Applying automatically parsed corpora to the study of language variation
Jelke Bloem, Arjen Versloot, Fred Weerman
Language variation

- Grammars often contain optionality
  - Same meaning, different form
- Grammatical variation phenomena are multivariate in nature
  - Rules can only outline the variation
- On what basis do we choose between the options?
Dative alternation

- A word order variation in English:
  1. He gave [his friend] [the ticket]
  2. He gave [the ticket] to [his friend]

- No simple rule on when to use one or the other
- Probabilistically modeled using 14 variables
  i.e. animacy of recipient, pronominality of recipient, given-ness (Bresnan et al., 2007)
- Switchboard corpus (3M words, 2360 instances)
Automatically parsed corpora

- Fewer annotation resources required
  - More data
- Exact definition of construction
- Flexible
- Contains errors (‘random’ or systematic)
- Annotation may be a limiting factor
Case study: Verbal clusters

- A word order variation in Dutch:
  1. ik denk dat ik het begrepen heb
     I think that I it understood have
  2. ik denk dat ik het heb begrepen
     I think that I it have understood

- Frisian, German: Only green order
Explaining the variation (Coussé et al, 2008)

- (Regional) linguistic background
  - Single speaker variation?
- Mode of communication
- Semantic factor
- Discourse factor
  - Priming by previous syntactic structures
Corpus study (de Sutter, 2009)

- “De Standaard” part of CONDIV corpus (3.2M words)
- Controlled for regional, register and diachronic variation
- Strict cluster criteria:
  - Only ‘hebben’ (to have) ‘zijn’ and ‘worden’ (to be) auxiliaries
  - Only complement clauses with ‘dat’ (that)
- Multivariate logistic regression model (10 variables)
- 2,390 manually verified clusters, 66.99% red order
Large-scale analysis

- Too limited definition of ‘verbal cluster’ by de Sutter (2009)
  - Unnecessary in a multivariate model
- Can be scaled up using large, automatically annotated corpora
  - Larger sample size
  - Coverage of more cluster types
Automatically annotated corpus

- Wikipedia part of “Lassy Large” corpus
- 145M tokens, 411.623 clusters, 71.65% red order
- Syntactic annotation lets us formally define various types of clusters using DACT (X-path queries)
  - Dependency trees (+ features)
  - May contain errors: 88.38% parser accuracy

understood have | have understood

Applying automatically parsed corpora to the study of language variation
Variables of the original study (de Sutter, 2009)

- Accented syllable distance
  - naar hun au-to is ge-lo-pen
  - n=4

- Separable main verb
  - heeft afgewassen (has washed up)

- Constituent after cluster
  - heeft gezien dat het gebeurde

- Length of the middle field
  - dat [hij naar hun auto] is gelopen
  - n=4

- Type of auxiliary
  - copular-zijn/passive-zijn/time/worden

- Syntactic persistence
  - afgewassen heeft en (...) weggelopen is

- Main verb frequency
  - naar hun auto is gelopen

- Pre-verbal constituent: Informativity and inherence
  - understood have | have understood
  - Applying automatically parsed corpora to the study of language variation
Variables in Lassy Large corpus

- Accented syllable distance
- Separable main verb
- Constituent after cluster
- Length of the middle field
- Type of auxiliary
- Syntactic persistence
- Main verb frequency
- Pre-verbal constituent: Informativity and inherence

understood have | have understood
Additional cluster types

- Main clause
  - Judges kunnen in principe niet worden ontslagen.
  - Judges can in principle not be dismissed.

- Infinitival clusters
  - ... waardoor de reclame weer op tv te zien was
  - ... thus the ad again on tv to see was

- Aux/mod worden hebben zijn / kunnen zullen willen laten mogen moeten blijven hoeven
  - ... dat iedereen hem ongestraft doden mocht
  - ... that everyone him with impunity kill may

understood have | have understood
## Model comparison

<table>
<thead>
<tr>
<th>Feature</th>
<th>De Sutter (2009)</th>
<th>This study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Separable main verb</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Yes</td>
<td>3.87</td>
<td>4.92</td>
</tr>
<tr>
<td>Constituent after cluster</td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Complement of main verb</td>
<td>0.47</td>
<td>-</td>
</tr>
<tr>
<td>Complement of preverbal noun</td>
<td>1.21</td>
<td>-</td>
</tr>
<tr>
<td>Comp. or adjunct of main verb</td>
<td>-</td>
<td>51.44</td>
</tr>
<tr>
<td>Comp. or adjunct of preverbal N.</td>
<td>-</td>
<td>0.44</td>
</tr>
<tr>
<td>Length of middle field</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-2 words</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>3-5 words</td>
<td>2.03</td>
<td>2.42</td>
</tr>
<tr>
<td>6-8 words</td>
<td>2.29</td>
<td>3.23</td>
</tr>
<tr>
<td>9-11 words</td>
<td>2.29</td>
<td>3.34</td>
</tr>
<tr>
<td>12-14 words</td>
<td>2.57</td>
<td>3.33</td>
</tr>
<tr>
<td>&gt;14 words</td>
<td>1.98</td>
<td>3.15</td>
</tr>
</tbody>
</table>

### Feature

<table>
<thead>
<tr>
<th>Feature</th>
<th>De Sutter (2009)</th>
<th>This study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of auxiliary</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Copular zijn</td>
<td>1.00</td>
<td>-</td>
</tr>
<tr>
<td>Auxiliary of time</td>
<td>18.30</td>
<td>-</td>
</tr>
<tr>
<td>Passive zijn</td>
<td>7.82</td>
<td>-</td>
</tr>
<tr>
<td>zijn</td>
<td>-</td>
<td>1.00</td>
</tr>
<tr>
<td>worden</td>
<td>11.73</td>
<td>1.19</td>
</tr>
<tr>
<td>hebben</td>
<td>-</td>
<td>2.19</td>
</tr>
<tr>
<td>modal</td>
<td>-</td>
<td>132.42</td>
</tr>
<tr>
<td>Main verb frequency</td>
<td></td>
<td></td>
</tr>
<tr>
<td>β</td>
<td>(from CELEX)</td>
<td>(Lassy Large)</td>
</tr>
<tr>
<td></td>
<td>$2.44_{-06}$</td>
<td>$3.73_{-08}$</td>
</tr>
<tr>
<td></td>
<td>$7.74_{-07}^{**}$</td>
<td>$1.05_{-08}^{***}$</td>
</tr>
<tr>
<td>Inherence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Yes</td>
<td>2.26</td>
<td>2.10</td>
</tr>
<tr>
<td>Information value</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Intermediate</td>
<td>1.41</td>
<td>1.21</td>
</tr>
<tr>
<td>High</td>
<td>1.94</td>
<td>1.11</td>
</tr>
</tbody>
</table>

The values are **odds ratios**, measuring effect size.

OR=2.00 means: if this feature is present instead of the baseline, red order is 2 times more probable.

understood have | have understood

---

Applying automatically parsed corpora to the study of language variation
# Model predictive power

Concordance index $c$

<table>
<thead>
<tr>
<th>Model</th>
<th>C-index</th>
<th>Nr. of features</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>De Sutter (2009)</td>
<td>0.8030</td>
<td>10</td>
<td>AUX/Sub only</td>
</tr>
<tr>
<td>Full model</td>
<td>0.8635</td>
<td>9</td>
<td>All clusters</td>
</tr>
<tr>
<td>Small model</td>
<td>0.7649</td>
<td>7</td>
<td>AUX/Sub only</td>
</tr>
</tbody>
</table>

Full model intercept = 0.6035

* Values actually not directly comparable
* $c=0.5$ is chance level, but the gold standard is not 1…
Stepwise regression

- Minimize Akaike Information Criterion (AIC)
- Indicates relative importance of the features

<table>
<thead>
<tr>
<th>Feature</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0. &lt;none&gt;</td>
<td>490828</td>
</tr>
<tr>
<td>1. Type of auxiliary</td>
<td>413913</td>
</tr>
<tr>
<td>2. Constituent after cluster</td>
<td>349852</td>
</tr>
<tr>
<td>3. Finiteness</td>
<td>338758</td>
</tr>
<tr>
<td>4. Length middle field</td>
<td>332781</td>
</tr>
<tr>
<td>5. Clause type</td>
<td>325857</td>
</tr>
<tr>
<td>6. Frequency main verb</td>
<td>324371</td>
</tr>
<tr>
<td>7. Inherence</td>
<td>323201</td>
</tr>
<tr>
<td>8. Separable verb</td>
<td>322519</td>
</tr>
<tr>
<td>9. Information value</td>
<td>322000</td>
</tr>
</tbody>
</table>

Applying automatically parsed corpora to the study of language variation
Replication summary

- Effect sizes largely similar to previous work
- Variables hold within a bigger model
- Cluster order is more or less affected by all variables
- Some variables could not be measured
### Additional features

<table>
<thead>
<tr>
<th>Feature</th>
<th>De Sutter (2009)</th>
<th>This study</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Infinitival clusters</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>-</td>
<td>1.00</td>
</tr>
<tr>
<td>Yes</td>
<td>-</td>
<td>0.03</td>
</tr>
<tr>
<td><strong>Clause type</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subordinate clause</td>
<td>-</td>
<td>1.00</td>
</tr>
<tr>
<td>Main clause</td>
<td>-</td>
<td>0.34</td>
</tr>
</tbody>
</table>

- **Red pure-infinitival cluster (but only with *hoeven* ‘need’)**
  
  ... zodat de machinist niet in de locomotief *zelf hoeft te zijn*
  
  ... so that the operator not in the locomotive itself need to be

- **Red main clause**
  
  Rechters *kunnen in principe niet worden ontslagen*.
  
  Judges can in principle not be dismissed.

| understood have | have understood |
Dutch Europarl corpus
as part of Lassy Large corpus

- European Parliament proceedings texts
- 138,304 clusters, 86.78% red order!
- Variable effects largely similar

<table>
<thead>
<tr>
<th>Feature</th>
<th>Europarl model</th>
<th>Wiki model</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>zijn</em></td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td><em>worden</em> (to be)</td>
<td>1.62</td>
<td>1.19</td>
</tr>
<tr>
<td><em>hebben</em> (to have)</td>
<td>2.57</td>
<td>2.19</td>
</tr>
<tr>
<td>modal</td>
<td>323.46</td>
<td>132.42</td>
</tr>
<tr>
<td>None</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Comp. or adjunct of main verb</td>
<td>31.22</td>
<td>51.44</td>
</tr>
<tr>
<td>Comp. or adjunct of preverbal noun</td>
<td>0.46</td>
<td>0.44</td>
</tr>
</tbody>
</table>

understood have | have understood
Semantic factor: Collostructional analysis

(Stefanowitsch & Gries, 2003)

- Relationship between a construction (red/green) and the words that fill its slots

```
... that I it VERB have
... dat ik het begrepen heb
... dat ik het gezien heb
... dat ik het gehoord heb
... dat ik het geschopt heb
```

- Calculate most strongly associated **collexemes**

  - Fisher’s Exact Test or other association measure

```
understood have | have understood
```

Applying automatically parsed corpora to the study of language variation
Collostructional analysis
Auxiliary, subordinate clause clusters only, cutoff=15

<table>
<thead>
<tr>
<th>Main verbs</th>
<th>Odds ratio - Red - Green</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - --- afkondigen</td>
<td>inf 29 0</td>
</tr>
<tr>
<td>2 - --- neerzetten</td>
<td>inf 24 0</td>
</tr>
<tr>
<td>3 - --- uitmaken</td>
<td>inf 21 0</td>
</tr>
<tr>
<td>4 - --- aanhouden</td>
<td>inf 21 0</td>
</tr>
<tr>
<td>5 - --- optekenen</td>
<td>inf 19 0</td>
</tr>
<tr>
<td>6 - --- overgeven</td>
<td>inf 18 0</td>
</tr>
<tr>
<td>7 - --- aanschaffen</td>
<td>inf 17 0</td>
</tr>
<tr>
<td>8 - --- uitschrijven</td>
<td>inf 16 0</td>
</tr>
<tr>
<td>9 - --- plaatsvinden</td>
<td>33.34 182 3</td>
</tr>
<tr>
<td>10--- indienen</td>
<td>22.95 42 1</td>
</tr>
</tbody>
</table>

understood have | have understood
### Collostructional analysis

Auxiliary, subordinate clause clusters only, cutoff=100, no particle verbs

<table>
<thead>
<tr>
<th>Main verbs</th>
<th>Odds ratio</th>
<th>Red</th>
<th>Green</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 -- verplichten (to oblige) ST</td>
<td>20.44</td>
<td>13</td>
<td>182</td>
</tr>
<tr>
<td>2 -- zien (to see) ST</td>
<td>17.36</td>
<td>148</td>
<td>1751</td>
</tr>
<tr>
<td>3 -- danken (to thank)</td>
<td>14.02</td>
<td>20</td>
<td>288</td>
</tr>
<tr>
<td>4 -- vinden (to find) ST</td>
<td>13.96</td>
<td>87</td>
<td>830</td>
</tr>
<tr>
<td>5 -- herkennen (to recognize) ST</td>
<td>7.08</td>
<td>20</td>
<td>97</td>
</tr>
<tr>
<td>6 -- relateren (to relate) ST</td>
<td>6.70</td>
<td>22</td>
<td>101</td>
</tr>
<tr>
<td>7 -- huwen (to marry)</td>
<td>5.94</td>
<td>28</td>
<td>11</td>
</tr>
<tr>
<td>8 -- besmetten (to infect)</td>
<td>4.87</td>
<td>24</td>
<td>80</td>
</tr>
<tr>
<td>9 -- wijten (to blame)</td>
<td>4.33</td>
<td>32</td>
<td>95</td>
</tr>
<tr>
<td>10-- bestemmen (to assign)</td>
<td>4.28</td>
<td>61</td>
<td>179</td>
</tr>
</tbody>
</table>

### Patterns: Semantic classes? Stative/dynamic verbs?

| understood have | have understood |

Applying automatically parsed corpora to the study of language variation
Conclusions

- Replicated and extended a linguistic study using an automatically annotated corpus
- Comprehensive model of Dutch verbal clusters
- Automatic approach is easily extended
  - Study regional/register/diachronic variation
- de Sutter (2009)’s variables generalize to another domain
- Larger sample allows more detailed analysis
Discussion

- Extend further:
  - Corpus of Spoken Dutch
  - A corpus with author/region/time metadata
  - Larger verbal clusters

- Semantics: more can be done

- Automatically annotated corpora for:
  - Dative alternation
  - ‘that’-optionality
  - Any other probabilistic phenomenon

- New types of corpora as NLP tools get better

understood have | have understood

Applying automatically parsed corpora to the study of language variation
References


