Using automatically annotated corpora in language variation research

Jelke Bloem, Arjen Versloot, Fred Weerman
Language variation

- Grammars often contain optionality
  - Same meaning, different form
  1. He said [that] he would do it

- 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
  - Dutch LASSY Small corpus: 1M tokens
  - Dutch LASSY Large corpus: 700M tokens
- Flexible
- Exact definition of construction
- Contains errors (‘random’ or systematic)
- Annotation may constrain what can be researched
Case study: Dutch 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
Manual corpus study (de Sutter, 2009)

- “De Standaard” part of CONDIV corpus (3.2M words)
- Controlled for regional, register and diachronic variation, specific cluster types

Select data -> Semi-automatically retrieve data -> Manually verify results

- Multivariate logistic regression model (10 variables)
- 2,390 manually verified clusters, 66.99% red order

understood have  |  have understood
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)
  - Limited to existing annotation
  - May contain errors: 88.38% parser accuracy
Automatic annotation

Limitations of the annotation:
- Accented syllable distance
- ‘to be’ = passive, copulative, temporal auxiliary…

Workflow: Choose corpus -> Define searches ->
Automatically retrieve sentences -> Automatically extract features

//node[@lemma="hebben" and (some $x in //node[@pt="ww"]
  satisfies (number(@begin) = $x/number(@end)))]

understood have | have understood

Using automatically annotated corpora in language variation research
Using GrETEL (Augustinus, Vandeghinste, and Van Eynde 2012)

■ Example-based treebank querying:

\[ \text{ik denk dat ik het heb begrepen} \]

understood have | have understood

Using automatically annotated corpora in language variation research
Using GrETEL: Output

Using automatically annotated corpora in language variation research

understood have   have understood
Using DACT: Query-based... querying
De Kok (2010)

Same query syntax: Insert the query that GrETEL produces

understood have | have understood
Verbal cluster study: Results

- 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>

Using automatically annotated corpora in language variation research
### Results: 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
* The gold standard is not 1…

<table>
<thead>
<tr>
<th>understood have</th>
<th>have understood</th>
</tr>
</thead>
</table>
Large-scale: Collostructional analysis
(Stefanowitsch & Gries, 2003)

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

... that I it **VERB** have

<table>
<thead>
<tr>
<th>Main verbs</th>
<th>Odds ratio</th>
<th>Red</th>
<th>Green</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 --- verplichten</td>
<td>20.44</td>
<td>13</td>
<td>182</td>
</tr>
<tr>
<td>(to oblige)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 --- zien</td>
<td>17.36</td>
<td>148</td>
<td>1751</td>
</tr>
<tr>
<td>(to see)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 --- danken</td>
<td>14.02</td>
<td>20</td>
<td>288</td>
</tr>
<tr>
<td>(to thank)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 --- vinden</td>
<td>13.96</td>
<td>87</td>
<td>830</td>
</tr>
<tr>
<td>(to find)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 --- herkennen</td>
<td>7.08</td>
<td>20</td>
<td>97</td>
</tr>
<tr>
<td>(to recognize)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

... that I it **have** **VERB**

<table>
<thead>
<tr>
<th>Main verbs</th>
<th>Odds ratio</th>
<th>Red</th>
<th>Green</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 --- staan</td>
<td>7.81</td>
<td>583</td>
<td>51</td>
</tr>
<tr>
<td>(to stand)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 --- gaan</td>
<td>6.74</td>
<td>751</td>
<td>76</td>
</tr>
<tr>
<td>(to go)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 --- hebben</td>
<td>6.40</td>
<td>882</td>
<td>94</td>
</tr>
<tr>
<td>(to have)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 --- zitten</td>
<td>5.70</td>
<td>200</td>
<td>24</td>
</tr>
<tr>
<td>(to sit)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 --- zijn</td>
<td>5.50</td>
<td>2583</td>
<td>317</td>
</tr>
<tr>
<td>(to be)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

understood have | have understood
Verbal cluster study summary

- Variable effects largely similar to previous work
- Variables hold within a bigger model
- Variables hold in other domain: Europarl corpus
- All variables in the study are associated with cluster word order
- Some variables could not be measured
- Detailed results on our poster or in Bloem, Versloot & Weerman (2014)
Conclusions

- Automatically annotated corpora are particularly useful for language variation studies
- Replicated and extended a manual linguistic study
- Larger sample allows more detailed analysis
- Automatic approach is easily extended
  - Study regional/register/diachronic variation
- Example-based querying, standard query syntax and standard annotation format help accessibility
Discussion

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

- Extend the study:
  - Corpus of Spoken Dutch
  - A corpus with writer/region/time metadata
  - Larger verbal clusters

- New types of corpora as NLP tools get better
References


understood have | have understood