Humanities Computing: A Federation of Disciplines

John Nerbonne
Alfa-Informatica
University of Groningen
nerbonne@let.rug.nl
http://www.let.rug.nl/alfa/

Series: Is Humanities Computing an Academic Discipline?
Organized by John Unsworth
University of Virginia
Oct. 29, 1999

RuG

Federation

Thesis: Humanities Computing (HC) is not a discipline (yet), but a federation of cooperating disciplines.

a discipline is demarcated by

- a subject matter
- a range of analytical techniques
- one or more competing theories
- where appropriate, practical applications

caution: these are claims about *fields*, not about every piece of work, or every scholar's cumulative work

HC has neither coherent subject matter nor theory (apart from its components)

Vision

Humanities Computing is "...die Fortsetzung der Geisteswissenschaften mit anderen Mitteln." (with apologies to Clausewitz, Battus).

- unsurprising
- but significant
  - HC must engage traditional Humanities
  - HC's primary value is within traditional Humanities

  which traditional Humanities problems have we solved?

- *opposed* to view that HC's purpose is to understand digital culture using humanities methods
  - studies comparing printing press to computers
    Electronic Incunabula (Nerbonne, 1995)
  - linguistic studies of computer-mediated communication
  - literary studies of hypertext vs. "planar" text
  - recent proposal from Dutch Science Council—WTR

interesting, but not HC's job

Parallel?

in general, scholarship makes use of all available technology

- optical magnification
  - astronomy
  - medicine
  - biology
- photography
  - astronomy
  - biology, biokinetics
  - ethnology
- phonograph
  - acoustics
  - linguistics
  - ethnology

an essential tool is insufficient to create a field of scholarship

common subject matter, theory is essential

"Promising"

HC ought to be past the stage of a "promising" development

Humanities colleagues

- are *not* anti-technical, certainly not in general (McCarty's essay)
- are not guarding noncomputational empires
- want *results*, like all good scholars
- HC vies for attention with
  - new theoretical discussion
  - broader views
  - other interdisciplinary perspectives
  - ⋮

HC is over 30 (*Computing and the Humanities*, 1967)

—no longer a "Wunderkind"

Which traditional Humanities problems have we solved?

Engaging the Humanities

George Welling (Groningen) digitized and organized the import records (*Paalgeld*) of Amsterdam 1771-1817

computational methods deployed for organization (database) and verification (consistency) and exploration (nominal record linkage)

historical results

- Baltic trade ("moedernegotie") eclipsed by American trade even in 1771 (*pace* Israel, de Vries)
- American shipping took over Dutch business prevented by British blockade in 4th Anglo-Dutch war (1780-84)
  - American shipping catapulted to world-wide second place

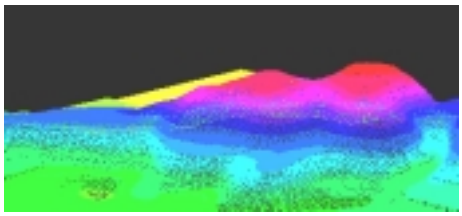
organizational effect

- active interest in HC by local historians

Art History

Elwin Koster (Groningen) has digitized and organized city maps

computational methods used to reconstruct architectural work for which plans (and buildings) were inaccessible



results in architectural history

- more complete reconstruction of urban development

a digital terrain model of 17th cent. Groningen

(Computational) Linguistics

Linguistics is universally part of Humanities

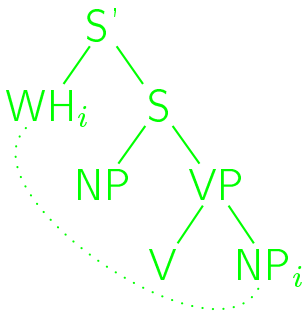
- language is a cultural product
- language is the vehicle for most elaborate and subtle cultural expression

aggressively interdisciplinary, esp. wrt psychology, cognitive science

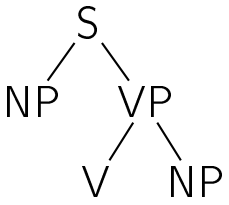
Computational Linguistics

- focus on computational language processing
- $1,500$-member prof. organization
- active collaboration with CS (50% members)

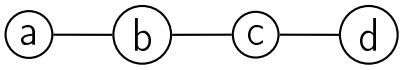
Computational Linguistics



Turing machine Undecidable



Push-down automaton Polynomial



Finite-state automaton Linear

Chomsky (1963):

> [. . .] we must conclude that the *competence* of the native speaker cannot be characterised by a finite automaton [. . .]. Nevertheless, the *performance* of the speaker or hearer must be representable by a finite automaton of some sort.

Van Noord *approximates* sophisticated grammars in FSA's

RuG

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).

Isoglosses

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

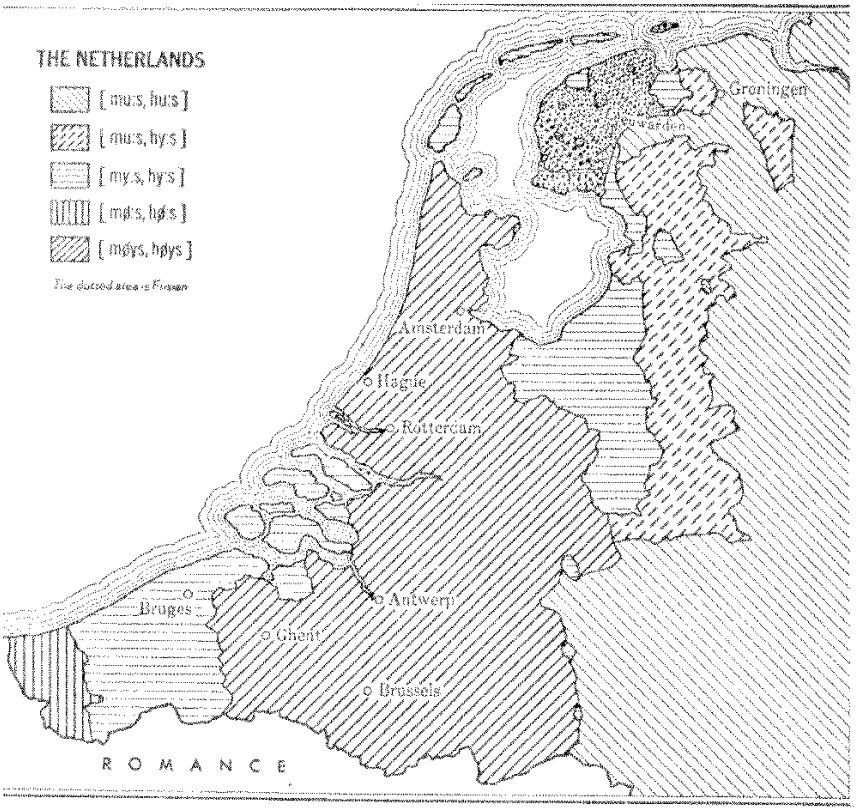


FIGURE 6. Distribution of syllabic sounds in the words *mouse* and *house* in the Netherlands. — After Kloeke.

Bloomfield: "every word has its history"

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

Linguistics

some unsolved problems in dialectology

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

Computational Perspective

need a way to AGGREGATE individual differences — a numerical view

- 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

biology 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 Pronunciations

- use 100-word sample in large number of varieties
- dialect distance is equal to the sum of the word distances — we've AGGREGATED over individual words!
- 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).
- American English example: 'saw a girl' is pronounced as [sɔːəgɪrl] (Standard American) and [sɔːrəgøːl] (Boston). Change the first pronunciation into the other.

sɔəgɪrl	delete r	1
sɔəgɪl	replace l/ø	2
sɔəgøl	insert r	1
sɔrəgøl		
		4

Levenshtein distance

- Calculate the cost of changing one string into another
- Refinement: by looking at the features the value of a replacement varies between 0 and 2. Diacritics [i̪,eː,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

Refinements

- By looking at the discrimination of the segments for each feature a weight can be calculated (Quinlan, 1993).
- Representation of diphthongs (one segment or two).
- Phonetic (Vierregge) vs. Phonological (SPE) Feature Systems

Using some feature system is a clear gain, as is attention to lexical structure (measuring distance word by word). Choice of feature system insignificant.

Two-segment representation of diphthongs preferable, but this may be artefactual.

Unclear results vis-à-vis frequency weighting.

Levenshtein

Average Levenshtein 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.

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 $r = 0.5$)

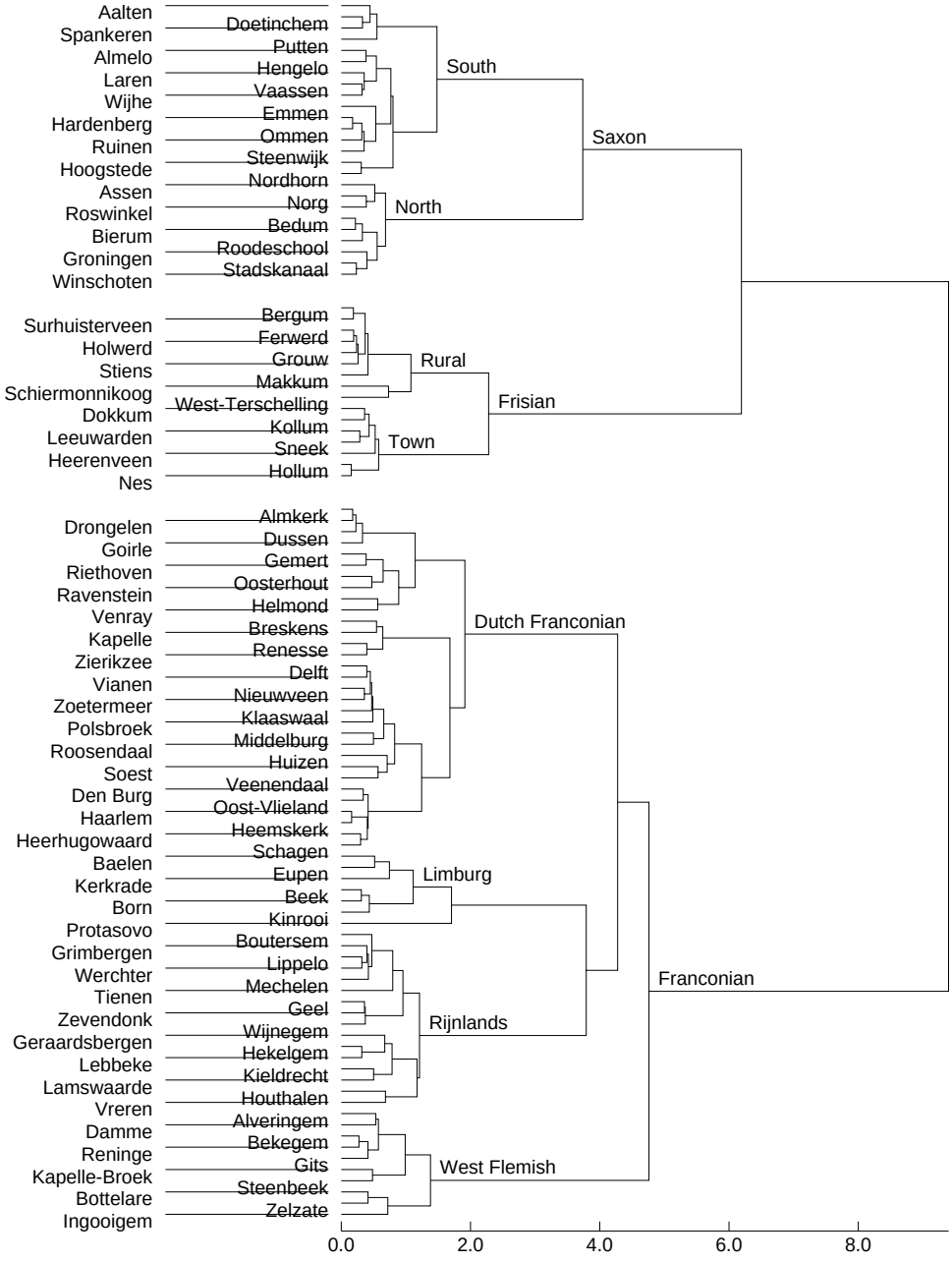
Clustering

	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 data points 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?

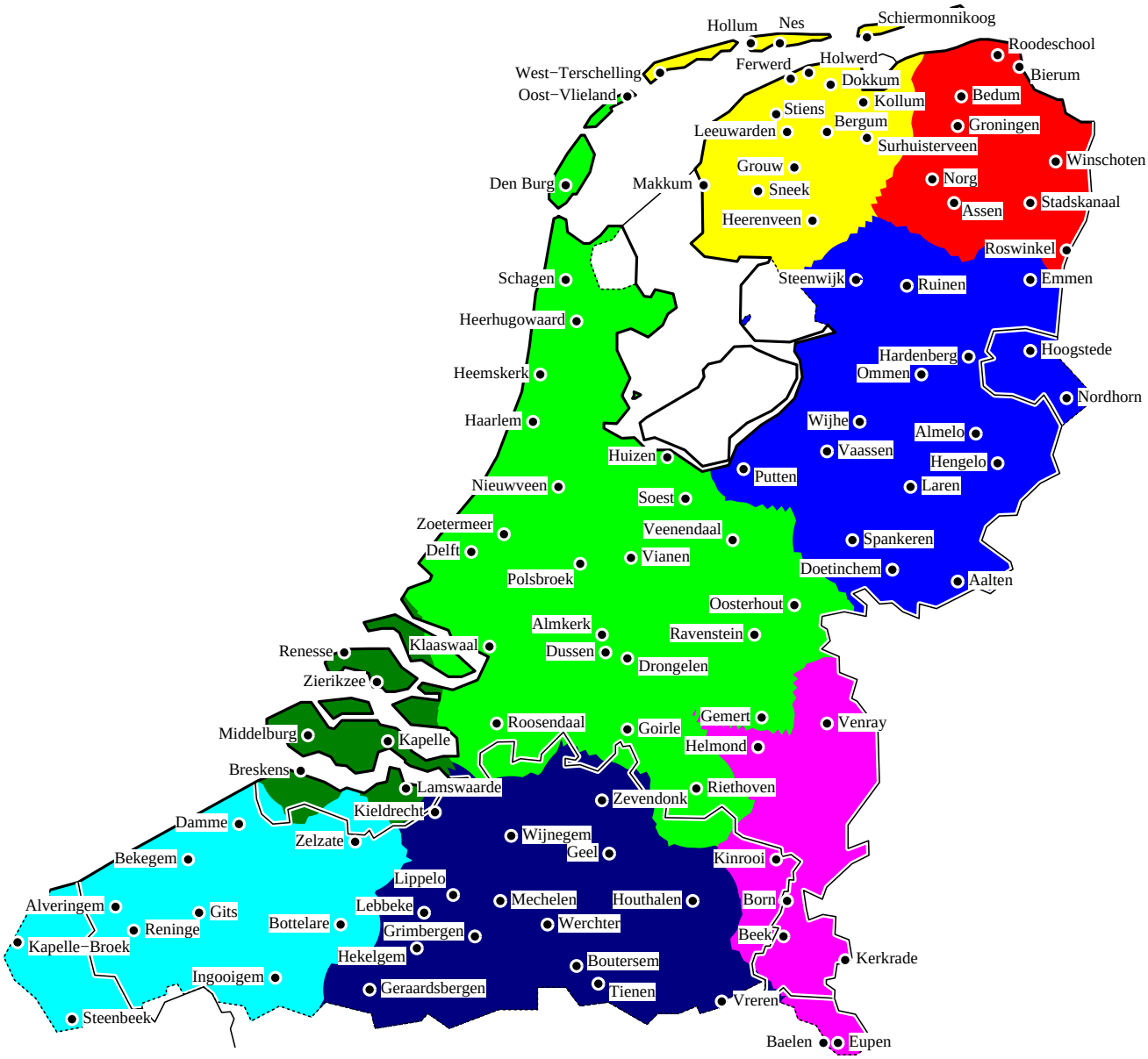
Clustering



Dendrogram derived from 104 × 104 matrix.

Clustering

Federation



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 $\times$ 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.

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.

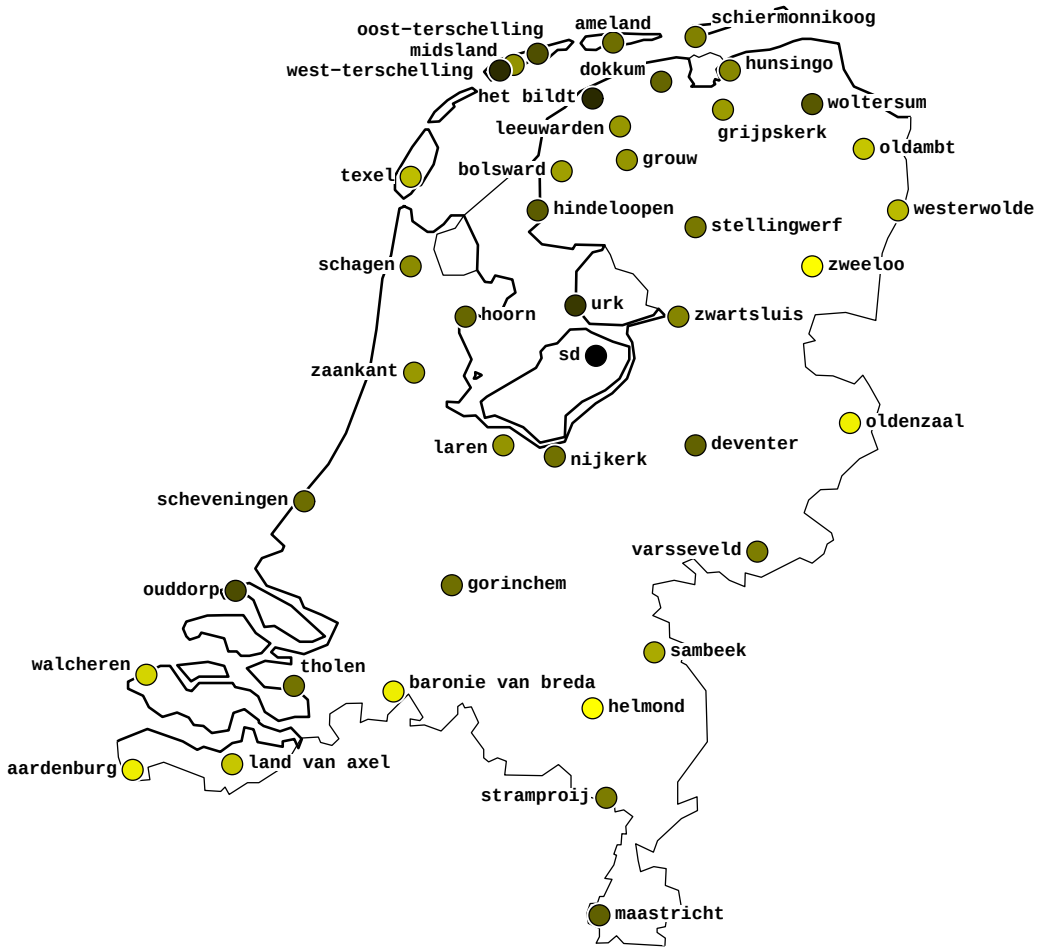
Current Topics of Investigation

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

History

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

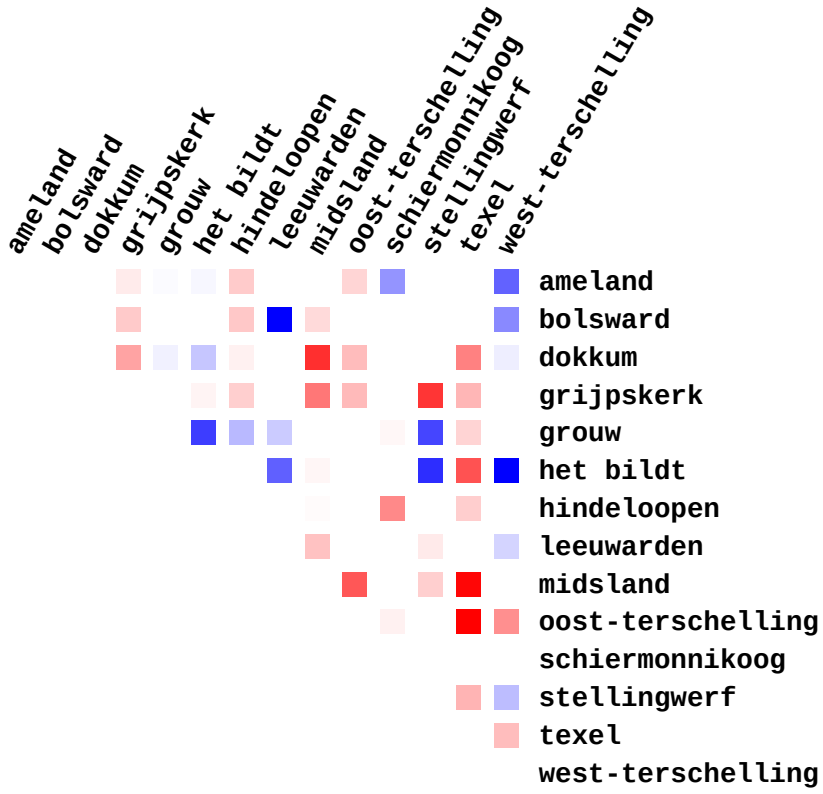
Winkler (1874) "dialect atlas" of Dutch, Flemish, Low German



yellow indicates most extreme changes

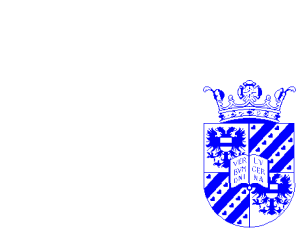
Convergence and Divergence

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



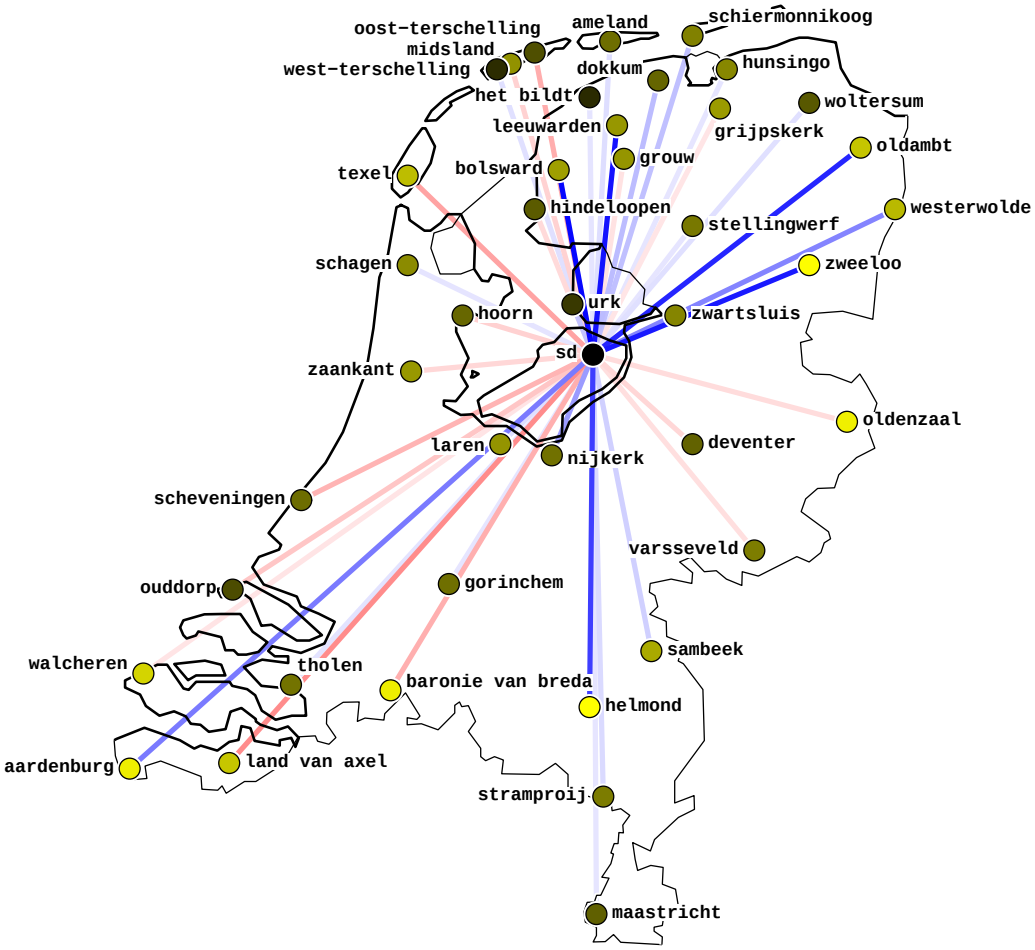
Blue convergence, red divergence.

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'.



An *Academic* Federation

Why do students study HC?

Groningen survey, 1999

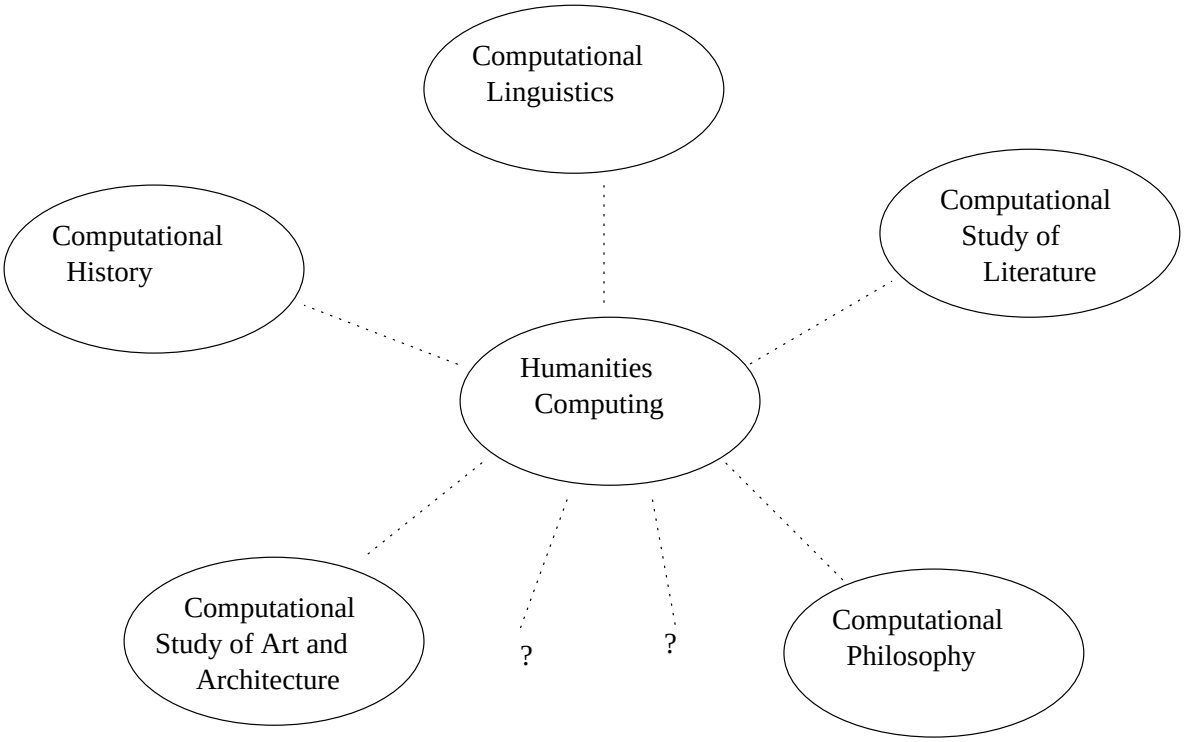
- general interest in computing, aversion to math
- attraction to more general, interdisciplinary approach
- preference for study with mix of practical and theoretical work

unflattering, but ...

- few students choose major based purely on subject matter
- students open to HC problems
- large potential pool of students

Vision

Federation of which disciplines?



Subdiscipline Dynamics

Some parent disciplines don't support the computational subdiscipline well.

- 80's grammars didn't allow good parsers (esp. because of incompleteness). Gazdar '87 (in Whitelock et al.)

  CL needed to supply its own more complete grammars
- CALL work baffled by evaluation — too little consensus in applied Linguistics about evaluating language learning. GLOSSER (see α-informatica web site).
- Unsworth's essay on the need for falsifiable statements on hypertext (see series web site).

Federation

John Hopcroft, *Turing Award Lecture*, 1986

> "The field of computer science [. . .] evolved from researchers from diverse backgrounds instead of emerging from an existing discipline." (*Comm.ACM* 30(3), '87, p.202)

Hopcroft notes contributions from Neurophysiology (McCulloch), Math (Rabin and Scott), Linguistics (Chomsky), and Electrical Engineering (his own).

More recently, Peter Denning has advocated that CS more explicitly embrace its multidisciplinary roots, which continue to be nourishing.

> "the common-sense interpretation of the computing professional [. . .] is too narrow [. . .] and it is misleading." (T.Greening (ed.) *Computer Science and Engineering Education*, reprinted in *Educom Review* Nov./Dec. '98)

> "the computing profession must embrace its boundaries with other fields to assure a constant stream of life-giving innovations" (ibid.)

Focus

If HC is to develop into a discipline, it needs more focus.

Some candidate problems to contribute to that.

What computational techniques, how much linguistic and textual structure (and how much additional expert knowledge) are needed

- to determine the language of a given text?
- to identify (candidate) glosses in bilingual texts?
- to justify treating spelling variants as the same in a set of documents (of a structure to be specified)?
- to determine the stemma of a set of manuscripts?
- to support authorship determination (among a sufficiently discriminated set of authors)?
- to improve search?
- to classify texts?
- to improve OCR?

Prospects

Hopcroft sketches his first task as ass't prof. at Princeton '64, to develop a course in automata theory, when *no one* could say exactly what it was!

> "At the time, I thought it strange that individuals were prepared to introduce courses into the curriculum without clearly understanding their content. In retrospect, I realize that people who believe in the future of a subject and who sense its importance will invest in the subject long before they can delineate its boundaries." (*ibid.*, p.199)