Background

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

**Heel erg zeer are** (near-)synonyms meaning ‘very’

**Heel** can modify adjectival (A) predicates only

**Erg** and **zeer** can modify A, verbal (V) and prepositional (P) predicates

1. (A) Hij is daar heel erg zeer blij over
2. (P) Hij is daar *heel / erg / zeer in zijn sas mee
3. (V) 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

Assessment of the facts

- Distinction is purely syntactic
- Cannot be derived from semantic differences
- No correlation found with other known facts
- Cannot be derived from general (universal) principles
- *must be acquired by L1 learners of Dutch*

Research Questions

- How can children acquire the fact that **erg** and **zeer** can modify A, V and P predicates (in L1 acquisition)?
- How can children acquire the fact that **heel** can modify A but cannot modify V and P predicates (in L1 acquisition)?
- What kind of evidence do children have access to for acquiring such properties?
- Is there a relation with the time of acquisition?
- Is there a role for indirect negative evidence (absence of evidence interpreted as evidence for absence)?

CHILDES corpora

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

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

PaQu

- Search for morpho-syntactic information and syntactic dependency relations
  - Distinction relevant ones v. irrelevant ones can now be made mostly automatically
  - [http://zardoz.service.rug.nl:8067/]

Small Experiment (was intended as a user test)

- Take all Dutch CHILDES corpora
- Select all adult utterances containing heel, erg or zeer
- Clean the utterances, e.g.:
  - ja, maar we bewaren (he) ook →
  - ja, maar we bewaren het ook
- Gather statistics and draw conclusions

Accuracy

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

<table>
<thead>
<tr>
<th></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>14</td>
<td>0</td>
<td>2</td>
<td>952</td>
<td></td>
</tr>
<tr>
<td>erg</td>
<td>347</td>
<td>27</td>
<td>109</td>
<td>14</td>
<td>0</td>
<td>2</td>
<td>675</td>
<td></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>138</td>
<td></td>
</tr>
</tbody>
</table>

Caveats

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

Interpretation

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

<table>
<thead>
<tr>
<th></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>14</td>
<td>0</td>
<td>2</td>
<td>952</td>
<td></td>
</tr>
<tr>
<td>erg</td>
<td>347</td>
<td>27</td>
<td>109</td>
<td>14</td>
<td>0</td>
<td>2</td>
<td>675</td>
<td></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>138</td>
<td></td>
</tr>
</tbody>
</table>

- 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?
- 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

- Similar experiments for the children’s speech (cf. [Odijk 2014:34])
- Similar experiments for te v. overmatig; worden v. roken and others
- Extend PaQu to include all relevant ‘metadata’
- Extend PaQu to natively support common formats such as CHAT, Folia, TEI, ...
- Make similar system for GrETEL
- Manually verify (parts of) parses for CHILDES corpora (UU AnnCor project)

References