Handle your verb clusters with care

Jeroen van Craenenbroeck

KU Leuven
jeroen.vancraenenbroeck@kuleuven.be

Dealing with bad data in linguistic theory
Meertens Institute, 17–18 March 2016

Introduction

Main goal of this research

To explore the interaction between formal-theoretical (generative) linguistics and quantitative-statistical approaches to language data.

Central data

Word order variation in two- and three-verb clusters in 267 Dutch dialects.

The focus of today's talk

To explore to what extent various types of 'data badness' affect the theoretical conclusions based on the quantitative analysis.

Outline

Verb clusters: the data

- in Dutch (like in many Germanic languages) verbs cluster 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:

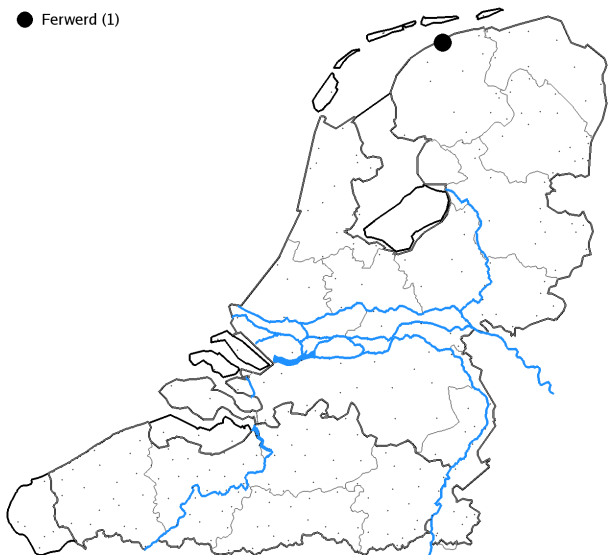
(2) dat hij gisteren tijdens de les heeft gelachen.
that he yesterday during the class had laughed
'that he laughed yesterday during class.' (12)

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

(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)

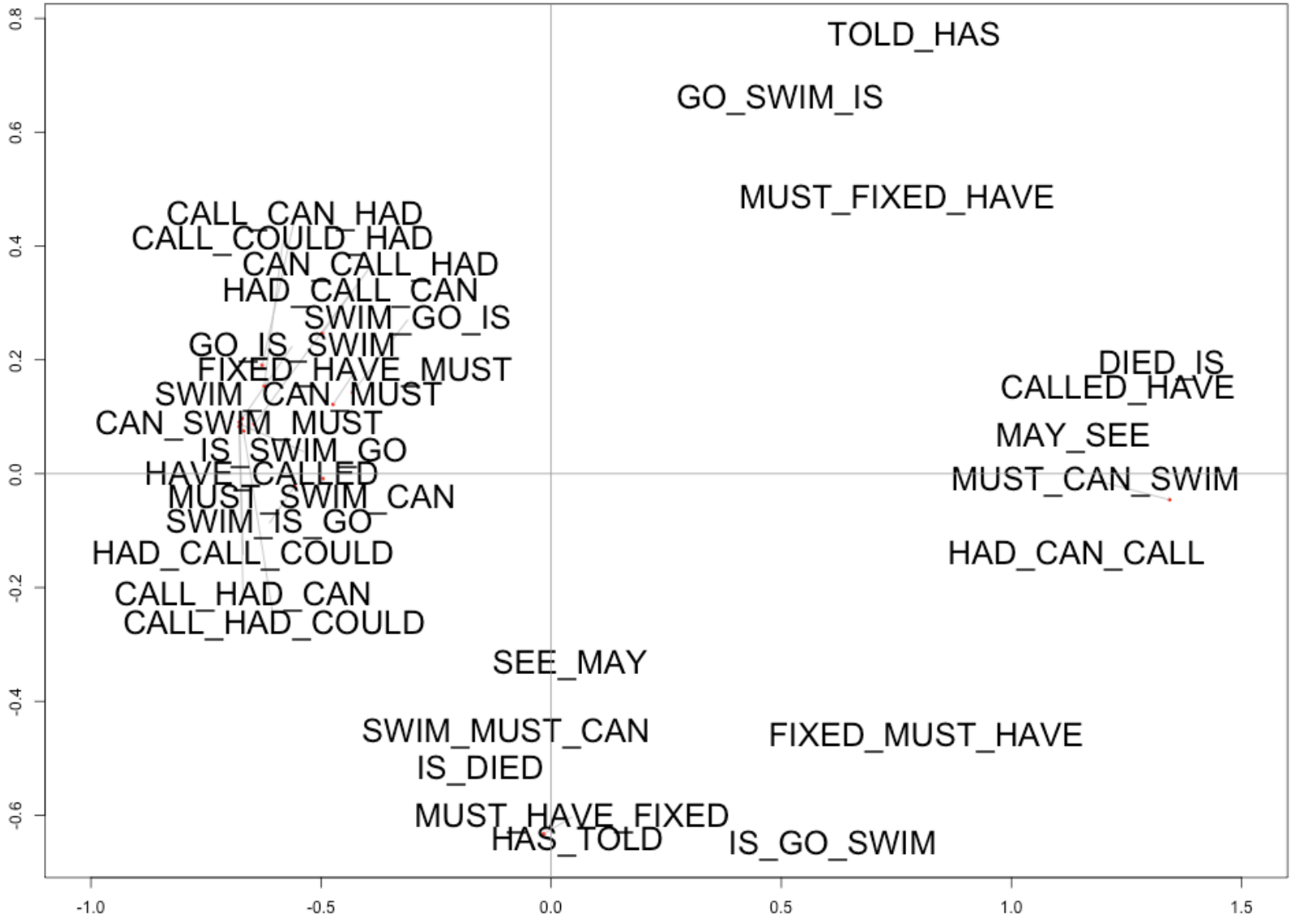
- the more dialects we look at and the more types of verb clusters we consider, the bigger (and messier) the variation becomes
- the data for this talk: 31 cluster orders from two- and three-verb clusters in 267 dialects of Dutch (Barbiers et al. 2008)

	Midsland	**Lies**	**W.-Terschelling**	**Oosterend**	…
IS_DIED	no	no	no	NA	…
DIED_IS	yes	yes	yes	NA	…
HAS_TOLD	no	no	no	no	…
TOLD_HAS	yes	yes	yes	yes	…
HAVE_CALLED	no	no	no	no	…
CALLED_HAVE	yes	yes	yes	yes	…
MAY_SEE	no	no	yes	no	…
SEE_MAY	yes	yes	yes	yes	…
CAN_SWIM_MUST	no	no	no	no	…
MUST_CAN_SWIM	no	no	no	yes	…
MUST_SWIM_CAN	yes	no	no	no	…
…	…	…	…	…	…

A quantitative-qualitative analysis

step #1 **Mutiple Correspondence Analysis**

- = an exploratory statistical technique for measuring the degree of correspondence between the rows and columns of a data table containing categorical data, and for visualizing those correspondences in a dimensionality that is smaller than that of the original data table
- applied to the case at hand: the 31 verb cluster orders are compared to one another based on their geographical distribution
- when two cluster orders typically co-occur in the same dialect locations, they are considered to be very similar; when they have a (near-)complementary distribution, they are very dissimilar
- this information can be represented on a two-dimensional graph

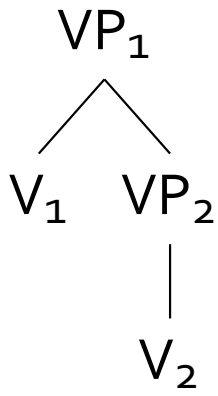

TOLD_HAS
GO_SWIM_IS
MUST_FIXED_HAVE
CALL_CAN_HAD
CALL_COULD_HAD
CAN_CALL_HAD
HAD_CALL_CAN
SWIM_GO_IS
GO_IS_SWIM
FIXED_HAVE_MUST
SWIM_CAN_MUST
CAN_SWIM_MUST
IS_SWIM_GO
HAVE_CALLED
MUST_SWIM_CAN
SWIM_IS_GO
HAD_CALL_COULD
CALL_HAD_CAN
CALL_HAD_COULD
DIED_IS
CALLED_HAVE
MAY_SEE
MUST_CAN_SWIM
HAD_CAN_CALL
SEE_MAY
SWIM_MUST_CAN
IS_DIED
FIXED_MUST_HAVE
MUST_HAVE_FIXED
HAS_TOLD
IS_GO_SWIM
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-1.0
-0.5
0.0
0.5
1.0
1.5

step #2 linguistic analyses as supplementary variables

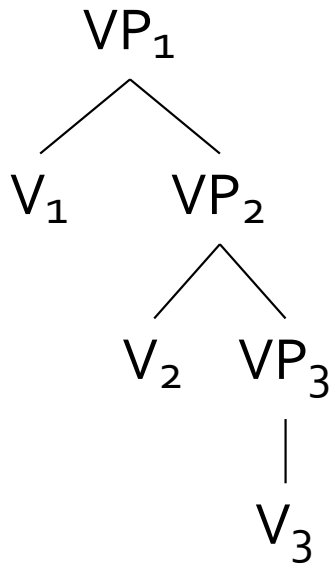
- supplementary variables are additional columns that are added to the data table
- they do not contribute to measuring the degree of correspondence between the rows (i.e. cluster orders), but can be used to interpret the data
- the supplementary variables used in this analysis are decomposed theoretical analyses of verb cluster orders
- example: Barbiers (2005)

- Barbiers (2005) derives verb cluster orders as follows:
  - base order is uniformly head-initial $\rightarrow$ derives 12 and 123

(6)

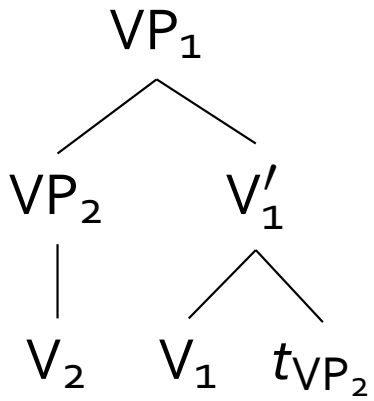


(7)

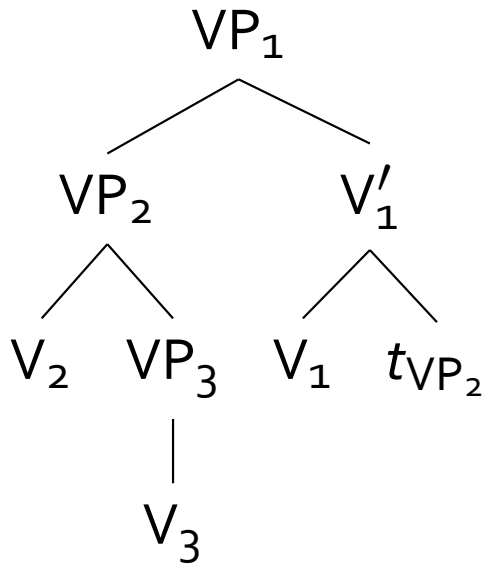


- Barbiers (2005) derives verb cluster orders as follows:
  - movement is VP-intraposition $\rightarrow$ derives 21 and 231, 312 and 132, and fails to derive 213

(8)

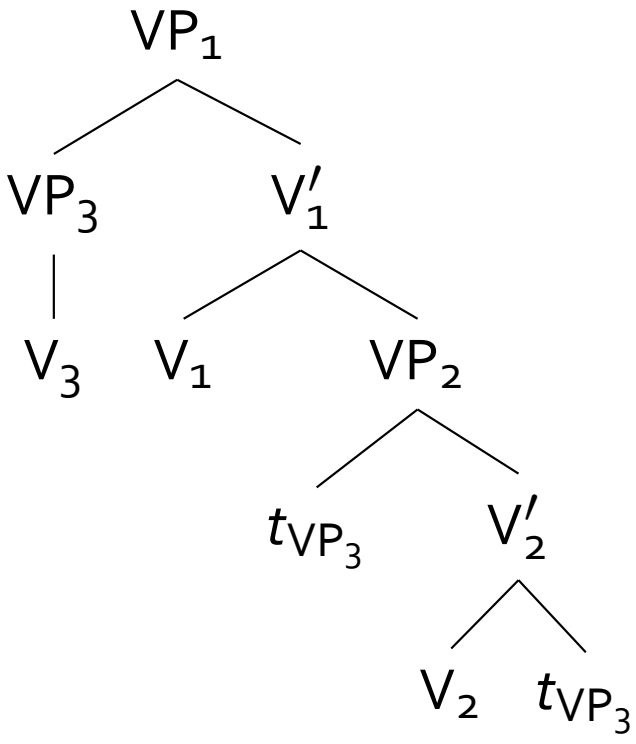


(9)

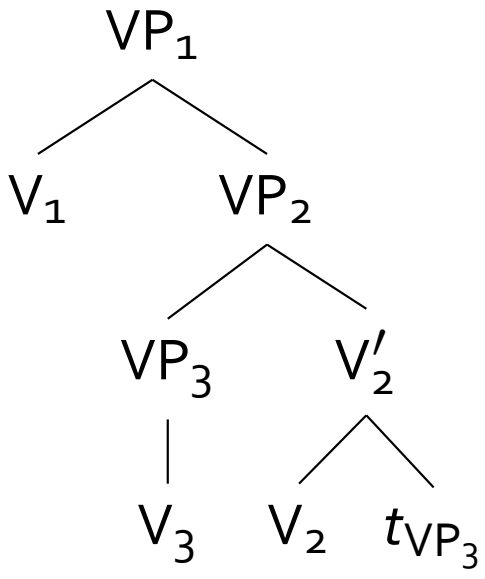


- Barbiers (2005) derives verb cluster orders as follows:
  - movement is VP-intraposition $\rightarrow$ derives 21 and 231, 312 and 132, and fails to derive 213

(10)

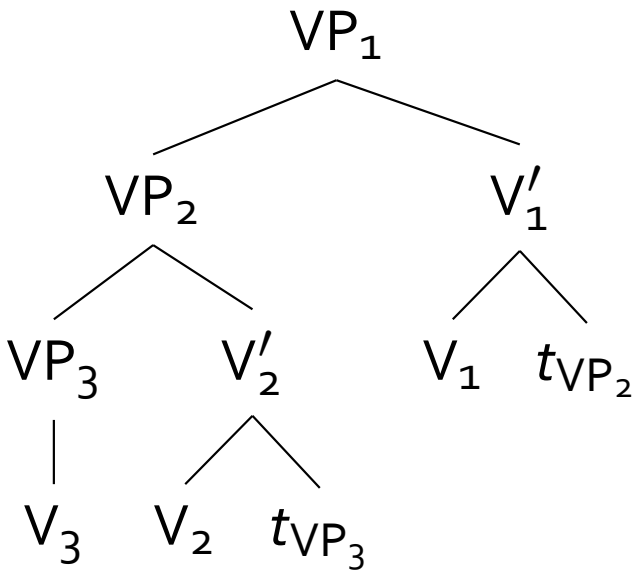


(11)



- Barbiers (2005) derives verb cluster orders as follows:
  - VP-intraposition can pied-pipe other material $\rightarrow$ derives 321 (movement of VP3 to specVP1 via specVP2 and with pied-piping of VP2)

(12)



- Barbiers (2005) derives verb cluster orders as follows:
  - VP intraposition is triggered by feature checking: modal and aspectual auxiliaries enter into a(n eventive) feature checking relation with the main verb, while perfective auxiliaries enter into a perfective checking relationship with their immediately selected verb $\rightarrow$ rules out 231 in the case of MOD-MOD/AUX-V-clusters and 312 in the case of AUX-AUX/MOD-V-clusters

(13) $[_{VP_1}\ \text{MUST}_{[u\text{Event}]}\ [_{VP_2}\ \text{CAN}_{[u\text{Event}]}\ [_{VP_3}\ \text{SWIM}_{[i\text{Event}]}\ ]]]$

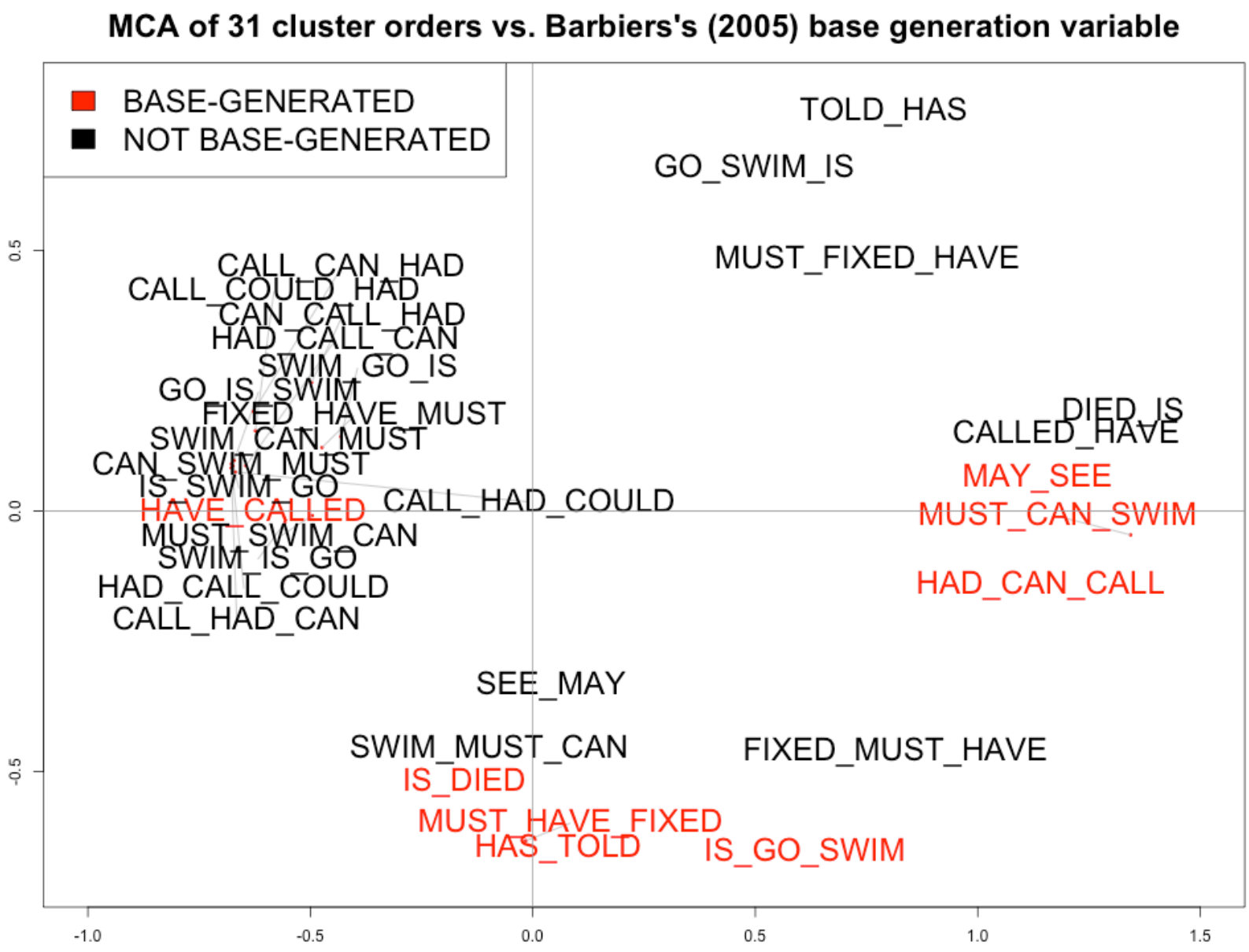
(14) $[_{VP_1}\ \text{HAD}_{[u\text{Perf}]}\ [_{VP_2}\ \text{CAN}_{[i\text{Perf},u\text{Event}]}\ [_{VP_3}\ \text{CALL}_{[i\text{Event}]}\ ]]]$

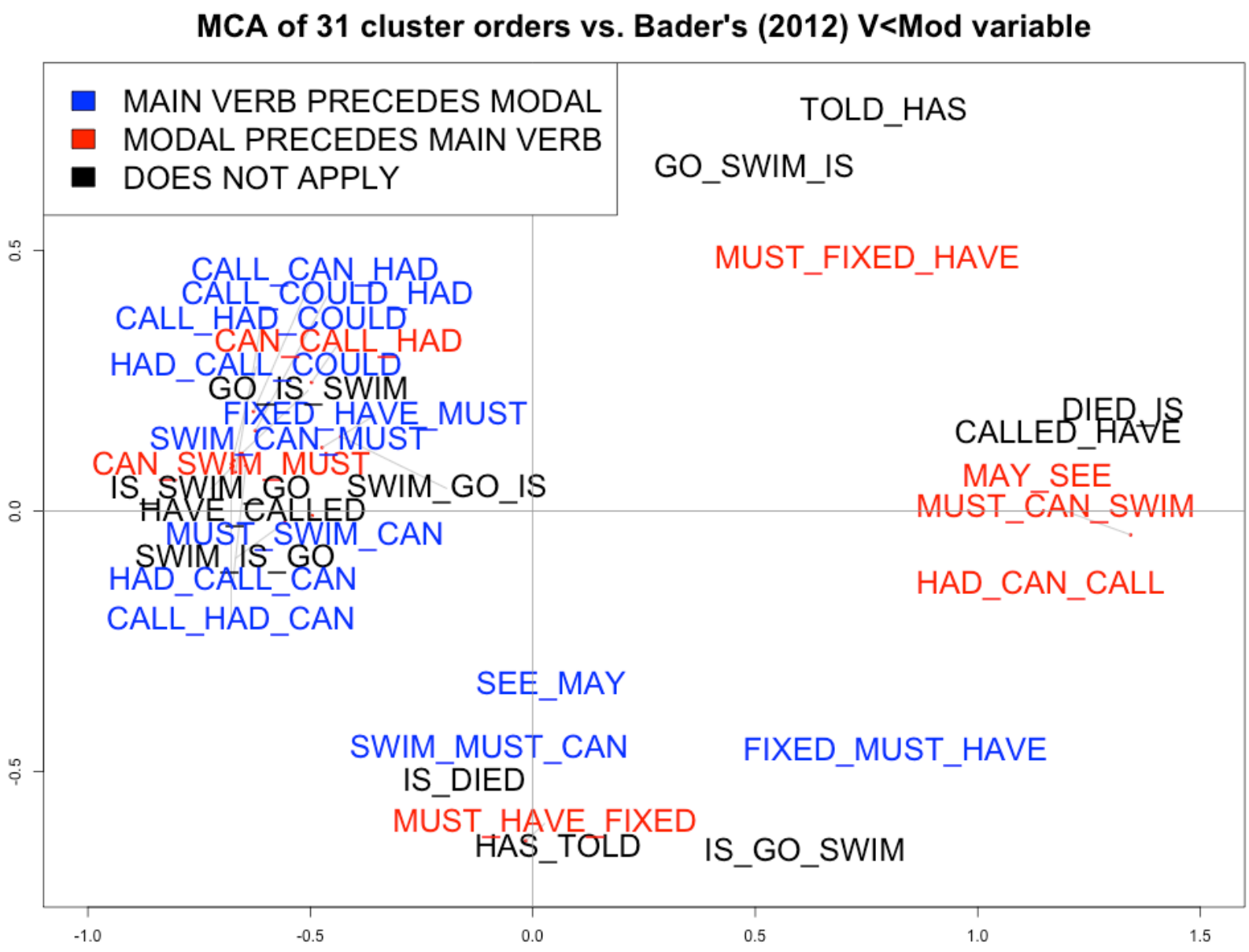
- from this theoretical account we can distill the following variables:
  - [BASE-GENERATION]: can the order be base-generated?
  - [MOVEMENT]: can the order be derived via movement?
  - [PIED-PIPING]: does the derivation involve pied-piping?
  - [FEATURE-CHECKING VIOLATION]: does the order involve a feature checking violation?

	BASE-GENERATION	**MOVEMENT**	**PIED-PIPING**	...
IS_DIED	yesBase	noMvt	noPiedP	...
DIED_IS	noBase	yesMvt	noPiedP	...
HAS_TOLD	yesBase	noMvt	noPiedP	...
TOLD_HAS	noBase	yesMvt	noPiedP	...
HAVE_CALLED	yesBase	noMvt	noPiedP	...
CALLED_HAVE	noBase	yesMvt	noPiedP	...
MAY_SEE	yesBase	noMvt	noPiedP	...
SEE_MAY	noBase	yesMvt	noPiedP	...
CAN_SWIM_MUST	noBase	yesMvt	noPiedP	...
MUST_CAN_SWIM	yesBase	noMvt	noPiedP	...
...	...	...	...	...

step #3 interpret the MCA-results using the linguistic variables

- the degree of correlation between a supplementary (i.e. linguistic) variable and a dimension of the MCA-plot can help to interpret that dimension and hence understand the underlying cause of variation in verb cluster ordering
- there are various ways of measuring/visualizing those correlations:
  - the plot can be color-coded according to specific variables


MCA of 31 cluster orders vs. Barbiers's (2005) base generation variable
BASE-GENERATED
NOT BASE-GENERATED
TOLD_HAS
GO_SWIM_IS
MUST_FIXED_HAVE
CALL_CAN_HAD
CALL_COULD_HAD
CAN_CALL_HAD
HAD_CALL_CAN
SWIM_GO_IS
GO_IS_SWIM
FIXED_HAVE_MUST
SWIM_CAN_MUST
CAN_SWIM_MUST
IS_SWIM_GO
HAVE_CALLED
CALL_HAD_COULD
MUST_SWIM_CAN
SWIM_IS_GO
HAD_CALL_COULD
CALL_HAD_CAN
DIED_IS
CALLED_HAVE
MAY_SEE
MUST_CAN_SWIM
HAD_CAN_CALL
SEE_MAY
SWIM_MUST_CAN
FIXED_MUST_HAVE
IS_DIED
MUST_HAVE_FIXED
HAS_TOLD
IS_GO_SWIM
0.5
0.0
-0.5
-1.0
-0.5
0.0
0.5
1.0
1.5

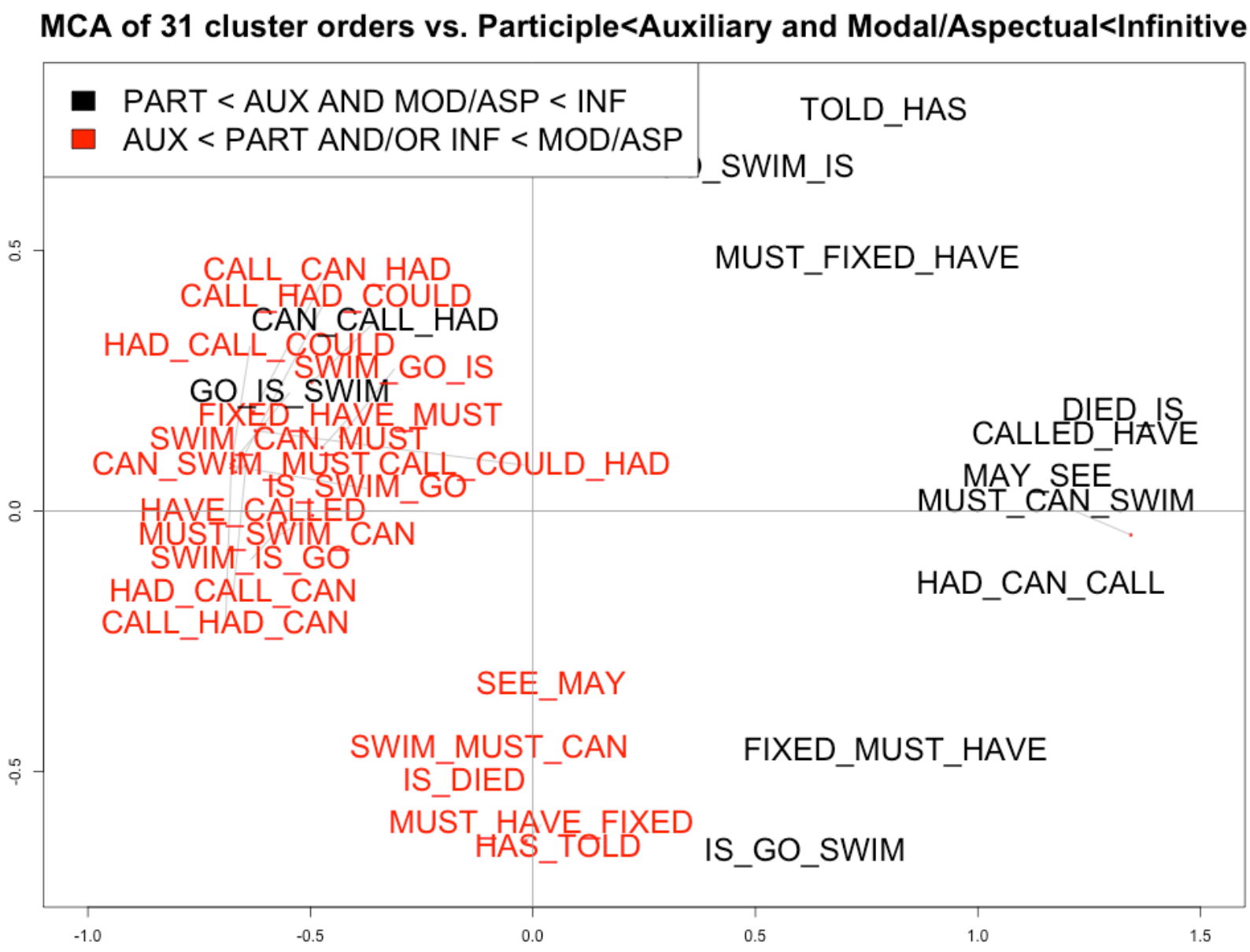

MCA of 31 cluster orders vs. Bader's (2012) V<Mod variable
MAIN VERB PRECEDES MODAL
MODAL PRECEDES MAIN VERB
DOES NOT APPLY
TOLD_HAS
GO_SWIM_IS
MUST_FIXED_HAVE
CALL_CAN_HAD
CALL_COULD_HAD
CALL_HAD_COULD
CAN_CALL_HAD
HAD_CALL_COULD
GO_IS_SWIM
FIXED_HAVE_MUST
SWIM_CAN_MUST
CAN_SWIM_MUST
IS_SWIM_GO
SWIM_GO_IS
HAVE_CALLED
MUST_SWIM_CAN
SWIM_IS_GO
HAD_CALL_CAN
CALL_HAD_CAN
DIED_IS
CALLED_HAVE
MAY_SEE
MUST_CAN_SWIM
HAD_CAN_CALL
SEE_MAY
SWIM_MUST_CAN
FIXED_MUST_HAVE
IS_DIED
MUST_HAVE_FIXED
HAS_TOLD
IS_GO_SWIM
0.5
0.0
-0.5
-1.0
-0.5
0.0
0.5
1.0
1.5

step #3 interpret the MCA-results using the linguistic variables

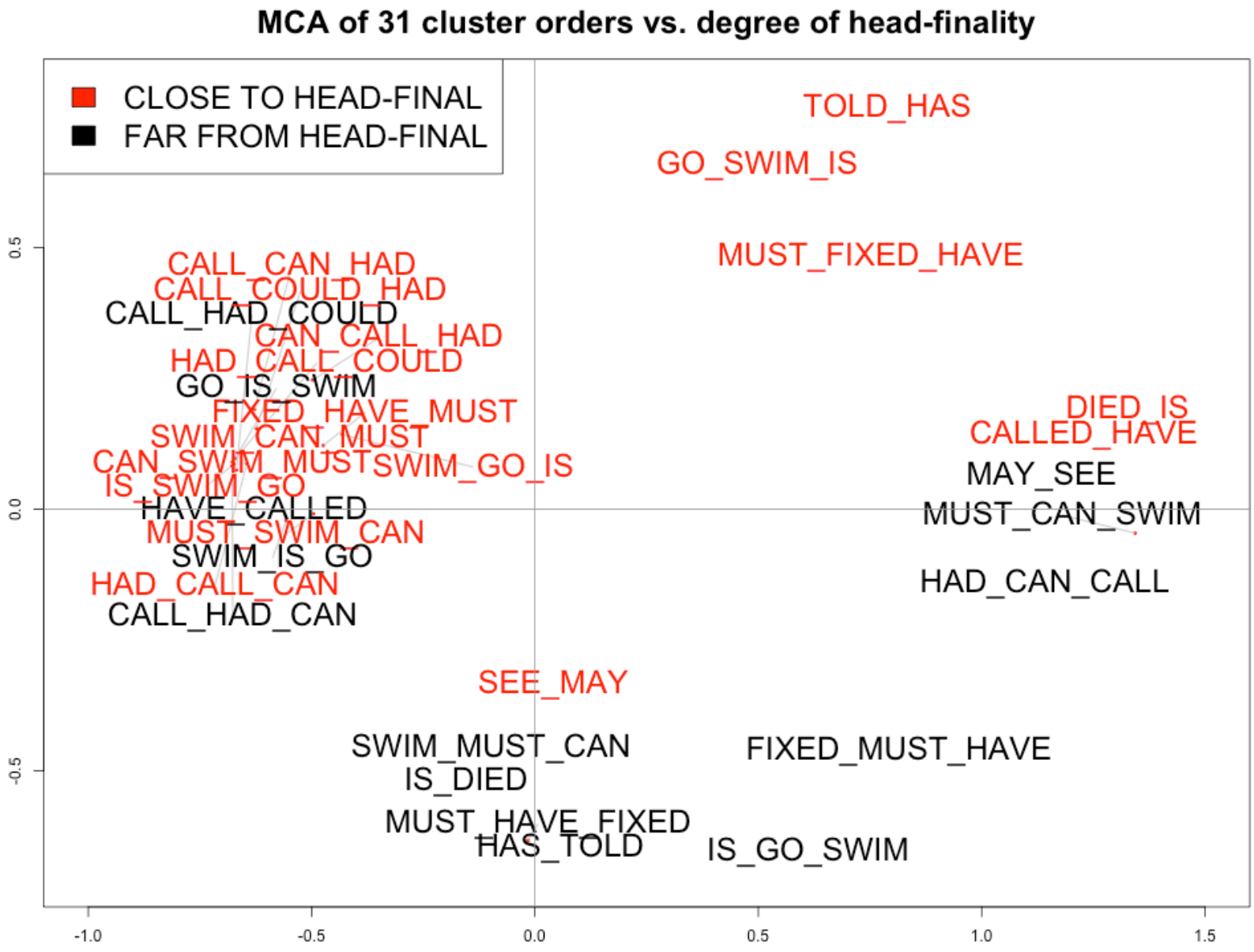
- the degree of correlation between a supplementary (i.e. linguistic) variable and a dimension of the MCA-plot can help to interpret that dimension and hence understand the underlying cause of variation in verb cluster ordering
- there are various ways of measuring/visualizing those correlations:
  - the plot can be color-coded according to specific variables
  - by calculating the squared correlation ratio (η^2):

	dimension 1	**dimension 2**
Barbiers (2005) base generation	0.126	0.309
Bader (2012) V<Mod	0.212	0.004

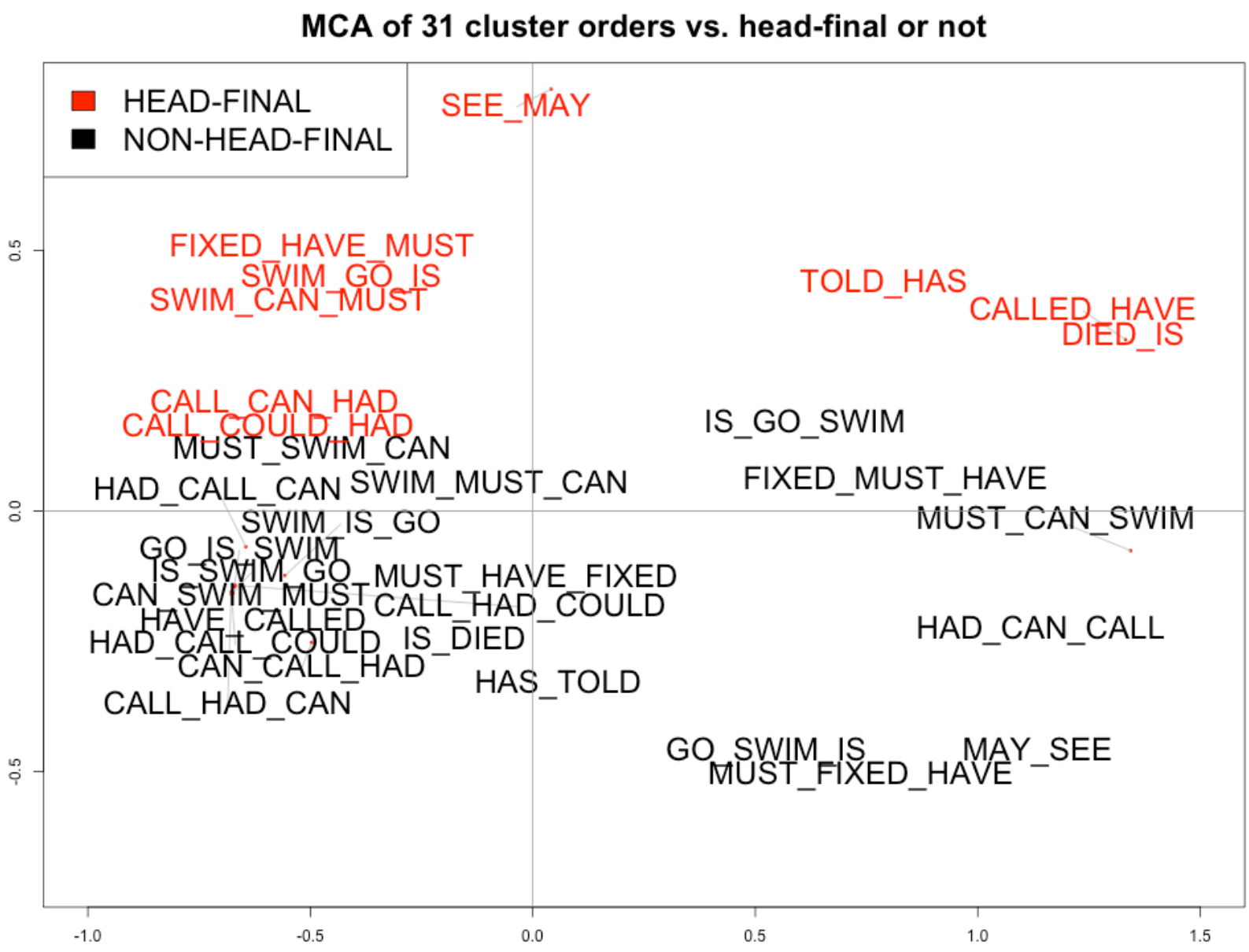
- using this methodology, we can arrive at an interpretation of the first three dimensions of the MCA-analysis, which together account for roughly 80% of the variance in the data set:
    1. **dimension #1:** sets apart dialects that consistently place infinitives to the right and participles to the left of their selecting verbs from those that don't


MCA of 31 cluster orders vs. Participle<Auxiliary and Modal/Aspectual<Infinitive
PART < AUX AND MOD/ASP < INF
AUX < PART AND/OR INF < MOD/ASP
TOLD_HAS
_SWIM_IS
MUST_FIXED_HAVE
CALL_CAN_HAD
CALL_HAD_COULD
CAN_CALL_HAD
HAD_CALL_COULD
SWIM_GO_IS
GO_IS_SWIM
FIXED_HAVE_MUST
SWIM_CAN_MUST
CAN_SWIM_MUST
CALL_COULD_HAD
IS_SWIM_GO
HAVE_CALLED
MUST_SWIM_CAN
SWIM_IS_GO
HAD_CALL_CAN
CALL_HAD_CAN
DIED_IS
CALLED_HAVE
MAY_SEE
MUST_CAN_SWIM
HAD_CAN_CALL
SEE_MAY
SWIM_MUST_CAN
IS_DIED
MUST_HAVE_FIXED
HAS_TOLD
FIXED_MUST_HAVE
IS_GO_SWIM
0.5
0.0
-0.5
-1.0
-0.5
0.0
0.5
1.0
1.5

- using this methodology, we can arrive at an interpretation of the first three dimensions of the MCA-analysis, which together account for roughly 80% of the variance in the data set:
    1. **dimension #1:** sets apart dialects that consistently place infinitives to the right and participles to the left of their selecting verbs from those that don't
    2. **dimension #2:** groups together cluster orders which are 0 or 1 movement operations away from a strictly head-final order (i.e. 132, 321, 231), from those that require at least two movement operations (123, 312, 213)


MCA of 31 cluster orders vs. degree of head-finality
CLOSE TO HEAD-FINAL
FAR FROM HEAD-FINAL
TOLD_HAS
GO_SWIM_IS
MUST_FIXED_HAVE
CALL_CAN_HAD
CALL_COULD_HAD
CALL_HAD_COULD
CAN_CALL_HAD
HAD_CALL_COULD
GO_IS_SWIM
FIXED_HAVE_MUST
SWIM_CAN_MUST
CAN_SWIM_MUST
SWIM_GO_IS
IS_SWIM_GO
HAVE_CALLED
MUST_SWIM_CAN
SWIM_IS_GO
HAD_CALL_CAN
CALL_HAD_CAN
DIED_IS
CALLED_HAVE
MAY_SEE
MUST_CAN_SWIM
HAD_CAN_CALL
SEE_MAY
SWIM_MUST_CAN
IS_DIED
FIXED_MUST_HAVE
MUST_HAVE_FIXED
HAS_TOLD
IS_GO_SWIM
0.5
0.0
-0.5
-1.0
-0.5
0.0
0.5
1.0
1.5

- using this methodology, we can arrive at an interpretation of the first three dimensions of the MCA-analysis, which together account for roughly 80% of the variance in the data set:
  1. **dimension #1:** sets apart dialects that consistently place infinitives to the right and participles to the left of their selecting verbs from those that don't
  2. **dimension #2:** groups together cluster orders which are 0 or 1 movement operations away from a strictly head-final order (i.e. 132, 321, 231), from those that require at least two movement operations (123, 312, 213)
  3. **dimension #3:** sets apart head-final from non-head-final cluster orders


MCA of 31 cluster orders vs. head-final or not
HEAD-FINAL
NON-HEAD-FINAL
SEE_MAY
FIXED_HAVE_MUST
SWIM_GO_IS
SWIM_CAN_MUST
TOLD_HAS
CALLED_HAVE
DIED_IS
CALL_CAN_HAD
CALL_COULD_HAD
IS_GO_SWIM
MUST_SWIM_CAN
HAD_CALL_CAN
SWIM_MUST_CAN
FIXED_MUST_HAVE
MUST_CAN_SWIM
SWIM_IS_GO
GO_IS_SWIM
IS_SWIM_GO
MUST_HAVE_FIXED
CAN_SWIM_MUST
CALL_HAD_COULD
HAVE_CALLED
HAD_CAN_CALL
HAD_CALL_COULD
IS_DIED
CAN_CALL_HAD
HAS_TOLD
CALL_HAD_CAN
GO_SWIM_IS
MAY_SEE
MUST_FIXED_HAVE
0.5
0.0
-0.5
-1.0
-0.5
0.0
0.5
1.0
1.5

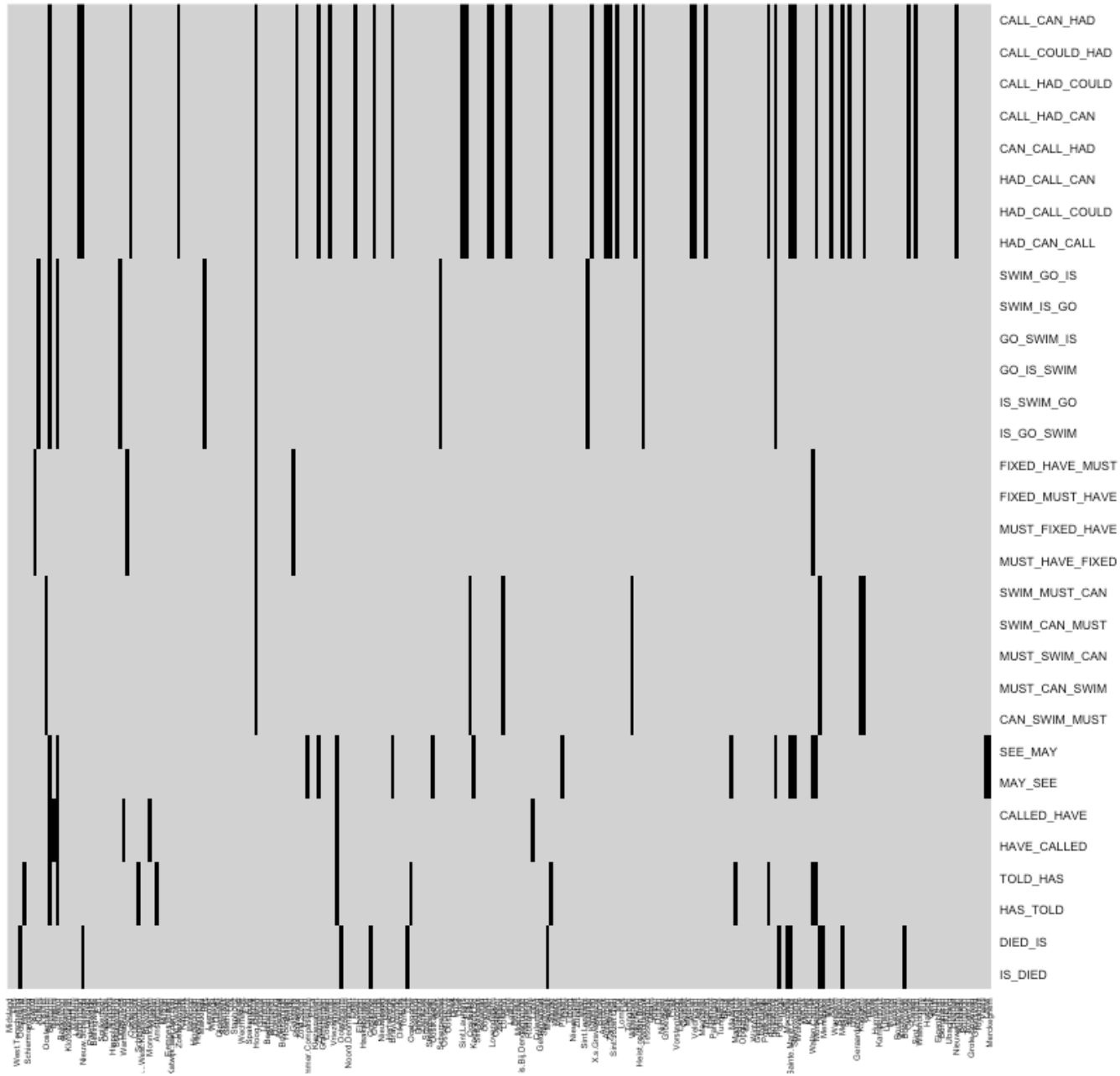
- using this methodology, we can arrive at an interpretation of the first three dimensions of the MCA-analysis, which together account for roughly 80% of the variance in the data set:
  1. **dimension #1:** sets apart dialects that consistently place infinitives to the right and participles to the left of their selecting verbs from those that don't
  2. **dimension #2:** groups together cluster orders which are 0 or 1 movement operations away from a strictly head-final order (i.e. 132, 321, 231), from those that require at least two movement operations (123, 312, 213)
  3. **dimension #3:** sets apart head-final from non-head-final cluster orders
- these interpretations can now be used to construct a (rough) parametric account of verb cluster ordering:
  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: [±Participle<Auxiliary and Modal/Aspectual<Infinitive]

Four data complications

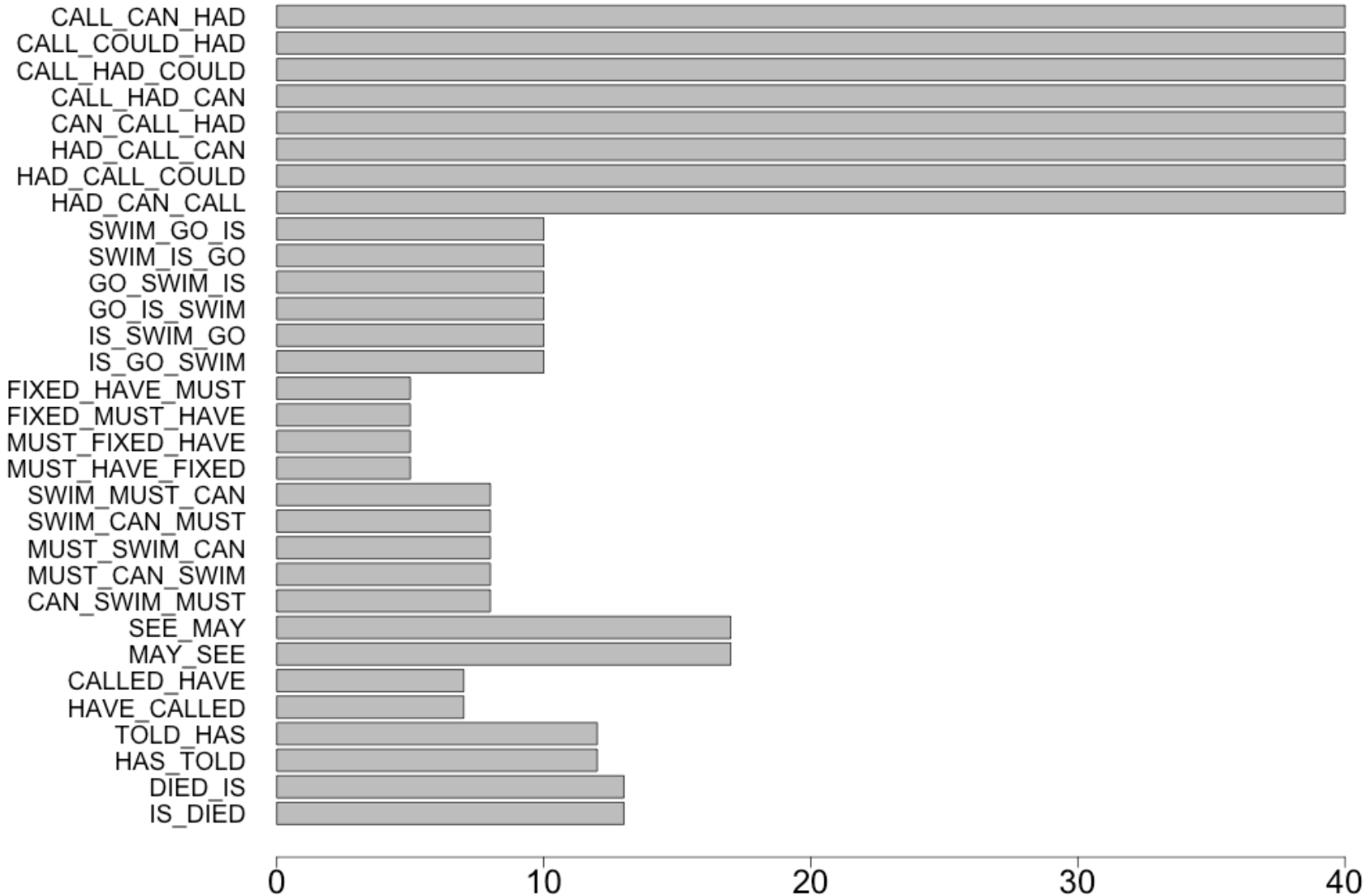
- the account as presented so far glosses over (or has made certain methodological choices concerning) four complications in the data:
  - **missing data:** for some combinations of cluster order and dialect location information about the (non-)occurrence of that cluster order in that dialect location is lacking
  - **different question types:** some cluster orders were presented in Standard Dutch in one order as translation questions, others were presented in the dialect in all possible orders as judgment tasks
  - **mixed data:** the data set contains both Dutch and Frisian dialects, but these groups are known to have very different verb cluster systems (strongly head-initial in Dutch, strongly head-final in Frisian)
  - **differences in frequency:** some cluster orders are highly frequent, while others are exceedingly rare

Missing values

- the data table contains $31 \times 267 = 8,277$ cells
- of those 8,277 cells, 538 (6.49%) are empty (NA)
- question: how are they distributed over the data?


CALL_CAN_HAD
CALL_COULD_HAD
CALL_HAD_COULD
CALL_HAD_CAN
CAN_CALL_HAD
HAD_CALL_CAN
HAD_CALL_COULD
HAD_CAN_CALL
SWIM_GO_IS
SWIM_IS_GO
GO_SWIM_IS
GO_IS_SWIM
IS_SWIM_GO
IS_GO_SWIM
FIXED_HAVE_MUST
FIXED_MUST_HAVE
MUST_FIXED_HAVE
MUST_HAVE_FIXED
SWIM_MUST_CAN
SWIM_CAN_MUST
MUST_SWIM_CAN
MUST_CAN_SWIM
CAN_SWIM_MUST
SEE_MAY
MAY_SEE
CALLED_HAVE
HAVE_CALLED
TOLD_HAS
HAS_TOLD
DIED_IS
IS_DIED

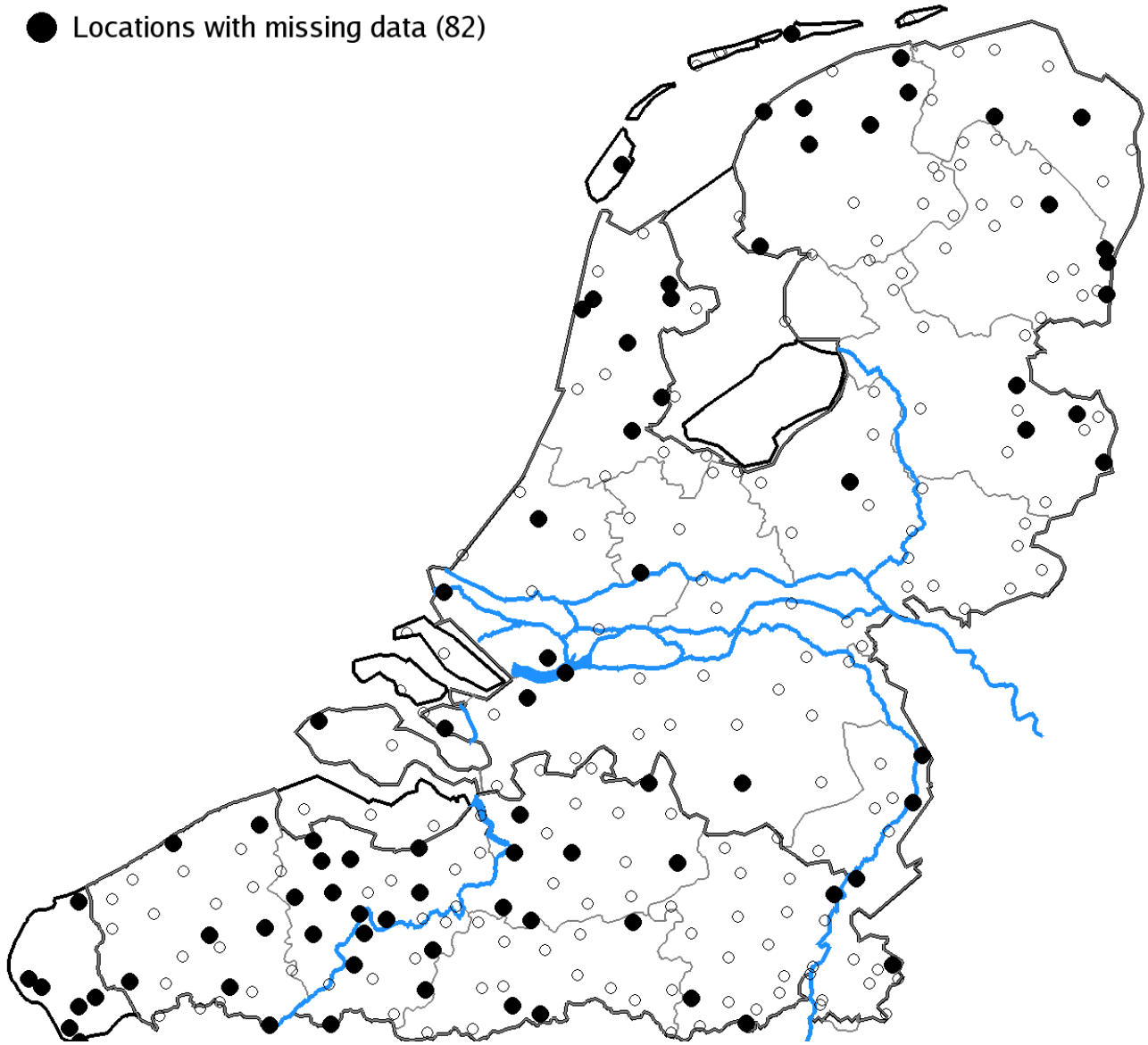
- the distribution of these NA's is not random across cluster types:
  - if a value for one order in a cluster is missing, the values for all other orders of that same cluster are also missing
  - some cluster orders are missing much more frequently than others


CALL_CAN_HAD
CALL_COULD_HAD
CALL_HAD_COULD
CALL_HAD_CAN
CAN_CALL_HAD
HAD_CALL_CAN
HAD_CALL_COULD
HAD_CAN_CALL
SWIM_GO_IS
SWIM_IS_GO
GO_SWIM_IS
GO_IS_SWIM
IS_SWIM_GO
IS_GO_SWIM
FIXED_HAVE_MUST
FIXED_MUST_HAVE
MUST_FIXED_HAVE
MUST_HAVE_FIXED
SWIM_MUST_CAN
SWIM_CAN_MUST
MUST_SWIM_CAN
MUST_CAN_SWIM
CAN_SWIM_MUST
SEE_MAY
MAY_SEE
CALLED_HAVE
HAVE_CALLED
TOLD_HAS
HAS_TOLD
DIED_IS
IS_DIED
0
10
20
30
40

(15) a. Vertaal: Vertel maar niet wie zij had kunnen roepen.
translate tell PRT not who she had can call
'Translate: Don't tell me who she would have been able to call.'

b. Vertel maar nie wie dassezij zou geropn en.
tell PRT not who that.she.she would called have
'Don't tell me who she would have called.' (Eeklo)

- the distribution of these NA's does appear to be fairly random across dialect locations:



- methodology for dealing with NAs in my analysis: **imputation of missing data**
  - regularized iterative multiple correspondence analysis (Josse et al. (2012), R-package `missMDA`)
  - fill in the NAs through an iterative procedure such that the newly filled in values have as little an effect as possible on the relations between the non-missing values

	Midsland	**Lies**	**Oosterend**	**Hollum** ...	
IS_DIED	no	no	NA	no	...
DIED_IS	yes	yes	NA	yes	...
HAS_TOLD	no	no	no	NA	...
TOLD_HAS	yes	yes	yes	NA	...
HAVE_CALLED	no	no	no	no	...
CALLED_HAVE	yes	yes	yes	yes	...
MAY_SEE	no	no	no	no	...
SEE_MAY	yes	yes	yes	yes	...
CAN_SWIM_MUST	no	no	no	no	...
...	...	...	...	...	...

- methodology for dealing with NAs in my analysis: **imputation of missing data**
  - regularized iterative multiple correspondence analysis (Josse et al. (2012), R-package `missMDA`)
  - fill in the NAs through an iterative procedure such that the newly filled in values have as little an effect as possible on the relations between the non-missing values

	Midsland	**Lies**	**Oosterend**	**Hollum** …	
IS_DIED	1	1	NA	1	…
DIED_IS	0	0	NA	0	…
HAS_TOLD	1	1	1	NA	…
TOLD_HAS	0	0	0	NA	…
HAVE_CALLED	1	1	1	1	…
CALLED_HAVE	0	0	0	0	…
MAY_SEE	1	1	1	1	…
SEE_MAY	0	0	0	0	…
CAN_SWIM_MUST	1	1	1	1	…
…	…	…	…	…	…

- methodology for dealing with NAs in my analysis: **imputation of missing data**
  - regularized iterative multiple correspondence analysis (Josse et al. (2012), R-package `missMDA`)
  - fill in the NAs through an iterative procedure such that the newly filled in values have as little an effect as possible on the relations between the non-missing values

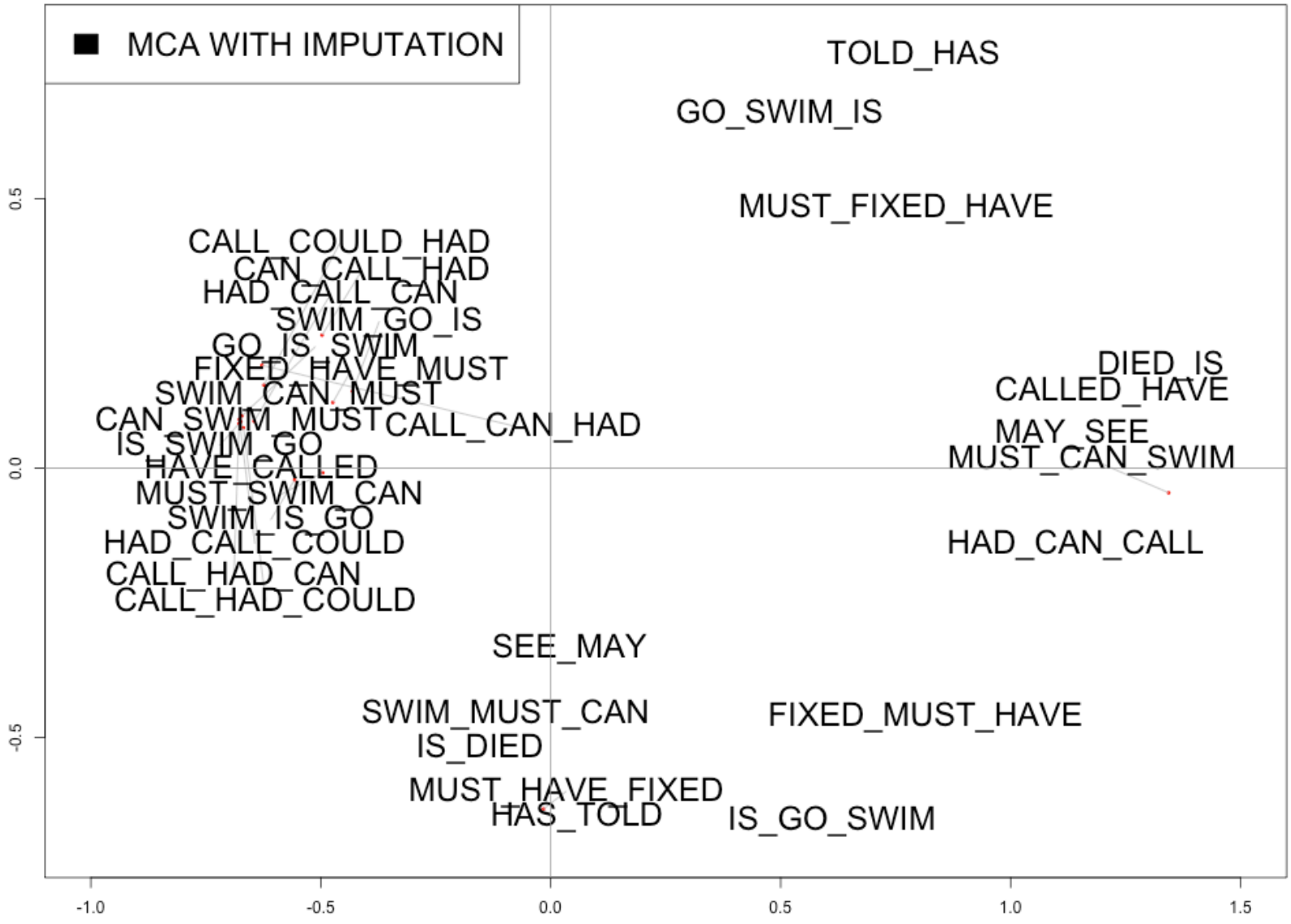
	Midsland	**Lies**	**Oosterend**	**Hollum** …	
IS_DIED	1	1	0.816	1	…
DIED_IS	0	0	0.088	0	…
HAS_TOLD	1	1	1	0.947	…
TOLD_HAS	0	0	0	0.255	…
HAVE_CALLED	1	1	1	1	…
CALLED_HAVE	0	0	0	0	…
MAY_SEE	1	1	1	1	…
SEE_MAY	0	0	0	0	…
CAN_SWIM_MUST	1	1	1	1	…
…	…	…	…	…	…

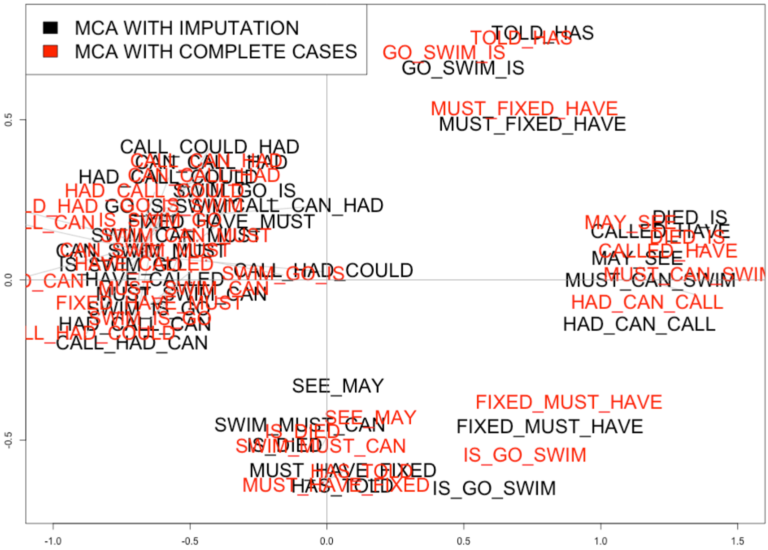
- methodology for dealing with NAs in my analysis: **imputation of missing data**
  - regularized iterative multiple correspondence analysis (Josse et al. (2012), R-package `missMDA`)
  - fill in the NAs through an iterative procedure such that the newly filled in values have as little an effect as possible on the relations between the non-missing values

	Midsland	**Lies**	**Oosterend**	**Hollum** …	
IS_DIED	1	1	1	1	…
DIED_IS	0	0	0	0	…
HAS_TOLD	1	1	1	1	…
TOLD_HAS	0	0	0	0	…
HAVE_CALLED	1	1	1	1	…
CALLED_HAVE	0	0	0	0	…
MAY_SEE	1	1	1	1	…
SEE_MAY	0	0	0	0	…
CAN_SWIM_MUST	1	1	1	1	…
…	…	…	…	…	…

- methodology for dealing with NAs in my analysis: **imputation of missing data**
  - regularized iterative multiple correspondence analysis (Josse et al. (2012), R-package `missMDA`)
  - fill in the NAs through an iterative procedure such that the newly filled in values have as little an effect as possible on the relations between the non-missing values

	Midsland	**Lies**	**Oosterend**	**Hollum** ...	
IS_DIED	no	no	no	no	...
DIED_IS	yes	yes	yes	yes	...
HAS_TOLD	no	no	no	no	...
TOLD_HAS	yes	yes	yes	yes	...
HAVE_CALLED	no	no	no	no	...
CALLED_HAVE	yes	yes	yes	yes	...
MAY_SEE	no	no	no	no	...
SEE_MAY	yes	yes	yes	yes	...
CAN_SWIM_MUST	no	no	no	no	...
...	...	...	...	...	...


MCA WITH IMPUTATION
TOLD_HAS
GO_SWIM_IS
MUST_FIXED_HAVE
CALL_COULD_HAD
CAN_CALL_HAD
HAD_CALL_CAN
SWIM_GO_IS
GO_IS_SWIM
FIXED_HAVE_MUST
SWIM_CAN_MUST
CAN_SWIM_MUST
IS_SWIM_GO
HAVE_CALLED
MUST_SWIM_CAN
SWIM_IS_GO
HAD_CALL_COULD
CALL_HAD_CAN
CALL_HAD_COULD
CALL_CAN_HAD
DIED_IS
CALLED_HAVE
MAY_SEE
MUST_CAN_SWIM
HAD_CAN_CALL
SEE_MAY
SWIM_MUST_CAN
IS_DIED
FIXED_MUST_HAVE
MUST_HAVE_FIXED
HAS_TOLD
IS_GO_SWIM
0.5
0.0
-0.5
-1.0
-0.5
0.0
0.5
1.0
1.5


MCA WITH IMPUTATION
MCA WITH COMPLETE CASES
TOLD_HAS
GO_SWIM_IS
MUST_FIXED_HAVE
CALL_COULD_HAD
CALL_CAN_HAD
FIXED_HAVE_MUST
CALL_HAD_COULD
HAVE_CALLED
MUST_SWIM_CAN
HAD_CALL_CAN
CALL_HAD_CAN
DIED_IS
CALLED_HAVE
MAY_SEE
MUST_CAN_SWIM
HAD_CAN_CALL
SEE_MAY
FIXED_MUST_HAVE
SWIM_MUST_CAN
IS_DIED
IS_GO_SWIM
MUST_HAVE_FIXED
HAS_TOLD
0.5
0.0
-0.5
-1.0
-0.5
0.0
0.5
1.0
1.5

Missing values: conclusion

- one way of dealing with missing data is through data imputation
- risk/downside: this effectively amounts to filling in values for linguistic variables (cluster orders) that were not originally recorded
- a comparison with a complete case analysis (i.e. deletion of dialect locations containing missing values) has revealed that the effect of data imputation on the overall analysis has been minimal

Differences in question type

- the methodology of the SAND-interviews was of two types when it came to verb cluster ordering:
  1. translation questions

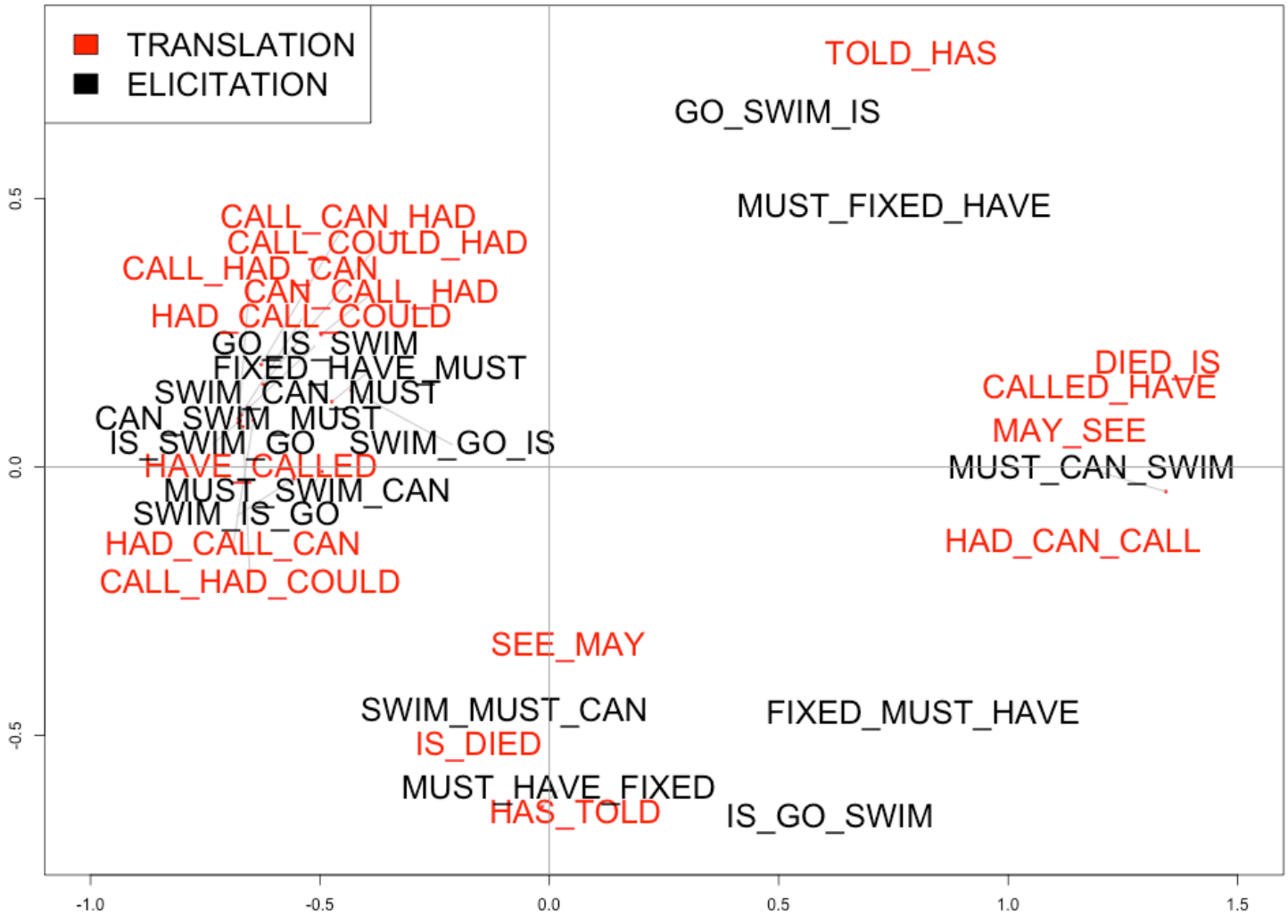
     (16) Vertaal: Ze weet niet dat Marie gisteren
     translate she knows not that Marie yesterday
     gestorven is.
     died is
     'Translate: She doesn't know that Mary died yesterday.'

  2. elicitation questions

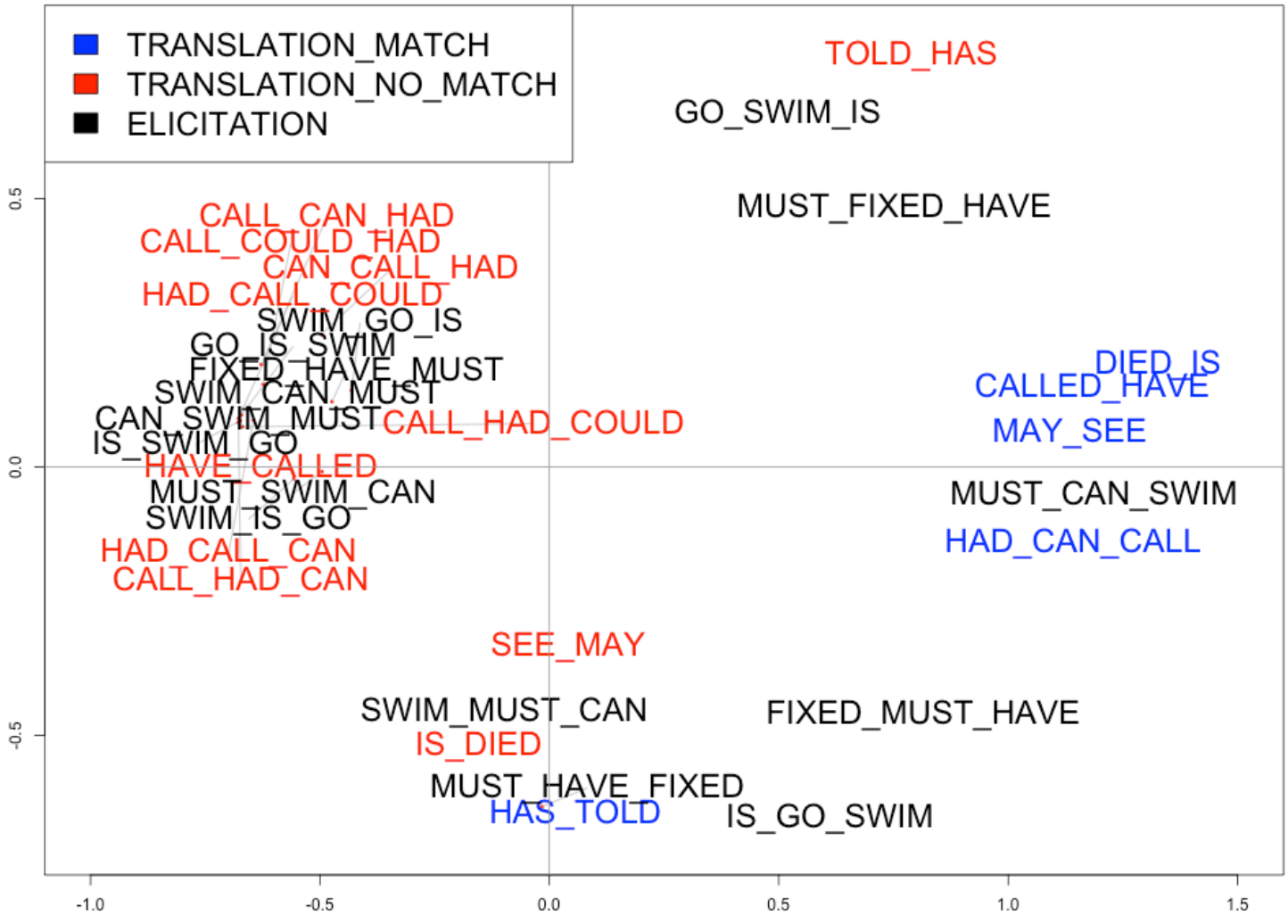
     (17) Komt deze zin voor in uw dialect?
     comes this sentences for in your dialect
     'Does this sentence occur in your dialect?'

Ik vind dat iedereen moet kunnen zwemmen.	yes/no
Ik vind dat iedereen moet zwemmen kunnen.	yes/no
Ik vind dat iedereen kunnen zwemmen moet.	yes/no
Ik vind dat iedereen zwemmen kunnen moet.	yes/no
Ik vind dat iedereen zwemmen moet kunnen.	yes/no

- question: to what extent has the difference in question methodology influenced the results?



- question: to what extent has the difference in question methodology influenced the results?

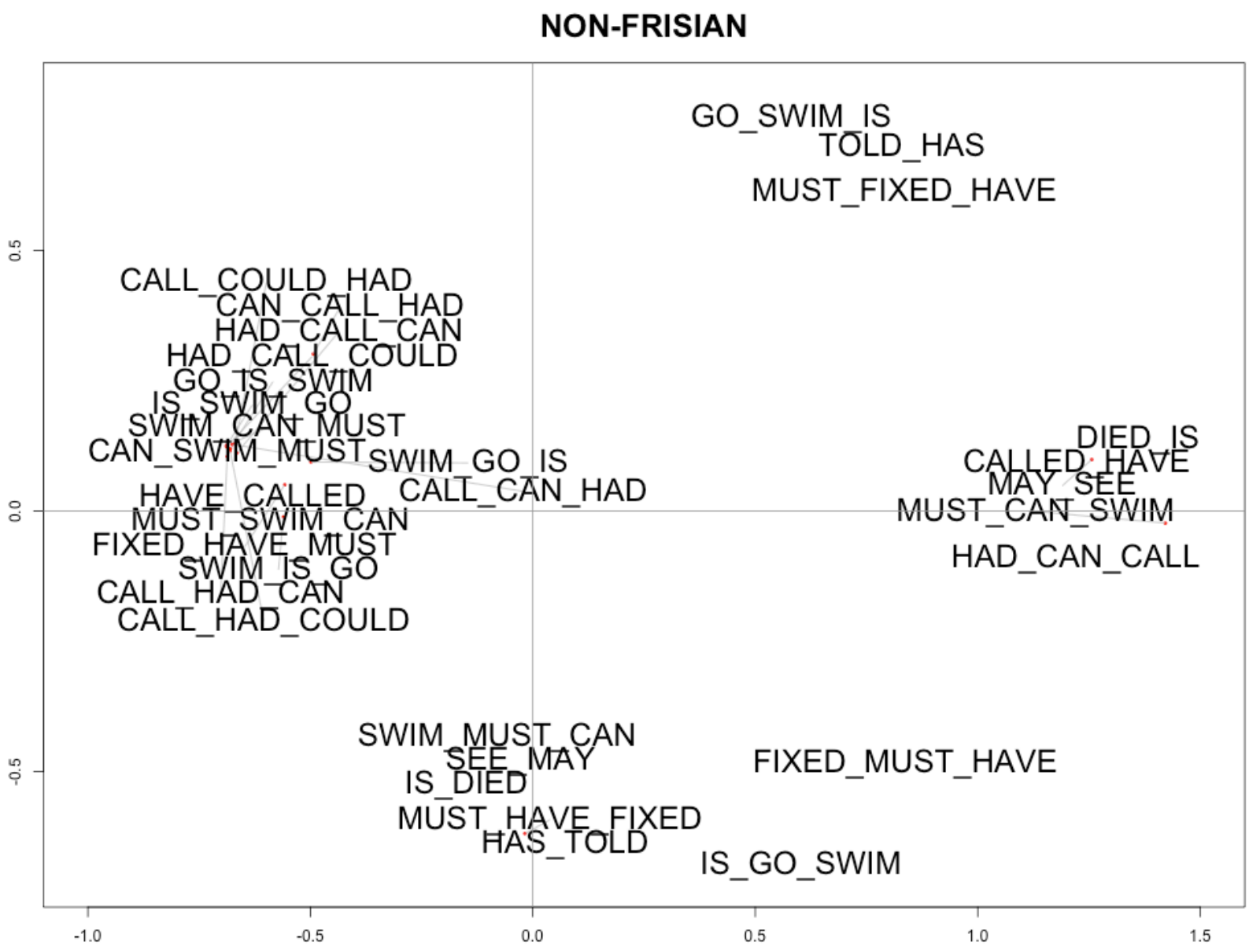


Differences in question type: conclusion

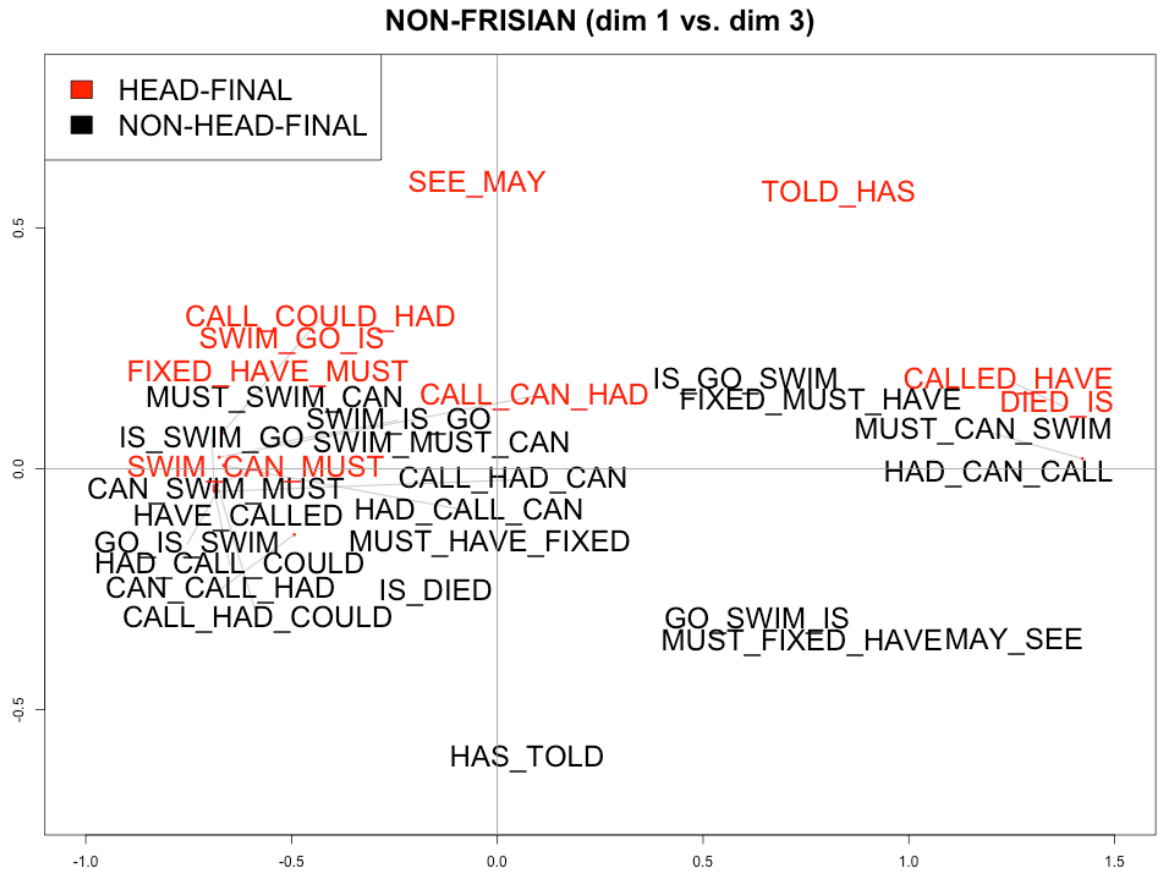
- rather than being an embarrassment for the analysis, the differences in question methodology prove to be additional sources of linguistically relevant information
- non-matching translations and elicitation questions tend to cluster together $\rightarrow$ they provide the strongest positive judgment by the native speaker
- from the perspective of question types, we can distinguish four patterns:
  1. a basic ascending pattern, except that it places the participle before the auxiliary
  2. a distinctly southern pattern including a 231-order in IPP-contexts, cluster interruption and a strong preference for participle preceding auxiliary
  3. a distinctly Netherlandic pattern, with a less strong preference to place the participle before the auxiliary, and with the possibility of placing the infinitive before the modal
  4. a rest category containing (among others) the typically Frisian orders

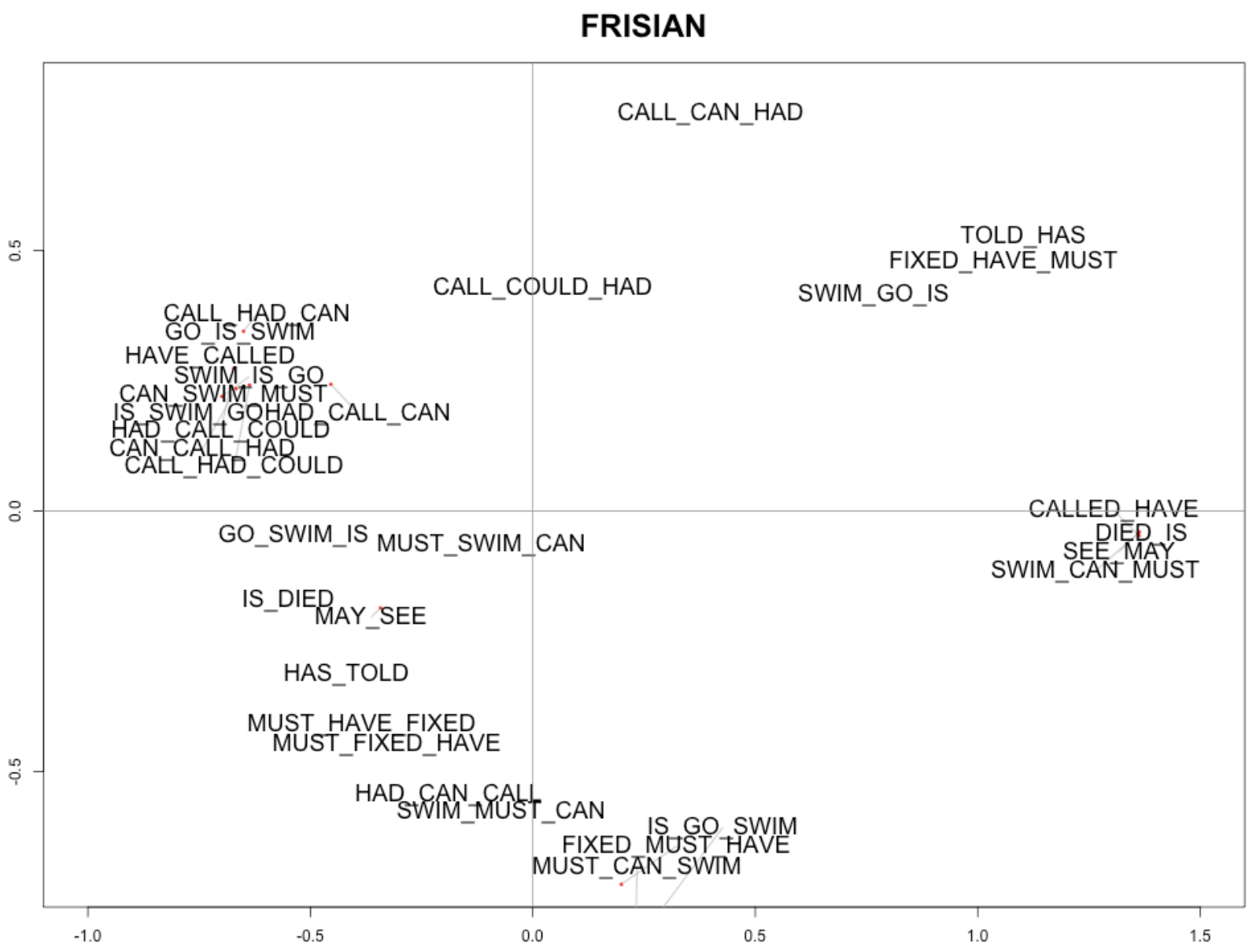
Mixed data

- the data set contains data from both Dutch (dialects) and Frisian (dialects)
- Barbiers et al. (2016): Dutch (dialects) and Frisian (dialects) differ in their base-generated verb cluster order: ascending (12 & 123) in Dutch, descending (21 & 321) in Frisian
- question: to what extent has this mixing of two types of data influenced our interpretation of the MCA-results (cf. in particular the third dimension)?
- let's split up the data set into a Frisian and a non-Frisian subset and redo the analysis

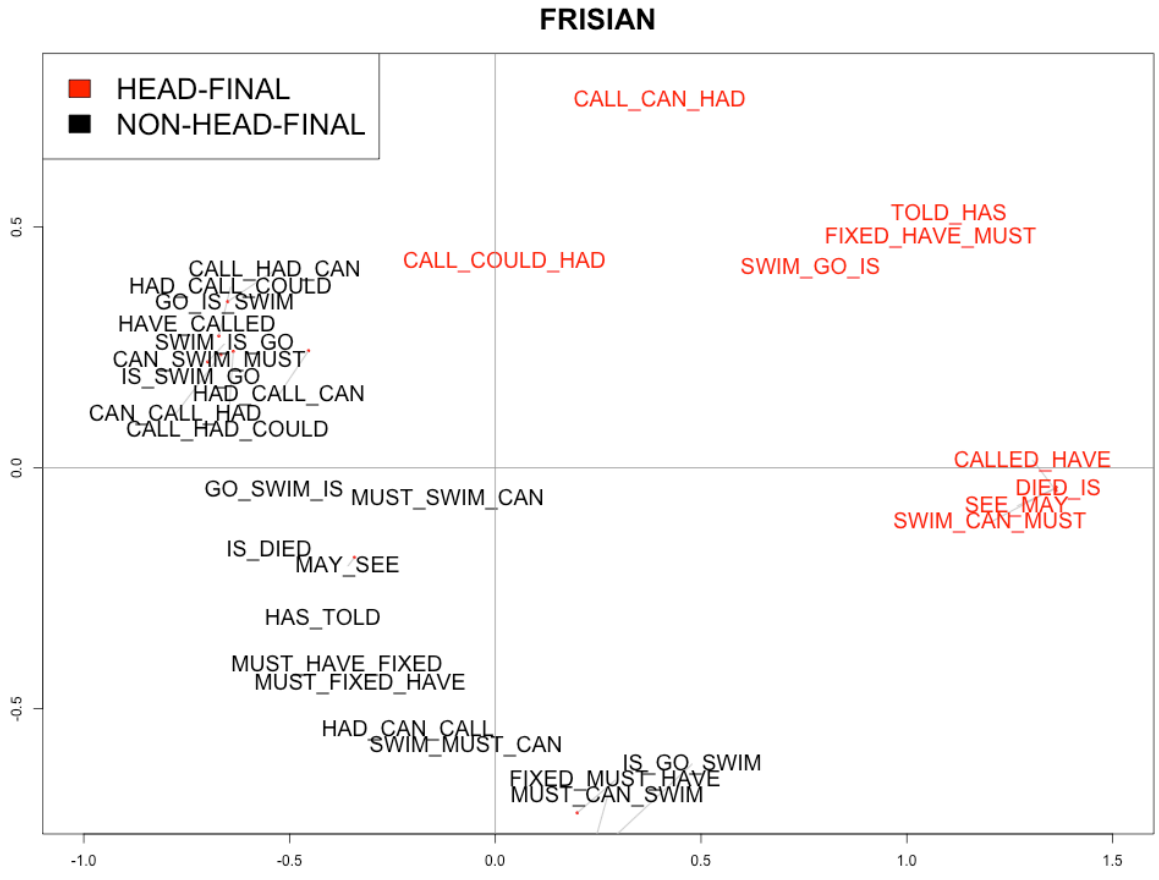

NON-FRISIAN
GO_SWIM_IS
TOLD_HAS
MUST_FIXED_HAVE
CALL_COULD_HAD
CAN_CALL_HAD
HAD_CALL_CAN
HAD_CALL_COULD
GO_IS_SWIM
IS_SWIM_GO
SWIM_CAN_MUST
CAN_SWIM_MUST
SWIM_GO_IS
CALL_CAN_HAD
HAVE_CALLED
MUST_SWIM_CAN
FIXED_HAVE_MUST
SWIM_IS_GO
CALL_HAD_CAN
CALL_HAD_COULD
DIED_IS
CALLED_HAVE
MAY_SEE
MUST_CAN_SWIM
HAD_CAN_CALL
SWIM_MUST_CAN
SEE_MAY
IS_DIED
MUST_HAVE_FIXED
HAS_TOLD
FIXED_MUST_HAVE
IS_GO_SWIM
0.5
0.0
-0.5
-1.0
-0.5
0.0
0.5
1.0
1.5

- the first two dimensions of the analysis are virtually unchanged
- the third dimension, however, no longer contains a strong effect of head-finality (and more generally, explains a smaller percentage of the overall variance)




FRISIAN
CALL_CAN_HAD
TOLD_HAS
FIXED_HAVE_MUST
CALL_COULD_HAD
SWIM_GO_IS
CALL_HAD_CAN
GO_IS_SWIM
HAVE_CALLED
SWIM_IS_GO
CAN_SWIM_MUST
IS_SWIM_GOHAD_CALL_CAN
HAD_CALL_COULD
CAN_CALL_HAD
CALL_HAD_COULD
CALLED_HAVE
DIED_IS
SEE_MAY
SWIM_CAN_MUST
GO_SWIM_IS
MUST_SWIM_CAN
IS_DIED
MAY_SEE
HAS_TOLD
MUST_HAVE_FIXED
MUST_FIXED_HAVE
HAD_CAN_CALL
SWIM_MUST_CAN
IS_GO_SWIM
FIXED_MUST_HAVE
MUST_CAN_SWIM
0.5
0.0
-0.5
-1.0
-0.5
0.0
0.5
1.0
1.5

- the first two dimensions of the Frisian analysis look very different
- there is a very strong effect of head-finality on the first (or first and second combined) dimension

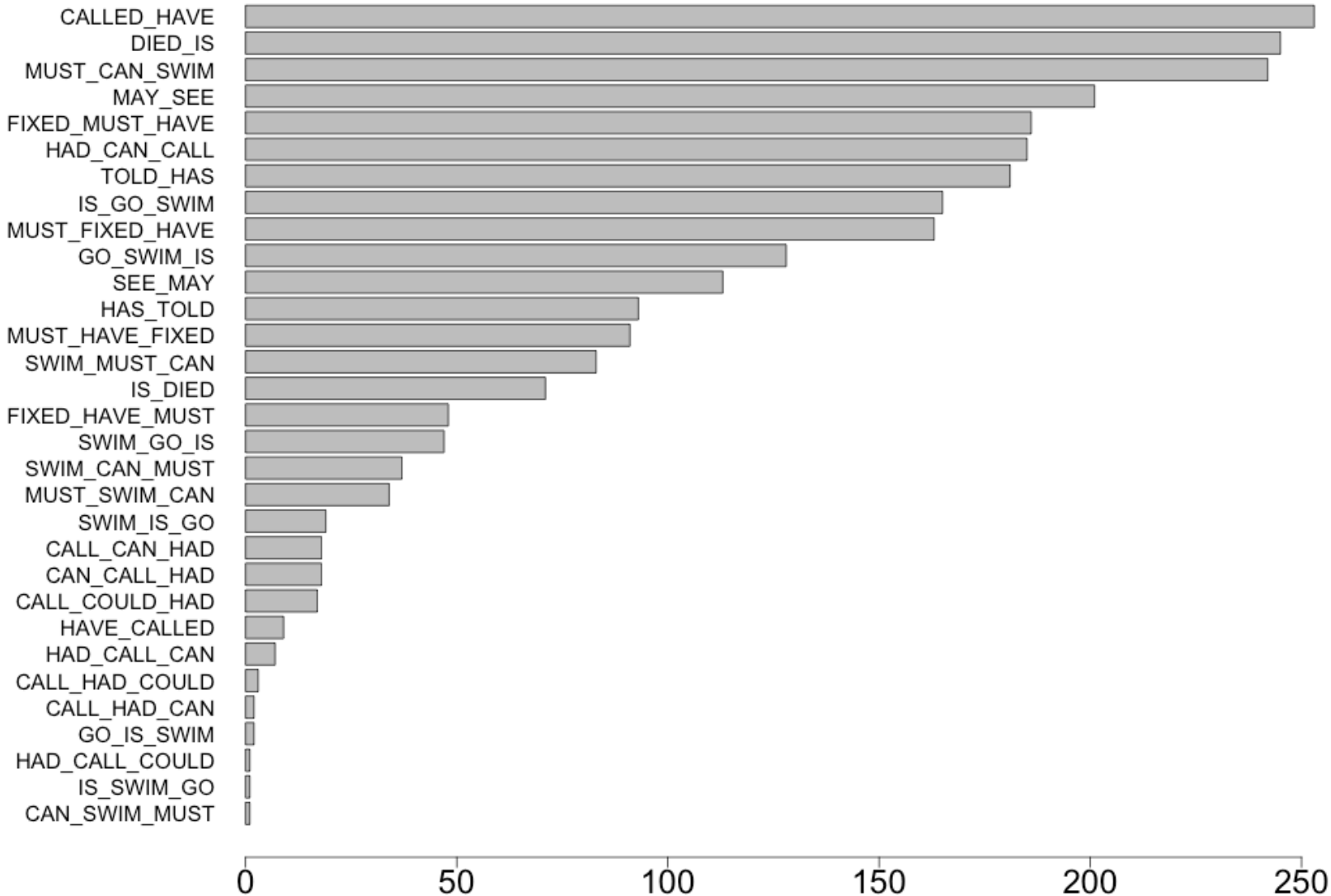


Mixed data: conclusion

- while the analysis in and of itself has no problems dealing with mixed data, we should be wary of subsets in the data when interpreting the results of the analysis
- splitting up the data into two separate subsets has revealed that the variance explained by the third dimension of the original MCA-analysis was almost entirely due to Frisian (dialects)
- this renders less likely our original interpretation of that dimension (a head-final base order underlying all varieties of Dutch)

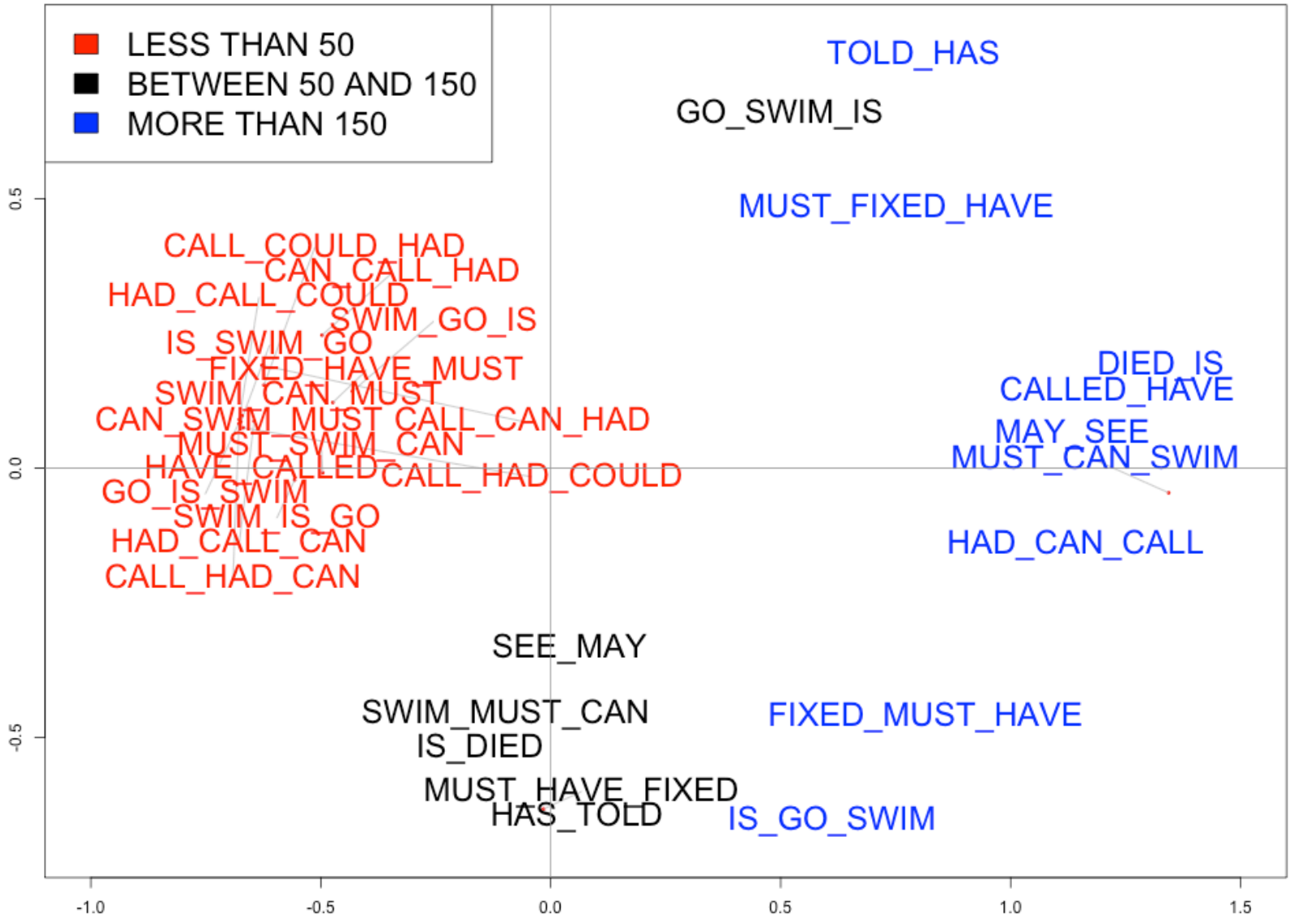
Differences in frequency

- there are massive differences in how frequent particular verb clusters are: some occur in nearly all of the 267 varieties of Dutch under investigation, others occur in virtually none


CALLED_HAVE
DIED_IS
MUST_CAN_SWIM
MAY_SEE
FIXED_MUST_HAVE
HAD_CAN_CALL
TOLD_HAS
IS_GO_SWIM
MUST_FIXED_HAVE
GO_SWIM_IS
SEE_MAY
HAS_TOLD
MUST_HAVE_FIXED
SWIM_MUST_CAN
IS_DIED
FIXED_HAVE_MUST
SWIM_GO_IS
SWIM_CAN_MUST
MUST_SWIM_CAN
SWIM_IS_GO
CALL_CAN_HAD
CAN_CALL_HAD
CALL_COULD_HAD
HAVE_CALLED
HAD_CALL_CAN
CALL_HAD_COULD
CALL_HAD_CAN
GO_IS_SWIM
HAD_CALL_COULD
IS_SWIM_GO
CAN_SWIM_MUST
0
50
100
150
200
250

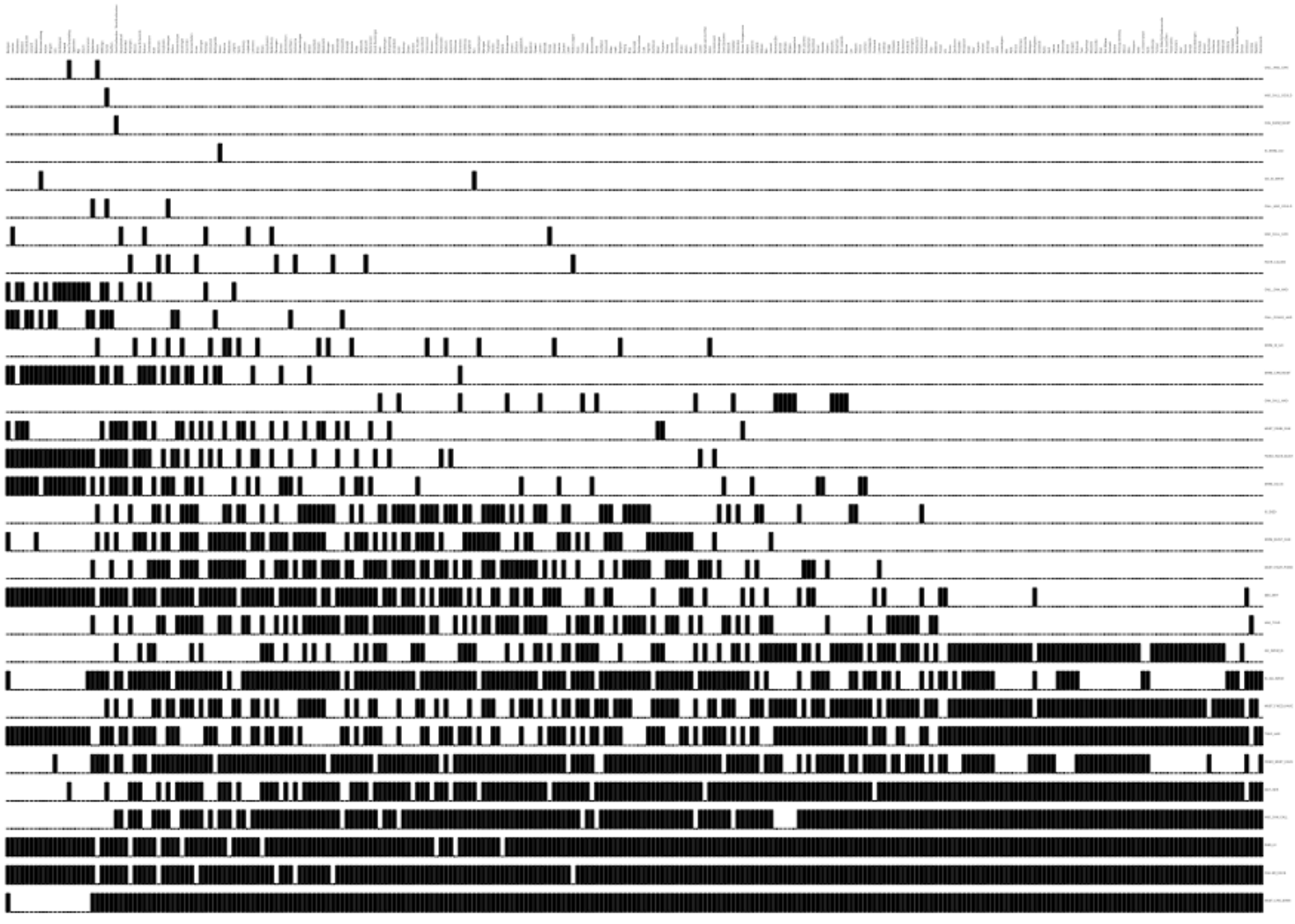
Differences in frequency

- there are massive differences in how frequent particular verb clusters are: some occur in nearly all of the 267 varieties of Dutch under investigation, others occur in virtually none
- question: to what extent have these differences in frequency influenced the MCA-results?

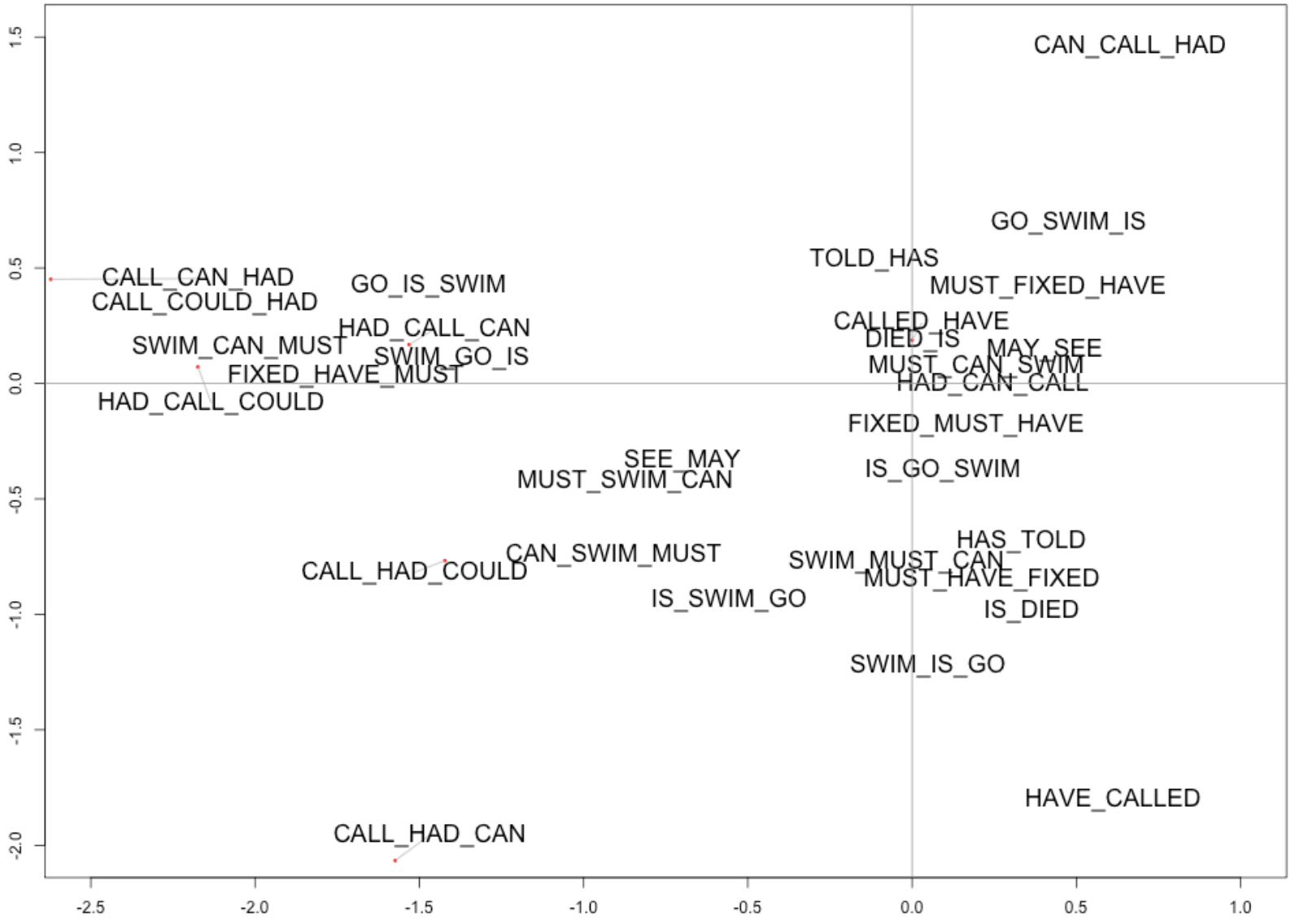

LESS THAN 50
BETWEEN 50 AND 150
MORE THAN 150
TOLD_HAS
GO_SWIM_IS
MUST_FIXED_HAVE
CALL_COULD_HAD
CAN_CALL_HAD
HAD_CALL_COULD
SWIM_GO_IS
IS_SWIM_GO
FIXED_HAVE_MUST
SWIM_CAN_MUST
CAN_SWIM_MUST
CALL_CAN_HAD
MUST_SWIM_CAN
HAVE_CALLED
CALL_HAD_COULD
GO_IS_SWIM
SWIM_IS_GO
HAD_CALL_CAN
CALL_HAD_CAN
DIED_IS
CALLED_HAVE
MAY_SEE
MUST_CAN_SWIM
HAD_CAN_CALL
SEE_MAY
SWIM_MUST_CAN
IS_DIED
FIXED_MUST_HAVE
MUST_HAVE_FIXED
HAS_TOLD
IS_GO_SWIM
0.5
0.0
-0.5
-1.0
-0.5
0.0
0.5
1.0
1.5

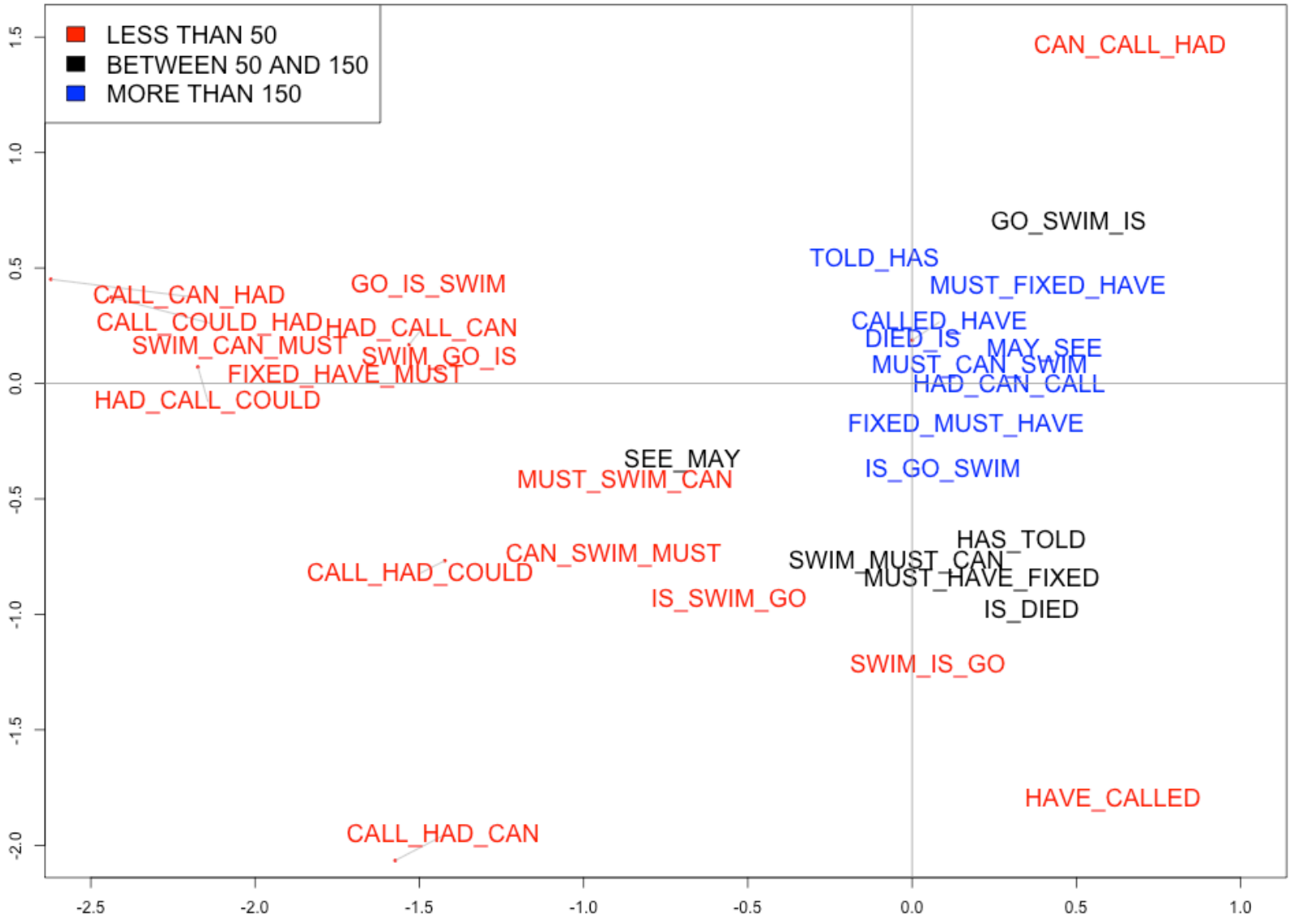
Differences in frequency

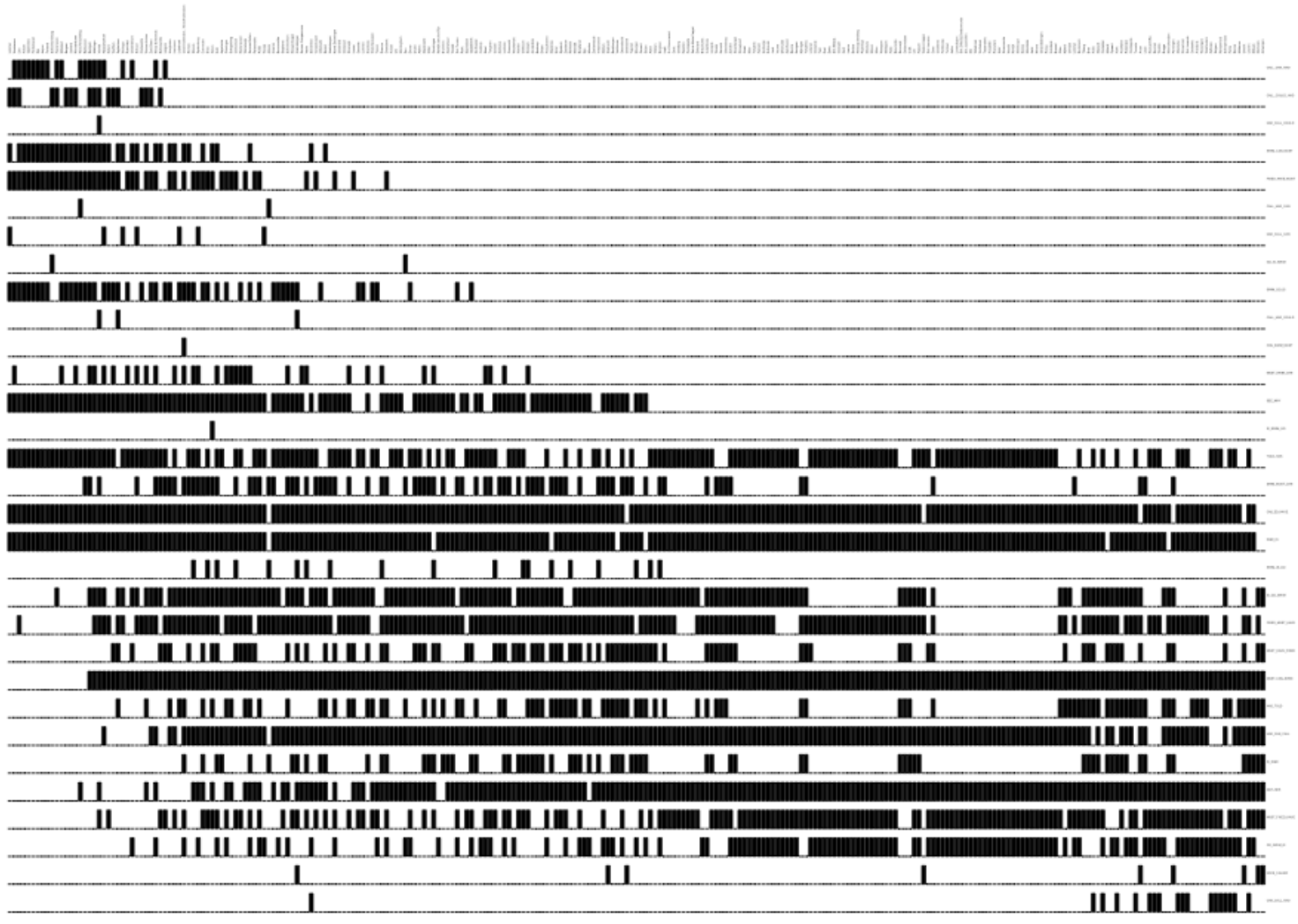
- there are massive differences in how frequent particular verb clusters are: some occur in nearly all of the 267 varieties of Dutch under investigation, others occur in virtually none
- question: to what extent have these differences in frequency influenced the MCA-results?
- answer: the first dimension is clearly dominated by frequency (η^2 of three-valued variable 'frequency' = 0.936)
- subquestions:
  1. what is causing this frequency effect?
  2. what can be done about it?



- what is causing this frequency effect?
  - highly frequent orders will be similar, regardless of where the (few) dialect locations not allowing for those orders are situated
  - highly infrequent orders will be similar, regardless of where the (few) dialect locations allowing for those orders are situated
- what can be done about it?
  - one possible solution: in measuring similarity between cluster orders, give more weight to two dialect locations sharing a rare cluster order than to two dialect locations sharing a highly frequent one
  - a statistical technique that will have this effect: correspondence analysis (CA) (instead of multiple correspondence analysis (MCA))


CAN_CALL_HAD
GO_SWIM_IS
TOLD_HAS
MUST_FIXED_HAVE
CALLED_HAVE
DIED_IS
MAY_SEE
MUST_CAN_SWIM
HAD_CAN_CALL
FIXED_MUST_HAVE
IS_GO_SWIM
HAS_TOLD
SWIM_MUST_CAN
MUST_HAVE_FIXED
IS_DIED
SWIM_IS_GO
HAVE_CALLED
CALL_CAN_HAD
CALL_COULD_HAD
GO_IS_SWIM
HAD_CALL_CAN
SWIM_CAN_MUST
SWIM_GO_IS
FIXED_HAVE_MUST
HAD_CALL_COULD
SEE_MAY
MUST_SWIM_CAN
CAN_SWIM_MUST
CALL_HAD_COULD
IS_SWIM_GO
CALL_HAD_CAN
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0


LESS THAN 50
BETWEEN 50 AND 150
MORE THAN 150
CAN_CALL_HAD
GO_SWIM_IS
TOLD_HAS
MUST_FIXED_HAVE
CALLED_HAVE
DIED_IS
MAY_SEE
MUST_CAN_SWIM
HAD_CAN_CALL
FIXED_MUST_HAVE
IS_GO_SWIM
HAS_TOLD
SWIM_MUST_CAN
MUST_HAVE_FIXED
IS_DIED
SWIM_IS_GO
HAVE_CALLED
CALL_HAD_CAN
SEE_MAY
MUST_SWIM_CAN
CAN_SWIM_MUST
CALL_HAD_COULD
IS_SWIM_GO
GO_IS_SWIM
CALL_CAN_HAD
CALL_COULD_HAD
HAD_CALL_CAN
SWIM_CAN_MUST
SWIM_GO_IS
FIXED_HAVE_MUST
HAD_CALL_COULD
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0



Differences in frequency: conclusion

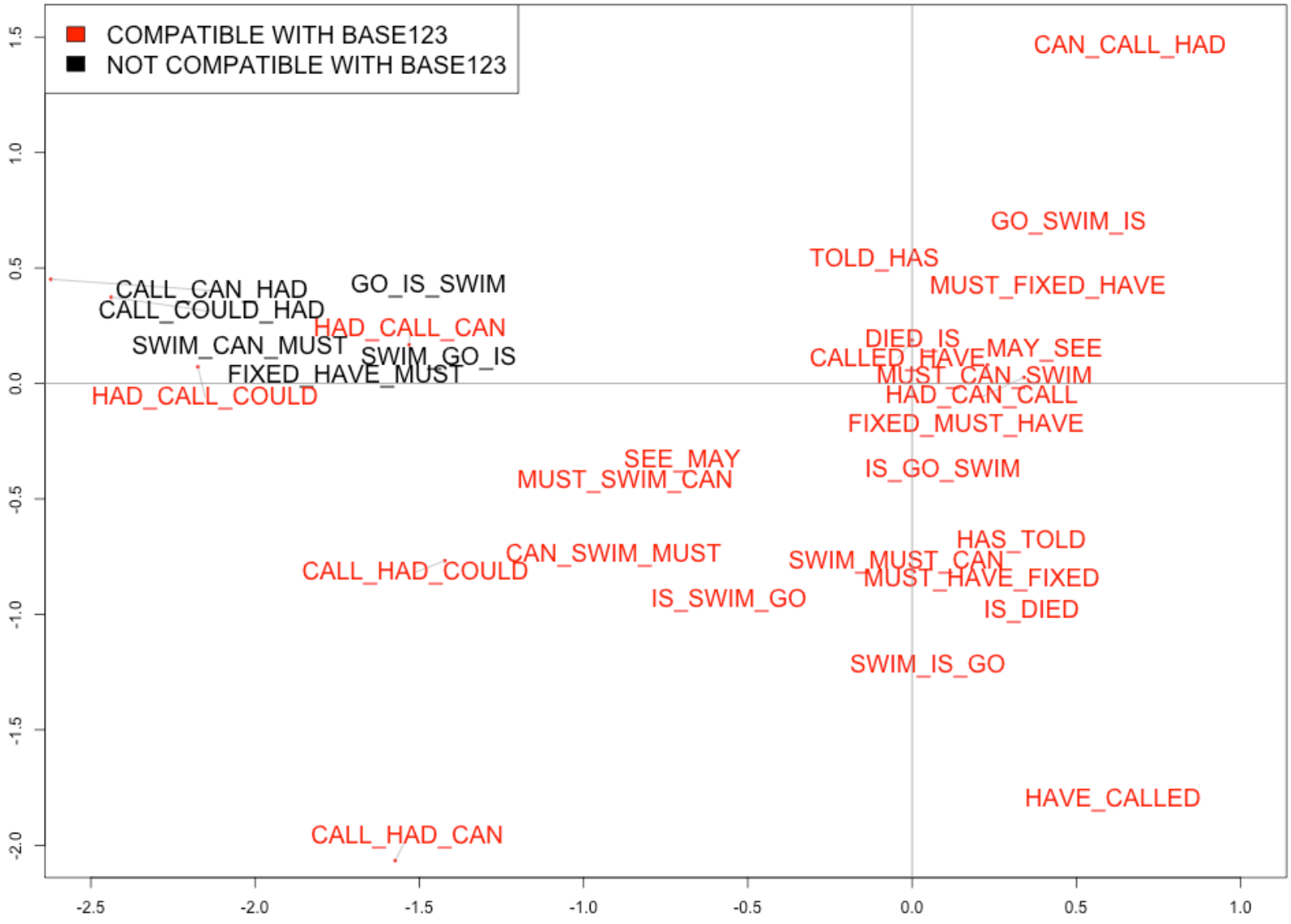
- the original MCA-based analysis was (in its first dimension) very heavily influenced by frequency: it grouped together verb cluster orders mainly based on how frequently they occur
- a CA-based analysis lends more weight to dialect locations sharing rare cluster orders than to dialect locations sharing very frequent ones
- the result is a measure of similarity between verb cluster orders that is less frequency-driven than an MCA-based one
- but to what extent do our original theoretical conclusions (based on the MCA-analysis) carry over to the new, CA-based account? put differently, does the CA-based analysis make sense from the perspective of theoretical linguistics?

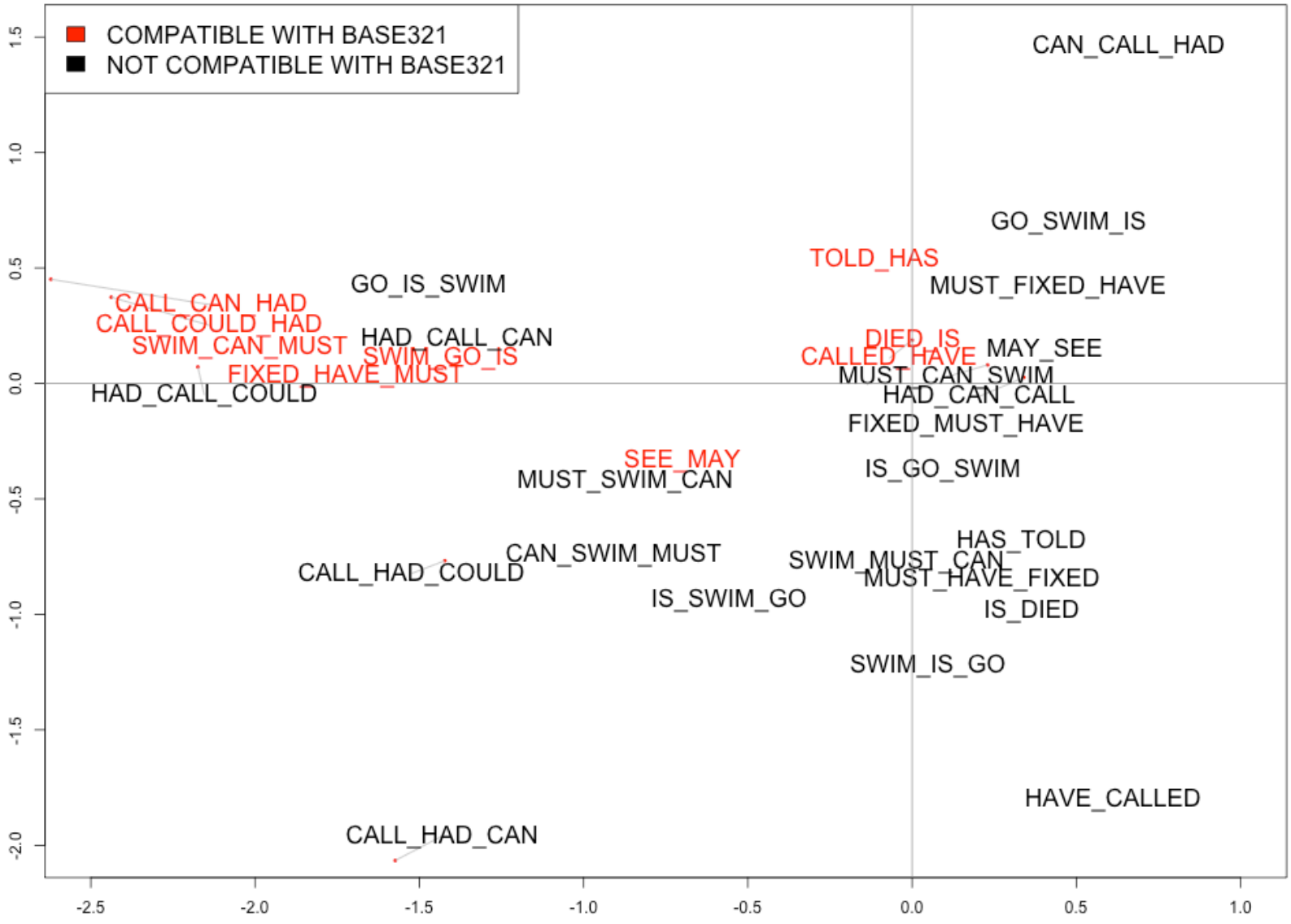
Towards a new analysis

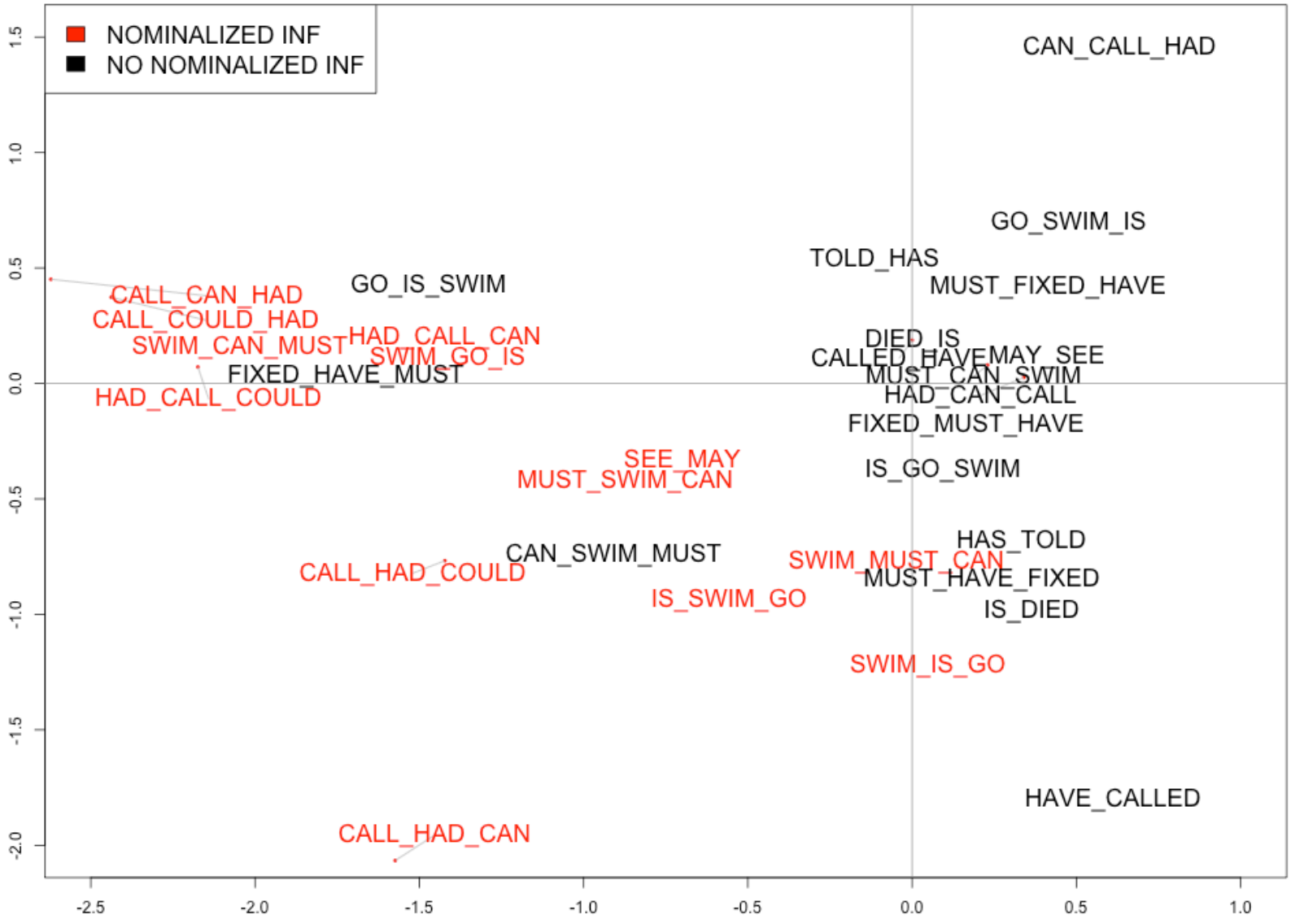
- recent new account of verb cluster variation in Dutch: Barbiers et al. (2016)
- their account is (a) based on the same data set, and (b) designed with an eye towards accounting for co-occurrence patterns in verb cluster word orders
- Barbiers et al. (2016) derive verb cluster orders as follows:
  1. there are two possible base orders: stricly ascending (12, 123) and strictly descending (21, 321)
  2. participles can be adjectivized or not: if they are, they precede the verb cluster (and hence their selecting verb): $PART_2$-AUX_1, $PART_3$-MOD_1-AUX_2, $INF_{IPP.2}$-INF_3-AUX_1, and (ambiguously) $PART_3$-AUX_2-MOD_1
  3. infinitives can be nominalized or not: if they are, they precede the verb cluster (and hence their selecting verb): INF_2-MOD_1, INF_3-MOD_1-MOD_2, and (ambiguously) INF_3-MOD_2-MOD_1
  4. dialects do/do not allow for interruption of the cluster by non-verbal material (requires an adjectival participle or a nominal infinitive) $\rightarrow$ yields the order 132

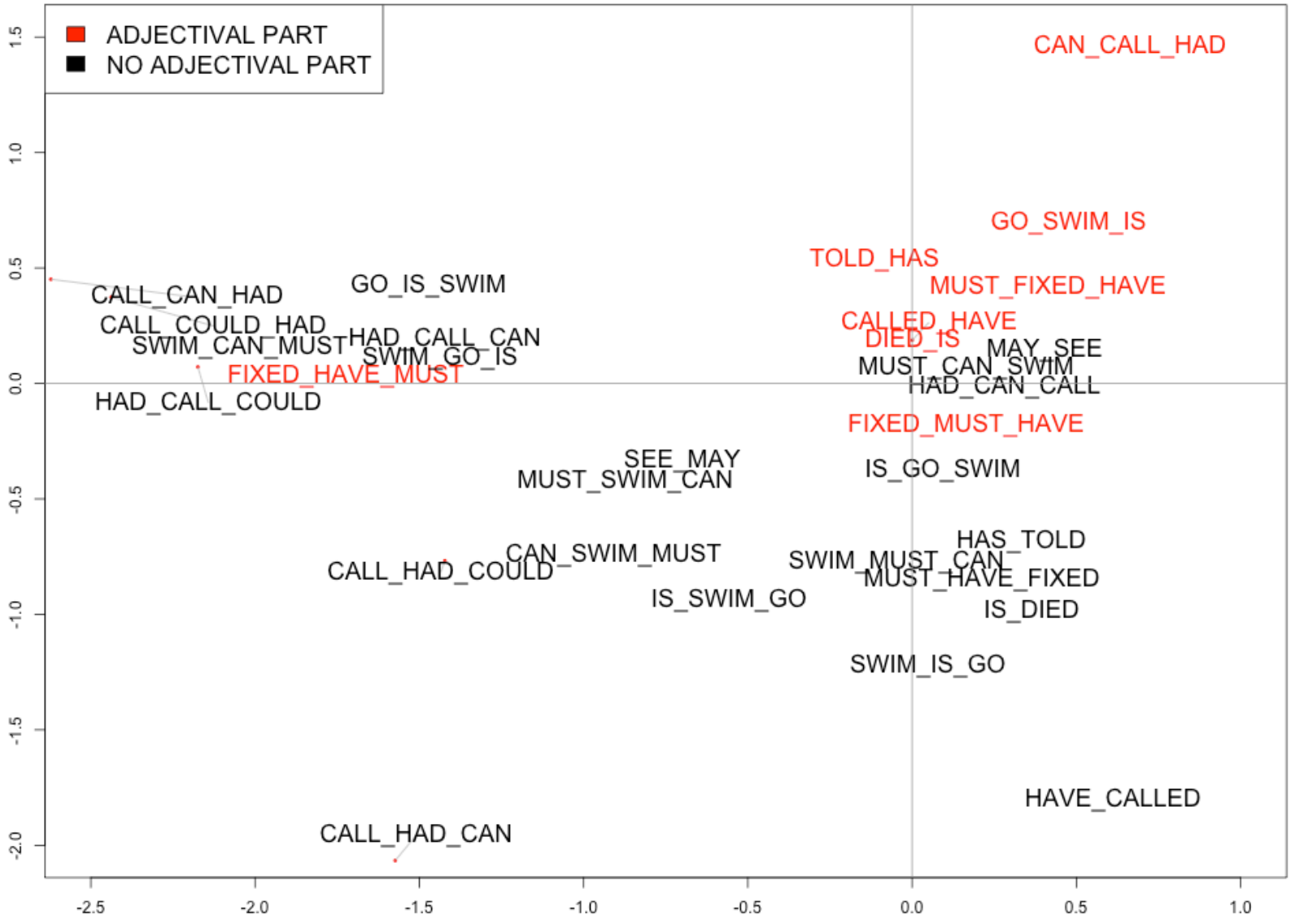
- as before, these linguistic properties can be coded as supplementary variables in the analysis:
  - [BASE123]: is the order compatible with an ascending base order?
  - [BASE321]: is the order compatible with a descending base order?
  - [ADJPART]: does the order involve an adjectivized participle?
  - [NOMINF]: does the order involve a nominalized infinitive?
  - [CLUSTINTERR]: does the order involve cluster interruption?
- this new analysis turns out to line up very nicely with the CA-based analysis of the data set:

η^2	dimension #1	dimension #2
BASE123	0.706	0.009
BASE321	0.312	0.096
ADJPART	0.007	0.321
NOMINF	0.454	0.073
CLUSTINTERR	0.003	0.028


COMPATIBLE WITH BASE123
NOT COMPATIBLE WITH BASE123
CAN_CALL_HAD
GO_SWIM_IS
TOLD_HAS
MUST_FIXED_HAVE
CALL_CAN_HAD
GO_IS_SWIM
CALL_COULD_HAD
HAD_CALL_CAN
SWIM_CAN_MUST
SWIM_GO_IS
FIXED_HAVE_MUST
HAD_CALL_COULD
DIED_IS
MAY_SEE
CALLED_HAVE
MUST_CAN_SWIM
HAD_CAN_CALL
FIXED_MUST_HAVE
SEE_MAY
MUST_SWIM_CAN
IS_GO_SWIM
HAS_TOLD
CAN_SWIM_MUST
SWIM_MUST_CAN
CALL_HAD_COULD
MUST_HAVE_FIXED
IS_SWIM_GO
IS_DIED
SWIM_IS_GO
HAVE_CALLED
CALL_HAD_CAN
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0


COMPATIBLE WITH BASE321
NOT COMPATIBLE WITH BASE321
CAN_CALL_HAD
GO_SWIM_IS
TOLD_HAS
MUST_FIXED_HAVE
GO_IS_SWIM
CALL_CAN_HAD
CALL_COULD_HAD
SWIM_CAN_MUST
HAD_CALL_CAN
SWIM_GO_IS
FIXED_HAVE_MUST
HAD_CALL_COULD
DIED_IS
CALLED_HAVE
MAY_SEE
MUST_CAN_SWIM
HAD_CAN_CALL
FIXED_MUST_HAVE
SEE_MAY
MUST_SWIM_CAN
IS_GO_SWIM
HAS_TOLD
CAN_SWIM_MUST
SWIM_MUST_CAN
CALL_HAD_COULD
MUST_HAVE_FIXED
IS_SWIM_GO
IS_DIED
SWIM_IS_GO
HAVE_CALLED
CALL_HAD_CAN
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0


NOMINALIZED INF
NO NOMINALIZED INF
CAN_CALL_HAD
GO_SWIM_IS
TOLD_HAS
MUST_FIXED_HAVE
GO_IS_SWIM
CALL_CAN_HAD
CALL_COULD_HAD
SWIM_CAN_MUST
HAD_CALL_CAN
SWIM_GO_IS
FIXED_HAVE_MUST
HAD_CALL_COULD
DIED_IS
CALLED_HAVEMAY_SEE
MUST_CAN_SWIM
HAD_CAN_CALL
FIXED_MUST_HAVE
SEE_MAY
MUST_SWIM_CAN
IS_GO_SWIM
HAS_TOLD
CAN_SWIM_MUST
SWIM_MUST_CAN
CALL_HAD_COULD
MUST_HAVE_FIXED
IS_SWIM_GO
IS_DIED
SWIM_IS_GO
HAVE_CALLED
CALL_HAD_CAN
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0


ADJECTIVAL PART
NO ADJECTIVAL PART
CAN_CALL_HAD
GO_SWIM_IS
TOLD_HAS
MUST_FIXED_HAVE
CALLED_HAVE
DIED_IS
MAY_SEE
MUST_CAN_SWIM
HAD_CAN_CALL
FIXED_MUST_HAVE
IS_GO_SWIM
HAS_TOLD
SWIM_MUST_CAN
MUST_HAVE_FIXED
IS_DIED
SWIM_IS_GO
HAVE_CALLED
GO_IS_SWIM
CALL_CAN_HAD
CALL_COULD_HAD
SWIM_CAN_MUST
HAD_CALL_CAN
SWIM_GO_IS
FIXED_HAVE_MUST
HAD_CALL_COULD
SEE_MAY
MUST_SWIM_CAN
CAN_SWIM_MUST
CALL_HAD_COULD
IS_SWIM_GO
CALL_HAD_CAN
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0

Conclusion

- this talk has looked at the effect of four sources of 'data badness' on a quantitative-qualitative analysis of word order variation in verb clusters in 267 varieties of Dutch:
  1. **missing data:** small effect $\rightarrow$ data imputation yielded comparable results to a complete cases analysis
  2. **differences in question type:** small effect $\rightarrow$ data from different question types can help guide the interpretation of the results
  3. **mixed data:** medium effect $\rightarrow$ effects caused by subsets of the data should be taken into account when interpreting the results of the whole data set
  4. **differences in frequency:** large effect $\rightarrow$ the statistical tools and techniques we use for determining similarities between linguistic phenomena should be selected such that they undercut the effect of frequency
- more general conclusion: there is ample opportunity for a fruitful collaboration between formal-theoretical linguistics on the one hand and quantitative-statistical analyses of large datasets on the other

References

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. Amsterdam: John Benjamins.

Barbiers, Sjef, Johan van der Auwera, Hans Bennis, Eefje Boef, Gunther De Vogelaer, and Margreet van der Ham. 2008. *Syntactische atlas van de Nederlandse dialecten. Deel II*. Amsterdam: Amsterdam University Press.

Barbiers, Sjef, Hans Bennis, and Lotte Hendriks. 2016. Merging verb cluster variation. Ms. Meertens Institute.

Josse, Julie, Marie Chavent, Benot Liquet, and François Husson. 2012. Handling missing values with regularized iterative Multiple Correspondence Analysis. *Journal of Classification* 29:91–116.