Discovery of association rules between syntactic variables

Seminar in Methodology and Statistics, Groningen, 23 May 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?”*

Syntactic variation & dialectometry

- Language variation dimensions

  { Macro, **Micro** }

  { Pronunciation, Lexis, Morphology, **Syntax** }

  { External, **Internal** }

  { Time, **Space** }

  { Qualitative, **Quantitative** }

- Research questions

  i. How can relevant associations between syntactic variables be discovered?

  ii. What are interesting associations between syntactic variables?

The big picture

- Generative syntax and functional typology share a primary interest in understanding the structural similarities and differences between language varieties
  - Ultimate goal: to characterise the superficial structural diversity of all language varieties as particular settings of relatively few parametric patterns
- This contribution: A computational method to automatically discover syntactic variable associations

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

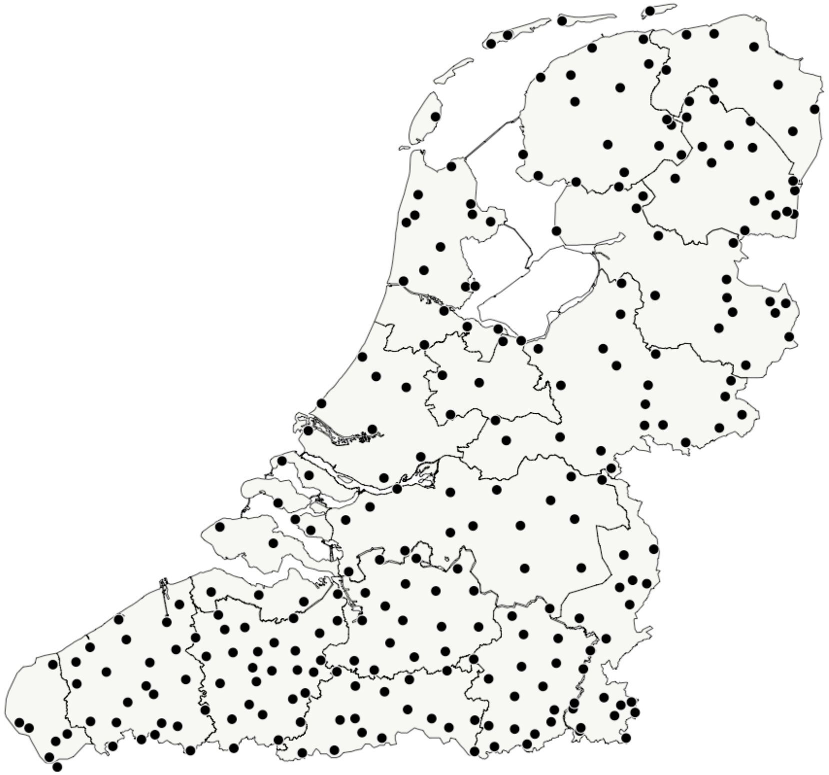
    - 106 syntactic contexts, 485 variables
  - SAND2: [Barbiers et al. 2007]

    Verbal clusters, Cluster interruption, Morphosyntactic variation, Negative particle, Negative concord and quantification

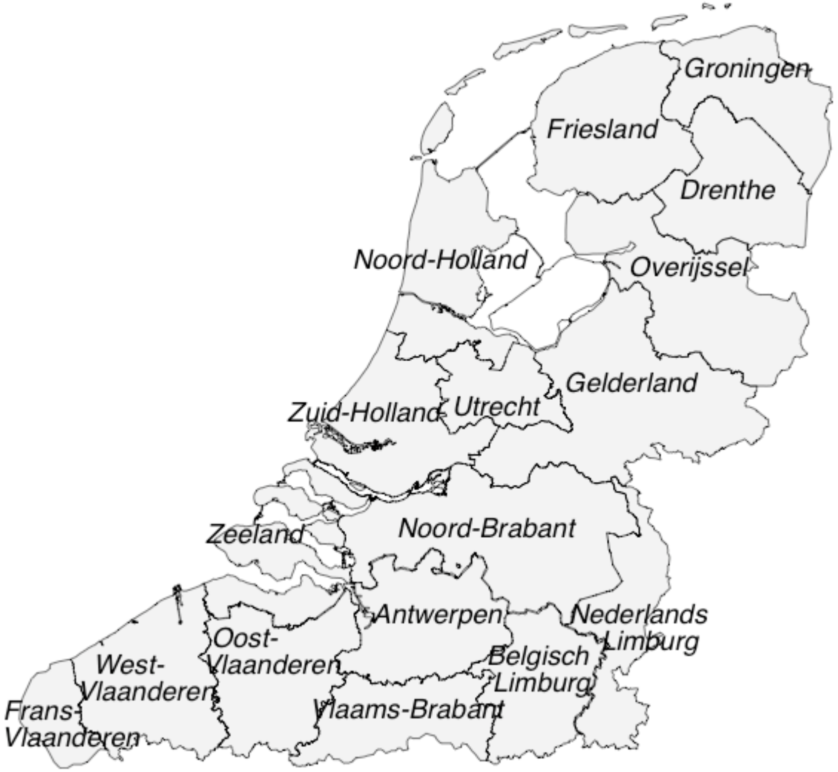
    - 65 syntactic contexts, 274 variables *(incomplete)*

Dutch language area

- Distribution of the 267 Dutch dialects in the SAND



- The provinces in the Dutch language area



“‘t lijkt wel ___ er iemand in de tuin staat.”

it looks AFFIRM ___ *there someone in the garden stands*

1. “Et lijk wel ofter een in den hof staat”

2. “Tis zo precies dater iemand in den hof staat”

3. “T lijk wel of datr iemand in den hof staat”

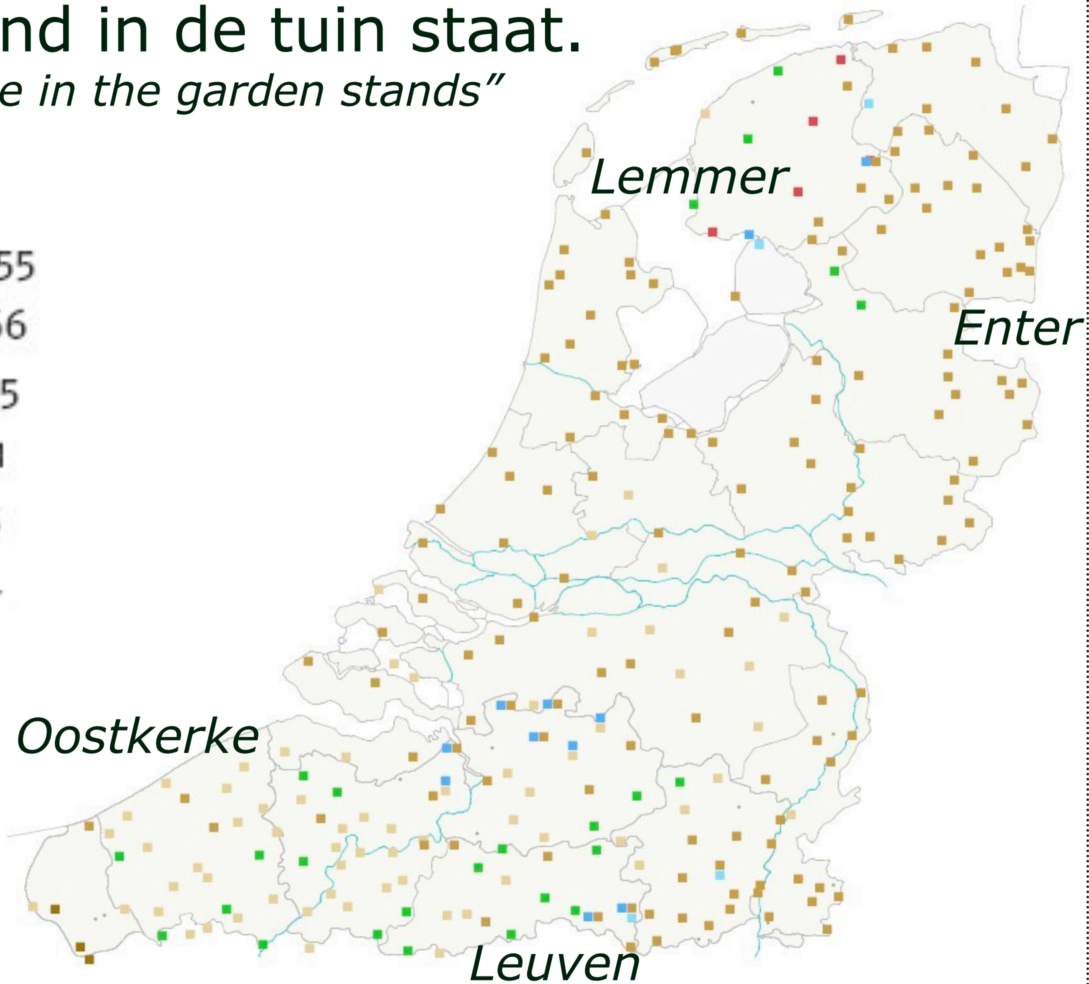
4. “It lijket wel as staat der een in de tuin”

SAND1 map 14b

‘t lijkt wel **of** er iemand in de tuin staat.

“it looks AFFIRM ***if*** *there someone in the garden stands”*

🟧	of	155
🟨	of dat	66
🟩	dat	25
🟦	as / of + ingebedde V2	11
🟥	at	5
🟦	as	4
🟫	et	3



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$_{weak}$ **you**$_{strong}$ healthily live, live you$_{weak}$ **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”*

Syntactic context & variables

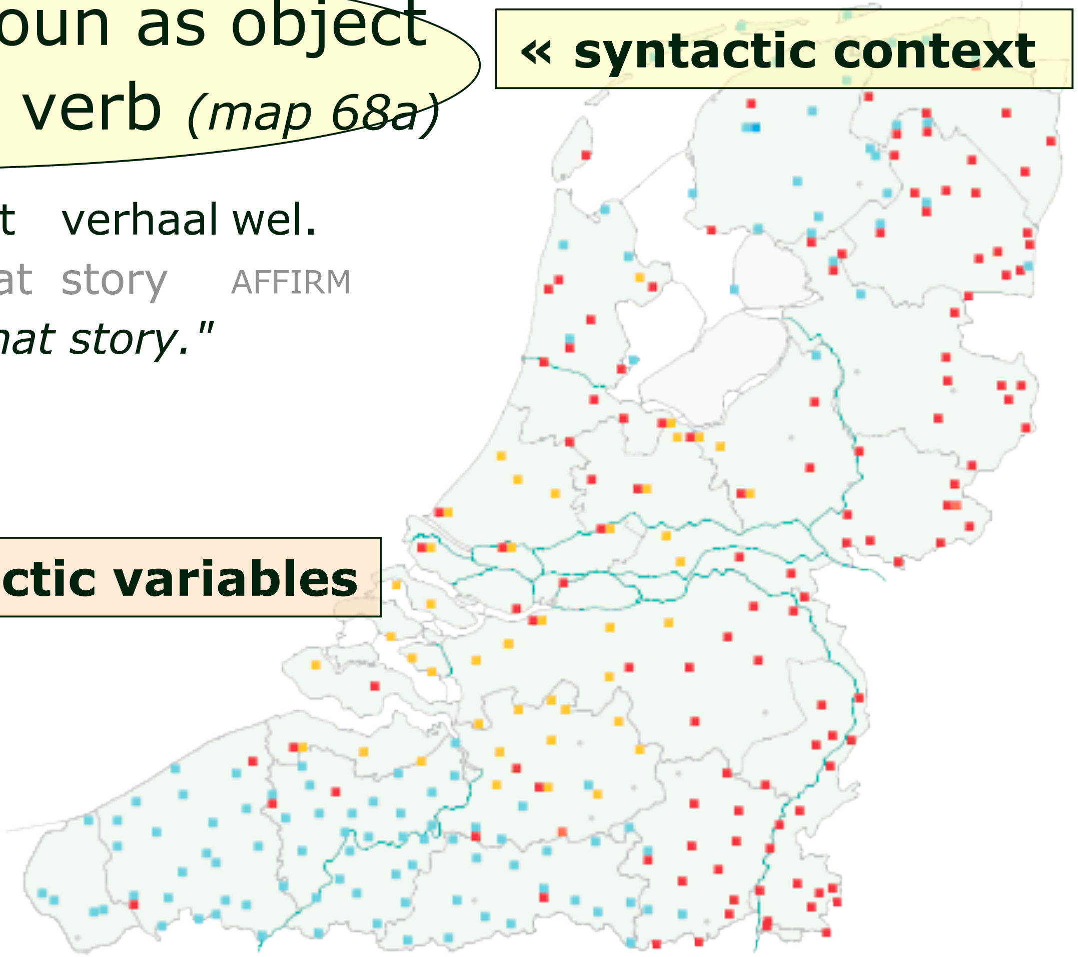
Weak reflexive pronoun as object of inherent reflexive verb *(map 68a)*

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

"John certainly remembers that story."

🟥	zich	121
🟦	hem	112
🟨	zijn eigen	43
🟧	zichzelf	2
🟦	hemzelf	1
🟦🟦 🟥🟨		
🟥🟧		

« syntactic variables



« syntactic context

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 → 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$_{\text{weak}}$	*believe*	*you*$_{\text{strong}}$	*certainly*	*not*	*that*	*he*	*stronger*	*is*	*than*	*you*$_{\text{weak}}$	*you*$_{\text{strong}}$

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

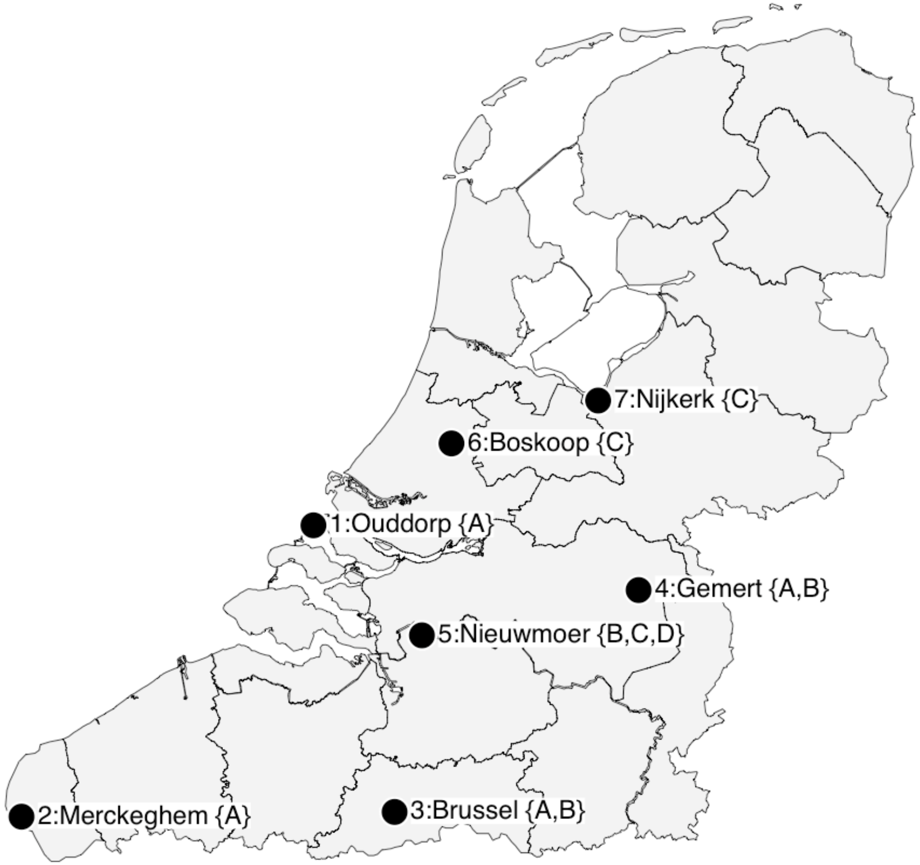
Jan	herinnert	**zijn**	**eigen**	dat	verhaal	wel.
John	*remembers*	*his*	*own*	*that*	*story*	*[affirmative]*

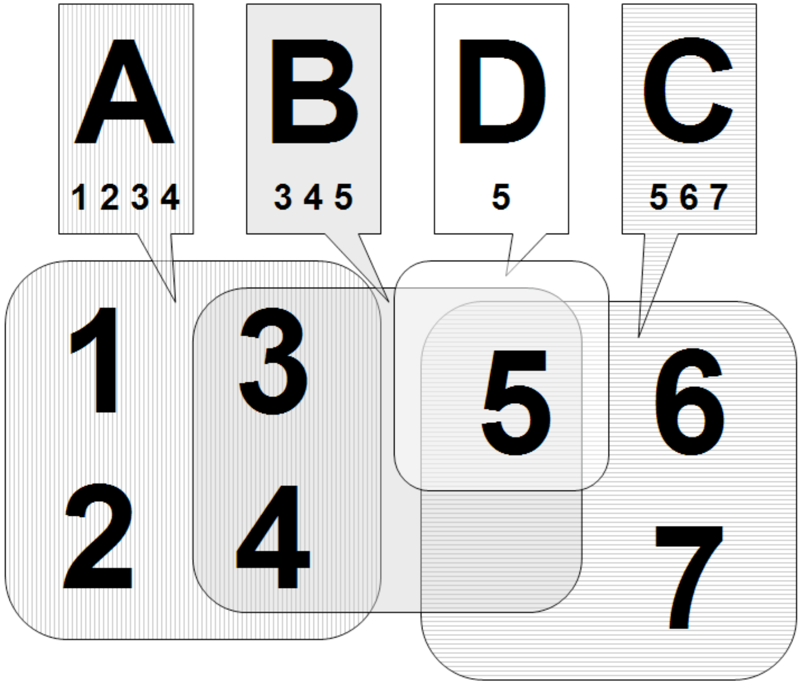
D. “Short subject relative, complementiser following relative pronoun” (84a)

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 → varB: (A ∩ B / A) * 100 = 2/4 * 100 = 50%
- **Coverage**: |A|

  How often does the rule apply?
  - varA → varB: A / N * 100 = 4/7 * 100 = 57%
- **Completeness**: |A&C| / |C|

  How much of the target class does the rule cover?
  - varA → varB: (A ∩ B / B) * 100 = 2/3 * 100 = 66%
- **Interestingness**: |A&C| - |A||C|/N

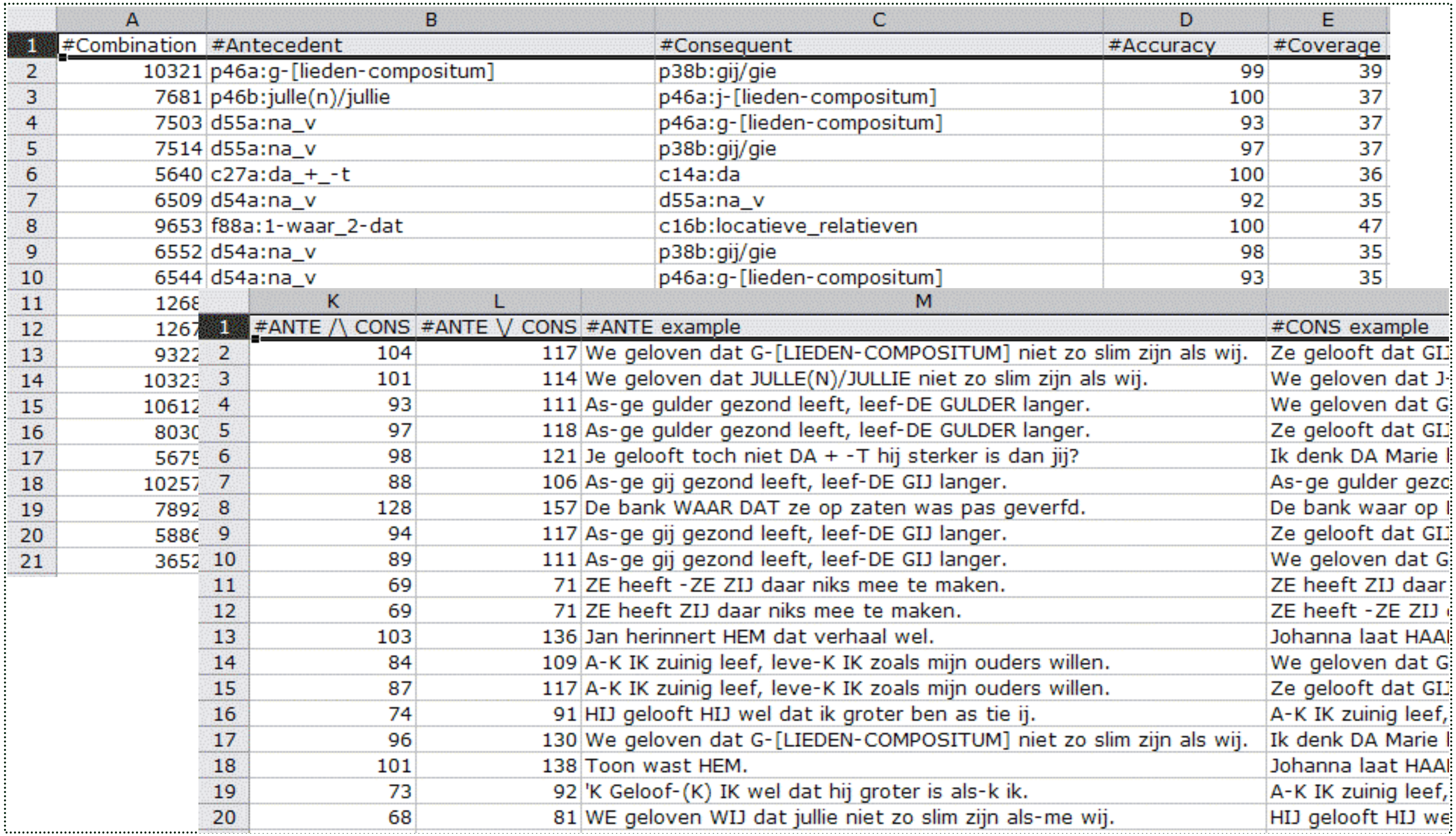
  Integrates the three factors above into one value...
  - varA → varB: (A ∩ B) - (A * B / N) = 2 – (4 * 3 / 7) = 0.28

Sample data results

The 8 highest ranked association rules:

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

Interactive exploration...



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				

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
7	88	106	As-ge gij gezond leeft, leef-DE GIJ langer.	As-ge gulder gezo
8	128	157	De bank WAAR DAT ze op zaten was pas geverfd.	De bank waar op
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
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

- 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*$_{plu ral,strong}$	*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 $\textit{you}_{\textit{plural,strong}}$ *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’.

Some conclusions

1. Association rule mining technique based on proportional overlap: *it works*.
   - Facilitates identification, validation and exploration of variable relationships
2. Reveals the existence of many potentially interesting associations within SAND1
3. Shows considerable overlaps between the geographical distributions of syntactic variable pairs
4. Results strongly indicate that many more potentially interesting associations between syntactic variables are likely to be uncovered

Discussion & future research

- Incorporate exception rules
- Alternative measures of interestingness / incorporation of additional rule quality evaluation factors (surprisingness, ...)
- Adding more data (SAND2)
  - Phonological data: discover potential associations between variables *among linguistic levels*
- Refine dialect area detection
- Comparison with methods such as Cramér's V and correspondence analysis