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

A quantitative-qualitative analysis

Four data complications
  Missing values
  Differences in question type
  Mixed data
  Differences in frequency

Towards a new analysis

Conclusion
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.’

- 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.’
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.’  

b. *dasto it ook net *meist zien.*  
that.you it also not may  see  
‘that you’re also not allowed to see it.’  

\[\text{✓ 21}\]  
\[\text{✗ 12}\]
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.’
   (*)

b. dajtgie ook nie meug zien.
   that.it.you also not may see
   ‘that you’re also not allowed to see it.’
   (√)
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)

<table>
<thead>
<tr>
<th></th>
<th>Midsland</th>
<th>Lies</th>
<th>W.-Terschelling</th>
<th>Oosterend</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS_DIED</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>NA</td>
</tr>
<tr>
<td>DIED_IS</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>NA</td>
</tr>
<tr>
<td>HAS_TOLD</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>TOLD_HAS</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>HAVE_CALLED</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>CALLED_HAVE</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>MAYSEE</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>SEE_MAY</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>CAN_SWIM_MUST</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>MUST_CAN_SWIM</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>MUST_SWIM_CAN</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
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
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 → derives 12 and 123
Barbiers (2005) derives verb cluster orders as follows:

- movement is VP-intraposition → derives 21 and 231, 312 and 132, and fails to derive 213

(8) \[ VP_1 \]
\[ VP_2 \]
\[ V_2 \]
\[ V_1 \]
\[ t_{VP_2} \]

(9) \[ VP_1 \]
\[ VP_2 \]
\[ VP_3 \]
\[ V_2 \]
\[ V_1 \]
\[ t_{VP_2} \]
\[ V_3 \]
Barbiers (2005) derives verb cluster orders as follows:

- movement is VP-intraposition → derives 21 and 231, 312 and 132, and fails to derive 213
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)

\[
(12)
\]

```
(12)
  VP_1
    VP_2
      VP_3
      V_2
      t_{VP_3}
    V'_2
    V_1
    t_{VP_2}
  V'_1
  V_1
```


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 → rules out 231 in the case of MOD-MOD/AUX-V-clusters and 312 in the case of AUX-AUX/MOD-V-clusters.

(13) \[ [\text{VP}_1 \text{MUST}_{u\text{Event}}] [\text{VP}_2 \text{CAN}_{u\text{Event}}] [\text{VP}_3 \text{SWIM}_{i\text{Event}}] ] \]

(14) \[ [\text{VP}_1 \text{HAD}_{u\text{Perf}}] [\text{VP}_2 \text{CAN}_{i\text{Perf},u\text{Event}}] [\text{VP}_3 \text{CALL}_{i\text{Event}}] ] \]
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?

<table>
<thead>
<tr>
<th>BASE-GENERATION</th>
<th>MOVEMENT</th>
<th>PIED-PIPING</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS_DIED</td>
<td>yesBase</td>
<td>noMvt</td>
</tr>
<tr>
<td>DIED_IS</td>
<td>noBase</td>
<td>yesMvt</td>
</tr>
<tr>
<td>HAS_TOLD</td>
<td>yesBase</td>
<td>noMvt</td>
</tr>
<tr>
<td>TOLD_HAS</td>
<td>noBase</td>
<td>yesMvt</td>
</tr>
<tr>
<td>HAVE_CALLED</td>
<td>yesBase</td>
<td>noMvt</td>
</tr>
<tr>
<td>CALLED_HAVE</td>
<td>noBase</td>
<td>yesMvt</td>
</tr>
<tr>
<td>MAYSEE</td>
<td>yesBase</td>
<td>noMvt</td>
</tr>
<tr>
<td>SEE_MAY</td>
<td>noBase</td>
<td>yesMvt</td>
</tr>
<tr>
<td>CAN_SWIM_MUST</td>
<td>noBase</td>
<td>yesMvt</td>
</tr>
<tr>
<td>MUST_CAN_SWIM</td>
<td>yesBase</td>
<td>noMvt</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

...
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
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 ($\eta^2$):

<table>
<thead>
<tr>
<th></th>
<th>dimension 1</th>
<th>dimension 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barbiers (2005) base generation</td>
<td>0.126</td>
<td>0.309</td>
</tr>
<tr>
<td>Bader (2012) V&lt;Mod</td>
<td>0.212</td>
<td>0.004</td>
</tr>
</tbody>
</table>
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. Particle<auxiliary and modal/aspectual<infinitive

- PART < AUX AND MOD/ASP < INF
- AUX < PART AND/OR INF < MOD/ASP

TOLD_HAS

MUST_FIXED_HAVE

DIED_IS

CALLED_HAVE

MAYSEE

MUST_CAN_SWIM

HAD_CAN_CALL

SEE_MAY

SWIM_MUST_CAN

IS_DIED

MUST_HAVE_FIXED

HAS_TOLD

IS_GO_SWIM

HAD_CALL_CAN

CALL_HAD_CAN

CALL_HAD_COULD

CAN_CALL_HAD

HAD_CALL_COULD

SWIM, GO_IS

GO IS SWIM

FIXED_HAVE MUST

CAN SWIM MUST CALL COULD_HAD

IS SWIM_GO

HAVE CALLED

MUST SWIM CAN

IS GO

HAD CALL CAN

CALL_HAD_CAN

SEE_MAY

SWIM MUST CAN

IS_DIED

MUST_HAVE_FIXED

HAS_TOLD

IS.GO.SWIM

HAD_CAN_CALL
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)
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 the scatter plot for visual representation of the clusters.
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?
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
<table>
<thead>
<tr>
<th>Phrase</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>CALL_CAN_HAD</td>
<td></td>
</tr>
<tr>
<td>CALL_COULD_HAD</td>
<td></td>
</tr>
<tr>
<td>CALL_HAD_COULD</td>
<td></td>
</tr>
<tr>
<td>CALL_HAD_CAN</td>
<td></td>
</tr>
<tr>
<td>CAN_CALL_HAD</td>
<td></td>
</tr>
<tr>
<td>HAD_CALL_CAN</td>
<td></td>
</tr>
<tr>
<td>HAD_CALL_COULD</td>
<td></td>
</tr>
<tr>
<td>HAD_CAN_CALL</td>
<td></td>
</tr>
<tr>
<td>SWIM_GO_IS</td>
<td></td>
</tr>
<tr>
<td>SWIM_IS_GO</td>
<td></td>
</tr>
<tr>
<td>GO_SWIM_IS</td>
<td></td>
</tr>
<tr>
<td>GO_IS_SWIM</td>
<td></td>
</tr>
<tr>
<td>IS_SWIM_GO</td>
<td></td>
</tr>
<tr>
<td>IS_GO_SWIM</td>
<td></td>
</tr>
<tr>
<td>FIXED_HAVE_MUST</td>
<td></td>
</tr>
<tr>
<td>FIXED_MUST_HAVE</td>
<td></td>
</tr>
<tr>
<td>MUST_FIXED_HAVE</td>
<td></td>
</tr>
<tr>
<td>MUST_HAVE_FIXED</td>
<td></td>
</tr>
<tr>
<td>SWIM_MUST_CAN</td>
<td></td>
</tr>
<tr>
<td>SWIM_CAN_MUST</td>
<td></td>
</tr>
<tr>
<td>MUST_SWIM_CAN</td>
<td></td>
</tr>
<tr>
<td>MUST_CAN_SWIM</td>
<td></td>
</tr>
<tr>
<td>CAN_SWIM_MUST</td>
<td></td>
</tr>
<tr>
<td>SEE_MAY</td>
<td></td>
</tr>
<tr>
<td>MAYSEE</td>
<td></td>
</tr>
<tr>
<td>CALLED_HAVE</td>
<td></td>
</tr>
<tr>
<td>HAVE_CALLED</td>
<td></td>
</tr>
<tr>
<td>TOLD_HAS</td>
<td></td>
</tr>
<tr>
<td>HAS_TOLD</td>
<td></td>
</tr>
<tr>
<td>DIED_IS</td>
<td></td>
</tr>
<tr>
<td>IS_DIED</td>
<td></td>
</tr>
</tbody>
</table>
(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 dasszij 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

<table>
<thead>
<tr>
<th></th>
<th>Midsland</th>
<th>Lies</th>
<th>Oosterend</th>
<th>Hollum</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS_DIED</td>
<td>no</td>
<td>no</td>
<td>NA</td>
<td>no</td>
</tr>
<tr>
<td>DIED_IS</td>
<td>yes</td>
<td>yes</td>
<td>NA</td>
<td>yes</td>
</tr>
<tr>
<td>HAS_TOLD</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>NA</td>
</tr>
<tr>
<td>TOLD_HAS</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>NA</td>
</tr>
<tr>
<td>HAVE_CALLED</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>CALLED_HAVE</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>MAYSEE</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>SEE_MAY</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>CAN_SWIM_MUST</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th></th>
<th>Midsland</th>
<th>Lies</th>
<th>Oosterend</th>
<th>Hollum</th>
<th>…</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS_DIED</td>
<td>1</td>
<td>1</td>
<td>0.816</td>
<td>1</td>
<td>…</td>
</tr>
<tr>
<td>DIED_IS</td>
<td>0</td>
<td>0</td>
<td>0.088</td>
<td>0</td>
<td>…</td>
</tr>
<tr>
<td>HAS_TOLD</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.947</td>
<td>…</td>
</tr>
<tr>
<td>TOLD_HAS</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.255</td>
<td>…</td>
</tr>
<tr>
<td>HAVE_CALLED</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>…</td>
</tr>
<tr>
<td>CALLED_HAVE</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>…</td>
</tr>
<tr>
<td>MAYSEE</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>…</td>
</tr>
<tr>
<td>SEE_MAY</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>…</td>
</tr>
<tr>
<td>CAN_SWIM_MUST</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>…</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th></th>
<th>Midsland</th>
<th>Lies</th>
<th>Oosterend</th>
<th>Hollum</th>
<th>…</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS_DIED</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>DIED_IS</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HAS_TOLD</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TOLD_HAS</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HAVE_CALLED</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CALLED HAVE</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MAY SEE</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SEE MAY</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CAN SWIM MUST</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th></th>
<th>Midsland</th>
<th>Lies</th>
<th>Oosterend</th>
<th>Hollum</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS_DIED</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>...</td>
</tr>
<tr>
<td>DIED_IS</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>...</td>
</tr>
<tr>
<td>HAS_TOLD</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>...</td>
</tr>
<tr>
<td>TOLD_HAS</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>...</td>
</tr>
<tr>
<td>HAVE_CALLED</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>...</td>
</tr>
<tr>
<td>CALLED_HAVE</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>...</td>
</tr>
<tr>
<td>MAYSEE</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>...</td>
</tr>
<tr>
<td>SEE_MAY</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>...</td>
</tr>
<tr>
<td>CAN_SWIM_MUST</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
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 gestorven is.
     translate she knows not that Marie yesterday died
     ‘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 → 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
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)
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.
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?
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 ($\eta^2$ 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))
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: strictly 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): \(\text{PART}_2\)-AUX\(_1\), \(\text{PART}_3\)-MOD\(_1\)-AUX\(_2\), INF\(_{\text{IPP},2}\)-INF\(_3\)-AUX\(_1\), and (ambiguously) \(\text{PART}_3\)-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) → 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:

<table>
<thead>
<tr>
<th></th>
<th>dimension #1</th>
<th>dimension #2</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASE123</td>
<td>0.706</td>
<td>0.009</td>
</tr>
<tr>
<td>BASE321</td>
<td>0.312</td>
<td>0.096</td>
</tr>
<tr>
<td>ADJPART</td>
<td>0.007</td>
<td>0.321</td>
</tr>
<tr>
<td>NOMINF</td>
<td>0.454</td>
<td>0.073</td>
</tr>
<tr>
<td>CLUSTINTERR</td>
<td>0.003</td>
<td>0.028</td>
</tr>
</tbody>
</table>
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 → data imputation yielded comparable results to a complete cases analysis
  2. **differences in question type**: small effect → data from different question types can help guide the interpretation of the results
  3. **mixed data**: medium effect → 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 → 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


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.
