Using PaQu for language acquisition research

Jan Odijk

CLARIN 2015 Conference

Wroclaw, 2015-10-16

Overview

- Introduction
- CHILDES Corpora
- PaQu
- Evaluation & Analysis
- Conclusions
- Future Work

Introduction

Cat	init	modifier	predicate	rest
A	Hij is daar	Heel / erg /zeer	blij	mee
gloss	He is there	very	happy	with
P	Hij is daar	*heel / erg / zeer	in zijn sas	mee
gloss	He is there	very	happy	with
V	…omdat dat mij	*heel / erg / zeer	verbaast	
gloss	…because that me	very	surprises	

(See [Odijk 2011, 2014] for more data and qualifications

Introduction

- Distinction is purely syntactic
- Cannot be derived from semantic differences
- Correlation with other known facts unlikely
- Cannot be derived from general (universal) principles
- ➔ must be acquired by L1 learners of Dutch

Introduction

- Minimal pair in acquisition
- Requires acquisition of negative property
  - No evidence in the input
  - No 'correction' or correction ignored
- May provide evidence for/against relevant hypotheses
  - E.g. Indirect Negative Evidence hypothesis
    - Absence of evidence ➔ evidence for absence

Corpus Analysis

- Problem: Ambiguity
  - *Heel* 7-fold ambiguous
  - *Erg* 4-fold ambiguous
  - *Zeer* 3-fold ambiguous
- (as any decent natural language word)
- For our purposes:
  - Morpho-syntactic and syntactic properties resolve the ambuigities

Corpus Analysis

- [Odijk 2014]
  - Automatic Corpus analysis: GrETEL, OpenSONAR, COAVA , LWRS, CMD
  - These apply to specific corpora only
  - **Manual** Corpus analysis of CHILDES Van Kampen Corpus
  - How can I apply these applications to my own corpus?
  - ➔ request for PaQu (extends LWRS), AutoSearch (extends CMD), …

PaQu

- PaQu= Parse and Query: https://dev.clarin.nl/node/4182
- Web application made by Groningen University
  - Upload corpus
    - Plain text or in Alpino format
  - Plain Text is automatically parsed by Alpino
  - Resulting treebank can be searched and analyzed
- Search
  - Word relations interface and XPATH Queries
- Analysis
  - User-definable statistics on search results (and metadata)

Experiments

- Take the Dutch CHILDES corpora
- Select all 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
- Upload it into PaQu
- Gather statistics and draw conclusions

Experiment 1

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

Experiment 1: Results

- Accuracy

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

Experiment 1: Interpretation

- Good for *heel, erg*
- Bad for *zeer,* but:
  - Completely due to *zeer doen* (lit. pain(ful) do, 'to hurt')
  - Can be identified very easily in PaQu
- Generalisability: Limited
  - It concerns (cleaned) adult speech
  - It concerns relatively short sentences, explicitly separated
  - It mostly concerns a very local grammatical relation

Experiment 2:

- All adults' utterances:

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**

Experiment 2: Interpretation

- *Heel* most frequent (almost 54%)

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**

Experiment 2: Interpretation

- *Heel* as mod A overwhelming: > 93%

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**

Experiment 2: Interpretation

- *Heel* as mod V, mod P wrong analysis

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**

Experiment 2: Interpretation

- Mod A and mod V more balanced for *erg*

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**

Experiment 2: Interpretation

- Evidence for *zeer* mostly lacking
- Cases of Mod V are mostly wrong analyses

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	3	0	19	21	7	**138**

Experiment 2: Interpretation

- Evidence for Mod P mostly lacking
- Some evidence for *erg, zeer* (4 occurrences)

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**

Experiment 3:

- Van Kampen Children's speech: Accuracy
- Similar to the Adults' speech but slightly lower

Word	Acc
heel	0.90
erg	0.73
zeer	0.17

Conclusions

- Linguistics:
  - No examples for mod P: how to explain *heel* v. *erg, zeer*?
  - Overwhelmingness of mod A for *heel* might be a relevant factor
  - Current Dutch CHILDES corpora probably too small to draw reliable conclusions

Conclusions

- PaQu:
  - PaQu is very useful for doing better and more efficient manual verification of hypotheses
  - In some cases its parses and their statistics can reliably be used directly (though care is required!)
  - Several small details were improved, small additions to functionality made through these experiments

Future Work

- More experiments for the children's speech (cf. [Odijk 2014:34])
- Similar experiments for other examples
  - *te* 'too' v. *overmatig 'excessively'; worden* 'become'v. *raken 'get'* 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, OpenSONAR
- Manually verify (parts of) parses for CHILDES corpora

(most is being done in CLARIAH-NL or UU AnnCor)

Thanks for Attention!

Visit the Demo at 16:30!

Visit the Bazaar at 14:30 for a completely different use of PaQu!

Correlation with other Differences?

Phenomenon	Opposes	Versus
Mod V,P	heel	erg, zeer
Meaning	erg	heel, zeer
Inflection	heel, erg	zeer
Comparative, Superlative	erg	heel, zeer
Modifiee	erg	heel, zeer
Pragmatics	zeer	heel, erg

➔ NO!

Ambiguity: *heel*

word	Morpho-syntax	Syntax	Meaning
heel	A	Mod N	(1)`whole’ (2) ‘in one piece’ (3)`large’
		Predc	‘in one piece’
		Mod A	`very’
	Vf		(1)`heal’ (2) `receive’

Ambiguity: *erg*

word	Morpho-syntax	Syntax	Meaning
erg	N utrum		`erg’
	N neutrum		`evil’
	A	Mod N, predc	‘bad’, ‘awful’
		Mod A V P	very

Ambiguity: *zeer*

word	Morpho-Syntax	Syntax	Meaning
zeer	N		`pain’
	A	Mod N, predc	‘painful’
		Mod A V P	‘very’