Syntactic variation from a quantitative perspective

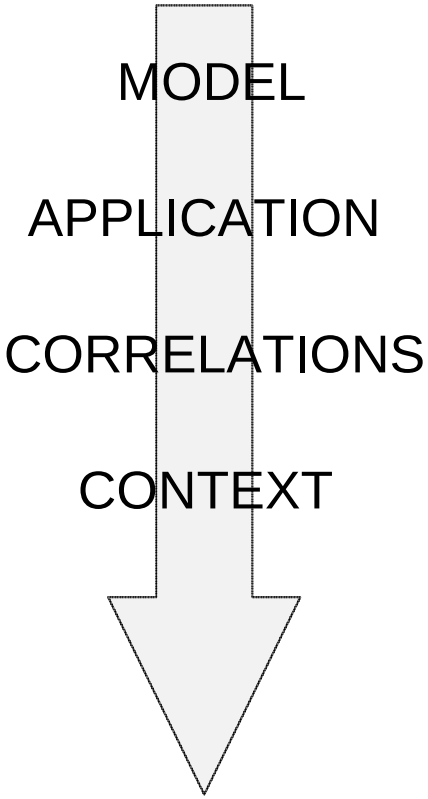
Marco René Spruit
Meertens Instituut / University of Amsterdam
http://www.meertens.knaw.nl/medewerkers/marco.rene.spruit
Trieste, May 3, 2006

1a) Research context

- The Determinants of Dialectal Variation project (DDV)
  - http://dialectometry.net
  - University of Groningen: information science
    - John Nerbonne
    - Wilbert Heeringa
  - Meertens Instituut: syntactic theory
    - Hans Bennis
    - Sjef Barbiers

1b) Research questions

1. What is a syntactic measure?
2. What are the syntactic distances between the Dutch dialects?
3. What are relevant dependencies between syntactic variables?
4. How does the geographic distribution of syntactic variables in Dutch dialects relate to the distribution of dialect differences on other linguistic levels?



1c) Some answers

- Dialectometric methods *can* be successfully applied to syntactic data
- There *is* geographic cohesion in syntactic variation
- There *are* significant correlations between the syntactic, perceptual and pronunciation levels

1d) Presentation outline

1. Introduction
2. Syntactic variation data
3. Dialectometric methods
4. Interpretation of results
5. Visualisation of dialect relationships
6. Reliability of results
7. Measure refinements
8. Syntactic variation in context
9. Variable correlations
10. Conclusions and future research

- Relevant software and data formats

2) Syntactic variation data

- Syntactic Atlas of the Dutch Dialects (SAND)
  - 267 Dutch dialects
  - SAND1: [Barbiers et al. 2005]

    Complementisers, Subject pronouns, Expletives, Subject doubling, Subject clitisation following yes/no, Reflexive and reciprocal pronouns, Fronting
    - 134 syntactic contexts, 507 variables
  - SAND2: [Barbiers et al. 2007]

    Verbal clusters, Cluster interruption, Morphosyntactic variation, Negative particle, Negative concord and quantification
    - 65 syntactic contexts, 274 variables *(incomplete)*

2a) SAND1 domains

SAND1: 134 syntactic contexts, 507 variables

1. Complementisers
   - "it looks AFFIRM **if** there someone in the garden stands"
2. Subject pronouns
   - "she believes that **you** earlier home are than I"
3. Expletives
   - "**there** sat a burglar in this closet"
4. Subject doubling
   - "if you$_{\text{weak}}$**you**$_{\textbf{strong}}$ healthily live,
     live you$_{\text{weak}}$**you**$_{\textbf{strong}}$ longer"

2a) SAND1 domains

SAND1: 134 syntactic contexts, 507 variables

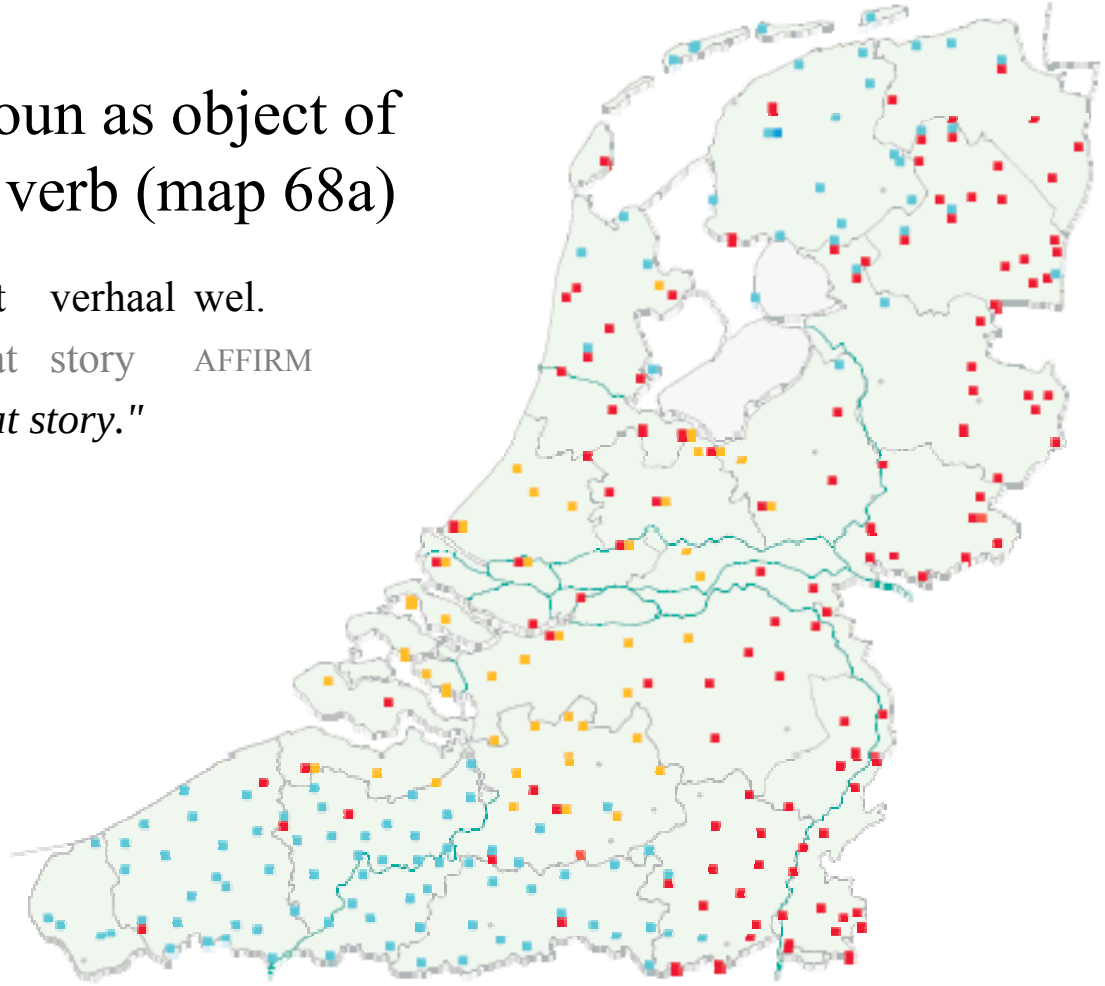
5. Subject clitisation following yes/no
   - "*Q*: did they already eat? *A*: yes**-3plur**"
6. Reflexive and reciprocal pronouns
   - "john remembers **himself** that story AFFIRM"
7. Fronting
   - "that is the man **who** they called have"

2a) SAND1 data sample

Weak reflexive pronoun as object of inherent reflexive verb (map 68a)

Jan	herinnert	**zich**	dat	verhaal	wel.
John	remembers	himself	that	story	AFFIRM

"John certainly remembers that story."



2b) SAND2 domains

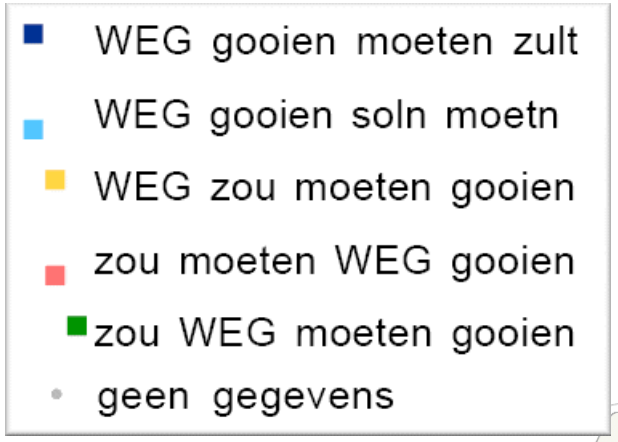
SAND2: 65 syntactic contexts, 274 variables

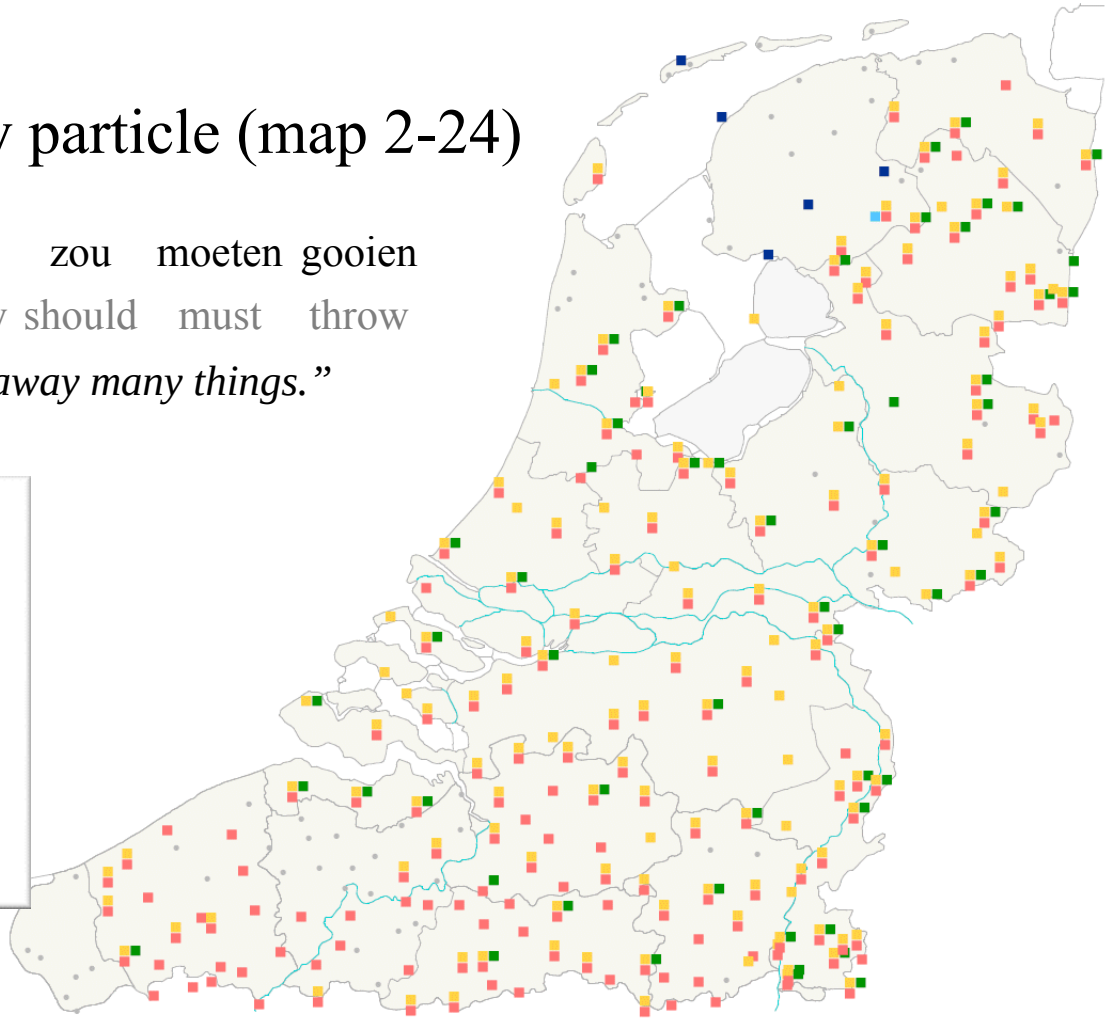
1. Verbal clusters (12, 62)
   - "Ik weet dat hij *is weeste zwemmen*"
2. Cluster interruption (10, 25)
   - "Ik denk dat je veel zou *weg* moeten gooien"
3. Morphosyntactic variation (12, 80)
   - "Niemand heeft dat ooit *wild* of *kund"*
4. Negative particle (8, 15)
   - "Els *en* wil *niet* zingen"
5. Negative concord & quantification (23, 92)
   - "Q: Wie heeft de auto meegenomen? A: *Niemand niet*."

2b) SAND2 data sample

Cluster interruption by particle (map 2-24)

Ik denk dat je veel **weg** zou moeten gooien
I think that you much away should must throw
"I think that you should throw away many things."





3) Dialectometric methods

- A *quantitative* research perspective
  - Assign *numerical* values to linguistic variables
  - Using a *measure* of linguistic distance
  - *Add up* individual variables to *objectively* arrive at more general description (versus interpreting isogloss bundles)
  - Examine *aggregated* differences between language varieties
- KEY: From measuring individual linguistic variables (qualitative) to aggregated differences between language varieties (quantitative)

3a) Syntactic variables

- Linguistic units in which two language varieties can vary
- In this work: Forms or word orders in a *syntactic context* in which two dialects can differ
  - Atomic variables

    syntactic variables as they have been recorded, without interpretations
  - Feature variables

    syntactic variables with manually annotated linguistic feature information (obtained after a syntactic analysis)
  - Composite variables

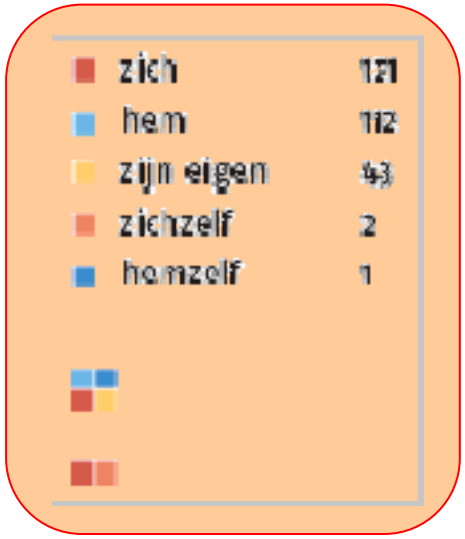
    collections of syntactic variables with nearly identical geographical distributions

3a) Syntactic context and variables

Weak reflexive pronoun as object of inherently reflexive verb (map 68a)

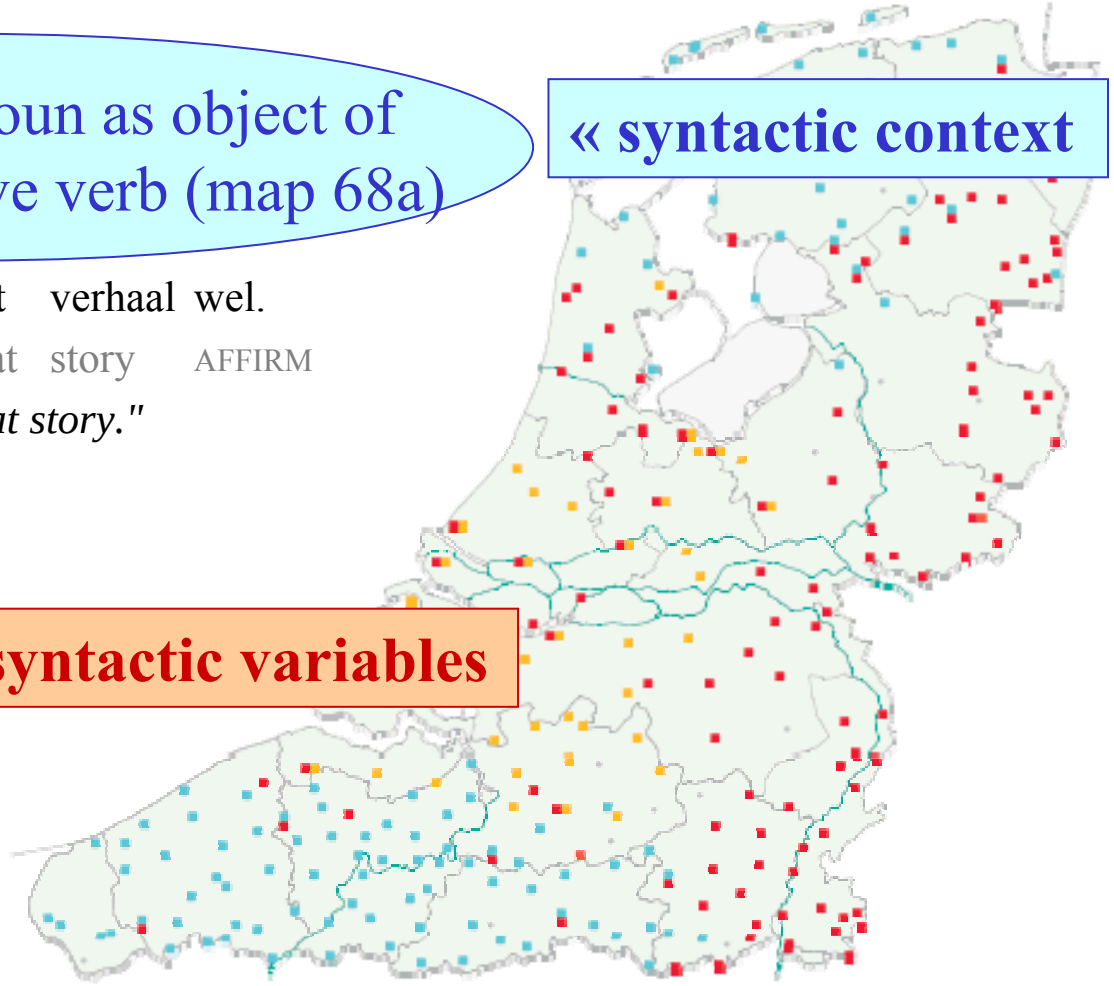
Jan	herinnert	**zich**	dat	verhaal	wel.
John	remembers	himself	that	story	AFFIRM

"John certainly remembers that story."



« syntactic context

« syntactic variables



3b) Measure of syntactic distance

- Hamming distance algorithm based on binary comparisons between atomic variables
- For example, the syntactic context *Cluster interruption by particle*:

"Ik denk dat je veel *weg* zou moeten gooien"/"*I think that you should throw away many things.*"

variables	*Lunteren*	*Veldhoven*	*distance*
WEG zou moeten gooien (1)	√	√	0
zou WEG moeten gooien (2)	√		1
zou moeten WEG gooien (3)	√	√	0
WEG gooien moeten zult (4)			0
WEG gooien soln moetn (5)			0

1

3c) Syntactic distance matrix

dialect	Lunteren	Bellingwolde	Hollum	Doel	Sint-Truiden	Veldhoven
Lunteren		37	50	47	40	**24**
Bellingwolde	37		31	50	55	41
Hollum	50	31		53	56	48
Doel	47	50	53		39	51
Sint-Truiden	40	55	56	39		44
Veldhoven	**24**	41	48	51	44	

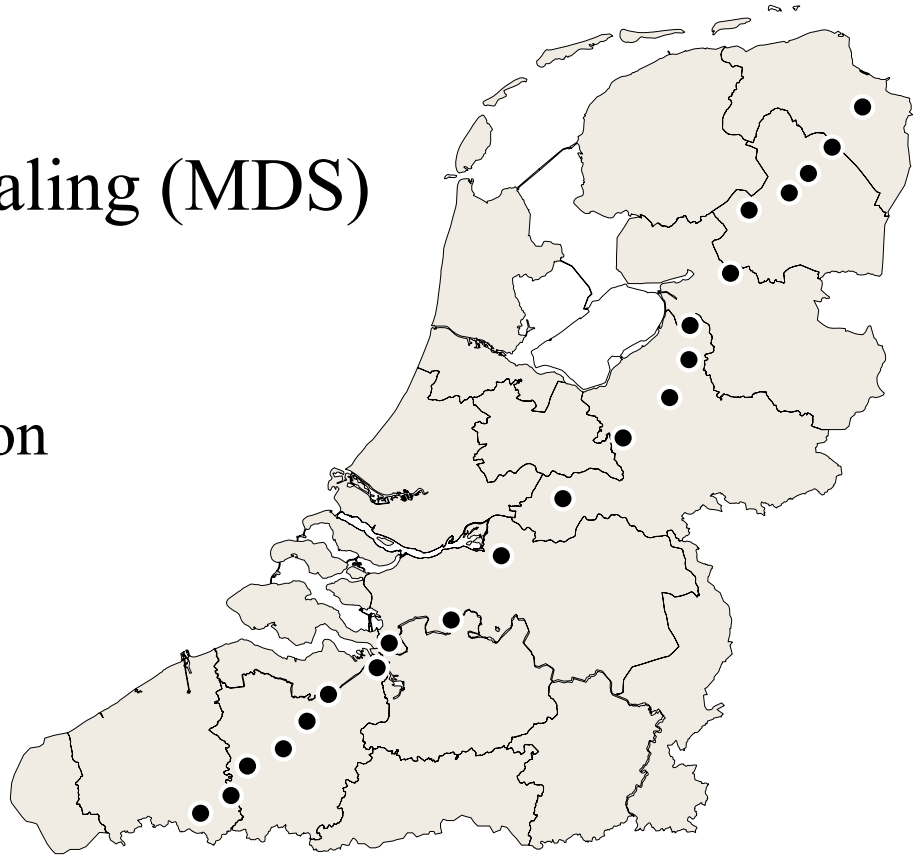
4) Interpretation of results

a) Cluster analysis
   - Dendrogram

b) Multidimensional scaling (MDS)
   - Generic MDS plot

c) Topological maps
   - Delauney triangulation
   - Voronoi polygons

1) Cluster maps
2) MDS maps
3) Hybrid maps
4) Barrier maps



4a) Cluster analysis

- Iterative procedure

  Each iteration, the two most similar/less distant locations are merged until one aggregated cluster of locations remains

  1. Find the smallest distance in the matrix
  2. Merge the two locations into one new location
  3. Calculate the distances from all dialect locations to the new location
- Using Ward's method for step 3

  equally sized clusters, suited for initial data exploration
- Analysis will *always* result in clusters

  whether they actually are "natural classes" or not..

4a) Cluster analysis iteration

Iteration *n*

location	A	B	C	D
A		1	3	6
B	1		2	5
C	3	2		3
D	6	5	3	

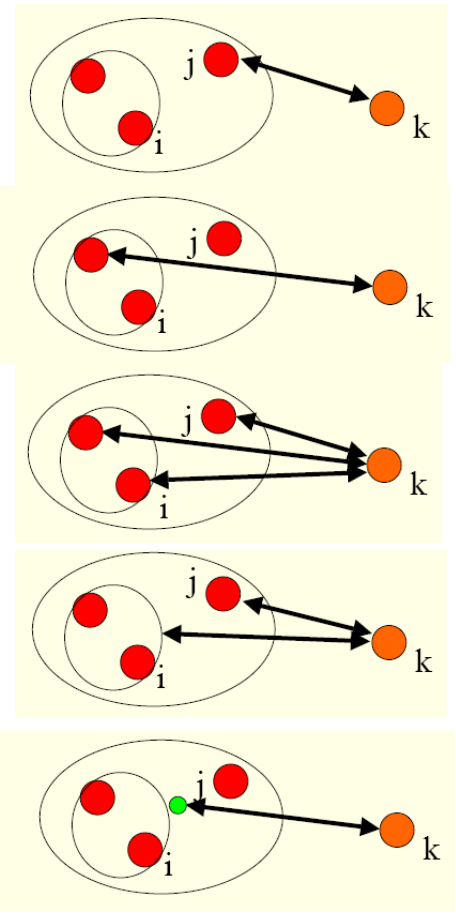
Next iteration

location	**AB**	C	D
AB		**2,5**	5,5
C	**2,5**		3
D	5,5	3	

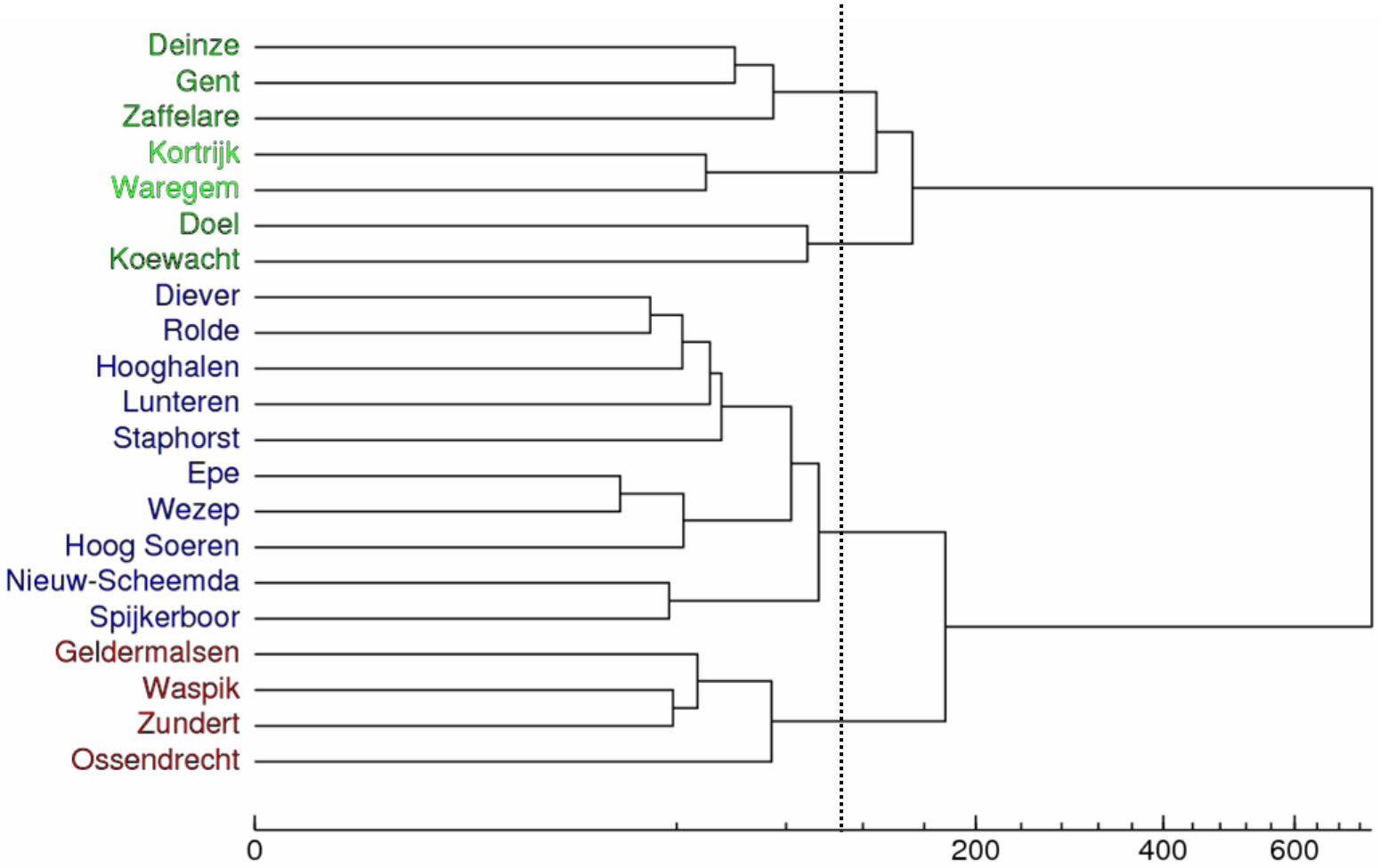
Iteration *n+1*

4a) Cluster update algorithms

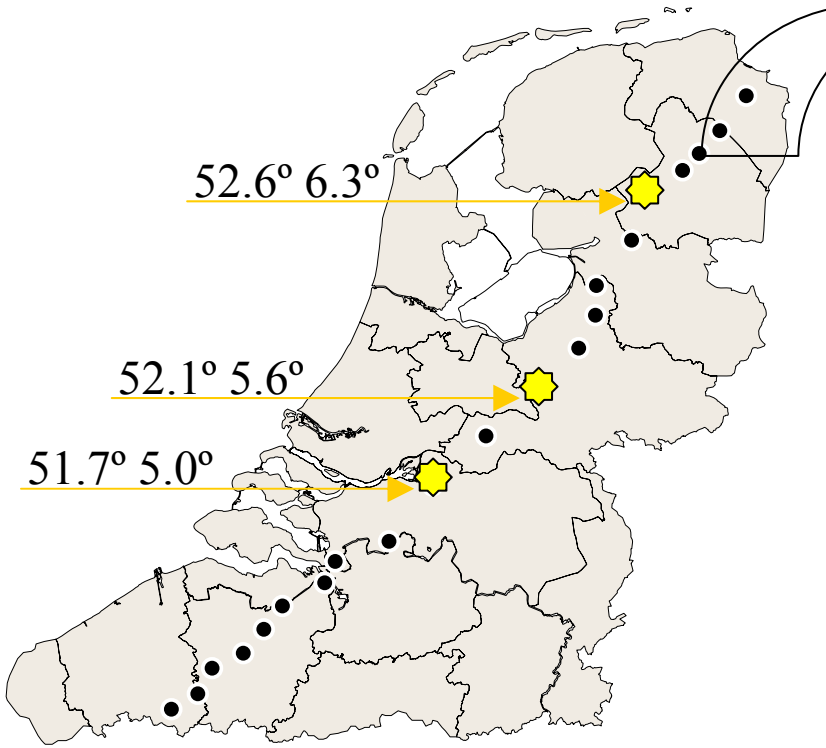
- Ways to calculate the distances from the merged locations to all other locations:
  - Single-link (nearest neighbour) / Complete-link (furthest neighbour)
  - (Un)weighted Pair Group Method using Centroids (UPGMC/WPGMC)
  - (Un)weighted Pair Group Method using Arithmetic averages (UPGMA/WPGMA)
  - Ward's method (minimum variance)



4a) Dendrogram



4b) Multidimensional scaling (MDS)

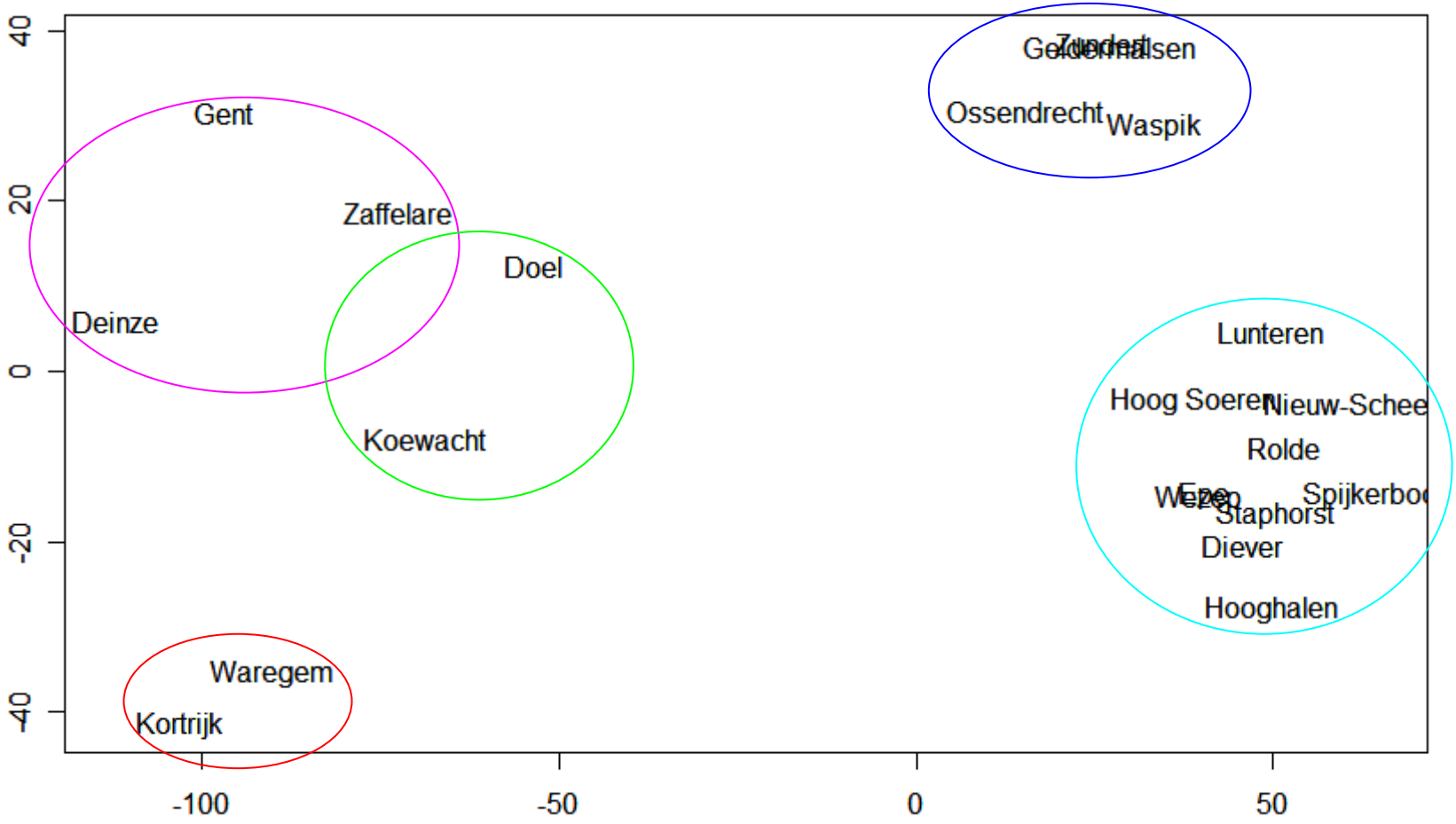


Get coordinates to determine the distance between locations

location	Diever	Lunteren	Waspik
Diever		114.8	199.0
Lunteren	114.8		86.4
Waspik	199.0	86.4	

In MDS: Get distance between locations to determine the coordinates...

4b) MDS plot

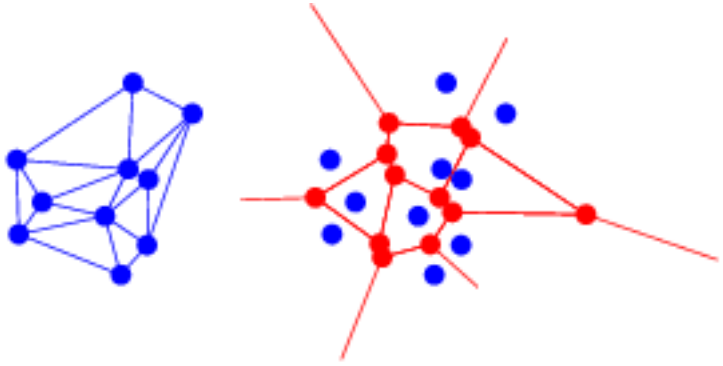

Gent
Zaffelare
Doel
Deinze
Koewacht
Ossendrecht
Waspik
Lunteren
Hoog Soeren
Rolde
Staphorst
Diever
Hooghalen
Waregem
Kortrijk
40
20
0
-20
-40
-100
-50
0
50

4b) MDS algorithms

- Classical Multidimensional Scaling
  - Metric: uses the actual distance values
- Kruskal's Non-metric Multidimensional Scaling
  - Non-metric: uses the ranks of the distances
  - Enlarges differences
- Sammon's Non-Linear Mapping
  - Also non-metric
  - Preserves small distances

4c) Topological maps

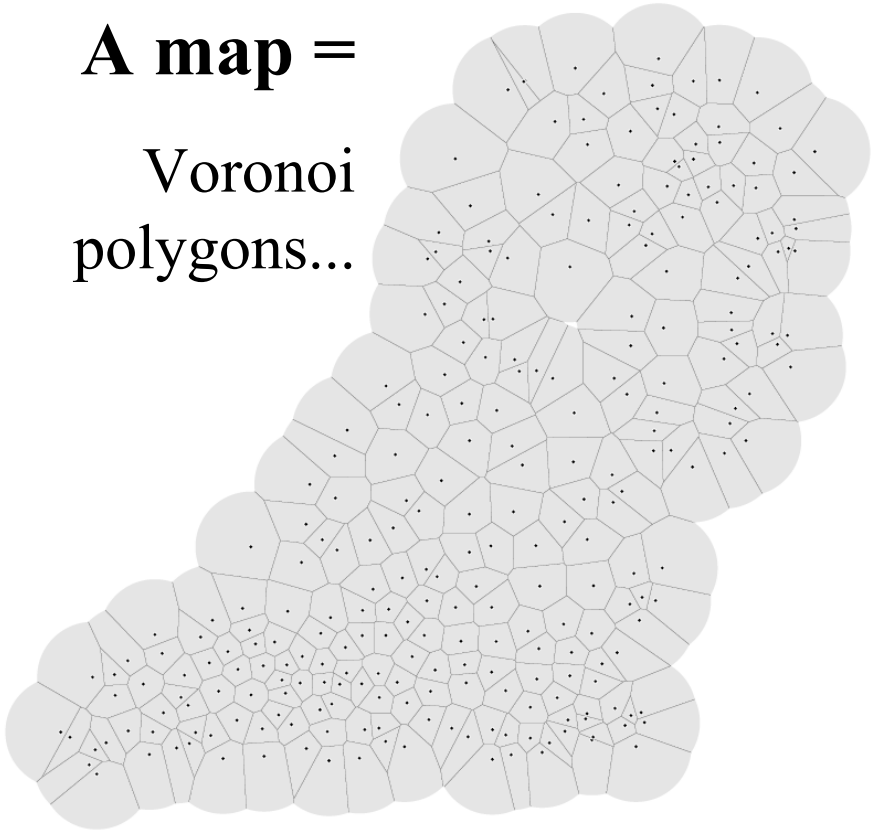
- Delaunay triangulation
  - Voronoi polygons
  - versus interpolation



Delaunay triangulation *Voronoi diagram*

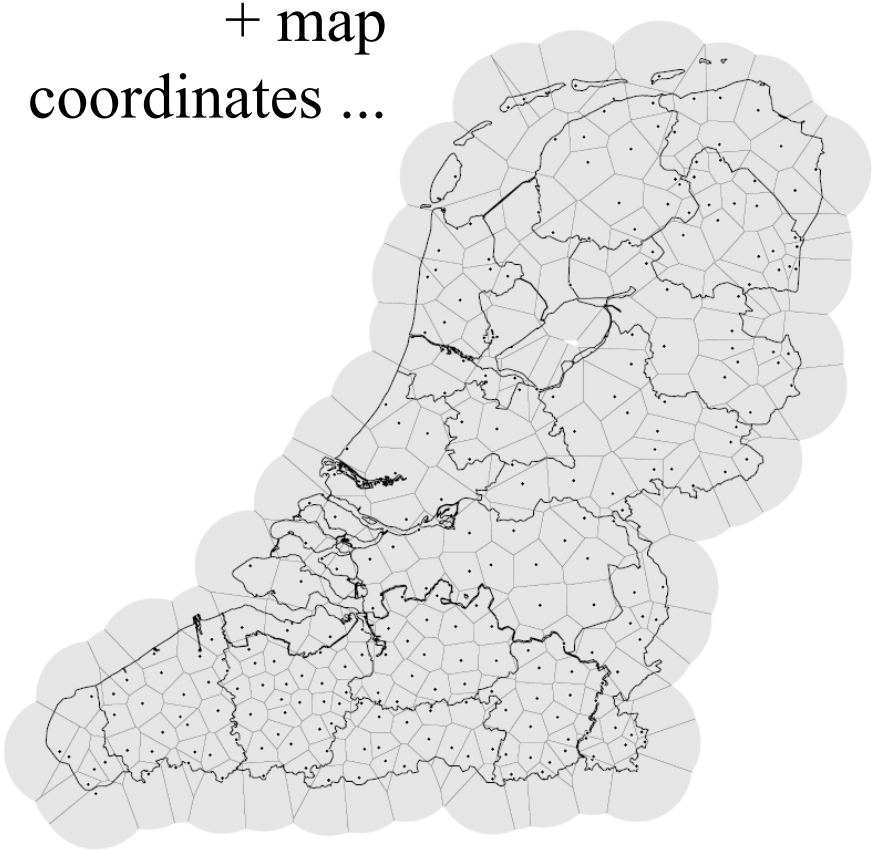
A map =

Voronoi polygons...



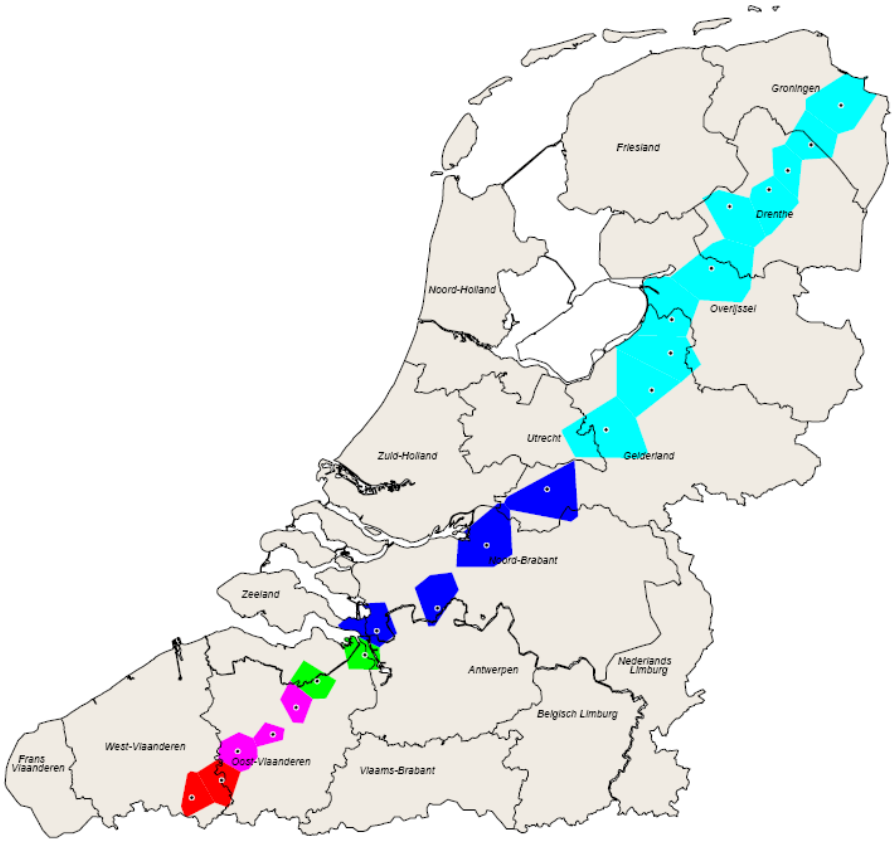
4c) Topological maps II

+ map coordinates ...



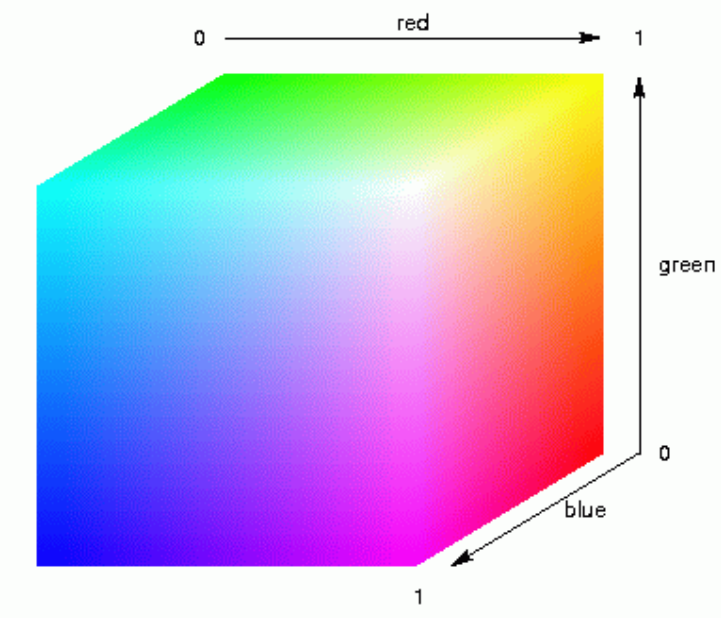
+ clipping coordinates.

4c) Cluster maps



4c) MDS maps

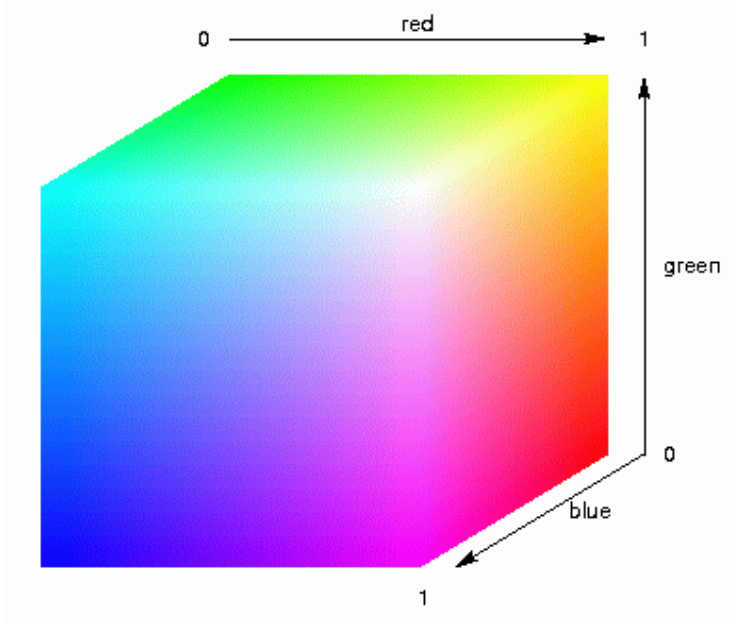
- MDS visualisation trick
  - Places the 267 dialect locations in a three-dimensional space, as faithful as possible to all dialect-pair relationships in the distance matrix
- Visualisation using colour maps
  - 3 dimensions ⇨
  - 3 primary colour components ⇨
  - each dialect has a unique colour
- Colour contrasts represent linguistic differences

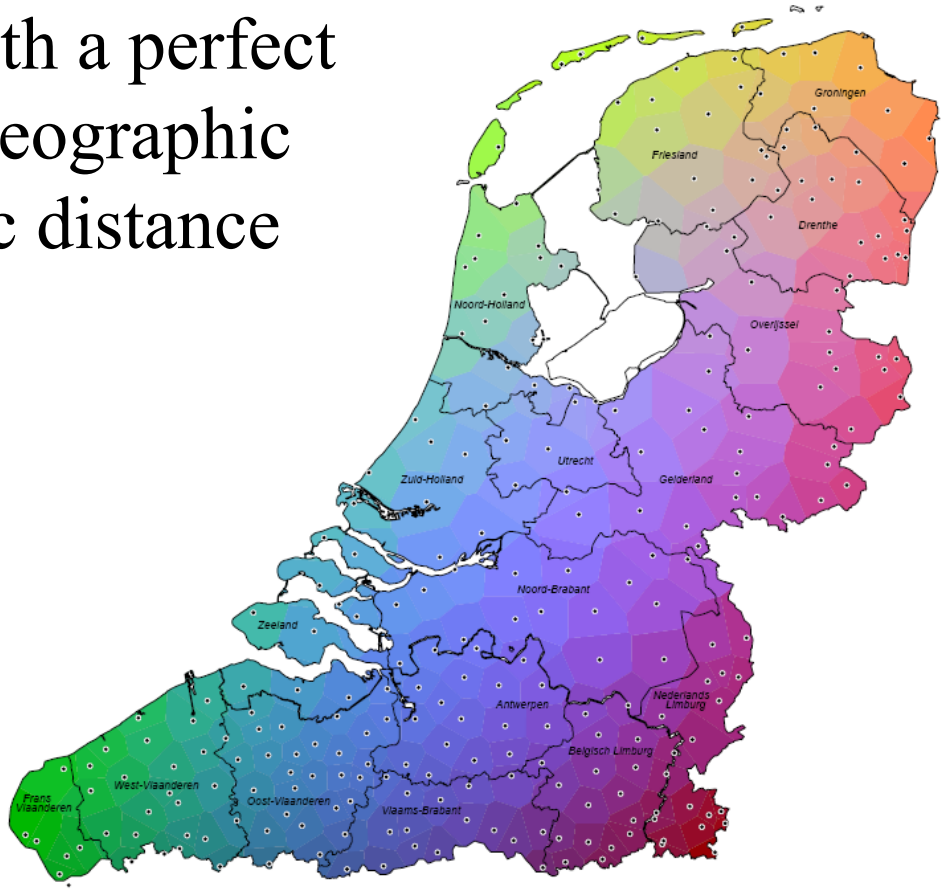


http://www.let.rug.nl/~kleiweg/kaarten/Afstanden.html.en

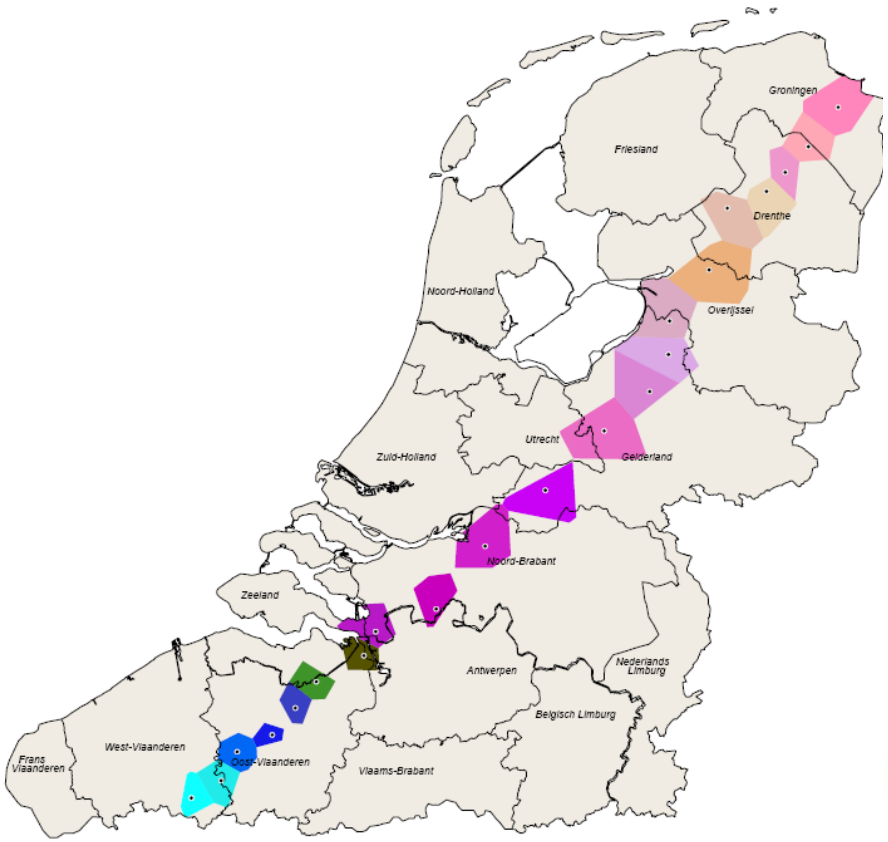
4c) MDS continuum map

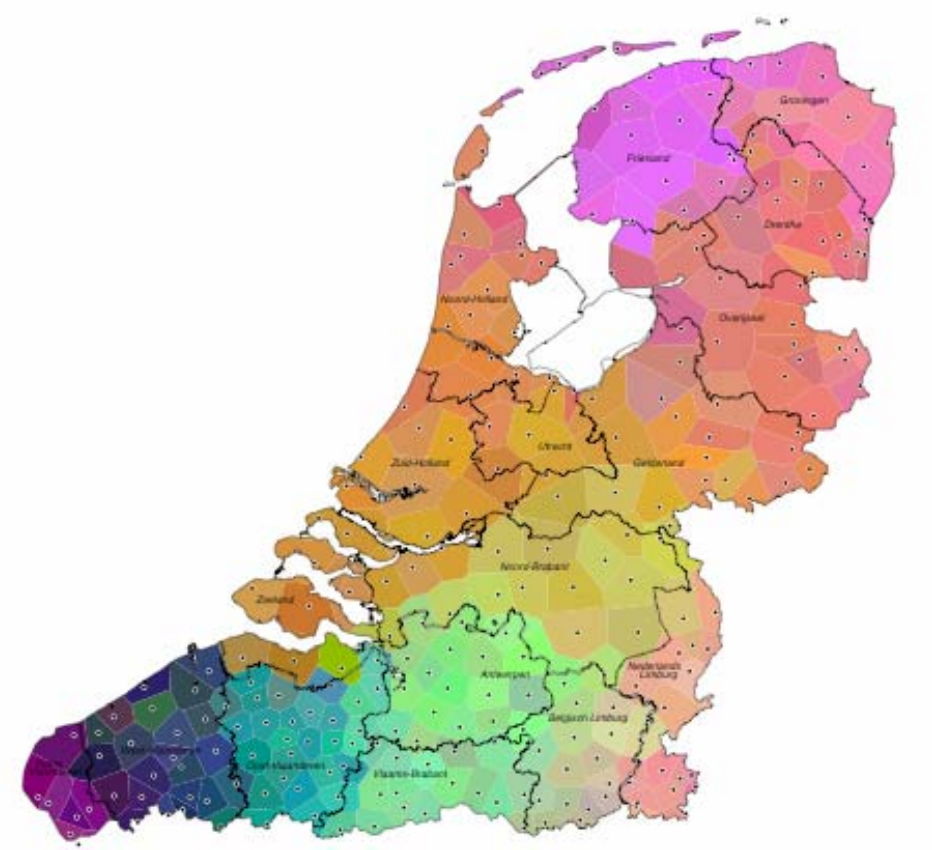
- A topological map with a perfect correlation between geographic distance and linguistic distance



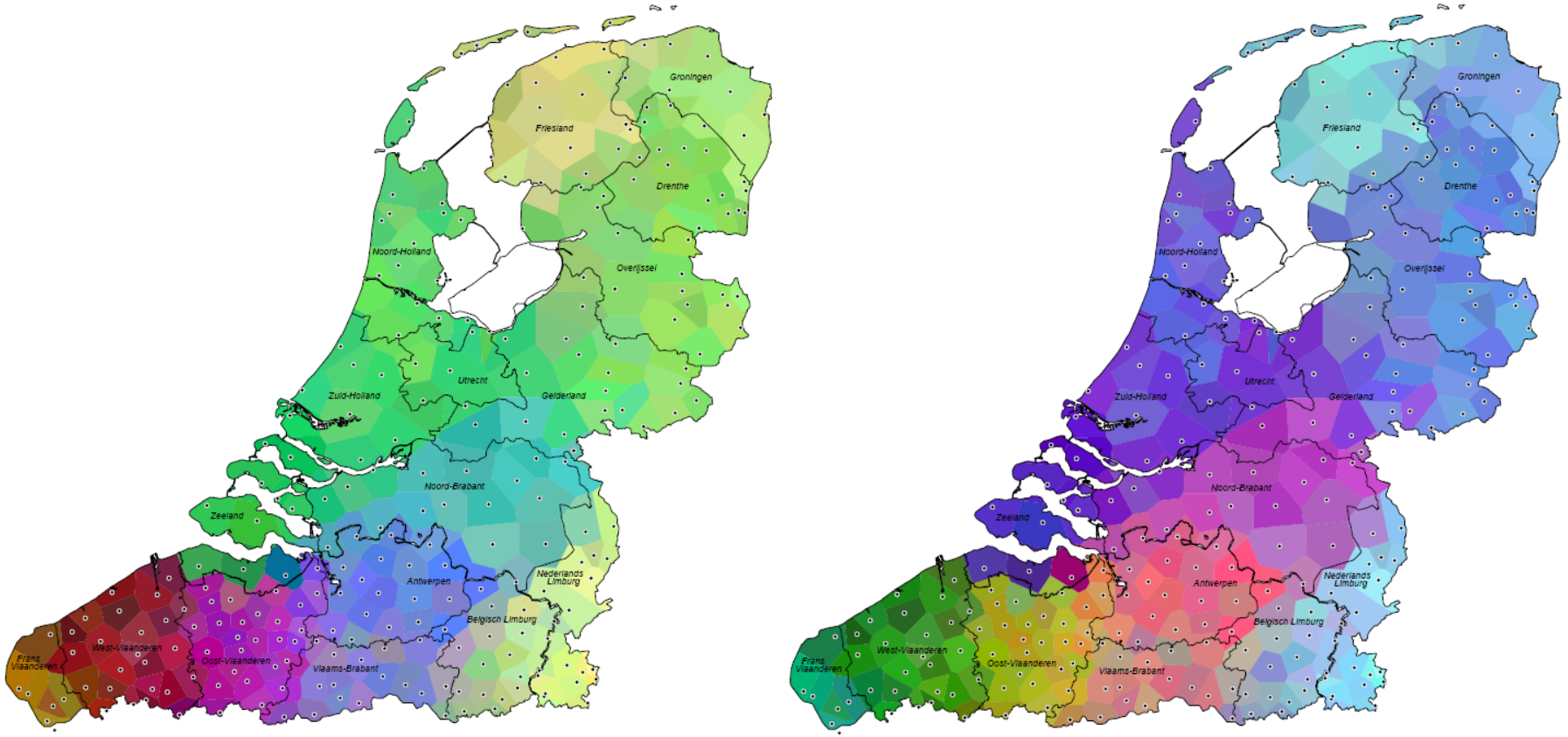


4c) MDS maps


Groningen
Friesland
Drenthe
Noord-Holland
Overijssel
Utrecht
Zuid-Holland
Gelderland
Noord-Brabant
Zeeland
Antwerpen
Nederlands Limburg
Belgisch Limburg
West-Vlaanderen
Frans Vlaanderen
Oost-Vlaanderen
Vlaams-Brabant



4c) MDS maps, rotated dimensions

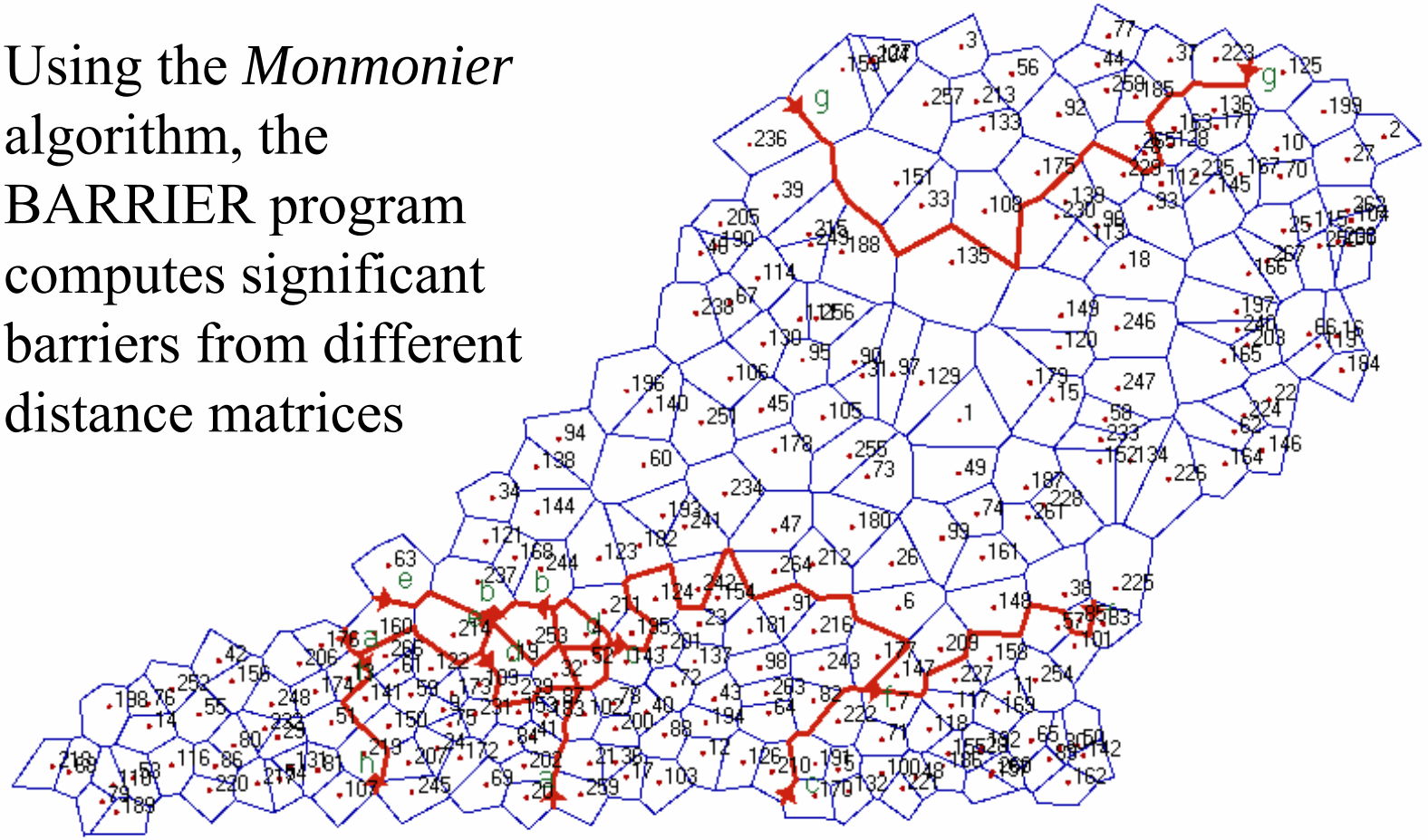


4c) Hybrid methods

- Cluster compositions
  - combinations of cluster analysis and MDS
- Clustering with additional noise
  - to make the cluster analysis more robust
- Visualisation of the distance between neighbouring areas by the thickness or colour of border*lines*

4c) Barrier maps

- Using the *Monmonier* algorithm, the BARRIER program computes significant barriers from different distance matrices



5) Visualisation of dialect relationships

- SAND1
  - Using MDS only
- SAND2
  - Using Cluster analysis and MDS as complementary techniques
- SAND MDS & Cluster map

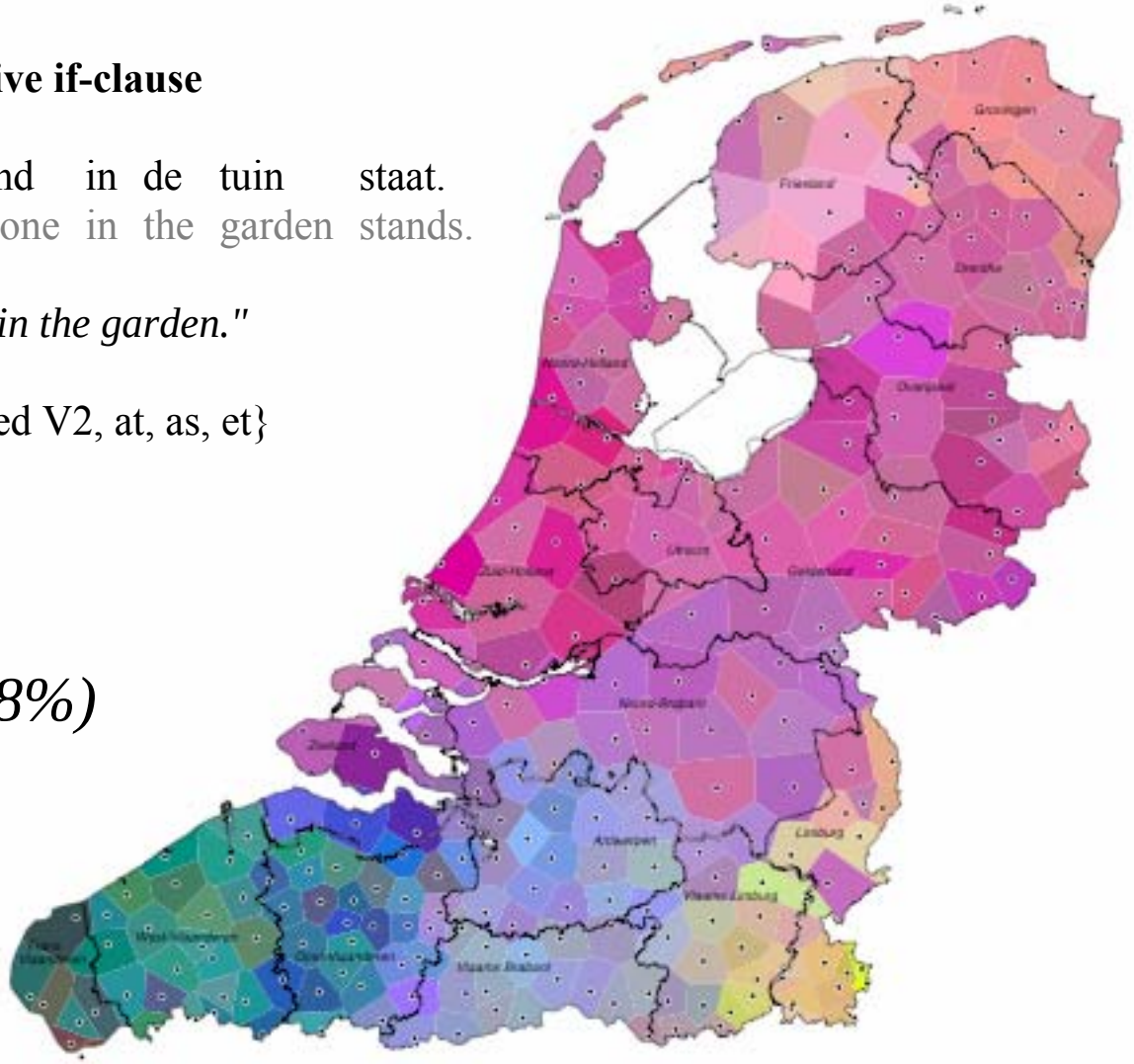
5a) Complementisers

Complementiser of comparative if-clause

't Lijkt wel **of** er iemand in de tuin staat.
It looks AFFIRM if there someone in the garden stands.

*"It looks as **if** there is someone in the garden."*

{of, of dat, dat, as/of + embedded V2, at, as, et}



- 101 variables *(19.8%)*
- $r = 0.94660937$

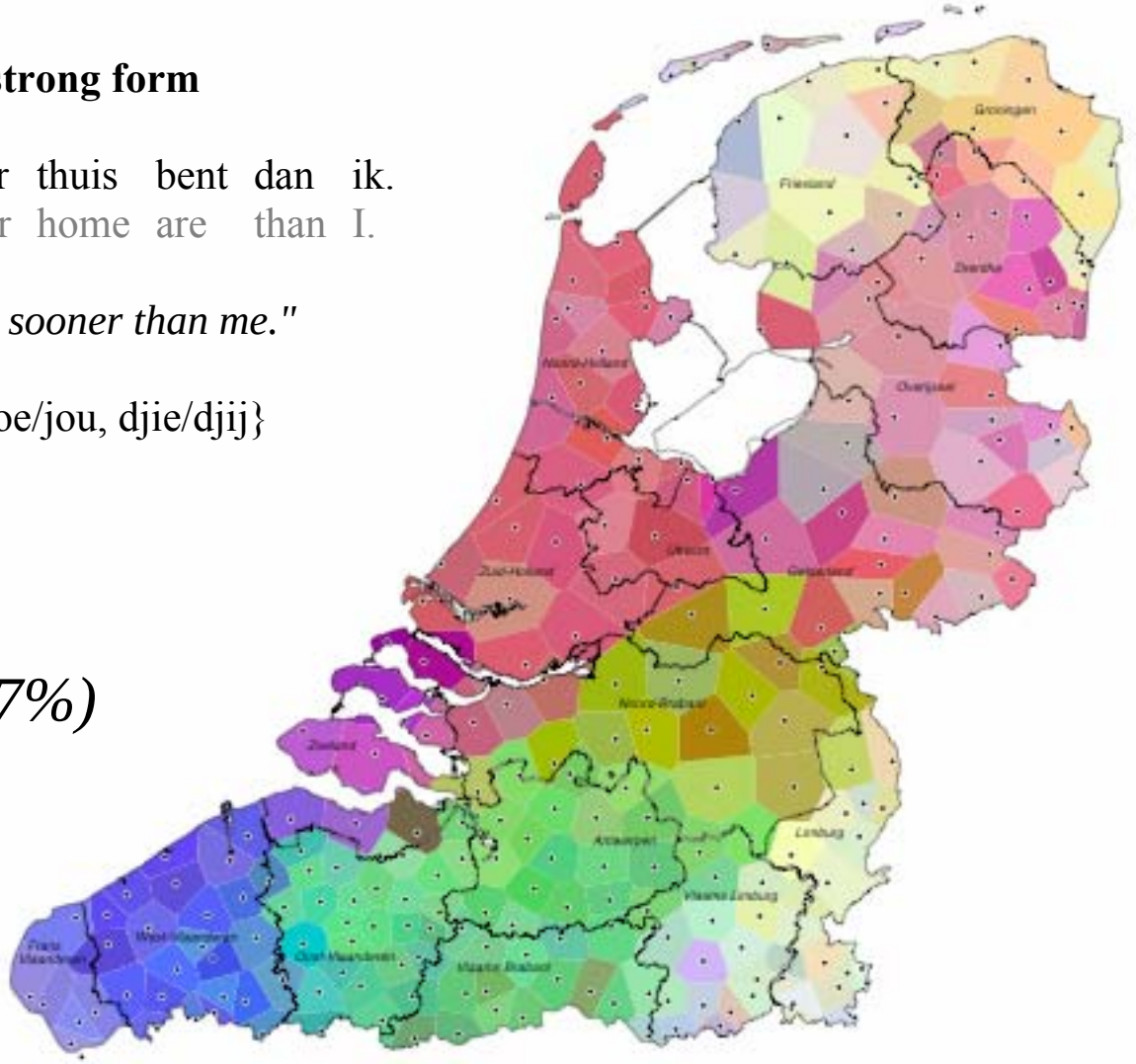
5a) Subject pronouns

Subject pronouns 2 singular strong form

Ze gelooft dat **jij** eerder thuis bent dan ik.
She believes that you earlier home are than I.

*"She thinks that **you**'ll be home sooner than me."*

{gij/gie, jij/jie, du, ie, dich, jo/joe/jou, djie/djij}



- 172 variables *(33.7%)*
- $r = 0.88065714$

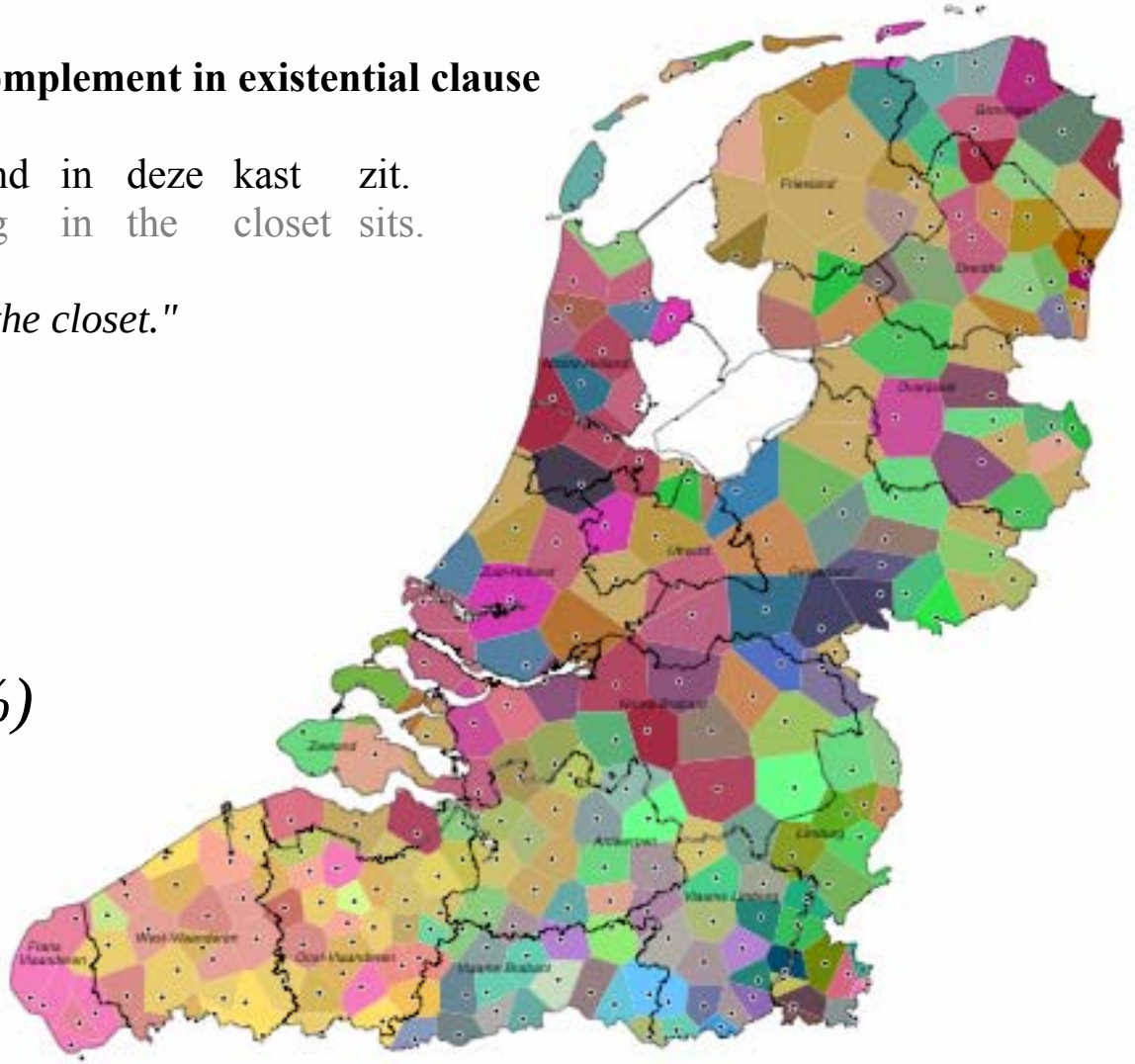
5a) Expletives

Expletive subject following complement in existential clause

't	Is	net	of	**er**	een	hond	in	deze	kast	zit.
It	is	just	if	there	a	dog	in	the	closet	sits.

*"It looks as if **there** is a dog in the closet."*

{er/d'r/t'r, daar, Ø}



- 13 variables *(2.5%)*
- $r = 0.8739387$

5a) Subject doubling

Subject doubling 2 singular

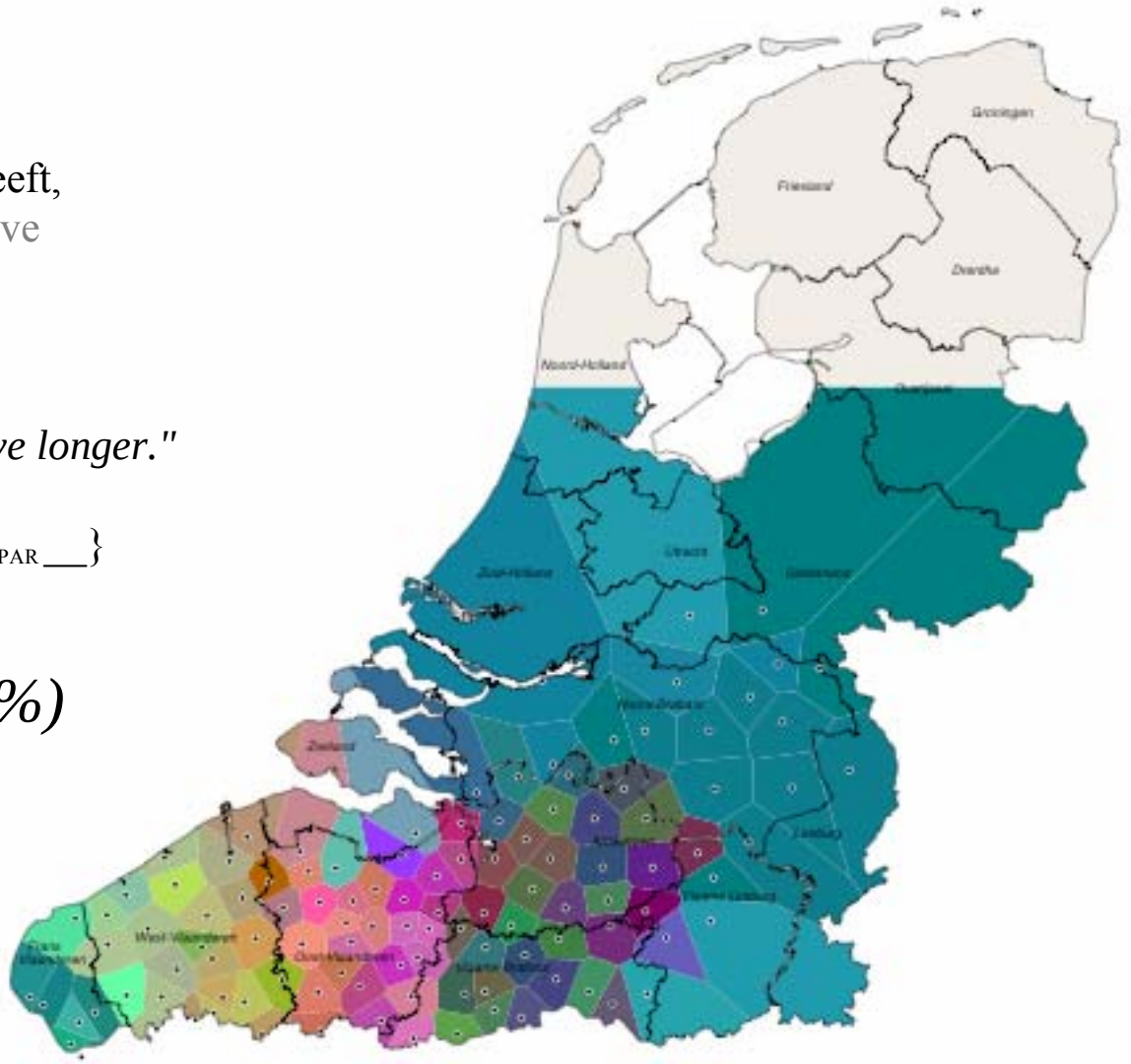
As-**ge gij** gezond leeft,
If-you$_{WEAK}$-you$_{STRONG}$ healthily live

leef-**de gij** langer.
live-you$_{WEAK}$-you$_{STRONG}$ longer.

*"If **you** live healthily **you** will live longer."*

{V_{FIN} __, __ V_{FIN} __, C __, C_{COMPAR} __}

- 54 variables *(10.6%)*
- r = 0.95438211



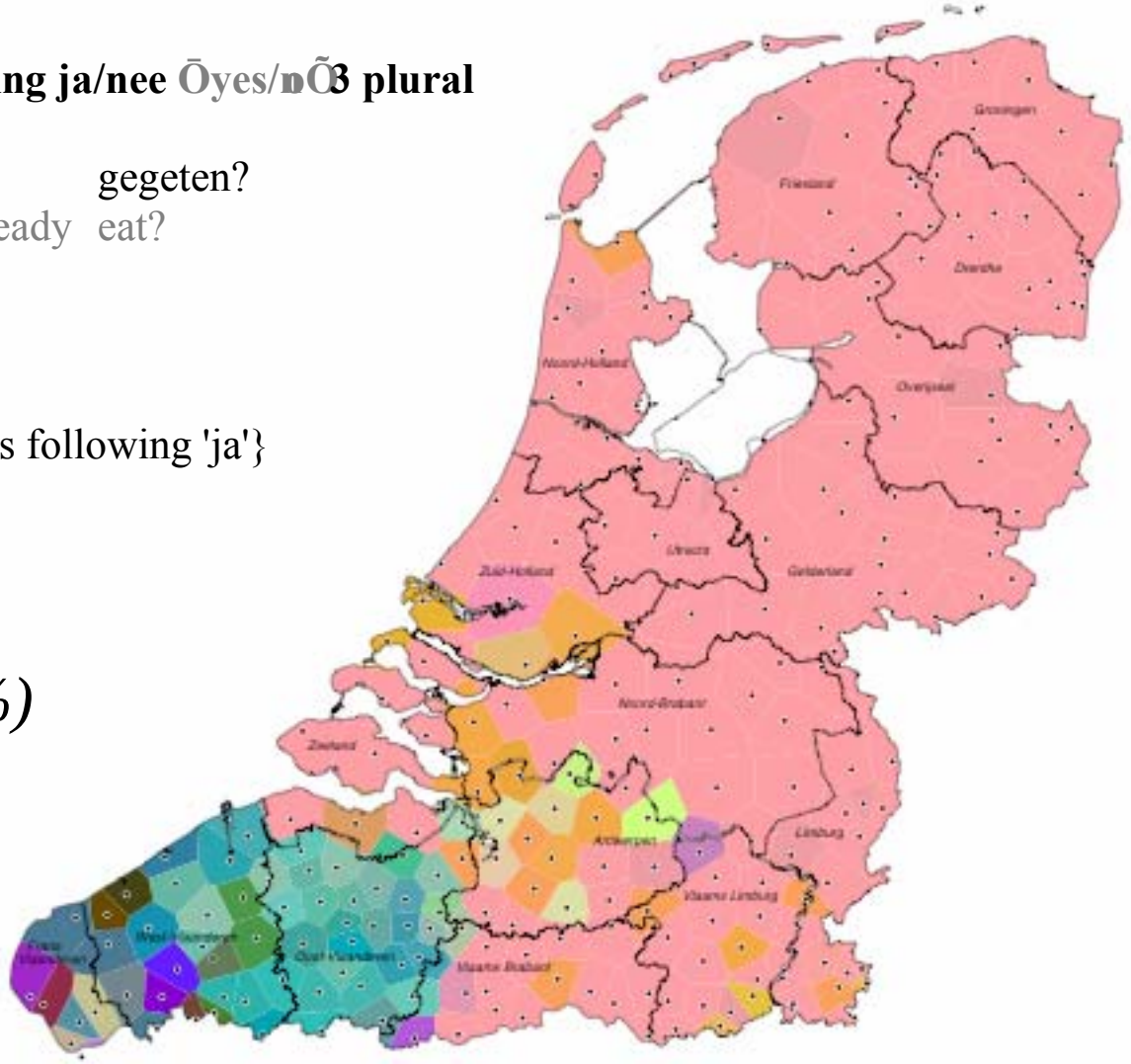
5a) Subject clitisation following yes/no

Clitics and agreement following ja/nee ōyes/nɔ̃3 plural

Vraag:	Hebben	ze	al	gegeten?
Question:	Did	they	already	eat?

Antwoord: ***Jaa-se***
Answer: Yes-3PLUR

{jaa-s(e), jaa-ns, jaa-t, no clitics following 'ja'}



- 30 variables *(5.9%)*
- $r = 0.99025193$

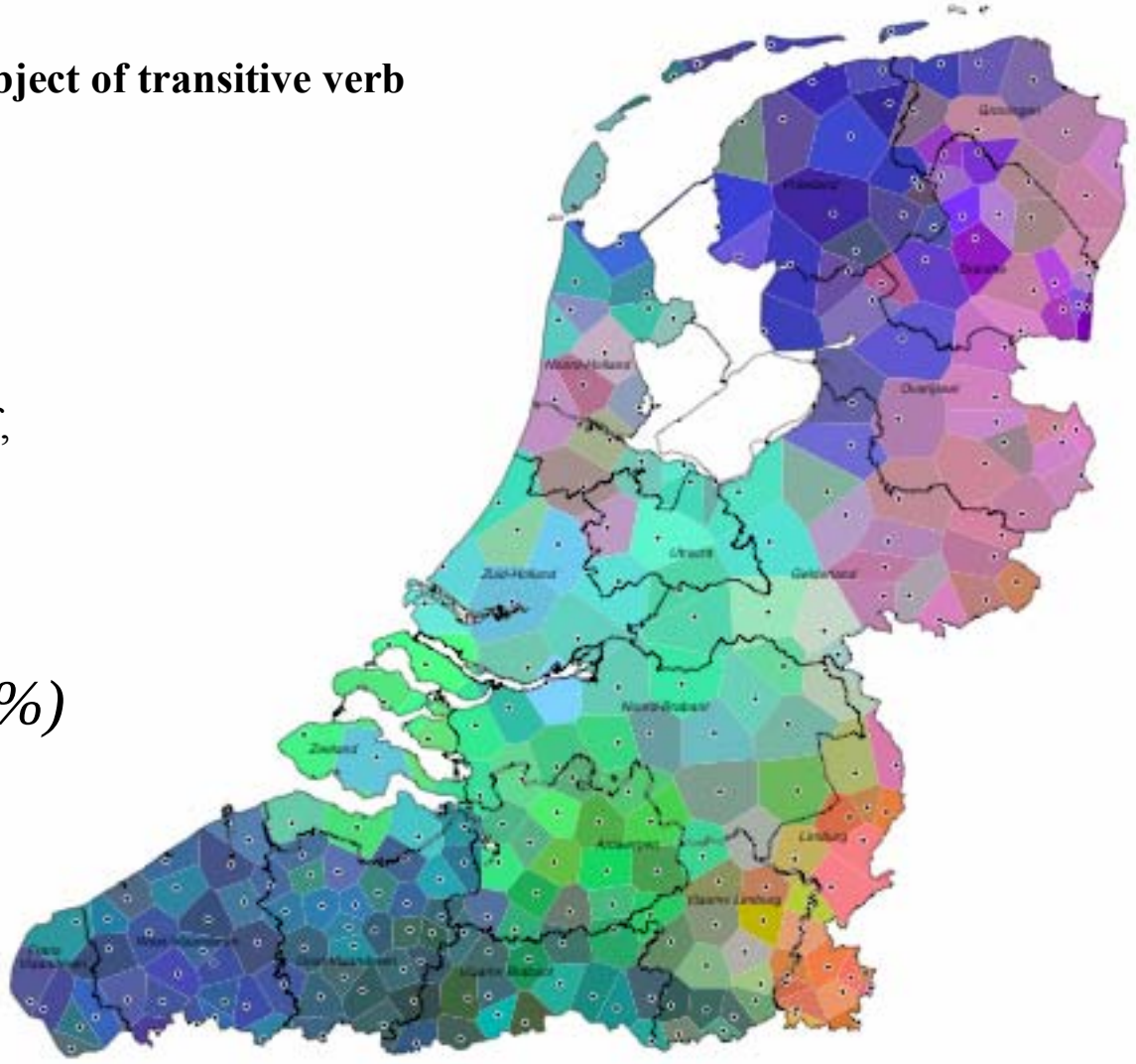
5a) Reflexive & reciprocal pronouns

Weak reflexive pronoun as object of transitive verb

Toon wast **zich**.
Toon washes himself.

*"Toon washes **himself**."*

{hem, zijn eigen, zich, zichzelf,
hemzelf, zijn eigen zelvere}



- 78 variables *(15.3%)*
- $r = 0.93453301$

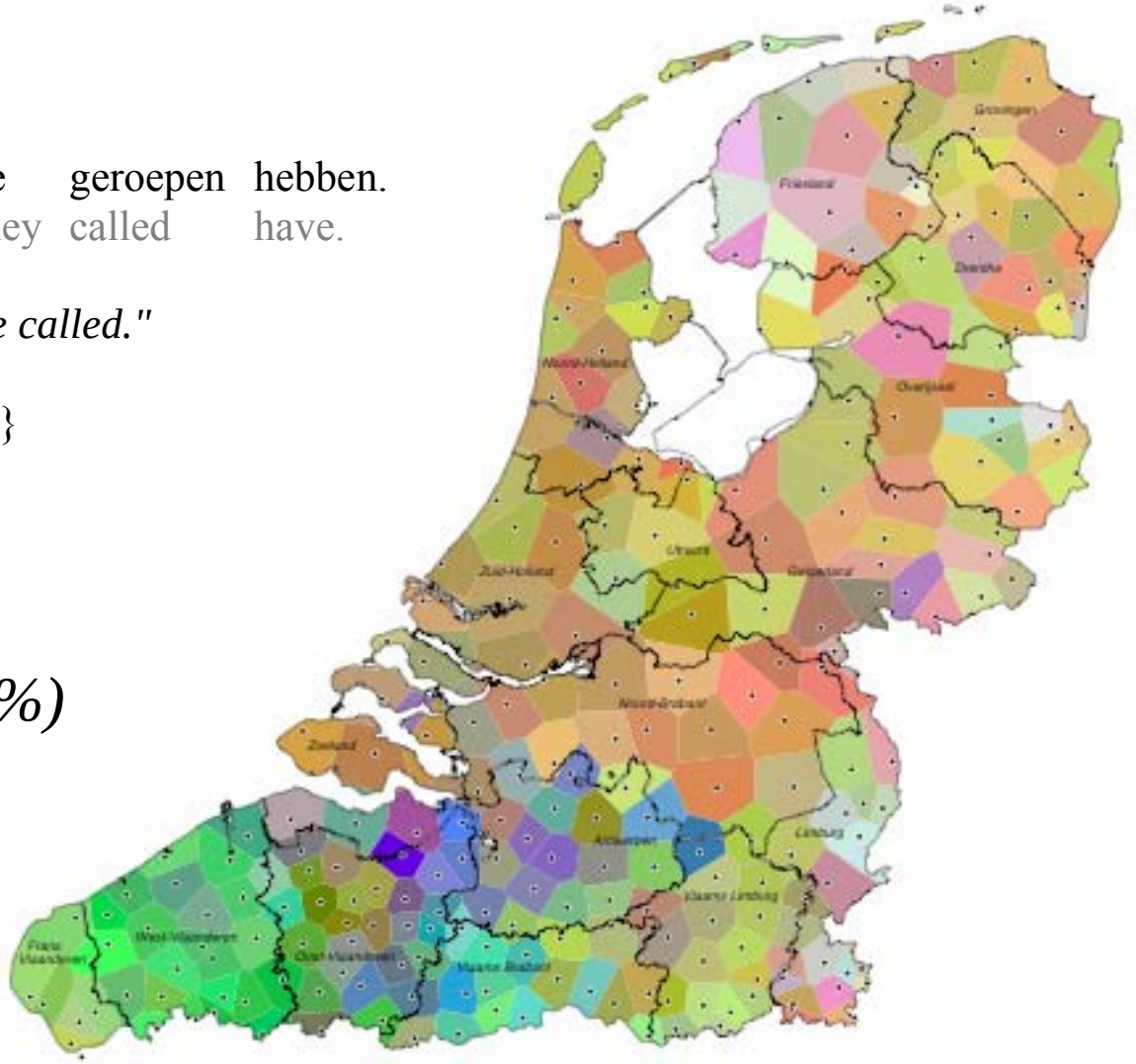
5a) Fronting

Short object relative

Dat	is	de	man	**die**	ze	geroepen	hebben.
That	is	the	man	who	they	called	have.

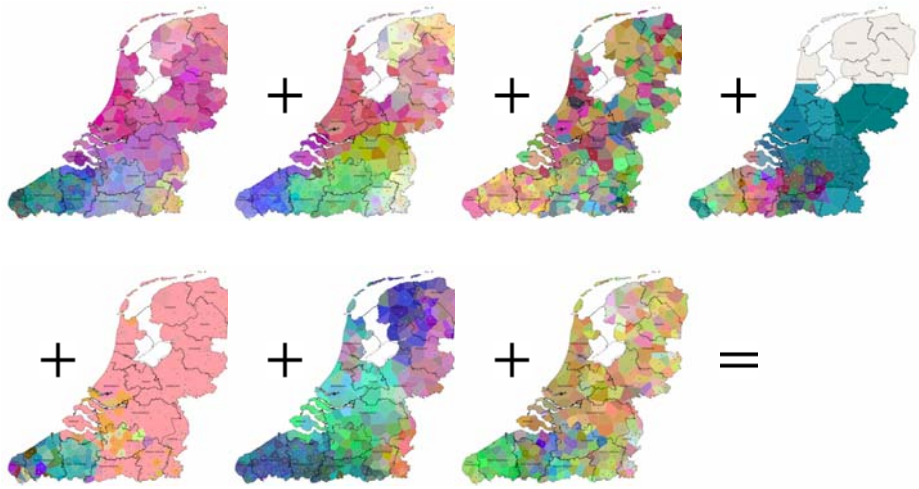
*"That is the man **who** they have called."*

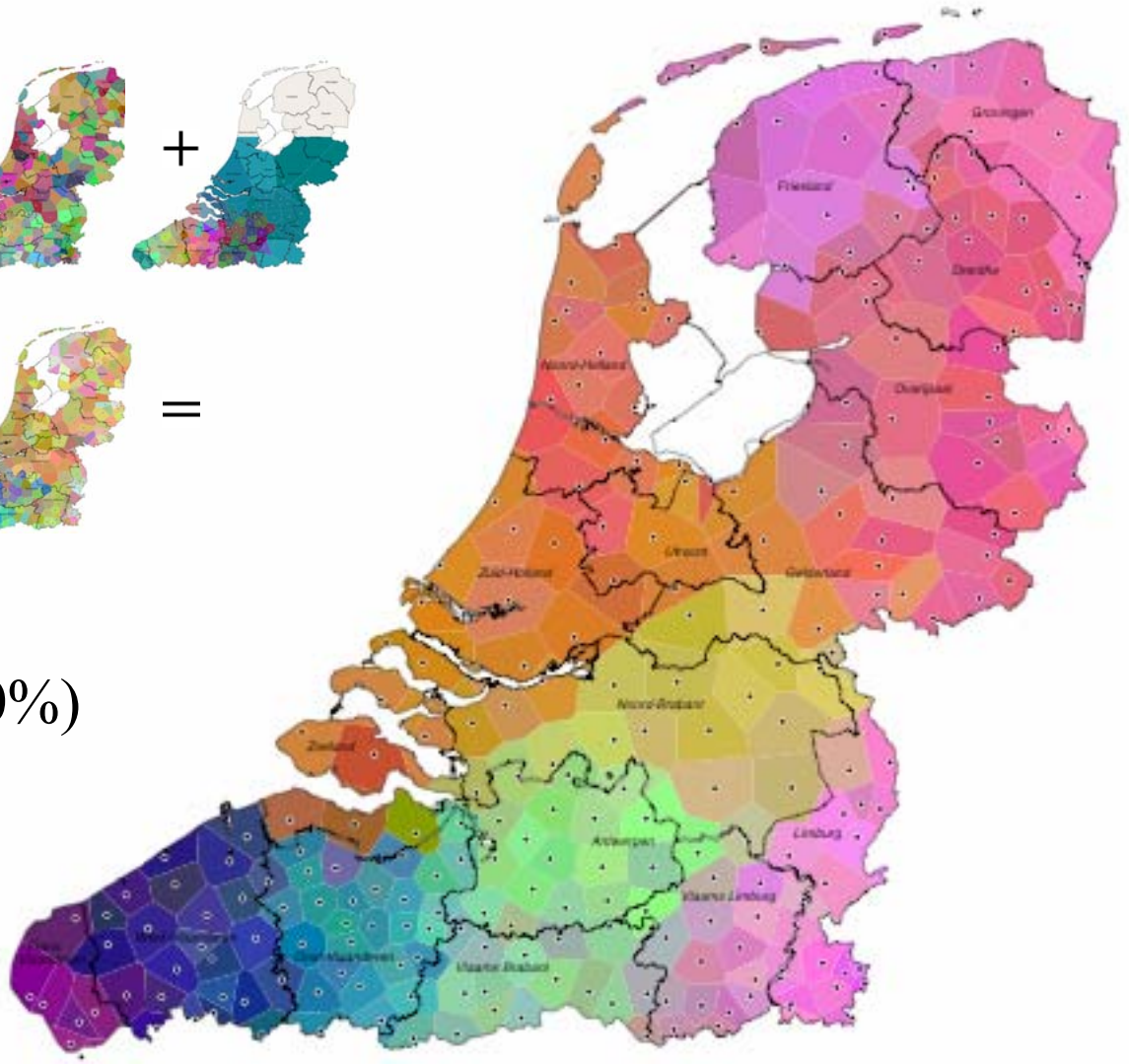
{die, dat, wie, der, den/dem, as}



- 62 variables *(12.2%)*
- $r = 0.77975377$

5a) SAND1 map

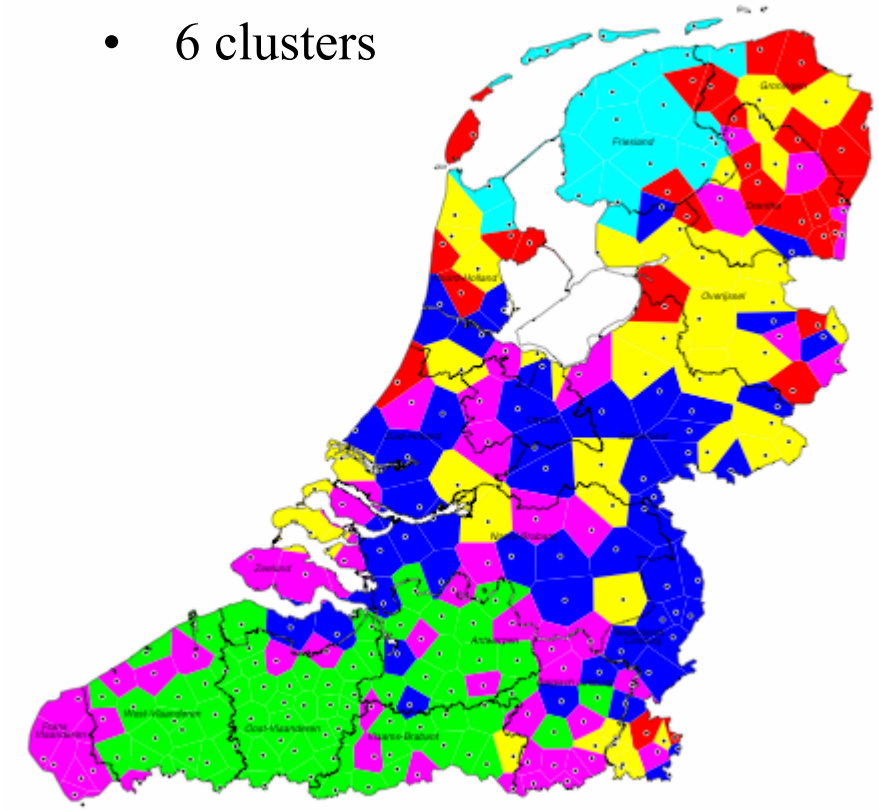




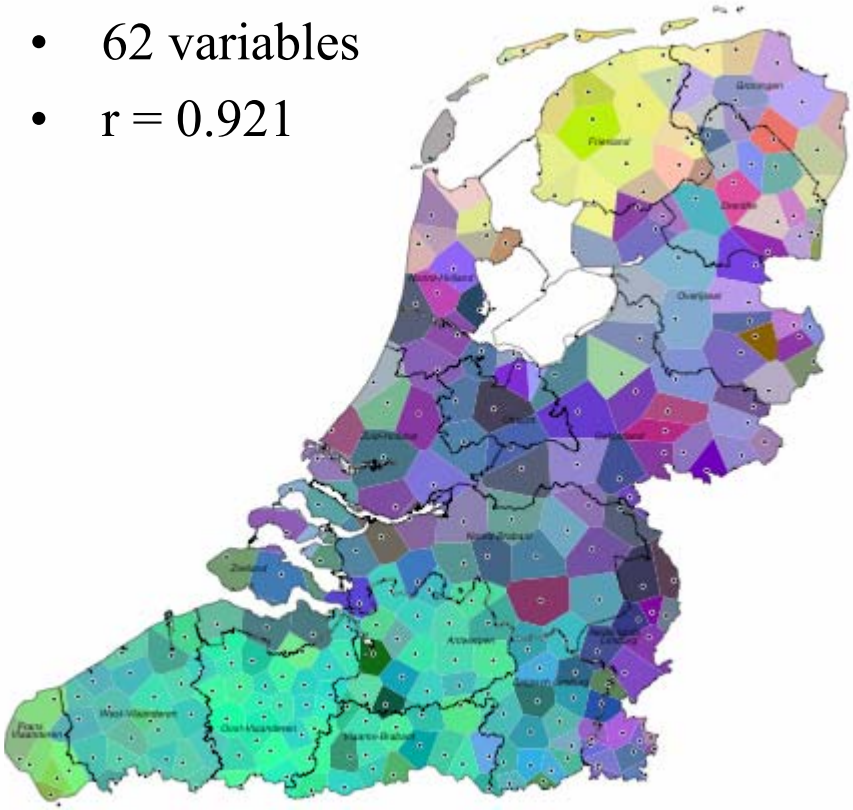
- 507 variables (100%)
- $r = 0.95905712$

5b) Verbal clusters (1)

- Ward's method
- 6 clusters

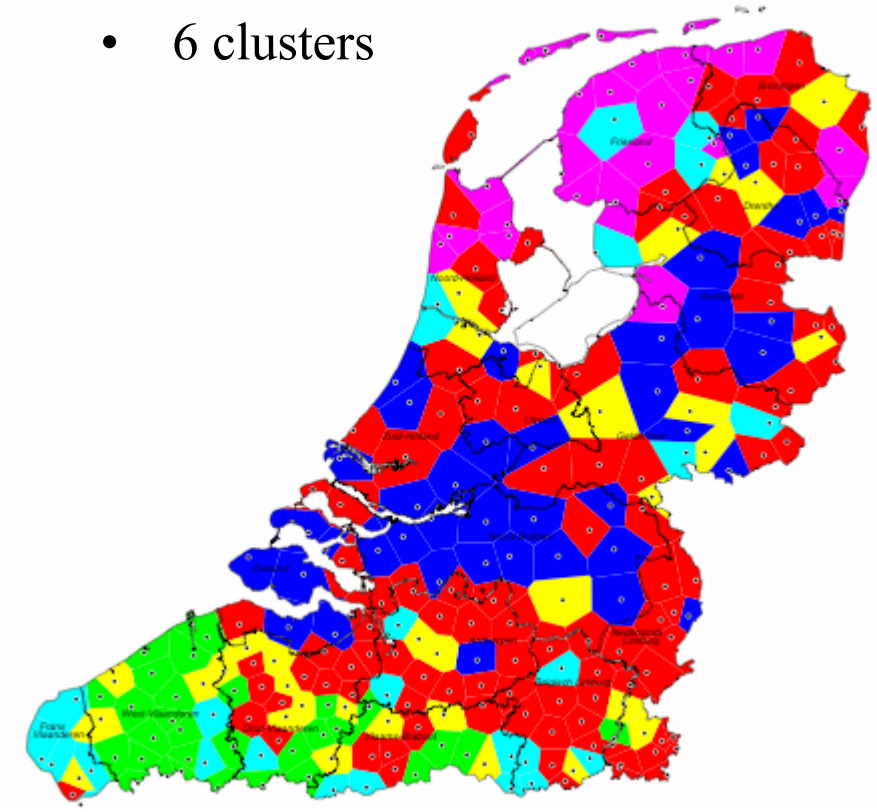


- Classical MDS
- 62 variables
- r = 0.921

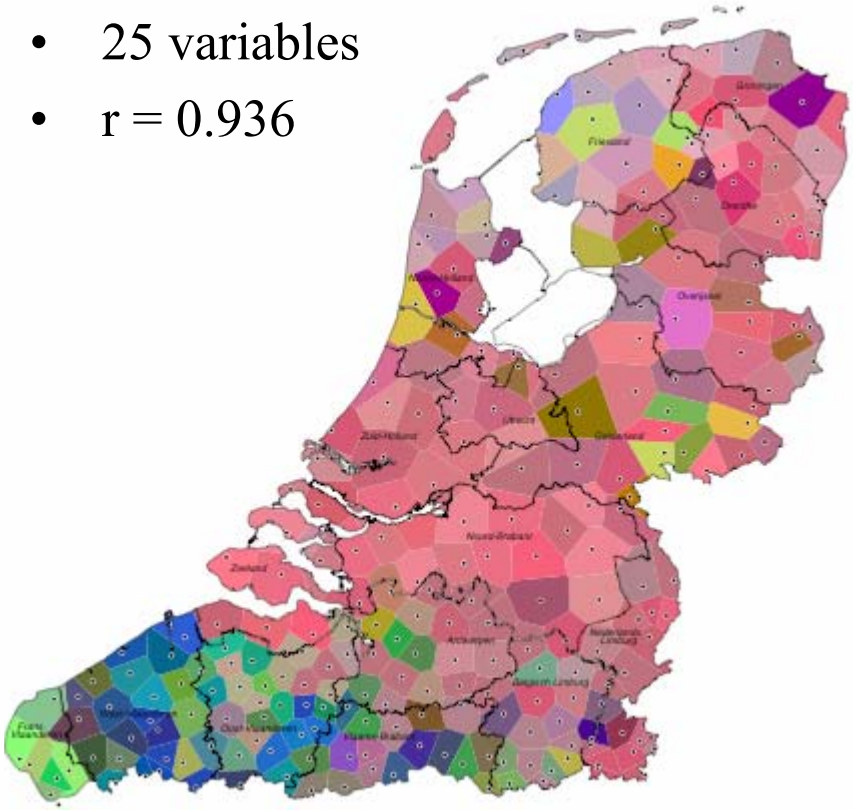


5b) Cluster interruption (2)

- Ward's method
- 6 clusters

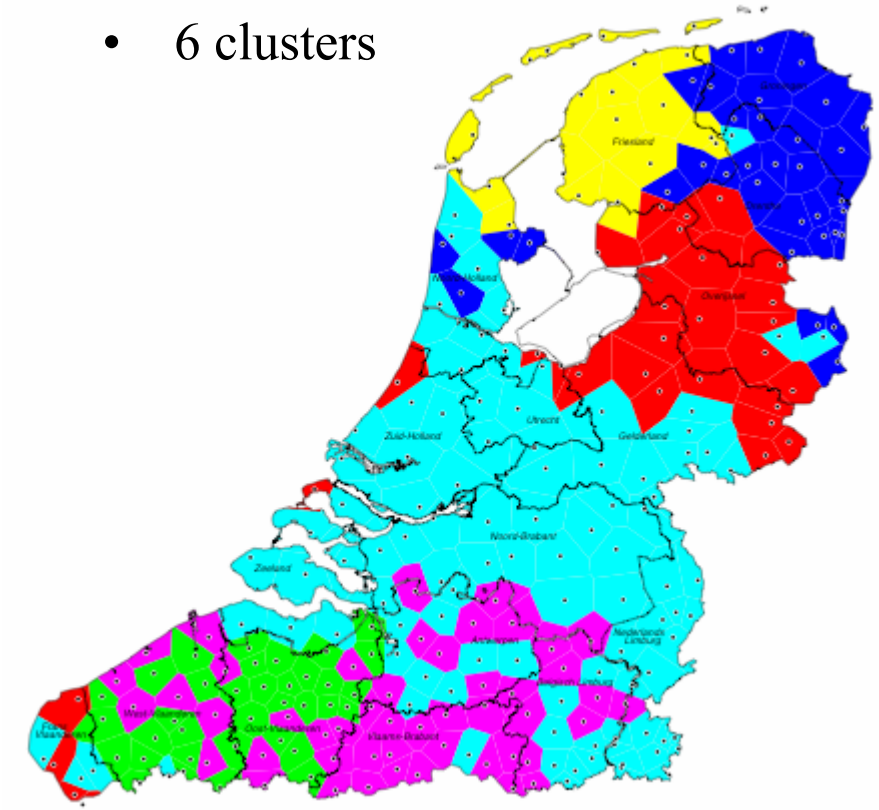


- Classical MDS
- 25 variables
- r = 0.936

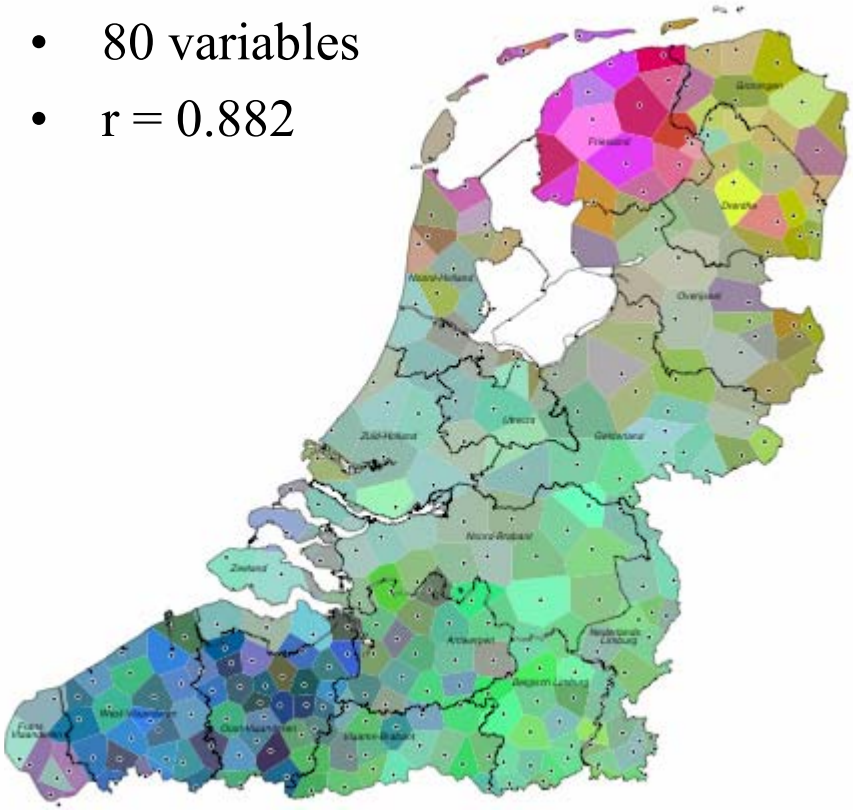


5b) Morphosyntactic variation (3)

- Ward's method
- 6 clusters

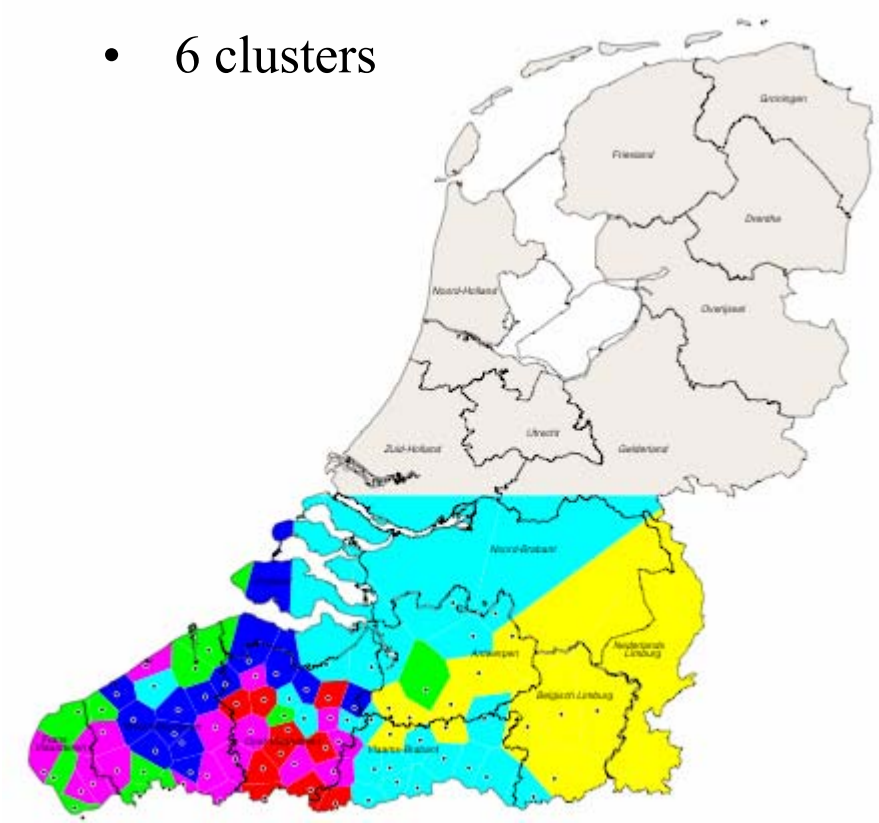


- Classical MDS
- 80 variables
- r = 0.882

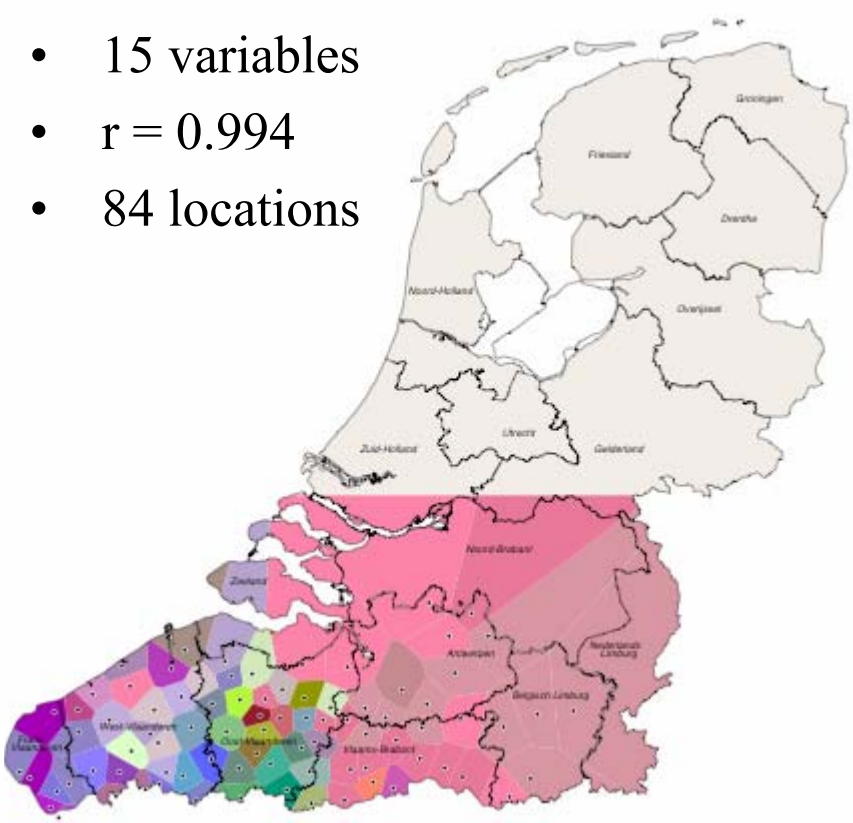


5b) Negative particle (4)

- Ward's method
- 6 clusters

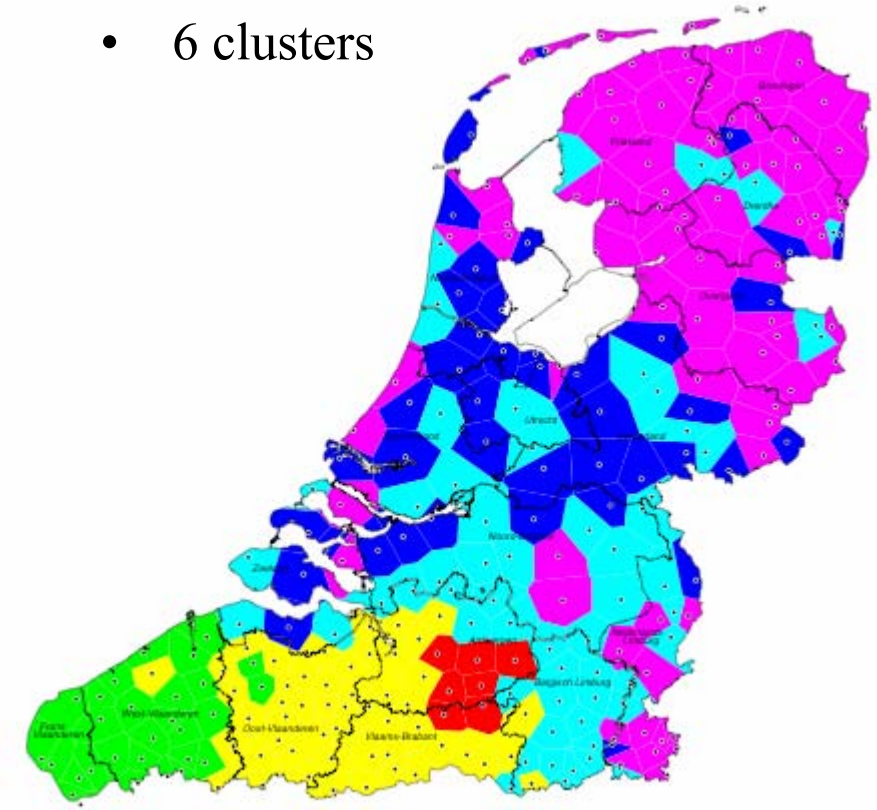


- Classical MDS
- 15 variables
- r = 0.994
- 84 locations

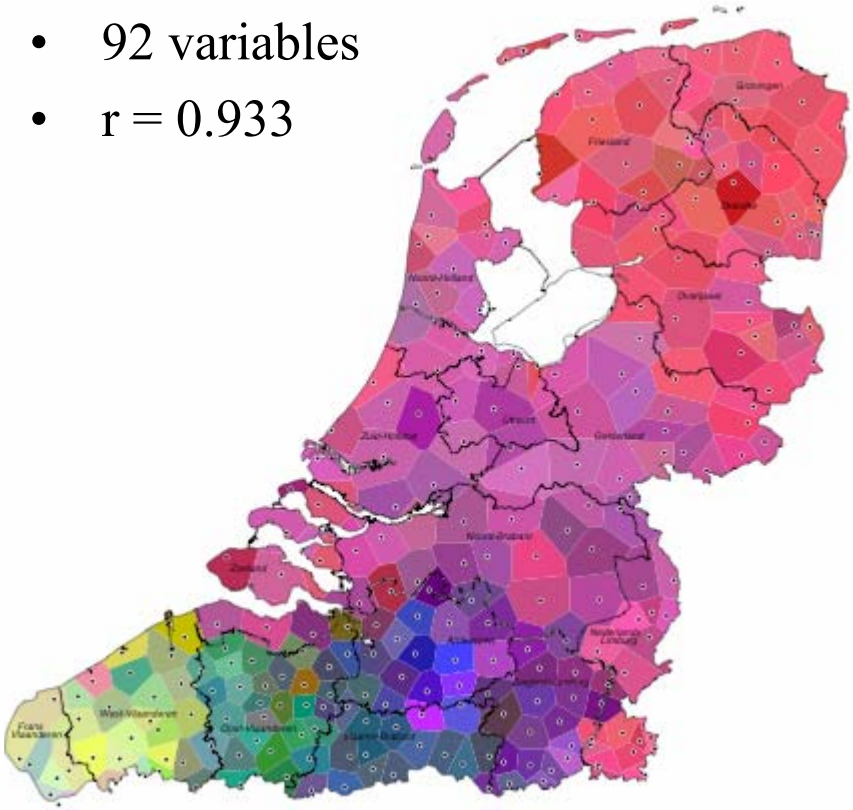


5b) Negative concord & quantifiers (5)

- Ward's method
- 6 clusters

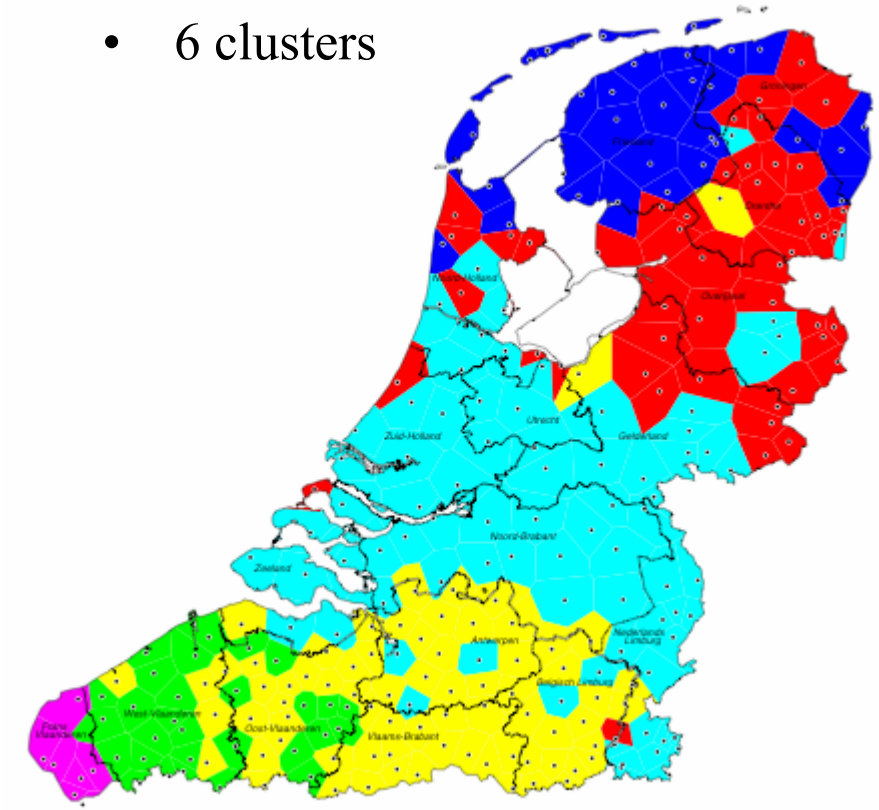


- Classical MDS
- 92 variables
- r = 0.933

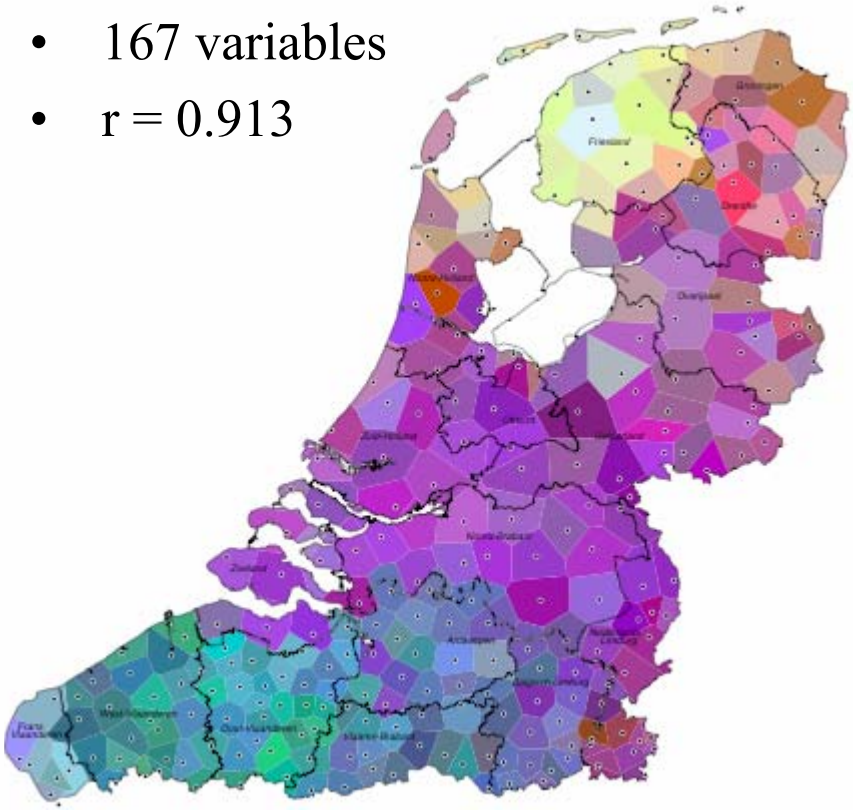


5b) Verbal cluster variation (1,2,3)

- Ward's method
- 6 clusters

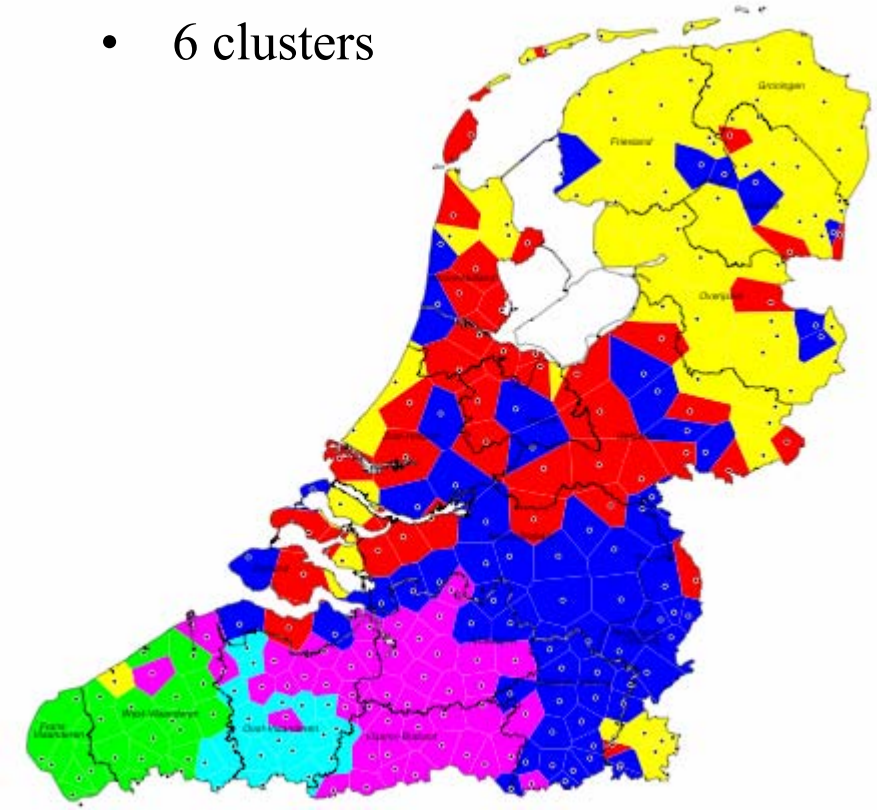


- Classical MDS
- 167 variables
- r = 0.913

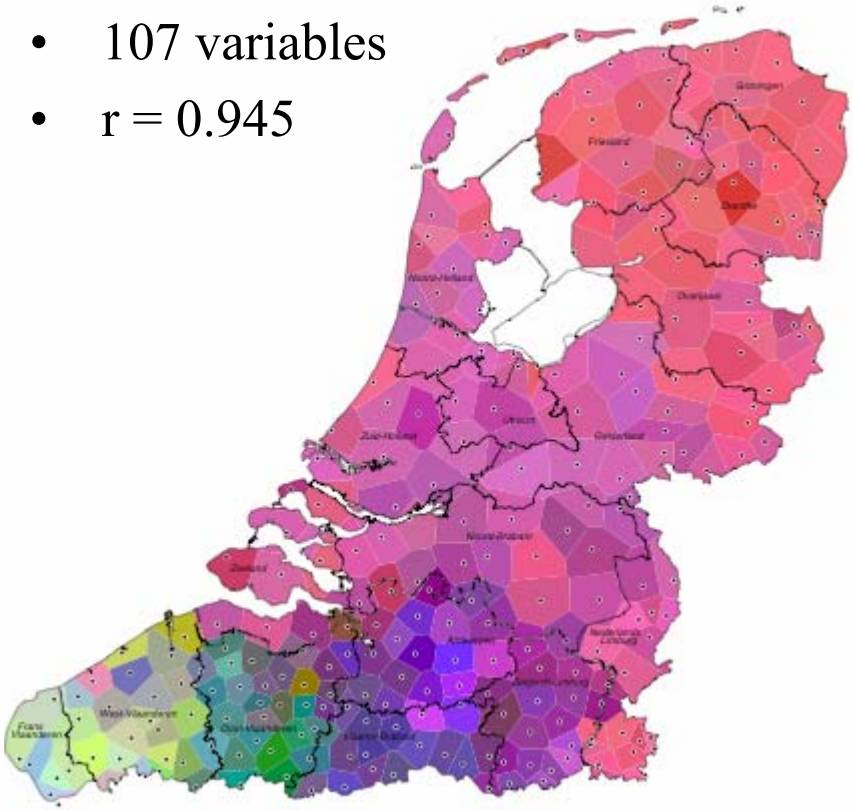


5b) Negative variation (4,5)

- Ward's method
- 6 clusters

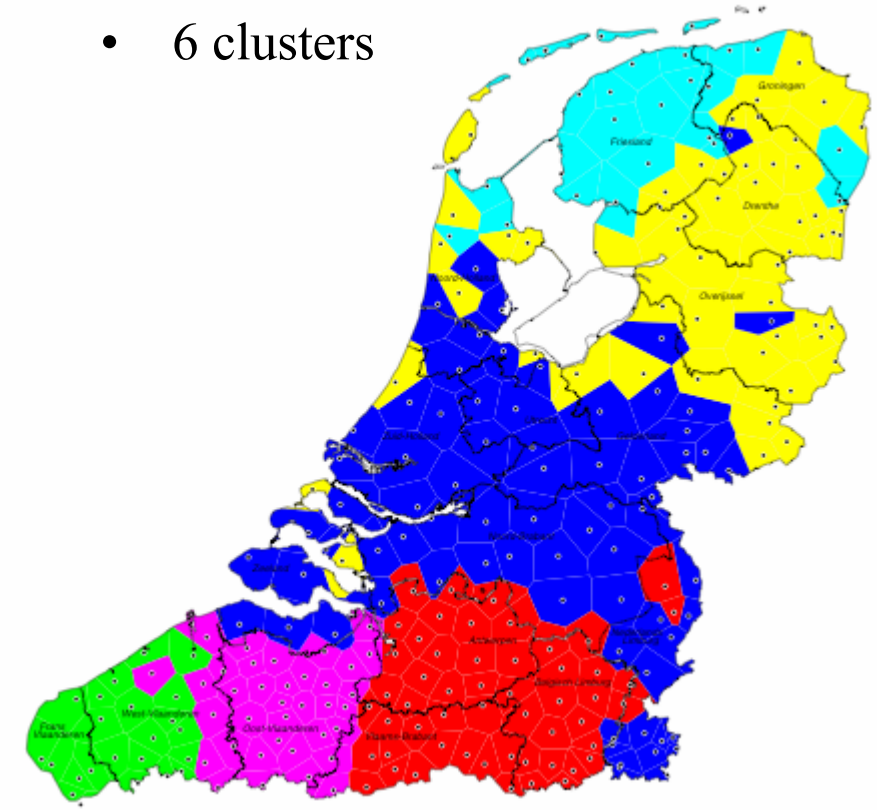


- Classical MDS
- 107 variables
- r = 0.945

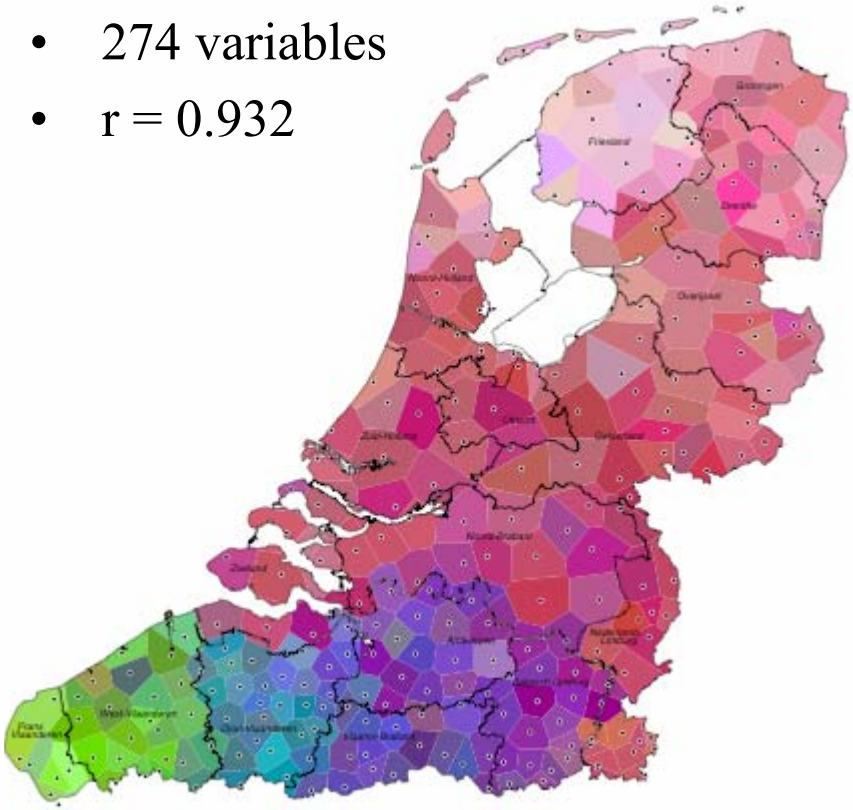


5b) SAND2 (1,2,3,4,5)

- Ward's method
- 6 clusters

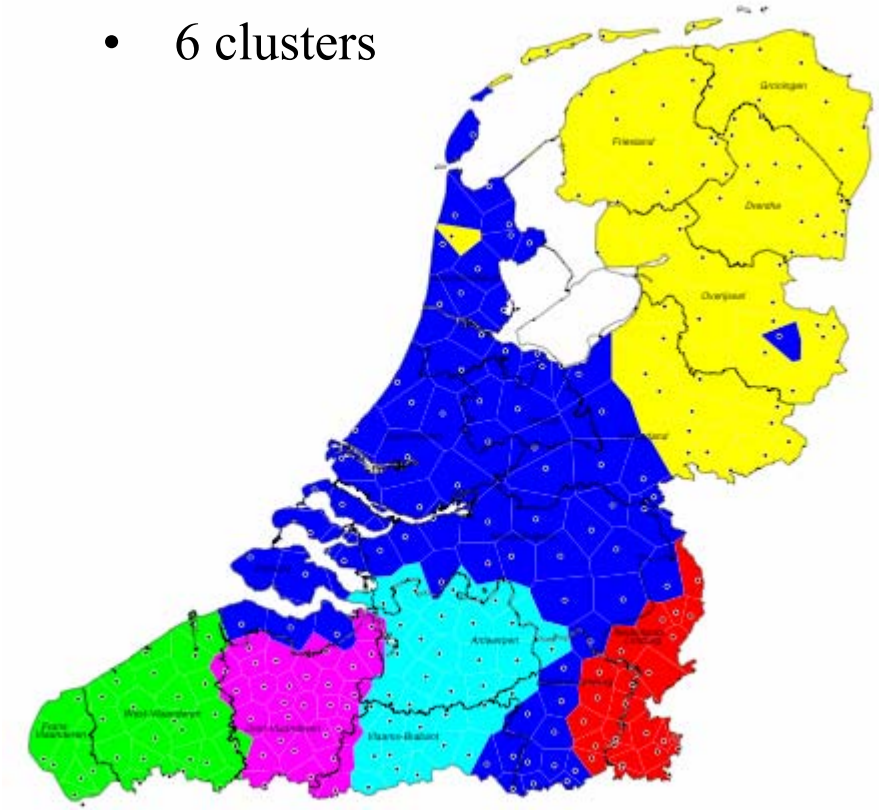


- Classical MDS
- 274 variables
- r = 0.932

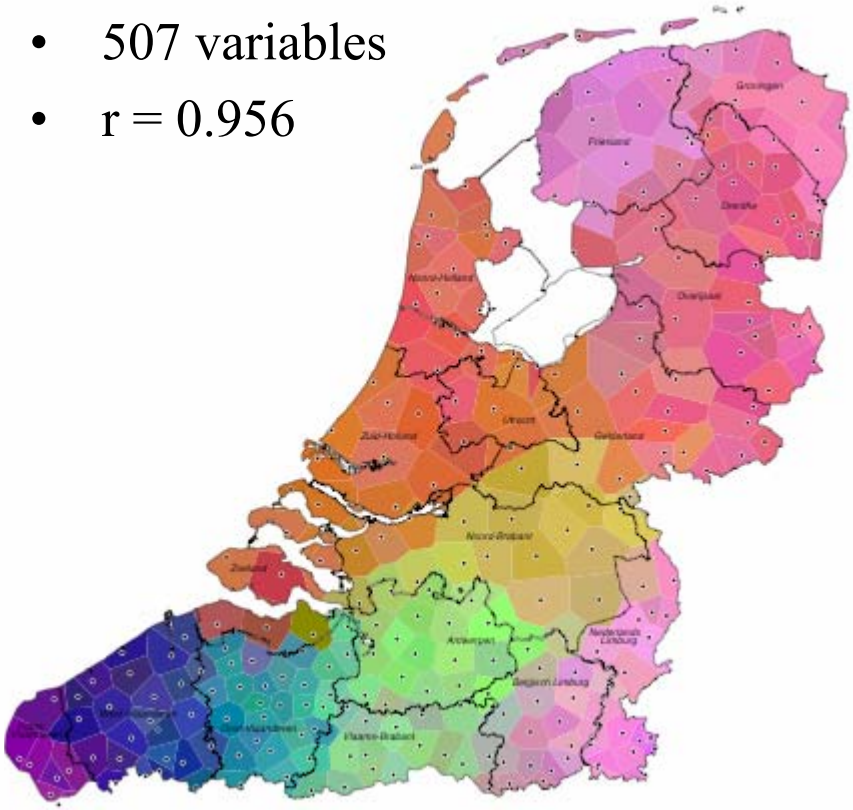


5c) SAND1

- Ward's method
- 6 clusters



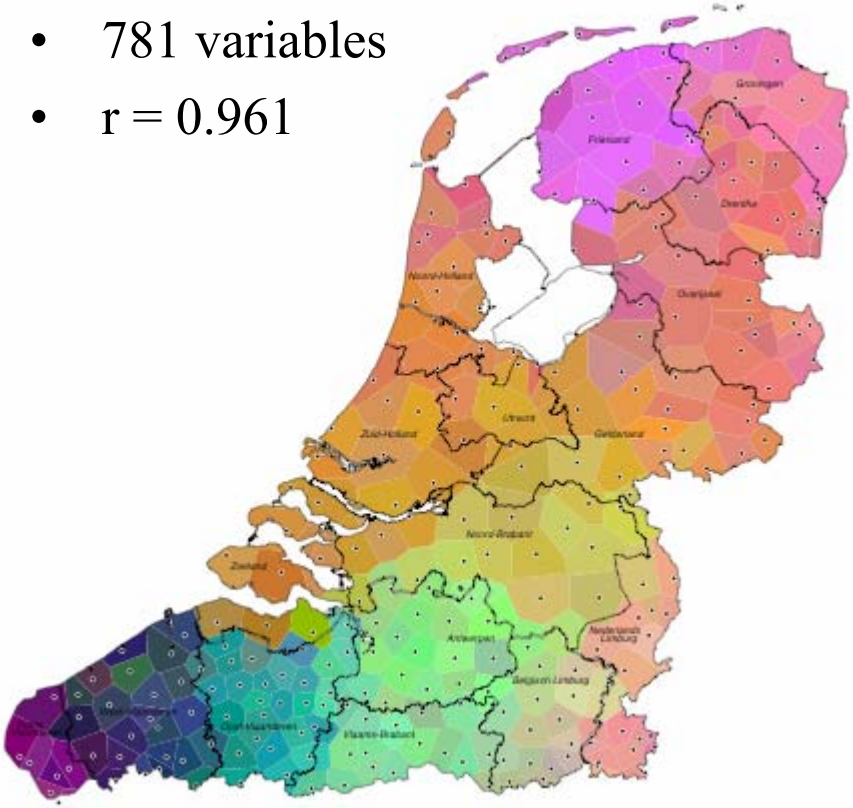
- Classical MDS
- 507 variables
- r = 0.956



5d) SAND

- Ward's method
- 12 clusters

- Classical MDS
- 781 variables
- r = 0.961



6a) Reliability of SAND1 results

Cronbach's alpha

- A coefficient of consistency
- Measures the minimum reliability
  - higher values are better

SAND1 Domain	***Cronbach's α***
Complementisers	0.905065
Subject pronouns and expletives	0.80833
Subject doubling and clitisation	0.901517
Reflexive pronouns	0.870652
Fronting	0.605801
SAND1	**0.952681**

6b) Reliability of SAND2 results

SAND2 Domain	*α*		
Verbal clusters	0.548583		
Cluster interruption	0.604349	0.8811578	
Morphosyntactic variation	0.480014		0.825129
Negative particle	0.67244	0.753002	
Negative concord and quantification	0.685998		

SAND 1+2: 0.954543

7) Measure refinements

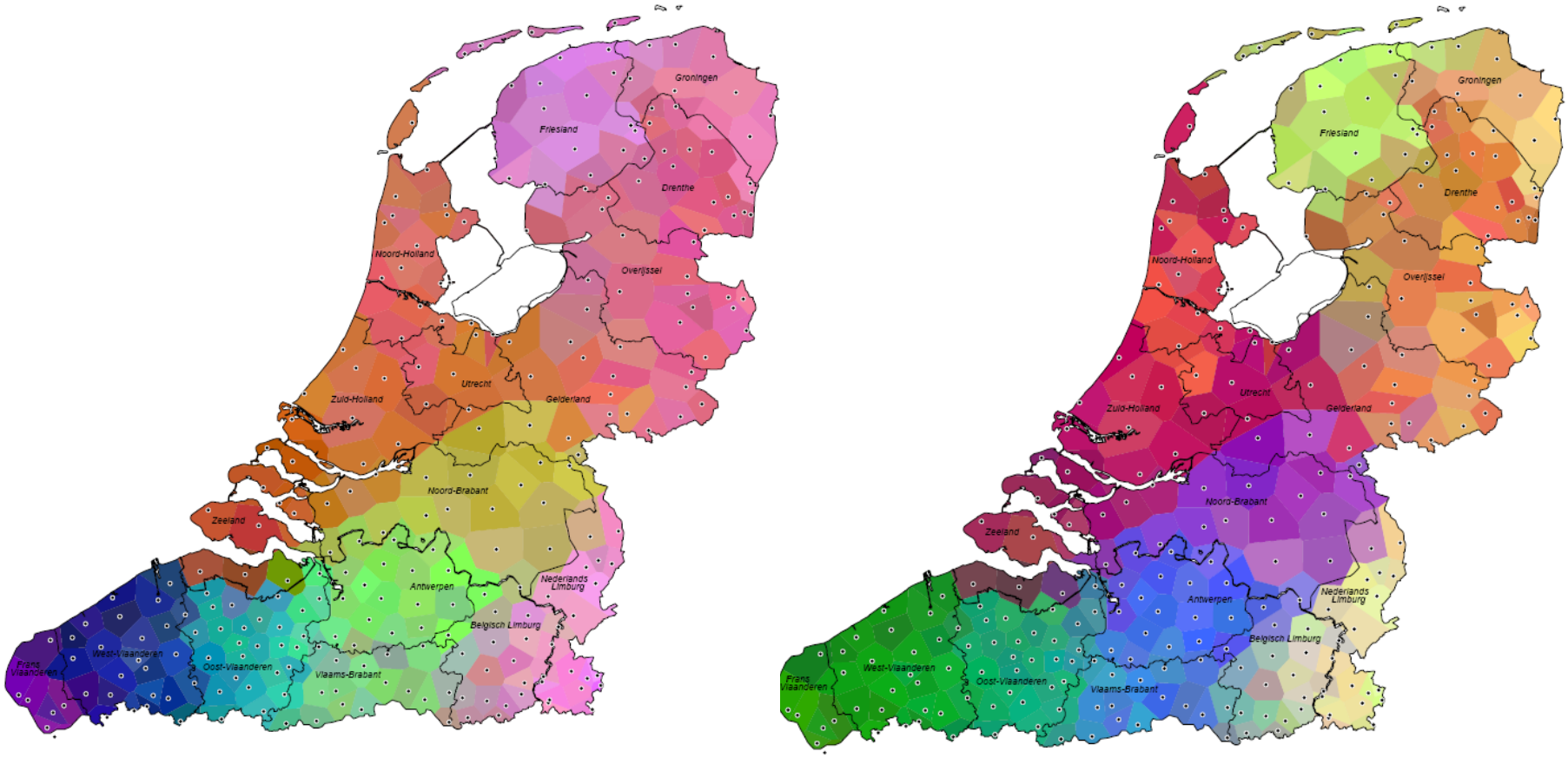
- Statistical additions
  - a) Frequency / GIW
  - 9a) Composite variables
- Linguistic improvements
  - b) Feature variables
  - c) Distance coefficients

7a) Frequency-weighted distances

- Goebl's Gewichteter Identitätswert (G.I.W.)
  - weighted dissimilarity by frequency of variables

variables	*Lunteren*	*Veldhoven*	*distance*
zich	√	√	121/267
hem			0
zijn eigen	√		1
zichzelf			0
hemzelf			0
			1.453

7a) Hamming versus GIW distances



7b) Feature variables

- Mapping from atomic variables (first column) to feature variables (first row) with respect to reflexive pronouns:

	personal *"hem"*	reflexive *"zich"*	possessive *"zijn"*	ownness" *eigen"*	focus *"zelf"*
hem	√				
hemzelf	√				√
zich		√			
zichzelf		√			√
zijn			√		
zijn zelf			√		√
zijn eigen			√	√	
zijn eigen zelf			√	√	√

7b) Measuring feature variables

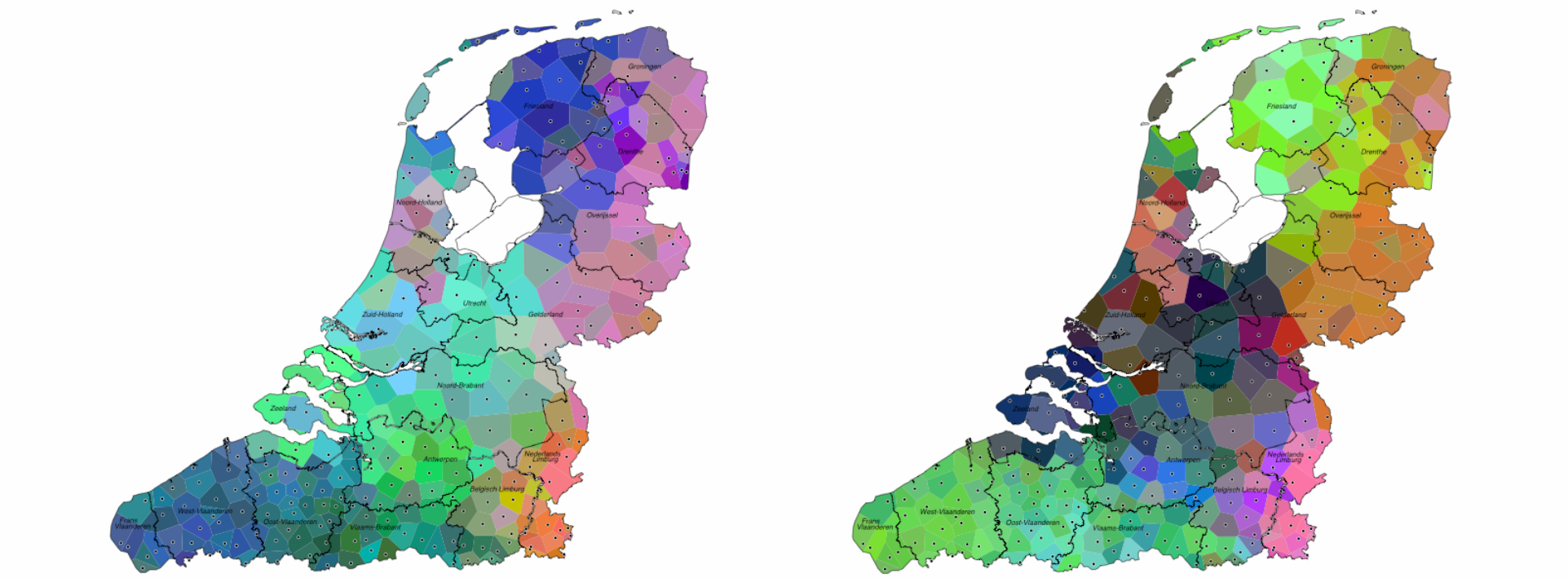
- Fragment of the distance measurement between two dialects using five feature variables (first column)

	Lunteren	Veldhoven	distance
	{zich, zijn eigen}	{zich}	
personal			0
reflexive	√	√	0
possessive	√		1
ownness	√		1
focus			0
			2

7b) Feature variable advantages

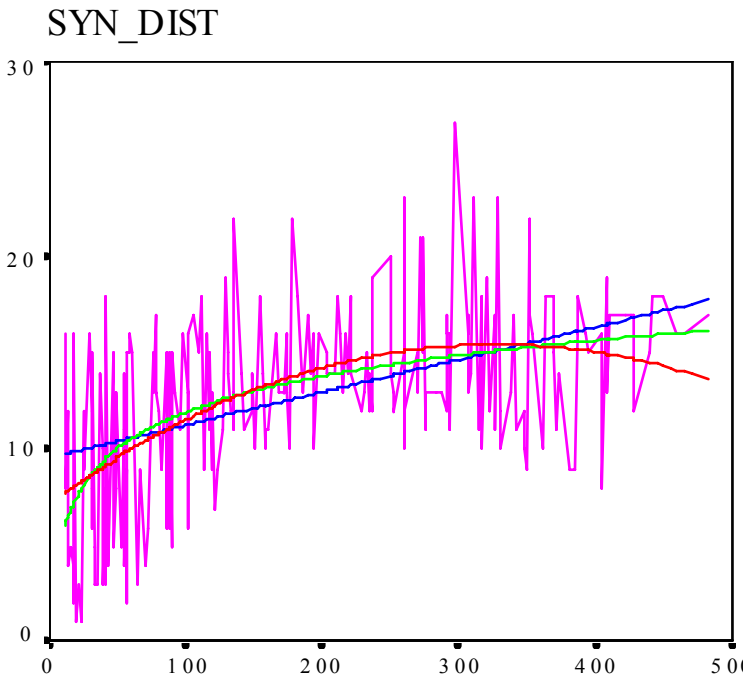
- No one-to-one mapping from atomic variables to feature variables
- Abstraction layer allows for more differentiation between dissimilar variable pairs
- Enables measurements based on distance coefficients (7c)
- *Downside*: requires analysis and annotation

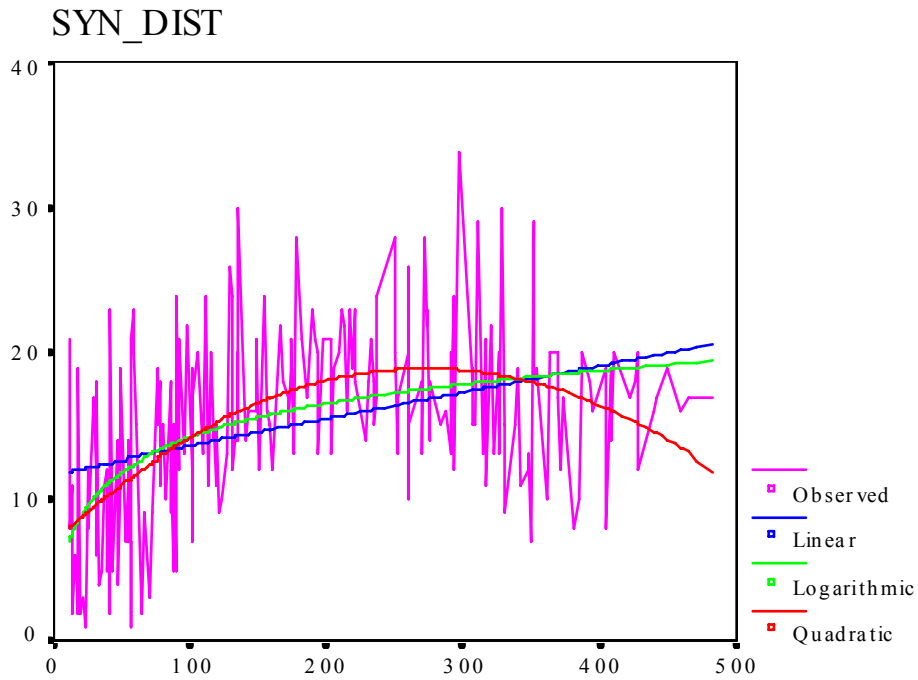
7b) Atomic versus feature variables



- MDS maps of the Reflexives domain (266 sites)
- *75 atomic variables:*
  $r = 0.93438533$
- *61 feature variables:*
  $r = 0.94461615$

7b) Atomic versus feature variables





Atomic variables:

r:	0.473	0.532	0.532
error:	*4.11*	*3.95*	*3.96*

Feature variables:

r:	0.377	0.484	0.532
error:	*5.86*	*5.54*	*5.37*

7b)Validation

- Local incoherence
  - A numerical probe to compare different matrices wrt the degree to which they reflect local geography faithfully
    - lower values are better
  - Idea: the relation between linguistic distance and geographic distance levels off
- Atomic variable: $li = 10.3$
- Feature variables: $li = 10.7$

7c) Distance coefficients

- <similarities, differences> pairs
  - separately maintain the number of identical & different variables for each dialect pair
  - Conversion from coefficient to distance value?
    - Use similarities or differences only
    - Incorporate a constant/weight for one dimension
    - Subtract similarities from differences
    - ...

Problem: how to avoid distances < 0

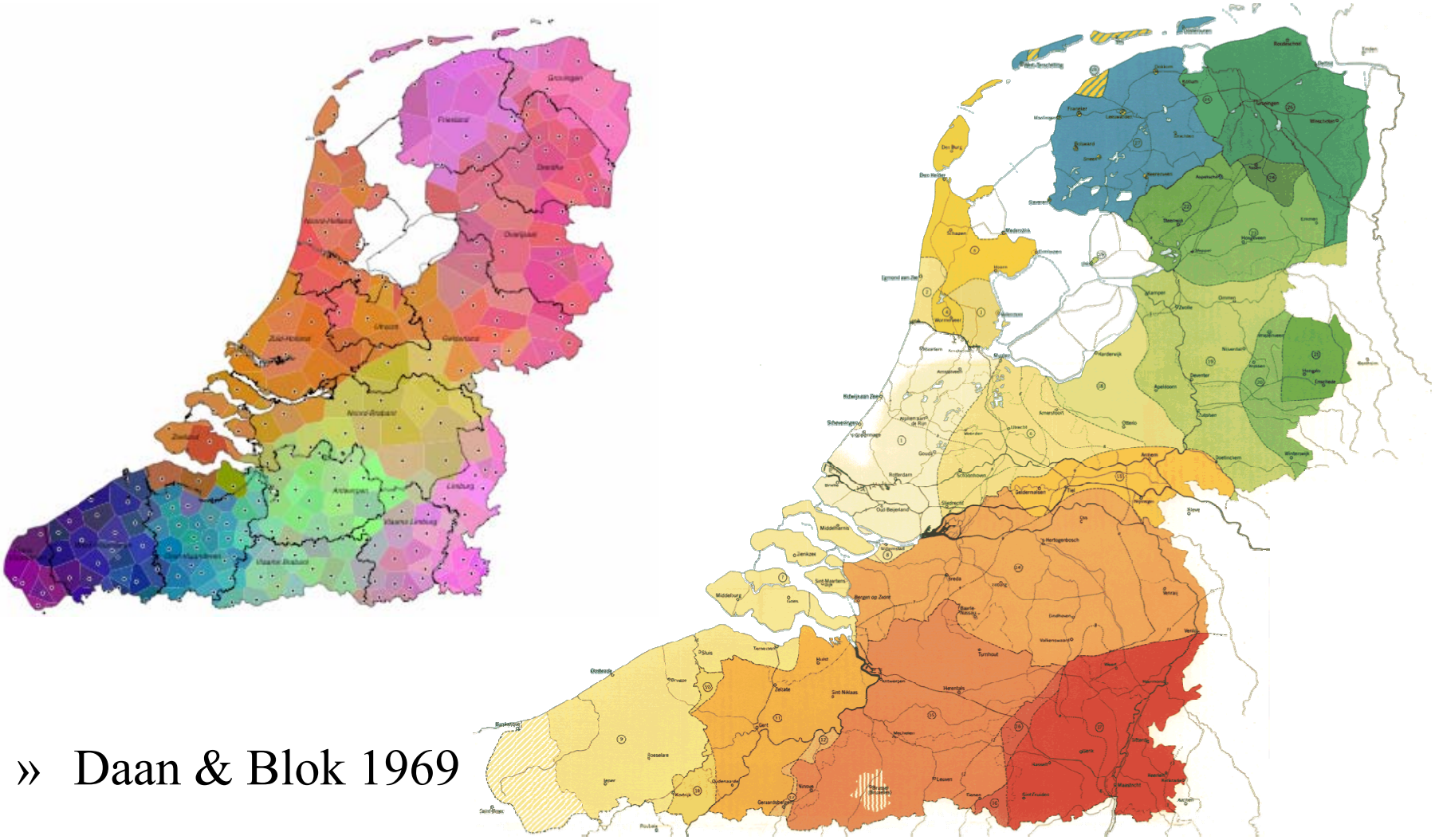
8) Syntactic variation in context

- Syntax versus:
  a) Perception
  b) Pronunciation
  c) Lexis
  d) Geography

8a) Syntax versus perception

- Daan & Blok map (1969)
  - Subjective judgements from 1500 local Dutch dialect speakers, collected in 1939
  - Determination of dialect borders
    - Netherlandic part: Arrow method
      - Neighbouring dialects which local speakers judge to be similar, are connected by arrows » clusters of localities
    - Flanders part: local experts, map designers

8a) Syntax versus perception

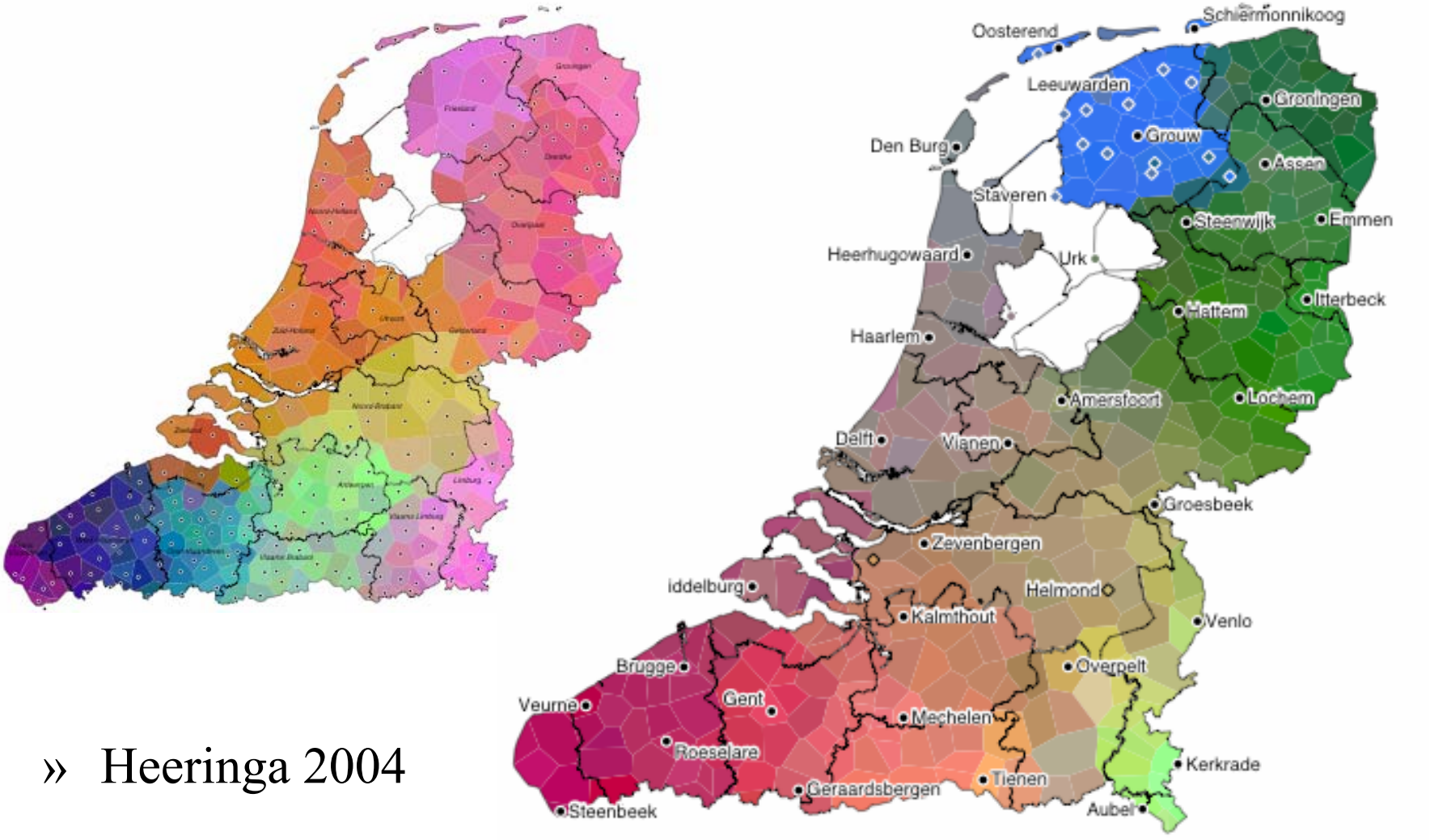


» Daan & Blok 1969

8b) Syntax versus pronunciation

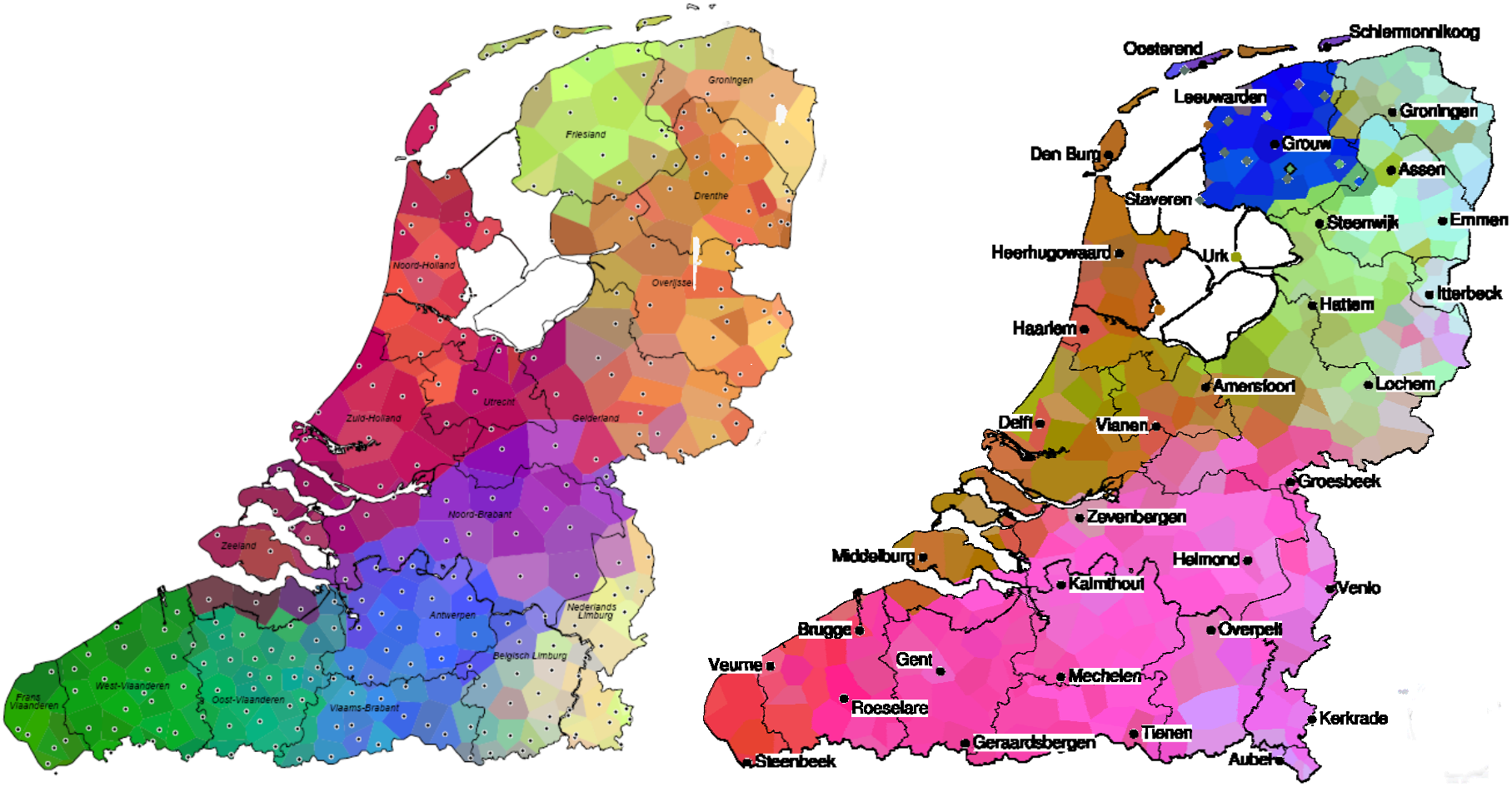
- Reeks Nederlandse Dialectatlassen (RND) / *"Series of Dutch Dialect Atlases"*
- Pronunciation data includes phonetic and morphologic variation
- Also lexical data
- Heeringa (2004)
- Uses Levenshtein distance measure
- Based on 125 words in 360 dialects
- RND ∩ SAND = 68 common dialects

8b) Syntax versus pronunciation

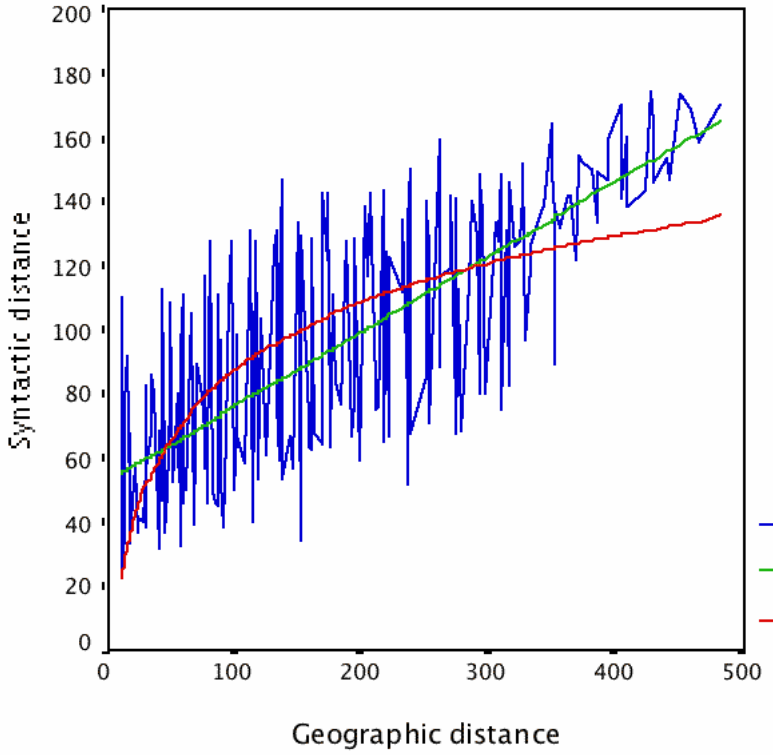


» Heeringa 2004

8c) Syntax versus lexis: GIW

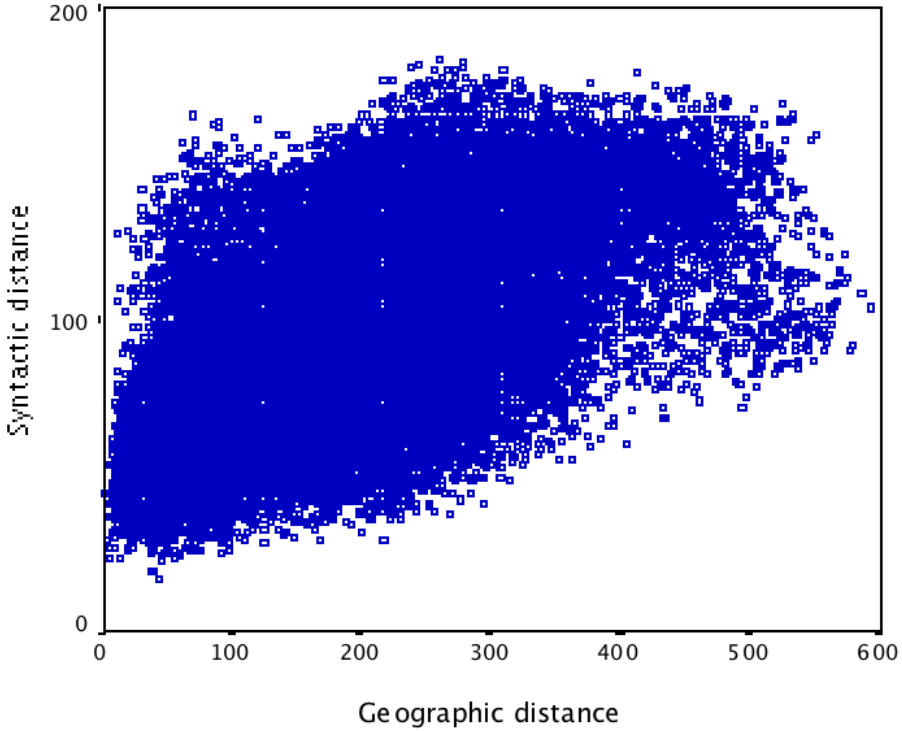


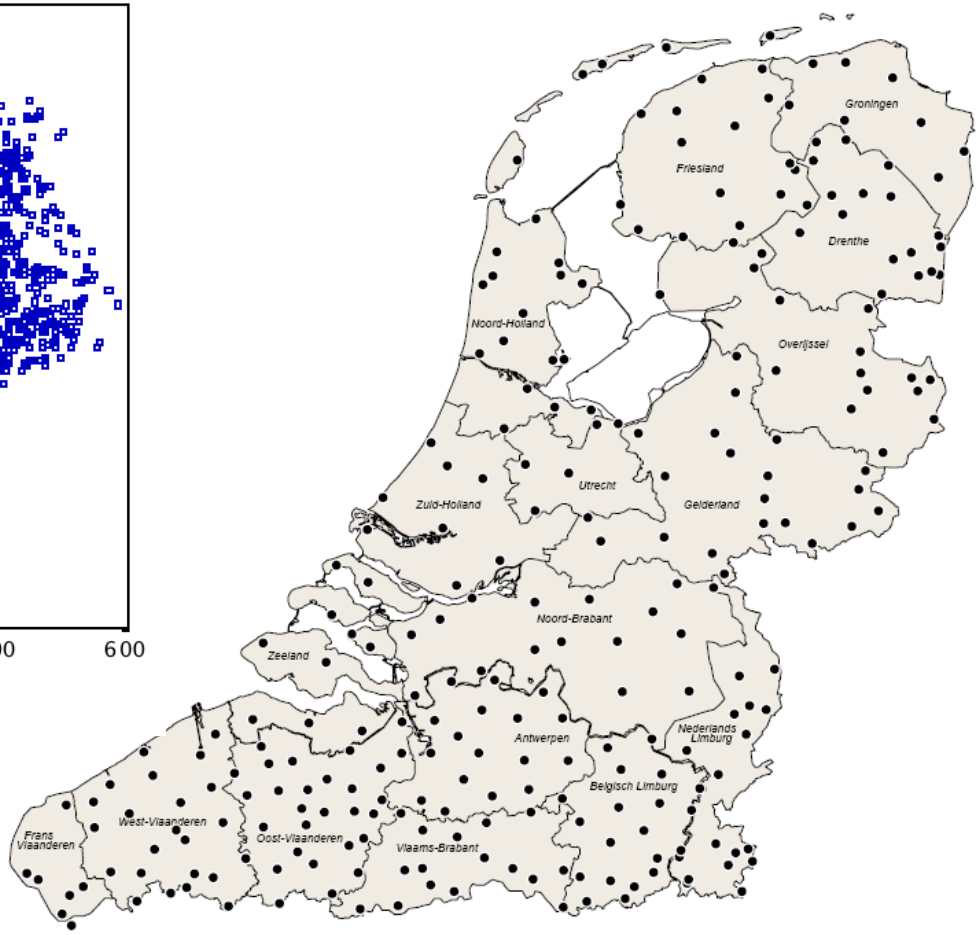
8d) Syntax versus geography



$r = 0.749$

8d) Syntax versus geography II





$r = 0.549$

8e) Correlations

- Upcoming paper on pronunciation, lexis and syntax correlations:
  - with geography
    - For all varieties (360 versus 267)
    - For common subset of 68 varieties
  - without geography
    - Using regression residuals

9a) Variable correlations

- Goal is to eliminate the redundancy in the data by reducing the number of variables
  a) Correspondence analysis
    - Factor analysis method with categorical variables
    - Matrix: variables in rows, locations in columns

<TODO>

9b) Variable correlations

b) Data mining

- Calculate **composite variables** for usage in refined measure
- Discover rules (within one iteration)
  - if variable X then variable Y, and
  - if variable Y then variable Z, then
  - if variable X then variable Z
    - These should automatically emerge when composite variables are calculated and merged recursively
- Secundary variables?

9b) Measure of variable distance

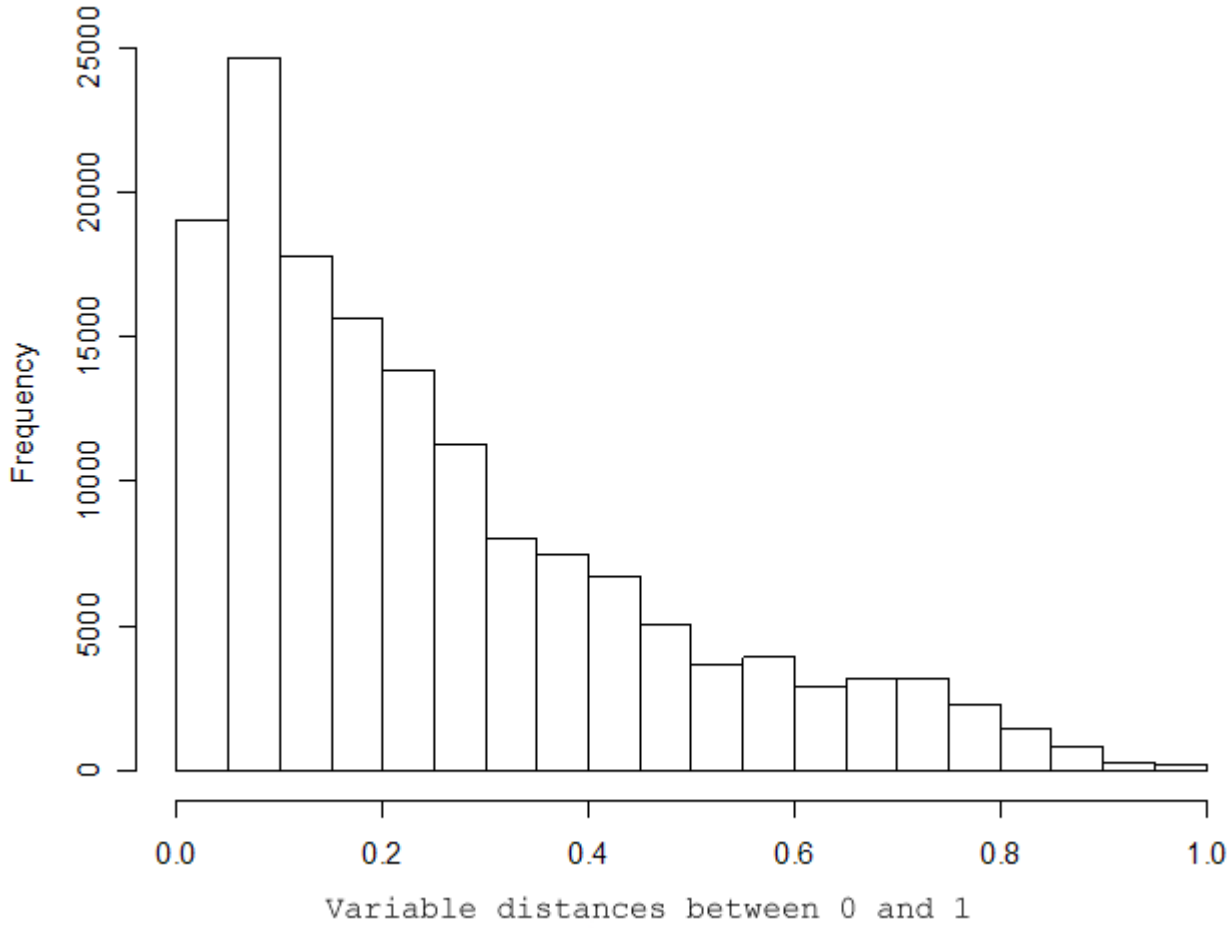
- Hamming distance algorithm based on binary comparisons of variable pairs between dialects
- For example, the syntactic context *Weak reflexive pronoun [...]:*

dialects	*zich*	*zijn eigen*	*distance*
Lunteren	√	√	0
Bellingwolde	√		1
Hollum			0
Doel			0
Sint-Truiden			0
Veldhoven	√		1

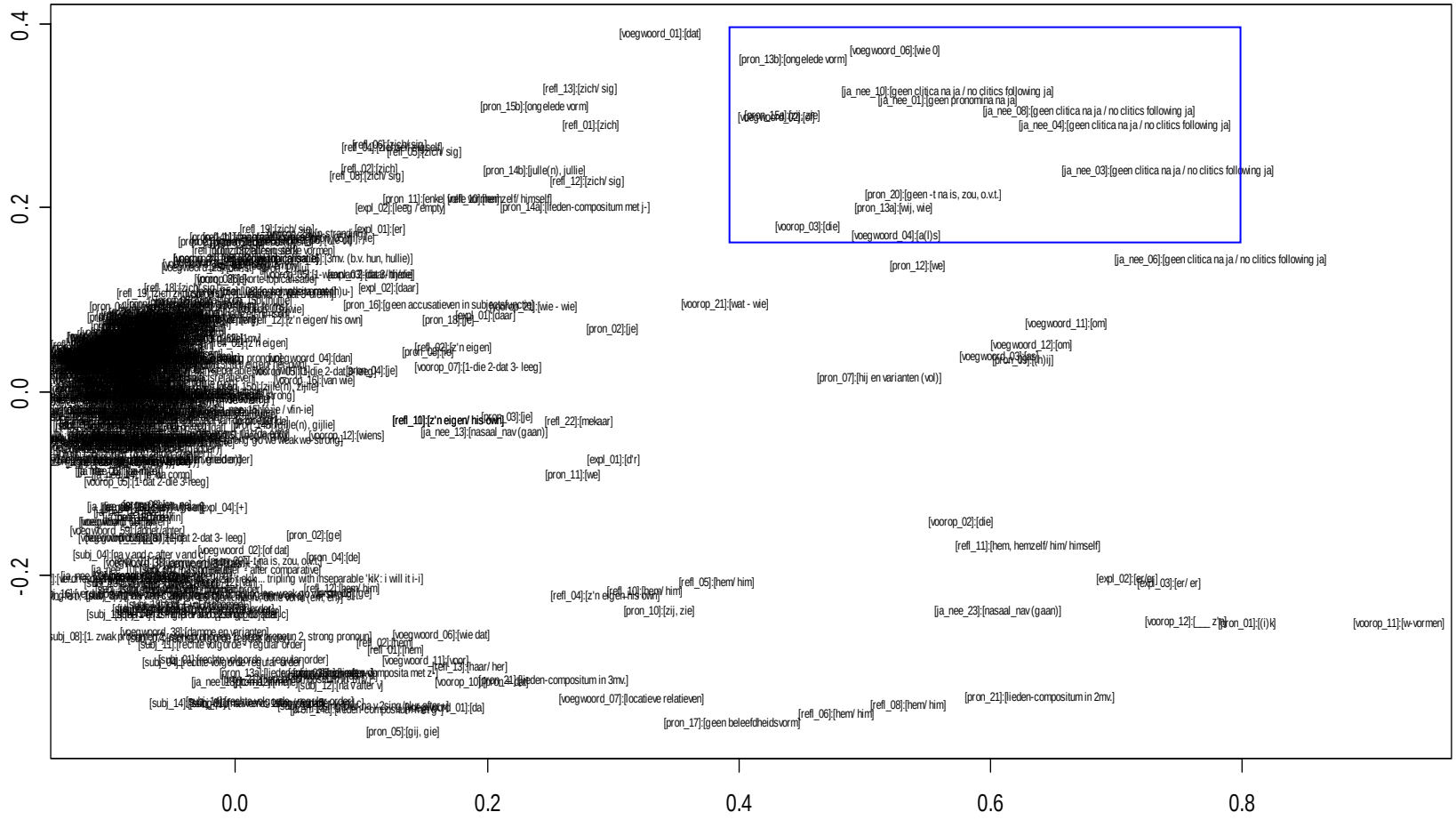
3c) Variable distance matrix

variables	[refl_01]:[hem]	[expl_02]:[er/er]	[voegwoord_62]:[2mv]	voorop_12]:[diens]	[ja_nee_03]:[jaa-ij]
[refl_01]:[hem]		0.520599	0.074906	0.116105	0.707865
[expl_02]:[er/er]	0.520599		0.014981	0.340824	0.014981
[voegwoord_62]:[2mv]	0.074906	0.014981		0.475655	0.805243
voorop_12]:[diens]	0.116105	0.340824	0.475655		0.352060
[ja_nee_03]:[jaa-ij]	0.707865	0.014981	0.805243	0.352060	

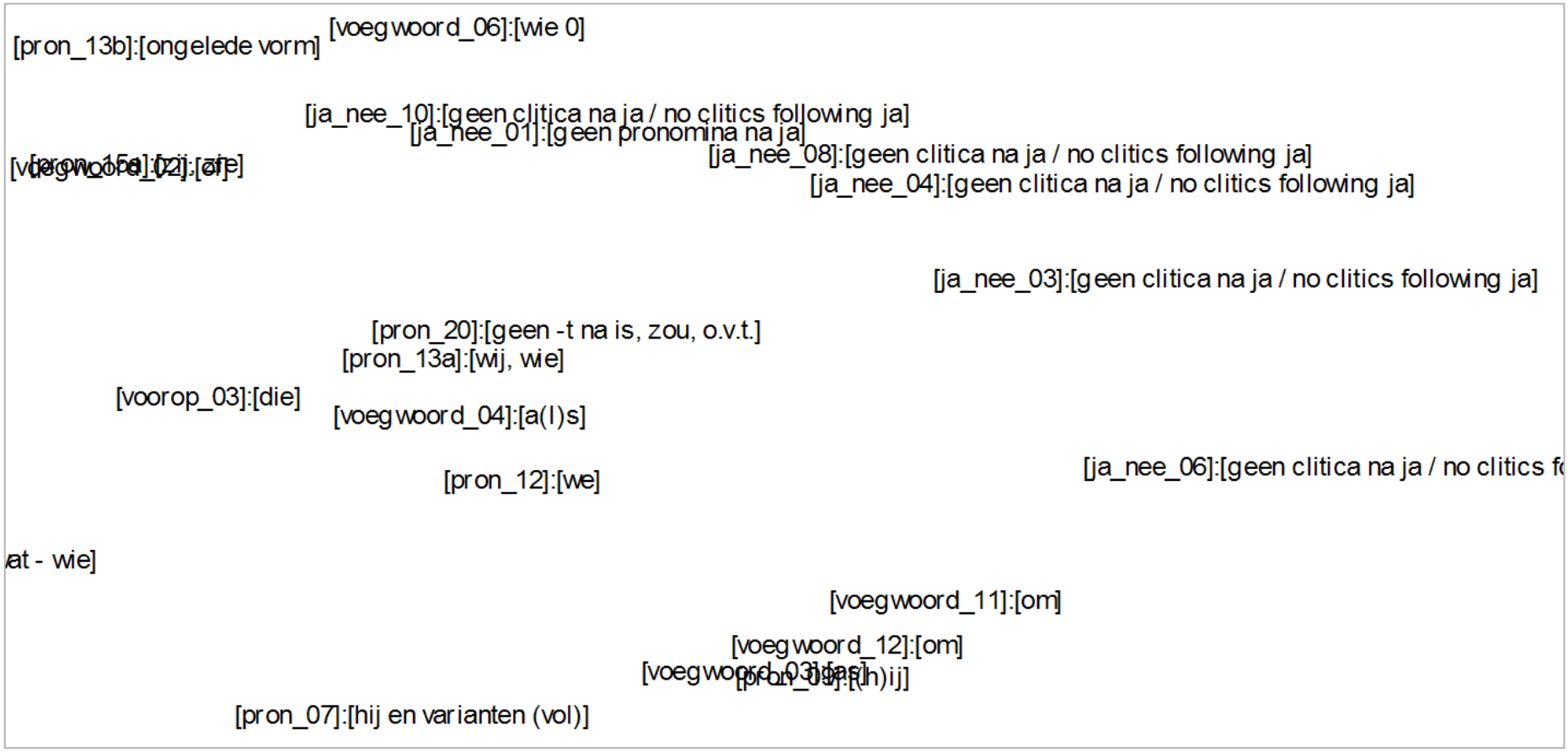
9b) Variable distances histogram



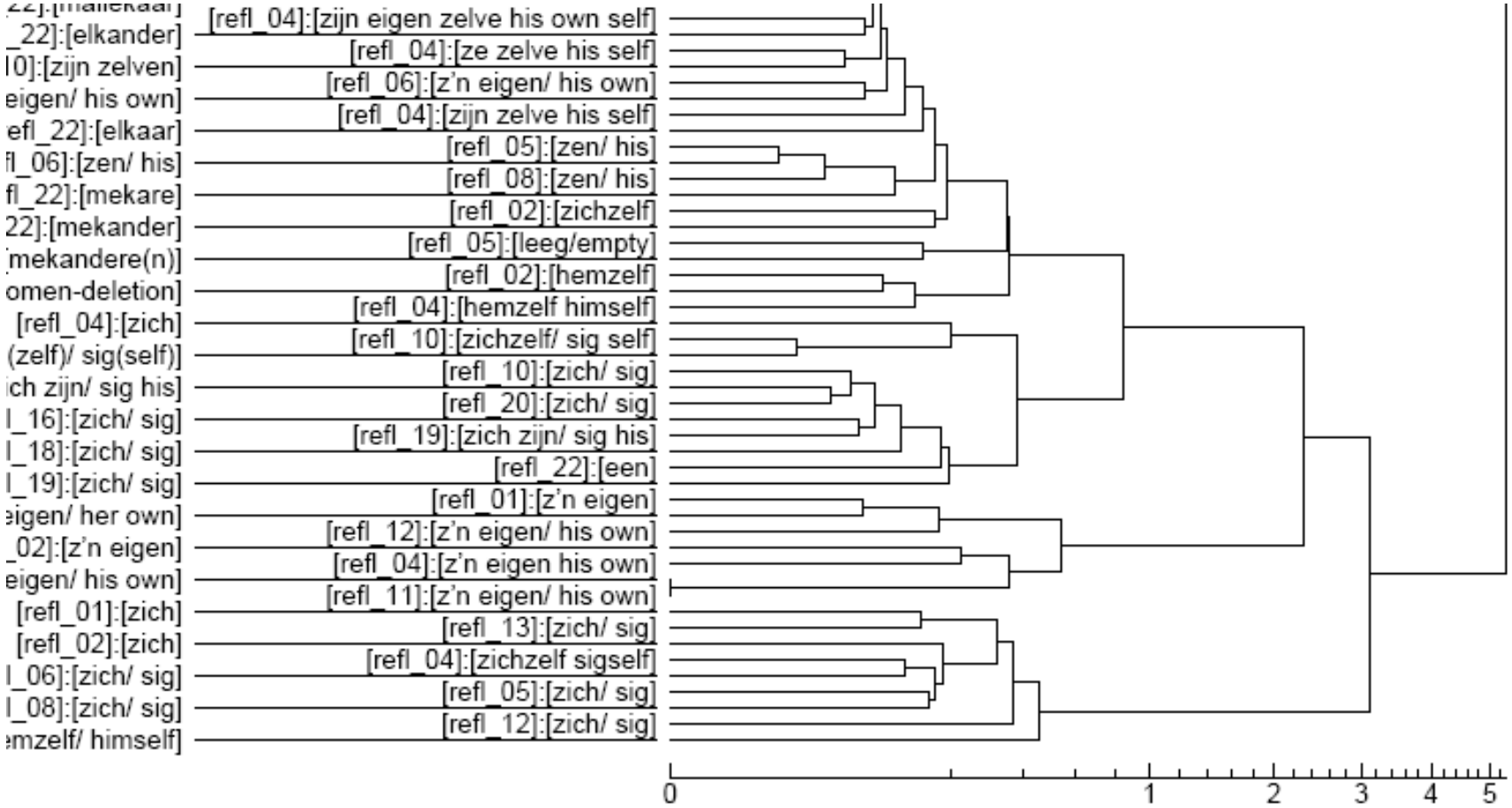
9b) Variable distances MDS plot



9b) Variable distances MDS plot

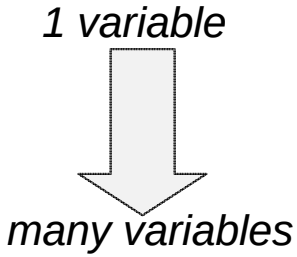


9b) Variable dendrogram



10a) Some conclusions

- There is geographic cohesion in syntactic variation
  Perspectives of syntactic variation in the aggregate
  - Symbolic representation
  - Mosaic-like variable distributions
  - Groups of geographic patterns
  - Continuum of geographic patterns



- Cluster analysis and multidimensional scaling can be used as complementary techniques to help interpret high-dimensional data
- Two-way analysis from both composite and feature variable perspectives to more accurate quantify linguistic variation

10b) Under construction

- Composite variables
  - incorporation in measurement procedure
  - variable dependency exploration
- Correlations between syntactic variation and pronunciation & lexis
- Further measure refinements
  - distance coefficients
  - secundary variables?

*) Relevant software

- RuG/L04
- dialectometry.net/syntax software (DiSS)
  - rd2wgs, consists of conversion routines for Rijksdriehoeksmeting to WGS84 coordinates
  - Geo Foundation Classes (GFC), a library for distance calculations between earth locations
- Visual Dialectometry (VDM)
- Barrier
- SPSS / R

RuG/L04

- Targets research areas:
  - Dialectometry
    - Levenshtein and GIW distance measurements at pronunciation and lexical level
  - Cartography
    - Geographical distribution maps resulting from cluster analysis and multidimensional scaling
- http://www.let.rug.nl/~kleiweg/L04
- *by Peter Kleiweg, University of Groningen*

dialectometry.net/syntax (dis)

- Dialect distance measurement *(sync)*
  - The SYNtactic measurement Console *(sync)* program measures the measures the distance between all dialect pairs based on the geographic distributions of syntactic variables
- Variable distance measurement *(varc)*
  - The VARiable Correlation program measures the correlation between all variable pairs based on their geographic distributions
- Data preparation *(dimp)*
  - The *Data IMport Program* transforms SAND output files into an appropriate XML file which can be processed by *sync*
- FINe-tune postscript map Console *(finc)*
- BAtch Script Console *(basc)*

Data format

- CSV » dimp » XML » sync/varc » CSV

```
− <data name="Syntactische Atlas van de Nederlandse Dialecten" project="Determinants of
  − <context name="refl_01" map="68a" description="Zwak reflexief pronome als object
    + <variable name="Hem"></variable>
    + <variable name="Hemzelf"></variable>
    + <variable name="z'n eigen"></variable>
    + <variable name="Zich"></variable>
    − <variable name="Zichzelf">
        <location name="G280p"/>
        <location name="K339p"/>
      </variable>
  </context>
  + <context name="refl_02" map="68b" description="Zwak reflexief pronome als object
```