Reverse dialectometry

Geography as a probe into linguistic theory

Jeroen van Craenenbroeck

KU Leuven/CRISSP

Maps and Grammar

September 17–18, 2014

Introduction: verbs, word order, and linguistic theory

Introduction: verbs, word order, and linguistic theory

- in Dutch (like in many Germanic languages) verbs tend to group together at the right edge of the (embedded) clause:

Introduction: verbs, word order, and linguistic theory

- in Dutch (like in many Germanic languages) verbs tend to group together at the right edge of the (embedded) clause:

(1) dat hij gisteren tijdens de les gelachen heeft.
that he yesterday during the class laughed has
'that he laughed yesterday during class.'

Introduction: verbs, word order, and linguistic theory

- in Dutch (like in many Germanic languages) verbs tend to group together at the right edge of the (embedded) clause:

(1) dat hij gisteren tijdens de les gelachen heeft.
that he yesterday during the class laughed has
'that he laughed yesterday during class.'

- moreover, such verbal clusters typically show a certain degree of freedom in their word order:

Introduction: verbs, word order, and linguistic theory

- in Dutch (like in many Germanic languages) verbs tend to group together at the right edge of the (embedded) clause:

(1) dat hij gisteren tijdens de les gelachen heeft.
that he yesterday during the class laughed has
'that he laughed yesterday during class.'

- moreover, such verbal clusters typically show a certain degree of freedom in their word order:

(2) dat hij gisteren tijdens de les heeft gelachen.
that he yesterday during the class had laughed
'that he laughed yesterday during class.'

Introduction: verbs, word order, and linguistic theory

- in Dutch (like in many Germanic languages) verbs tend to group together at the right edge of the (embedded) clause:

(1) dat hij gisteren tijdens de les gelachen heeft.
that he yesterday during the class laughed has
'that he laughed yesterday during class.' (21)

- moreover, such verbal clusters typically show a certain degree of freedom in their word order:

(2) dat hij gisteren tijdens de les heeft gelachen.
that he yesterday during the class had laughed
'that he laughed yesterday during class.' (12)

- this word order freedom is typically a source of interdialectal variation:

- this word order freedom is typically a source of interdialectal variation:

(3) **Ferwerd Dutch**

a. dasto it ook net zien meist.
that.you it also not see may
'that you're also not allowed to see it.' (✓21)

b. *dasto it ook net meist zien.
that.you it also not may see
'that you're also not allowed to see it.' (*12)

- this word order freedom is typically a source of interdialectal variation:

(4) **Gendringen Dutch**

a. dat ee et ook nie zien mag.
that you it also not see may
'that you're also not allowed to see it.' (✓21)

b. dat ee et ook nie mag zien.
that you it also not may see
'that you're also not allowed to see it.' (✓12)

- this word order freedom is typically a source of interdialectal variation:

(5) **Poelkapelle Dutch**

a. *dajtgie ook nie zien meug.
that.it.you also not see may
'that you're also not allowed to see it.' (*21)

b. dajtgie ook nie meug zien.
that.it.you also not may see
'that you're also not allowed to see it.' (✓12)

- and the more complex the verbal cluster, the more variation there is: in verbal clusters consisting of two modal auxiliaries and one main verb, out of the six orders that are theoretically possible, four are attested in Dutch dialects:

- and the more complex the verbal cluster, the more variation there is: in verbal clusters consisting of two modal auxiliaries and one main verb, out of the six orders that are theoretically possible, four are attested in Dutch dialects:

(6) Ik vind dat iedereen moet kunnen zwemmen.
I find that everyone must can swim
'I think everyone should be able to swim.' (123)

- and the more complex the verbal cluster, the more variation there is: in verbal clusters consisting of two modal auxiliaries and one main verb, out of the six orders that are theoretically possible, four are attested in Dutch dialects:

(6) Ik vind dat iedereen moet kunnen zwemmen.
I find that everyone must can swim
'I think everyone should be able to swim.' (123)

(7) a. Ik vind dat iedereen moet zwemmen kunnen. (132)
b. Ik vind dat iedereen zwemmen moet kunnen. (312)
c. Ik vind dat iedereen zwemmen kunnen moet. (321)
d. *Ik vind dat iedereen kunnen zwemmen moet. (231)
e. *Ik vind dat iedereen kunnen moet zwemmen. (213)

- but once again, it is not the case that each of the four allowed orders is attested in all dialects:

- but once again, it is not the case that each of the four allowed orders is attested in all dialects:

(8) *Midsland Dutch*

a. *dat elkeen mot kanne zwemme.
that everyone must can swim
'that everyone should be able to swim.' (*123)

b. dat elkeen mot zwemme kanne. (✓132)

c. *dat elkeen zwemme mot kanne. (*312)

d. dat elkeen zwemme kanne mot. (✓321)

e. *dat elkeen kanne zwemme mot. (*231)

f. *dat elkeen kanne mot zwemme. (*213)

- but once again, it is not the case that each of the four allowed orders is attested in all dialects:

(9) *Langelo Dutch*

a. dat iedereen moet kunnen zwemmen.
that everyone must can swim
'that everyone should be able to swim.' (✓123)

b. *dat iedereen mot zwemmen kunnen. (*132)

c. dat iedereen zwemmen mot kunnen. (✓312)

d. *dat iedereen zwemmen kunnen mot. (*321)

e. *dat iedereen kunnen zwemmen mot. (*231)

f. *dat iedereen kunnen mot zwemmen. (*213)

- more generally, the four possible cluster orders yield a total of 16 possible combinations, of which 12 are attested in Dutch dialects:

- more generally, the four possible cluster orders yield a total of 16 possible combinations, of which 12 are attested in Dutch dialects:

example dialect	**123**	**132**	**321**	**312**
Beetgum	✓	✓	✓	✓
Hippolytushoef	✓	✓	✓	*
Warffum	✓	✓	*	*
Oosterend	✓	*	*	*
Schermerhorn	✓	✓	*	✓
Visvliet	✓	*	✓	✓
Kollum	✓	*	✓	*
Langelo	✓	*	*	✓
Midsland	*	✓	✓	*
Lies	*	*	✓	*
Bakkeveen	*	*	✓	✓
Waskemeer	*	✓	*	*

- in order to get a more complete picture of the variation, we can look at the results from the SAND-project:

- in order to get a more complete picture of the variation, we can look at the results from the SAND-project:
  - Syntactic Atlas of the Dutch Dialects (2000–2004)

- in order to get a more complete picture of the variation, we can look at the results from the SAND-project:
  - Syntactic Atlas of the Dutch Dialects (2000–2004)
  - dialect interviews in 267 dialect locations in Belgium, France, and the Netherlands

- in order to get a more complete picture of the variation, we can look at the results from the SAND-project:
  - Syntactic Atlas of the Dutch Dialects (2000–2004)
  - dialect interviews in 267 dialect locations in Belgium, France, and the Netherlands
- the SAND-questionnaire contained eight questions on word order in verb clusters for a total of 31 cluster orders

- in order to get a more complete picture of the variation, we can look at the results from the SAND-project:
  - Syntactic Atlas of the Dutch Dialects (2000–2004)
  - dialect interviews in 267 dialect locations in Belgium, France, and the Netherlands
- the SAND-questionnaire contained eight questions on word order in verb clusters for a total of 31 cluster orders
- if we map, for each of the 267 SAND-dialects, which dialect has which combination of cluster orders, we find 137 different combinations of verb cluster orders

- in order to get a more complete picture of the variation, we can look at the results from the SAND-project:
  - Syntactic Atlas of the Dutch Dialects (2000–2004)
  - dialect interviews in 267 dialect locations in Belgium, France, and the Netherlands
- the SAND-questionnaire contained eight questions on word order in verb clusters for a total of 31 cluster orders
- if we map, for each of the 267 SAND-dialects, which dialect has which combination of cluster orders, we find 137 different combinations of verb cluster orders
- in other words, there are 137 different types of dialects when it comes to word order in verbal clusters

- **question:** how can we make sense of this massive variation from the point of view of theoretical linguistics?

- **question:** how can we make sense of this massive variation from the point of view of theoretical linguistics?
- e.g. Principles & Parameters: natural language is the result of the interplay between:

- **question:** how can we make sense of this massive variation from the point of view of theoretical linguistics?
- e.g. Principles & Parameters: natural language is the result of the interplay between:
  1. Principles: innate properties that are invariant across all languages

- **question:** how can we make sense of this massive variation from the point of view of theoretical linguistics?
- e.g. Principles & Parameters: natural language is the result of the interplay between:
  1. Principles: innate properties that are invariant across all languages
  2. Parameters: simple, often binary choices ('switches') which are responsible for interlinguistic differences, and which determine the space of variation in natural language

- **question:** how can we make sense of this massive variation from the point of view of theoretical linguistics?
- e.g. Principles & Parameters: natural language is the result of the interplay between:
  1. Principles: innate properties that are invariant across all languages
  2. Parameters: simple, often binary choices ('switches') which are responsible for interlinguistic differences, and which determine the space of variation in natural language
- **so:**

- **question:** how can we make sense of this massive variation from the point of view of theoretical linguistics?
- e.g. Principles & Parameters: natural language is the result of the interplay between:
  1. Principles: innate properties that are invariant across all languages
  2. Parameters: simple, often binary choices ('switches') which are responsible for interlinguistic differences, and which determine the space of variation in natural language
- **so:**
  - what are the parameters of word order variation in verb clusters?

- **question:** how can we make sense of this massive variation from the point of view of theoretical linguistics?
- e.g. Principles & Parameters: natural language is the result of the interplay between:
  1. Principles: innate properties that are invariant across all languages
  2. Parameters: simple, often binary choices ('switches') which are responsible for interlinguistic differences, and which determine the space of variation in natural language
- **so:**
  - what are the parameters of word order variation in verb clusters?
  - is this variation even parameter-related? how much noise is there in these data? is some of the variation extra-grammatical (cf. Barbiers (2005))?

- **question:** how can we make sense of this massive variation from the point of view of theoretical linguistics?
- e.g. Principles & Parameters: natural language is the result of the interplay between:
  1. Principles: innate properties that are invariant across all languages
  2. Parameters: simple, often binary choices ('switches') which are responsible for interlinguistic differences, and which determine the space of variation in natural language
- **so:**
  - what are the parameters of word order variation in verb clusters?
  - is this variation even parameter-related? how much noise is there in these data? is some of the variation extra-grammatical (cf. Barbiers (2005))?
  - related methodological question: how do we go about finding those parameters?

- **in this talk** I argue that a quantitative-statistical analysis of the data enriched with insights from formal-theoretical linguistics can separate the wheat from the chaff

- **in this talk** I argue that a quantitative-statistical analysis of the data enriched with insights from formal-theoretical linguistics can separate the wheat from the chaff
- more specifically, I will argue that roughly 80% of the variation found in Dutch verb cluster orders can be reduced to three grammatical parameters

Dialect variation and quantitative methods: dialectometry

Dialect variation and quantitative methods: dialectometry

- **dialectometry** is a subdiscipline of linguistics that uses computational and quantitative techniques in dialectology (Nerbonne and Kretzschmar Jr., 2013)

Dialect variation and quantitative methods: dialectometry

- **dialectometry** is a subdiscipline of linguistics that uses computational and quantitative techniques in dialectology (Nerbonne and Kretzschmar Jr., 2013)
- in a typical dialectometric analysis locations are used as individuals and linguistic phenomena as variables $\rightarrow$ we're measuring similarities and differences between dialect locations based on their linguistic profile

Dialect variation and quantitative methods: dialectometry

- **dialectometry** is a subdiscipline of linguistics that uses computational and quantitative techniques in dialectology (Nerbonne and Kretzschmar Jr., 2013)
- in a typical dialectometric analysis locations are used as individuals and linguistic phenomena as variables $\rightarrow$ we're measuring similarities and differences between dialect locations based on their linguistic profile
- often used method: Multidimensional Scaling (MDS)

Dialect variation and quantitative methods: dialectometry

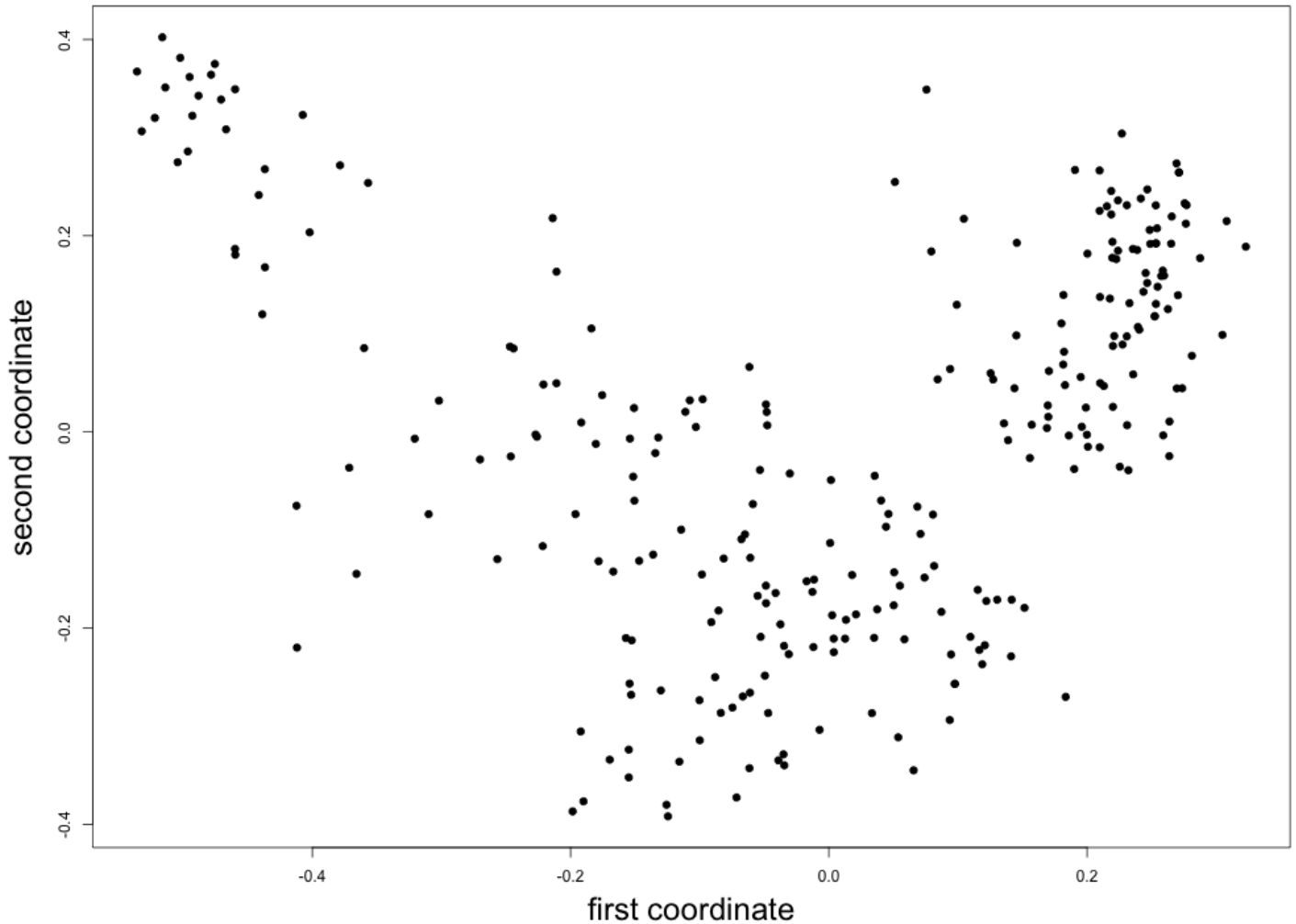
- **dialectometry** is a subdiscipline of linguistics that uses computational and quantitative techniques in dialectology (Nerbonne and Kretzschmar Jr., 2013)
- in a typical dialectometric analysis locations are used as individuals and linguistic phenomena as variables $\rightarrow$ we're measuring similarities and differences between dialect locations based on their linguistic profile
- often used method: Multidimensional Scaling (MDS)
- starting point: data table with dialects in rows and cluster orders in columns

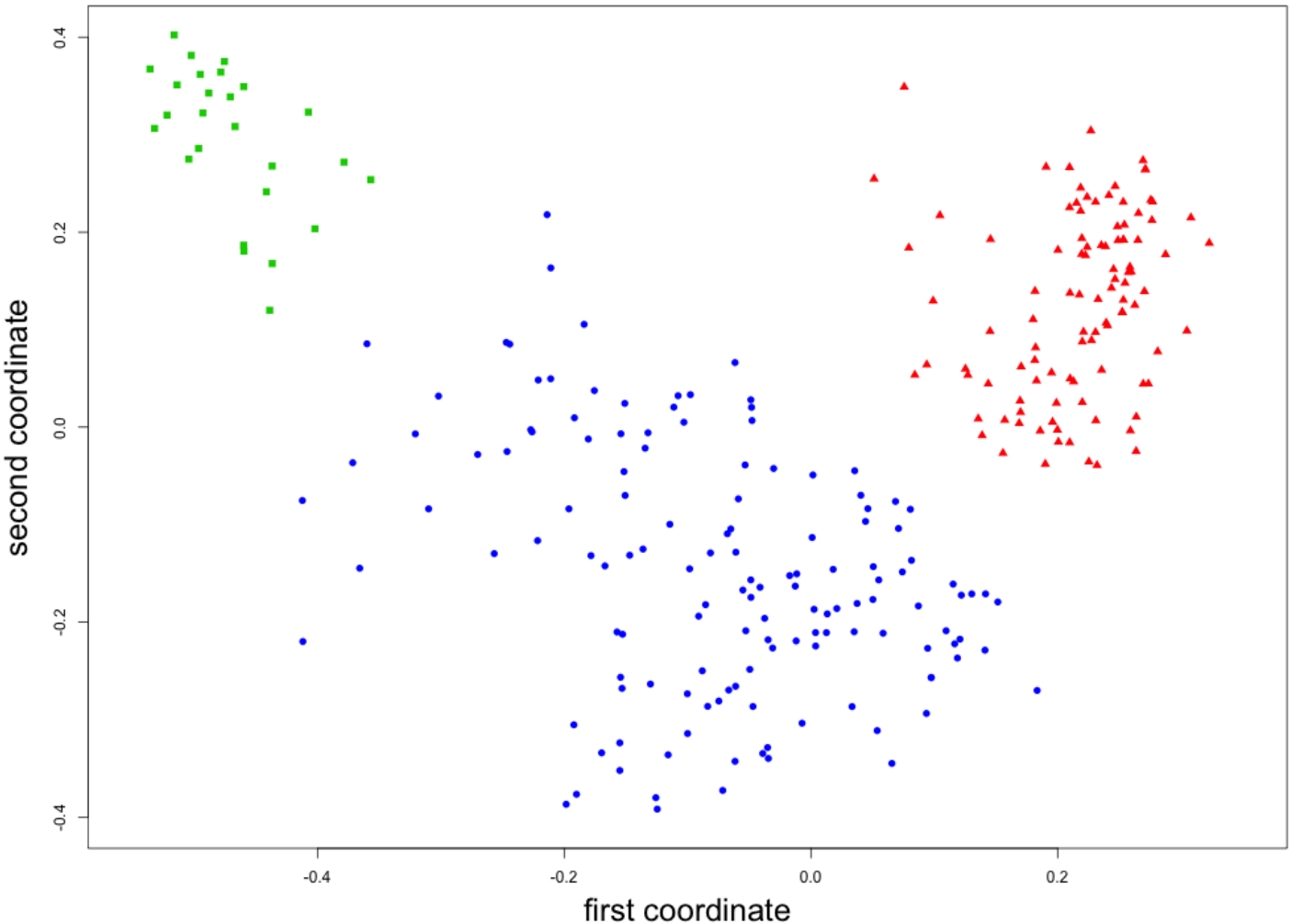
	AUX1(be.sg)-PART2	PART2-AUX1(be.sg)	AUX1(have.sg)-PART2	PART2-AUX1(have.sg)	AUX1(have.pl)-PAR
Midsland / Midslân	no	yes	no	yes	
Lies	no	yes	no	yes	
West-Terschelling	no	yes	no	yes	
Oosterend	NA	NA	no	yes	
Hollum	no	yes	NA	NA	
Schiermonnikoog	no	yes	no	yes	
Ferwerd / Ferwert	no	yes	no	yes	
Anjum / Eanjum	no	yes	no	yes	
Kollum	no	yes	no	yes	
Visvliet	no	yes	no	yes	
Oosterbierum / Ea	no	yes	no	yes	
Beetgum / Bitgum	no	yes	NA	NA	
Bergum / Burgum	no	yes	no	yes	
Jorwerd / Jorwert	no	yes	NA	NA	
Bakkeveen / Bakk	no	yes	no	yes	
Waskemeer / De V	no	yes	no	yes	
Kloosterburen	no	yes	no	yes	
Warffum	no	yes	no	yes	
Leermens	no	yes	no	yes	
Groningen	no	yes	yes	no	
Nieuw-Scheemda	NA	NA	no	yes	
Langelo	no	yes	no	yes	

- step 1: convert the data table into a 267×267 (symmetric) distance matrix, whereby for each pair of locations a distance between them is calculated based on the linguistic features they share

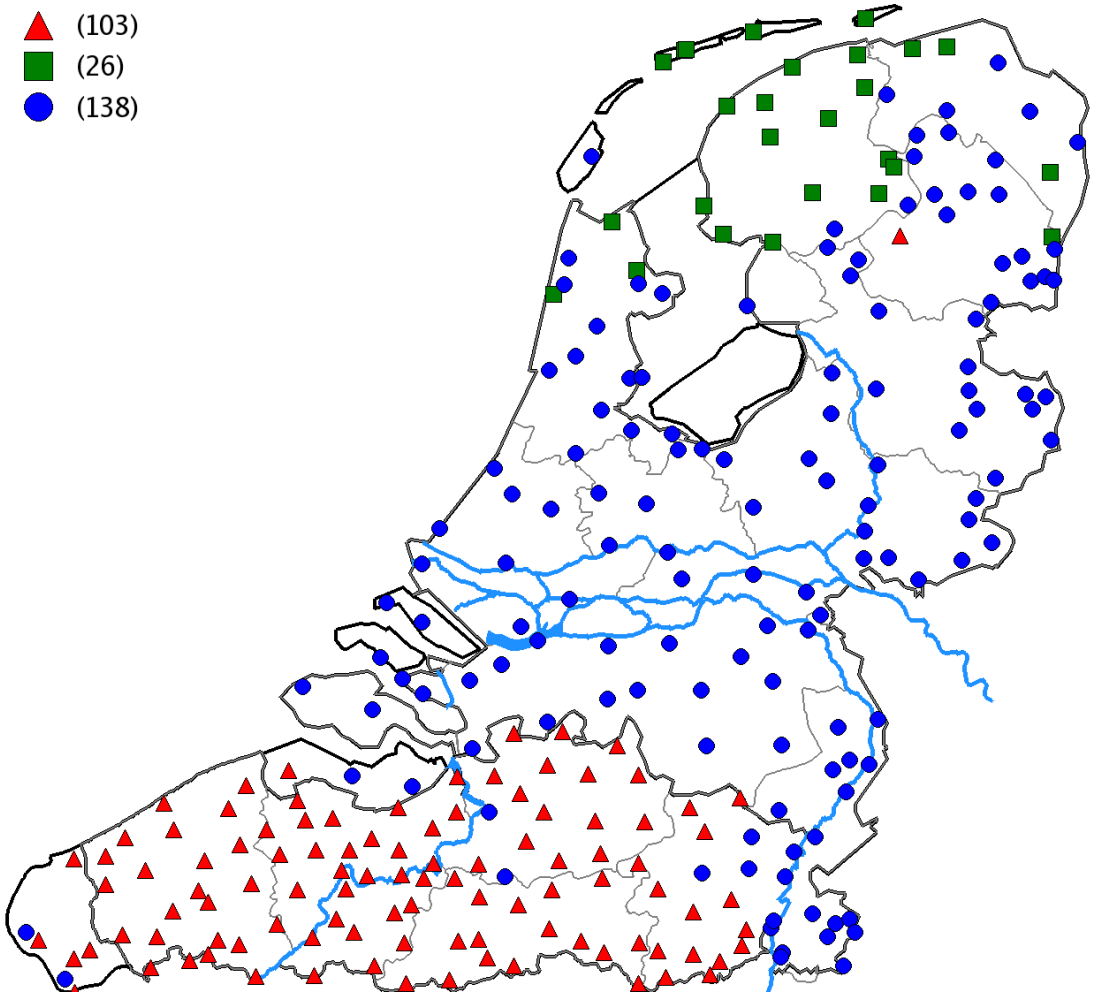
	Midsland	Lies	West-Ters	Oosteren	Hollum	Schiermo	Ferwerd /	Anjum /	Kollum	Visvliet	Oosterbie	Beetgum	Bergum /	Jorwerd /
Midsland / Mi	0,000	0,500	0,333	0,706	0,250	0,647	0,357	0,250	0,611	0,650	0,533	0,545	0,500	0,692
Lies	0,500	0,000	0,444	0,750	0,588	0,375	0,471	0,563	0,444	0,444	0,632	0,714	0,500	0,667
West-Terschel	0,333	0,444	0,000	0,789	0,429	0,667	0,286	0,429	0,632	0,600	0,500	0,500	0,429	0,583
Oosterend	0,706	0,750	0,789	0,000	0,706	0,765	0,737	0,538	0,563	0,600	0,600	0,727	0,813	0,846
Hollum	0,250	0,588	0,429	0,706	0,000	0,667	0,167	0,000	0,625	0,714	0,462	0,500	0,500	0,545
Schiermonnik	0,647	0,375	0,667	0,765	0,667	0,000	0,625	0,667	0,400	0,556	0,706	0,750	0,571	0,667
Ferwerd / Fer	0,357	0,471	0,286	0,737	0,167	0,625	0,000	0,182	0,588	0,682	0,308	0,333	0,333	0,400
Anjum / Eanju	0,250	0,563	0,429	0,538	0,000	0,667	0,182	0,000	0,571	0,625	0,417	0,556	0,500	0,600
Kollum	0,611	0,444	0,632	0,563	0,625	0,400	0,588	0,571	0,000	0,353	0,625	0,643	0,429	0,571
Visvliet	0,650	0,444	0,600	0,600	0,714	0,556	0,682	0,625	0,353	0,000	0,588	0,500	0,667	0,692
Oosterbierum	0,533	0,632	0,500	0,600	0,462	0,706	0,308	0,417	0,625	0,588	0,000	0,167	0,571	0,500
Beetgum / Bit	0,545	0,714	0,500	0,727	0,500	0,750	0,333	0,556	0,643	0,500	0,167	0,000	0,500	0,455
Bergum / Bur	0,500	0,500	0,429	0,813	0,500	0,571	0,333	0,500	0,429	0,667	0,571	0,500	0,000	0,222
Jorwerd / Jorv	0,692	0,667	0,583	0,846	0,545	0,667	0,400	0,600	0,571	0,692	0,500	0,455	0,222	0,000
Bakkeveen / E	0,400	0,500	0,438	0,706	0,385	0,563	0,357	0,385	0,438	0,579	0,533	0,545	0,385	0,583
Waskemeer /	0,438	0,526	0,556	0,818	0,500	0,588	0,471	0,533	0,471	0,652	0,588	0,667	0,429	0,500
Kloosterburen	0,500	0,412	0,611	0,810	0,563	0,357	0,529	0,600	0,333	0,636	0,706	0,667	0,385	0,583
Warffum	0,563	0,438	0,667	0,737	0,625	0,429	0,588	0,643	0,400	0,652	0,600	0,636	0,571	0,750
Leermens	0,667	0,652	0,739	0,550	0,773	0,650	0,739	0,722	0,389	0,455	0,667	0,571	0,684	0,765
Groningen	0,714	0,682	0,714	0,636	0,783	0,762	0,800	0,778	0,471	0,476	0,684	0,714	0,737	0,786
Nieuw-Scheer	0,650	0,682	0,650	0,652	0,773	0,762	0,739	0,722	0,556	0,368	0,647	0,615	0,667	0,786
Langelo	0,727	0,524	0,739	0,652	0,792	0,650	0,760	0,647	0,550	0,500	0,700	0,824	0,810	0,950

- step 2: reduce this 267-dimensional matrix to a two- or three-dimensional one, so that it can easily be visualized


second coordinate
first coordinate
0.4
0.2
0.0
-0.2
-0.4
-0.4
-0.2
0.0
0.2


second coordinate
first coordinate
0.4
0.2
0.0
-0.2
-0.4
-0.4
-0.2
0.0
0.2

- step 3: project back onto a geographical map


(103)
(26)
(138)

- shortcomings of this approach for my current purposes:

- shortcomings of this approach for my current purposes:
  1. the linguistic constructions themselves play only an indirect role in the outcome of the analysis: we can see when two dialects differ, but we don't see which cluster orders are responsible for this difference or how they cluster or correlate

- shortcomings of this approach for my current purposes:
  1. the linguistic constructions themselves play only an indirect role in the outcome of the analysis: we can see when two dialects differ, but we don't see which cluster orders are responsible for this difference or how they cluster or correlate
  2. there is no link between the data that feed into the quantitative analysis and the formal theoretical literature on verb clusters

Reverse dialectometry

Reverse dialectometry

- **proposal:** two changes to the classical dialectometric setup:

Reverse dialectometry

- **proposal:** two changes to the classical dialectometric setup:
  1. cluster orders are *individuals* rather than variables, i.e. instead of calculating differences between dialect locations, we measure differences between linguistic constructions

Reverse dialectometry

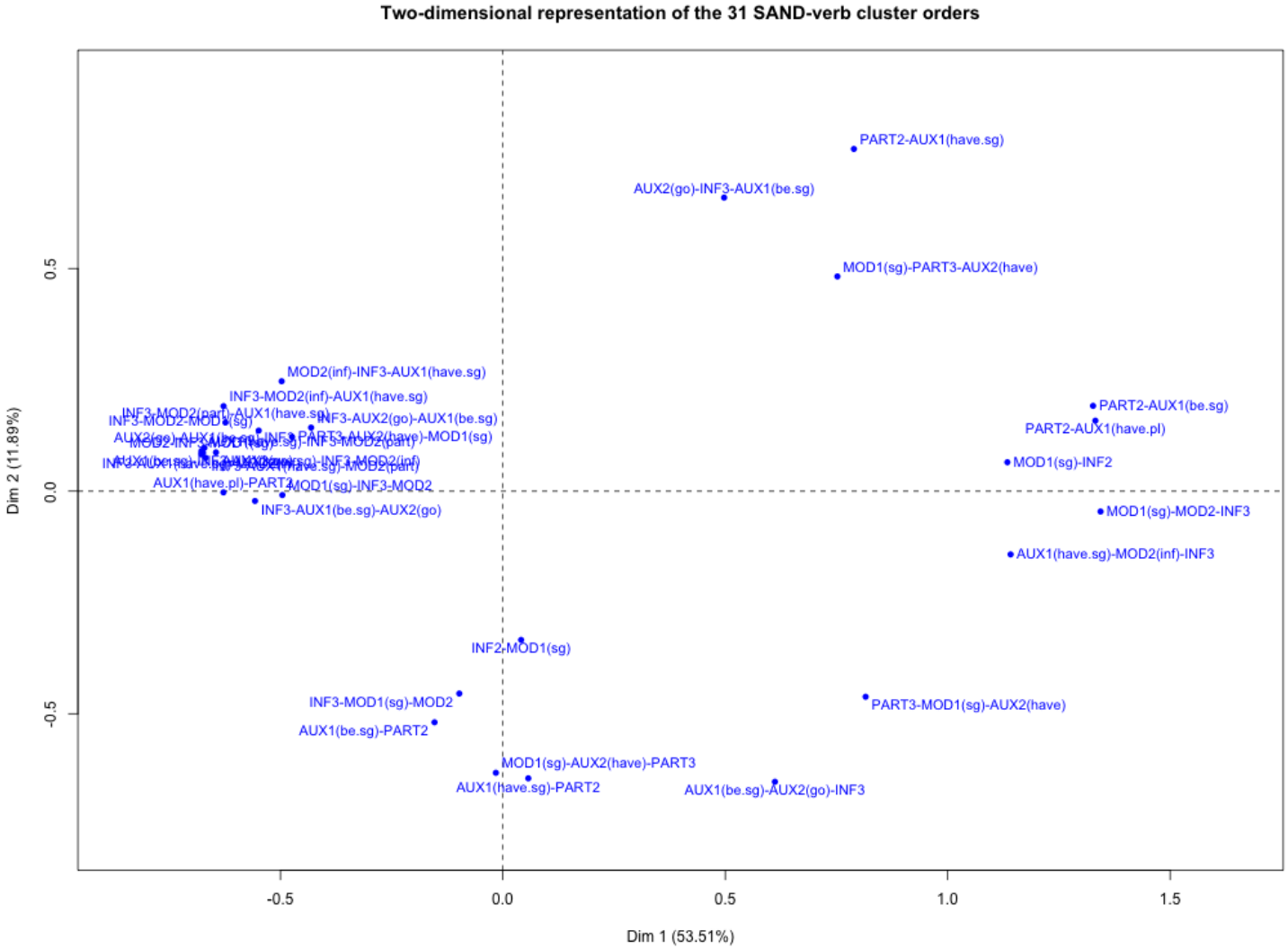
- **proposal:** two changes to the classical dialectometric setup:
  1. cluster orders are *individuals* rather than variables, i.e. instead of calculating differences between dialect locations, we measure differences between linguistic constructions
  2. Multiple Correspondence Analysis (MCA) instead of Multidimensional Scaling (MDS): involves the same kind of dimension reduction, but applied simultaneously to individuals and variables $\rightarrow$ will allow for the inclusion of formal theoretical variables alongside geographical ones

Reverse dialectometry

- **proposal:** two changes to the classical dialectometric setup:
  1. cluster orders are *individuals* rather than variables, i.e. instead of calculating differences between dialect locations, we measure differences between linguistic constructions
  2. Multiple Correspondence Analysis (MCA) instead of Multidimensional Scaling (MDS): involves the same kind of dimension reduction, but applied simultaneously to individuals and variables $\rightarrow$ will allow for the inclusion of formal theoretical variables alongside geographical ones
- starting point: a data table with cluster orders as rows and dialect locations as columns

	Midsland	Lies	West.Tersch	Oosterend	Hollum	Schiermonni	Ferwerd	Anjum	Kollum	Visvliet
AUX1(be.sg)-PART2	no	no	no	NA	no	no	no	no	no	no
PART2-AUX1(be.sg)	yes	yes	yes	NA	yes	yes	yes	yes	yes	yes
AUX1(have.sg)-PART2	no	no	no	no	NA	no	no	no	no	no
PART2-AUX1(have.sg)	yes	yes	yes	yes	NA	yes	yes	yes	yes	yes
AUX1(have.pl)-PART2	no	no	no	no	no	no	no	no	no	no
PART2-AUX1(have.pl)	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
MOD1(sg)-INF2	no	no	yes	no	no	no	no	no	no	yes
INF2-MOD1(sg)	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
MOD2-INF3-MOD1(sg)	no	no	no	no	no	no	no	no	no	no
MOD1(sg)-MOD2-INF3	no	no	no	yes	no	no	no	no	yes	yes
MOD1(sg)-INF3-MOD2	yes	no	no	no	no	no	no	no	no	no
INF3-MOD2-MOD1(sg)	yes	yes	yes	no	yes	yes	yes	yes	yes	yes
INF3-MOD1(sg)-MOD2	no	no	no	no	no	no	no	no	no	yes
MOD1(sg)-AUX2(have)-PART3	no	no	no	no	no	no	no	NA	no	no
MOD1(sg)-PART3-AUX2(have)	no	no	no	no	no	no	no	NA	yes	yes
PART3-MOD1(sg)-AUX2(have)	no	yes	no	yes	no	no	no	NA	yes	yes
PART3-AUX2(have)-MOD1(sg)	yes	yes	yes	no	yes	yes	yes	NA	yes	yes
AUX1(be.sg)-AUX2(go)-INF3	no	no	no	yes	no	no	no	no	NA	yes
AUX1(be.sg)-INF3-AUX2(go)	no	no	no	no	no	no	no	no	NA	no
AUX2(go)-AUX1(be.sg)-INF3	no	no	no	no	no	yes	no	no	NA	no
AUX2(go)-INF3-AUX1(be.sg)	no	no	no	no	no	no	no	no	NA	no
INF3-AUX1(be.sg)-AUX2(go)	no	no	no	no	no	no	no	no	NA	no
INF3-AUX2(go)-AUX1(be.sg)	yes	yes	yes	no	yes	no	yes	yes	NA	no
AUX1(have.sg)-MOD2(inf)-INF3	no	no	no	yes	no	no	no	no	no	no
AUX1(have.sg)-INF3-MOD2(part)	no	no	no	no	no	no	no	no	no	yes
AUX1(have.sg)-INF3-MOD2(inf)	no	no	no	no	no	no	no	no	no	no
MOD2(inf)-INF3-AUX1(have.sg)	no	no	no	no	no	no	no	no	no	no
INF3-AUX1(have.sg)-MOD2(inf)	no	no	yes	no	no	no	no	no	no	no
INF3-AUX1(have.sg)-MOD2(part)	no	no	no	no	no	no	no	no	no	yes
INF3-MOD2(part)-AUX1(have.sg)	no	yes	no	no	no	yes	no	no	yes	yes
INF3-MOD2(inf)-AUX1(have.sg)	yes	yes	yes	no	yes	no	yes	yes	no	yes

- transform to a distance matrix and reduce its dimensionality


Two-dimensional representation of the 31 SAND-verb cluster orders
PART2-AUX1(have.sg)
AUX2(go)-INF3-AUX1(be.sg)
MOD1(sg)-PART3-AUX2(have)
MOD2(inf)-INF3-AUX1(have.sg)
INF3-MOD2(inf)-AUX1(have.sg)
INF3-AUX2(go)-AUX1(be.sg)
PART2-AUX1(be.sg)
PART2-AUX1(have.pl)
MOD1(sg)-INF2
AUX1(have.pl)-PART2
MOD1(sg)-INF3-MOD2
INF3-AUX1(be.sg)-AUX2(go)
MOD1(sg)-MOD2-INF3
AUX1(have.sg)-MOD2(inf)-INF3
INF2-MOD1(sg)
INF3-MOD1(sg)-MOD2
PART3-MOD1(sg)-AUX2(have)
AUX1(be.sg)-PART2
MOD1(sg)-AUX2(have)-PART3
AUX1(have.sg)-PART2
AUX1(be.sg)-AUX2(go)-INF3
Dim 2 (11.89%)
0.5
0.0
-0.5
-0.5
0.0
0.5
1.0
1.5
Dim 1 (53.51%)

- **note:** each point now represents a particular cluster order and closeness of points indicates how alike two verb cluster orders are based on their geographical spread

- **note:** each point now represents a particular cluster order and closeness of points indicates how alike two verb cluster orders are based on their geographical spread
- if this likeness is the result of grammatical parameters, then verb cluster orders that are 'closeby' should be the result of the same parameter setting, i.e. parameters create **natural classes** of verb cluster orders

- **note:** each point now represents a particular cluster order and closeness of points indicates how alike two verb cluster orders are based on their geographical spread
- if this likeness is the result of grammatical parameters, then verb cluster orders that are 'closeby' should be the result of the same parameter setting, i.e. parameters create **natural classes** of verb cluster orders
- in order to find those parameters, we can also encode the cluster orders in terms of their theoretical linguistic analyses

- theoretical accounts differ in which analysis they assign to which cluster order ⇒ cluster orders have their own specific 'fingerprint' in each analysis, some of them very similar to one another and others very different

- theoretical accounts differ in which analysis they assign to which cluster order ⇒ cluster orders have their own specific 'fingerprint' in each analysis, some of them very similar to one another and others very different
- we can encode the SAND cluster orders in our database in terms of those fingerprints and then compare them to the geographical clustering

- theoretical accounts differ in which analysis they assign to which cluster order ⇒ cluster orders have their own specific 'fingerprint' in each analysis, some of them very similar to one another and others very different
- we can encode the SAND cluster orders in our database in terms of those fingerprints and then compare them to the geographical clustering
- e.g. in Barbiers (2005)'s analysis cluster orders can differ from one another on four counts:

- theoretical accounts differ in which analysis they assign to which cluster order ⇒ cluster orders have their own specific 'fingerprint' in each analysis, some of them very similar to one another and others very different
- we can encode the SAND cluster orders in our database in terms of those fingerprints and then compare them to the geographical clustering
- e.g. in Barbiers (2005)'s analysis cluster orders can differ from one another on four counts:
  - [±base-generation]: can the order be base-generated?

- theoretical accounts differ in which analysis they assign to which cluster order ⇒ cluster orders have their own specific 'fingerprint' in each analysis, some of them very similar to one another and others very different
- we can encode the SAND cluster orders in our database in terms of those fingerprints and then compare them to the geographical clustering
- e.g. in Barbiers (2005)'s analysis cluster orders can differ from one another on four counts:
  - [±base-generation]: can the order be base-generated?
  - [±movement]: can the order be derived via movement?

- theoretical accounts differ in which analysis they assign to which cluster order ⇒ cluster orders have their own specific 'fingerprint' in each analysis, some of them very similar to one another and others very different
- we can encode the SAND cluster orders in our database in terms of those fingerprints and then compare them to the geographical clustering
- e.g. in Barbiers (2005)'s analysis cluster orders can differ from one another on four counts:
  - [±base-generation]: can the order be base-generated?
  - [±movement]: can the order be derived via movement?
  - [±pied-piping]: does the derivation involve pied-piping?

- theoretical accounts differ in which analysis they assign to which cluster order ⇒ cluster orders have their own specific 'fingerprint' in each analysis, some of them very similar to one another and others very different
- we can encode the SAND cluster orders in our database in terms of those fingerprints and then compare them to the geographical clustering
- e.g. in Barbiers (2005)'s analysis cluster orders can differ from one another on four counts:
  - [±base-generation]: can the order be base-generated?
  - [±movement]: can the order be derived via movement?
  - [±pied-piping]: does the derivation involve pied-piping?
  - [±feature-checking violation]: does the order involve a feature checking violation?

	Barbiers-base.generation	Barbiers-movement	Barbiers-spec-pied-piping	Barbiers-feature.checking-failure
AUX1(be.sg)-PART2	yesBase	noMvt	noPiedP	noFeatCheckFail
PART2-AUX1(be.sg)	noBase	yesMvt	noPiedP	noFeatCheckFail
AUX1(have.sg)-PART2	yesBase	noMvt	noPiedP	noFeatCheckFail
PART2-AUX1(have.sg)	noBase	yesMvt	noPiedP	noFeatCheckFail
AUX1(have.pl)-PART2	yesBase	noMvt	noPiedP	noFeatCheckFail
PART2-AUX1(have.pl)	noBase	yesMvt	noPiedP	noFeatCheckFail
MOD1(sg)-INF2	yesBase	noMvt	noPiedP	noFeatCheckFail
INF2-MOD1(sg)	noBase	yesMvt	noPiedP	noFeatCheckFail
MOD2-INF3-MOD1(sg)	noBase	yesMvt	noPiedP	yesFeatCheckFail
MOD1(sg)-MOD2-INF3	yesBase	noMvt	noPiedP	noFeatCheckFail
MOD1(sg)-INF3-MOD2	noBase	yesMvt	noPiedP	noFeatCheckFail
INF3-MOD2-MOD1(sg)	noBase	yesMvt	yesPiedP	noFeatCheckFail
INF3-MOD1(sg)-MOD2	noBase	yesMvt	noPiedP	noFeatCheckFail
MOD1(sg)-AUX2(have)-PART3	yesBase	noMvt	noPiedP	noFeatCheckFail
MOD1(sg)-PART3-AUX2(have)	noBase	yesMvt	noPiedP	noFeatCheckFail
PART3-MOD1(sg)-AUX2(have)	noBase	yesMvt	noPiedP	noFeatCheckFail
PART3-AUX2(have)-MOD1(sg)	noBase	yesMvt	yesPiedP	noFeatCheckFail
AUX1(be.sg)-AUX2(go)-INF3	yesBase	noMvt	noPiedP	noFeatCheckFail
AUX1(be.sg)-INF3-AUX2(go)	noBase	yesMvt	noPiedP	noFeatCheckFail
AUX2(go)-AUX1(be.sg)-INF3	noBase	noMvt	noPiedP	noFeatCheckFail
AUX2(go)-INF3-AUX1(be.sg)	noBase	yesMvt	noPiedP	noFeatCheckFail
INF3-AUX1(be.sg)-AUX2(go)	noBase	yesMvt	noPiedP	yesFeatCheckFail
INF3-AUX2(go)-AUX1(be.sg)	noBase	yesMvt	noPiedP	noFeatCheckFail
AUX1(have.sg)-MOD2(inf)-INF3	yesBase	noMvt	noPiedP	noFeatCheckFail
AUX1(have.sg)-INF3-MOD2(part)	noBase	yesMvt	noPiedP	noFeatCheckFail
AUX1(have.sg)-INF3-MOD2(inf)	noBase	yesMvt	noPiedP	noFeatCheckFail
MOD2(inf)-INF3-AUX1(have.sg)	noBase	yesMvt	noPiedP	noFeatCheckFail
INF3-AUX1(have.sg)-MOD2(inf)	noBase	yesMvt	noPiedP	yesFeatCheckFail
INF3-AUX1(have.sg)-MOD2(part)	noBase	yesMvt	noPiedP	yesFeatCheckFail
INF3-MOD2(part)-AUX1(have.sg)	noBase	yesMvt	noPiedP	noFeatCheckFail
INF3-MOD2(inf)-AUX1(have.sg)	noBase	yesMvt	noPiedP	noFeatCheckFail

- **in total:** 70 additional variables distilled from the theoretical literature on verb clusters have been added to the data table:

- **in total:** 70 additional variables distilled from the theoretical literature on verb clusters have been added to the data table:
  - the analyses of Barbiers (2005), Barbiers and Bennis (2010), Abels (2011), Haegeman and Riemsdijk (1986), Bader (2012), and Schmid and Vogel (2004)

- **in total:** 70 additional variables distilled from the theoretical literature on verb clusters have been added to the data table:
  - the analyses of Barbiers (2005), Barbiers and Bennis (2010), Abels (2011), Haegeman and Riemsdijk (1986), Bader (2012), and Schmid and Vogel (2004)
  - four analyses from Wurmbrand (2005): a head-initial head movement analysis, a head-final head movement analysis, a head-initial XP-movement analysis, a head-final XP-movement analysis

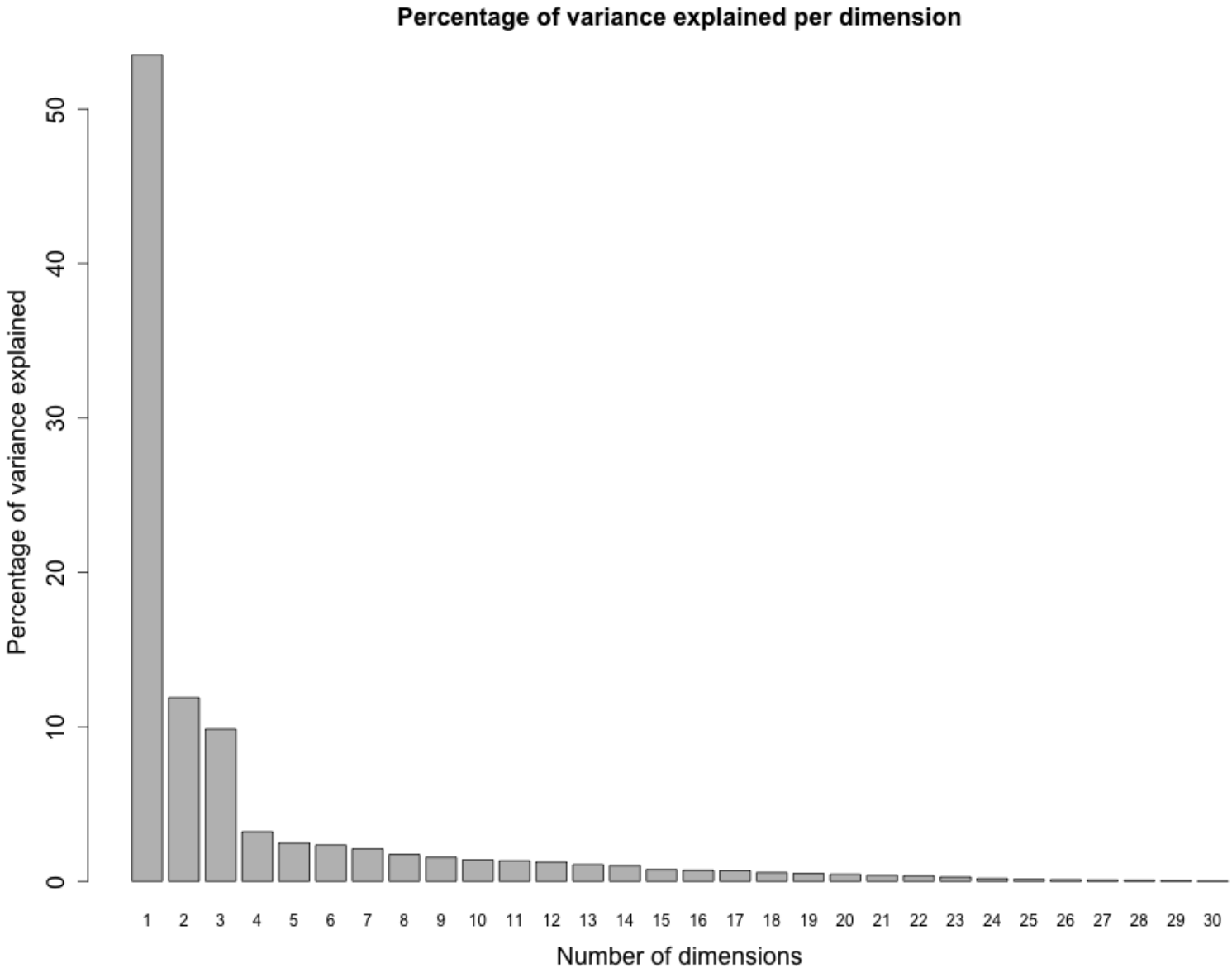
- **in total:** 70 additional variables distilled from the theoretical literature on verb clusters have been added to the data table:
  - the analyses of Barbiers (2005), Barbiers and Bennis (2010), Abels (2011), Haegeman and Riemsdijk (1986), Bader (2012), and Schmid and Vogel (2004)
  - four analyses from Wurmbrand (2005): a head-initial head movement analysis, a head-final head movement analysis, a head-initial XP-movement analysis, a head-final XP-movement analysis
  - 17 additional variables based on the theoretical literature, but not linked to a specific analysis

- **in total:** 70 additional variables distilled from the theoretical literature on verb clusters have been added to the data table:
  - the analyses of Barbiers (2005), Barbiers and Bennis (2010), Abels (2011), Haegeman and Riemsdijk (1986), Bader (2012), and Schmid and Vogel (2004)
  - four analyses from Wurmbrand (2005): a head-initial head movement analysis, a head-final head movement analysis, a head-initial XP-movement analysis, a head-final XP-movement analysis
  - 17 additional variables based on the theoretical literature, but not linked to a specific analysis
- in the analysis, these 70 variables are used as **supplementary variables**: they do not contribute to the dimension reduction, but they are mapped against its output, in order to interpret the results

- **recall:** we are trying to determine if the variation in word order in verbal clusters is determined by grammatical parameters, and if so to what extent

- **recall:** we are trying to determine if the variation in word order in verbal clusters is determined by grammatical parameters, and if so to what extent
- this means we need to determine **how many** parameters there are and **what they are**

- **recall:** we are trying to determine if the variation in word order in verbal clusters is determined by grammatical parameters, and if so to what extent
- this means we need to determine **how many** parameters there are and **what they are**
- **proposal (I):** the number of parameters responsible for the verb cluster variation = the number of dimensions we reduce our data set to


Percentage of variance explained per dimension
Percentage of variance explained
0
10
20
30
40
50
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
Number of dimensions

- **note:** there seems to be a clear cut-off point after the third dimension

- **note:** there seems to be a clear cut-off point after the third dimension
- together, the first three dimensions account for 78.46% of the variation in the SAND verb cluster data

- **note:** there seems to be a clear cut-off point after the third dimension
- together, the first three dimensions account for 78.46% of the variation in the SAND verb cluster data
- this means that roughly 80% of the variation in verb cluster ordering in SAND can be reduced to three parameters

- **note:** there seems to be a clear cut-off point after the third dimension
- together, the first three dimensions account for 78.46% of the variation in the SAND verb cluster data
- this means that roughly 80% of the variation in verb cluster ordering in SAND can be reduced to three parameters
- in order to know what those parameters are, we need to *interpret* the first three dimensions

- **proposal (I):** the number of parameters responsible for the verb cluster variation = the number of dimensions we reduce our data set to

- **proposal (I):** the number of parameters responsible for the verb cluster variation = the number of dimensions we reduce our data set to
- **proposal (II):** the identity of those parameters = the interpretation of the dimensions

- **proposal (I):** the number of parameters responsible for the verb cluster variation = the number of dimensions we reduce our data set to
- **proposal (II):** the identity of those parameters = the interpretation of the dimensions
- the degree of similarity/correlation between a dimension and a linguistic variable can be determined by:
  1. visual inspection of a color-coded map

- **proposal (I):** the number of parameters responsible for the verb cluster variation = the number of dimensions we reduce our data set to
- **proposal (II):** the identity of those parameters = the interpretation of the dimensions
- the degree of similarity/correlation between a dimension and a linguistic variable can be determined by:
  1. visual inspection of a color-coded map
  2. calculating the squared correlation ratio (η^2): value between 0 and 1 indicating the strength of the link between a dimension and a particular categorical variable; can be interpreted as the percentage of variation on the dimension that can be explained by that categorical variable

Dimension 1

Dimension 1

- is related to the **morphological form** of the verb: infinitive (*will* **see**) or auxiliary (*have* **seen**)

Dimension 1

- is related to the **morphological form** of the verb: infinitive (*will* **see**) or auxiliary (*have* **seen**)
- this dimension separates dialects where the infinitive follows the auxiliary it combines with (*will see*) and the participle precedes the auxiliary it combines with (*seen have*) from dialects where at least one of those orders differs

Dimension 1

- is related to the **morphological form** of the verb: infinitive (*will* **see**) or auxiliary (*have* **seen**)
- this dimension separates dialects where the infinitive follows the auxiliary it combines with (*will see*) and the participle precedes the auxiliary it combines with (*seen have*) from dialects where at least one of those orders differs
- more specifically, the variable InfMod.AuxPart:

Dimension 1

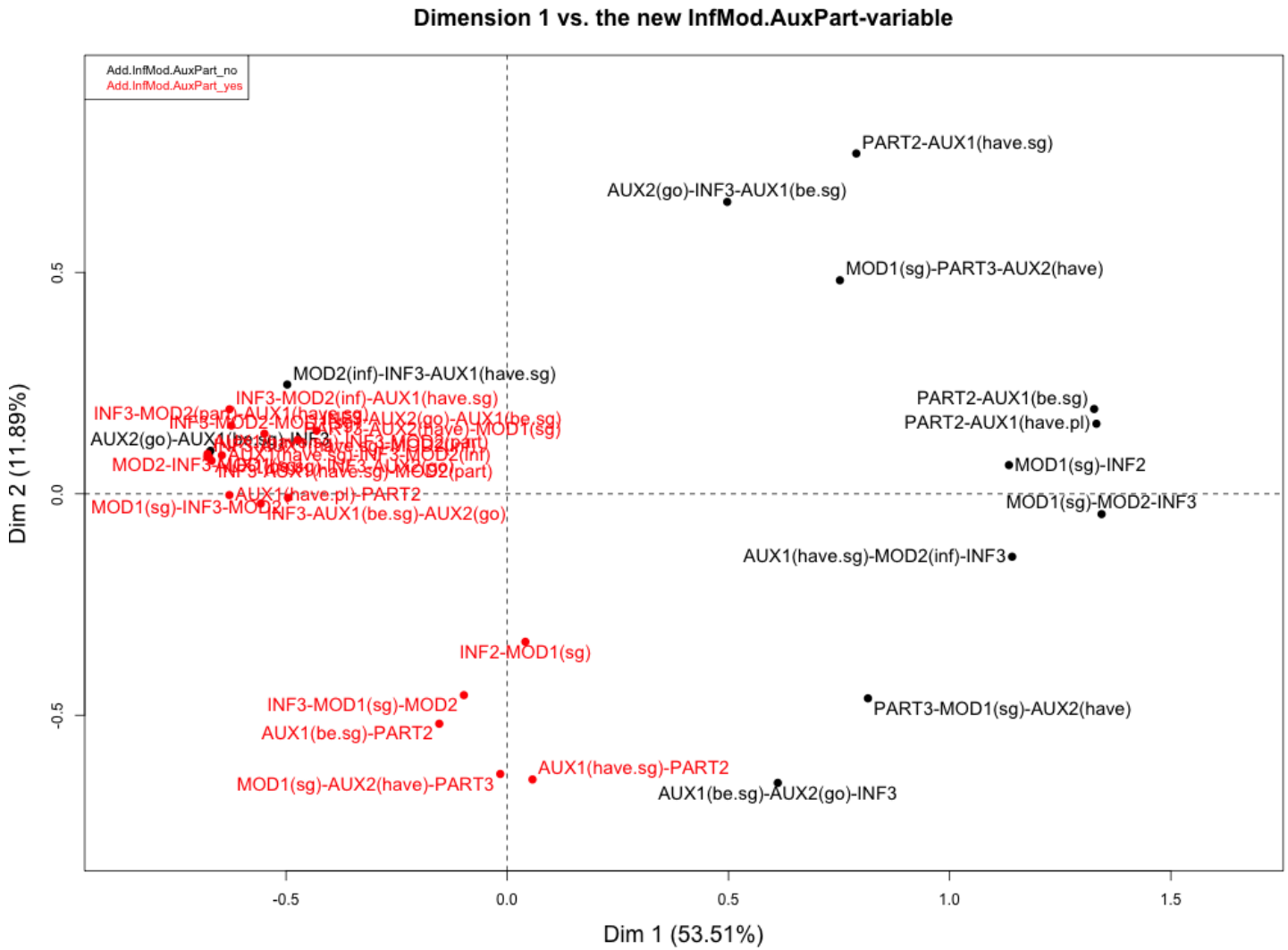
- is related to the **morphological form** of the verb: infinitive (*will* **see**) or auxiliary (*have* **seen**)
- this dimension separates dialects where the infinitive follows the auxiliary it combines with (*will see*) and the participle precedes the auxiliary it combines with (*seen have*) from dialects where at least one of those orders differs
- more specifically, the variable InfMod.AuxPart:
  - set to 'no' when the modal precedes the infinitive (when present) and the participle precedes the auxiliary (when present)

Dimension 1

- is related to the **morphological form** of the verb: infinitive (*will* **see**) or auxiliary (*have* **seen**)
- this dimension separates dialects where the infinitive follows the auxiliary it combines with (*will see*) and the participle precedes the auxiliary it combines with (*seen have*) from dialects where at least one of those orders differs
- more specifically, the variable InfMod.AuxPart:
  - set to 'no' when the modal precedes the infinitive (when present) and the participle precedes the auxiliary (when present)
  - set to 'yes' when at least one of these conditions is not met

Dimension 1

- is related to the **morphological form** of the verb: infinitive (*will* **see**) or auxiliary (*have* **seen**)
- this dimension separates dialects where the infinitive follows the auxiliary it combines with (*will see*) and the participle precedes the auxiliary it combines with (*seen have*) from dialects where at least one of those orders differs
- more specifically, the variable InfMod.AuxPart:
  - set to 'no' when the modal precedes the infinitive (when present) and the participle precedes the auxiliary (when present)
  - set to 'yes' when at least one of these conditions is not met
- this variable has a η^2 of 0.6142


Dimension 1 vs. the new InfMod.AuxPart-variable
Add.InfMod.AuxPart_no
Add.InfMod.AuxPart_yes
PART2-AUX1(have.sg)
AUX2(go)-INF3-AUX1(be.sg)
MOD1(sg)-PART3-AUX2(have)
MOD2(inf)-INF3-AUX1(have.sg)
INF3-MOD2(inf)-AUX1(have.sg)
PART2-AUX1(be.sg)
PART2-AUX1(have.pl)
MOD1(sg)-INF2
MOD1(sg)-MOD2-INF3
AUX1(have.sg)-MOD2(inf)-INF3
INF2-MOD1(sg)
PART3-MOD1(sg)-AUX2(have)
INF3-MOD1(sg)-MOD2
AUX1(be.sg)-PART2
AUX1(have.sg)-PART2
MOD1(sg)-AUX2(have)-PART3
AUX1(be.sg)-AUX2(go)-INF3
Dim 2 (11.89%)
Dim 1 (53.51%)
0.5
0.0
-0.5
-0.5
0.0
0.5
1.0
1.5

Dimension 2

- is related to the 'slope' of the cluster:
  ascending (e.g. $1 \nearrow 2 \nearrow 3$) or descending (e.g. $3 \searrow 2 \searrow 1$)

Dimension 2

- is related to the 'slope' of the cluster:
  ascending (e.g. 1 $\nearrow$ 2 $\nearrow$ 3) or descending (e.g. 3 $\searrow$ 2 $\searrow$ 1)
- more specifically, the variable FinalDescent:

Dimension 2

- is related to the 'slope' of the cluster:
  ascending (e.g. 1 $\nearrow$ 2 $\nearrow$ 3) or descending (e.g. 3 $\searrow$ 2 $\searrow$ 1)
- more specifically, the variable FinalDescent:
  - set to 'yes' if the cluster ends in a descending order

Dimension 2

- is related to the 'slope' of the cluster:
  ascending (e.g. 1 $\nearrow$ 2 $\nearrow$ 3) or descending (e.g. 3 $\searrow$ 2 $\searrow$ 1)
- more specifically, the variable FinalDescent:
  - set to 'yes' if the cluster ends in a descending order
  - set to 'no' if it ends in an ascending order

Dimension 2

- is related to the 'slope' of the cluster:
  ascending (e.g. 1 ↗ 2 ↗ 3) or descending (e.g. 3 ↘ 2 ↘ 1)
- more specifically, the variable FinalDescent:
  - set to 'yes' if the cluster ends in a descending order
  - set to 'no' if it ends in an ascending order

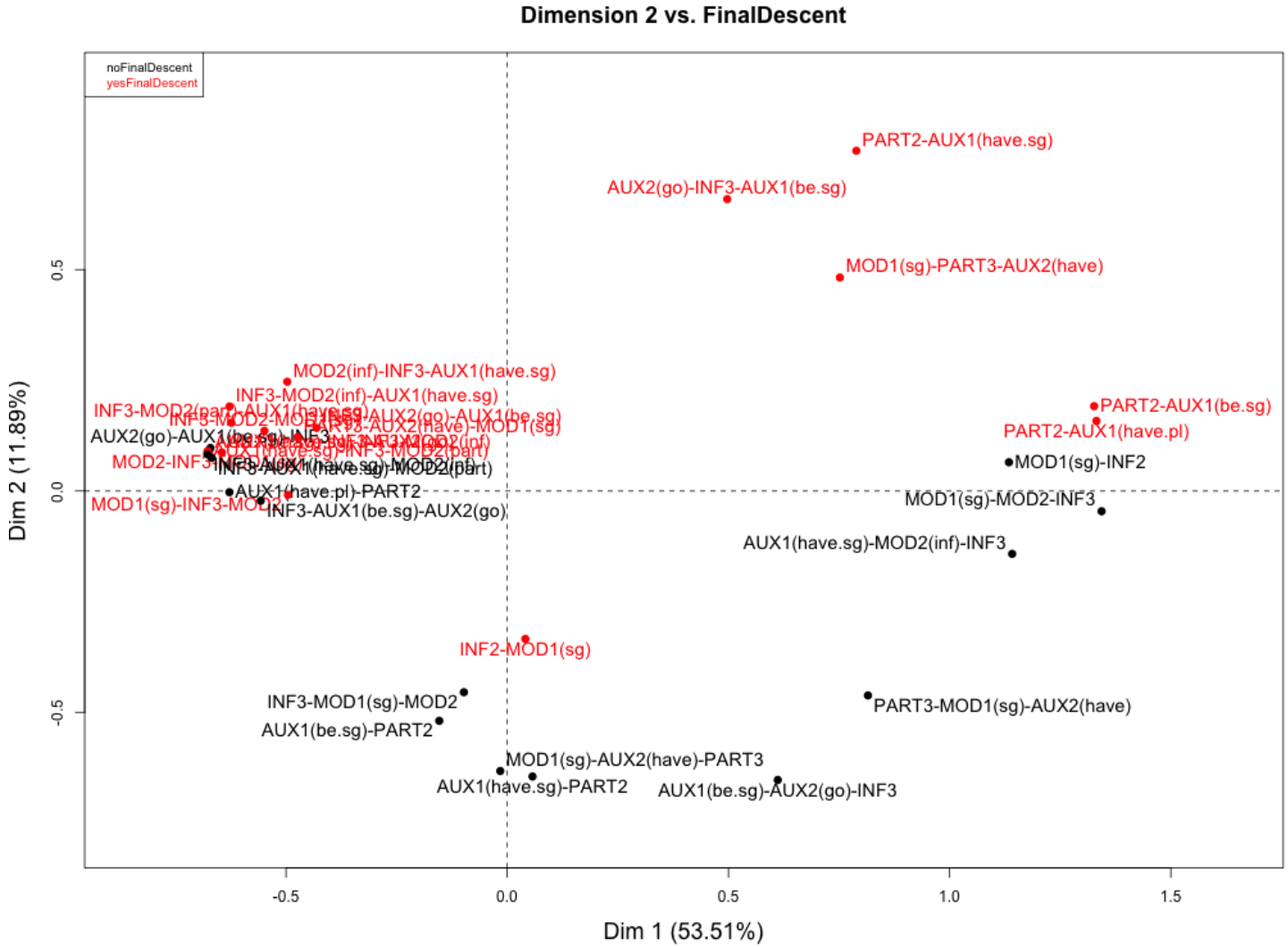
FinalDescent_yes	**FinalDescent_no**
21	12
132	123
321	312
231	213

Dimension 2

- is related to the 'slope' of the cluster:
  ascending (e.g. 1 ↗ 2 ↗ 3) or descending (e.g. 3 ↘ 2 ↘ 1)
- more specifically, the variable FinalDescent:
  - set to 'yes' if the cluster ends in a descending order
  - set to 'no' if it ends in an ascending order

FinalDescent_yes	**FinalDescent_no**
21	12
132	123
321	312
231	213

- this variable has a η^2 of 0.382

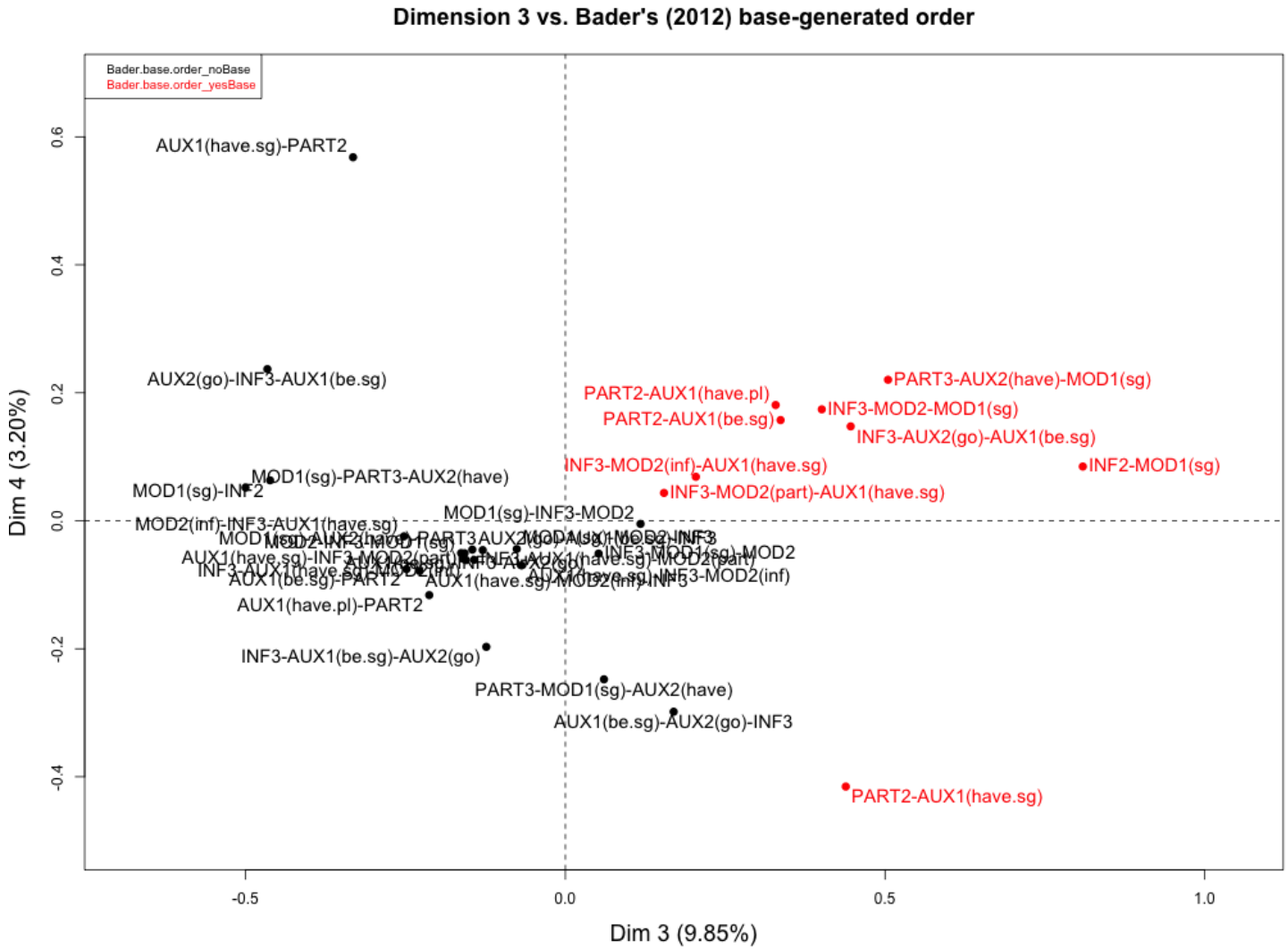

Dimension 2 vs. FinalDescent
noFinalDescent
yesFinalDescent
PART2-AUX1(have.sg)
AUX2(go)-INF3-AUX1(be.sg)
MOD1(sg)-PART3-AUX2(have)
MOD2(inf)-INF3-AUX1(have.sg)
INF3-MOD2(inf)-AUX1(have.sg)
PART2-AUX1(be.sg)
PART2-AUX1(have.pl)
MOD1(sg)-INF2
MOD1(sg)-MOD2-INF3
AUX1(have.sg)-MOD2(inf)-INF3
AUX1(have.pl)-PART2
INF3-AUX1(be.sg)-AUX2(go)
INF2-MOD1(sg)
INF3-MOD1(sg)-MOD2
AUX1(be.sg)-PART2
PART3-MOD1(sg)-AUX2(have)
MOD1(sg)-AUX2(have)-PART3
AUX1(have.sg)-PART2
AUX1(be.sg)-AUX2(go)-INF3
Dim 2 (11.89%)
Dim 1 (53.51%)
0.5
0.0
-0.5
-0.5
0.0
0.5
1.0
1.5

Dimension 3

- is again related to the slope of the cluster (and strongly so)

Dimension 3

- is again related to the slope of the cluster (and strongly so)
- it separates the strictly descending orders (i.e. 21 and 321) from all the others (12, 123, 132, 312, 213, 231): $\eta^2 = 0.686$


Dimension 3 vs. Bader's (2012) base-generated order
Bader.base.order_noBase
Bader.base.order_yesBase
AUX1(have.sg)-PART2
AUX2(go)-INF3-AUX1(be.sg)
MOD1(sg)-PART3-AUX2(have)
MOD1(sg)-INF2
MOD2(inf)-INF3-AUX1(have.sg)
MOD1(sg)-INF3-MOD2
AUX1(have.sg)-INF3-MOD2(part)
INF3-AUX1(have.sg)-MOD2(inf)
AUX1(be.sg)-PART2
AUX1(have.pl)-PART2
INF3-AUX1(be.sg)-AUX2(go)
PART3-MOD1(sg)-AUX2(have)
AUX1(be.sg)-AUX2(go)-INF3
PART2-AUX1(have.pl)
PART2-AUX1(be.sg)
PART3-AUX2(have)-MOD1(sg)
INF3-MOD2-MOD1(sg)
INF3-AUX2(go)-AUX1(be.sg)
INF3-MOD2(inf)-AUX1(have.sg)
INF2-MOD1(sg)
INF3-MOD2(part)-AUX1(have.sg)
INF3-MOD2(inf)
PART2-AUX1(have.sg)
Dim 4 (3.20%)
Dim 3 (9.85%)
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.5
0.0
0.5
1.0

Combining the dimensions into a theoretical analysis

Combining the dimensions into a theoretical analysis

- the quantitative-statistical analysis thus yields three ingredients which theoretical linguists can use to base their analysis on

Combining the dimensions into a theoretical analysis

- the quantitative-statistical analysis thus yields three ingredients which theoretical linguists can use to base their analysis on
- for example, a possible parametrized analysis of verb clusters:

Combining the dimensions into a theoretical analysis

- the quantitative-statistical analysis thus yields three ingredients which theoretical linguists can use to base their analysis on
- for example, a possible parametrized analysis of verb clusters:
  1. a head-final base order

Combining the dimensions into a theoretical analysis

- the quantitative-statistical analysis thus yields three ingredients which theoretical linguists can use to base their analysis on
- for example, a possible parametrized analysis of verb clusters:
  1. a head-final base order
  2. which dialects can diverge from or not: [±Movement] (dimension 3)

Combining the dimensions into a theoretical analysis

- the quantitative-statistical analysis thus yields three ingredients which theoretical linguists can use to base their analysis on
- for example, a possible parametrized analysis of verb clusters:
  1. a head-final base order
  2. which dialects can diverge from or not: [±Movement] (dimension 3)
  3. those that diverge can diverge strongly or not: Economy of Movement (dimension 2)

Combining the dimensions into a theoretical analysis

- the quantitative-statistical analysis thus yields three ingredients which theoretical linguists can use to base their analysis on
- for example, a possible parametrized analysis of verb clusters:
  1. a head-final base order
  2. which dialects can diverge from or not: [±Movement] (dimension 3)
  3. those that diverge can diverge strongly or not: Economy of Movement (dimension 2)
  4. above and beyond all this, a headedness parameter regulates the order of infinitives and participles *vis-à-vis* their selecting verbs: [±ModInf&PartAux] (dimension 1)

Conclusion

- roughly 80% of the variation found in Dutch verb cluster orders can be reduced to three grammatical parameters by applying a statistical analysis to the data

Conclusion

- roughly 80% of the variation found in Dutch verb cluster orders can be reduced to three grammatical parameters by applying a statistical analysis to the data
- more generally, there is room for fruitful collaboration between formal-theoretical and quantitative-statistical linguistics:

Conclusion

- roughly 80% of the variation found in Dutch verb cluster orders can be reduced to three grammatical parameters by applying a statistical analysis to the data
- more generally, there is room for fruitful collaboration between formal-theoretical and quantitative-statistical linguistics:
  - the former can guide the interpretation of results from the latter

Conclusion

- roughly 80% of the variation found in Dutch verb cluster orders can be reduced to three grammatical parameters by applying a statistical analysis to the data
- more generally, there is room for fruitful collaboration between formal-theoretical and quantitative-statistical linguistics:
  - the former can guide the interpretation of results from the latter
  - the latter can help evaluate and test hypotheses of the former

References I

Abels, Klaus. 2011. Hierarchy-order relations in the germanic verb cluster and in the noun phrase. *GAGL* 53:1–28.

Bader, Markus. 2012. Verb-cluster variations: a harmonic grammar analysis. Handout of a talk presented at "New ways of analyzing syntactic variation", November 2012.

Barbiers, Sjef. 2005. Word order variation in three-verb clusters and the division of labour between generative linguistics and sociolinguistics. In *Syntax and variation. Reconciling the biological and the social*, ed. Leonie Cornips and Karen P. Corrigan, volume 265 of *Current issues in linguistic theory*, 233–264. John Benjamins.

Barbiers, Sjef, and Hans Bennis. 2010. De plaats van het werkwoord in zuid en noord. In *Voor Magda. Artikelen voor Magda Devos bij haar afscheid van de Universiteit Gent*, ed. Johan De Caluwe and Jacques Van Keymeulen, 25–42. Gent: Academia.

References II

Haegeman, Liliane, and Henk van Riemsdijk. 1986. Verb projection raising, scope, and the typology of verb movement rules. *Linguistic Inquiry* 17:417–466.

Nerbonne, John, and William A. Kretzschmar Jr. 2013. Dialectometry++. *Literary and Linguistic Computing* 28:2–12.

Schmid, Tanja, and Ralf Vogel. 2004. Dialectal variation in German 3-Verb clusters. *The Journal of Comparative Germanic Linguistics* 7:235–274.

Wurmbrand, Susanne. 2005. Verb clusters, verb raising, and restructuring. In *The Blackwell Companion to Syntax*, ed. Martin Everaert and Henk van Riemsdijk, volume V, chapter 75, 227–341. Oxford: Blackwell.