Linguistic Research with PaQu

www.clarin.nl

Jan Odijk, Utrecht University

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* en *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

word	Morphosyntax	Syntax	Meaning
heel	A	Mod N	(1) `whole' (2) `large'
		Mod A	`very'
	Vf		(1) `heal' (2) `receive'
erg	N utrum		`erg'
	N neutrum		`evil'
	A	Mod N, predc	'bad', 'awful'
		Mod A V P	very
zeer	N		`pain'
	A	Mod N, predc	'painful'
		Mod A V P	'very'

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 be> [//] we bewaren (he)t 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)

word	Acc	Rev Acc
heel	0.94	0.95
erg	0.88	0.91
zeer	0.21	0.21

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*: *zeer doen*

Results	mod A	mod N	Mod V	mod P	predc	other	unclear	Total
heel	886	46	*2*	*2*	14	0	2	**952**
erg	347	27	109	0	187	5	0	**675**
zeer	7	1	*83*	0	19	21	7	**138**

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

Conclusions

- 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. *raken* 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

[**Odijk 2011**] Odijk, J. , "User Scenario Search", internal CLARIN-NL document, April 13, 2011. [docx]

[**Odijk 2014**] Odijk, J. ,'CLARIN: What's in it for Linguists?', Uilendag lecture, Utrecht, Mar 27, 2014. [pptx]