Boosting Linguistic Research with CLARIN

Jan Odijk

ESSLLI 2018 Summerschool

Sofia, 2018-08-14

Overview

- Introduction
- CLARIN
- Corpus Search and Analysis
  - CHILDES Corpora, LASSY, CGN, BASILEX, SoNaR
- Towards Analysis
- Concluding Remarks
- 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 differences 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 ‘corrections’ or corrections ignored
- May provide evidence for/against relevant hypotheses
  - E.g. Indirect Negative Evidence hypothesis
    - Absence of evidence ➔ evidence for absence

Introduction

- Problem is not specific to these words

word1	selection1	word 2	selection2
even 'as'	Mod A	*even zeer* 'as much'	Mod A V P
te 'too'	Mod A	*te zeer* 'too much'	Mod A V P
vrij 'rather'	Mod A	*nogal* 'rather'	Mod A V P
worden 'become'	AP,NP	*raken* 'get'	AP, PP

Introduction

- Problem is not specific to Dutch:

word1	selection1	word 2	selection2
very	Mod A	*very much*	Mod A V P
as	Mod A	*as much*	Mod A V P
too	Mod A	*too much*	Mod A V P
become	AP,NP	*get*	AP, PP

First Language Acquisition

- Input is speech: speech → symbol sequence
- Spontaneous speech:
- No word boundaries → tokenization
- Find out (know?) that selection by PoS is relevant
- Determine the PoS tags of this language
- Assign PoS to each word
- Each of the 3 words is ambiguous: disambiguation
- Many of the candidate modifiees are ambiguous (esp. participles): disambiguation

First Language Acquisition

- Here we avoid most of these problems:
  - We start with orthographic transcriptions
  - Enriched with annotations for hesitations, filled pauses, retracings, … (not always correct)
- Ambiguity of the words is unavoidable
  - But we focus on only one meaning

CLARIN

- Analysis of these phenomena in corpora is desirable and partially required.
- CLARIN offers many applications to search in corpora
  - On-line web applications
  - Accessible from anywhere
  - No login or login via your own account (on-going)
  - Dedicated interfaces

Corpus Analysis

- Our interest: *heel*, *zeer*, *erg* meaning ‘very’
- Problem: Ambiguity
  - *Heel* 6-fold ambiguous
  - *Erg* 4-fold ambiguous
  - *Zeer* 3-fold ambiguous
- (as most natural language words)
- For our purposes:
  - Morpho-syntactic and syntactic properties resolve the ambiguities
  - ➔Searching in Treebanks / Parsebanks most desirable

PaQu

- PaQu= Parse and Query: https://dev.clarin.nl/node/4182
- Web application made by Groningen University
  - Search (in Dutch Treebanks: LASSY, CGN)
    - XPATH Queries
    - Search for dependencies (d, rel, h)
  - Analysis
    - User-definable statistics on search results (and metadata)
  - Upload corpus (login required)
    - Plain text, TEI, FoLIA, Alpino format, CHAT (soon)
    - Text is automatically parsed by Alpino
    - Resulting parsebank can be searched and analyzed

[Odijk et al. 2017]

GrETEL

- GrETEL= Greedy Extraction of Trees for Empirical Linguistics: http://gretel.ccl.kuleuven.be/gretel3/
- Web application made by Leuven University
  - Search (in Dutch Treebanks: LASSY, CGN)
    - XPATH Queries
    - Query by Example (QBE)
  - Analysis
    - None

[Augustinus et al. 2012]

GrETEL 4

- GrETEL 4: http://gretel.hum.uu.nl/gretel4/ng/home
- Extensions to GrETEL (on-going,Utrecht University):
  - Upload corpus
    - Plain text, CHAT, Alpino format; TEI, FoLIA (soon)
    - Text is automatically parsed by Alpino
    - Resulting parsebank can be searched and analyzed
  - Analysis
    - User-definable statistics on search results (and metadata)

[Odijk et al. 2018]

Uploaded Corpora

- Uploaded corpora parsed fully automatically
  - ➔Will contain errors
  - ➔Must assess the parsing accuracy for the phenomena being investigated
  - Or must do a manual analysis of a superset of the query results

CHILDES Parsing Accuracy: Adults

- Accuracy of parsing *heel, erg, zeer* in CHILDES Van Kampen Adult Speakers (compared a Gold standard)

word	Acc
heel	0.95
erg	0.91
zeer	0.21

CHILDES Parsing Accuracy: Children

- 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

CHILDES Parsing Accuracy

- 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 relatively short sentences, explicitly separated
  - It mostly concerns a very local grammatical relation

Parsing Accuracy

- All results are from manually verified treebanks or have initially been generated by PaQu and or GrETEL
- All results have been mapped to grammatical labels I want to distinguish
  - E.g. *vnw* (pronoun) ➔ A or N depending on the word
  - *Bw (adverb)* ➔ *A* or *P*
  - Past participles: always *ww* (verb) ➔ V or A
  - *Mwu* ➔ *N,A,V,* or *P*
- All results have been manually checked and if needed corrected (for some only a sample)
  - Analysis options make it possible to do this efficiently

Corpora

Corpus	#utts (K)	#tokens (M)	modality	spontaneity	formality
LASSY-Small	65	1	written	prepared	formal
CGN	130	1	spoken	mixed	mixed
VanKampenJAC	61	0,3	spoken	spontaneous	informal
VanKampen LAUorSAR	47	0,15	spoken	spontaneous	informal
CHILDES Dutch	545	1,9	spoken	spontaneous	informal
Basilex		13,5	written	prepared	formal
Wikipedia	8707	145	written	prepared	very formal
SoNaR	*41000*	*542*	*written*	*mixed*	*mixed*

Heel

Corpus /m tokens	mod A	mod V	mod P
LASSY-Small	295,6	0,0	0,0
CGN	2899,4	0,0	7,9
VanKampenJAC	2191,1	3,3	3,3
VanKampen LAUorSAR	1616,9	0,0	6,5
CHILDES Dutch	2512,4	3,2	8,5
Basilex	172,0	0,0	1,7
Wikipedia	90,5	0,0	0,3

Erg

Corpus / m tokens	mod A	mod V	mod P
LASSY-Small	156,0	13,7	5,5
CGN	324,6	78,1	13,2
VanKampenJAC	112,5	49,6	0,0
VanKampen LAUorSAR	77,6	6,5	6,5
CHILDES Dutch	189,7	44,1	2,1
Basilex	324,5	73,8	3,5
Wikipedia	128,3	12,2	2,1

Zeer

Corpus / m tokens	mod A	mod V	mod P
LASSY-Small	307,4	7,3	2,7
CGN	207,0	7,9	1,8
VanKampenJAC	6,6	6,6	0,0
VanKampen LAUorSAR	6,5	0,0	0,0
CHILDES Dutch	6,4	2,7	1,6
Basilex	26,7	1,7	0,3
Wikipedia	342,0	18,3	1,9

Main findings (1)

- *Heel 'very'*
  - modifies adjectival phrases and not generally V, P
  - But it can modify a small number of adverbial PPs
- Heel 'completely' (= *geheel*)
  - Is ill-formed in the standard language (except in some specific combinations (e.g. *ander(s)* 'other, different'*)*
  - Occurs in informal language among Flemish speakers
  - Is irrelevant for assessing the status of *heel* 'very'

Main findings(2)

- *Erg* and *zeer 'very'*
  - modify adjectival, verbal and adpositional phrases
- Frequency of *zeer* mod P
  - is very low
  - Comparable to frequency of *heel* modifying adverbial PPs
  - Modification potential of *heel* is not generalized to all PPs, but for *zeer* it is

SoNaR

- Main findings confirmed by (limited) searching in token-annotated SoNaR corpus, using OpenSONAR
- (Parsebank exists but cannot be searched yet efficiently)
- Investigated using adjacency + manual check
- *heel* mod P: 1.6 / million tokens
- *zeer* mod P: 3.0 / million tokens
- *heel* Mod P with Flemish speakers

CLARIN-provided Treebank Apps

- Can be used by ***all*** linguists:
  - Online available
    - No downloading and installing of software
    - No downloading and storing (increasingly larger) corpora
  - Dedicated interfaces:
    - No need for programming or even using query languages (dependency search, QBE)
    - No need to know the structure of the treebank in every fine detail

Towards Analysis

- CLARIN's Treebank Search Applications prepare the ground for
  - the actual analysis of the facts
  - formation of new hypotheses
  - comparison of existing hypotheses / theories, e.g.
    - Baker 1979, Berwick 1985, Bowerman 1988, Pinker 1989, Xu and Tenenbaum 2007, Stefanowich 2008, Yang 2016, …

Towards Analysis

- Some considerations:
  - Overwhelming amount of mod A for *heel:*
    - very high frequency, esp. in spoken language
    - very high proportion of mod A among mod A, mod V and Mod P: >99%)
    - Cf. in CHILDES:
      heel: 99.5 : 0.1 : 0.3; *erg*: 80.4 : 18.7 : 0.9; *zeer*: 60 : 25 : 15
  - *heel* mod P involves a small number of fixed combinations:
    - these but not these (with synonyms, closely related words, different Ps)
    - Probably to be described as exceptions, e.g. *heel*: mod PP[*aan het begin*]

Towards Analysis

- Tentative Proposal (unconscious acquisition)
  - a phenomenon must occur more than Θ_{min} to be taken into consideration
  - If one of the options for syntactic selection occurs more than Θ_{win} , it is made the only option.
  - Constructions violating the syntactic selection restrictions are stored as exceptions
  - Values of Θ_{min} and Θ_{win} to be determined:
    - $80.4\% < \Theta_{win}$;
    - $\Theta_{min} = 5$ (?)

Towards Analysis

- Tentative Proposal (continued)
  - Applied to *heel*, *erg*, *zeer*
    - *Heel*: mod A + exceptions
    - *erg*, *zeer:* mod A V P
  - To be tested and refined by
    - Analysis per child in CHILDES
      - VanKampenJac is compatible:
      - *Heel* = 99.7 : 0.2 : 0.2
      - *Erg* = 69.4 : 30.6 : 0.0
      - *Zeer* = 50.0 : 50.0 : 0.0

      Though no evidence for *zeer* mod P
    - Analysis of other examples
  - To be compared with proposals by others

Towards Analysis

- Yang's (2016:177) *Sufficiency Principle:*
  - Let *R* be a generalization over *N* items, of which *M* items are attested to follow *R*. *R* can be extended to all *N* items iff $N - M < \theta_N$, where $\theta_N = N / \ln N$
- Is probably inapplicable, or if applicable, false
- It requires a generalization *R* for a class of items
- The class of items cannot be characterized morphologically
  - v. the '*asleep*' class in English, Yang 2016:180

Towards Analysis

- Item class can be characterized semantically
  - Cf. double object verbs: ‘verbs of caused possession that involves the transfer of objects, entities, or abstract information’ (Yang 2016: 201)
  - As ‘degree modifiers’
- R1: Degree modifiers select for A
- R2: Degree modifiers select for A, V, or P.

Towards Analysis

- There are about 115 degree modifiers in Dutch
  - By the way, found with a different CLARIN application: Cornetto
  - Initial, tentative, assessment: Max. 15 combine with only with A, the rest with A, V, or P (N=115, for R2: M=100)
  - 115-100= 15< θ_{115} (= *115* / ln 115 ≈ 24)
    - but R2 does not generalize to all items

Concluding Remarks

- CLARIN makes search and analysis from corpora easy for all linguists
  - No downloading, installing software
  - No downloading, storing (huge) data needed
  - Dedicated interfaces avoid the need
    - To program
    - Or even to write queries
    - To have fine-grained knowledge of the exact structure of the treebank
- Linguistic Research can be
  - Accelerated
  - Based on many more data than ever before
- Illustrated here for one problem from Dutch

Concluding Remarks

- What about other languages?
- CLARIN offers a lot:
  - PMLTQ > 320 treebanks, > 70 lgs
  - Tündra >55 treebanks, > 45 lgs
  - INESS > 470 treebanks, > 70 lgs
  - All offer search applications, analysis tools, annotation tools, etc.
- BulTreeBank offers resources for Bulgarian
  - integrated into CLARIN .

Future Work

- Gather statistics in CHILDES per child
- Search in a parsebank for the SoNaR Corpus
- Similar experiments for other examples
  - *te* 'too' v. *overmatig/te zeer 'excessively'; worden* 'become'v. *raken 'get', even* 'as', and others
- Test and refine (or adapt) the tentative hypothesis suggested in this presentation

Future Work

- Manually verify parses for (parts of) CHILDES corpora
  - A start has been made in UU AnnCor project
- Further extensions to the Treebank Apps
  - Support for Universal Dependencies, Generating more Treebank Statistics, improved interface, set operations on search results, manual annotation of search results, full programming for analysis, …

Thanks for Your Attention!

References (1)

[Augustinus 2012*]* Liesbeth Augustinus, Vincent Vandeghinste, and Frank Van Eynde (2012). **"Example-Based Treebank Querying"**. In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012).* Istanbul, Turkey. pp. 3161-3167.

[Baker 1979] C.L. Baker. 1979. Syntactic theory and the projection problem. *Linguistic Inquiry,* 10(4):533–581.

[Berwick 1985] Robert Berwick. 1985. *The Acquisition of Syntactic Knowledge*. MIT Press, Cambridge, MA.

[Bowerman 1988] Melissa Bowerman. 1988. The 'no negative evidence' problem: how do children avoid constructing on overly general grammar? In Hawkins, John A. (ed.), *Explaining Language Universals* . Oxford and Cambridge, MA: Blackwell, 73–101.

[Odijk et al. 2017] Odijk, Jan, Noord, Gertjan van, Kleiweg, Peter & Tjong Kim Sang, Erik (28.12.2017). The Parse and Query (PaQu) Application. In Jan Odijk & Arjan van Hessen (Eds.), *CLARIN in the Low Countries* (pp. 281-297) (17 p.). London, UK: Ubiquity Press, DOI: http://dx.doi.org/10.5334/bbi.23.

References (2)

[Odijk et al. 2018] Jan Odijk, Martijn van der Klis and Sheean Spoel. 2018. 'Extensions to the GrETEL Treebank Query Application', in: Eduard Bejček (Ed.) *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16)*, pp. 46-55. Prague, Czech Republic. http://aclweb.org/anthology/W/W17/W17-7608.pdf

[Pinker 1989] Steven Pinker. 1989. *Learnability and Cognition*. MIT Press, Cambridge, MA

[Stefanowich 2008] Anatol Stefanowich. 2008. Negative entrenchment: A usage-based approach to negative evidence. *Cognitive Linguistics*, 19(3): 513-533.

[Tellings et al 2014] A. Tellings, Hulsbosch, M., Vermeer, A., and van den Bosch, A., `BasiLex: an 11.5 million words corpus of Dutch texts written for children', *Computational Linguistics in the Netherlands Journal*, vol. 4, pp. 191-208, 2014.

[Xu & Tenenbaum 2007] F. Xu and J.B. Tenenbaum. 2007. Word learning as Bayesian inference. *Psychological Review*, 114(2):245.

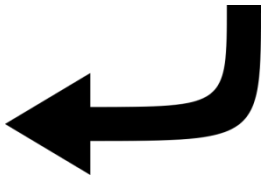
[Yang 2016] Charles Yang. 2016. *The Price of Productivity: How Children Learn to Break the Rules of Language*. MIT Press, Cambridge, Mass.

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

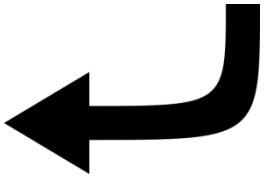
➔ NO!

Ambiguity: *HEEL*



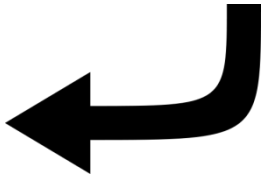
word	Morpho-syntax	Syntax	Meaning
heel	A	Mod N	(1)`whole’ (2) ‘in one piece’ (3)`large’
		Predc	‘in one piece’
		Mod A	`very’
	Vf		(1)`heal’ (2) `receive’

Ambiguity: *ERG*



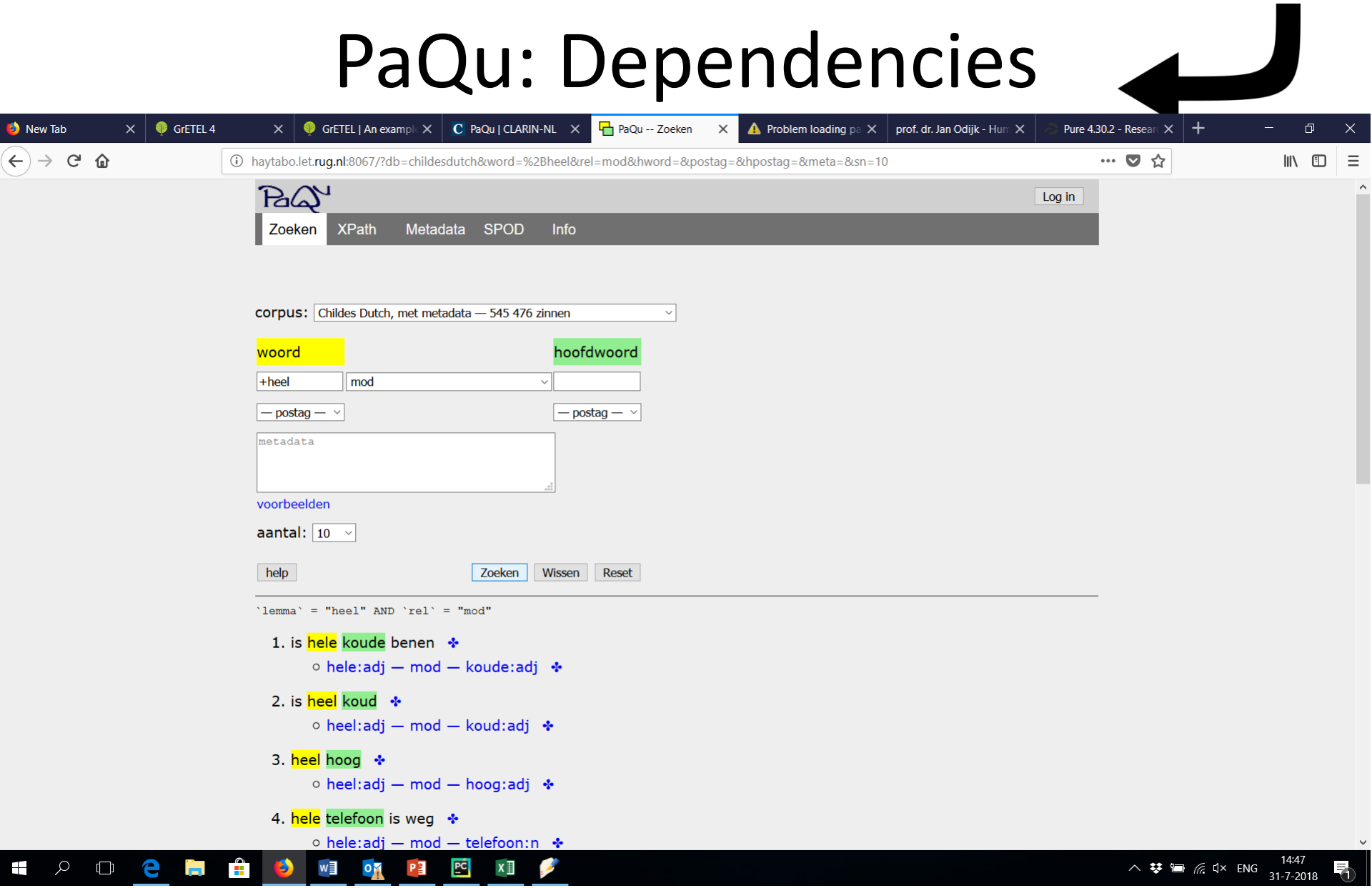
word	Morpho-syntax	Syntax	Meaning
erg	N utrum		`erg’
	N neutrum		`evil’
	A	Mod N, predc	‘bad’, ‘awful’
		Mod A V P	very

Ambiguity: *ZEER*

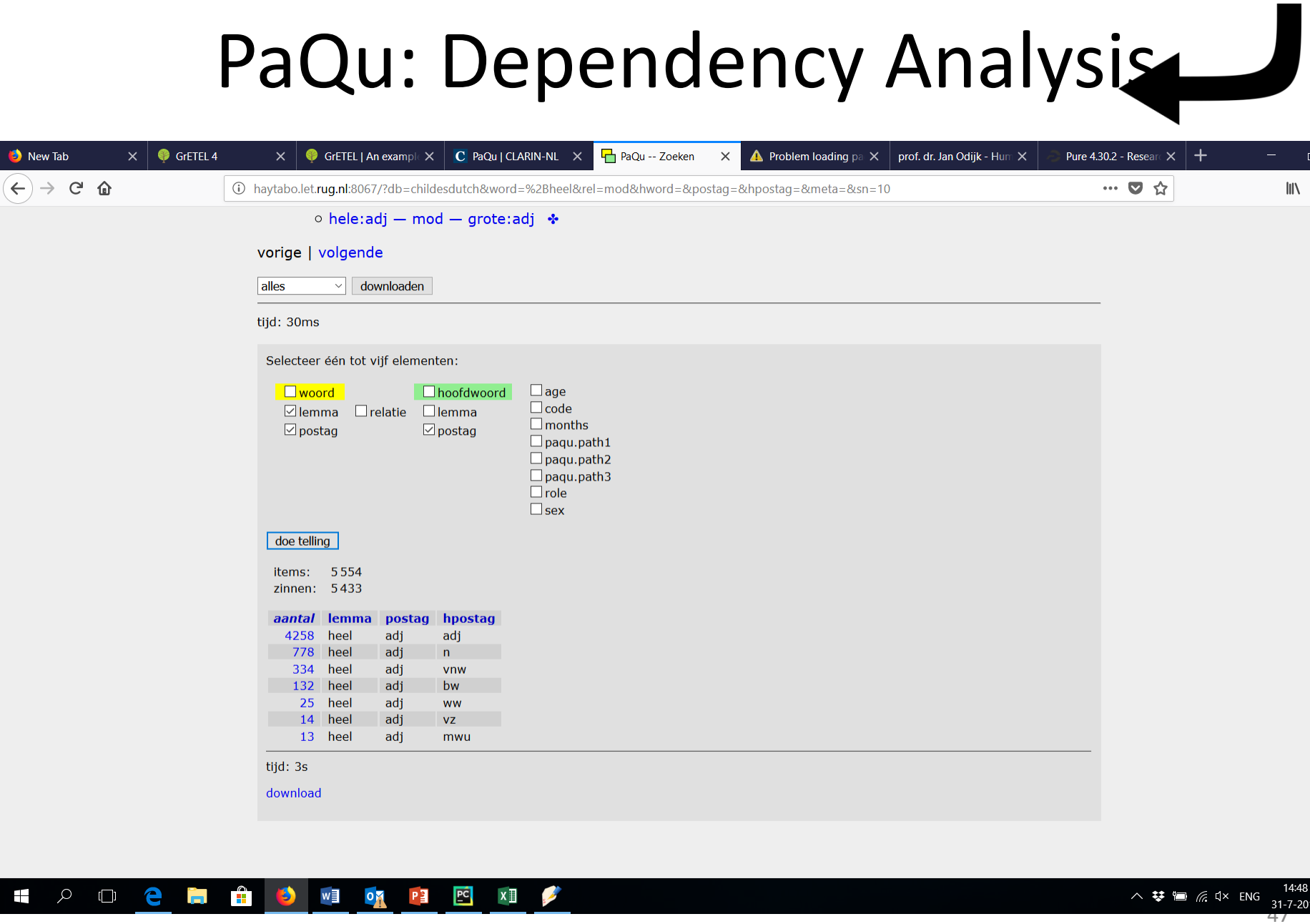


word	Morpho-Syntax	Syntax	Meaning
zeer	N		`pain’
	A	Mod N, predc	‘painful’
		Mod A V P	‘very’

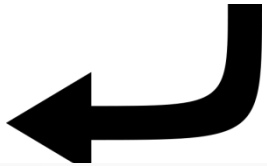
PaQu: Dependencies



PaQu: Dependency Analysis



PaQu: Upload Corpus



haytabo.let.rug.nl:8067/corpora

✣ ✗	gereed	PARSEME MWE test	35	openbaar	
✣ ✗	gereed	VanKampen-child-heelergzeer	327	openbaar	
✣ ✗	gereed	VanKampenHeel	445	openbaar	
✣ ✗	gereed	VanKampenJAC	61 538	privé	afgeleid corpus, beschermd
✣ ✗	gereed	VanKampenZeer	41	openbaar	
✣ ✗	gereed	wikizeermodww	8 169	privé	afgeleid corpus, beschermd
✣ ✗	gereed	zeerModwwBasilex	175	privé	afgeleid corpus
✣ ✗	gereed	zeerModwwWiki	8 169	privé	afgeleid corpus, beschermd

Nieuw corpus maken

Heb je een corpus in **FoLiA**-formaat, met interne of externe **metadata**, klik dan **hier**.

De tekst die je uploadt moet platte tekst zijn, zonder opmaak (geen Word of zo), gecodeerd in utf-8.
Daarnaast worden een aantal andere formaten herkend, zie uitleg.

Titel:

Upload document:

Browse... No file selected.

Soort document (uitleg):

Automatisch bepaald of ander formaat

Submit Query

LASSY-Small: *heel* as modifier

- Query
- Results

aantal	lemma	hpostag	mod A	mod V	mod P	mod N	predc	other	unclear		Total
401	heel	n	0	0	0	401	0	0	0		**401**
277	heel	adj	277	0	0	0	0	0	0		**277**
104	heel	vnw	27	0	0	77	0	0	0		**104**
19	heel	mwu	2	0	0	17	0	0	0		**19**
14	heel	ww	11	0	0	3	0	0	0		**14**
7	heel	bw	7	0	0	0	0	0	0		**7**
2	heel	tw	0	0	0	2	0	0	0		**2**
824			**324**	**0**	**0**	**500**	**0**	**0**	**0**		**824**

LASSY-Small: *erg* as modifier

- Query
- Results

aantal	lemma	hpostag	mod A	mod V	mod P	mod N	predc	other	unclear		Total
146	erg	adj	146	0	0	0	0	0	0		**146**
35	erg	ww	17	15	3		0	0	0		**35**
14	erg	n	0	0	0	14	0	0	0		**14**
7	erg	vnw	7	0	0	0	0	0	0		**7**
3	erg	mwu	0	0	3	0	0	0	0		**3**
1	erg	bw	1	0	0	0	0	0	0		**1**
206			**171**	**15**	**6**	**14**	**0**	**0**	**0**		**206**

LASSY-Small: *zeer* as modifier

- Query; Results

aantal	lemma	hpostag	mod A	mod V	mod P	mod N	predc	other	unclear	Total
274	zeer	adj	274	0	0	0	0	0	0	**274**
50	zeer	ww	42	7	1	0	0	0	0	**50**
20	zeer	vnw	20	0	0	0	0	0	0	**20**
2	zeer	mwu	0	1	1	0	0	0	0	**2**
1	zeer	bw	1	0	0	0	0	0	0	**1**
1	zeer	n	0	0	0	1	0	0	0	**1**
1	zeer	vz	0	0	1	0	0	0	0	**1**
349			**337**	**8**	**3**	**1**	**0**	**0**	**0**	**349**

CGN: *heel* as modifier

- Query; Results

aantal	lemma	hpostag	mod A	mod V	mod P	mod N	predc	other	unclear		Total
2668	heel	adj	2668	0	0	0	0	0	0		**2668**
767	heel	n	5	0	0	762	0	0	0		**767**
511	heel	vnw	451	0	0	58	0	1	1		**511**
115	heel	bw	115	0	0	0	0	0	0		**115**
61	heel	ww	52	0	2	4	0	2	1		**61**
24	heel	let	0	0	0	0	0	0	24		**24**
23	heel	spec	0	0	0	1	0	0	22		**23**
14	heel	vg	11	0	0	2	0	0	1		**14**
9	heel	mwu	3	0	2	4	0	0	0		**9**
6	heel	vz	0	0	5	0	0	1	0		**6**
3	heel	tsw	0	0	0	0	0	3	0		**3**
1	heel	lid	0	0	0	1	0	0	0		**1**
4202			**3305**	**0**	**9**	**832**	**0**	**7**	**49**		**4202**

CGN: *erg* as modifier

- Query; Results

aantal	lemma	hpostag	mod A	mod V	mod P	mod N	predc	other	unclear		Total
301	erg	adj	301	0	0	0	0	0	0		**301**
124	erg	ww	21	86	9	0	4	0	4		**124**
40	erg	vnw	38	0	0	2	0	0	0		**40**
14	erg	n	0	1	0	12	1	0	0		**14**
6	erg	vz	0	0	6	0	0	0	0		**6**
5	erg	bw	5	0	0	0	0	0	0		**5**
5	erg	let	0	0	0	0	0	0	5		**5**
5	erg	vg	4	1	0	0	0	0	0		**5**
2	erg	mwu	1	1	0	0	0	0	0		**2**
2	erg	spec	0	0	0	0	0	0	2		**2**
1	erg	tsw	0	0	0	0	0	0	1		**1**
505			**370**	**89**	**15**	**14**	**5**	**0**	**12**		**505**

CGN: *zeer* as modifier

- Query; Results:

aantal	lemma	hpostag	mod A	mod V	mod P	mod N	predc	other	unclear		Total
193	zeer	adj	193	0	0	0	0	0	0		**193**
34	zeer	ww	22	9	2	0	0	1	0		**34**
15	zeer	vnw	15	0	0	0	0	0	0		**15**
4	zeer	bw	4	0	0	0	0	0	0		**4**
3	zeer	n	1	0	0	2	0	0	0		**3**
1	zeer	vg	1	0	0	0	0	0	0		**1**
250			**236**	**9**	**2**	**2**	**0**	**1**	**0**		**250**

VanKampenJAC: *heel* as modifier

- Query; Results:

aantal	lemma	hpostag	mod A	mod V	mod P	mod N	predc	other	unclear		Total
619	heel	adj	619	0	0	0	0	0	0		**619**
130	heel	n	0	0	0	130	0	0	0		**130**
28	heel	vnw	24	0	0	4	0	0	0		**28**
20	heel	bw	19	1	0	0	0	0	0		**20**
1	heel	mwu	0	0	1	0	0	0	0		**1**
1	heel	vz	0	0	0	0	0	0	1		**1**
799			**662**	**1**	**1**	**134**	**0**	**0**	**1**		**799**

VanKampenJAC : *erg* as modifier

- Query; Results:

aantal	lemma	hpostag	mod A	mod V	mod P	mod N	predc	other	unclear		Total
29	erg	adj	29	0	0	0	0	0	0		**29**
15	erg	ww	0	12	0	0	3	0	0		**15**
6	erg	n	1	3	0	2	0	0	0		**6**
4	erg	vnw	4	0	0	0	0	0	0		**4**
											0
											0
54			**34**	**15**	**0**	**2**	**3**	**0**	**0**		**54**

VanKampenJAC : *zeer* as modifier

- Query; Results:

aantal	lemma	hpostag	mod A	mod V	mod P	mod N	predc	other	unclear		Total
53	zeer	ww	0	2	0	0	0	51	0		**53**
2	zeer	adj	2	0	0	0	0	0	0		**2**
1	zeer	n	0	0	0	1	0	0	0		**1**
56			**2**	**2**	**0**	**1**	**0**	**51**	**0**		**56**

VK/LAUorSAR : *heel* as modifier

- Query; Results:

aantal	lemma	hpostag	mod A	mod V	mod P	mod N	predc	other	unclear		Total
193	heel	adj	193	0	0	0	0	0	0		**193**
45	heel	n	12	0	0	31	0	0	2		**45**
30	heel	vnw	28	0	0	2	0	0	0		**30**
17	heel	bw	15	0	0	0	0	1	1		**17**
3	heel	mwu	2	0	0	0	0	0	1		**3**
1	heel	tw	0	0	0	0	0	0	1		**1**
1	heel	vz	0	0	1	0	0	0			**1**
1	heel	ww	0	0	0	0	0	0	1		**1**
291			**250**	**0**	**1**	**33**	**0**	**1**	**6**		**291**

VK/LAUorSAR: *erg* as modifier

- Query; Results

aantal	lemma	hpostag	mod A	mod V	mod P	mod N	predc	other	unclear		Total
11	erg	adj	11	0	0	0	0	0	0		**11**
6	erg	ww	1	1	1	0	0	0	0		**3**
1	erg	n	0	0	0	0	1	0	0		**1**
											0
18			**12**	**1**	**1**	**0**	**1**	**0**	**0**		**15**

VK/LAUorSAR : *zeer* as modifier

- Query; Results:

aantal	lemma	hpostag	mod A	mod V	mod P	mod N	predc	other	unclear		Total
4	zeer	ww	0	0	0	0	0	4	0		**4**
1	zeer	adj	1	0	0	0	0	0	0		**1**
5			**1**	**0**	**0**	**0**	**0**	**4**	**0**		**5**

CHILDES: *heel* as modifier

- Query; Results:

aantal	lemma	hpostag	mod A	mod V	mod P	mod N	predc	unclear	other	Total
4258	heel	adj	4257	0	1	0	0	0	0	**4258**
778	heel	n	22	0	1	748	0	4	3	**778**
334	heel	vnw	304	0	0	23	2	5	0	**334**
132	heel	bw	121	2	1	0	0	4	4	**132**
25	heel	ww	17	4	0	1	1	2	0	**25**
14	heel	vz	1	0	8	0	1	4	0	**14**
13	heel	mwu	5	0	5	1	0	2	0	**13**
5554			**4727**	**6**	**16**	**773**	**4**	**21**	**7**	**5554**

CHILDES: *erg* as modifier

- Query; Results:

aantal	lemma	hpostag	mod A	mod V	mod P	mod N	predc	unclear	other	Total
318	erg	adj	317	0	1	0	0	0	0	**318**
113	erg	ww	14	82	3	1	12	1	0	**113**
34	erg	n	1	0	0	30	0	3	0	**34**
21	erg	vnw	21	0	0	0	0	0	0	**21**
4	erg	bw	2	1	0	0	1	0	0	**4**
1	erg	mwu	1	0	0	0	0	0	0	**1**
1	erg	tsw	1	0	0	0	0	0	0	**1**
492			**357**	**83**	**4**	**31**	**13**	**4**	**0**	**492**

CHILDES: *zeer* as modifier

- Query; Results:

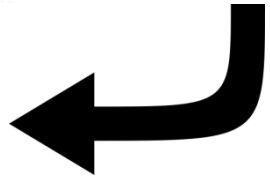
aantal	lemma	hpostag	mod A	mod V	mod P	mod N	predc	unclear	other		Total	
135	zeer	ww	1	5	3	0	0	1	125		**135**	
10	zeer	adj	10	0	0	0	0	0	0		**10**	
8	zeer	n				7			1		**8**	
1	zeer	vnw	1	0	0	0	0	0	0		**1**	
154			**12**	**5**	**3**	**7**	**0**	**1**	**126**		**154**	

Basilex: *Heel* as modifier

- Query
- Results

#	lemma	rel	hpostag
13099	heel	mod	adj
6157	heel	mod	n
1595	heel	mod	vnw
666	heel	mod	bw
347	heel	mod	ww
36	heel	mod	mwu
18	heel	mod	vz

Basilex: *Heel*: Analysis



Summary	vnw	bw	ww	mwu	vz
adj	1316	664	336	2	4
n	278		7	20	
adverbial PP				13	10
unclear		1		1	1
open slot	1		1		3
predm		1	1		

Basilex: *Erg* as modifier

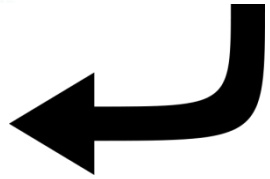
- Query
- Results

#	lemma	rel	hpostag
4091	erg	mod	adj
1312	erg	mod	ww
277	erg	mod	n
259	erg	mod	vnw
44	erg	mod	bw
21	erg	mod	mwu
7	erg	mod	vz
1	erg	mod	tw

Basilex*: Erg*: Analysis

Summary	adj	ww	n	vnw	bw	mwu	vz	tw	predc	ld
adj	4090	*326*	70	192	43				8	1
vz					1	18	7		17	4
n			180	67		2				
ww		986	10							
tsw			2							
indep			8						8	
unclear	1		7			1			5	
illformed								1		
VP										2

Basilex: *Zeer* as modifier



- Query
- Results

#	lemma	rel	hpostag
281	zeer	mod	adj
179	zeer	mod	ww
65	zeer	mod	n
6	zeer	mod	vnw
1	zeer	mod	mwu

Basilex: *Zeer*: Analysis

Summary	adj	ww	n	vnw	mwu	ld	predc
adj	281	68	3	6			3
ww		23					
n			62				
vz		3			1		
unclear		5					2
indep		80					

Wikipedia: *Heel* as modifier

- Query
- Results

#	lemma	rel	hpostag
27900	heel	mod	n
10652	heel	mod	adj
4029	heel	mod	vnw
690	heel	mod	bw
596	heel	mod	mwu
576	heel	mod	ww
41	heel	mod	tw
29	heel	mod	vz
19	heel	mod	spec

Wikipedia: *Heel*: Analysis

Summary	adj	ww	vnw	n	mwu	bw	vz	spec	tw	/ m
adj	10652	569	1199		8	690	5	0	0	90.5
ww		0	0				0	0	0	0.0
n		7	2829	27907	562		0	19	41	216.3
vz		0	0		26		16	0	0	0.3
pred		0	0				3	0	0	0.0
unclear		0	1				5	0	0	0.0
tbd		0	0		0		0	0	0	0.0

Wikipedia: *Erg* as modifier

- Query
- Results

#	lemma	rel	hpostag
15034	erg	mod	adj
4645	erg	mod	ww
826	erg	mod	vnw
792	erg	mod	n
134	erg	mod	mwu
47	erg	mod	bw
8	erg	mod	vz
4	erg	mod	spec
3	erg	mod	tw

Wikipedia*: Erg*: Analysis

Summary	adj	ww	vnw	n	mwu	bw	vz	erg + ld	erg + predc	spec	tw	/ m
adj	15034	2708	798		8	46	2	0	0	4	3	128.3
ww		1937	1		0	0	3	-48	-128	0	0	12.2
n		0	27	792	0	0		0	0	0	0	5.6
vz		0	0		125	1	1	48	128	0	0	2.1
pred		0	0		1	0	2	0	0	0	0	0.0

Wikipedia: *Zeer* as modifier

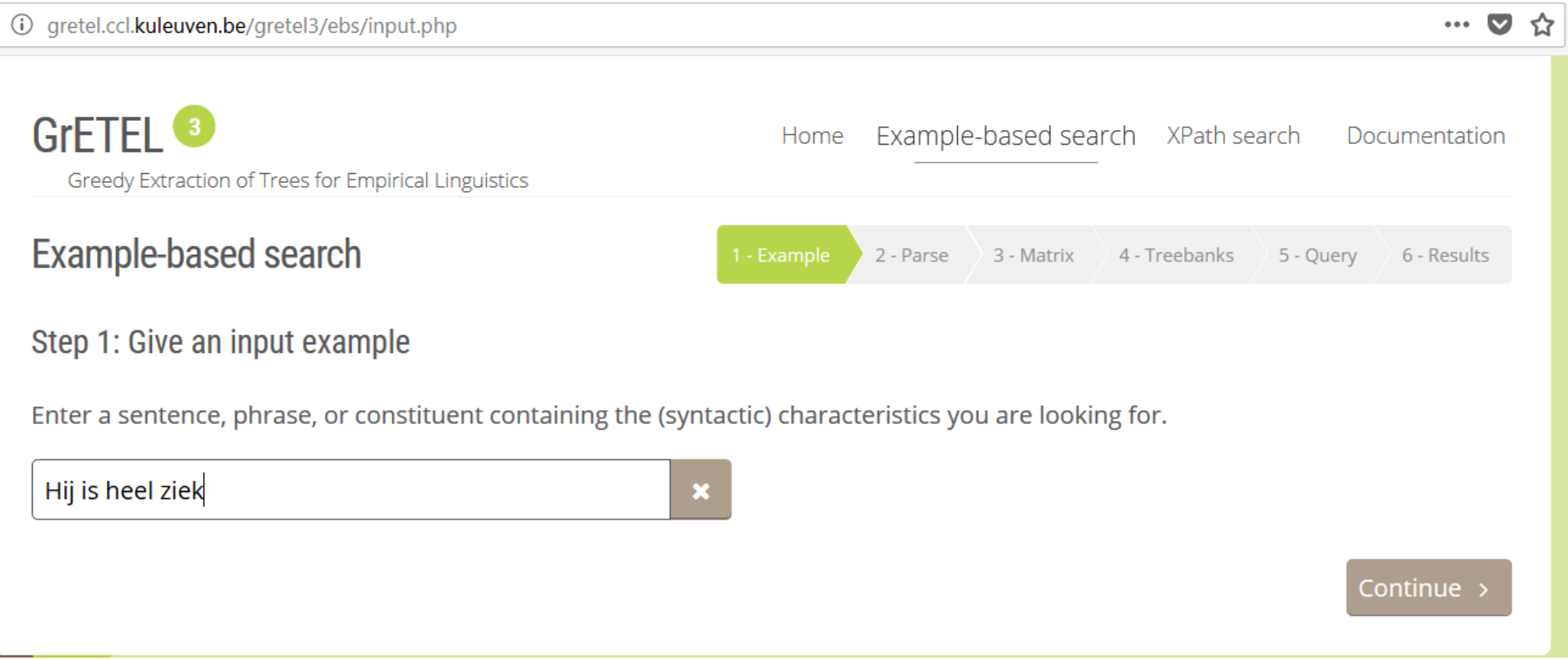
- Query
- Results

#	lemma	rel	hpostag
41664	zeer	mod	adj
8319	zeer	mod	ww
2035	zeer	mod	vnw
260	zeer	mod	bw
159	zeer	mod	n
137	zeer	mod	mwu
22	zeer	mod	vz
12	zeer	mod	spec
1	zeer	mod	tw

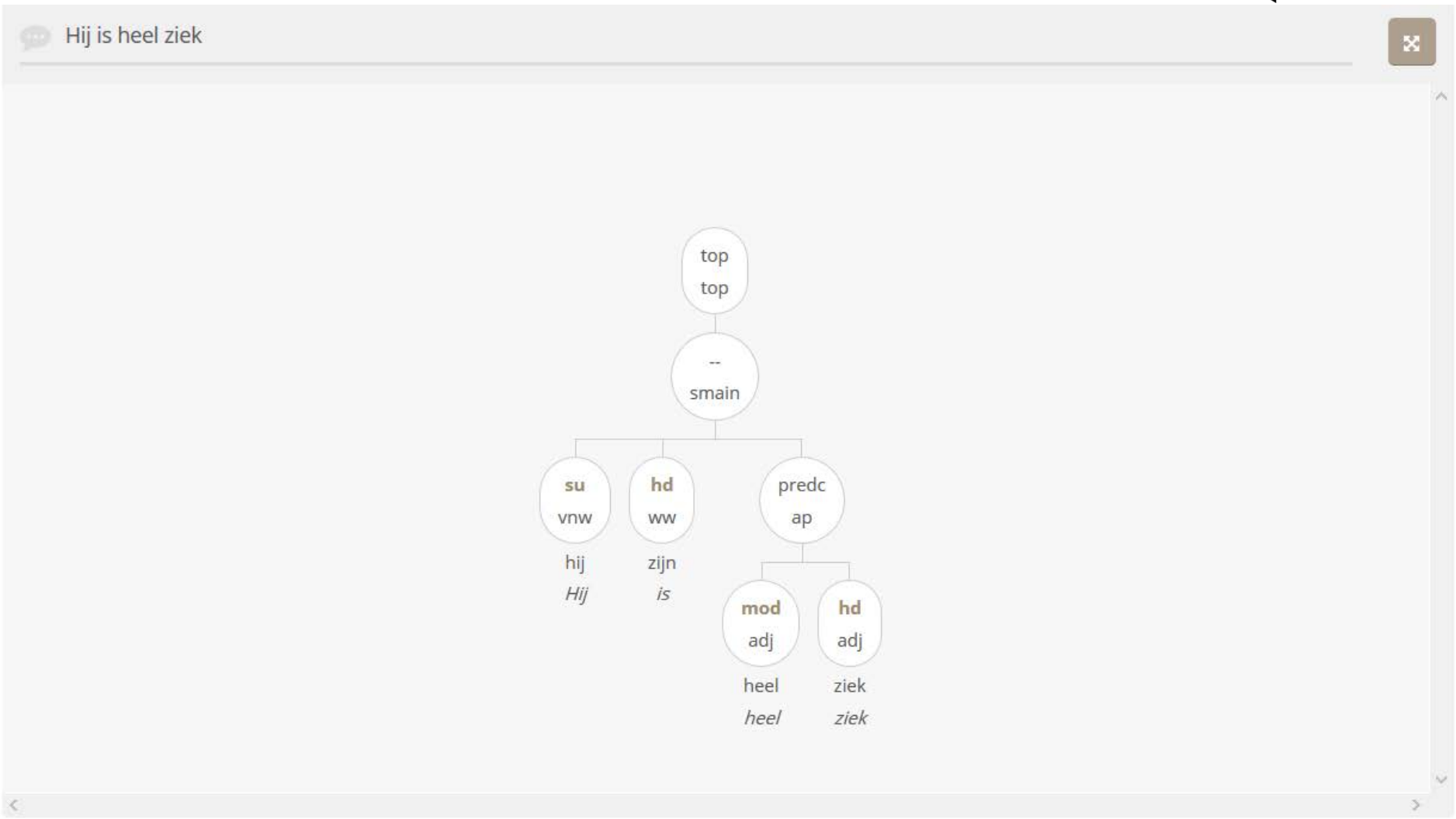
Wikipedia: *Zeer*: Analysis

Summary	adj	ww	vnw	n	mwu	bw	vz	spec	Zeer + ld	Zeer + predc	tw	/ m
adj	41663	5498	2033	89	26	260	9	12			1	342.0
ww	0	2802	1	0	0	0	1	0	-47	-102	0	18.3
n	0		0	63	0	0	0	0			0	0.4
vz	1		0	4	109	0	12	0	47	102	0	1.9
pred	0		0	0	0	0	0	0			0	0.0
unclear	0		1	3	2	0	0	0			0	0.0
zeer doen	0	19	0	0	0	0	0	0	0	0	0	0.1

GRETEL Query By Example



GRETEL Query By Example



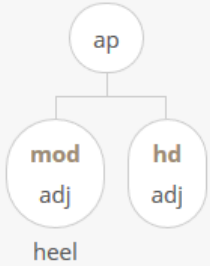
GRETEL Query By Example

Show advanced options

sentence	Hij	is	heel	ziek
word	○	○	○	○
lemma	○	○	◉	○
word class	○	○	○	◉
optional in search	◉	◉	○	○

Search options

- ☐ Respect word order
- ☐ Include context in results
- ☐ Ignore properties of the dominating node
- ☑ Include sentence IDs in downloadable results



GRETEL Query By Example

Individual results

Click on a sentence ID to view the tree structure. The part of the sentence matching your input structure is set in bold.

# of results: **266** / 266 Filter components

#	ID	Component	Sentence
1	WS-U-E-A-0000000007.p.45.s.1	WSUE	Meer hierover straks in het 10 uur journaal op Nederland 3 Einde van dit acht-uur-journaal , op dit net volgt Netwerk en ik wens u nog een **heel prettige** avond
2	WS-U-E-A-0000000228.p.6.s.4	WSUE	Toen werkten de Spanjaarden **heel nauw** samen met de Italianen .
3	WS-U-E-A-0000000245.p.7.s.3	WSUE	**Heel eigentijds** was de actie van Bram en Freek in 1978 : ze riepen op tot een boycot van het WK Voetbal in Argentinië .
4	WS-U-E-A-0000000207.p.21.s.2	WSUE	En dat hield Acebes **heel lang** vol , zelfs toen de rest van de wereld er al van overtuigd was dat Al-Qaida verantwoordelijk was .
5	WS-U-E-A-0000000040.p.53.s.1	WSUE	Hier NETWERK zo meteen en ik wens u nog een **heel prettige** avond ....
6	WS-U-E-A-0000000005.p.37.s.13	WSUE	Hier presenteren vakbroeders feestelijk hun waar aan elkaar . geluid op Op de Horecava zijn de prijzen wel **heel redelijk** .
7	WS-U-E-A-0000000021.p.6.s.3	WSUE	Zij ervaart het nieuws als ' **heel onwerkelijk** ' .
8	WS-U-E-A-0000000021.p.20.s.1	WSUE	Brabantse nachten mogen dan misschien lang zijn , **voor illegale escortbedrijven in Eindhoven** zijn ze tegenwoordig ook **heel onzeker** .
9	WS-U-E-A-0000000201.p.38.s.1	WSUE	Hier NETWERK zo meteen en ik wens u nog een **heel prettige** avond ....
10	WS-U-E-A-0000000205.p.8.s.9	WSUE	Tot slot , Amerika is nu een **heel ander** land **dan vier jaar geleden** .

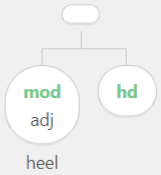
GRETEL 4 Analysis

Not secure | gretel.hum.uu.nl/gretel4/ng/xpath-search?currentStep=3&xpath=%2F%2Fnode%5B%0A%20%20%20%20node%5B@rel%3D"mod"%20and%20@pt%3D"adj"%20and%20@lemma%3D"heel"%5D%20...

XPath

```
//node[
    node[@rel="mod" and @pt="adj" and @lemma="heel"] and
    node[@rel="hd" ]]
```

Tree



Analysis Table

Table ▾ | $node1.pt ▾

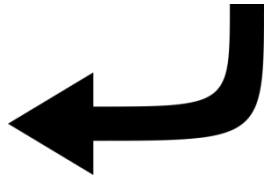
Count ▾ ↕ ↔

$node1.lemma ▾

$node2.pt ▾

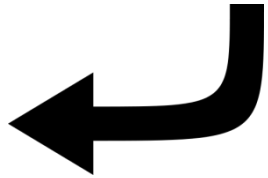
$node1.lemma	$node2.pt	Totals
heel	(none)	14
	adj	132
	bw	2
	n	206
	vnw	66
	ww	5

LASSY-Small *heel* mod V



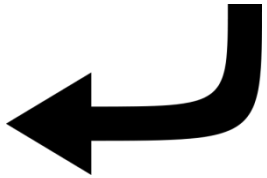
- *Heel* mod ww does occur in the query results, but
  - Most examples are participles, which all must be analysed as A, e.g.
    - *heel gecompliceerd* ‘very complicated’
    - *heel overtuigend* ‘very convincing’
    - *heel vervelend* ‘very boring’
  - Some substantivised infinitives, *heel* actually modifies N:
    - *Het hele … gebeuren* ‘the whole … happening’
    - *Hun hele hebben en houden* ‘all their possessions’
    - their whole have and hold

CGN *heel* mod V



- *Heel* mod V does occur, but
  - Ill-formed for me
  - Almost exclusively by Flemish speakers
  - *Heel* must mean *geheel* or *helemaal* ‘completely’ here
- Examples:
- … heel te verdwalen … ‘to get completely lost’
- … heel omgebouwd … ‘completely rebuilt’

CGN *heel* mod P

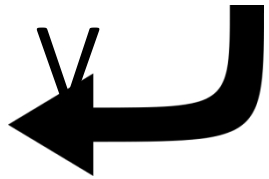


- Adverbial PPs
- One other case (ill-formed for me)
  - …'t heel voor de hand ligt …
  - … it very before the hand lies …
  - '… it is very obvious …'
  - (*voor de hand liggen**d*** is adjectival in nature and often occurs with *heel*)
  - I assume it is a performance error
  - It might also be a case of *heel* mod V
- One example where *heel* means 'completely' (not counted)
  - *heel beneden* 'completely downstairs' (Flemish speaker)

VanKampen Children *heel* Mod P

- One example:
  - *heel buiten* ‘very (completely?) outside’
  - xxx Laura is heel [?] buiten.
  - Sarah34.cha, speaker SAR, age 3;5.30
  - [?] marks that ‘heel’ is the best guess of the transcriber of what was said

VanKampen Adults *heel* Mod V



– One example:
  - ik kijk heel uit. I watch very out ‘I am very careful’
  - laura30.cha, speaker JAC (adult)

VanKampen Adults *heel* Mod P

– One example:
  - *Heel af en toe*
  - Instance of *heel* + adverbial PPs

CHILDES *heel* Mod P

Adverbial PPs	heel af en toe	5
	heel in de verte	2
Unclear Cases	heel naar buiten	1
	heel buiten	classified as predc
	heel uit	1
	hele boov	1
	heel in	1

CHI iri30211.325 3;2.11 heel op ə dobbeltje duikelen very on eh somersault do 'do a somersault' (?)	CHI tom21010.50 2;10.10 Heel in Completely in (?) (of a boat in a lock)	CHI daa21028.78 3;10.28 hele boov very above 'high above' (?)
SAR sarah34.91 P3;5.30 xxx Laura is heel buiten xxx Laura is very outside 'Laura is completely outside' (?)	BOU mat20501.429 (father) heel iets naar buiten very something to outside 'a little bit to the outside'	

CHILDES *heel* Mod V

CHI abe20409.856 2;4.9 **Heel schaatsen** **Very skate** **??**	**CHI abe20506.495 2;5.6** **Heel bedankt** **Very thanked** **'Thanks very much'**	**CHI 2;7.29 abe20729.943** **heel praten** **very talk** **??**
CHI Wijnen/30608.206 3;6.8 **hier kan heel bomen op** **here can very trees on** **'Many(?) trees can be put on top of this'**	LAU laura60.790 4;7.02 … als je het heel vertellen heb … … if you it very tell have … '…if you have told it all…' (?)	CHI abe30107.1343 3;1.7 dat is heel pas op doen that is very be careful do 'that means being very careful'

heel + Adverbial PPs

Heel af en toe very off and to 'very occasionally'	*heel in de verte* very in the distance 'very far in the distance'	*heel in het algemeen* very in the general 'very generally'
heel in het bijzonder very in the particular 'more particularly'	*heel in het kort* very in the short 'very briefly'	*heel in het begin* very in the beginning 'at the very beginning'
heel op het laatst Very at the latest 'at the very end'	*heel uit de verte* very from the distance 'from very far in the distance '	*heel aan het eind* very at the end 'at the very end'

Ill-formed heel + Adverbial PPs

**heel in de hoogte [v. verte]* very in the height 'at very great height'	**heel in de nabijheid [v. verte]* very in the closeness 'very close by'	**heel in de diepte [v. verte]* very in the depth 'at very great depth'
**heel naar de verte [v. in]* very to the distance 'towards very far in the distance' (v. *naar heel in de verte*)	*Heel aan de kop *[v. begin]* Very at the head 'at the very beginning'	*Heel aan de start *[v. begin]* *Very at the start* 'at the very beginning'
*heel aan de staart [v. *eind*] Very at the tail 'at the very end'	*heel in het lang [v. *kort*] very in the long 'very extensively'	

VanKampen Children *erg* Mod P

- VanKampen/Laura50 LAU 3;11.16
  - Gaat ie heel erg over heen
  - Goes he very very over PRT
  - ‘He goes very much over that’
  - (unclear from the context what could be intended)

SoNaR *heel* Mod P

heel buiten zijn verwachting

HEEL outside his expectation

'completely unexpectedly' (?)

heel in het zwart

HEEL in the black

'fully in black' (?)

heel in orde

HEEL in order

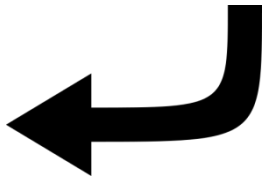
'fully OK' (?)

- Most probably meaning ‘completely’
- Maybe some ‘very’
- All speakers Flemish

Mod P: *heel* v. *zeer*

Corpus / m tokens	*heel* mod P	*zeer* mod P
LASSY-Small	0,0	2,7
CGN	7,9	1,8
VanKampenJAC	3,3	0,0
VanKampen LAUorSAR	6,5	0,0
CHILDES Dutch	5,8	1,6
Basilex	1,7	0,3
Wikipedia	0,3	1,9

Spontaneous Speech (all from adults in CHILDES)



- Hesitations, filled pauses (mat30603)
  - *en ehm (.) gaan we nog ehm (.) +...*
  - And hmm go we still hmm
- Repetitions (laura01)
  - *een molen [/] molen.*
  - A mill mill
- False starts and retracing (tom01302)
  - *<geef jij> [//] kom jij op mijn verjaardag ?*
  - Give you come you on my birthday ?
- Unfinished utterances (see filled pauses)

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), …