Measuring syntactic variation in Dutch dialects

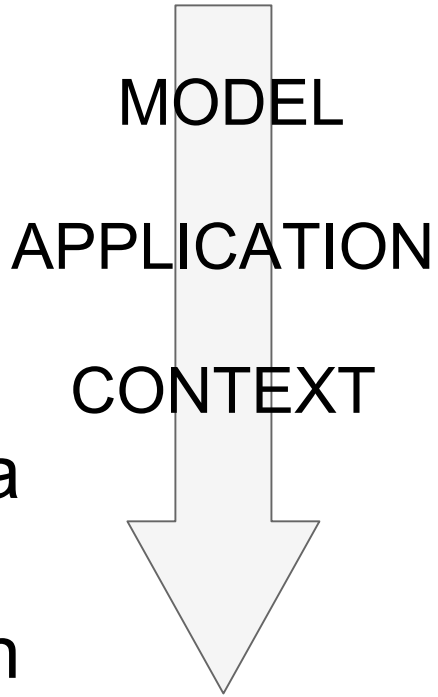
Marco René Spruit
Meertens Instituut, Amsterdam, The Netherlands
http://www.meertens.knaw.nl/medewerkers/marco.rene.spruit

the *Determinants of Dialectal Variation* project
http://dialectometry.net

Methods XII workshop *Progress in Dialectometry: Toward Explanation,* August 5th, 2005.

Research questions

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



Syntactic Atlas of the Dutch Dialects

SAND1:

[Barbiers et al. 2005]

- **267** Dutch dialects
- **134** unique geographic distributions / syntactic features
- **507** variant distributions

Domains:

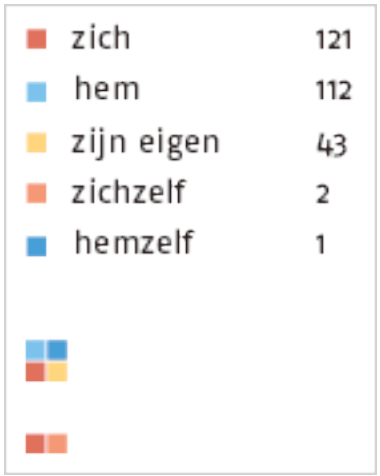
1. Complementisers
2. Subject pronouns
3. Expletives
4. Subject doubling
5. Subject clitisation following yes/no
6. Reflexive and reciprocal pronouns
7. Fronting

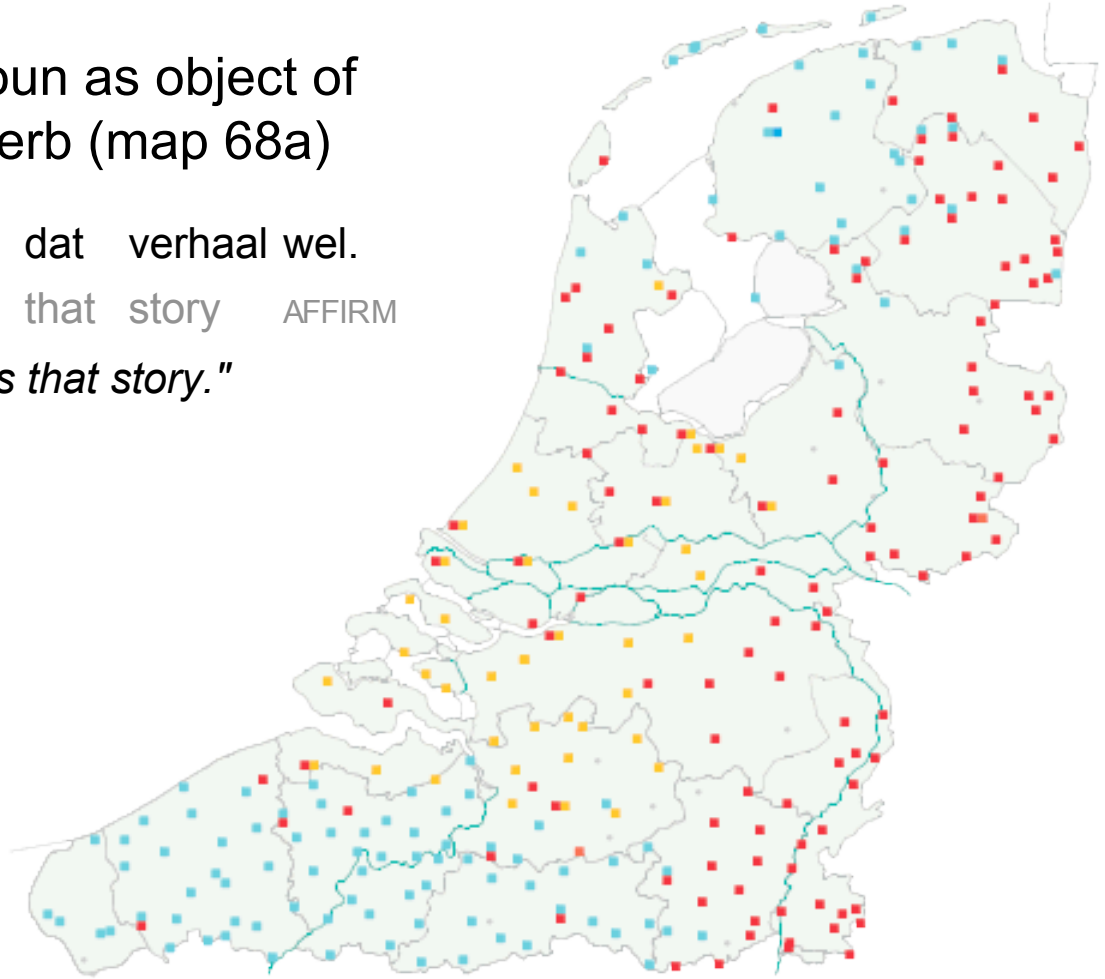
SAND1 data sample

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."





Variant-based measurement

- Hamming distance algorithm -> based on binary comparisons between feature variants
- For example: *Weak reflexive pronoun* as [...]:

variants	*Lunteren*	*Veldhoven*	*distance*
zich	+	+	0
hem	-	-	0
zijn eigen	+	-	1
zichzelf	-	-	0
hemzelf	-	-	0
			1

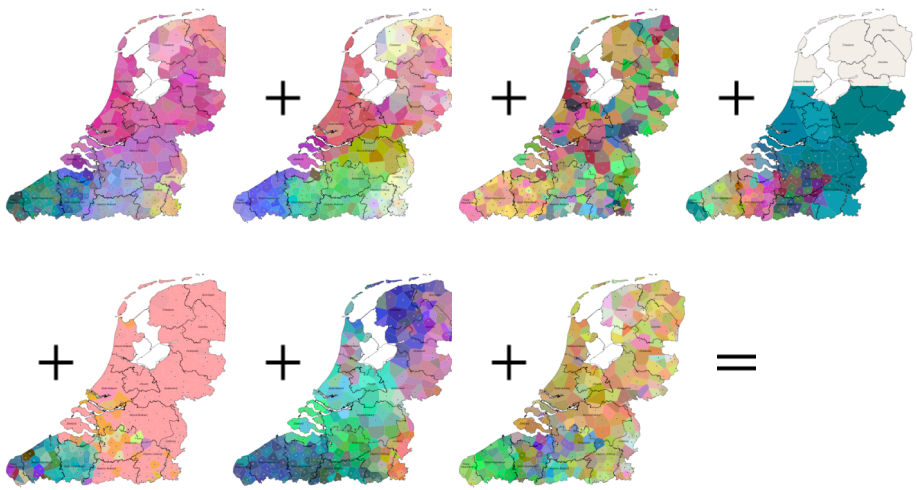
Distance matrix

dialect	Lunteren	Bellingwolde	Hollum	Doel	Sint-Truiden	Veldhoven
Lunteren		66	52	122	77	**47**
Bellingwolde	66		56	134	81	51
Hollum	52	56		116	63	59
Doel	122	134	116		115	111
Sint-Truiden	77	81	63	115		72
Veldhoven	**47**	51	59	111	72	

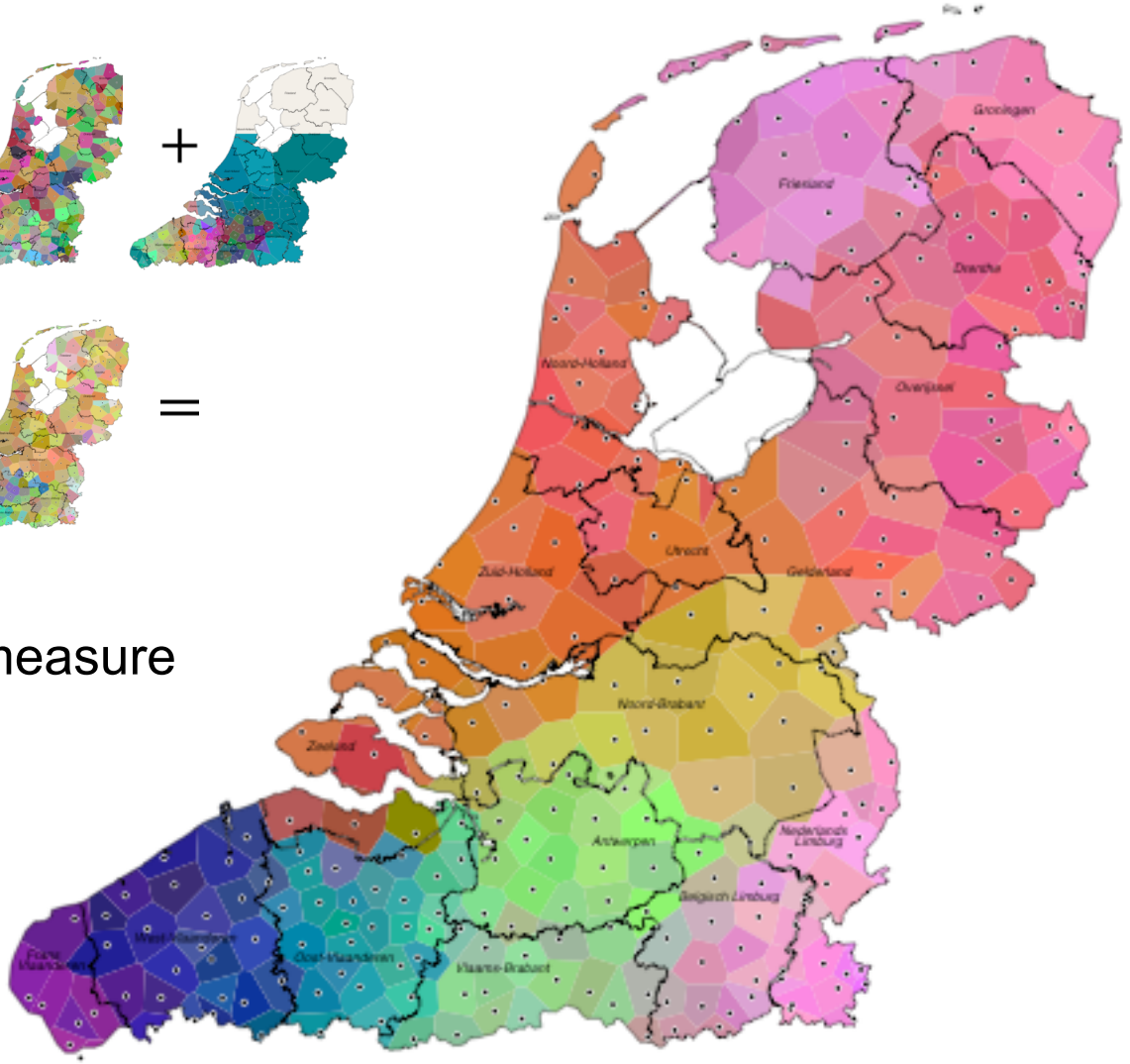
Analysis & Visualisation

- Multidimensional scaling (Classical MDS)
  - Goal: to optimally represent the most differentiating feature variants for each dialect in relation to all other dialects
- MDS dialect colour maps
  - Each dialect has its own unique composite colour
  - Colour contrasts represent linguistic differences
    - a perfect correlation results in a colour continuum
    - a low correlation results in a mosaic-like map

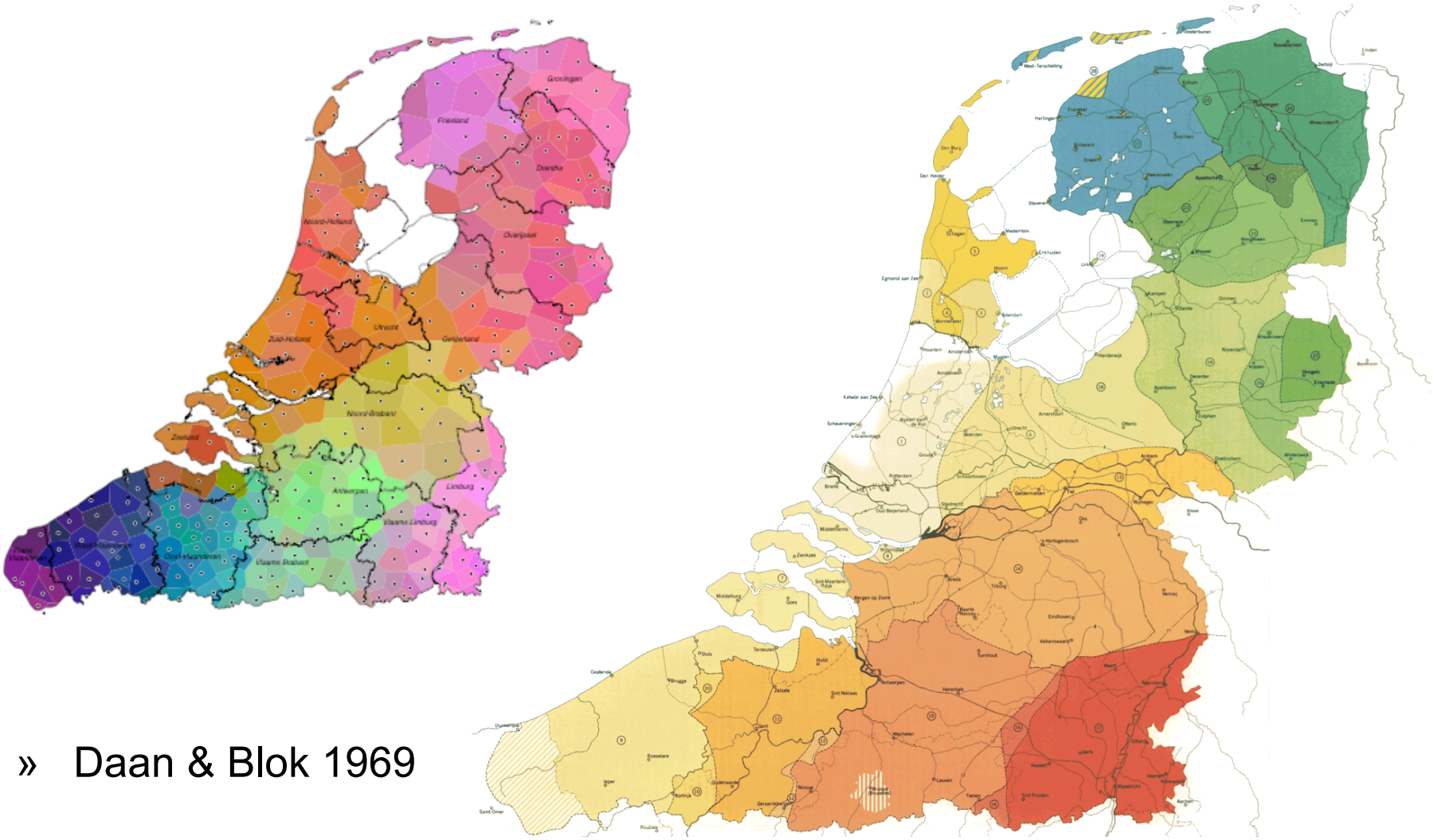
SAND1 MDS map



- Hamming distance measure
- 267 dialect locations
- 507 feature variants (in 134 features)
- r = 0.959

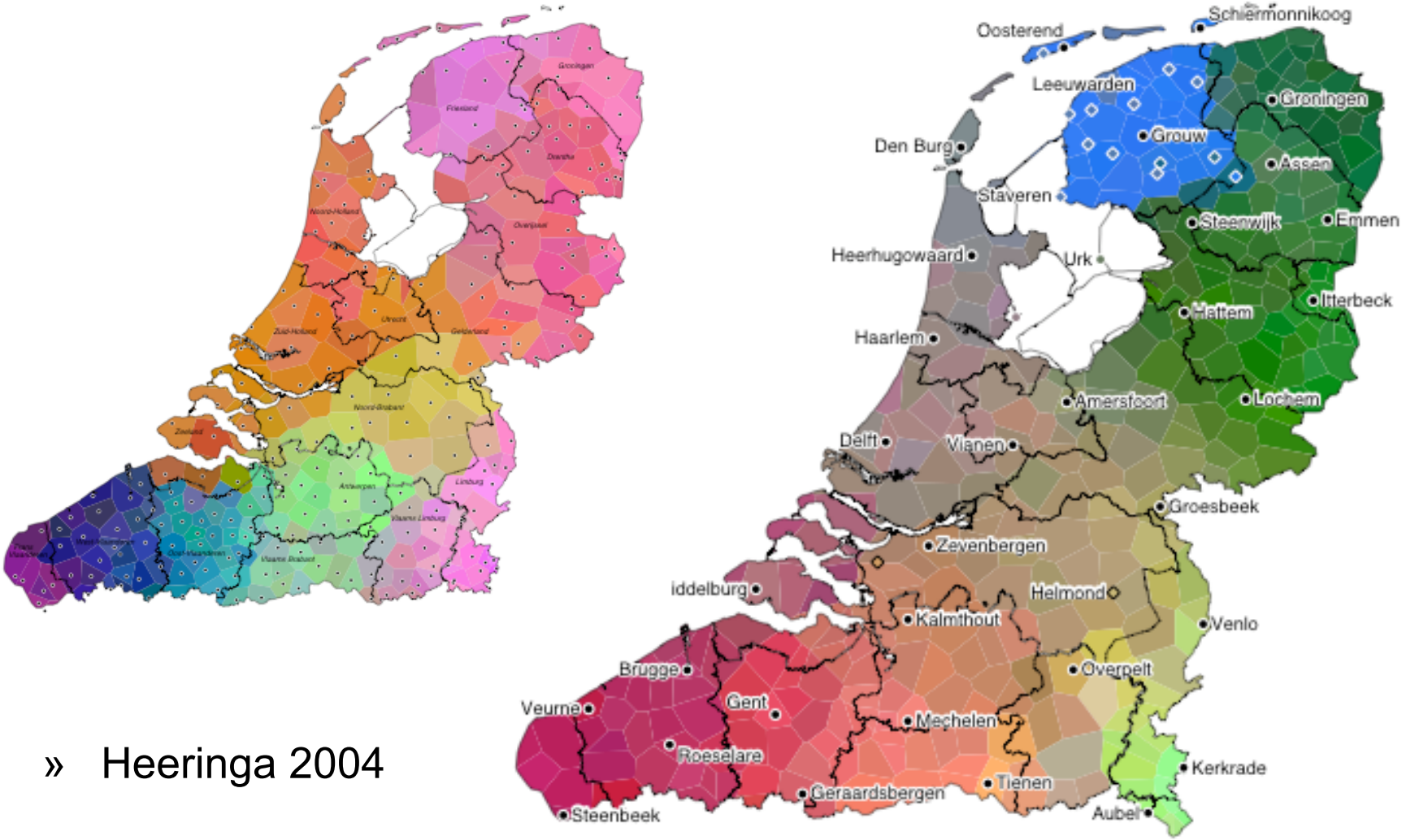


Syntax versus perception



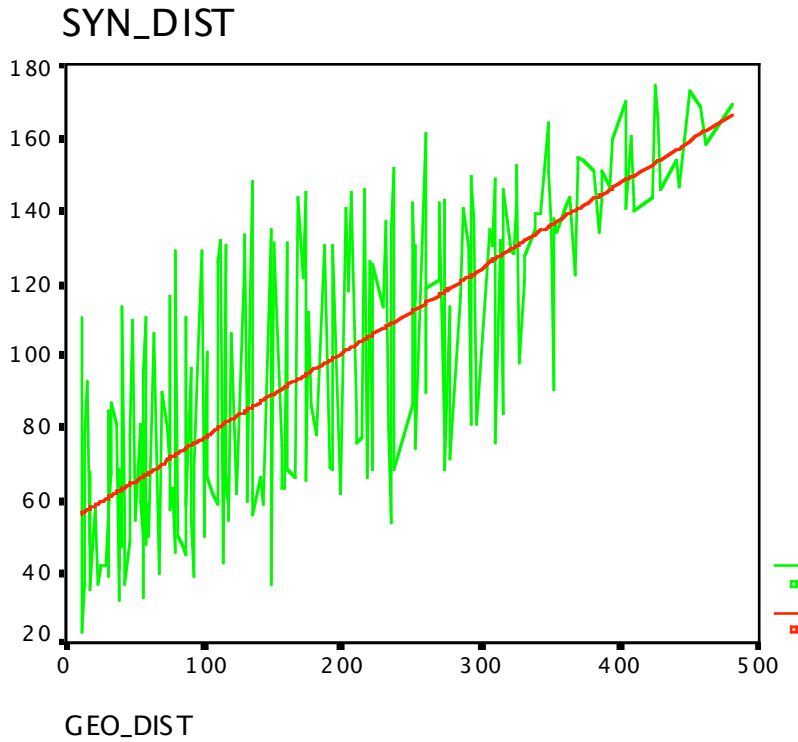
» Daan & Blok 1969

Syntax versus pronunciation



» Heeringa 2004

Syntax versus geography



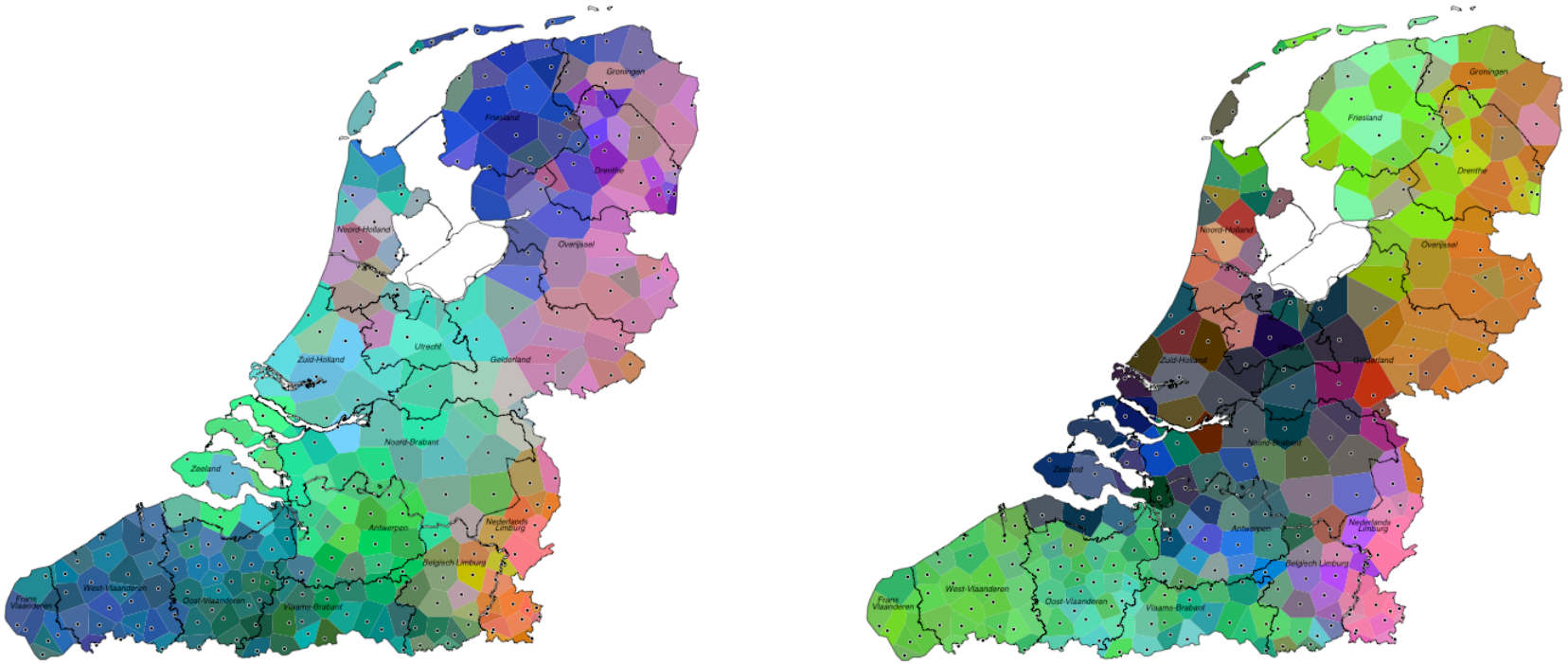
- r = 0.749
- Standard error = 26.657

Attribute-based measurement

- Binary comparisons between variant attributes
- *For example:* Attributes in the Reflexives domain:

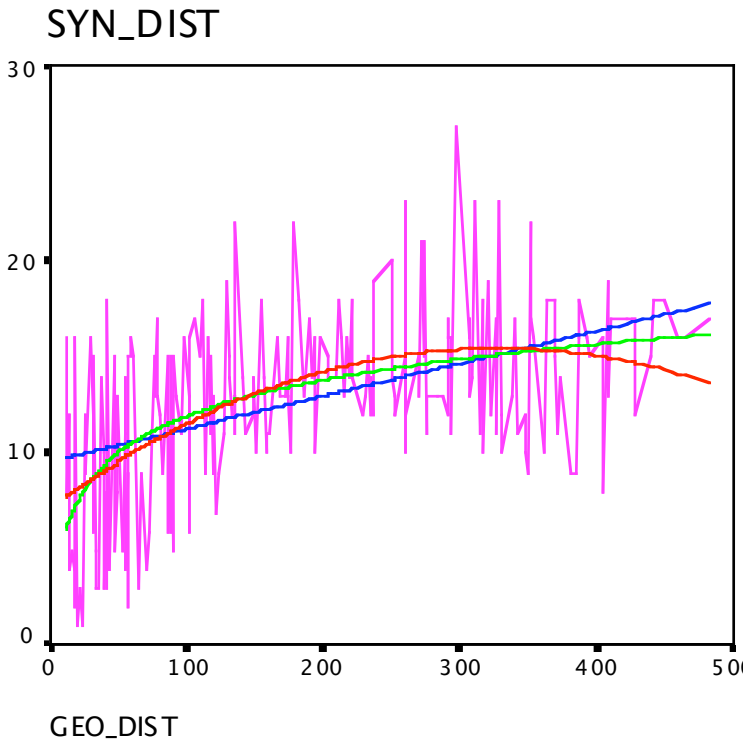
	personal *"hem"*	reflexive *"zich"*	possessive *"zijn"*	ownness *"eigen"*	focus *"zelf"*
hem	+				
hemzelf	+				+
zich		+			
zichzelf		+			+
zijn			+		
zijn zelf			+		+
zijn eigen			+	+	
zijn eigen zelf			+	+	+

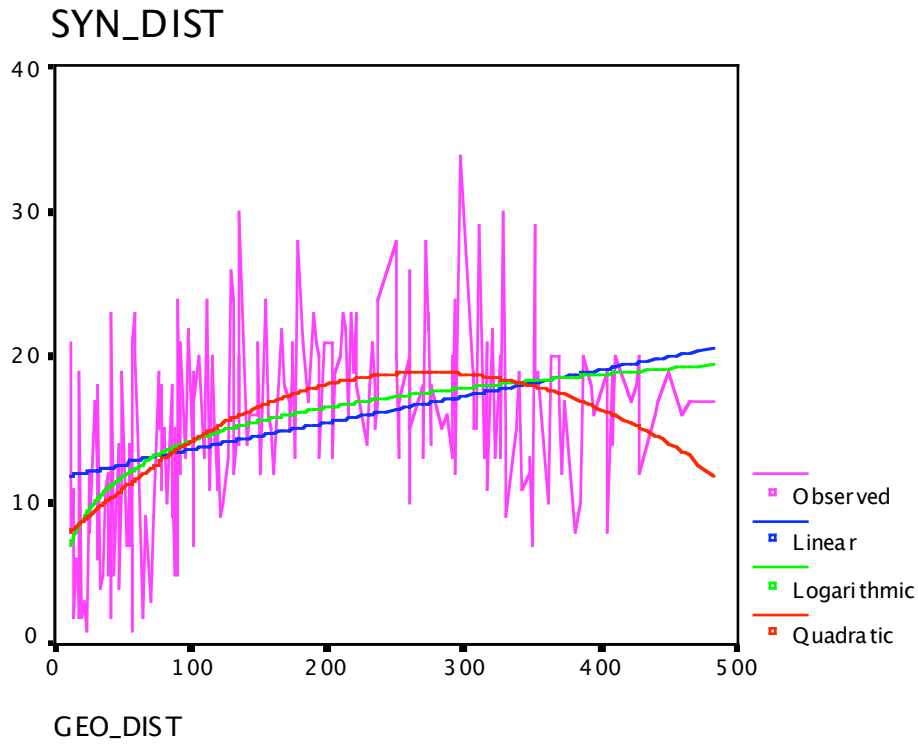
Variants versus attributes: Visualisation



- MDS maps of the Reflexives domain (266 sites)
- *75 variants:*
  r = 0.93438533
- *61 attributes:*
  r = 0.94461615

Variants versus attributes: Analysis





Variant-based:

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

Attribute-based:

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

Conclusions

1. Dialectometry works well on purely syntactic data
2. Syntactic variation coheres geographically
3. High correspondence between syntactic and perceptual dialect maps
4. Good subjective overlap with dialect map based on pronunciation (more precise analysis soon)
5. High correspondence between variant-based and attribute-based measurements with respect to Reflexives

Future research

- More advanced measurements
  - Attribute-based measurements
    - Continuing data annotation
    - Incorporating similarity-difference distance coefficients
  - Experimentation with variant frequency (G.I.W.)
- Aggregating SAND2 data
- Analysis of feature dependencies between syntactic phenomena (P.C.A.)
- Analysis of correlation between linguistic levels