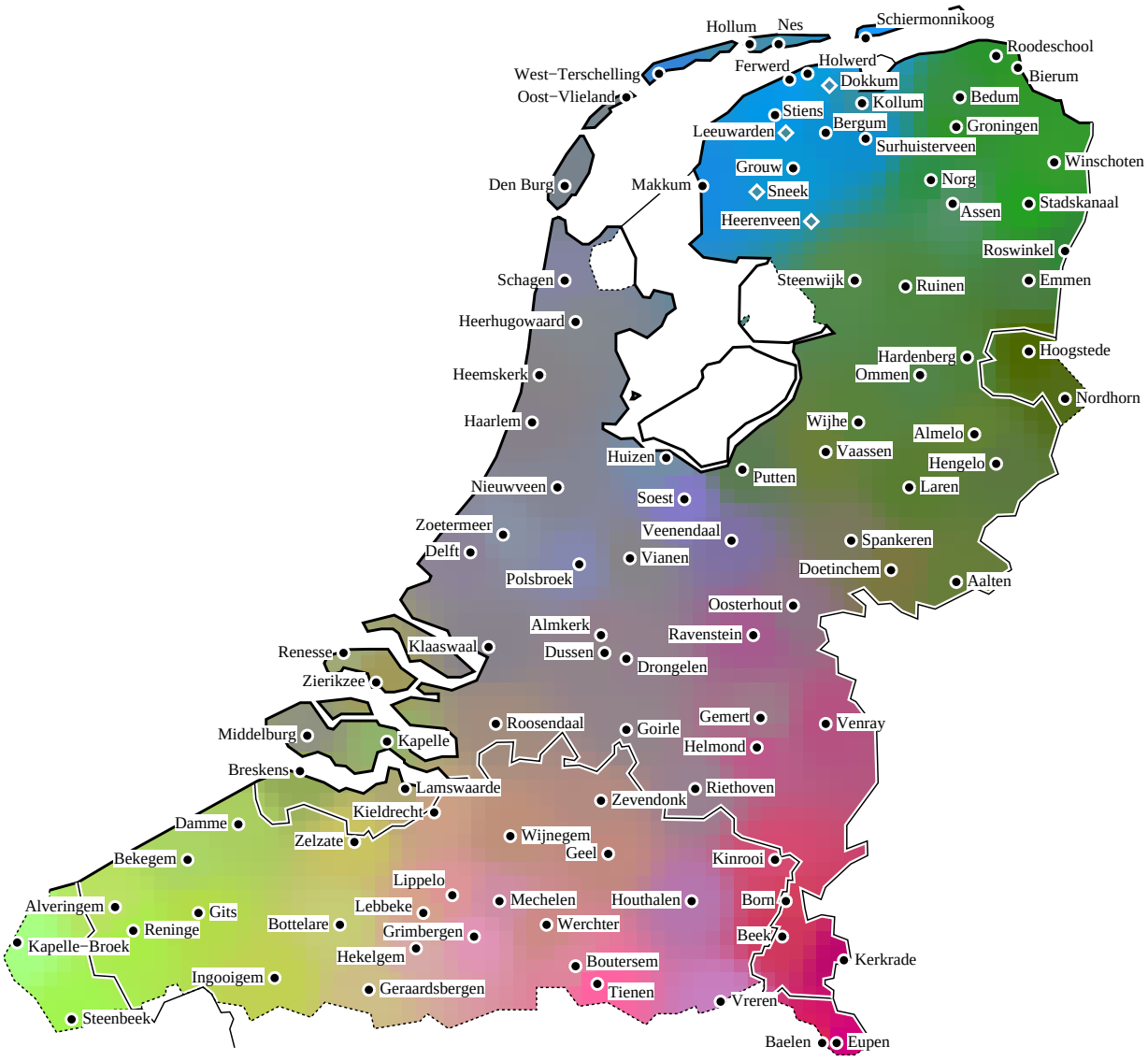
Multidimensional scaling

- 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 $\times$ 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.

Dialect Continuum?



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

Dialectology Problems

- What is the analytical basis of 'dialect areas'?
  Areas are coherent geographical regions in which language variety tends to differ (from other regions).

  Levenshtein Contribution: aggregatable measure of difference.

- In what sense are dialectal differences "cumulative" (Chalmers and Trudgill)?
  Dialects show a **strong, positive** correlation between geographical and phonetic distance.

  Levenshtein contribution: numeric measure of phonetic distance (on which to base correlation).

- How to reconcile the notions 'dialect area' and "dialect continuum"?
  Areas and **continua** are two perspectives on underlying continuously varying dialect reality.

  Levenshtein contribution: distance measure yielding continuum directly and subject to clustering to obtain areas.

Variation Linguistics

Dialectology has given way to variation linguistics, study of how language variation depends on social class, sex, age, ...

Edit distance is neutral about the external correlates of variation — a measurement, not a theory of what causes measurement differences.

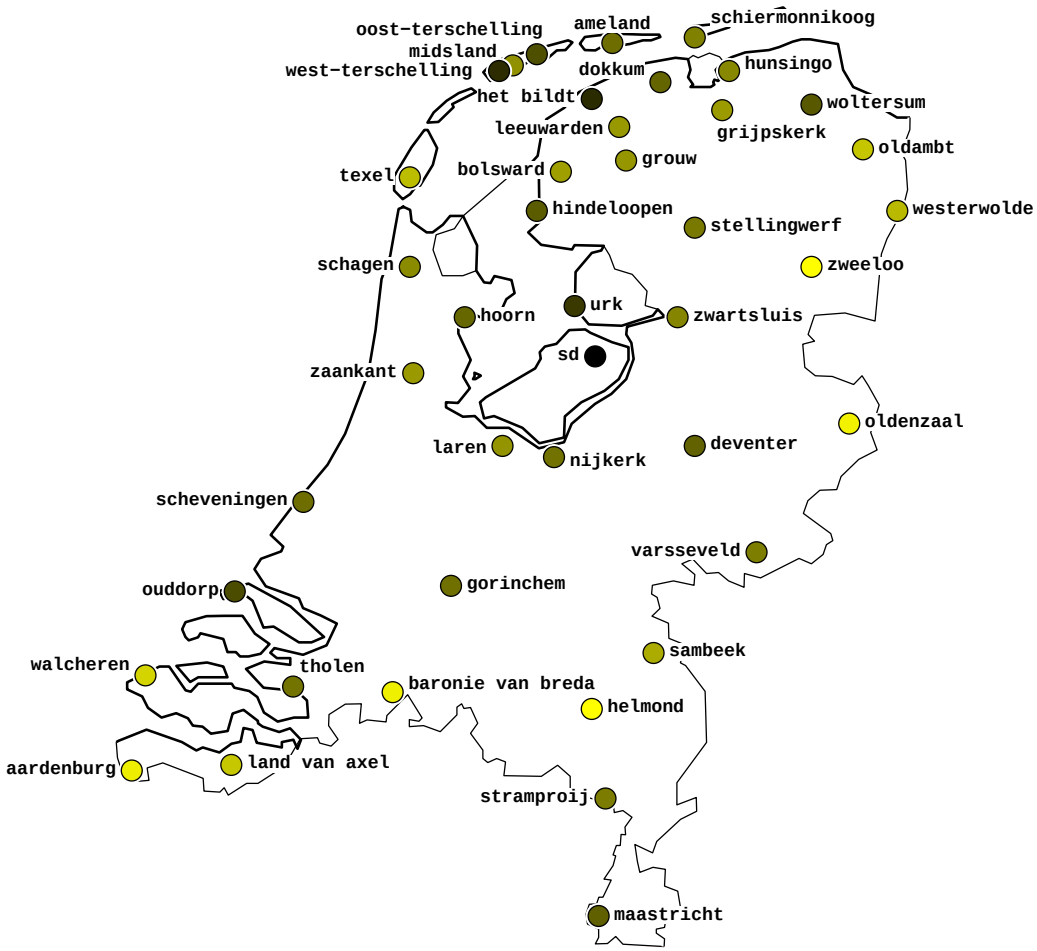
Current Topics of Investigation

- effect of standard language
- effect of political border (Bentheim)

History

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

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

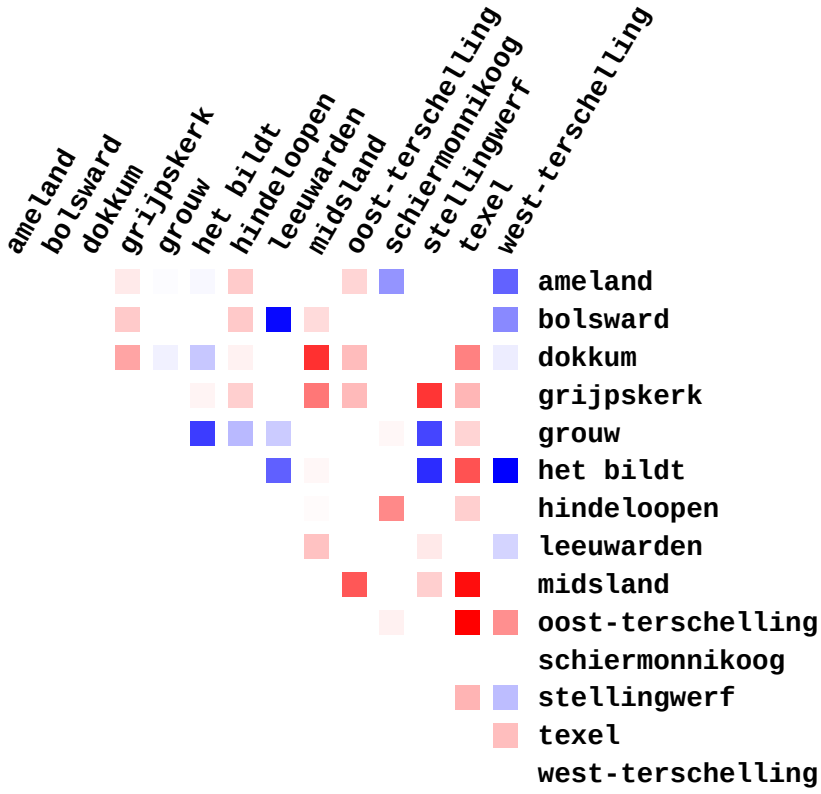


yellow indicates most extreme changes

RuG

Convergence and Divergence

We can also examine more generally which varieties became more or less alike?



Blue convergence, red divergence.

Note volatile rows (showing red and blue).

Combining Views

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