Lexical preferences in Dutch verbal cluster ordering

Jelke Bloem, Arjen Versloot, Fred Weerman
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

- No apparent meaning difference
Word order choice

- Optionality in the grammar of Dutch verb clusters
- On what basis do we choose between options?

- Information structure (i.e. focus)
- Processing efficiency
- Meaning
- Stored patterns/preferences?

understood have | have understood

Lexical preferences in Dutch verbal cluster ordering | 3
Word order choice

- Previous corpus studies tested general factors:
  - Choice of auxiliary verb
  - Length of sentence
  - Morphological complexity of the verb

- Lexical associations not studied
  - ‘bedoeld heeft’ or ‘heeft bedoeld’?
  - ‘gestaan heeft’ or ‘heeft gestaan’?
Lexical associations

■ Can some of the variation be explained by studying lexical associations?

Take a very big corpus and find:
Verbs that occur significantly more in the red order
Verbs that occur significantly more in the green order
Lexical associations and preferences

- **Lexical association**: Association in corpus

- **Lexical preference**: Association in mental lexicon
  A word that is more often heard in one of the two possible constructions, may also be produced more often in that construction.

- **Semantic preference**: Association in mental lexicon
  A construction may prefer verbs with certain meaning properties: constructions are pairs of form and meaning

- **Other factors**
  e.g. morphological complexity, length of sentence
Automatically annotated corpus

- Wikipedia part of “Lassy Large” corpus
- 145M tokens, 411.623 clusters, 71.65% red order
- Necessary to find enough instances of lexical verbs
- May contain errors: 88.38% parser accuracy

understood have | have understood

Lexical preferences in Dutch verbal cluster ordering
Collostructional analysis (Stefanowitsch & Gries, 2003)

- Method for identifying the meaning of a construction
- Measures the 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
...

... that I it have VERB
... dat ik het heb gemaakt
... dat ik het heb bedacht
... dat ik het heb gehoord
... dat ik het heb beschreven
...
Collostructional analysis (Stefanowitsch & Gries, 2003)

… 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 …

- If some words are more strongly associated with one order, we found a lexical association

- Associated words are called collexemes

-> Like collocations, but with a word and a construction, instead of a word and a word
Collostructional analysis (Stefanowitsch & Gries, 2003)

… that I it VERB AUX … that I it AUX VERB

- Distinctive collexeme analysis
- Calculate most strongly associated collexemes
  - Fisher’s Exact Test
- Semantically analyze lists of collexemes
  - Semantic classes?

"verbs denoting ways of speaking cleverly and deviously"

understood have | have understood
### Collostructional analysis

**Auxiliary, subordinate clause clusters only, cutoff=15**

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

**Pattern of morphologically complex verbs**

**understood have** | **have understood** | **Lexical preferences in Dutch verbal cluster ordering**
Morphological complexity factor

- Control for this by excluding them from analysis

understood have   |   have understood  
Lexical preferences in Dutch verbal cluster ordering
Collostructional analysis
Subordinate clause clusters only, ‘have’ only, cutoff=50

<table>
<thead>
<tr>
<th>Main verbs</th>
<th>Odds ratio</th>
<th>Red</th>
<th>Green</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 --- kampen</td>
<td>71.75</td>
<td>6</td>
<td>157</td>
</tr>
<tr>
<td></td>
<td>(to struggle)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 --- zeggen</td>
<td>44.43</td>
<td>4</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>(to say)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 --- danken</td>
<td>25.11</td>
<td>44</td>
<td>400</td>
</tr>
<tr>
<td></td>
<td>(to thank)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 --- lijden</td>
<td>12.16</td>
<td>32</td>
<td>142</td>
</tr>
<tr>
<td></td>
<td>(to suffer)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 --- munten</td>
<td>6.26</td>
<td>24</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>(to target)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 --- bedoelen</td>
<td>5.67</td>
<td>26</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>(to intend)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 --- bieden</td>
<td>5.43</td>
<td>69</td>
<td>137</td>
</tr>
<tr>
<td></td>
<td>(to offer)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Patterns: Semantic classes?
understood have | have understood

Lexical preferences in Dutch verbal cluster ordering | 13
Prevalence of the lexical associations

- How common are these associations?
  758 verbs, 59% is significant at $\alpha = 0.05$

  ‘worden’ and ‘zijn’ (to be) may have adjectival semantics

- Clusters with ‘have’ only:
  218 verbs, 51% is significant at $\alpha = 0.05$
  Bonferroni correction: 25.6% significant, $\alpha = 0.000229$
Semantic classes

- Semantic preferences of word orders?
- Usually said to be no meaning difference
- Pardoen (1991): Stative, dynamic interpretation

- Verb semantic classes from Cornetto database
  Analysis of top 20 red order and green order verb collexemes
  Properties of top senses of each verb

understood have | have understood
Semantic classes

<table>
<thead>
<tr>
<th>Property</th>
<th>Red order</th>
<th>Green order</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>16</td>
<td>15</td>
</tr>
<tr>
<td>Dynamic</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td>Attributive</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Spatial</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Cognition</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>Negative polarity</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>Positive polarity</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

Some semantic tendencies

Some similar words are adjacent:
‘give’+’take’, ‘eat’+’drink’
Conclusions

- Clear lexical associations observed in corpus data
- Even after controls, a large portion is significant
  - Indicates lexical preference
- Some semantic associations also observed
  - Not the ones predicted by Pardoen (1991)
  - Word orders are distinct constructions?
- Large corpus sample allows analysis at the lexical level

understood have | have understood
Discussion

- Extend semantic analysis:
  - Telicity not included, for example
  - Compute difference in distributional semantic network
- Include these results in a multifactorial model
- Method can be applied to many near-synonymous constructions
- Can the main verbs (non-heads) really influence cluster order variation?
References


understood have | have understood

Lexical preferences in Dutch verbal cluster ordering
<table>
<thead>
<tr>
<th>Property</th>
<th>Red order</th>
<th>Green order</th>
</tr>
</thead>
<tbody>
<tr>
<td>Val:Mono</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Val:Di</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>Val:Tri</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td>Intransitive</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Transitive</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td>Ditransitive</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>