Three quantitative perspectives on syntactic variation

ACLC lecture, Amsterdam, 23 March 2007, Marco René Spruit
http://www.meertens.knaw.nl/medewerkers/marco.rene.spruit

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
  - *“What are the determinants of dialectal variation?”*

Presentation outline

Three quantitative approaches on syntactic variation:

1. "Classifying Dutch dialects using a syntactic measure"/ "Measuring syntactic variation in Dutch dialects"
2. "Associations among linguistic levels"
3. "Discovery of association rules between syntactic variables"

1

"Classifying Dutch dialects using a syntactic measure"

Syntactic variation, dialectometry, MDS, dialect area classifications

Syntactic variation data

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

    Complementisers, Subject pronouns, Subject doubling, Reflexive and reciprocal pronouns, Fronting
    - 106 syntactic contexts, 485 variables
  - SAND2: [Barbiers et al. 2007]

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

SAND1 domains

1. Complementisers
   - ‘t lijkt wel **of** er iemand in de tuin staat.
     *“it looks AFFIRM **if** there someone in the garden stands”*
2. Subject pronouns
   - Ze gelooft dat **jij** eerder thuis bent dan ik.
     *“she believes that **you** earlier home are than I”*
3. Subject doubling
   - As-**ge gij** gezond leeft, leef-**de gij** langer.
     “if you_{weak} $\mathbf{you_{strong}}$ healthily live, live you_{weak} $\mathbf{you_{strong}}$ longer”
4. Reflexive and reciprocal pronouns
   - Jan herinnert **zich** dat verhaal wel.
     *“john remembers **himself** that story AFFIRM”*
5. Fronting
   - Dat is de man **die** het verhaal heeft verteld.
     *“that is the man **who** the story has told”*

Dialectometric methods

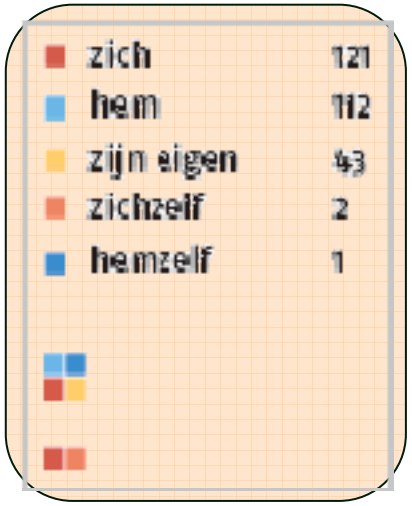
- A *quantitative* research perspective
  - Assign *numerical* values to linguistic variables
  - Using a *measure* of linguistic distance
  - *Add up* individual variables to *objectively* arrive at more general description (versus interpreting isogloss bundles)
  - Examine *aggregated* differences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

Syntactic context & variables

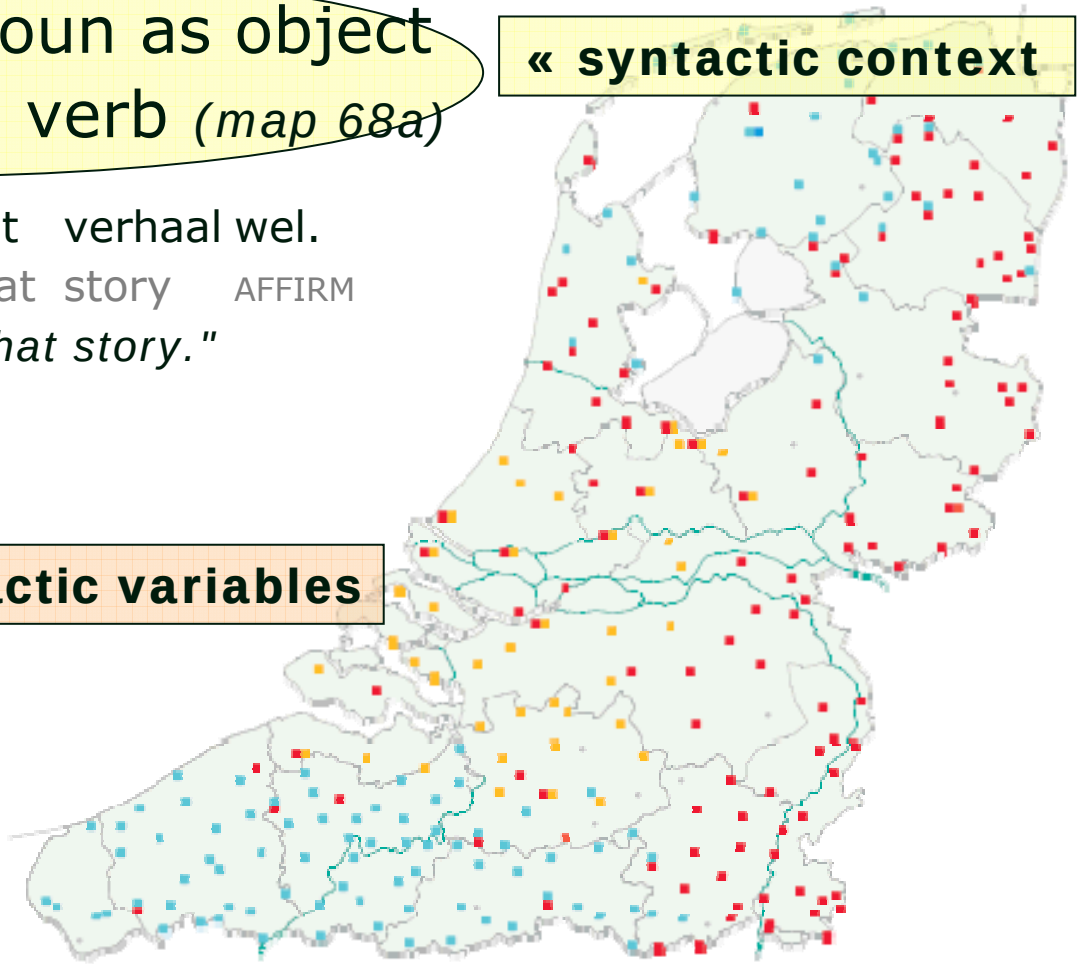
Weak reflexive pronoun as object of inherent reflexive verb *(map 68a)*

« **syntactic context**

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



« **syntactic variables**



Hamming distance

- Syntactic context in SAND1 map 68a

  Weak reflexive pronoun as object of inherent reflexive verb:

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

"John certainly remembers that story."

variable	***Lunteren***	***Veldhoven***	***distance***
r68a:zich	√	√	0
r68a:hem			0
r68a:zijn_eigen	√		1
r68a:zichzelf			0
r68a:hemzelf			0

Distance between the dialects of Lunteren and Veldhoven = 1

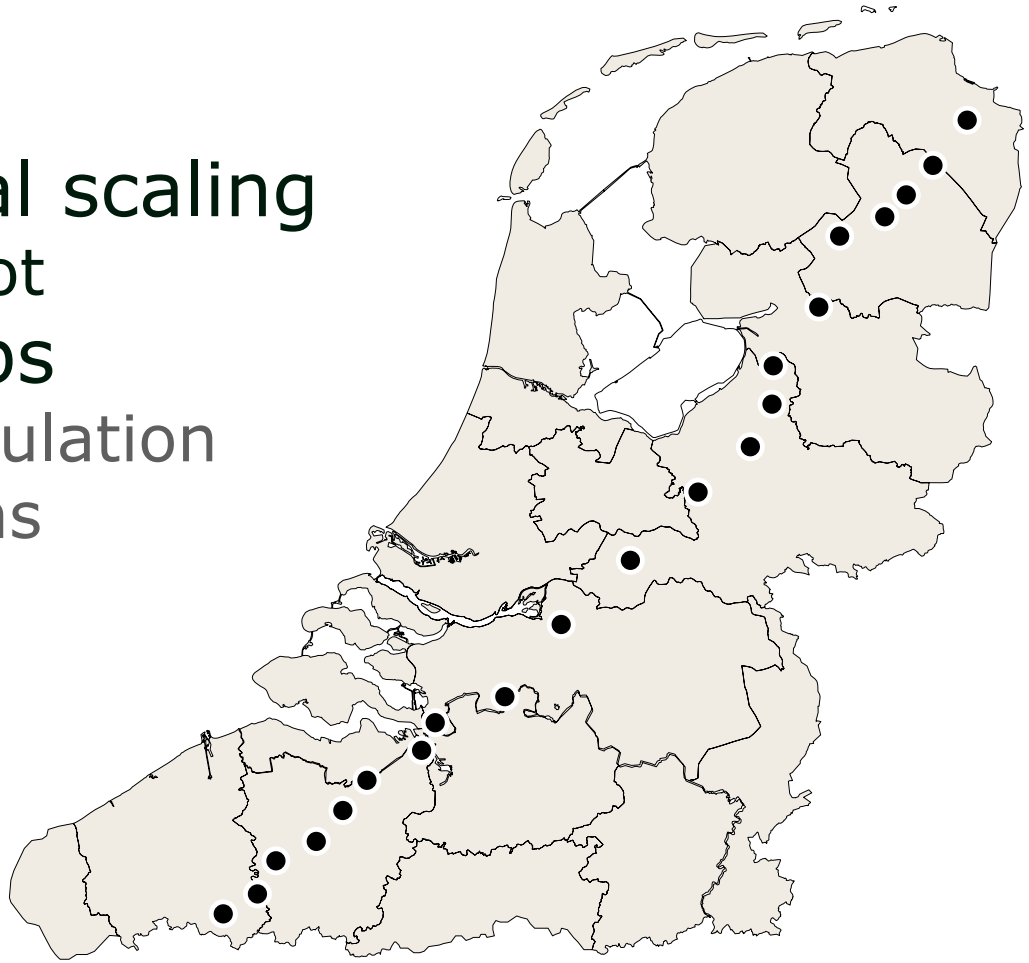
(1 / 5) * 100 = 20 %

Distance matrix

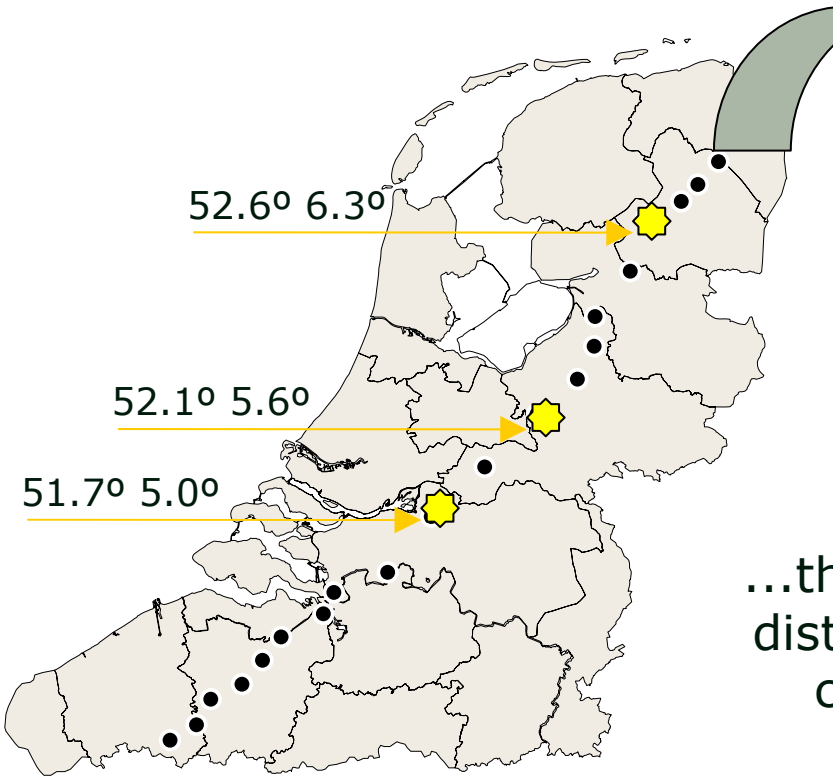
dialect	Lunteren	Bellingwolde	Hollum	Doel	Sint-Truiden	Veldhoven
Lunteren		0.128	0.109	0.237	0.153	**0.095**
Bellingwolde	0.128		0.109	0.258	0.153	0.099
Hollum	0.109	0.109		0.227	0.126	0.122
Doel	0.237	0.258	0.227		0.225	0.216
Sint-Truiden	0.153	0.153	0.126	0.225		0.140
Veldhoven	**0.095**	0.099	0.122	0.216	0.140	

Interpretation of results

1. Cluster analysis
   - Dendrogram
2. Multidimensional scaling
   - Generic MDS plot
3. Topological maps
   - Delauney triangulation
   - Voronoi polygons
   - Cluster maps
   - **MDS maps**
   - Hybrid maps
   - Barrier maps



Multidimensional scaling *(MDS)*

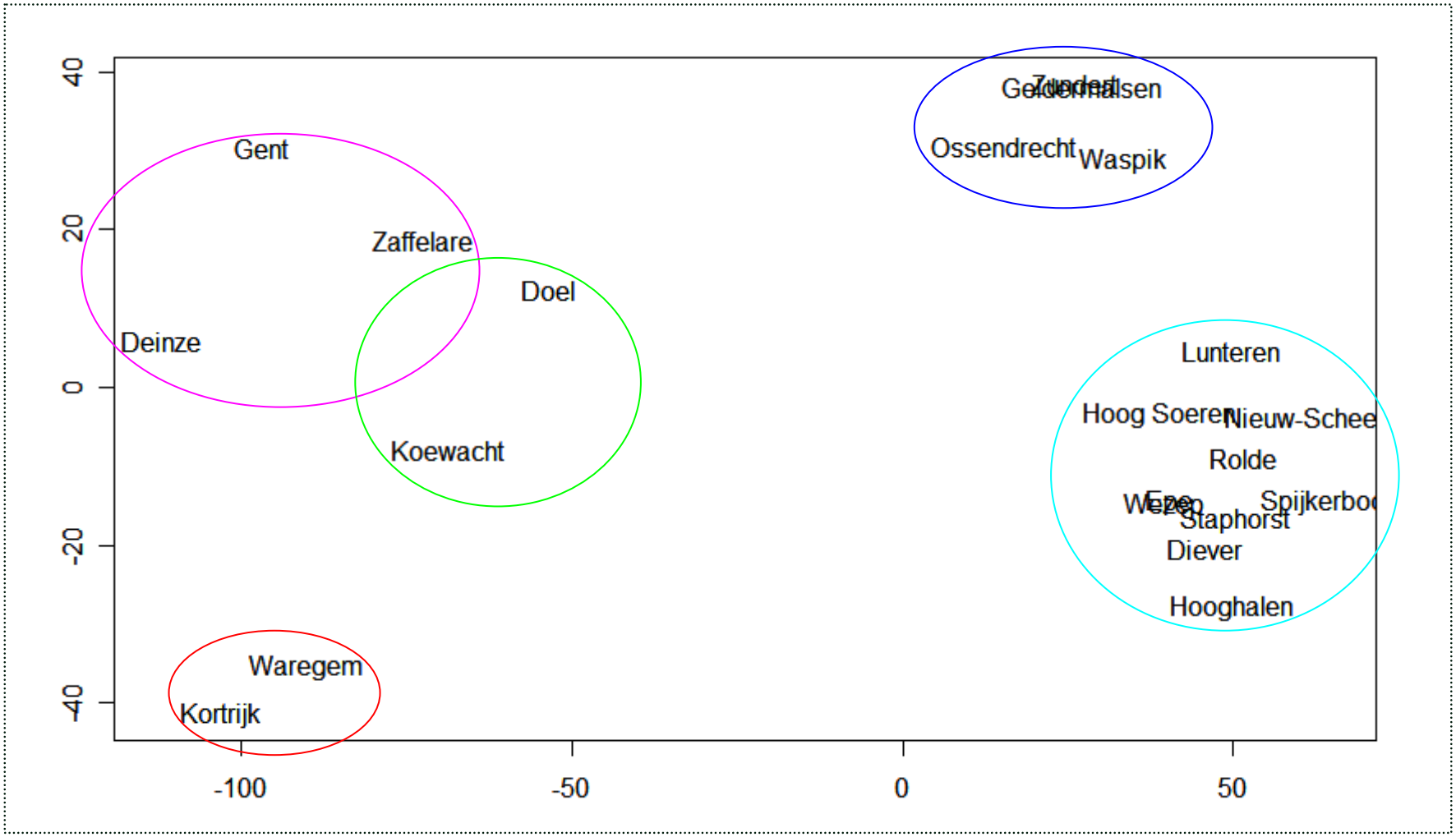


Instead of using coordinates to calculate the distance between locations...

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

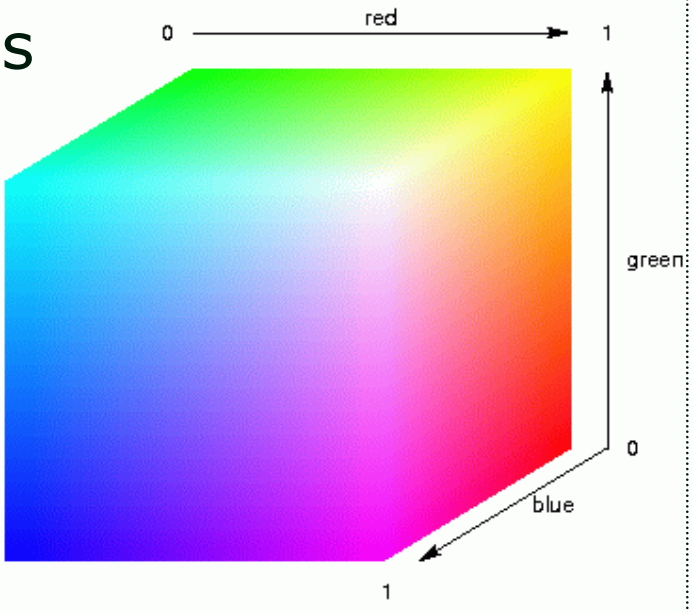
...the MDS algorithm uses the distance between locations to calculate the coordinates...

MDS plot


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

Map colours using MDS

- 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

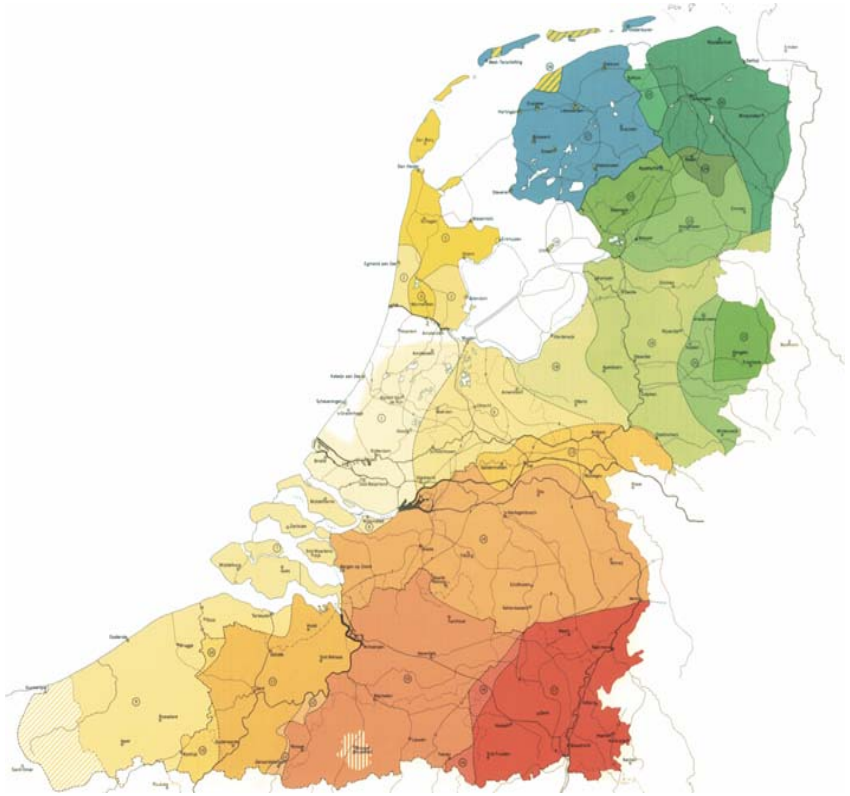
Continuum versus mosaic maps

- Continuum map

- Mosaic map

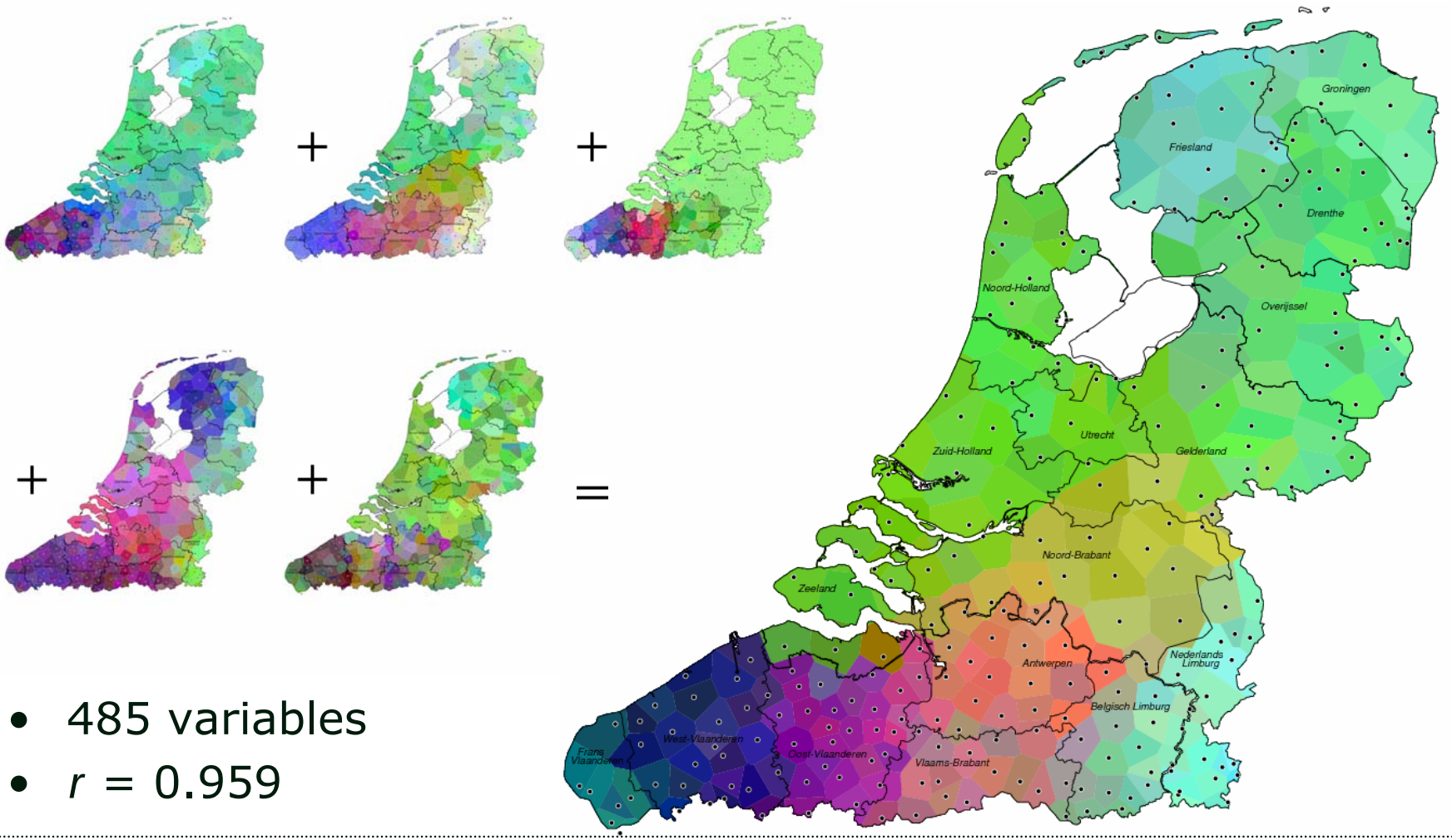
External reference maps

- Daan & Blok map
  (based on Perception)



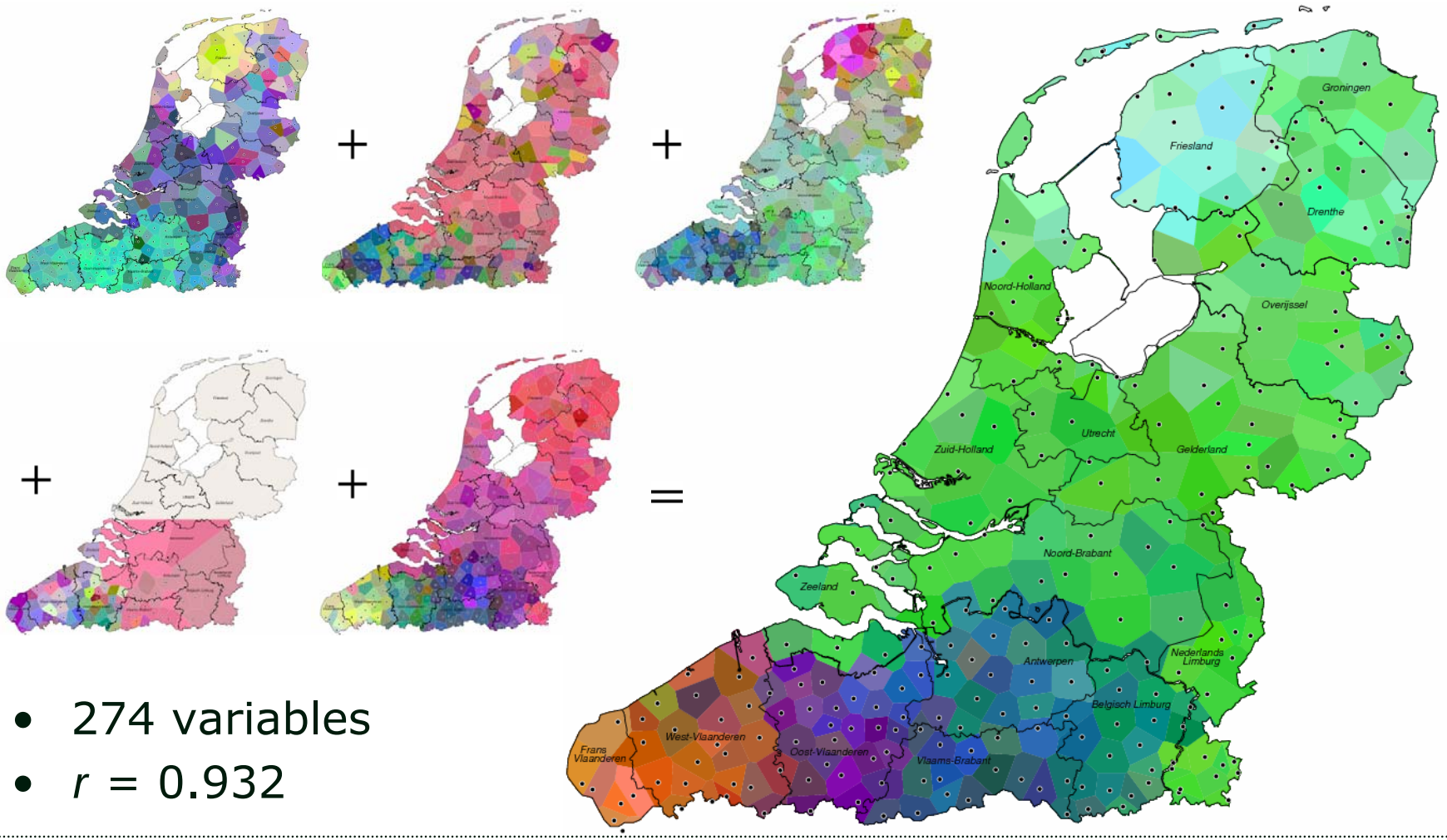
- De Schutter map
  (based on expert opinion)

SAND1



- 485 variables
- r = 0.959

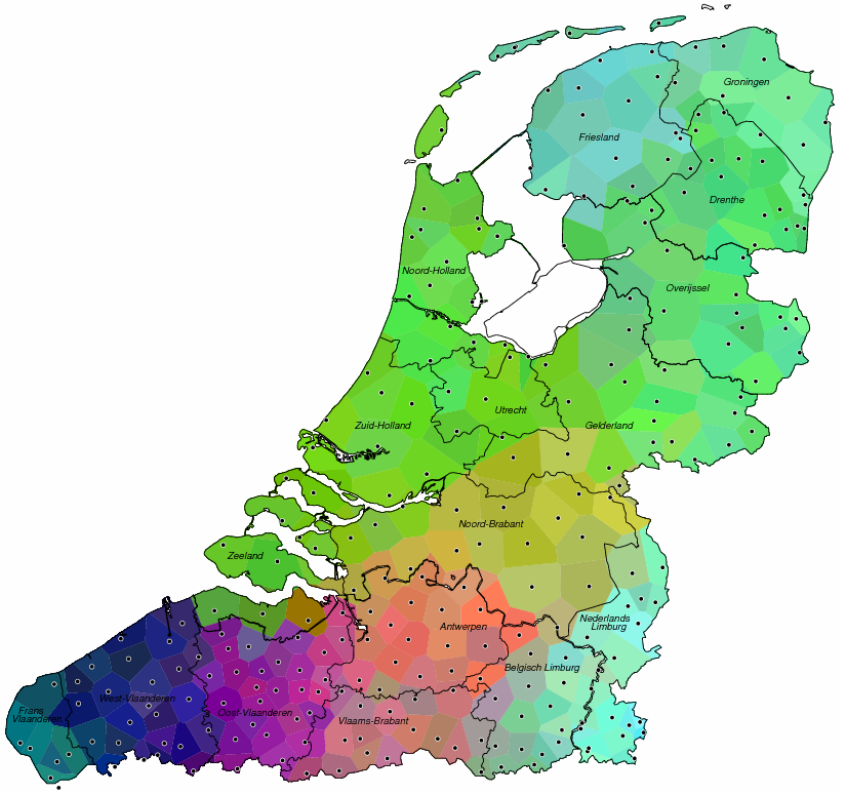
SAND2



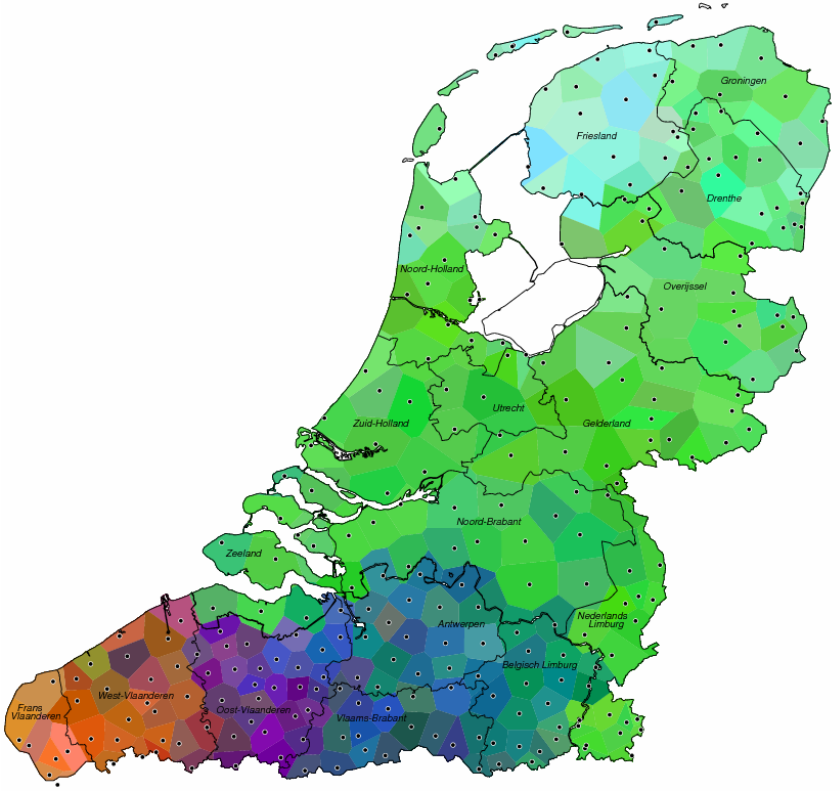
- 274 variables
- r = 0.932

SAND1 versus SAND2

SAND1 +



SAND2 = ...



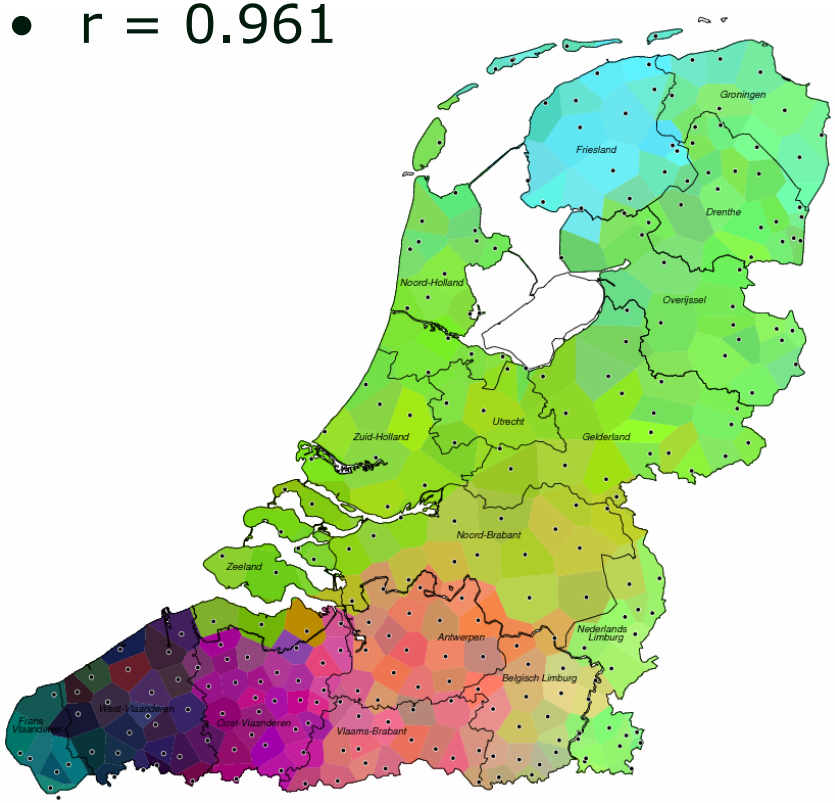
SAND

Cluster analysis animation

- Ward's method
- 12 clusters

Classical MDS

- 759 variables
- r = 0.961



Method reliability & measure refinements

Cronbach's α, Jaccard & GIW distances, feature & composite variables,...

Consistency in SAND1

Syntactic domain	*# variables*	*Cronbach's α*
Complementisers	84	0.867
Subject pronouns and expletives	189	0.791
Subject doubling and clitisation	78	0.748
Reflexive pronouns	74	0.872
Fronting	59	0.589
SAND1	**484**	**0.94**

Consistency in SAND2

Syntactic domain	***Cronbach's α***		
Verbal clusters	0.549	0.881	0.825
Cluster interruption	0.604		
Morphosyntactic variation	0.480		
Negative particle	0.672	0.753	
Negative concord and quantification	0.686		
SAND 1 + 2	**0.955**		

Jaccard distance

- Jaccard distance = 1 - (intersection/union)

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

"John certainly remembers that story."

variable	***Lunteren***	***Veldhoven***	***distance***
r68a:zich	√	√	0
r68a:hem r68a:zijn_eigen	√		1
r68a:zichzelf r68a:hemzelf			

Distance between the dialects of Lunteren and Veldhoven = 1

(1 - (1 / 2)) * 100 = 50 %

GIW distance

- GIW (Goebl 1984): Frequency-weighted similarity
  - Infrequent matches count more heavily

variable	***Lunteren***	***Veldhoven***	***distance***
r68a:zich	√	√	121/266 = 0.45
r68a:hem			
r68a:zijn_eigen	√		= 1
r68a:zichzelf			
r68a:hemzelf			
			= 1.45

Distance between the dialects of Lunteren and Veldhoven

(1.45 / 2) * 100 = 73 %

Lunteren	zich	zijn_eigen	
Veldhoven	zich	zich	
GIW distance	0.45	1	**= (1.45 / 2) * 100 = 73 %**

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			√	√	√

Measuring feature variables

- Using Hamming distance on atomic variables on SAND1 map 68a: 1/5 * 100 = 20%

	Lunteren	Veldhoven	distance
	{zich, zijn eigen}	{zich}	
r68a: personal			0
r68a: reflexive	√	√	0
r68a: possessive	√		1
r68a: ownness	√		1
r68a: focus			0
			2 differences
		Hamming distance:	**2 / 5 = 0.4**
		Jaccard distance:	**2 / 3 = 0.66**

2

"Associations among linguistic levels"

with Wilbert Heeringa and John Nerbonne

Degrees of association between pronunciation, lexis and syntax

Association questions

1. To what degree are aggregate pronunciational, lexical and syntactic distances associated with one another when measured among varieties of a single language?
    - Are syntax and pronunciation more strongly associated with one another than either is associated with lexical distance?
2. Is there evidence for influence among the linguistic levels, even once we control for the effect of geography?
    - Do syntax and pronunciation more strongly influence one another than either (taken separately) influences or is influenced by lexical distance?

Data sources

- Pronunciational variation & Lexical variation:
  - Series of Dutch Dialect atlasses

    [**RND**: Blancquaert & Peé 1925-1982]
    - 360 dialects, 125 words in phonetic transcription

      RND contains 1956 translations of 139 sentences
- Syntactic variation:
  - SAND1

RND ∩ SAND

RND ∩ SAND

» 360 ∩ 267 locations

= 70 common dialects

Distance measures

- Levenshtein distance { $0 \leq d \leq 1$ }
  - Minimum cost of optimal alignment between words
  - Measures variation in pronunciation numerically
  - *To measure pronunciational differences*
- G.I.W. distance { $0 \leq d \leq 1$ }
  - Frequency-weighted comparisons between nominal variables
  - Rarely used variables count more heavily than more frequent ones
  - Measures lexical & syntactic variation at a nominal level
  - *To measure lexical and syntactic differences*

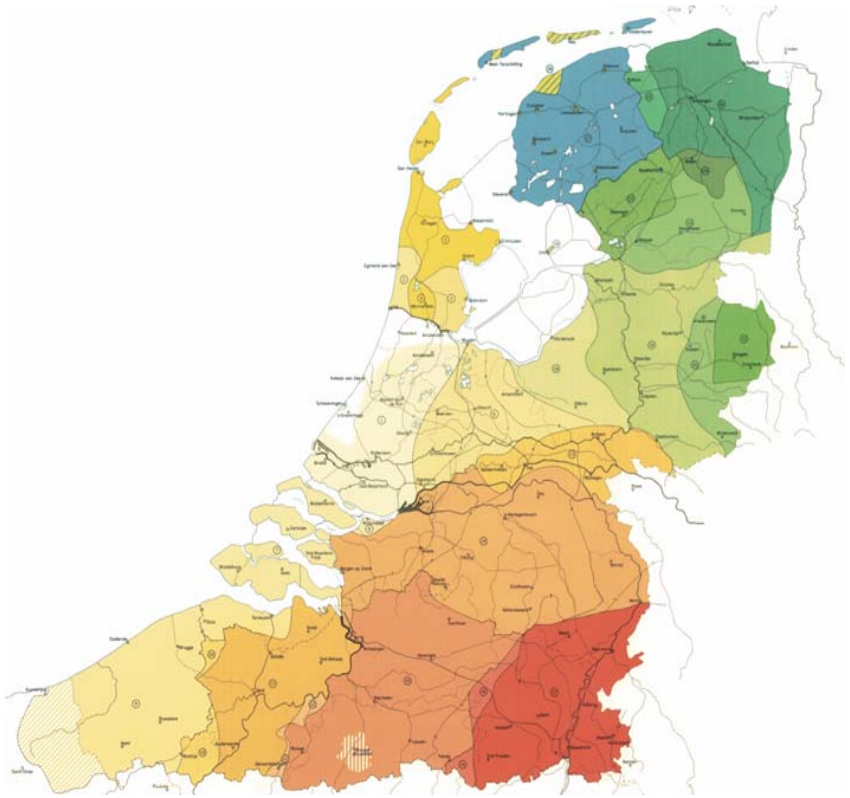
Levenshtein distance

- String alignment and Levenshtein distance calculation between two pronunciations of the Dutch word *hart* 'heart'.

Alignment	**[hart]**	**[ærtə]**	***Edit operation***	***Cost***
1	h		delete h	1
2	a	æ	substitute æ for a	1
3	r	r		0
4	t	t		0
5		ə	insert ə	1
				—
			Levenshtein distance =	3 / 5 = 0.6

Perception versus expert opinion

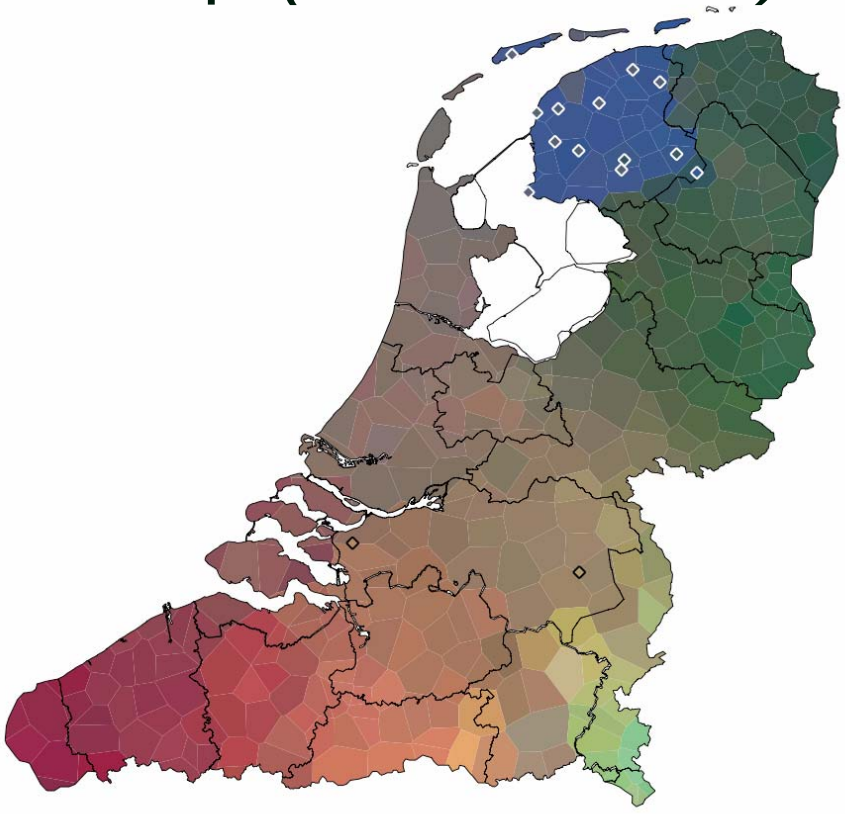
- Daan & Blok map *(Arrow method)*



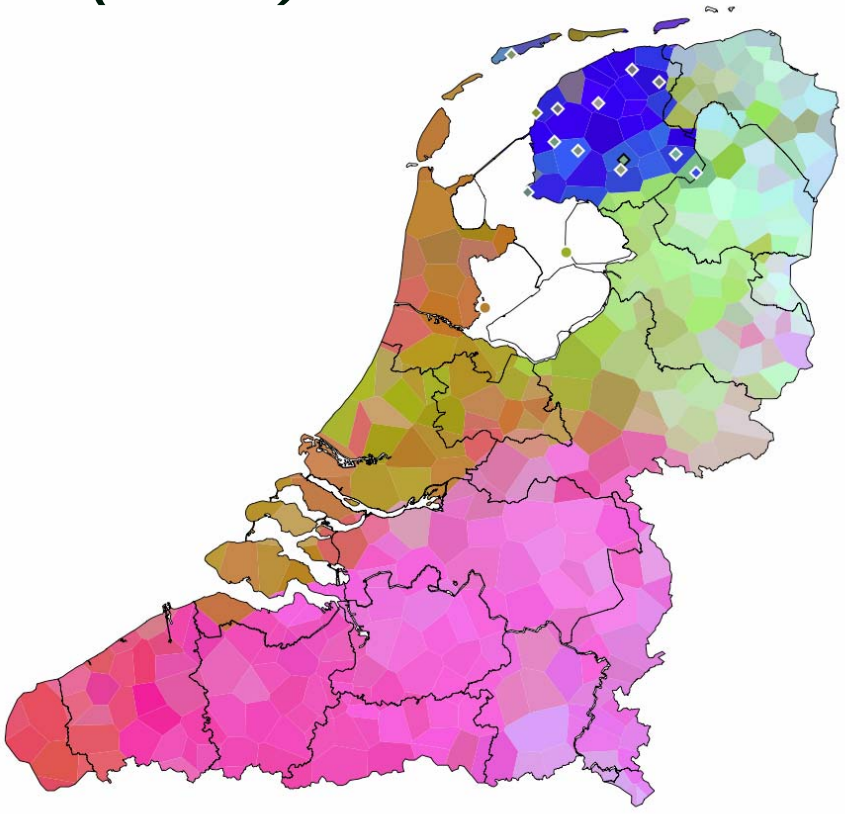
- De Schutter map *("expert opinion")*

Pronunciation versus lexis

- Pronunciation MDS map *(Levenshtein)*

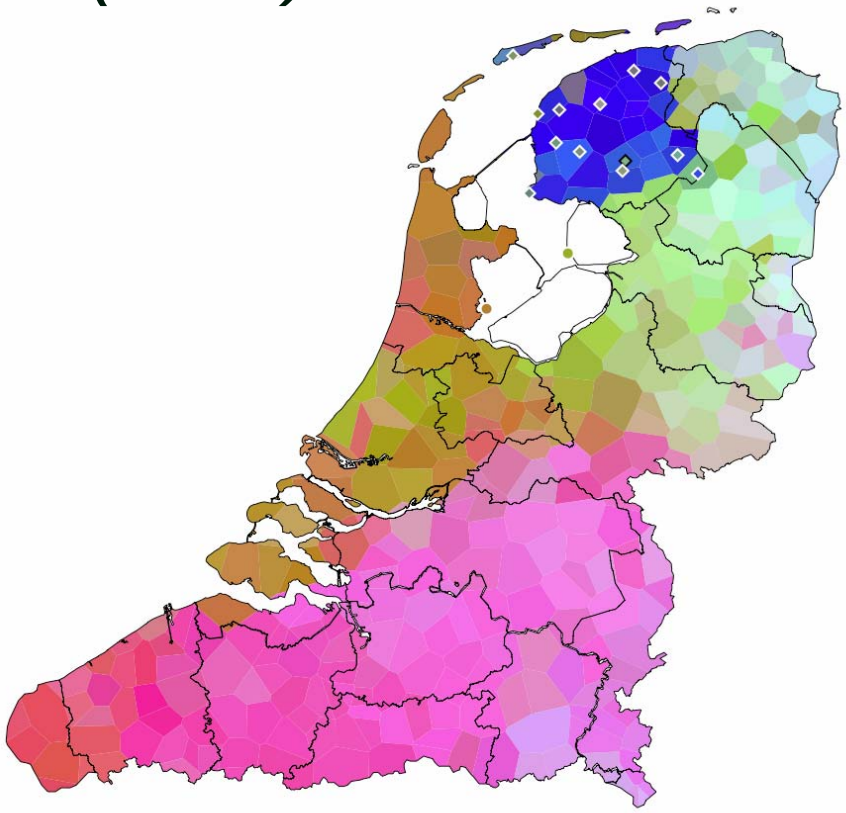


- Lexis MDS map *(GIW)*

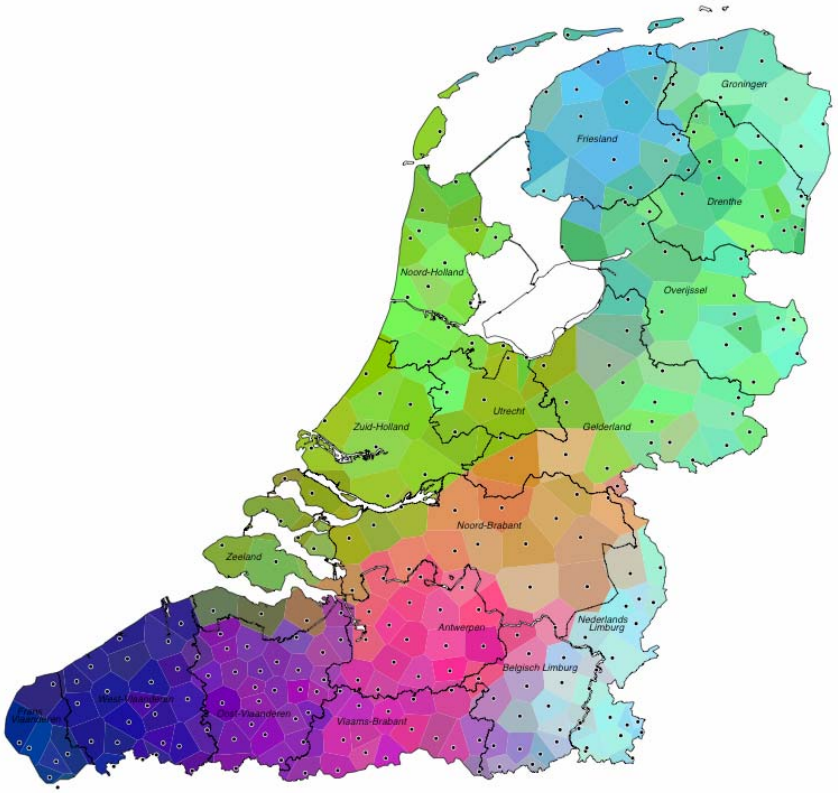


Lexis versus syntax

- Lexis MDS map *(GIW)*

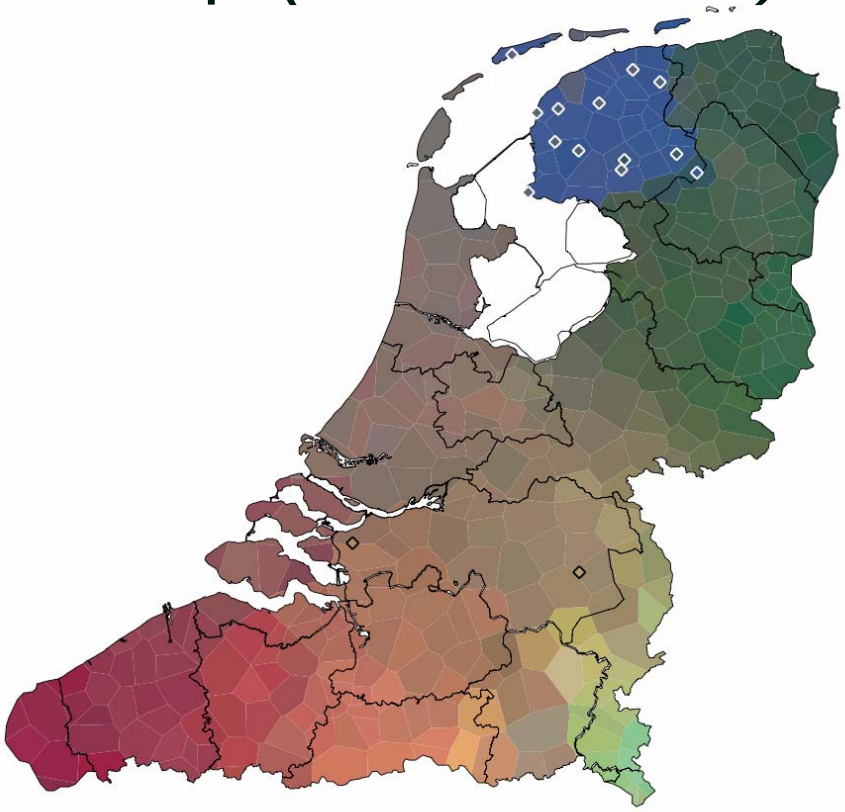


- Syntax MDS map *(GIW)*

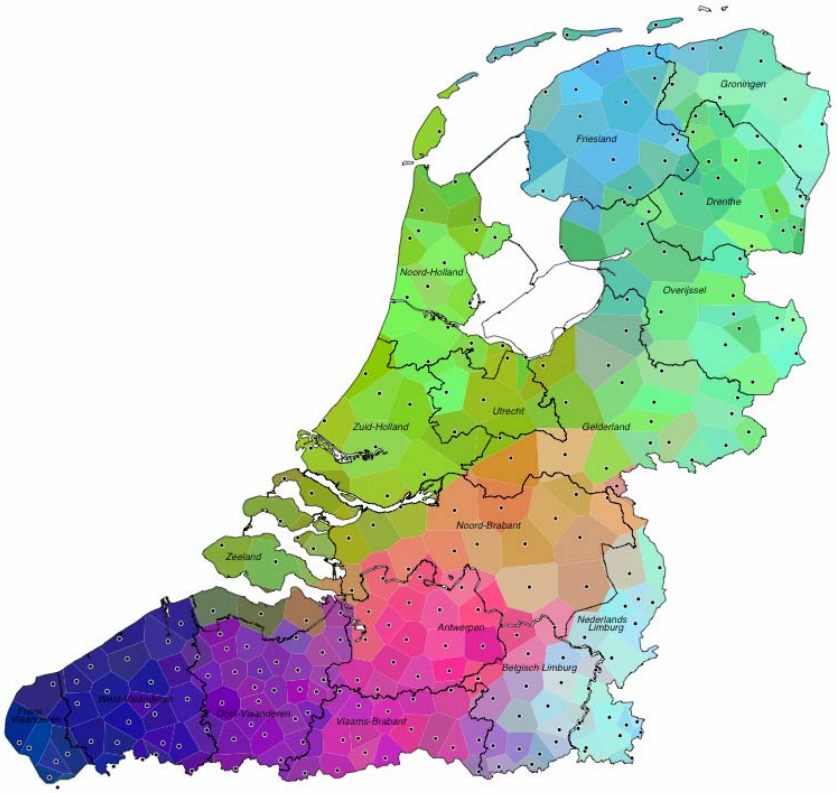


Pronunciation versus syntax

- Pronunciation MDS map *(Levenshtein)*



- Syntax MDS map *(GIW)*



Consistency

- Cronbach's alpha: A coefficient of consistency to measure the minimum reliability (0 <= α <= 1)

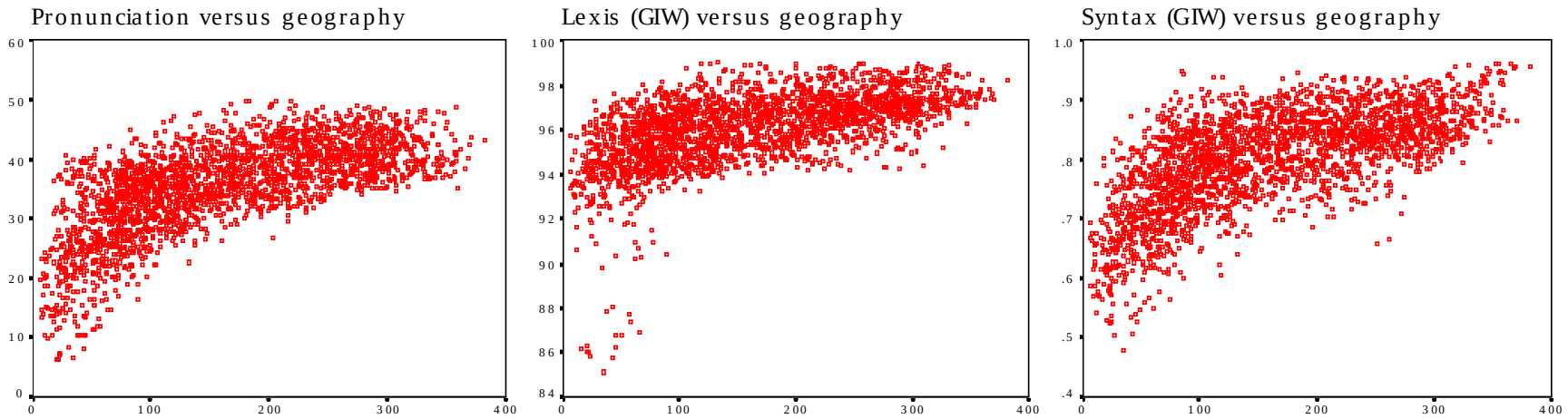
Linguistic level	***# variables***	***Cronbach's α***
Pronunciation	125	0.97
Lexis	107	0.75
Syntax	106	0.94

Correlations among linguistic levels I

- Based on the 70 common varieties
- Using Levenshtein (pronunciation) and GIW (lexis and syntax) distance measures
- For all correlation coefficients: $p < 0.001$

Linguistic level 1	⇔	***Linguistic level 2***	***r***	***r^2 * 100***
Pronunciation	⇔	Lexis	0.617	38 %
Lexis	⇔	Syntax	0.496	25 %
Syntax	⇔	Pronunciation	0.648	42 %

Geographic distributions



- Based on the 70 varieties
- From left to right:
  - Pronunciation versus geography
  - Lexis versus geography
  - Syntax versus geography

Correlations with geography

- Using the 70 common varieties
- Using Levenshtein (pronunciation) and GIW (lexis and syntax) distance measures
- For all correlation coefficients: $p < 0.001$

Linguistic level	⇔	***Geography***	***r***	***r^2 * 100***
Pronunciation	⇔	Geography	0.685	47 %
Lexis	⇔	Geography	0.575	33 %
Syntax	⇔	Geography	0.669	45 %

Correlations among linguistic levels II

- *Without the influence of geography as third factor*
- Based on the 70 common varieties
- Using Levenshtein (pronunciation) and GIW (lexis and syntax) distance measures
- For all correlation coefficients: $p < 0.001$

Linguistic level 1	⇔	***Linguistic level 2***	***r***	***r^2 * 100***
Pronunciation	⇔	Lexis	0.374	14 %
Lexis	⇔	Syntax	0.183	3 %
Syntax	⇔	Pronunciation	0.350	12 %

Influence of geograpy as third factor

- *Geography as a factor of influence underlying the associations between linguistic levels:*

  (1 - (corr_without_geography / corr_with_geography)) * 100

Linguistic level 1	⇔	***Linguistic level 2***	***Geographic Influence***
Pronunciation	⇔	Lexis	39 %
Lexis	⇔	Syntax	63 %
Syntax	⇔	Pronunciation	46 %

3

“Discovery of association rules between syntactic variables”

Data mining the Syntactic atlas of the Dutch dialects

Data mining the SAND

- Knowledge Discovery in Databases (KDD)
  - "the science of extracting useful information from large data sets or databases" (Hand *et al.*, 2001)
  - An umbrella term for techniques like association *rules*, decision *trees*, neural *networks*, ...
- Association rule mining: A $\rightarrow$ C
  - *A*: predicting attribute value(s) ("antecedent")
  - *C*: predicted class ("consequent")
- Based on proportional overlap
  - Geographical co-occurrences of variables

Sample variables

A. "Complementiser of comparative if-clause" *(14b)*

't	lijkt	wel	**of**	**dat**	er	iemand	in	de	tuin	staat.
it	*looks*	*[affirm]*	*if*	*that*	*there*	*someone*	*in*	*the*	*garden*	*stands*

B. "Subject doubling 2 singular" *(54a)*

Ge	gelooft	gij	zeker	niet	dat	hij	sterker	is	as	**-ge**	**gij.**
you$_{weak}$	*believe*	*you$_{strong}$*	*certainly*	*not*	*that*	*he*	*stronger*	*is*	*than*	*you$_{weak}$*	*you$_{strong}$*

C. "Weak reflexive pronoun as object of inherent reflexive verb" *(68a)*

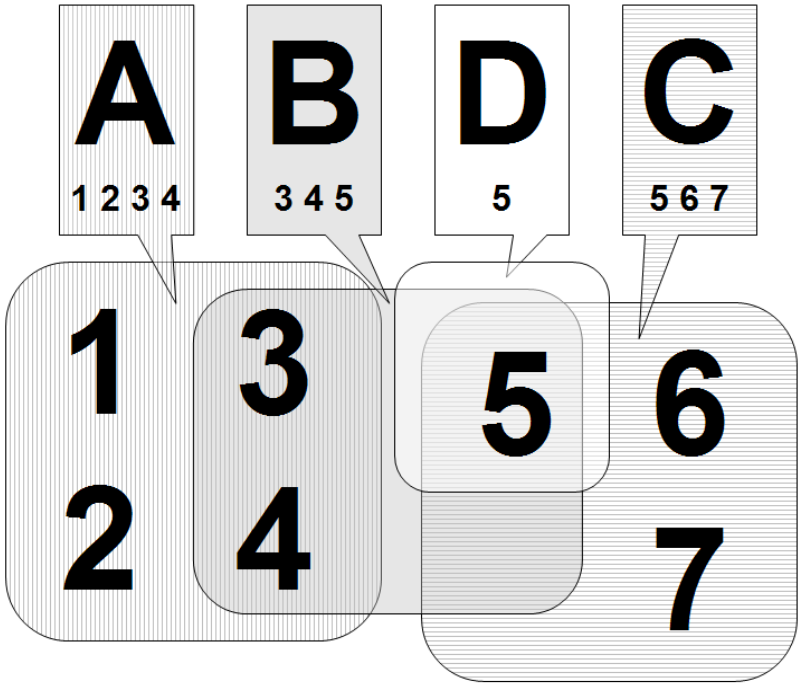
Jan	herinnert	**zijn**	**eigen**	dat	verhaal	wel.
John	*remembers*	*his*	*own*	*that*	*story*	*[affirmative]*

D. "Short subject relative, complementiser following relative pronoun" (84a)

Dat	is	de	man	**die**	**dat**	het	verhaal	verteld	heeft.
that	*is*	*the*	*man*	*who*	*that*	*the*	*story*	*told*	*has*

Sample data illustration

- *Example*: **4** variables (A-D) in **7** locations (1-7)



Evaluation factors of rule quality

- **Accuracy**: |A&C| / |A|

  How often is the rule correct?

  – varA $\rightarrow$ varB: (A $\cap$ B / A) * 100 = 2/4 * 100 = 50%

- **Coverage**: |A|

  How often does the rule apply?

  – varA $\rightarrow$ varB: A / N * 100 = 4/7 * 100 = 57%

- **Completeness**: |A&C| / |C|

  How much of the target class does the rule cover?

  – varA $\rightarrow$ varB: (A $\cap$ B / B) * 100 = 2/3 * 100 = 66%

- **Interestingness**: |A&B| - |A||B|/N

  Integrates the three factors above into one value...

  – varA $\rightarrow$ varB: (A $\cap$ B) - (A * B / N) = 2 – (4 * 3 / 7) = 0.28

Sample data results

The 8 highest ranked association rules:

#	*Antecedent* → *Consequent*	*Interestingness*	*Complexity*	*Accuracy*	*Coverage*	*Completeness*
1.	$B \rightarrow A \vee D$	0.86	1	100	42	60
2.	$A \vee D \rightarrow B$	0.86	1	60	71	100
3.	$D \rightarrow B$	0.57	0	100	14	33
4.	$D \rightarrow C$	0.57	0	100	14	33
5.	$B \rightarrow D$	0.57	0	33	42	100
6.	$C \rightarrow D$	0.57	0	33	42	100
7.	$B \rightarrow A$	0.29	0	66	42	50
8.	$A \rightarrow B$	0.29	0	50	57	66

Interactive exploration...

	A	B	C	D	E
1	#Combination	#Antecedent	#Consequent	#Accuracy	#Coverage
2	10321	p46a:g-[lieden-compositum]	p38b:gij/gie	99	39
3	7681	p46b:julle(n)/jullie	p46a:j-[lieden-compositum]	100	37
4	7503	d55a:na_v	p46a:g-[lieden-compositum]	93	37
5	7514	d55a:na_v	p38b:gij/gie	97	37
6	5640	c27a:da_+_-t	c14a:da	100	36
7	6509	d54a:na_v	d55a:na_v	92	35
8	9653	f88a:1-waar_2-dat	c16b:locatieve_relatieven	100	47
9	6552	d54a:na_v	p38b:gij/gie	98	35
10	6544	d54a:na_v	p46a:g-[lieden-compositum]	93	35
11	1268				
12	1267				
13	9322				
14	10323				
15	10612				
16	8030				
17	5675				
18	10257				
19	7892				
20	5886				
21	3652				

	K	L	M	
1	#ANTE /\ CONS	#ANTE \/ CONS	#ANTE example	#CONS example
2	104	117	We geloven dat G-[LIEDEN-COMPOSITUM] niet zo slim zijn als wij.	Ze gelooft dat GI
3	101	114	We geloven dat JULLE(N)/JULLIE niet zo slim zijn als wij.	We geloven dat J-
4	93	111	As-ge gulder gezond leeft, leef-DE GULDER langer.	We geloven dat G
5	97	118	As-ge gulder gezond leeft, leef-DE GULDER langer.	Ze gelooft dat GI
6	98	121	Je gelooft toch niet DA + -T hij sterker is dan jij?	Ik denk DA Marie l
7	88	106	As-ge gij gezond leeft, leef-DE GIJ langer.	As-ge gulder gezc
8	128	157	De bank WAAR DAT ze op zaten was pas geverfd.	De bank waar op l
9	94	117	As-ge gij gezond leeft, leef-DE GIJ langer.	Ze gelooft dat GI
10	89	111	As-ge gij gezond leeft, leef-DE GIJ langer.	We geloven dat G
11	69	71	ZE heeft -ZE ZIJ daar niks mee te maken.	ZE heeft ZIJ daar
12	69	71	ZE heeft ZIJ daar niks mee te maken.	ZE heeft -ZE ZIJ
13	103	136	Jan herinnert HEM dat verhaal wel.	Johanna laat HAA
14	84	109	A-K IK zuinig leef, leve-K IK zoals mijn ouders willen.	We geloven dat G
15	87	117	A-K IK zuinig leef, leve-K IK zoals mijn ouders willen.	Ze gelooft dat GI
16	74	91	HIJ gelooft HIJ wel dat ik groter ben as tie ij.	A-K IK zuinig leef,
17	96	130	We geloven dat G-[LIEDEN-COMPOSITUM] niet zo slim zijn als wij.	Ik denk DA Marie l
18	101	138	Toon wast HEM.	Johanna laat HAA
19	73	92	'K Geloof-(K) IK wel dat hij groter is als-k ik.	A-K IK zuinig leef,
20	68	81	WE geloven WIJ dat jullie niet zo slim zijn als-me wij.	HIJ gelooft HIJ we

No. 1 association rule in SAND1

Ante: p46a:g-lieden (Subject pronouns 2 plural, strong forms)

We geloven dat **g-lieden** niet zo slim zijn als wij.
we believe that *you*$_{plural,strong}$ *not so smart are as we.*
'We believe that you are not as smart as we are.'

Cons: p38b:gij/gie (Subject pronouns 2 singular, strong forms)

Ze gelooft dat **gij / gie** eerder thuis bent dan ik.
she believes that *you*$_{singular,strong}$ *earlier home are than I*
'She thinks that you'll be home sooner than me.'

Stat: Rank=1, Combination=10,321, Interestingness=58.38, Accuracy=99%, Coverage=39%, Completeness=89%, Complexity=0, A-Locations=105, C-Locations=116, AC-Overlap=104, AC-Disjunction=117

Interp: The plural pronoun 'g-lieden' belongs to the same paradigm as the singular pronoun 'gij'.

More associated rules for...

- We geloven dat g-lieden niet zo slim zijn als wij.
  'we believe that you$_{strong}$ not so smart are as we'

  a) Ze gelooft dat gij/gie eerder thuis bent dan ik.
     'she believes that you earlier home are than I'
  b) Ik denk da Marie hem zal moeten roepen.
     'I think that Mary him will must call'
  c) U [niet-beleefdh] gelooft dat Lisa even mooi is als Anna.
     'you [non-honorific] believe that Lisa as beautiful is as Anna'
  d) Fons zag een slang naast hem.
     'Fons saw a snake next to him'
  e) Erik liet mij voor hem werken.
     'Erik let me for him work'
  f) De jongen wie/die z'n moeder gisteren hertrouwd is.
     'the boy who/that his mother yesterday remarried is'

Implicational chain of rules

1/4: d54a:after_v (Subject doubling 2 singular)
As gij gezond leeft, leef- **de** **gij** langer.
if *you*$_{sing}$ *healthily live, live-* *you*$_{sing,weak}$ *you*$_{sing,strong}$ *longer*

2/4: d55a:after_v (Subject doubling 2 plural)
As gulder gezond leeft, leef- **de** **gulder** langer.
if *you*$_{plural}$ *healthily live, live-* *you*$_{plural,weak}$ *you*$_{plural,strong}$ *longer*

3/4: p46a:g-lieden (Subject pronouns 2 plural, strong forms)
We geloven dat **g-lieden** niet zo slim zijn als wij.
we believe that *you*$_{plural,strong}$ *not so smart are as we.*

4/4: p38b:gij/gie (Subject pronouns 2 singular, strong forms)
Ze gelooft dat **gij / gie** eerder thuis bent dan ik.
she believes that *you*$_{singular,strong}$ *earlier home are than I*

A higher complexity rule

- "if either antecedent variable A1 or A2 occurs in a dialect, then syntactic variable C also occurs"

A1:	p46b:julle(n)/jullie (Subject pronouns 2 plural, strong forms, complex) We geloven dat **julle(n)/jullie** niet zo slim zijn als wij. *we believe that you$_{plural,strong}$ not so smart are as we.* 'We believe that you are not as smart as we are.'
A2:	p46b:julder/jielder (Subject pronouns 2 plural, strong forms, complex) We geloven dat **julder/jielder** niet zo slim zijn als wij.
C:	p46a:j-[lieden-compositum] (Subject pronouns 2 plural, strong forms) We geloven dat **j-lieden** niet zo slim zijn als wij.
Int:	The infrequent pronoun 'julder/jielder' perfects the implicational association of the frequent 'julle(n)/jullie' variant with the pronoun 'j-lieden'.

Conclusions

1. Dialectometric methods can be successfully applied to syntactic data and the results clearly show geographically coherent patterns
2. There are significant associations among the syntactic, pronunciational and lexical levels, but geographic distance plays a very important role as an underlying structuring factor
3. Association rule mining based on proportional overlap can contribute to the identification, exploration and validation of associations between syntactic variables