Quantity and quality in linguistic variation
The case of verb clusters

Jeroen van Craenenbroeck

KU Leuven/CRISSP

Séminaire
INALCO – SeDyL
Paris, February 6, 2015
This talk in one slide

- **main topic:** interaction between formal-theoretical and quantitative-statistical linguistics

- **starting point:** the massive amount of variation attested in Dutch verb clusters necessitates a collaboration between formal and quantitative approaches

- **traditional dialectometry** measures (dis)similarities between dialect locations based on their linguistic profile

- **reverse dialectometry** measures (dis)similarities between *linguistic constructions* based on their geographical distribution and maps these results against formal-theoretical parameters

- **result:** a method that can detect and identify grammatical parameters in a large and highly varied data set
Introduction

- quantitative and qualitative linguistics typically represent two non-communicating, non-intersecting subfields

- **qualitative linguistics:**
  - theory-driven
  - works with small and/or simplified data sets
  - makes little or no use of mathematical or statistical methods

- **quantitative linguistics:**
  - data-driven
  - works with (very) large and highly varied data sets
  - makes extensive use of mathematical or statistical methods
▶ **this talk** explores the possibility of a collaboration between the two approaches:
  ▶ to what extent can a quantitative analysis of large datasets lead to new theoretical insights?
  ▶ to what extent can theoretical analyses guide and inform quantitative analyses of language data?

▶ **proposal:**
  ▶ quantitative analyses can be used to test, evaluate, and compare hypotheses from the theoretical literature
  ▶ theoretical analyses can be used to interpret and narrow down the space of hypotheses generated by quantitative analyses

▶ **main topic for today:** word order variation in dialect Dutch verb clusters:
  ▶ lots of data available and lots of interdialectal variation
  ▶ extensively researched in the formal-theoretical (esp. generative) literature for at least four decades
The data: dialect Dutch verb clusters

- 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:

(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:

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

(7) a. ...dat iedereen moet zwemmen kunnen. \(\checkmark\) 132  
b. ...dat iedereen zwemmen moet kunnen. \(\checkmark\) 312  
c. ...dat iedereen zwemmen kunnen moet. \(\checkmark\) 321  
d. *...dat iedereen kunnen zwemmen moet.  (*231)  
e. *...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:

(8) **Midsland Dutch**

   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 mot 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.* (*321)

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

e. *dat iedereen kunnen zwemmen mot.* (*213)
more generally, the four possible cluster orders yield a total of 16 possible combinations, of which 12 are attested in Dutch dialects:

<table>
<thead>
<tr>
<th>example dialect</th>
<th>123</th>
<th>132</th>
<th>321</th>
<th>312</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beetgum</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Hippolytushoef</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>*</td>
</tr>
<tr>
<td>Warffum</td>
<td>✓</td>
<td>✓</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Oosterend</td>
<td>✓</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Schermerhorn</td>
<td>✓</td>
<td>✓</td>
<td>*</td>
<td>✓</td>
</tr>
<tr>
<td>Visvliet</td>
<td>✓</td>
<td>*</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Kollum</td>
<td>✓</td>
<td>*</td>
<td>✓</td>
<td>*</td>
</tr>
<tr>
<td>Langelo</td>
<td>✓</td>
<td>*</td>
<td>*</td>
<td>✓</td>
</tr>
<tr>
<td>Midsland</td>
<td>*</td>
<td>✓</td>
<td>✓</td>
<td>*</td>
</tr>
<tr>
<td>Lies</td>
<td>*</td>
<td>*</td>
<td>✓</td>
<td>*</td>
</tr>
<tr>
<td>Bakkeveen</td>
<td>*</td>
<td>*</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Waskemeer</td>
<td>*</td>
<td>✓</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>
in order to get a more complete picture of the variation, we can look at the results from the SAND-project:

- 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
Research questions

- this state of affairs raises (at least) two fundamental questions for grammatical theory:
  1. to what extent is this variation due to grammar-internal properties, and what part of it is grammar-external?
  2. how can we draw the line between the two?

- one extreme position: Barbiers (2005): the grammar rules out 231 and 213 in MOD-MOD-V-cluster, but all other orders are freely available to all speakers; the choice between them is determined by sociolinguistic factors (geographical and social norms, register, context, . . . )

- in this talk I use quantitative-statistical methods to identify three grammatical (micro)parameters that together are responsible for the bulk of the variation
Theoretical background: dialectometry

- **dialectometry** is a subdiscipline of linguistics that uses computational and quantitative techniques in dialectology (Nerbonne and Kretzschmar Jr., 2013)
- this section presents a prototypical dialectometric analysis, which will serve as a stepping stone for the actual analysis in the next section
- **starting point:** the raw data from 13 SAND-maps:
  - 4 about two-verb clusters (3×AUXILIARY-PARTICIPLE, 1×MODAL-INFINITIVE)
  - 4 about three-verb clusters (MODAL-MODAL-INFINITIVE, MODAL-AUXILIARY-PARTICIPLE, AUXILIARY-AUXILIARY-INFINITIVE, AUXILIARY-MODAL-INFINITIVE)
  - 3 about particle placement inside the cluster
  - 2 about morphology of the past participle
- for a total of 67 linguistic variables in 267 locations
this yields a $267 \times 67$ matrix with one row per location and one column per linguistic variable, i.e. locations $=$ individuals and linguistic phenomena $=$ variables
<table>
<thead>
<tr>
<th>Location</th>
<th>AUX1(be.sg)-PART2</th>
<th>PART2-AUX1(be.sg)</th>
<th>AUX1(have.sg)-PART2</th>
<th>PART2-AUX1(have.sg)</th>
<th>AUX1(have)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Midsland / Midlārs</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Lies</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>West-Terschelling</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Oosterend</td>
<td>NA</td>
<td>NA</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Hollum</td>
<td>no</td>
<td>yes</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Schiermonnikoog</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Ferwerd / Ferwert</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Anjum / Eanjum</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Kollum</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Visvilet</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Oosterbierum / Ee</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Beetgum / Bitgum</td>
<td>no</td>
<td>yes</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Bergum / Burgum</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Jorwerd / Jorwert</td>
<td>no</td>
<td>yes</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Bakkeveen / Bakk</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Waskemeer / De W</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Kloosterburen</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Warffum</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Leermens</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Groningen</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Nieuw-Scheemda</td>
<td>NA</td>
<td>NA</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Langelo</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
</tbody>
</table>
this yields a $267 \times 67$ matrix with one row per location and one column per linguistic variable, i.e. locations = individuals and linguistic phenomena = variables

step 1: convert the table into a $267 \times 267$ (symmetric) distance matrix, whereby for each pair of locations a distance between them is calculated based on the linguistic features they share
Quantity and quality in linguistic variation

Jeroen van Craenenbroeck

This talk in one slide

Introduction

The data: dialect Dutch verb clusters

Research questions

Theoretical background: dialectometry

Methodology: reverse dialectometry

Results

Main conclusion

Extensions

References

<table>
<thead>
<tr>
<th></th>
<th>Midsland</th>
<th>Lies</th>
<th>West-Tersch.</th>
<th>Oosterend</th>
<th>Hollum</th>
<th>Schiermonniko</th>
<th>Ferwerd</th>
<th>Anjum</th>
<th>Kollum</th>
<th>Visvliet</th>
<th>Oosterbier</th>
<th>Beet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Midsland / Midsland</td>
<td>0,000</td>
<td>0,500</td>
<td>0,333</td>
<td>0,706</td>
<td>0,250</td>
<td>0,647</td>
<td>0,357</td>
<td>0,250</td>
<td>0,611</td>
<td>0,650</td>
<td>0,533</td>
<td>0,000</td>
</tr>
<tr>
<td>Lies / Midsland</td>
<td>0,500</td>
<td>0,000</td>
<td>0,444</td>
<td>0,750</td>
<td>0,588</td>
<td>0,375</td>
<td>0,471</td>
<td>0,563</td>
<td>0,444</td>
<td>0,444</td>
<td>0,632</td>
<td>0,000</td>
</tr>
<tr>
<td>West-Tersch. / Midsland</td>
<td>0,333</td>
<td>0,444</td>
<td>0,000</td>
<td>0,789</td>
<td>0,429</td>
<td>0,667</td>
<td>0,286</td>
<td>0,429</td>
<td>0,632</td>
<td>0,600</td>
<td>0,500</td>
<td>0,000</td>
</tr>
<tr>
<td>Oosterend / Midsland</td>
<td>0,706</td>
<td>0,750</td>
<td>0,789</td>
<td>0,000</td>
<td>0,706</td>
<td>0,765</td>
<td>0,737</td>
<td>0,538</td>
<td>0,563</td>
<td>0,600</td>
<td>0,600</td>
<td>0,000</td>
</tr>
<tr>
<td>Hollum / Midsland</td>
<td>0,250</td>
<td>0,588</td>
<td>0,429</td>
<td>0,706</td>
<td>0,000</td>
<td>0,667</td>
<td>0,167</td>
<td>0,000</td>
<td>0,625</td>
<td>0,714</td>
<td>0,462</td>
<td>0,000</td>
</tr>
<tr>
<td>Schiermonniko / Midsland</td>
<td>0,647</td>
<td>0,375</td>
<td>0,667</td>
<td>0,765</td>
<td>0,667</td>
<td>0,000</td>
<td>0,625</td>
<td>0,667</td>
<td>0,400</td>
<td>0,556</td>
<td>0,706</td>
<td>0,000</td>
</tr>
<tr>
<td>Ferwerd / Midsland</td>
<td>0,357</td>
<td>0,471</td>
<td>0,286</td>
<td>0,737</td>
<td>0,167</td>
<td>0,625</td>
<td>0,000</td>
<td>0,182</td>
<td>0,588</td>
<td>0,682</td>
<td>0,308</td>
<td>0,000</td>
</tr>
<tr>
<td>Anjum / Midsland</td>
<td>0,250</td>
<td>0,563</td>
<td>0,429</td>
<td>0,538</td>
<td>0,000</td>
<td>0,667</td>
<td>0,182</td>
<td>0,000</td>
<td>0,571</td>
<td>0,625</td>
<td>0,417</td>
<td>0,000</td>
</tr>
<tr>
<td>Kollum / Midsland</td>
<td>0,611</td>
<td>0,444</td>
<td>0,632</td>
<td>0,563</td>
<td>0,625</td>
<td>0,400</td>
<td>0,588</td>
<td>0,571</td>
<td>0,000</td>
<td>0,353</td>
<td>0,625</td>
<td>0,000</td>
</tr>
<tr>
<td>Visvliet / Midsland</td>
<td>0,650</td>
<td>0,444</td>
<td>0,600</td>
<td>0,600</td>
<td>0,714</td>
<td>0,556</td>
<td>0,682</td>
<td>0,625</td>
<td>0,353</td>
<td>0,000</td>
<td>0,588</td>
<td>0,000</td>
</tr>
<tr>
<td>Oosterbierum / Midsland</td>
<td>0,533</td>
<td>0,632</td>
<td>0,500</td>
<td>0,600</td>
<td>0,462</td>
<td>0,706</td>
<td>0,308</td>
<td>0,417</td>
<td>0,625</td>
<td>0,588</td>
<td>0,000</td>
<td>0,000</td>
</tr>
<tr>
<td>Beetgum / Midsland</td>
<td>0,545</td>
<td>0,714</td>
<td>0,500</td>
<td>0,727</td>
<td>0,500</td>
<td>0,750</td>
<td>0,333</td>
<td>0,556</td>
<td>0,643</td>
<td>0,500</td>
<td>0,167</td>
<td>0,000</td>
</tr>
<tr>
<td>Bergum / Midsland</td>
<td>0,500</td>
<td>0,500</td>
<td>0,429</td>
<td>0,813</td>
<td>0,500</td>
<td>0,571</td>
<td>0,333</td>
<td>0,500</td>
<td>0,429</td>
<td>0,567</td>
<td>0,751</td>
<td>0,000</td>
</tr>
<tr>
<td>Jorwerd / Midsland</td>
<td>0,592</td>
<td>0,667</td>
<td>0,583</td>
<td>0,846</td>
<td>0,454</td>
<td>0,667</td>
<td>0,400</td>
<td>0,600</td>
<td>0,571</td>
<td>0,592</td>
<td>0,500</td>
<td>0,000</td>
</tr>
<tr>
<td>Bakkeveen / Midsland</td>
<td>0,400</td>
<td>0,500</td>
<td>0,438</td>
<td>0,706</td>
<td>0,385</td>
<td>0,563</td>
<td>0,357</td>
<td>0,385</td>
<td>0,438</td>
<td>0,579</td>
<td>0,533</td>
<td>0,000</td>
</tr>
<tr>
<td>Waskemeer / Midsland</td>
<td>0,438</td>
<td>0,526</td>
<td>0,556</td>
<td>0,818</td>
<td>0,500</td>
<td>0,588</td>
<td>0,471</td>
<td>0,533</td>
<td>0,471</td>
<td>0,552</td>
<td>0,588</td>
<td>0,000</td>
</tr>
<tr>
<td>Kloosterburen / Midsland</td>
<td>0,500</td>
<td>0,412</td>
<td>0,611</td>
<td>0,810</td>
<td>0,563</td>
<td>0,357</td>
<td>0,529</td>
<td>0,600</td>
<td>0,333</td>
<td>0,636</td>
<td>0,706</td>
<td>0,000</td>
</tr>
<tr>
<td>Warffum / Midsland</td>
<td>0,563</td>
<td>0,438</td>
<td>0,667</td>
<td>0,737</td>
<td>0,625</td>
<td>0,429</td>
<td>0,588</td>
<td>0,643</td>
<td>0,400</td>
<td>0,652</td>
<td>0,600</td>
<td>0,000</td>
</tr>
<tr>
<td>Leermens / Midsland</td>
<td>0,667</td>
<td>0,652</td>
<td>0,739</td>
<td>0,550</td>
<td>0,773</td>
<td>0,650</td>
<td>0,739</td>
<td>0,722</td>
<td>0,389</td>
<td>0,455</td>
<td>0,667</td>
<td>0,000</td>
</tr>
<tr>
<td>Groningen / Midsland</td>
<td>0,714</td>
<td>0,682</td>
<td>0,714</td>
<td>0,636</td>
<td>0,783</td>
<td>0,762</td>
<td>0,800</td>
<td>0,778</td>
<td>0,471</td>
<td>0,476</td>
<td>0,684</td>
<td>0,000</td>
</tr>
<tr>
<td>Nieuw-Scheemda / Midsland</td>
<td>0,650</td>
<td>0,682</td>
<td>0,650</td>
<td>0,652</td>
<td>0,773</td>
<td>0,762</td>
<td>0,739</td>
<td>0,722</td>
<td>0,556</td>
<td>0,368</td>
<td>0,647</td>
<td>0,000</td>
</tr>
<tr>
<td>Langeio / Midsland</td>
<td>0,727</td>
<td>0,524</td>
<td>0,739</td>
<td>0,652</td>
<td>0,792</td>
<td>0,650</td>
<td>0,760</td>
<td>0,647</td>
<td>0,550</td>
<td>0,500</td>
<td>0,700</td>
<td>0,000</td>
</tr>
</tbody>
</table>
this yields a $267 \times 67$ matrix with one row per location and one column per linguistic variable, i.e. locations = individuals and linguistic phenomena = variables

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

step 2: apply multidimensional scaling (MDS) to the distance matrix

MDS is a mathematical technique for reducing a multidimensional distance matrix to a low dimensional space in which each point represents an object from the distance matrix, and distances between points represents, as well as possible, dissimilarities between objects (Borg and Groenen, 2005)
2-dimensional MDS-representation 67 verb cluster phenomena

---

This talk in one slide
Introduction
The data: dialect Dutch verb clusters
Research questions
Theoretical background: dialectometry
Methodology: reverse dialectometry
Results
Main conclusion
Extensions
References
Quantity and quality in linguistic variation

Jeroen van Craenenbroeck

This talk in one slide

Introduction

The data: dialect Dutch verb clusters

Research questions

Theoretical background: dialectometry

Methodology: reverse dialectometry

Results

Main conclusion

Extensions

References
this yields a $267 \times 67$ matrix with one row per location and one column per linguistic variable, i.e. locations $=$ individuals and linguistic phenomena $=$ variables

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

step 2: apply multidimensional scaling (MDS) to the distance matrix

step 3: project the data back onto a geographical map
Introduction

The data: dialect Dutch verb clusters

Research questions

Theoretical background: dialectometry

Methodology: reverse dialectometry

Results

Main conclusion

Extensions

References
note: the linguistic variables (i.e. cluster orders) are used to determine the degree of similarity/difference between dialect locations

these similarities and differences are then projected back onto a geographical map, which makes it possible to discern dialect regions based on what verb cluster phenomena they possess

shortcomings of this approach for our 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 and to what extent

2. there is no link between the data that feed into the quantitative analysis and the formal theoretical literature on verb clusters
Methodology: 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. theoretical analyses of verb cluster orders are decomposed in their constitutive parts, which makes it possible to include them as supplementary variables in the analysis
starting point: a $31 \times 267$ data table whereby each cluster order represents a row and each dialect location a column
Quantity and quality in linguistic variation

Jeroen van Craenenbroeck

This talk in one slide

Introduction

The data: dialect Dutch verb clusters

Research questions

Theoretical background: dialectometry

Methodology: reverse dialectometry

Results

Main conclusion

Extensions

References

<table>
<thead>
<tr>
<th></th>
<th>Midsland</th>
<th>Lies</th>
<th>West.Tersch</th>
<th>Oosterend</th>
<th>Hollum</th>
<th>Schiermonni</th>
<th>Ferwerd</th>
<th>Anjum</th>
<th>Kollum</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUX1(be.sg)-PART2</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>NA</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>PART2-AUX1(be.sg)</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>NA</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>AUX1(have.sg)-PART2</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>NA</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>PART2-AUX1(have.sg)</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>NA</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>AUX1(have.pl)-PART2</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>PART2-AUX1(have.pl)</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>MOD1[sg]-INF2</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>INF2-MOD1[sg]</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>MOD2-INF3-MOD1[sg]</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>MOD1[sg]-MOD2-INF3</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>MOD1[sg]-INF3-MOD2</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>INF3-MOD2-MOD1[sg]</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>INF3-MOD1[sg]-MOD2</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>MOD1[sg]-AUX2(have)-PART3</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>NA</td>
<td>no</td>
</tr>
<tr>
<td>PART3-MOD1[sg]-AUX2(have)</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>NA</td>
<td>yes</td>
</tr>
<tr>
<td>PART3-AUX2(have)-MOD1[sg]</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>NA</td>
<td>yes</td>
</tr>
<tr>
<td>AUX1(be.sg)-AUX2(go)-INF3</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>NA</td>
<td>no</td>
</tr>
<tr>
<td>AUX1(be.sg)-INF3-AUX2(go)</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>NA</td>
<td>no</td>
</tr>
<tr>
<td>AUX2(go)-AUX1(be.sg)-INF3</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>NA</td>
<td>no</td>
</tr>
<tr>
<td>AUX2(go)-INF3-AUX1(be.sg)</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>NA</td>
<td>no</td>
</tr>
<tr>
<td>INF3-AUX1(be.sg)-AUX2(go)</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>NA</td>
<td>no</td>
</tr>
<tr>
<td>INF3-AUX2(go)-AUX1(be.sg)</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>NA</td>
</tr>
<tr>
<td>AUX1(have.sg)-MOD2(inf)-INF3</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>AUX1(have.sg)-INF3-MOD2(part)</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>AUX1(have.sg)-INF3-MOD2(inf)</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>MOD2(inf)-INF3-AUX1(have.sg)</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>INF3-AUX1(have.sg)-MOD2(inf)</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>INF3-AUX1(have.sg)-MOD2(part)</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>INF3-MOD2(part)-AUX1(have.sg)</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>INF3-MOD2(inf)-AUX1(have.sg)</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
</tbody>
</table>
starting point: a $31\times 267$ data table whereby each cluster order represents a row and each dialect location a column

the dialect locations are now used to determine the degree of difference/similarity between the various cluster orders $\rightarrow$ each of the 31 cluster orders is compared to each other cluster order on 267 variables (i.e. as many as there are dialect locations)

when we reduce the 31-dimensional distance matrix to a two-dimensional space, we can plot the differences and similarities between the cluster orders from the SAND-project
Quantity and quality in linguistic variation

Jeroen van Craenenbroeck

This talk in one slide

Introduction

The data: dialect
Dutch verb clusters

Research questions

Theoretical background: dialectometry

Methodology: reverse dialectometry

Results

Main conclusion

Extensions

References
- **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 grammar, i.e. grammatical microparameters, then verb cluster orders that are near one another 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

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

(10) \[ \begin{align*}
\text{VP}_1 \quad & \quad \text{VP}_2 \\
\text{V}_1 \quad & \quad \text{V}_2
\end{align*} \]

(11) \[ \begin{align*}
\text{VP}_1 \quad & \quad \text{VP}_2 \\
\text{V}_1 \quad & \quad \text{V}_2 \quad \text{VP}_3 \\
\text{V}_2 \quad & \quad \text{V}_3
\end{align*} \]
Barbiers (2005) derives verb cluster orders as follows:

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

\[
\begin{align*}
(12) & \\
& \text{VP}_1 \\
& | \\
& \text{VP}_2 \qquad \text{V}'_1 \\
& | \\
& \text{V}_2 \quad \text{V}_1 \quad t_{\text{VP}_2} \\
(13) & \\
& \text{VP}_1 \\
& | \\
& \text{VP}_2 \qquad \text{V}'_1 \\
& | \\
& \text{V}_2 \quad \text{VP}_3 \quad \text{V}_1 \quad t_{\text{VP}_2}
\end{align*}
\]
Barbiers (2005) derives verb cluster orders as follows:

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

(14)

(15)
Barbiers (2005) derives verb cluster orders as follows:

- VP-intraposition can pied-pipe other material → derives 321 (movement of VP3 to specVP1 via specVP2 and with pied-piping of VP2)

(16)
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.

\[
(17) \quad \left[ \text{VP}_1 \text{mod}[u_{\text{Event}}] \left[ \text{VP}_2 \text{mod}[u_{\text{Event}}] \left[ \text{VP}_3 \text{inf}[i_{\text{Event}}] \right] \right] \right]
\]

\[
(18) \quad \left[ \text{VP}_1 \text{aux}[u_{\text{Perf}}] \left[ \text{VP}_2 \text{mod}[i_{\text{Perf}},u_{\text{Event}}] \left[ \text{VP}_3 \text{inf}[i_{\text{Event}}] \right] \right] \right]
\]
from this theoretical account we can distill the following micro-parameters:

- [±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?

and the 31 SAND cluster orders can be encoded in terms of these micro-parameters
<table>
<thead>
<tr>
<th>AUX1(be.sg)-PART2</th>
<th>Barbiers-base</th>
<th>Barbiers-movement</th>
<th>Barbiers-spec-pied-piping</th>
<th>Barbiers-feature checking-failure</th>
</tr>
</thead>
<tbody>
<tr>
<td>PART2-AUX1(be.sg)</td>
<td>noBase</td>
<td>yesMvt</td>
<td>noPiedP</td>
<td>noFeatCheckFail</td>
</tr>
<tr>
<td>AUX1(have.sg)-PART2</td>
<td>yesBase</td>
<td>noMvt</td>
<td>noPiedP</td>
<td>noFeatCheckFail</td>
</tr>
<tr>
<td>PART2-AUX1(have.sg)</td>
<td>noBase</td>
<td>yesMvt</td>
<td>noPiedP</td>
<td>noFeatCheckFail</td>
</tr>
<tr>
<td>AUX1(have.pl)-PART2</td>
<td>yesBase</td>
<td>noMvt</td>
<td>noPiedP</td>
<td>noFeatCheckFail</td>
</tr>
<tr>
<td>PART2-AUX1(have.pl)</td>
<td>noBase</td>
<td>yesMvt</td>
<td>noPiedP</td>
<td>noFeatCheckFail</td>
</tr>
<tr>
<td>MOD1(sg)-INF2</td>
<td>yesBase</td>
<td>noMvt</td>
<td>noPiedP</td>
<td>noFeatCheckFail</td>
</tr>
<tr>
<td>INF2-MOD1(sg)</td>
<td>noBase</td>
<td>yesMvt</td>
<td>noPiedP</td>
<td>noFeatCheckFail</td>
</tr>
<tr>
<td>MOD1(sg)-MOD2-INF3</td>
<td>yesBase</td>
<td>noMvt</td>
<td>noPiedP</td>
<td>noFeatCheckFail</td>
</tr>
<tr>
<td>MOD1(sg)-INF3-MOD1(sg)</td>
<td>noBase</td>
<td>yesMvt</td>
<td>noPiedP</td>
<td>noFeatCheckFail</td>
</tr>
<tr>
<td>MOD3-MOD2-MOD1(sg)</td>
<td>noBase</td>
<td>yesMvt</td>
<td>yesPiedP</td>
<td>noFeatCheckFail</td>
</tr>
<tr>
<td>INF3-MOD1(sg)-MOD2</td>
<td>noBase</td>
<td>yesMvt</td>
<td>noPiedP</td>
<td>noFeatCheckFail</td>
</tr>
<tr>
<td>MOD1(sg)-AUX2(have)-PART3</td>
<td>yesBase</td>
<td>noMvt</td>
<td>noPiedP</td>
<td>noFeatCheckFail</td>
</tr>
<tr>
<td>MOD1(sg)-PART3-AUX2(have)</td>
<td>noBase</td>
<td>yesMvt</td>
<td>noPiedP</td>
<td>noFeatCheckFail</td>
</tr>
<tr>
<td>PART3-MOD1(sg)-AUX2(have)</td>
<td>noBase</td>
<td>yesMvt</td>
<td>noPiedP</td>
<td>noFeatCheckFail</td>
</tr>
<tr>
<td>AUX1(have.go)-AUX2(go)-INF3</td>
<td>yesBase</td>
<td>noMvt</td>
<td>noPiedP</td>
<td>noFeatCheckFail</td>
</tr>
<tr>
<td>AUX1(have.go)-INF3-AUX2(go)</td>
<td>noBase</td>
<td>yesMvt</td>
<td>noPiedP</td>
<td>noFeatCheckFail</td>
</tr>
<tr>
<td>AUX2(go)-AUX1(be.sg)-INF3</td>
<td>noBase</td>
<td>yesMvt</td>
<td>noPiedP</td>
<td>noFeatCheckFail</td>
</tr>
<tr>
<td>AUX2(go)-INF3-AUX1(be.sg)</td>
<td>noBase</td>
<td>yesMvt</td>
<td>noPiedP</td>
<td>noFeatCheckFail</td>
</tr>
<tr>
<td>AUX3-AUX2(go)-AUX1(be.sg)</td>
<td>noBase</td>
<td>yesMvt</td>
<td>noPiedP</td>
<td>noFeatCheckFail</td>
</tr>
<tr>
<td>AUX1(have.sg)-MOD2(inf)-INF3</td>
<td>yesBase</td>
<td>noMvt</td>
<td>noPiedP</td>
<td>noFeatCheckFail</td>
</tr>
<tr>
<td>AUX1(have.sg)-INF3-MOD2(part)</td>
<td>noBase</td>
<td>yesMvt</td>
<td>noPiedP</td>
<td>noFeatCheckFail</td>
</tr>
<tr>
<td>AUX1(have.sg)-INF3-MOD2(inf)</td>
<td>noBase</td>
<td>yesMvt</td>
<td>noPiedP</td>
<td>noFeatCheckFail</td>
</tr>
<tr>
<td>MOD2(inf)-INF3-AUX1(have.sg)</td>
<td>noBase</td>
<td>yesMvt</td>
<td>noPiedP</td>
<td>noFeatCheckFail</td>
</tr>
<tr>
<td>INF3-AUX1(have.sg)-MOD2(inf)</td>
<td>noBase</td>
<td>yesMvt</td>
<td>noPiedP</td>
<td>noFeatCheckFail</td>
</tr>
<tr>
<td>INF3-AUX1(have.sg)-MOD2(part)</td>
<td>noBase</td>
<td>yesMvt</td>
<td>noPiedP</td>
<td>noFeatCheckFail</td>
</tr>
<tr>
<td>INF3-MOD2(part)-AUX1(have.sg)</td>
<td>noBase</td>
<td>yesMvt</td>
<td>noPiedP</td>
<td>noFeatCheckFail</td>
</tr>
<tr>
<td>INF3-MOD2(inf)-AUX1(have.sg)</td>
<td>noBase</td>
<td>yesMvt</td>
<td>noPiedP</td>
<td>noFeatCheckFail</td>
</tr>
</tbody>
</table>
Barbiers’s microparameters thus serve as additional variables in our data table

**in total:** 70 additional variables distilled from the theoretical literature on verb clusters:

- a head-initial head movement analysis, a head-final head movement analysis, a head-initial XP-movement analysis, a head-final XP-movement analysis (all based on Wurmbrand (2005))
- 17 additional variables based on the theoretical literature, but not linked to a specific analysis

**note:** all these variables are encoded as supplementary variables, i.e. they do not contribute to the construction of the verb cluster plot, but they can be mapped/matched against it
there are three ways of testing how well a supplementary (linguistic) variable lines up with the output of the geographical analysis:

1. visual inspection of a color-coded plot
This talk in one slide

Introduction

The data: dialect Dutch verb clusters

Research questions

Theoretical background: dialectometry

Methodology: reverse dialectometry

Results

Main conclusion

Extensions

References
there are three ways of testing how well a supplementary (linguistic) variable lines up with the output of the geographical analysis:

1. visual inspection of a color-coded plot
2. $\eta^2$ (squared correlation ratio): value between 0 and 1 indicating the strength of the link between the dimension and a particular categorical variable; can be interpreted as the percentage of variation on the dimension that can be explained by the categorical variable

<table>
<thead>
<tr>
<th>Hypothetical Variable</th>
<th>dimension 1</th>
<th>dimension 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.861</td>
<td>0.043</td>
</tr>
</tbody>
</table>

- **word of caution:** $\eta^2$ also goes up if the number of possible values of the categorical variable goes up (Richardson (2011)) → safest option is to look for variables with a high $\eta^2$ and only two or three possible values
there are three ways of testing how well a supplementary (linguistic) variable lines up with the output of the geographical analysis:

1. visual inspection of a color-coded plot
2. $\eta^2$ (squared correlation ratio): value between 0 and 1 indicating the strength of the link between the dimension and a particular categorical variable; can be interpreted as the percentage of variation on the dimension that can be explained by the categorical variable
3. v-test: value indicating whether a value of a categorical variable varies significantly from 0 on a particular dimension; significant values are $<-2$ or $>2$

<table>
<thead>
<tr>
<th>HypotheticalVariable</th>
<th>dimension 1</th>
<th>dimension 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>-5.082</td>
<td>-1.130</td>
</tr>
<tr>
<td>no</td>
<td>5.082</td>
<td>1.130</td>
</tr>
</tbody>
</table>
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

- 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

- **proposal (II)**: the identity of those parameters = the interpretation of the dimensions
The number of relevant dimensions

- recall: the analysis reduces a multi-dimensional distance matrix into a low-dimensional space, *while retaining as much as possible of the information present in the original object*

- we can plot the percentage of variance explained per dimension (= scree plot)
Percentage of variance explained per dimension

Number of dimensions

Percentage of variance explained

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

Quantity and quality in linguistic variation

Jeroen van Craenbroeck

This talk in one slide

Introduction

The data: dialect Dutch verb clusters

Research questions

Theoretical background: dialectometry

Methodology: reverse dialectometry

Results

Introduction

The number of relevant dimensions

Dimension 1

Dimension 2

Dimension 3

Conclusion

Main conclusion

Extensions

References
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
Dimension 1

- highest $\eta^2$-values:

<table>
<thead>
<tr>
<th>dimension 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>BarBen.NomInf 0.425</td>
</tr>
<tr>
<td>Bader.VMod 0.398</td>
</tr>
</tbody>
</table>

- BarBen.NomInf: Barbiers and Bennis (2010): the infinitival main verb is nominalized
- Bader.VMod: Bader (2012): the complement of a modal verb precedes the modal

- confirmed by $v$-test:

<table>
<thead>
<tr>
<th>$v$-test dimension 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>yesNomInf -3.303</td>
</tr>
<tr>
<td>yesVMod -3.090</td>
</tr>
</tbody>
</table>
Quantity and quality in linguistic variation

Jeroen van Craenbroeck

This talk in one slide

Introduction

The data: dialect

Dutch verb clusters

Research questions

Theoretical background: dialectometry

Methodology: reverse dialectometry

Results

Introduction

The number of relevant dimensions

Dimension 1

Dimension 2

Dimension 3

Conclusion

Main conclusion

Extensions

References
Introduction

The number of relevant dimensions

Dimension 1

Dimension 2

Dimension 3

Conclusion

Main conclusion

Extensions

References
note: while both variables propose a general split that seems well represented on dimension 1, there are a number of verb clusters orders for which they are irrelevant (because the cluster doesn’t contain the relevant configuration)

let’s try to strengthen our interpretation of dimension 1 by incorporating those points

Abels (2011): there seems to be a correlation between the position of infinitives vis-à-vis modals on the one hand and the position of participles vis-à-vis auxiliaries on the other
new 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

η² of InfMod.AuxPart: 0.6142

<table>
<thead>
<tr>
<th>v-test dimension 1</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>InfMod.AuxPart_yes</td>
<td>-4.292</td>
</tr>
<tr>
<td>InfMod.AuxPart_no</td>
<td>4.292</td>
</tr>
</tbody>
</table>
Quantity and quality in linguistic variation

Jeroen van Craenenbroeck

This talk in one slide

Introduction

The data: dialect Dutch verb clusters

Research questions

Theoretical background: dialectometry

Methodology: reverse dialectometry

Results

Introduction

The number of relevant dimensions

Dimension 1

Dimension 2

Dimension 3

Conclusion

Main conclusion

Extensions

References

Dimension 1 vs. the new InfMod.AuxPart-variable
- **note:** this new variable aligns very nicely with the first dimension
  - two recalcitrant cluster orders:
    - AUX2(go)-AUX1(be.sg)-INF3: possibly spurious: this is an order that seems excluded in any cluster in Dutch (only two hits in the whole of SAND)
    - MOD2(inf)-INF3-AUX1(have.sg)

- **this means** that the first (and most important) source of variation in Dutch verb clusters—i.e. the first microparameter—concerns the placement of modals and auxiliaries vs. the verbs they select

- it sets apart dialects that consistently place infinitives to the right and participles to the left from those that don’t
Dimension 2

- highest $\eta^2$-values:
  
<p>| |</p>
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
</table>
  | SchmiVo.MAPhc            | 0.379  
  | Barbiers.base.generation | 0.309  

- SchmiVo.MAPhc: Schmid and Vogel (2004): “If A and B are sister nodes at LF, and A is a head and B is a complement, then the correspondent of A precedes the one of B at PF.”
- Barbiers.base.generation: Barbiers (2005): head-initial base structure

- confirmed by v-test:
  
<p>| |</p>
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
</table>
  | Barbiers.base.generation_no | 3.044  
  | SchmiVo.MAPhc_starMAPhc    | 2.855  
  | SchmiVo.MAPhc_okMAPhc      | -3.044  
  
References
Quantity and quality in linguistic variation

Jeroen van Craenenbroeck

This talk in one slide

Introduction

The data: dialect Dutch verb clusters

Research questions

Theoretical background: dialectometry

Methodology: reverse dialectometry

Results

Introduction

The number of relevant dimensions

Dimension 1

Dimension 2

Dimension 3

Conclusion

Main conclusion

Extensions

References
Introduction
The number of relevant dimensions
Dimension 1
Dimension 2
Dimension 3
Conclusion
Main conclusion
Extensions
References
note: just as was the case with dimension 1, the variables culled from the literature leave room for improvement in interpreting dimension 2

another variable that does well is slope ($\eta^2 = 0.422$): is the order ascending, descending, first-ascending-then-descending, or first-descending-then-ascending?
Quantity and quality in linguistic variation

Jeroen van Craenenbroeck

This talk in one slide

Introduction

The data: dialect Dutch verb clusters

Research questions

Theoretical background: dialectometry

Methodology: reverse dialectometry

Results

Introduction

The number of relevant dimensions

Dimension 1

Dimension 2

Dimension 3

Conclusion

Main conclusion

Extensions

References

---

Dimension 2 vs. slope

Dimension 2 (11.89%)

Dimension 1 (53.51%)
- **note:** ascDesc and desc pattern towards the positive values of dimension 2, while asc and descAsc tend to yield negative values for this dimension
- **new variable:** FinalDescent:
  - set to ‘yes’ if the cluster ends in a descending order
  - set to ‘no’ if it ends in an ascending order

<table>
<thead>
<tr>
<th>FinalDescent_yes</th>
<th>FinalDescent_no</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>12</td>
</tr>
<tr>
<td>132</td>
<td>123</td>
</tr>
<tr>
<td>321</td>
<td>312</td>
</tr>
<tr>
<td>231</td>
<td>213</td>
</tr>
</tbody>
</table>
Quantity and quality in linguistic variation

Jeroen van Craenenbroeck

This talk in one slide

Introduction

The data: dialect Dutch verb clusters

Research questions

Theoretical background: dialectometry

Methodology: reverse dialectometry

Results

Introduction

The number of relevant dimensions

Dimension 1

Dimension 2

Dimension 3

Conclusion

Main conclusion

Extensions

References
- $\eta^2$ of FinalDescent: 0.382

<table>
<thead>
<tr>
<th></th>
<th>v-test dimension 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>FinalDescent_yes</td>
<td>3.387</td>
</tr>
<tr>
<td>FinalDescent_no</td>
<td>-3.387</td>
</tr>
</tbody>
</table>

- possible theoretical interpretation of FinalDescent: it 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)
  - caveat: two-verb clusters $\rightarrow$ there are only two possible orders, so you can always get from one to the other with one movement operation

- **this means** that the second source of variation in Dutch verb clusters—i.e. the second microparameter—concerns the degree to which a cluster order diverges from a strictly head-final order
Dimension 3

- highest $\eta^2$-values:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>dimension 3</td>
<td></td>
</tr>
<tr>
<td>SchmiVo.MAPch</td>
<td>0.701</td>
</tr>
<tr>
<td>Bader.base.order</td>
<td>0.686</td>
</tr>
</tbody>
</table>

- SchmiVo.MAPch: Schmid and Vogel (2004): “If A and B are sister nodes at LF, and A is a head and B is a complement, then the correspondent of B precedes the one of A at PF.”

- Bader.base.order: Bader (2012): a strictly head-final base order

- confirmed by v-test:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>v-test dimension 3</td>
<td></td>
</tr>
<tr>
<td>SchmiVo.MAPch.okMAPch</td>
<td>4.537</td>
</tr>
<tr>
<td>Bader.base.order.yes</td>
<td>4.537</td>
</tr>
</tbody>
</table>
Introduction
The data: dialect Dutch verb clusters
Research questions
Theoretical background: dialectometry
Methodology: reverse dialectometry
Results
Introduction
The number of relevant dimensions
Dimension 1
Dimension 2
Dimension 3
Conclusion
Main conclusion
Extensions
References
there is a very strong correlation between a head-final base order and the third dimension in the analysis

this means that the third source of variation in Dutch verb clusters—i.e. the third microparameter—concerns the question of whether a dialect diverges from a strictly head final order or not
Results: Conclusion

- interpreting the first three dimensions of the quantitative analysis of the verb cluster data in the Syntactic Atlas of Dutch Dialects allows us to construct the following (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: [±ModInf&PartAux] (dimension 1)
Main conclusion

- the considerable variation found in Dutch verb cluster orders can be reduced to three grammatical microparameters:
  1. the order of modals and auxiliaries vs. the verbs they select: $[\pm \text{ModInf}\&\text{PartAux}]$
  2. the degree of divergence from a head-final order: $[\text{EconomyOfMovement}]$
  3. adherence to a head-final order or not: $[\pm \text{Movement}]$

- 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
Extensions

▶ **empirical (I):**
  ▶ extend the analysis with more data from Dutch (written SAND-questionnaire) and other Germanic languages and varieties such as German and Swiss German

▶ **empirical (II):**
  ▶ apply these same techniques to other empirical domains

▶ **theoretical:**
  ▶ combine the three microparameters uncovered by the quantitative analysis into one, coherent theoretical account of verb clusters
  ▶ use additional statistical techniques (cluster analysis, association rule mining) to further explore the data
  ▶ find a way to compare not just individual theoretical hypotheses but rather entire analyses of verb cluster ordering
References I


References II


