

University of Groningen

Topics in Corpus-Based Dutch Syntax

Beek, Leonoor Johanneke van der

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version

Publisher's PDF, also known as Version of record

Publication date:

2005

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Beek, L. J. V. D. (2005). *Topics in Corpus-Based Dutch Syntax*. s.n.

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

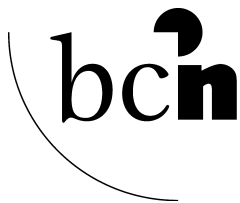
Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Topics in Corpus-Based Dutch Syntax

Leonoor Johanneke van der Beek



The work in this thesis has been carried out under the auspices of the Behavioral and Cognitive Neurosciences (BCN) research school, Groningen, and has been part of the Pionier project *Algorithms for Linguistic Processing* supported by grant number 220-70-001 from the Netherlands Organization for Scientific Research (NWO).



Groningen Dissertations in Linguistics 54

ISSN 0928-0030

Document prepared with L^AT_EX 2_ε

Printed by PrintPartners Ipskamp.

Rijksuniversiteit Groningen

Topics in Corpus-Based Dutch Syntax

Proefschrift

ter verkrijging van het doctoraat in de
Letteren
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. F. Zwarts,
in het openbaar te verdedigen op
donderdag 10 november 2005
om 13.15 uur

door

Leonoor Johanneke van der Beek

geboren op 5 februari 1978
te Beuningen

Promotor: Prof. dr. ir. J. Nerbonne

Copromotores: Dr. G.J.M. van Noord
Dr. G. Bouma

Beoordelingscommissie: Prof. dr. J. Bresnan
Prof. dr. F. Van Eynde
Prof. dr. J. Hoeksema

Preface

I am indebted to many people for being able to finish this thesis. First of all my supervisor Gertjan van Noord. This dissertation has benefitted greatly from his comments on drafts and from the discussions we had, especially in the last months of the project. Moreover, I would like to thank him for allowing me to follow my own interests and make my own mistakes at all times.

I owe debt also to my promotor John Nerbonne and co-promotor Gosse Bouma for their valuable comments on my writings, and John also for his organizational and financial support as the head of the department. Joan Bresnan, Frank Van Eynde and Jack Hoeksema kindly agreed to be on my reading committee, for which I would like to thank them. This thesis benefitted particularly from the comments of Frank Van Eynde, who corrected mistakes, suggested improvements and offered additional data. I tried to incorporate these comments in this final version as well as possible, but obviously I remain responsible for all remaining flaws.

The seeds for almost all of this work were planted during my stay at Stanford University. I owe thanks to the CSLI people, Timothy Baldwin, John Beavers, Dan Flickinger, Stephan Oepen and Ivan Sag, for welcoming me in the LinGO Project and providing me with a stimulating work environment. I'm especially grateful to Timothy Baldwin. I learned a lot from our collaboration, the results of which form part of this thesis. Another highlight of my stay was Joan Bresnan's introduction to LFG. I certainly owe her my gratitude for discussion and support. I am thankful also to Rob Malouf for sending that one email that got me in Stanford in the first place.

Returning home was made easy by the many people who contributed to work and life in the Alfa-Informatica department. Wyke van der Meer, Anna Hausdorf and the secretaries ensured things ran smoothly in the department. Tanja Gaustad was always there to share everyday ups and downs with and Begoña Villada often initiated professional discussions or after-hour activities. Gerlof Bouma was always easily persuaded to share his linguistic intuitions or a coffee with me, and I thoroughly enjoyed our collaboration.

Stasinos Konstantopoulos made life easier by putting together the RuG thesis stylefile, and more agreeable by dragging me to *het Paard* every now and then. Eleonora Rossi helped me out with various practical issues, gave moral support when needed and makes a great “body of language” on the cover. Besides work, there were *aio*-dinners, movies and Gaioo/Grasp! meetings, pancake lunches, potluck dinners and (Friday) afternoon drinks—a big thank you to all who made life in Groningen so much fun.

Stasinos, Tanja and Begoña should be thanked once more for giving me the opportunity to practice the defense ceremony as a *paranimf*. I feel fortunate and proud that Irene Jansen and Eleonora Rossi agreed to stand by my side when it’s my turn, at last.

I would have never started this project if it weren’t for Víctor Sánchez-Valencia. His faith in me convinced me that I could do it and through our discussions on various linguistic topics I learned that I might very well love it. I wish I could still discuss linguistics, politics and life with you—or say thank you.

Finally, many thanks to my parents and my sisters. For always supporting me, for being there. *Dank je mam, dank je pap.*

Contents

1	Introduction	1
1.1	Corpus Linguistics	1
1.1.1	Data collection	1
1.1.2	Gradient patterns and probability	4
1.1.3	Resources for natural language processing	6
1.1.4	Evaluation	7
1.1.5	Corpus linguistics and this thesis	7
1.2	Methodological Preliminaries	9
1.2.1	Corpora	9
1.2.2	Tools	10
1.2.3	Statistics	11
1.3	Theoretical Framework	13
1.3.1	Lexical Functional Grammar	14
1.3.2	Optimality Theory	23
1.3.3	Stochastic OT	25
1.3.4	OT and LFG	26
1.4	Overview	29
2	Clefts	31
2.1	Introduction	31
2.2	Transitive Clefts	33
2.2.1	Differences between cleft clauses and other relative clauses	33
2.2.2	The c-focus	36
2.2.3	Agreement	37
2.2.4	The relative clause	44
2.2.5	Formalization	47
2.3	Intransitive Clefts	52
2.3.1	Differences between transitive and intransitive clefts . .	52
2.3.2	Differences between intransitive clefts and other com- plementizer constructions	56
2.3.3	The intransitive analysis	59

2.3.4	Formalization	62
2.4	Conclusion	64
3	Dative Alternations	67
3.1	Introduction	67
3.2	Previous Work	69
3.2.1	Linearization Constraints	70
3.2.2	The NP/PP alternation in English	73
3.3	Preliminaries	74
3.3.1	Resources and methodology	74
3.4	Linearization: the Double Object Construction	76
3.4.1	Pronominality	77
3.4.2	Gradient patterns	83
3.4.3	Weight	84
3.5	Linearization: Dative PP Shift	85
3.5.1	Weight	86
3.5.2	Pronominality and definiteness	88
3.6	The NP/PP Alternation	90
3.6.1	Lexical preferences	90
3.6.2	Weight and pronominality	92
3.6.3	Implementation in OT	94
3.7	More Factors in the Dative Construction?	99
3.8	Additional Evidence: the AcI	101
3.9	Conclusion	106
4	Determinerless PPs	109
4.1	Introduction	109
4.2	The Syntax of Determinerless PPs	110
4.2.1	Fixed determinerless PPs	111
4.2.2	Independent bare noun NPs	111
4.2.3	Compositional determinerless PPs	114
4.2.4	Prepositions selecting for determinerless NPs	117
4.2.5	Determinerless PPs as dependents	120
4.3	Extraction of PP-Ds	121
4.3.1	Introduction	121
4.3.2	Preliminaries	123
4.3.3	Prepositions selecting for determinerless NPs	125
4.3.4	Fixed determinerless PPs	126
4.3.5	Compositional determinerless PPs	128
4.3.6	Dependent determinerless PPs	131
4.4	Evaluation and Distribution of PP-Ds	132

4.5	Conclusion and Discussion	135
5	Countability	137
5.1	Introduction	137
5.2	Preliminaries	141
5.2.1	Countability classes	141
5.2.2	Lexical resources	144
5.2.3	Past research	145
5.3	Corpus-based Classification	147
5.3.1	Feature space	148
5.3.2	Methodology	151
5.3.3	Monolingual classifiers: design	152
5.3.4	Monolingual classifiers: results and discussion	155
5.3.5	Crosslingual classifiers	157
5.3.6	Crosslingual classification: results and discussion	159
5.3.7	Binary vs. three-way classifiers	162
5.3.8	Corpus-based approach: conclusion	162
5.4	Ontology-based Classification	163
5.4.1	Lexical resources for WordNet-based classification	165
5.4.2	Classifier design	166
5.4.3	Results and discussion	169
5.4.4	Ontology-based classification: conclusion	175
5.5	Conclusion	176
6	Conclusions and Future Work	179
6.1	Conclusions	179
6.2	Future work	182
	Appendix	185
	Bibliography	189
	Samenvatting	203
	Groningen Dissertations in Linguistics	209

