Investigating *heel, erg, zeer* with PaQu

*Jan Odijk, Tin-dag, Utrecht, 2015-02-07*

(1) **Background**
   a. AutoSearch: (enrich,) upload & search (expected March 2015, INL)
      - PoS-tags, Corpus Modern Dutch interface
   b. PaQu: upload (, enrich) and search (July 2015, V1 available, RUG)
      - syntactic structures, Groningen Word Relations Search Application, User tests on-going

(2) **Basic Facts**
   a. *Heel erg zeer* are (near-)synonyms meaning *very*
   b. *Heel* can modify adjectival (A) predicates only
   c. *Erg en zeer* can modify A, verbal (V) and prepositional (P) predicates

(3) **Illustration**
   a. Hij is daar heel /erg / zeer blij over
   b. Hij is daar *heel / erg /zeer* in zijn sas mee
   c. Dat verbaast mij *heel / erg /zeer* 
      *(very in English is like Dutch *heel* (v. *very much*))
      See [Odijk 2011, 2014] for more data and qualifications

(4) **Assessment of the facts**
   a. Distinction is purely syntactic
   b. Cannot be derived from semantic differences
   c. No correlation found with other known facts (see (5))
   d. Cannot be derived from general (universal) principles
   e. ➔ must be acquired by L1 learners of Dutch

(5) **Correlation with other differences?** Unlikely! Then we must find *{heel} v. {erg, zeer}.*
   But we only find other oppositions:
   a. Meaning? *{heel, zeer} v. {erg}*
   b. Adverbial Inflection? *{heel, erg} v. {zeer}*
      - Hele / erge / *zere* grote handen
   c. Comparative / Superlative? *{heel, zeer} v. {erg}*
      - *heler / erger / *zeerder* ziek (dan Piet)
      - *Het heelst / het ergst / *het zeerst* ziek
   d. Modification by *heel erg zeer*? *{heel, zeer} v. {erg}*
      - *heel heel* / *erg heel / * zeer heel ziek
      - Heel erg / *erg zeer / * zeer erg ziek
      - *Heel zeer / *erg zeer / *zeer zeer* ziek
   e. Early use by children? *{heel} v. {erg} v. {zeer} [Odijk 2014]*
   f. Formality? *{zeer}{(formal)} v. *{heel, erg}*(neutral)

\[1\text{ Well-formed as repeated modifiers}\]
(6) Research Questions
a. How can children acquire the fact that *erg* and *zeer* can modify A, V and P predicates (in L1 acquisition)?
b. How can children acquire the fact that *heel* can modify A but can NOT modify V and P predicates (in L1 acquisition)?
c. What kind of evidence do children have access to for acquiring such properties?
d. Is there a relation with the time of acquisition?
e. Is there a role for indirect negative evidence (absence of evidence interpreted as evidence for absence)?
f. …

(7) CHILDES corpora
a. Use Dutch CHILDES corpora to investigate this
b. Problem: ambiguity of the relevant words
c. Dutch CHILDES corpora do NOT have (reliable) pos-tags and no syntactic parses at all
d. Done manually for Van Kampen Corpus [Odijk 2014:91]
e. PaQu (Parse and Query) automates this

(8) Ambiguity of *heel, erg, zeer*

<table>
<thead>
<tr>
<th>word</th>
<th>Morphosyntax</th>
<th>Syntax</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>heel</td>
<td>A Mod N</td>
<td>(1) ‘whole’ (2) ‘large’</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mod A</td>
<td>‘very’</td>
<td></td>
</tr>
<tr>
<td>Vf</td>
<td>Mod N</td>
<td>(1) ‘heal’ (2) ‘receive’</td>
<td></td>
</tr>
<tr>
<td>erg</td>
<td>N trutum</td>
<td>‘erg’</td>
<td></td>
</tr>
<tr>
<td></td>
<td>N neutrum</td>
<td>‘evil’</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A Mod N, predc</td>
<td>‘bad’, ‘awful’</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mod A V P</td>
<td>‘Very’</td>
<td></td>
</tr>
<tr>
<td>zeer</td>
<td>N</td>
<td>‘pain’</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A Mod N, predc</td>
<td>‘painful’</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mod A V P</td>
<td>‘very’</td>
<td></td>
</tr>
</tbody>
</table>

(9) PaQu
a. Search for morpho-syntactic information and syntactic dependency relations
b. Distinction relevant ones v. irrelevant ones can now be made mostly automatically


Small Experiment (was intended as a user test)

a. Take all Dutch CHILDES corpora
b. Select all adult utterances containing heel, erg or zeer

c. Clean the utterances, e.g.
   • ja, maar <we be> /// we bewaren (he)t ook ➔
   • ja, maar we bewaren het ook

d. Gather statistics and draw conclusions

Accuracy

a. Manual annotation of Van Kampen corpus used as gold standard (Acc)
b. Alpino makes finer distinctions: I mapped these
c. Annotation errors in the gold standard: revised gold standard (Rev Acc)

<table>
<thead>
<tr>
<th>word</th>
<th>Acc</th>
<th>Rev Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>heel</td>
<td>0.94</td>
<td>0.95</td>
</tr>
<tr>
<td>erg</td>
<td>0.88</td>
<td>0.91</td>
</tr>
<tr>
<td>zeer</td>
<td>0.21</td>
<td>0.21</td>
</tr>
</tbody>
</table>

Caveats

a. It concerns (cleaned) adult speech
b. It concerns relatively short sentences, explicitly separated
c. It mostly concerns a very local grammatical relation
d. Most problematic for zeer: zeer doen

Results

<table>
<thead>
<tr>
<th>Results</th>
<th>mod A</th>
<th>mod N</th>
<th>Mod V</th>
<th>mod P</th>
<th>predc</th>
<th>other</th>
<th>unclear</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>heel</td>
<td>886</td>
<td>46</td>
<td>2</td>
<td>2</td>
<td>14</td>
<td>0</td>
<td>2</td>
<td>952</td>
</tr>
<tr>
<td>erg</td>
<td>347</td>
<td>27</td>
<td>109</td>
<td>0</td>
<td>187</td>
<td>5</td>
<td>0</td>
<td>675</td>
</tr>
<tr>
<td>zeer</td>
<td>7</td>
<td>1</td>
<td>83</td>
<td>0</td>
<td>19</td>
<td>21</td>
<td>7</td>
<td>138</td>
</tr>
</tbody>
</table>

Interpretation

a. Overwhelming # examples for mod A for heel
b. Large # examples for mod A and mod V for erg
c. Very few examples for zeer (mod V mostly wrong parses )
d. No examples of mod P / mod V for heel at all (the 4 are wrong parses)
e. PP predicates with zeer, erg: op prijs stellen, in de smaak vallen only (analyzed as mod V) – 3 occurrences
Conclusions

a. Linguistics
   • No examples for mod P: how to explain *heel* v. *erg*, *zeer*?
     Overwhelmingness of mod A for *heel*?
   • Are the current Dutch CHILDES corpora representative enough to draw reliable conclusions?

b. PaQu
   • PaQu is very useful for doing better and more efficient manual verification of hypotheses
   • In some cases its fully automatically generated parses and their statistics can reliably be used directly (though care is required!)

Future Work

a. Similar experiments for the children’s speech (cf. [Odijk 2014:34])
b. Similar experiments for *te* v. *overmatig*; *worden* v. *raken* and others
c. Extend PaQu to include all relevant ‘metadata’
d. Extend PaQu to natively support common formats such as CHAT, Folia, TEI, …
e. Make similar system for GrETEL (http://nederbooms.ccl.kuleuven.be/eng/gretel)
f. Manually verify (parts of) parses for CHILDES corpora (UU AnnCor project)

References