

Finding focus

a study of the historical development of focus
in English

Published by
LOT
Trans 10
3512 JK Utrecht
The Netherlands

phone: +31 30 253 6006

e-mail: lot@uu.nl
<http://www.lotschool.nl>

Cover illustration: Sharon Komen
ISBN: 978-94-6093-112-3
NUR 616

Copyright © 2013 by Erwin Ronald Komen. All rights reserved.

Finding focus

a study of the historical development of focus
in English

Doctoral Thesis

to obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the Rector Magnificus prof. dr. S.C.J.J. Kortmann,
according to the decision of the Council of Deans
to be defended in public on Tuesday, June 25, 2013
at 10.30 hours
by

Erwin Ronald Komen

Born on September 8, 1960
in Utrecht, Netherlands

Promotor: Prof. dr. Ans van Kemenade

Co-promotor: Prof. dr. Bettelou Los (University of Edinburgh)

Doctoral Thesis Committee:

Prof. dr. Antal van den Bosch

Prof. dr. Helen de Hoop

Prof. dr. Johanna Nichols (University of California, Berkeley)

Prof. dr. Ann Taylor (University of York)

Prof. dr. Suzanne Winkler (University of Tübingen)



Acknowledgements

I would like to acknowledge those who have made it possible for me to do the research described in this dissertation. Top on the list is God, whom I thank for his 24/7 support, for health and for moments of inspiration, which I happened to have especially in the mornings.

Bettelou Los has supported and encouraged me from the very beginning. She has always been there with her encyclopaedic knowledge of articles and with her insight into the English language. It was she who coined the term “Inert” for one of the five referential categories in the Pentaset. Professor Ans van Kemenade has not only provided expert supervision, but also demonstrated her creativity with words by introducing a new Dutch verb, “zeezakken”, which derives from the computer program “Cesac” (the predecessor of “Cesax”). I am grateful to both Bettelou and Ans for introducing me to the intriguing world of English historical linguistics.

I shared an office with Gea Dreschler, Rosanne Hebing, Sanne van Vuuren and Meta Links, who have made working at Radboud University a pleasure, and I have enjoyed many a lunch talk with Astrid Bracke, Griet Coupé and Nynke de Haas. My other colleagues at the English language department have always made me feel part of the group. I have enjoyed getting to know Janine Berns, with whom I have been organising a series of inter-departmental talks.

Gea, Rosanne, Bettelou, as well as Monique Tangelder and Lieke Verheijen were involved in annotating some of the texts used in this research. Sándor Chardonnens helped me understand several Old English passages. Carlos Gussenhoven introduced me to tone and intonation and then helped me analyze the grammar of Chechen intonation. Eva D’hondt gave helpful suggestions concerning the computational linguistics side of my research. Helen de Hoop, Robert Van Valin, Frans van der Slik, Petra Hendriks, Ann Taylor, and Alice Harris have contributed through their teaching. I have consulted with Pieter Muysken on several occasions, as well as with Frans van der Slik, who I first met when we got stuck in the elevator.

I am grateful for the feedback I have received on the various articles I have written over the years. Much of the feedback was given anonymously, but I have enjoyed closer interaction with fellow-linguists at several workshops: Kristin Bech, Marco Corniglio, Hanne Eckhoff, Kristine Eide, Rik van Gijn, Jeremy Hammond, Dag Haug, Vadim Kimmelman, Dejan Matić, Svetlana Petrova, Susan Pintzuk, Saskia van Putten, Eva Schlachter, Gerold Schneider, Mark de Vries, Eirik Welo and Hedde Zeijlstra.

Many thanks to Antal van den Bosch, Helen de Hoop, Johanna Nichols, Ann Taylor and Suzanne Winkler for agreeing to be on the manuscript committee, for taking the time to read this book and for the feedback I have received so far.

Several colleagues from SIL-International have indirectly contributed to my research. Mick Foster laid the foundations for my interest in corpus linguistics. John Clifton was one of the first to help me in my academic writing efforts. Linda Humnick, one of my SIL supervisors, read several of my papers, helped me with her enthusiasm and shared her deep insight. Stephen Levinsohn demonstrated his practical approach to information structure and discourse analysis, and much of what I do in chapter 4 is directly inspired by his excellent and helpful scholarly work.

My children Irina, Benjamin, Sharon and Ariel listened to my ideas, and even though I did not always manage to get them across successfully, I appreciate their input. Sharon was my partner in entering the world of the mental models, and I've made good use of the High school graduation project on the brain which Irina did a few years ago.

My wife Liesbeth has always been there with support and enthusiasm, and my appreciation of her cannot be captured adequately enough in words.

I gratefully acknowledge the Netherlands Organisation for Scientific Research (NWO) for funding the four-year research project I have been part of (project no. 360-70-370).



Abstract

A vital skill for anyone who wants to communicate in written form is the manipulation of word order to convey emphasis in such a way, that a reader understands what is focused. Word order is influenced, and sometimes dictated, by syntax, but also by the desire to start off with what is known, and introduce new matters only in relation to that background. This dissertation addresses the question how syntax relates to information structure in general by investigating the development of constituent focus and presentational focus in English against the background of its changing syntax, while part of the analysis is substantiated by data from present-day Chechen.

The introduction (chapter 1) introduces the notion of “syntax” that I use as the expression of grammatical functions and relations as well as the definition of default word order. When the linguistic realisation of syntax is not possible through morphology, a language resorts to using word order. Related to this view of syntax is the hypothesis that changes in English syntax correlate with changes in focus: where the syntax of English increasingly requires word order, focus needs other ways to express itself.

Against the background of the latest developments in psycholinguistics, which are related to the way in which humans process a text they read or a narrative they hear (chapter 2), chapter 3 offers a working definition of focus. The three different focus articulations (which differ due to the size of the domain in which the focus occurs) can co-occur with points of departure, and word order in general is also influenced by the “Principle of Natural Information Flow”. Chapter 4 introduces my working hypothesis about the relation between syntax, pragmatics and text-organization: any linguistic realization (including word order) can be seen as a combination of (at least) these three factors. The chapter continues by touching upon a change in English syntax that seems to have been the main trigger for the changes in focus that come up later in this study. The syntactic change is the loss of the V2 system, which ultimately led to a reduction of three subject positions to just one. I attempt to tease apart word order variation caused by syntactic and text-structural factors from variation that relates to focus. The chapter contains a detailed analysis of two texts (one from Old English and one from late Modern English), and finishes with initial observations as to the changes that took place in the expression of presentational and constituent focus. What chapter 4 also finishes with is the clear realization that detailed analyses of individual texts does not give us the generalizations in tendencies we are looking for, since they involve too little data (which means that we either miss phenomena, or have too little examples of phenomena to gain enough significance). This is why a corpus approach is called for, and the planning and execution of this approach spans chapters 5-9.

The corpus approach makes use of syntactically parsed texts, and enriches them with a small set referential state primitives derived in chapter 5 using the semi-automatic program Cesax described in chapter 6. The result of this enrichment process is a set of texts in *xml* format, and chapter 7 describes how the texts can effectively be searched for combinations of syntactic and referential information.

The main idea behind the corpus approach is that it should be possible, given the syntactic and referential information, to determine the focus domains, and, consequently, see if a clause contains presentational or constituent focus. This strategy is used in chapter 8 to detect presentational focus, and we learn that Old English expressed this kind of focus by putting the syntactic subject after the verb (in the post-core slot), but this approach became increasingly infelicitous, and it was taken over by the expletive *there* approach.

Chapter 9 tests a variety of potential diagnostics for constituent focus, and those that pass the test are then used to see how the expression of this focus articulation has changed over time. It appears that one important OE strategy, the use of the clause-initial slot as one that could host contrastively focused elements, was jeopardized by the same syntactic change that also led to a gradual but almost complete loss in subject-auxiliary inversion. Constituent focus is often accompanied by overt contrast in the form of an emphatic adverb (such as “only”) or local contrast (“not X but Y”), and by watching the placement of these diagnostics we saw that the *it*-cleft gradually took over as the method to express constituent focus. This led to the question whether the *it*-cleft is a syntactic focusing device par excellence.

The answer to the *it*-cleft question spans chapters 10-12, and starts with a thorough definition of what can and what cannot be recognized as an *it*-cleft construction. A subsequent review of the function of *it*-clefts as reported in the literature brought to light several recent synchronic studies on Scandinavian languages, which claim that the major function of the *it*-cleft lies in text-organization. The synchronic study described in chapter 11 shows that Chechen, a language that does not use prosody to signal focus, uses word order and *wh*-clefts instead. Most striking is the fact that the language has an *it*-cleft construction, but only uses it for text organization. The diachronic study of English in chapter 12 underscores the evolving picture. The first *it*-clefts in English were used almost exclusively in text organization strategies, while their function as a constituent focus device only evolved once the privileged clause-initial position for emphasis was disappearing.

Chapter 13 looks back at the results found in the area of focus, arguing that focus cannot be part of syntax, nor can syntax be part of focus, and that syntax depends on referentiality. The most intriguing claim, perhaps, that this study results in is the hypothesis that focus is “compositional” in nature: if having syntactic and referential information about the constituents in a clause is sufficient to determine the focus domain, and, consequently, the focus articulation of that clause, then the higher order notion of focus can be seen as built up by syntax and referentiality. This claim is only partly validated by the research described in this book, and further work should seek to look into the hypothesis in more detail.



Content

Acknowledgements	i
Abstract	iii
List of Tables	xv
List of Figures	xvii
Abbreviations	xix
Part I – Introduction	1
1 Introduction	3
1.1 Syntax	4
1.2 English word order changes	7
1.2.1 Verb-second	7
1.2.2 Subject-finite-verb inversion	8
1.2.2.1 The plan	9
1.2.2.2 The search	10
1.2.2.3 The outcome	11
1.2.3 Expressing emphasis	13
1.3 Aim of this study	13
1.4 Methodological issues	14
1.4.1 The corpus as a “corpulect”	14
1.4.2 Translated texts	15
1.4.3 Significance of corpus findings	17
1.5 Corpora used	18
1.6 Outline of the study	19
Part II – Theory	23
2 A discourse processing model	25
2.1 Thinking about the mind	25
2.2 Insights from psycholinguistics	26
2.3 A mental model for discourse processing	28
2.3.1 Mental entity	28
2.3.2 Mental model	29
2.4 Conclusions	31
3 Focus types	33
3.1 Defining focus	33
3.2 Focus articulations	35
3.2.1 Topic-comment	36
3.2.2 Constituent focus	37
3.2.2.1 Equative clauses	37
3.2.2.2 Explicit contrastive focus	39

3.2.2.3	Emphatic prominence	40
3.2.3	Thetic sentences	41
3.2.4	Focus domain generalizations	42
3.3	Interactions with focus articulations	43
3.3.1	The principle of natural information flow	43
3.3.2	Point of departure	44
3.3.3	Dominant focal element	46
3.4	Marked versus unmarked focus	48
3.5	Focus and newness	50
3.6	Discussion	51
4	Narrative text word orders	55
4.1	A model for word order variations	57
4.1.1	Text-structure and word order	58
4.1.2	Modelling word orders: the slot-structure model	59
4.2	Old English syntax and focus	63
4.2.1	Syntactic triggers of V2	63
4.2.2	Pa-initial as V2 trigger	64
4.2.3	Pragmatic triggers of V2/V3	65
4.2.4	Adverbs as topic-domain dividers	68
4.2.5	Late subjects	69
4.3	Syntactic changes	69
4.4	Changes in the expression of focus	71
4.5	The text-charting approach	72
4.5.1	Choosing texts to chart	72
4.5.2	Text-charting as a technique	73
4.5.3	Automatically charted texts	75
4.6	Old English narrative	77
4.6.1	Narrative text	78
4.6.2	Word orders motivated by syntax or text organization	82
4.6.3	Syntactic variation	85
4.6.3.1	Default	85
4.6.3.2	Subordinate clauses	87
4.6.3.3	V-initial	87
4.6.4	Discourse variation	88
4.6.4.1	Referential point of departure	88
4.6.4.2	T-initial	90
4.6.4.3	AP-initial	92
4.6.4.4	T-correlated	94
4.6.4.5	AP-correlated	95
4.6.4.6	Logical	96
4.6.4.7	Conjunct	97
4.6.5	Focus in Old English	99
4.6.5.1	Split constituents	101
4.6.5.2	Apposition and focus	102
4.6.5.3	Unestablished information as DFE	104
4.6.5.4	Established information as DFE	106
4.6.5.5	Adverbial DFEs	106
4.6.5.6	Preposing	107
4.6.5.7	The it-cleft	108

4.7	Late Modern English narrative	109
4.7.1	Narrative text	109
4.7.2	Pragmatically neutral word orders	111
4.7.3	Syntactic variation in Modern English	113
4.7.3.1	Default word order and complementation	113
4.7.3.2	V-initial	114
4.7.4	Discourse variation in Modern English	115
4.7.4.1	Referential point of departure	115
4.7.4.2	AP-initial	117
4.7.4.3	Logical	118
4.7.4.4	Conjunct	118
4.7.5	Focus in Modern English	119
4.7.5.1	Expletive constructions	119
4.7.5.2	T-initial	121
4.7.5.3	Apposition and focus	122
4.7.5.4	Preposing	124
4.7.5.5	Established information as DFE	125
4.7.5.6	The it-cleft	126
4.8	Discussion	126
5	Referential state primitives	133
5.1	Criteria for referential state primitives	133
5.2	Existing taxonomies	135
5.2.1	A taxonomy of given and new	135
5.2.2	The topic acceptability scale	137
5.2.3	The givenness hierarchy	138
5.2.4	Coreference resolution	141
5.2.5	The PROIEL tagset	142
5.3	The Pentaset as referential state primitives	143
5.3.1	Identity	144
5.3.2	Inferred	145
5.3.3	Assumed	147
5.3.4	Inert	148
5.3.5	New	150
5.4	Is the Pentaset sufficient?	151
5.4.1	Pentaset categories versus alternatives	151
5.4.2	Deriving other categories from the Pentaset	152
5.4.3	Generics	154
5.4.4	Referential islands	155
5.4.5	Conclusions	160
5.5	Deriving topic and focus	160
5.5.1	Topic guessing	161
5.5.2	Centering theory	162
5.5.3	Deriving focus domains	164
5.5.3.1	Copula clauses in general	164
5.5.3.2	Copula clauses and variable creating expressions	167
5.6	Discussion	171

Part III - Methodology	175
6 Corpus development	177
6.1 How to add referential state primitives	177
6.2 The data and the task	179
6.3 The coreference resolution algorithm	183
6.3.1 Gathering NP features	184
6.3.2 Divide the text into sections	185
6.3.3 Identify discourse new noun phrases in the current section	186
6.3.4 Process the clauses of each sentence in chunk order	186
6.3.5 Collect the source NPs	187
6.3.6 Perform local coreference resolution	187
6.3.7 Determine the order of treating source NPs	188
6.3.8 Get the best antecedent for each source NP	188
6.3.9 Check for suspicious coreference solutions	192
6.3.10 Move the NP from the source to the antecedent collection	194
6.4 Case study: a history book from 1866	194
6.5 Discussion	196
7 Querying corpora	199
7.1 Choosing a text format and a query language	200
7.2 Accessing constituents' antecedents	203
7.3 CorpusStudio: a wrapper around Xquery	204
7.3.1 Antecedents and coreferential chains	205
7.3.2 Preceding and following sentences	207
7.3.3 Matching strings	208
7.3.4 Returning output	209
7.3.5 Returning numbers	210
7.4 Querying coreferenced corpora	212
7.5 Discussion	218
Part IV – Results	221
8 Presentational focus	223
8.1 Newness and presentational focus	224
8.2 Looking for presentational focus	226
8.3 Subject positions	228
8.4 Presentational focus with “New” subjects	229
8.4.1 Subject chain length differences	230
8.4.2 Subject position differences	231
8.4.3 Two strategies for postverbal new subjects	233
8.4.4 The other postverbal subjects	236
8.4.5 Preverbal new subjects	237
8.4.6 Constituent focus versus presentational focus	239
8.5 Presentational focus with unanchored “New” subjects	241
8.6 Presentational focus with reintroduced subjects	242
8.7 Discussion	244
9 Constituent focus in diachronic English	249
9.1 Looking for constituent focus in the main clause	252

9.2	Adverbs as diagnostics for constituent focus	253
9.2.1	Adverbs for focus and emphasis	253
9.2.2	Determining the position of constituents with a focus adverb	254
9.2.3	Results for the position of constituents with a focus adverb	255
9.3	Negation as diagnostic for constituent focus	258
9.4	Positive negation as diagnostic for constituent focus	260
9.5	Local contrast as diagnostics for constituent focus	261
9.5.1	Finding local contrast	262
9.5.2	An experiment with local contrast	262
9.6	Emphatic pronouns as diagnostics for constituent focus	264
9.7	Apposition as diagnostics for constituent focus	265
9.8	Split constituents as diagnostics for constituent focus	266
9.9	Contrastive left dislocation	268
9.9.1	Finding CLD resumptives	269
9.9.2	An experiment with CLD resumptives	270
9.10	Constituent answers as diagnostics for constituent focus	271
9.11	Clefts as diagnostics for constituent focus	272
9.11.1	The information status of free relatives	273
9.11.2	Constituent focus and <i>wh</i> -clefts	274
9.11.3	Constituent focus and reversed <i>wh</i> -clefts	274
9.11.4	The development of <i>wh</i> -clefts	275
9.12	Discussion	277
10	Cleft constructions	281
10.1	Defining clefts	281
10.1.1	Cleft definitions	282
10.1.2	The status of adjunct <i>it</i> -clefts	284
10.1.3	Specification and predication	286
10.1.4	Complements versus clefts	287
10.1.5	Referential status of the pronoun	289
10.1.6	Towards a definition	291
10.1.7	Cleft diagnostics	293
10.1.8	Testing the cleft diagnostics	295
10.2	The function of clefts	296
10.2.1	Obligatory clefts	296
10.2.2	Clefts for focus	298
10.2.3	Clefts as an avoidance strategy	300
10.2.4	Clefts to introduce presupposition	302
10.2.5	Clefts as a discourse strategy	302
10.2.6	Conclusions	306
11	Clefts in present-day Chechen	309
11.1	Focus in Chechen	309
11.2	Chechen intonation	312
11.2.1	Intonational phrases	313
11.2.2	Accentual phrases	314
11.2.3	Lexical tone	316
11.2.4	Intonation and focus	320
11.3	Chechen <i>it</i> -clefts	323
11.3.1	The Chechen <i>it</i> -cleft construction	323

11.3.2	Looking for Chechen <i>it</i> -clefts	325
11.3.2.1	A corpus of Chechen texts	325
11.3.2.2	Defining queries for Chechen <i>it</i> -clefts	326
11.3.2.3	Transforming query results into a database of Chechen clefts	327
11.3.2.4	Working with the Chechen cleft database	329
11.3.3	Discussion of the corpus findings	331
11.3.4	The function of Chechen <i>it</i> -clefts	332
11.4	Conclusions and implications	337
12	Clefts in diachronic English	341
12.1	Research on the history of clefts in English	341
12.2	Making a historical cleft database	344
12.2.1	Requirements for a cleft database	344
12.2.1.1	CleftType	345
12.2.1.2	CleftedCat	346
12.2.1.3	CleftedType	347
12.2.1.4	CleftedCoref	347
12.2.1.5	ClauseStatus	348
12.2.1.6	FocusType	348
12.2.2	Gathering initial data for the cleft database	349
12.2.3	Identifying additional candidates for the <i>it</i> -cleft database	353
12.2.3.1	Locating additional candidates for <i>it</i> -clefts	353
12.2.3.2	Clefts tagged as complement clauses	354
12.2.3.3	Clefts tagged as relative clauses	356
12.3	Results from the historical cleft database	357
12.3.1	The number of <i>it</i> -clefts in English time periods	357
12.3.2	Syntactic features	358
12.3.2.1	Category of the clefted constituent	358
12.3.2.2	Position of the clefted constituent	360
12.3.3	Information status	361
12.3.4	Information structure status	364
12.3.5	Emphatic cleft types	367
12.4	Clefts and emphasis	369
12.5	Conclusions	373
Part V	- Implications	377
13	Theoretical implications and conclusions	379
13.1	Background	379
13.2	Methodology	380
13.3	Focus changes	381
13.3.1	Presentational focus	381
13.3.2	Constituent focus	383
13.4	Implications for grammar	384
13.4.1	Syntax and referential information conspire for word order	385
13.4.2	Multi-phrasal prefields	386
13.4.3	Syntax may depend on referentiality	388
13.4.4	Mappings between syntax and focus	388
13.4.5	Grammar may have avoidance strategies	389
13.5	Focus is compositional	391
13.6	Future work	391

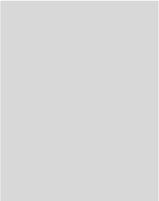
Bibliography	397
Samenvatting (Summary in Dutch)	413
De theorie	413
Methode 1: Met de hand	414
Methode 2: Automatisch	415
Wat heeft het opgeleverd?	418
Part VI - Appendices	421
14 Appendix	423
14.1 Working with CorpusStudio	423
14.1.1 Corpus research projects	423
14.1.2 Defining queries	424
14.1.3 Combining queries	425
14.1.4 Research project results	427
14.2 A selection of queries	429
14.2.1 Copula clauses	429
14.2.2 Presentational focus	430
14.2.3 Focus adverb constituent position	431
14.2.4 Local contrast	433
14.2.5 Contrastive left dislocation	434
14.2.6 Occurrence of <i>wh</i> -clefts	435
14.3 Statistics of tables and figures	436
14.3.1 The decline of subject-finite-verb inversion in main clauses	436
14.3.2 Chain-starting PPs in main clauses	436
14.3.3 New and chain-starting PPs found in main clauses and subclauses	436
14.3.4 New subject presentational focus per chainlength category	437
14.3.5 New subject presentational focus per clause type	437
14.3.6 New subject presentational focus for medium and large subject chains	437
14.3.7 The decline of subjects occurring after the finite verb in main clauses	438
14.3.8 Postverbal presentational focus with syntactic subjects versus expletives	438
14.3.9 Main clause subjects that occur after the finite verb and that are linked	438
14.3.10 Unanchored non-quantified subjects occurring after the finite verb	438
14.3.11 NPs and PPs modified by a focus adverb	439
14.3.12 Postverbal subject location in main clauses with the verb <i>have</i>	439
14.3.13 Preverbal noun phrases with local contrast	439
14.3.14 The position of CLD resumptive demonstrative pronouns	440
14.3.15 Syntactic category of the clefted constituent	440
14.3.16 Clefted constituents preceding the copula	440
14.3.17 Information status of the clefted constituent	441
14.3.18 Information status of the cleft clause	441
14.3.19 Combined information status of clefted constituent and cleft clause	442
14.3.20 Information structure status of the cleft	442
14.3.21 Emphatic cleft types	443
14.3.22 Subject-auxiliary inversion for clause-initial focused PPs	443
Index	445
Curriculum Vitae	449



List of Tables

Table 1 Subject-finite-verb main clauses with neutral and inverted word orders	11
Table 2 Old English slot-structure	60
Table 3 Late Modern English slot-structure	62
Table 4 Charted representation of the sentences in (3a-c)	74
Table 5 Change in the position of three crucial slots of automatically charted texts	77
Table 6 Word orders in Old English motivated by syntax or text-organization.....	83
Table 7 Division of Old English into slots	84
Table 8 Narrative divisioning by special clause-types.....	92
Table 9 Focus types per focus articulation	100
Table 10 Pragmatically neutral word orders in late Modern English.....	112
Table 11 Division of late Modern English into slots	112
Table 12 Division of late Modern English “there” clauses into slots	121
Table 13 A late Modern English dominant focal element	125
Table 14 Comparison of information status category sets	152
Table 15 Deriving other information status categories from the Pentaset	153
Table 16 Types of “XP be YP” copula clauses depending on the referential and syntactic categories of their components (surface word orders).....	165
Table 17 Constraints used to determine the best antecedent.....	189
Table 18 Suspicious situations	193
Table 19 Reference types used in the case study	195
Table 20 Crucial constraints in the “Long” text	196
Table 21 The result of using the <code>ru:ard()</code> function	211
Table 22 Prepositional phrases in main clauses found by query (179)	214
Table 23 Length distribution of chains started out by main clause and subclause PPs	216
Table 24 Texts that have been enriched with referential information	227
Table 25 Word order categories for subjects	228
Table 26 Coreferential chain length categories	228
Table 27 Late subjects that are linked and unlinked	237
Table 28 Reintroduction of subjects after an absence of more than 50 clauses	243
Table 29 Possible constituent focus diagnostics	251
Table 30 Word order categories for main clause constituents	252
Table 31 Adverbs for focus and emphasis found in the parsed English corpora	254
Table 32 Positional distribution of adverb-marked constituent focus.....	256
Table 33 Positional distribution of local-contrast marked constituent focus	263
Table 34 The position of CLD resumptive demonstrative pronouns	270

Table 35 Occurrence of wh-clefts versus reversed wh-clefts	276
Table 36 Johansson's discourse functions of clefts related to information states	304
Table 37 Word orders in sentences containing wh-question words	310
Table 38 Looking for Chechen it-clefts	327
Table 39 Results of the Chechen it-cleft database	331
Table 40 Syntactic and information state features used in the English cleft database	344
Table 41 Clefts that were mistakenly taken for complement clause constructions	354
Table 42 Clefts that were mistakenly taken for simple relative clause constructions	356
Table 43 Number of it-clefts found in the parsed corpora	357
Table 44 Cleft type categories	363
Table 45 Information Structure Status categories	365
Table 46 Emphatic cleft types per 100,000 main clauses	369
Table 47 Focus-particle it-clefts and those of them that are adjuncts	371
Table 48 A query execution table defined in the constructor editor	426
Table 49 Table with results provided by CorpusStudio	428



List of Figures

Figure 1 The decline of subject-auxiliary inversion in main clauses from OE (Old English) until LmodE (late Modern English)	11
Figure 2 Discourse and situation model	29
Figure 3 Visualization of the three-dimensional syntax-focus-text space.....	57
Figure 4 The decline of subject-auxiliary inversion in main clauses from OE (Old English) until LmodE (late Modern English)	70
Figure 5 Prince's (1981) taxonomy of given-new information.....	135
Figure 6 Information states based on discourse and hearer	137
Figure 7 The topic acceptability scale (Lambrecht, 1994).....	138
Figure 8 Ariel's accessibility marking scale (Ariel, 1999)	139
Figure 9 The givenness hierarchy (Gundel et al., 1993)	139
Figure 10 The manual annotation program "Cesac"	142
Figure 11 The referential state primitives in the Pentaset.....	144
Figure 12 Suspicious context in a text from the parsed English corpora	159
Figure 13 Conversion from (a) psd format (right) to (b) psdx format (left).....	182
Figure 14 The semi-automatic coreference resolution algorithm	184
Figure 15 Chain-starting PPs in main clauses.....	214
Figure 16 New and chain-starting PPs found in main clauses and subclauses.....	215
Figure 17 PP-initiated chains with at least one subject.....	217
Figure 18 New subject presentational focus per chainlength category	230
Figure 19 New subject presentational focus per clause type (see Table 25)	231
Figure 20 New subject presentational focus for medium and large subject chains.....	232
Figure 21 The decline of subjects occurring after the finite verb in main clauses	233
Figure 22 Postverbal presentational focus with syntactic subjects versus expletives	235
Figure 23 Main clause subjects that occur after the finite verb and that are linked	236
Figure 24 Unanchored non-quantified subjects occurring after the finite verb in main clauses.....	242
Figure 25 The proportion of NPs and PPs modified by a focus adverb occurring before the finite verb in main clauses	255
Figure 26 Percentage of "have" clauses with postverbal subject.....	258
Figure 27 Percentage of local contrast noun phrases before the finite verb.....	263
Figure 28 Pitch track and speech waveforms of an utterance of (267)	313
Figure 29 Pitch track and speech waveform of an utterance of (268).....	315
Figure 30 Pitch track and speech waveform of an utterance of (271).....	317
Figure 31 Pitch track and speech waveform of an utterance of (272).....	318
Figure 32 Pitch track and speech waveform of an utterance of (274).....	319

Figure 33 Pitch track and speech waveform of an utterance of (275).....	320
Figure 34 The chechen it-cleft database seen from Cesax	330
Figure 35 Coding of a typical cleft	350
Figure 36 Editing of cleft features within Cesax	352
Figure 37 Number of it-clefts normalized per 100,000 main clauses	358
Figure 38 Syntactic category of the clefted constituent	359
Figure 39 Clefted constituents preceding the copula	360
Figure 40 Information status of the clefted constituent	361
Figure 41 Information status of the cleft clause.....	362
Figure 42 Combined information states of clefted constituent and cleft clause.....	364
Figure 43 Information Structure Status of it-clefts in selected sub periods	367
Figure 44 Division of emphatic clefts into types	368
Figure 45 Emphatic it-clefts compared with clause-initial focus marking.....	370
Figure 46 Subject-auxiliary inversion for clause-initial PPs with a focus adverb or particle	372
Figure 47 De vijf basismogelijkheden om te verwijzen	415
Figure 48 Main definitions of a corpus research project in CorpusStudio.....	423

Abbreviations

General

3ms	third person masculine singular	FP	focus phrase
3ns	third person neuter singular	GEN	genitive case
3s	third person singular	H*	high tone pitch accent; lexical high tone
ACC-gen	Accessible from general world knowledge	H _a	accentual phrase boundary high done
ACC-inf	Accessible from inference to items in prior discourse	INF	infinitive
ACC-sit	Accessible from the situation	InP	intonation phrase
AcP	accentual phrase	IP	inflectional phrase
ADJ	adjective	IP-MAT	main clause
ADV	adverb	IP-SUB	subordinate clause
AgrSP	subject agreement phrase	L _a	accentual phrase boundary low tone
AP	adverbial phrase	LFD	left dislocation
ARG	argument	L _i	intonation phrase boundary low tone
B1	LmodE (1700-1769)	LmodE	late modern English (1700-1914)
B2	LmodE (1770-1839)	M1	ME (1150-1250)
B3	LmodE (1840-1914)	M2	ME (1250-1350)
BNC	British national corpus	M3	ME (1350-1420)
C	complementizer	M4	ME (1420-1500)
C ⁰	head of the CP	ME	middle English (1150-1500)
CF	constituent focus	MENT	mental entity
CP	complementizer phrase	N	noun
D	determiner	NEST	non-established information
DAT	dative	NLP	natural language processing
DET	determiner	NONSPEC	non-specific
DFE	dominant focal element	NP	noun phrase
DO	direct object	NP-OB1	direct object NP
E1	eModE (1500-1569)	NP-OB2	indirect object NP
E2	eModE (1570-1639)	NP-PRD	predicative NP
E3	eModE (1640-1710)	NPR	proper noun
EEG	electroencephalogram	NPR\$	possessive proper noun
eModE	early modern English (1500-1710)	NPRS	plural proper noun
ERP	event-related brain potential	NP-RSP	resumptive NP
EST	established information	NP-SBJ	subject NP
FD	focus domain		

NP-VOC	vocative NP	RefPoD	referential point of departure
NS	plural noun	RFL	reflexive
O1	OE (450-850)	RRG	role and reference grammar
O2	OE (850-950)	S	subject
O3	OE (950-1050)	SC	Shali-Chechen dialect
O4	OE (1050-1150)	SOV	subject-object-verb
OE	old English (450-1150)	SPEC	specifier
OVS	object-subject-verb	SV	subject-verb
OVS	object-verb-subject	SVO	subject-verb-object
PDE	present-day English	TC	topic-comment articulation
PF	presentational focus	TEI	text-encoding initiative
PGN	person-gender-number	V2	verb-second
PL	plural	V _{fin}	finite verb (also V _f , V _{finite})
PoD	point of departure	V _{non-finite}	non-finite verb
POSS	possessive	VP	verb phrase
PP	prepositional or postpositional phrase	WH	<i>wh</i> -constituent
PRO\$	possessive pronoun	XP, YP	any kind of constituent
PRS	present tense	XVS	XP-verb-subject
PST	past tense		
QUANT	quantifier NP		

Chechen

ALL	allative
B	noun class “B”
D	noun class “D”
ERG	ergative
IMPF	imperfective past
J	noun class “J”
LOC	locative
NMLZ	nominalizer
OBL	oblique case
PLSE	polite request
PSTN	past tense with the <i>-na</i> suffix
PSTR	past tense with the <i>-ra</i> suffix
PTC	predicational participle
QM	polar question
REL	relativizer (attributive participle)
SG	singular (for verbs)
V	noun class “V”

Part I

Introduction

This dissertation concentrates on a vital aspect of written communication: focus. Where others have tried to gain understanding of the rules by which we know that a particular word or phrase should be read with emphasis, the research presented in this book addresses the fundamental question how rules for emphasis, focus rules, *interact* with syntactic rules—the rules used to determine the grammatical relations between words and constituents.

There are several reasons why one would want to know how focus interacts with syntax. The first one is related to second language learning. Learning a language might seem nothing more than learning its vocabulary and its syntax, but research has shown that near native speaker abilities can only be reached by a proper understanding and a proper use of focus rules (see Hannay and Mackenzie, 2002 on information-structure influenced word order in English, and see Lozano, 2006 on the proper acquisition of discourse-sensitive Spanish word order). It is us, linguists, who need to find these rules, investigate how they interact with syntax rules and then make a description of the system we have found available to the language learners, so that they can reach a higher level of proficiency.

Another reason why the *interaction* between focus and syntax warrants research relates to the publication of grammatical descriptions of languages. If the rules by which we understand that something is focused are *part of* the syntax rules, this would mean that a proper grammatical description of a language must include them, and this, in turn, means we have to find out what they are.¹ If focus rules are independent of syntax, but can change from language to language, we too would need to describe them. Only—and now I am speaking hypothetically—if focus rules for any given language automatically would derive from more general rules for the structuring of information combined with the specific rules governing the syntax of that language would we be free to abstain from the tedious task of investigating and describing focus for every individual language. To be sure: the quest for the interaction between focus and syntax rules has far-reaching consequences.

The strategy this dissertation takes to understand the interaction between focus and syntax is to consider the changes that have taken place in the syntax and focus in the history of one well-documented language: English. This language has undergone major syntactic changes in the course of its 1000+ years of history, and documents in it are available from before 1000 A.D. until now, and many of these have been digitized and syntactically parsed, so that we have a well-sized corpus available that we can use to seek answers to the questions we have. The main idea, then, is that we take note of the changes in English syntax, investigate the changes in the way focus is expressed, and evaluate the interchange between these parallel developments: have changes in syntax, for instance, led to changes in the way focus is expressed?

One of the major contributions to the English historical line of research is Ball (1991), who investigated the development of the cleft construction, and concludes that the *it*-cleft emerged in late Middle English and early Modern English (around 1500 A.D.). Since this construction is often perceived as a prototypical focusing device, one could envision a scenario whereby Old English, the predecessor of Middle English, had different means to express focus, but when the syntax of the language changed, the traditional way of expressing focus became less appropriate (or perhaps even unavailable), so that language speakers had to “create” new ways to express emphasis, which then resulted in the birth of the *it*-cleft for this particular purpose.

While it will be shown that this scenario goes some way to account for the data, we will also see that it has its problems. Contrary to what has been claimed by Ball (1991) and later by Patten (2010), I hypothesize that the *it*-cleft did not suddenly emerge out of nowhere in Middle English, but was already present in Old English. The function of the construction was mainly to support text organization, but then its ability to express narrow focus made it an ideal candidate at the time when changes in English syntax jeopardized earlier focus strategies. The hypothesis that the *it*-cleft may have text organization as its main function will be shown to be plausible, when we observe the role it plays in a language like Chechen, which has this construction, but never uses it for focusing.

Part of this thesis is devoted to shedding more light on the role played by *it*-clefts in expressing focus, but, given our overall goal of establishing what the relation between focus and syntax is, we will not stop there. Specifically, we will consider how two important kinds of focus (constituent focus and presentational focus; see chapter 4) were expressed in English over time, and how the focusing strategies relate to the changing syntax.

In this introduction, we briefly consider the nature of syntax (1.1) and then we will have a, necessarily brief, look at some of the major changes in English syntax (section 1.2; chapter 4 contains more on syntax). Against the background of the overall research question about the relation between syntax and focus, section 1.3 presents the aims of the present study more specifically. Section 1.4 briefly discusses some of the methodological challenges attached to a corpus-based study, and section 1.5 lists the corpora used. Section 1.6 discusses the organization of this book.

1.1 Syntax

Given the aim of this study to look at the relationship between syntax and focus, I would like to briefly touch upon the question what should be considered as belonging to “syntax” (I postpone a discussion on the nature of focus to chapter 3).

One dictionary of linguistic terms defines syntax as “... *the way words are put together in a language to form phrases, clauses, or sentences*” (Loos, 2003). This definition stresses a major role of syntax, which is to define which words belong to one phrase, which phrases form larger constituents such as clauses, and which

clauses combine together into sentences, and several other definitions stress this same “vertical” hierarchical role of syntax (Crystal, 1980, Tesnière, 1959).

Other definitions of syntax, however, are broader: they regard dependency relations between words and constituents as one area within a larger definition of syntax as “the study of the principles and processes by which sentences are constructed in particular languages” (Chomsky, 1957). This last definition of syntax seems to incorporate *everything* that contributes to the “construction” of a sentence in a language, and this, necessarily, includes word order. This brings us to an important matter we need to resolve at the start of this book: how does *word order* relate to syntax? Consider what Dryer writes:

- (1) “Languages also vary in the extent to which the order of elements is fixed. In some languages (e.g. English), only certain orders of S, O, and V are grammatical, and one order is the dominant one; but other languages allow all six orders. In some of the latter group (e.g. Russian), one order is dominant; in others (e.g. Cayuga, an Iroquoian language), the order is sufficiently flexible that no single pattern is dominant. The degree of flexibility is related to the function of word order in the language. In some languages, like English, order indicates which noun phrase is subject and which is object; in others, order does not mark grammatical function, but varies with discourse properties of the different elements in the clause (cf. Givón 1983, Mithun 1987).” (Dryer, 2003)

Dryer observes that some languages use word order to encode “grammatical function”, whereas others, having other strategies to signal grammatical relations, use word order for marking “discourse properties”. The approach I will be using in this book is based on Dryer’s observations: I consider word order to be *partly* part of syntax, and I do this by adopting the following definition of syntax:

- (2) *Definition of syntax*
The syntax of a language is the set of rules describing the way by which grammatical functions or relations are signalled.

What the definition above says is that syntax aims at signalling to the language user what the “grammatical” functions or relations are of words, phrases and clauses. The signalling of grammatical functions and relations may be done by methods such as case, agreement, juxtaposition and word order. Agreement in case can be used to signal that words belong to one and the same constituent, as for instance in the Old English phrase *halgum gewirtum* ‘holy writings’, where dative case agreement signals that the two words are part of one and the same phrase. If we accept that case agreement combines with *adjacency* here to indicate that the two words belong to one and the same phrase, then the question is what determines whether the *word order* of the phrase is [Adj-N] or [N-Adj]. A look at the electronically available OE texts reveals that full noun phrases almost exclusively have the word order [Adj-N], which indicates that there is a kind of “default” word order.² Such default word orders facilitate the processing of language, allowing the parts that do not need much attention to be automatized. Occurrences of the adjective following the noun in OE

occur when another attributive element precedes the noun, such as *þreom wicum fullum* ‘for three full weeks’. Word order regularities such as “an adjective precedes the noun it modifies” facilitate the recognition of constituent boundaries, and are also part of the syntax of the English language.³ Other languages, however, may not have the same word order restrictions. French apparently allows both [Det-Adj-N] (such as *un domestique simple* ‘a SIMPLE servant’, with the focus on the adjective “simple”) as well as [Det-N-Adj] (such as *un simple domestique* ‘a simple servant’, with the focus on the noun “servant”); both orders are acceptable, but one of them is the default or unmarked order, and the other order has a slightly different meaning. Without expanding on the significance of word order within noun phrases (which is beyond the scope of this study), the point I would like to make here is that word order is part of syntax when it is needed to indicate a grammatical function or relationship. OE apparently requires some kind of attributive element to precede the Noun; French *syntax* allows adjectives to follow or precede the Noun, but with a difference in meaning.

Another example on the level of the constituent is that of Prepositional Phrases (PPs) in OE: these constituents may vary in the relative order of the preposition and the Noun Phrase (NP) that is being modified. In the PP *to sumum mynstre* ‘to a minster’ (in the clause “he came to a minster”), for instance, the preposition precedes the NP, whereas in the PP *him to* ‘to him’ (in the clause “she began to speak to him”) the preposition follows the NP.⁴ Adjacency of a preposition and an NP is, apparently, enough to indicate the fact that they combine into a PP, and there are situations where even adjacency is not needed (preposition stranding). The variation in word order (P-NP versus NP-P) can then be used to signal meaning differences that are not strictly syntactic in nature, but perhaps more semantically or pragmatically related.

Let us now turn to the clause level. When grammatical case is used to signal a grammatical function like “subject”, “direct object” or “indirect object”, then constituent order (location relative to a verb—finite or infinite—in the clause) need not be used to signal this grammatical function. Word order variation can then be used to signal pragmatics matters (of which chapter 3 will discuss more). If, on the other hand, grammatical case cannot be used to indicate the fact that a particular NP is the subject (which is the case in Present-day English, unless the NP is a pronoun that has a form unambiguously signaling its case), then constituent order has to be used to convey the proper grammatical relation: the subject is the NP that precedes the finite verb, and the object is the one that follows it.

In sum, syntax involves the rules to convey grammatical functions and relations, these rules can make use of strategies such as case, agreement, adjacency, and where necessary also word order. A language usually also contains “default” word orders, which serve to ease the processing burden, and which can also be regarded as part of a language’s syntax. Where word order is *not* necessary for syntax, variations can sometimes be used for pragmatics.

1.2 English word order changes

After the invasion of the Jutes, the Angles and the Saxons in 449 A.D., English started developing as a separate language (see Baugh and Cable, 2002 for a detailed history of the language). The first extant manuscripts in this language are several centuries later.⁵ The numerous manuscripts that appear from then on provide insight into the development of English into its present form. What began as a collection of tribal languages similar in many respects to present-day German and Dutch grew to its present form, which differs in many respects from its predecessor. The subsequent sections touch upon some of the syntactic changes that have taken place, inasmuch as they are relevant for this current study, and they also contain a short introduction into the corpus research methodology used in this study. A fuller account of changes in English word order phenomena is included in chapter 4.

1.2.1 Verb-second

Old English, the ancient predecessor of Present-day English, can in some sense be regarded as a “verb-second” language: a language where the finite verb (the verb inflected for tense, person and number) appears as the second constituent in the main clause, even if the first constituent is not the subject, but, for instance, a prepositional phrase or an adverb (Los, 2012, van Kemenade, 2012). There are sentences that seem to indicate Old English is a verb-third language, since the finite verb only appears as the third constituent in the main clause, as illustrated by (3), where the finite verb has been set out in bold-face.

- (3) a. Ða **wurdon** hire ylðran swiðlice geblissode þurh hi. [coeuphr:25]
 then were her parents greatly blessed through her
‘Then her parents rejoiced exceedingly on her account.’
- b. Ongemang þissum, **com** ham Pafnuntius. [coeuphr:88]
 in.the.midst of.this came home Paphnutius
‘In the midst of this, Paphnutius came home.’
- c. Ða se cniht **bæd** hine þæt he come mid him
 then that servant asked him that he come with him
 to Eufrosinan. [coeuphr:98]
 to Euphrosyne
‘Then the servant prayed him to come with him to Euphrosyne.’

The example in (3a) is typical of verb-second: the finite verb *wurdon* ‘were’ follows in second position after the initial constituent, the time adverbial *þa* ‘then’. Along the same line is example (3b), where the first constituent is a prepositional phrase *ongemang þissum* ‘in the midst of this’, followed by the finite verb *com* ‘came’. An example that seems to contradict a strict verb-second analysis of Old English is (3c), where the verb appears in third position. There are *two* constituents preceding the finite verb *bæd* ‘asked’ in (3c): the adverbial phrase of time *þa* ‘then’ and the subject *se cniht* ‘that servant’. One explanation that has been given for this kind of deviation to the verb-second regularities that are observed, is that pronominal or otherwise referentially established subjects have a dedicated subject position on the left of the finite verb (van Kemenade, 2012). Old English, then, is regarded as having two

positions for the subject: established (given) subjects precede the finite verb, and non-established ones follow it. Such a structure can be seen as a grammaticalization of the “Principle of Natural Information Flow” (which will be explained more fully in 3.3.1), where more established material precedes less established information.⁶ Further variation includes (4a), where three constituents precede the finite verb: the adverb *weninga* ‘perhaps’, the subject *God*, and the indirect object pronoun *him* ‘to him’.

- (4) a. Weninga God him **hæfd** be me sum þing onwriġen. [coephr:295]
 Perchance God him has by me some thing revealed
 ‘Perhaps God has revealed something to him about me.’

One possible explanation for the word order in (4a) could, again, be related to the Old English tendency to put established (given) material before unestablished (non-given) information, which could include not only subjects, but objects as well. A full study of word order variation is beyond the scope of this book. I focus on the expression of focus, and where possible indicate how it interacts with syntax.

1.2.2 Subject-finite-verb inversion

I would like to illustrate the changes in English syntax by looking at the decline of a phenomenon called “subject-finite-verb inversion”: a subject that would normally precede the finite (since the neutral word order in English is Sbj-V_{finite}), now follows it, so that the word order V_{finite}-Sbj results. Subject-finite-verb inversion was relatively frequent in earlier English, but Present-day English has retained it in a few clearly recognizable contexts (where it is restricted to auxiliaries), some of which are illustrated in (5), where the finite verb (the auxiliary) is in bold-face, with the subject underlined (examples are from the “British National Corpus”; see section 1.5).

- (5) a. Who **did** you rob for this? [BNC HTY:160]
 b. **Does** the pattern seem satisfactory in the longer term? [BNC K8Y:808]
 c. In no way **did** she wish ill health on the woman. [BNC JXS:3195]
 d. Not a tear **did** she shed. [BNC EFP:35]

The auxiliary in Present-day English obligatorily precedes the subject in *wh*-questions (5a), in polar questions (5b) and with negated adverbials, such as the negated prepositional phrase *in no way* in (5c). There is a tendency too for negated objects, such as *not a tear* in (5d), to appear clause-initially, giving rise to subject-auxiliary inversion.

Historically speaking, subject-finite-verb inversion can be seen as a remnant of the Old English verb-second rule, which, in its strictest sense (but there are exceptions as we have seen above in section 1.2.1), has the verb as the second constituent, with the first constituent reserved for contexts that sometimes are syntactic in nature (*wh*-question placement, for instance), and sometimes pragmatically motivated (as we will start to see in chapter 4).

How are we to visualize the suggested decline in subject-finite-verb inversion? I would like to take an excursion here and attempt to give a partial answer to this

question. There are two reasons for this excursion: subject-finite-verb inversion is one of the clearest syntactic changes in English, and an extended excursion on *how* we visualize a linguistic change gives a clearer idea of the kind of research described in chapters (8)-(12), to which the theoretical groundwork explained in chapters (2)-(7) builds up.

1.2.2.1 The plan

What we will do, to summarize the general plan, is look for subject-finite-verb inversion in an available set of English texts that are taken from time periods ranging from Old English (starting just before 900) until late Modern English (ending roughly at 1900). These texts belong to four different corpora, which are described and referred to in section 1.5. What is important for now is to understand that these texts have all been annotated syntactically: the category (verb, noun, adjective etc) of each word has been added, and the hierarchy of words into phrases and phrases into clauses has been made clear:

(6) *Syntactic annotation of sentence (3b)*

```
(IP-MAT
  (PP (P Ongemang) (NP-DAT (D^D pissum)))
  (, ,)
  (VBDI com)
  (ADVFP-DIR (ADV^D ham))
  (NP-NOM (NR^N Pafnuntius))
  (. ,))
```

The example in (6) is a “labelled bracketing” representation of (3b), which is one line from the Old English text called “Euphrosyne”, and the hierarchy provided by the brackets shows that the whole sentence is a main clause (an “IP-MAT” in the annotation language) containing four constituents: a PP *Ongemang pissum* ‘meanwhile’ (with internal structure), a finite verb *com* ‘came’, an adverbial phrase of direction *ham* ‘home’ and a nominative case NP *Pafnuntius*.

With an idea of what the syntactically annotated English texts look like, we can now formulate the strategy to look for subject-finite-verb inversion more clearly: we will need to look for the relative occurrence of two different main clause word orders:

- (7) a. XP – Subject – Auxiliary – ... – V_{non-finite}
 b. XP – – Auxiliary – Subject – ... – V_{non-finite}

The word order in (7a) is the regular one where the subject precedes the finite verb (the auxiliary) and the one in (7b) is the inverted word order, where the subject follows the finite verb. We can look for the word orders in (7a) and (7b) by identifying all main clauses (those that are tagged IP-MAT), and see if they have the necessary components: a subject NP, an auxiliary, a non-finite verb, and some constituent that precedes both subject and auxiliary.

1.2.2.2 The search

We will trace how subject-finite-verb inversion changed in the history of English by looking at a collection of texts that roughly span the period 900-1900 AD. The surface structure oriented syntactic annotation of these texts allows us, in principle, to look for two specific word orders: one neutral word order (7a) and one that contains subject-finite-verb inversion (7b).

The question now at hand is *how* we can go about doing this search for sentences with particular word orders in the available texts. The question of “querying” (that is: searching through) annotated text corpora will be dealt with extensively in chapter 7, especially in the light of the information added in order to identify focus, as described in chapters 3-6, but we will take the opportunity here to look ahead, in order to clarify the task at hand.

We will conduct our search for sentences with the word orders specified in (7) by means of two algorithms with which a computer program (the program “CorpusStudio”) will work its way through the available texts.⁷ The algorithms will ultimately have to be written in computer-readable format, but it is enough for this moment to look at the strategy of one of them: the one that looks for the inverted word order:

(8) *Algorithm that finds the inverted word order in (7b)*

Step 1: Consider each constituent in the text; select any that is a main clause

Step 2: Check if it has the following “child” constituents:

First constituent, Subject (as different constituent),

Finite verb, Non-finite verb

Step 3: Check word order conditions:

Condition a: the “First constituent” precedes “Finite verb”

Condition b: the “Finite verb” precedes “Subject”

Condition c: the “Subject” precedes the “Non-finite verb”

Step 4: Output:

If all Conditions are met, add this line to the output

Step 1 in algorithm (8) selects only those constituents that are main clauses (this can be done by looking at the “label” of the constituent, which is IP-MAT for main clauses). Step 2 looks at the “children” of the main clause that has been found: those constituents that are hierarchically directly under the main clause. At least four children must be found, and they need to have the labels that match those of a subject, a finite verb and a non-finite verb (the label of the “first constituent” does not need checking). Knowing that we have identified a main clause with the correct child constituents, step 3 checks for the correct word order: first constituent, finite verb, subject and then non-finite verb, as in (7b). The last step marks the line that has been found as belonging to the “output” of the program, which, as we will see in chapter 7, consists of two parts: (a) the total number of sentences that fulfil the algorithm’s conditions, as divided over different English time periods, and (b) the text and syntax of each sentence, accompanied by a little bit of context.

The actual algorithms that have been used in order to get the results shown in Table 1 and Figure 1 in the next section are slightly more complex, since they also

determine the kind of first constituent that should be there according to (7), so that we get separate results for “Object” first constituents, “PPs” and “Adverbs”.

1.2.2.3 The outcome

The outcome of the algorithm defined in (8) and the algorithm that finds the “uninverted” word order from (7a), comes first of all in the form of a table, where each row is intended for one word order, and each cell in that row gives the number of main clauses found in a particular time period that satisfy the word order of that row:

Table 1 Subject-finite-verb main clauses with neutral and inverted word orders

FirstConst	Type	900-1150	1150-1500	1500-1710	1700-1914
Obj	neutral	89	141	205	74
Obj	inverted	96	188	126	29
PP	neutral	310	2274	3619	2342
PP	inverted	283	911	409	89
Adv	neutral	764	1394	2091	834
Adv	inverted	1191	1028	484	78

The absolute number of occurrences reported in Table 1 shows that there are considerable differences in the total number of main clauses found in the English texts that satisfy a particular word order at a particular time period. We get a much better idea of the trends if we look at the proportion (the percentage) of main clauses with subject-finite-verb inversion for each particular time period (and for each particular first constituent type: object, prepositional phrase or adverbial), which is what Figure 1 shows.

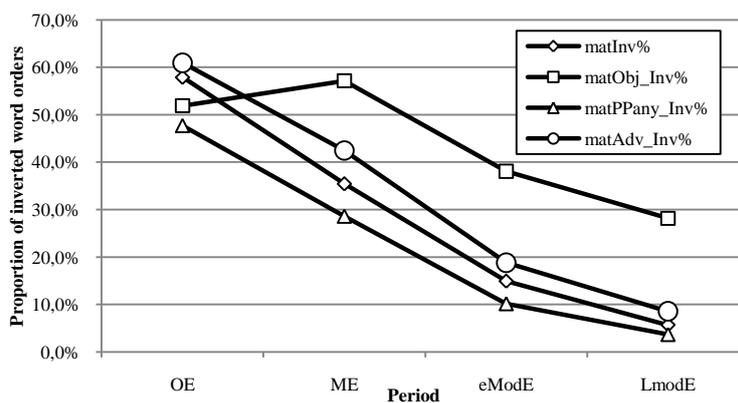


Figure 1 The decline of subject-auxiliary inversion in main clauses from OE (Old English) until LmodE (late Modern English)

The total percentage of main clauses with inversion, those with the pattern XP-Aux-Sbj-V_{non-finite}, are indicated by the line “matInv%” in Figure 1, and this shows a clear

trend from almost 60% in Old English (OE), to 35% in Middle English (ME), then 15% in early Modern English (eModE), finishing with 5% in late Modern English (LmodE).⁸ The other lines in Figure 1 show the individual patterns for subject-finite-verb inversion where the clause-initial constituent is an object (matObj_Inv%), a prepositional phrase (matPPany_Inv%) and an adverbial (matAdv_Inv%).

Subject-finite-verb inversion in Present-day English is, as illustrated above, restricted to well-defined syntactic situations like questions and negation. Subject-finite-verb inversion in Old English, however, could occur under different circumstances, and (9) shows a few examples that have been supplied as the second “outcome” by running the implementation of algorithm (8) in the program CorpusStudio.

- (9) a. Ða **weard** he gehyrt þurh þas word. [coeuphr:236]
 then was he heartened through that word
‘These words consoled him.’
- b. (An Antiochia þare ceastre wæs sum cyningc Antiochus gehaten:)
 æfter þæs cyninges naman **wæs** seo ceaster Antiochia geciged.
 after that king’s name was this city Antioch called
*‘(In the city of Antioch there was a king named Antiochus,) [coapollo:3-4]
 from whom the city itself took the name Antioch.’*
- c. (Ða ic ongear com, þa sædon hi me þæt min dohtor wære forðfaren,)
 and me **wæs** min sar eal geedniwod. [coapollo:506-507]
 and me was my pain completely renewed
*‘(When I returned, they told me that my daughter was dead,)
 and my pain was all renewed to me.’*

One common trigger for subject-finite-verb inversion in Old English is the clause-initial time adverbial *þa* ‘then’ in (9a), which we have also seen in example (3a). The translation into Present-day English no longer works with inversion. The next example in (9b) illustrates a subject-finite-verb inversion triggered by the clause-initial prepositional phrase *æfter þæs cyninges naman* ‘by the king’s name’. A literal translation into Present-day English would be: “By that king’s name, the city was called Antioch”, but this would put heavy contrastive emphasis on the prepositional phrase, suggesting that there were other kings whose names could have served as the basis for the naming of Antioch. The translation provided by Archibald (1991), which is the free translation given above, circumvents the problem of contrastive emphasis on prepositional phrases that precede the subject in main clauses by placing the sentence in a subordinate clause.

The third example in (9c) illustrates yet another Old English trigger for subject-finite-verb inversion, which may very well have a pragmatic motivation: emphasis on *me*. If this is a correct interpretation, then we see a shift from using the preverbal domain to express emphatic constituents in Old English to the clause-final position in Present-day English, as in the translation of (9c). This will be discussed extensively in chapter 4.

1.2.3 Expressing emphasis

The previous section on subject-finite-verb inversion hinted at the idea that one of the functions of the first constituent in older forms of English has been to host emphatic constituents: those with contrastive focus. An example from Middle English where we can see this first constituent feature in action is (10), which is taken from an “Abbreviated history of England” written by John Capgrave.

- (10) a. (That same Gilbert was ryth affectuous onto þe Heremites of Seynt Austin, for, as it is seid, he was aqweyntid with Doctour Gilis in Frauns.)
 and **at his request** Gylis was meued to make [cmcapchr:2686-8]
 and at his request Gillis was moved to make
 þat bok OfGouernauns of Princes.
 that book of governance of princes
‘(This same Gilbert had a true affinity for the Heremites of saint Austin, because, as it is said, he was acquainted with Doctor Gillis in France,) and it was at his request that Gillis was inspired to publish the book about the governance of princes.’

Highlighting of the clause-initial prepositional phrase *at his request* can in Present-day English be achieved by using an *it*-cleft construction: “it was at this request that ...”. The end of the Middle English period (which is around 1500 A.D.) sees a clear rise in the number of *it*-clefts that are being used to express emphasis. As we consider how the ways to express narrow focus have changed throughout the history of English (a search that starts from chapter 9), the *it*-cleft will be one of the constructions studied in more detail.

1.3 Aim of this study

The study described in this book is quite a broad one, connecting such diverse areas as psycholinguistics, corpus linguistics, syntax and information structure. All of these disciplines are called for in order to address one overall research question:

- (11) *Research question in this study*
 What can we learn about the interaction between syntax and focus, when we look at the development of the English language as visible in the available syntactically parsed corpora?

This question confines the study to the interaction between syntax and focus, which means that we will not look at matters that either touch only on syntax or only on focus. A further restriction is that we concentrate on the English language—although one chapter (11) considers a totally different language to support one of my main arguments. One final, and perhaps most important, restriction of this study is that we will only look at data from available syntactically parsed corpora. The details of these corpora will be presented in section 1.5, but what is important when it comes to the research question addressed in this study is that the material we base this study on is *written* and it is *limited*. The fact that we look at written data means that we will, necessarily, abstain from considering the influence of intonation on focus in the history of English, but instead emphasize on word order, particles and

constructions. Another implication of working with a limited amount of written data is that we will have to do with whatever we have got: we are not in a position to get more data.

I have touched upon the limitations inherent in the formulation of the research question in (11), but there are several advantages that, in my opinion, more than justify the direction taken in this study. The history of English syntax has been studied extensively by a wide group of scholars, which means that this current study is able to build on a considerable body of work that has been done. A study of the syntax and the changes it underwent throughout its development can, for instance, be found in Fischer et al (2000), while work is continuously progressing, witness the diversity in authors and topics appearing in recent handbooks (Nevalainen and Traugott, 2012, van Kemenade and Los, 2006b).

Information structure in English at its various stages of development is another topic that is approached by present-day scholars from various angles of research (Biberauer and Kemenade, 2011, Los, 2009, van Kemenade and Milicev, 2012), and it will be difficult to mention all the research going on into present-day English information structure (Birner and Ward, 1998, Ward, 1985, Ward et al., 2002).

Mentioning research done by others may raise the suspicion that the current study will be an introspective one, but this is not the case. Even though I aim to mention and make use of the results gained by others in the past, the core of this study involves original corpus research work undertaken by myself, in close cooperation with my colleagues. The kind of data that proved to be necessary to answer questions concerning the interaction between syntax and focus is an enrichment of already existing corpora, and in the process of annotating the enrichment in order to do my original corpus research work, I have made an important discovery about the nature of “focus” itself: focus is compositional (in the sense that syntactic and referential information can be combined to determine the “focus”). But I will leave the decomposition of focus into its more fundamental components to chapters 2 and 5, while the research described in chapters 8-12 will serve to underline the practical usefulness of this fundamental principle.

1.4 Methodological issues

A large part of the research described in this dissertation is based on corpus data. In what follows I will explain why we can make use of such data, and what quantitative measures we can use to evaluate the significance of our results.

1.4.1 The corpus as a “corpulect”

The main reason to use corpus data is that our quest for the relation between focus and syntax should aim at deciphering the mechanics of *real* language, as it has been used by people for communicative purposes, and texts that are sampled in a corpus represent real text. Nevertheless, I do realise that working with corpora has several pitfalls we need to be aware of. The first issue is to what extent a corpus is representative of the language as it is used by native speakers in a natural environment. I propose a corpus (or a subcorpus) represents a “**corpulect**”: a cross-

section of a language with clear boundaries, on a par with “idiolect” (the language of one individual) “dialect” (the language spoken in one place) and “ethnolect” (the language spoken by one ethnic group). A corpus is a selection of texts, and sometimes even parts of texts, since the actual texts are too long to fit into a corpus of manageable size. This selection is a subjective choice: the corpus developers decide which texts and which parts of those texts are included in the corpus, and which are not. They do so for very valid reasons: they usually aim for a corpus that contains a good mixture of text genres and time periods. While I think everyone would agree that selecting texts on the basis of time periods is good and objective, the mixture of text genres is not only subjective, but may also lead one astray in a sense. Text genres may appear in a language with vastly different frequencies. That is why a corpus aimed at giving a clear picture of a language should reflect those frequencies. But this is an impossible requirement, if we want a corpus to be of manageable size. It remains unclear how these frequencies should be measured in the first place: do we check for number of publications (and then what do we do with *manuscripts* in OE/ME?) or for number of actually printed copies? And how do we know certain book styles were actually read at all, and did not end up sitting on shelves? Ultimately I think we don’t know these details, and any corpus is only a subjective estimate of a language. But if we accept the content of a corpus to be a “corpulect”, a cross-section of the actual language, we may decide that what we find in such a “corpulect” is to some extent representative of the actual language.

1.4.2 Translated texts

One notable problem with corpora in general (not only diachronic ones) is the presence of translated texts within a corpus. Depending on the translation method used, such texts may be more or less representative of the language we are investigating. Several Old English texts, for example, are translations from Latin originals. And a large part of the Chechen texts are translations from English.

I think we can still make use of translated texts, but we have to be aware of their nature. If a text is a translation, we should be beware of outliers. There is one notable outcome that I would like to bring to the attention of the reader right from the start, since it is crucial for large part of this dissertation. There is one text from Old English that contains a huge number of *it*-clefts with a temporal adjunct as clefted constituent (see chapter 10 for details on the terminology). It is this text that contains the vast majority of *it*-clefts in the whole period of Old English. Bede’s ecclesiastical history of the English church is a translation from Latin. However, the investigations of Ball (1991: 94-95) as well as my own research have shown that the instances of the construction I claim to be *it*-clefts in King Alfred’s translation of Bede do not have a matching cleft construction in the Latin original.⁹ This is why I argue that the cleft results from this text are representative of Old English. The Old English rendering of Bede is not always very literal, but sometimes summarizes the Latin (12a-b), expands it (12c), or only roughly conveys it (12d).

- (12) a. **Pa wæs sumedæge, þætthe sorgende bæd hwonne seo adl to**
 then was some day that he worrying asked when this fit to
him cwome, þa wæs gongende in to him sum þara broðra
 him come then was going in to him one their brothers
 ‘Qui cum die quadam sollicitus horam accessionis exspectaret, ingressus
 ad eum quidam de fratribus.’ [cobede:1879]
*‘He was one day anxiously expecting the hour that his fit was to come on,
 when one of the brothers, coming in to him, said: ...’*
- b. **Ða wæs þy æfterangere his rices, þættese arwyrða**
 then was the next year.DAT his of.rule that the honorable
fæder Paulinus, se wæs geo in Eoferwicceastre biscop, þa
 father Paulinus who was earlier in York-city bishop then
wæs in Hrofesceastre, forðgewat & to Drihtne ferde þy syxtan
 was in Rochester departed and to Lord went the sixth
dæge Iduum Octobrium, æfter þon þe he \$nigontyne winter &
 day.DAT - October after that that he nineteen winter and
twegen monað & an & twentig daga biscophade onfeng.
 two months and one and twenty days bishopric started
 ‘Cuius anno secundo, hoc est ab incarnatione dominica anno DCXLIII,
 reuerentissimus pater Paulinus, quondam quidem Eburacensis, sed tunc
 Hrofensis episcopus ciuitatis, transiuit ad Dominum sexto Iduum
 Octobrium die; qui X et VIII annos, menses duos, dies XXI episcopatum
 tenuit.’ [cobede:1948]
*‘In his second year, that is, in the year of our Lord 644, the most reverend
 Father Paulinus, formerly bishop of York, but then of the city of Rochester,
 departed to our Lord, on the 10th day of October, having held the
 bishopric nineteen years, two months, and twenty-one days.’*
- c. **Ða wæs þy æfterangeare, cwom sum monn,**
 then was the next year.DAT came one man
in Nordanhymbra mægðe, wæs his noma Eomær. [cobede:1152]
 in Northumbria district was his name Eomar
 ‘Deus te incolumem custodiat, dilectissime frater. (No Latin available)
 Quo tempore etiam gens Nordanhymbrorum, hoc est ea natio Anglorum.’
*‘The next year some man came into Northumbria, and his name was
 Eomar.’*
- d. **Ða wæs sona, þæs þe heo þæt gefeoht ongunnon,** [cobede:2409]
 then was soon that that they that fight started
þætthe þa hæðnan wæron slegene & geflemde, ond þritigaldormonna
 that the pagans were slain and fled and 30 noblemen
& heretogena, þa ðe þam cyninge to fultome cwomon,
 and those.who.went.there those who the king.DAT to assistance came
 ‘Inito ergo certamine fugati sunt et caesi pagani, duces regii XXX, qui ad
 auxilium uenerant, pene omnes interfecti.’
*‘The engagement beginning, the pagans were defeated, the thirty
 commanders, and those who had come to his assistance were put to flight,
 and almost all of them slain.’*

Comparison of the Latin and OE reveals that none of the *it*-clefts are translations of a cleft in Latin. The Latin original has different constructions instead of *it*-clefts:

either nothing, or time references in a dative-case noun phrase or discourse markers such as *ergo*, *at uero*, *autem*, *nec*, *qui cum* and *et*.

Another argument against viewing *it*-clefts as an idiolectal phenomenon, and in favour of accepting them as representative of Old English is that they occur in 23 different Old English texts of the parsed corpus (the whole corpus has approximately 100 texts). Even though they usually don't occur more than once or twice in one text, their occurrence in the whole corpus should still be regarded as statistically significant, since they occur in a significant number of different texts.

The Chechen corpus is a different matter. The texts that have been translated from English into Chechen reveal a much scantier use of *it*-clefts, and where they do occur, the word order is not as expected. This is why these texts need to be treated with much more care.

1.4.3 Significance of corpus findings

One final matter related to corpus research is that of statistics and significance, a matter that is immediately relevant for the subject-finite-verb decline findings reported in section 1.2.2.

We would ideally only be able to say something about a phenomenon in a language with enough statistical significance if we would investigate a *random* selection of texts. Texts from corpora are definitely *not* a random selection: they have been carefully chosen. This means that our standard statistical techniques may not readily apply to the corpus data we work with, which brings us to the problem that if we find, say, 18.8% of a particular phenomenon in the corpus data, we can (a) not be sure that this 18.8% translates to 18.8% of “the language” (as per the discussion of “corpulect” in section 1.4.1), and (b) what the error range of this 18.8% is. Is it 18-19%, or 15-25%? Is the 18.8% (484 out of 2875 according to Table 1) found for subject-finite-verb inversion in early Modern English sentences that start with an adverb statistically more significant than the 38% (126 out of 331) found for sentences that start with an object? We just don't know. This is a serious problem for corpus research in general, but I think we should try to give some significance measures to our findings.

Several researchers make use of the Chi-square test or, if the amount of data is too low, its equivalent Fisher's exact test. I too will use these tests, even though they have a large drawback: they only tell us whether there is a significant difference between two points in our data. They would, if we turn back to the numbers found for subject-finite-verb inversion in Table 1, tell us whether the change from 38% in eModE to 28% in LmodE for subject-finite-verb inversion in sentences starting with an *object* is significant or not.¹⁰ They cannot tell us how significant *one* point in our data is, and they don't take into account the size of the corpus, the number of texts in it, and the number of different texts a phenomenon we measure occurs in.

If we accept the corpus as representing a world of its own, a “corpulect” as in 1.4.1, then I suggest that there is at least one additional measure of significance we can calculate, and I will refer to this measure as the “corpulect distribution”. The measure of *corpulect distribution* I propose depends on N_{corp} the number of texts that

are present in a corpus and N_{occ} the number of texts the phenomenon we are measuring is observed in, and it is defined as in (13).

$$(13) D_{corp} = \frac{N_{occ}}{N_{corp}}$$

So if we observe a phenomenon in 10 out of 20 texts, its D_{corp} equals 0.50, but if we observe it in 10 out of 100 texts, its D_{corp} equals 0.10—it is much less significant.

The subject-finite-verb inversion reported in section 1.2.2 can serve as an example here. The inverted word order is found in 76 of 100 texts, while the neutral order is found in 72 of 100 texts in Old English. By late Modern English the corpulect distribution has become more diverse: it is 100% for the neutral word order, against 65% for the inverted order.

The *it*-clefts in Old English serve as a second example. The whole Old English corpus has 100 different texts, and where the cleft occurs, it does so mostly only once or twice in a text. This last fact means that it is a relatively rare phenomenon. But what is its corpulect distribution? It occurs in 24 out of 100 texts, so its corpulect distribution is 0.24 (or 24%). This tells us the phenomenon, though rare, still occurs in a relatively significant distribution.

1.5 Corpora used

The English diachronic data used for this study are taken from four syntactically parsed corpora:

- YCOE: the York-Toronto-Helsinki Parsed Corpus of Old English Prose, which contains approximately 1.5 million words, divided over 100 texts (Taylor et al., 2003). Its earliest manuscripts are from the 9th century, and the time range runs from 450 until 1150 A.D.
- PPCME2: the Penn-Helsinki Parsed Corpus of Middle English, second edition (Kroch and Taylor, 2000). This corpus contains about 1.2 million words, which are divided over 55 text samples, and it covers a period from 1150 to 1500 A.D.
- PPCEME: the Penn-Helsinki Parsed Corpus of Early Modern English (Kroch et al., 2004). It contains about 1.7 million words, which are divided over 448 text samples. The period it covers runs from 1500 to 1710 A.D.
- PPCMBE: the Penn Parsed Corpus of Modern British English (Kroch et al., 2010). This corpus contains about 950.000 words, which are divided over 101 text samples, covering the period from 1700 until 1914 A.D.

Examples taken from these corpora are referred to in square brackets, by the filename, which is followed by a colon, and then by the line number, which is the number following the last period in the ID field of the *psd* files, since this number is a consecutively running line number that uniquely identifies each line in the file. There are two Present-day English corpora that have also been used on occasion:

- BNC: the British National Corpus (BNC, 2007). This corpus contains 100 million words and covers the time period from the 1980s until 1993.

- ICE-GB: the British component of the International Corpus of English (ICE-GB, 2011). It contains one million words of spoken and written British English from the 1990s.

Examples taken from these corpora are also referred to in square brackets, but they start with the abbreviation of the corpus name: “BNC” for the British National Corpus and “ICE-GB” for the British component of the ICE. These are then followed by the text and line number identifiers as defined by their individual corpora.

Part of the Chechen data used in chapter 11 is taken from a corpus of newspaper and journal texts that have initially been gathered under the auspices of the New Mexico State University, and are now freely downloadable (Zacharsky and Cowie, 2011). This corpus divides into two parts: a monolingual Chechen part that consists of 315,000 words, which are divided over 608 texts, and a parallel Chechen-English part that consists of 155,000 words, which are divided over 323 texts. Examples from these texts are referred to within square brackets, using a name that starts with “m” for the monolingual part and with “p” for the parallel part, followed by the name of the text (which mainly consists of numbers), a colon, and the line number of the example.

1.6 Outline of the study

The study that follows in this dissertation roughly divides into four parts:

- (14) a. Theory (chapters 2-5)
- b. Methodology (chapters 6-7)
- c. Results (chapters 8-12)
- d. Implications (chapter 13)

Both the theoretical and the methodological parts are interspersed with practical examples and shorter studies, but a thorough foundation is needed in terms of *what* (theory) and *how* (methodology), before the question of the interaction between focus and syntax can be addressed.

The theoretical part starts in chapter 2 by presenting a model of discourse processing, which is in line with the findings of the latest psycholinguistic research. The definition of focus in chapter 3, which builds on the discourse model, discerns three focus articulations based on focus domains: presentational focus (broad focus), the topic-comment articulation, and constituent focus (narrow focus).

Chapter 4 is a thorough practical exercise that, on the one hand, illustrates the theory of chapters 2 and 3 by a detailed investigation of an Old English and a late Modern English narrative text, and, on the other hand, offers the first insights into the changes that took place in the English strategies to convey presentational and constituent focus.

Chapter 5 is a key theoretical chapter in that it develops the hypothesis that “focus” can be determined on the basis of the syntax of a clause, the referential states of the clause’s components, and knowledge of the antecedents of the constituents in the clause.

With the hypothesis that focus can be determined by syntax and referential state in hand, the methodological chapters 6 and 7 describe how referential information of constituents can be added to the syntactically annotated English texts, and how the existing query tools can be extended to incorporate this information into powerful corpus searches.

Chapter 8 starts the first serious application that takes into account all the knowledge that has been gained so far, and applies it to find out how strategies to express presentational focus have changed in English. Chapter 9 continues this line of research, but now for constituent focus.

There is one prototypical candidate for the expression of constituent focus in Present-day English, and that is the *it*-cleft, and the remaining corpus research chapters concentrate on this construction, in order to address claims about the reasons for its appearance. Chapter 10 provides a definition of the *it*-cleft, and it explores its function as noted in the literature for English and other languages. This leads to the hypothesis that the *it*-cleft need not automatically be linked to a constituent focus function in all languages, a claim that is validated in chapter 11: Chechen has an *it*-cleft, but does not use it for constituent focus at all; it uses it for text-organization. Chapter 12 considers the development of the *it*-cleft in English, recognizing that Old English started out with this construction mainly as a text-organization device. It was the loss of the Old English “first position” (the pre-core slot) for conveying contrastive focus, that led to a stark increase of the *it*-cleft as a strategy to express constituent focus.

The last chapter, 13, discusses the implications of the findings in this study: the strategies used to express presentational and constituent focus have changed over time, syntax cannot be part of focus, nor can focus be part of syntax, and the referential status of a sentence’s constituents is an even more fundamental property influencing both syntax and focus.

¹ The results of this study indicate that, even though there are general rules or tendencies for focus, the expression of focus differs per language, so that it is imperative to know the focus rules of a language.

² There are several instances of type [N-Adj] such as: *Gode sylfum* ‘God.DAT self/same’ and *Gode ælmihtigum* ‘God.DAT almighty’. But instances such as these are potentially ambiguous, since the second word could be interpreted as a nominalization, so that they might represent a [N-N] word order.

³ Other languages may be less dependent on word order to signal constituent boundaries. Take for instance Russian, where words within a constituent do not only agree in case, but also in (grammatical) gender. This leads to an increased freedom in word order, so that occurrences of *dusha moja* and *moja dusha* ‘my soul’ can co-occur without difficulty in establishing constituent boundaries. The difference between the two word order possibilities of this example are probably more related to style and register: the former is a likely variant for poetry, while the latter is the more unmarked variant.

⁴ This book is, again, not the place to expand on the semantic, syntactic or pragmatic differences between [P NP] and [NP P]. I am assuming here that there is no syntactic difference: the adjacency of the P next to an NP is enough to indicate that the combination of the two is a PP. In fact, adjacency sometimes is not even required, witness the possibility of preposition stranding in, for example, this sentence: “The boy I spoke with yesterday lives around the corner”. The full PP is [_{PP} with [_{NP} the boy]], but the NP is positioned clause-initially, and we can “recognize” the full PP since we know that every P needs an NP object, and “the boy” is the best matching candidate.

⁵ Some English-Latin glossaries appear around 700-800 A.D., and then the “Pastoral Care” (Cura Pastoralis), written by king Alfred in Old English prose, appears, with a manuscript date that has been determined to be just before 900 A.D. (Ker, 1956).

⁶ Compare French *Ils suivirent Jean* (‘They followed John’) with *Ils le suivirent* (‘They followed him’). French seems to allow several clitic pronouns (representing established information) to precede the finite verb, as in: *Je le lui donne* ‘I give it to him’ (literally: I – it – to.him – give).

⁷ The computer program “CorpusStudio” does not accept “plain text”, so the algorithms need to be re-formulated into computer readable code. This matter is discussed more fully in chapter 7.

⁸ The significance of the transitions between periods of the four lines, in accordance with the two-tailed Fisher’s exact test, are as follows (see for details the appendix, section 14.3.1):

First constituent Obj/Adv/PP:	all transitions are significant
First constituent Obj:	only the transition from ME to eModE is significant
First constituent Adv:	all transitions are significant
First constituent PP:	all transitions are significant

⁹ Ball (1991) remarks: “With a few exceptions, this construction represents additional structure not in the source”.

¹⁰ The two-tailed Fisher exact test gives a *p*-value of 0.0774, which means that the association between the periods (eModE to LmodE) is “not quite statistically significant”.

Part II

Theory

In order to investigate the changes in the relation between focus and syntax in English, the notion of “focus” needs to be defined, and I would like to work towards such a definition by considering how the *mind* plays a role in understanding what we read or hear. This chapter defines a model of how the mind processes discourse, building on the models posited in the past (Johnson-Laird, 1983, Zwaan and Radvansky, 1998), and making use of the results of psycholinguistic findings.

The discourse processing model presented in this chapter forms the background for the treatment of the different focus articulations and focus types discussed in chapter 3, as well as the information state primitives posited in chapter 4.

2.1 Thinking about the mind

Even before advanced tools like MRI scans and EEG’s became available, researchers have been thinking about how the mind processes incoming information. Chafe (1976) was one of the first to define the notion of “given” in relation to the *mind*. He notes that given information is “that knowledge which the speaker assumes to be in the consciousness of the addressee at the time of the utterance”, where the term “consciousness” points to part of the hearer’s mind that is involved in processing the current discourse. Chafe emphasizes that something “new” is not necessarily completely new to an addressee, but is new in relation to what a hearer is currently “thinking about”. The “restaurant” in (15b), for example, already existed before it was mentioned.¹ The labels “new” and “given” do not, strictly speaking, pertain to noun phrases (e.g. “a nice little restaurant”), but to the “referents” they refer to (that is: to the restaurant itself). Chafe defines a “referent” as the “idea a noun expresses”, but we will make slightly different definitions in our model in section 2.3, so that Chafe’s “referent” becomes equal to what we will call a “mental entity”. Chafe also notes that the capacity of “consciousness” is limited, which causes “given” items to leave the addressee’s consciousness after some time, although reference to the item may still be “recovered”.

- (15) a. Once upon a time there was a boy named Jack.
b. It was evening when he saw a nice little restaurant.
c. As soon as he came in, the owner approached him.
d. The man stared at him, and Jack was desperately looking for words of wisdom.
e. Then the doorbell rung, and in came the mayor.
f. The door was wide open, and the sun shone straight at them.

Chafe (1987: 29) argues that humans unconsciously construct “schemata”, which are “clusters of interrelated expectations”. A schema can, for instance, be everything that is expected to take place in a restaurant. As soon as a restaurant (or one of the

actions of items closely connected with a restaurant) is encountered in a discourse, the “restaurant schema” is recovered from long-term memory, and the items connected with a restaurant (waiter, table, reservation, owner) become available to be filled in. If Chafe is right, a model of the mind should facilitate the creation and storage of such schemata.

The assumption of the existence of a large long-term memory (relatively slow) as well as a smaller-sized short-term memory (relatively fast) is based on experiments that show that only a small amount of information may be ‘active’ at any given time—presumably to increase processing speed. It is the most active information that is in the focus of our attention, and most readily available. Less active information may be less available or less accessible. A logical step would be to make a distinction between different types of the information on the basis of accessibility. A noteworthy example of a theory of “activation states” or “accessibility states” is Ariel (Ariel, 1994, Ariel, 1999), which has over fifteen noun phrase types, each of which corresponding to a “degree of accessibility associated with the mental entity in one’s memory” (see chapter 5.2.3, Figure 8). But if the goal of distinguishing accessibility levels is to instruct the addressee where the mental entity referred to can be found, then a two-way distinction should suffice, since a mental entity can be either in the short-term memory or in the long-term one.²

Connected to the discussion of accessibility is the issue of whether noun phrases refer to entities in the real world or not (the restaurant and persons in (15), for instance, are all fictional, so do not have real-world counterparts). Gívon (1982) remarks that referents do not necessarily have to exist in the “real” physical world, but they do so in a “universe of discourse”: a universe that is created, and in which the participants of a discourse are present. He also notes that speakers may refer to a non-existing entity, such as “book” in “I didn’t read any book today”. This is in line with the observations about different nonreferential noun phrase categories made by Hopper and Thompson (1984).

2.2 Insights from psycholinguistics

In the fields of psycholinguistics and neurolinguistics, the notion of “situation model” or “mental model” has gained general acceptance (Dijk and Kintsch, 1983, Johnson-Laird, 1983, Zwaan and Radvansky, 1998). This model is continually being refined by experimental data from brain research and psycholinguistics. Assume there is a discourse (written or spoken) and a person, which we will refer to as the “addressee”, who has the task of making sense of this text. Zwaan and Radvansky posit that an addressee transforms such a discourse dynamically into a model. This “situation model” consists of a set of participants as well as propositions involving these participants. Every piece of incoming information may contain elements that are divided into five independent dimensions: *time*, *space*, *causation*, *intentionality* and *protagonist*. It would be beyond the scope and focus of this current study to explain all of these dimensions, except for the one involving the participants and objects in a discourse, the dimension labelled as “protagonist”.

The existence of mental models (or situation models) seems to be confirmed by neurolinguistic experiments that have been conducted. Such experiments monitor the activation levels of parts of the brain in parallel with tasks that participants have to perform. Some of the results are summarized in (16).

- (16) a. The same brain areas that are activated when certain physical actions are involved, are also activated when one *reads* about these actions (Zwaan, 2004).
 b. Information that is “in” the situation described in a text is more active in the comprehender’s mind than information that is not in the situation (Zwaan, 2004).

Zwaan and Radvansky argue for the dynamic creation and updating of a model of the discourse we make in our mind, and they distinguish several kinds of memories. At a particular time t_n a person reads a clause or a sentence, transforming this in a “current model”, which he stores in short-term working memory (STWM). This current model is then combined with the “integrated model” from steps t_1 - t_{n-1} that is kept in long-term working memory (LTWM) in a step called “updating”. The complete model at the end of the process is stored in long-term memory (LTM).

Cognitive, psycholinguistic and neurolinguistic research make use of “event related brain potentials” (ERP), which are “an averaged measure of electroencephalogram (EEG) activity associated with particular critical events” (Hagoort and van Berkum, 2007). Peaks in this waveform with negative and positive potential at certain time intervals are indicators of strong activity and may coincide with mismatches, as illustrated in (17).

- (17) a. N400: a negative peak after 400 ms indicates a semantics-related effect.
 b. P600: a positive peak after 600 ms indicates a syntax-related effect.
 c. Nref: a sustained negative offset after 300 ms indicates a problem with resolving the correct reference.

EEG related experiments are, through the measurement of ERP waveforms, able to show several characteristics of the architecture and operation of the mind, such as the ones in (18), all of which also support the existence of mental models.

- (18) a. Isolated sentences with inanimate objects engaged in conversation (such as: *the girl comforted the clock*) cause an N400 effect, but no such effect occurs when the sentence is embedded in the context of a story involving inanimate participants (Nieuwland and van Berkum, 2006).
 b. An addressee tries to determine whether a noun phrase has a unique referent within 300 ms, as indicated by the Nref effect that occurs when an ambiguous noun phrase is introduced (van Berkum et al., 2007: 160).
 c. Addressees generally look for a possible participant in the immediate context, and if an appropriate one cannot be found (as for instance in *Anna shot at Linda as he jumped over the fence*), then a P600 effect occurs (van Berkum et al., 2007: 162).

The observation in (18a) fits a model of the brain where the addressee builds a situation model involving participants that come with certain prototypical properties (supplied by default from long-term memory?), and that can receive non-standard properties (e.g. that an inanimate object such as a clock can be sad). The observations in (18b-c) indicate that reference resolution takes place very quickly, and is initially concerned with the small set of available referents within the situation model built so far.

Van Berkum et al (2007) found support for the idea of continuous processing and updating of a mental model. They established that listeners initially detect a syntactic error in a sentence like: “David praised Linda because *he*...” The hearer expects discourse continuation with the female pronoun *she*, because of the nature of the verb “praise” (compare the opposite effect of the verb “apologize”). They also established that participants who leave the scene, within the world evoked by the discourse, no longer serve as candidates for antecedents in coreference resolution, which further supports the idea of a mental model. Martin et al (2012) found that a structurally impossible antecedent candidate intervening between a pronoun and its actual antecedent will be considered first, and is only rejected after error detection has taken place.

Such investigations demonstrate that there are sufficient grounds for assuming some kind of mental model that is continuously being updated by a reader or hearer.

2.3 A mental model for discourse processing

This study assumes that for each new discourse act (e.g. a text that is being read or a piece of oral communication that is being listened to) *a new* mental model is being created.³ This model dynamically creates and updates *mental entities*, which are representations in the mind of objects or persons.

2.3.1 Mental entity

The *mental entities* constitute the basic components of the mental model we will be working with, and the question is how these mental entities relate to real-world concepts (that is, physical entities like for instance a person named “Jack”) and imaginary concepts (such as fairy tale figures) on the one hand and linguistic expression (the noun phrases) in the discourse on the other hand.

Mental entities are a kind of in-betweens, leading a solitary and confined live within the mind of comprehenders. A linguistic expression first of all refers to a mental entity, and it is only mental entities that then refer either to real-world concepts or to imaginary ones.

(19) Mental entity

Given a constituent XP, the *mental entity* of XP, written as $\text{MEnt}(XP, i)$, is the entity in Situation Model(i), which the addressee builds of Discourse(i), and which is uniquely associated to the constituent XP.

The definition of mental entities in (19) places it specifically in a particular Situation Model indexed with i , which the addressee is creating as a result of a particular

Discourse indexed with i . This discourse may, as explained before, refer to a particular oral comprehension act (watching a movie, listening to the radio, listening to someone speaking to you), or the reading of a letter, a book, a chapter, or any other document.

The *mental entities* of a particular discourse roughly equate to one's mental pictures of the participants in a discourse. One mental entity uniquely matches with one real-world or imaginary-world concept.

2.3.2 Mental model

With the concept of *mental entities* in place, the discourse processing model adopted in this dissertation can now be defined, and Figure 2 serves to help with this definition.

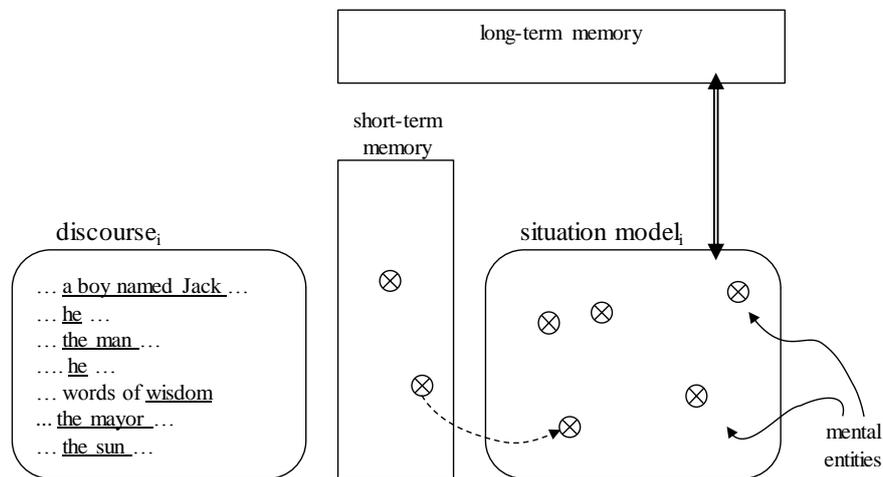


Figure 2 Discourse and situation model

If we start telling a story that starts like (15a): “Once upon a time there was a boy named Jack”, then the referring expression *a boy named Jack* becomes a mental entity within the current situation model. Making use of mental entities has the advantage that there is no prerequisite for “referents”, which are the objects or persons these mental entities refer to, to actually exist in the real world.

The discourse processing model above assumes that an addressee sequentially parses a particular Discourse _{i} , seeking mental entities for each noun phrase s/he encounters.⁴ The first challenge the addressee is faced with is the decision whether to create a new mental entity or use an existing one. The model used in this study assumes that the addressee does not decide this on the basis of the *form* of the noun phrase – if this were the case, it would mean that the decision is language-dependent, since the inventory of NP types differs from language to language. Instead, we opt for the following scenario:

- (20)
- a. Every occurrence of a noun phrase leads to the creation of a mental entity in short-term memory.
 - b. A process of reference resolution determines whether this mental entity matches with an already existing mental entity in the “situation model” (in a memory area between the short and long term memory).
 - c. If there is a match, then the features (or characteristics) of the noun phrase are added to the existing mental entity within the situation model.
 - d. If there is no match, the mental entity is copied to the situation model.
 - e. The entity in short-term memory is now deleted. (Alternatively a small cache of pointers to the n most recently accessed mental entities is kept here.)

The second decision has to do with the nature and operation of inferences. Suppose we encounter the following story: “James sat down in his car and turned on *the lights*.” The noun phrase *the lights* is new to the addressee’s situation model, because it has never been mentioned before. But it is not entirely new, because it piggy-backs on the situation evoked by the noun phrase *his car*. What is the psycholinguistic reality: does the mention of *his car* immediately trigger a reference to a default model of “car” stored somewhere in long-term memory, which comes with obvious “slots” for “wheels”, “windows”, “ignition”, “lights” etc?⁵ This would be a kind of “proactive” understanding of matters: as soon as we encounter something (such as a “car”) we know more about, we fetch its “model”, which comes with certain fillable “slots”.

The alternative would be a “retroactive” understanding: it is only when “the lights” are met that we start looking for an antecedent, and when we don’t find one, we start expanding mental entities present in the current situation model with the kind of “model” information stored away in long-term memory. We start this process by evaluating the mental entities in our situation model starting with the most salient one (that is: the one that has been referred to most recently). The mental processes involved in proactive versus retroactive inferences seem to be quite different, which is why we *will* make a difference between them later on. (Chapter 5 will adopt the proactive inferences as having a separate referential state, whereas the retroactive inferences group together with completely new entities in the mental model.)

Another question regarding situation models is that of lifetime. Do participants always stay within the situation model, or can they “leave”? Psycholinguistic experiments show that addressee’s, in a sense, build a model of the physical situation in their mind, and if something occurs that is in conflict with the model built so far, this leads to increased processing difficulties (see Zwaan and Radvansky, 1998 and references therein). This could entail that when a participant “leaves the scene” in a discourse, he is no longer present within the addressee’s situation model. Alternatively one could argue that his presence within the situation model is “tagged” with a time label (indicating “enter stage” and “exit stage” times) or that he gets an “availability status” assigned. The mental entity then still belongs

to the situation model, but in a more restricted way, making him inaccessible for certain references.

2.4 Conclusions

As soon as a person starts reading a text or starts listening to someone telling a story, a situation model is being formed in that person's mind. The items and persons referred to in the text or story result in the creation of mental entities within the addressee's mind, and this is irrespective of their actual real-world existence. Processing of noun phrases encountered in a text is such, that an addressee will first seek to connect them with entities that are already established in the discourse model, slots created by schemata, entities available in the situation or world knowledge items stored in one's long-term memory.

The mental entities that have been defined in this chapter will form the building blocks for the various focus articulations and focus types discussed in chapter 3, as well as the information state primitives posited in chapter 4.

¹ The sample story is mine—it was not used by Chafe to illustrate his ideas.

² Information can be “further away” or “more to the surface” in one's memory, which would imply a “stack” model of the mind (first-in-last-out). It is unclear whether such a stack model would only apply to short-term memory, or also to long-term memory. But in both cases the question remains how and why the signals indicating various degrees of accessibility that are supposedly conveyed by the noun phrase types could help one to recover something from memory.

³ One may alternatively assume that a new situation-indexical is created for each new discourse act, and that participants within this discourse act receive such a situation-indexical. We have chosen to use the notion of a “situation model” with its own set of participants for explanatory purposes. The end result should not differ.

⁴ We restrict ourselves to *noun phrases* here, but there are obviously more categories for which mental entities will be sought, such as location and time adverbials.

⁵ The idea that we have some kind of default models in our mind (in long-term memory) for things like “car”, “robin” etc has been argued for repeatedly, and has been the subject of several studies (see for instance: Brewer and Treyens, 1981, Garnham, 2001: chapter 6, Johnson-Laird, 1983, Rumelhart et al., 1986).

Since the aim of this study, as stated in (11), is to understand the interaction between syntax and focus, we need to have a clear definition of the latter notion, and this chapter aims to do so by referring to the discourse processing model discussed in the previous chapter.

Focus in spoken English is often connected to particular intonation patterns (Gussenhoven, 2007). But since there are languages for which intonation is not a means to achieve focus (e.g. Chechen, chapter 11), and since we will be looking at the historical development of the English language, which comes to us in the form of documents, we will only pay attention to non-prosodic means of realizing focus.

Against the background of my own definition of “focus” (3.1), we recognize three sentence types, or “focus articulations”, which differ as to their focus domain (3.2). These focus articulations interact with other (non-pragmatic) factors that contribute to the word order in a sentence (3.3). Recognizing these factors helps us establish when focus is pragmatically marked (3.4). The review of focus in this chapter ends with a discussion on the relation between focus and “newness” (3.5).

3.1 Defining focus

The term “focus” has been given a number of different interpretations over the years: context independent (versus context dependent) information, new (versus background, given or presupposed) information, contrastive (versus non-contrastive) information; see Kruijff-Korbayová and Steedman (2003), especially their Figure 1, for a semantic map and references. The practical definition of focus in (21) couches focus in the notion of the mental model, as introduced in chapter 2. An addressee is continuously making and updating a mental model of the information received by reading or listening.¹

(21) *Definition of focus*

Focus is the part of the sentence that should be understood as most highlighted or salient by the addressee, because it is new with respect to the current mental model, or contrasts with presupposed information, or is unpredictable, non-recoverable or of high communicative interest.

The definition starts by saying that focus is part of a sentence, which means that focus is encoded in the linguistic form of a sentence. Strictly speaking, focus can be in the form of a constituent, an event or a relation between constituents (Lambrecht, 1994). The focused part of a sentence distinguishes itself from the rest by receiving some kind of emphasis or highlighting. This is not necessarily intonational. In fact, (22) groups several different linguistic means available to focus a constituent (for some of these see: Dooley and Levinsohn, 2001, Féry and Krifka, 2008).

- (22) a. **Intonation.**
A particular configuration of tones (for example a high tone or a low tone) may be associated with the beginning or the end of a focus domain, which is the part of the sentence that is most informative (Gussenhoven, 2007).
- b. **Morphology.**
Some languages use a morpheme to indicate a particular kind of focus (see below on focus types). The morpheme may be a suffix (the *-i* suffix attached after the perfective suffix *-go* on the verb in Chadic, for instance, signals VP focus (Hartmann and Zimmermann, 2004)). Morphemes that are used for focus may piggyback on a morpheme with a different function (Weber (as quoted by van Valin, 2005: 73) reports on the evidential marker *-shi* in Huallaga Quechua doing double duty as focus marker).
- c. **Particles.**
Particles are small function words, and some languages use them for focus, such as the particle *ga* in Japanese (see Kuno (1973), and Lambrecht's (1994) discussion) and the word *only* in English (Rooth, 1992); Sornicola (2006) identifies particle focus as one of the main strategies in European languages.
- d. **Word order.**
The focused constituent may be highlighted syntactically by moving it to a particular part of the sentence. In African Bantu languages, which are SVO, the focused constituent often occurs immediately after the finite verb (as for instance for Zulu: Cheng and Downing, 2009). In SOV languages like Turkic and Chechen, the focus position is immediately preceding the finite verb (Komen, 2007b).
- e. **Special constructions.**
Another strategy that is sometimes used in languages to emphasize one particular constituent is the use of special constructions. Examples of these constructions are: *it*-cleft, *wh*-cleft, left dislocation, right dislocation, particle preposing.
- f. **Ellipsis.**
A constituent can be focused by leaving out the elements around it that are *not* focused (see for instance Winkler, 2005).²

There is, then, quite a spectrum of morphosyntactic means to mark focus, but should we also distinguish different *types* of focus? Opinions are divided on this question. One could argue for a unification under one umbrella (Krifka, 2007), recognizing a common feature of the different kinds of focus: any highlighting or focus implies the presence of *alternatives*. One could, on the other hand, divide focus up into subtypes, according to the *functions* these fulfil (Gussenhoven, 2007).³ And there are probably many more ways to divide focus into categories. The approach taken in this dissertation is based on Lambrecht (1994) and on Levinsohn (2009), who, in turn, base their work on others (e.g. Drubig, 2000, Gundel, 1988, Jacobs, 2001, van Valin, 1999). It involves the following steps:

- (23) *Focus detection approach*
- a. Divide clauses into one of three “focus articulations”
 - b. Recognize linguistic phenomena interacting with these focus articulations
 - c. Recognize the difference between marked and unmarked forms

By recognizing clause types based on the focus domain they contain (section 3.2), we incorporate universal differences that are predictable. Interactions with focus articulations (section 3.3) result in form differences, which are not necessarily related to differences in focus meaning, which is why they need to be taken into account too. The marked versus unmarked distinction, finally, allows us to differentiate between what is default (pragmatically unmarked), and what not (marked). We may expect differences to occur in both of these categories.

3.2 Focus articulations

There are several ways to look at focus, and one of them is based on the size of the focus *domain*—the part of the sentence or clause that is being highlighted. This idea is touched upon by scholars such as Prince (1981) and Gundel (1974), but it is Lambrecht (1994) who combines crucial parts of the research into a framework for dealing with information structure. Lambrecht argues that cross-linguistically, languages make use of three kinds of focus domains: (a) the whole clause, (b) the predicate, or (c) one constituent only. This universal distinction serves as a basis for Lambrecht to posit three corresponding focus “structures” (which are also referred to as “focus articulations”): (a) sentence focus, (b) predicate focus, and (c) argument focus.

In this dissertation, the three focus articulations that have been recognized by Lambrecht will also be used, but with slightly different labels that should fit them better. The terms used for the different focus articulations in this dissertation are given in (24).

- (24) *Focus articulations*
- a. Topic-comment (also known as: *predicate focus*)
 - b. Constituent focus (also known as: *argument focus, focus-background*)
 - c. Thetic sentence (also known as: *sentence focus*)

Lambrecht’s term “sentence focus” is slightly misleading, since it suggests that the focus domain contains the *whole* clause, which is not entirely true. We will use the term “thetic sentence” or “thetic clause” to refer to clauses where the focus domain includes the subject and the predicate, which is in line with Sasse (1987, 2006) and Bailey (2009). Thetic clauses usually have a temporal or locational point of departure, which grounds the newly presented information in the established information. This point of departure links to the established information, and so is not new, and is not part of the focus domain.

The *predicate focus* articulation is often referred to as a “topic-comment” structure, since its main function is to provide (new) information on an established topic. We will adopt that name, since it is closer to the function of this focus articulation.

The articulation called “argument focus” by Lambrecht is referred to as the “focus-background” division by Prince (1981). Prince’s term is understandable, since this focus articulation highlights one constituent, with the result that the rest of the clause serves as presupposition or background. Since this articulation restricts the focus domain to one constituent, whether it is a verbal argument or an adjunct, we will use the term “constituent focus”, in line with for instance Dooley and Levinsohn (2001).

3.2.1 Topic-comment

The “**topic-comment**” articulation in (24a) is the default one in a narrative, since it is used to make a comment (that is: introduce a new development) about an already established referent. The referent is prototypically represented by the grammatical subject, and the comment is in the predicate. Since the comment is the “new” information (where we take “new” in the sense of adding information about the topic to the mental model of the addressee), the focus domain is the predicate. The narrative in (25) serves as an example for the topic-comment articulation.

- (25) a. **My father** killed the captain of the privateer, [fayrer-1900:23-33]
 b. and **0** had, with other wounds, his right arm shattered by a bullet.
 c. For his services on this occasion **he** was promoted to the rank of lieutenant in 1808.
 d. When **0** lying unconscious from fever in Malta Hospital, some one hung a gold cross and chain round his neck with an inscription:
 e. **he** never knew the donor.
 f. **He** recovered after a long illness, with his right arm badly crippled,
 g. and **0** remained for some time on half pay.
 h. **He** served afterwards,
 i. and **0** was first lieutenant of the Orpheus, Captain Hugh Pigott, during the American war.
 j. Subsequently **he** obtained permission to command an Indiaman,
 k. and for many years **0** sailed in ships of that class, the Lady Flora being the last.

The narration starts in (25a) by stating the topic in a lexical NP *my father*, anchoring it to the main participant of this autobiography, the “I” person. Most of the other sentences in (25b-k) have a topic-comment articulation, which is achieved by keeping a reference to the topic *my father* as the grammatical subject (realized either as 3rd person pronoun *he* or as a zero).⁴ The “comment” part, containing the information about the topical referent that the author wants the reader to have, is the VP in all these cases. (There is more that can be said about the information structure of (25), in particular the role of the clause-initial adjuncts in lines (25c,j,k), but that will come in section 3.3.2, where we will talk about “points of departure”.)

3.2.2 Constituent focus

In a clause with “**constituent focus**” articulation, as in (24b), the focus domain is that of the constituent that is being highlighted. Examples of constituent focus are given in (26).

- (26) a. I was especially interested in Mr Wharton Jones' lectures on Physiology.
At these lectures T. H. Huxley sat by my side, and **he** it was who first directed my attention to their great interest and importance.
[fayrer-1900:563-4]
- b. We had a little difference of opinion about the base of the skull, Guthrie listening with interest. It appeared the examiner meant the inside, while I was describing the **outside**. [fayrer-1900:597-8]
- c. **Not One Gleam of Comfort** will I afford him, I'll assure you Lucy.
[Stevens-1745:556]

In the *it*-cleft construction of (26a), the constituent *he* contrasts with all members of the set of people fulfilling the condition that they “first directed my attention to their (=lectures) great interest and importance”. The constituent *he* is syntactically singled out through the use of an equative clause (see section 3.2.2.1), and provides the variable of the open proposition in the relative clause *who was the first ... importance*.⁵ Another case of contrastive focus is in (26b), where *outside* is set against the already established *inside*. Example (26c) has focus on the direct object of the verb *afford*, the constituent *not one gleam of comfort*. The focus here is not necessarily one of contrast (that is to say: a little bit of comfort as opposed to some more comfort), but it is emphatic prominence (unmarked *no* versus emphatic *not one*). The focus effect is the result of a combination of two mechanisms: negation and word order. (The particular word order that is used here is the subject-finite-verb inversion that has been discussed in section 1.2.2 of the introduction.)

Constituent focus overrides a topic-comment structure. The clause in (26b), for example, could easily be understood as a comment *was describing the outside* that goes with the topic *I*. Since there is explicit contrast between *outside* and *inside*, however, the focus domain really restricts itself to one constituent only (the contrastive one), which is why it has the *constituent focus* articulation.

Some clauses with a constituent-focus articulation can only be recognized from the context in which they occur, since they do not distinguish themselves syntactically from topic-comment clauses. There are a few types of constituent-focus, however, that can be recognized relatively easy, and these will now be treated.

3.2.2.1 Equative clauses

The first recognizable type is that of the copula clause with an NP subject and an NP complement: $NP_{Sbj} + be + NP_{Compl}$.⁶ Equative clauses can, in general, be specificational or predicational (Akmajian, 1979, Declerck, 1984, Patten, 2010).⁷ If the equative clause is specificational, the complement, even though it is syntactically an NP, provides a description of a set that can only contain one member, and the

subject NP states who the member of this set is. Some examples of equative constructions are in (27).

- (27) a. The murderer is **John**.
 b. What I wanted to tell you is **this**.
 c. **Mary** is the one who borrowed my computer.

The set that can only contain one member in (27a) consists of the one person who did the killing, and the unique member of this set is provided by the subject “John”. The set in (27b), which is a *wh*-cleft construction, is the one thing that the speaker had in mind to tell the hearer, and this one thing is neatly summarized as the subject “this”. The set in (27c), which is an example of a reversed *wh*-cleft construction, consists of the one person who borrowed a particular computer, and the unique member of this set is “Mary”.

Equative clauses of the type illustrated in (27a-c) *can* have constituent focus force whether they are *wh*-cleft constructions or not, but there are two additional requirements that need to be met for them to be of the constituent-focus type, and the first one has to do with the relative newness of the subject and the complement. True constituent-focused equative clauses have an NP complement that is relatively newer than the NP subject. As soon as the complement is newer than the subject, equative clauses are instances of the default (unmarked) articulation, the topic-comment one, as in (28).

- (28) a. There is one thing I want her to know. She is the sunshine in my life!
 b. Do you know Harry? He is my brother-in-law.

In (28a) the equative clause subject “sh” has already been established in the preceding clause, whereas the identity of “one thing” is not yet disclosed. The result is a topic-comment clause with “you” as topic, to which the addressee in his mental model adds the characteristic of “[he says she is] the sunshine of his life” (which is the value of the variable “one thing” introduced in the first clause). The person named “Harry” in (28b) enters the addressee’s mental model in the first sentence, so that the second sentence, the equative one, does not serve to answer the question “who is my brother-in-law”, but provides a characteristic of “Harry”, which is to be added to the mental representation of him in the addressee’s mental model.

The second requirement for equative clauses to actually have a constituent focus articulation is that it needs to be specificational and not predicational. A few examples of predicational copula clauses are in (29):

- (29) a. The **runner** was a beautiful lady.
 b. The **runner** was quite nice.

The examples in (29a-b) are not specificational but predicational: the complements add descriptive characteristics to the “runner”, whose identity must have already been established in the previous context.⁸ No separate mental entity is created for “a beautiful lady” – the NP complement only serves to add the characteristics “beautiful” and “lady” to the mental entity of “runner” that is available in the mental

model of the addressee. The equative clause in (29a) with the structure $NP_{Sbj} + be + NP_{Compl}$ has the same force as the one in (29b) that has the structure $NP_{Sbj} + be + Adj$, where it is syntactically unambiguous, in the sense that the complement only provides a characteristic for the subject. Predicational copula clauses such as (29a-b) have a topic-comment articulation.

Copula clauses in general, and equative (NP-be-NP) clauses in particular, are able to have a wide range of different meanings, and how their type (specificational, predicational and so on) derives from the syntax and semantics of their components still is a matter of research (Cann, 2003, Mikkelsen, 2005). The examples above and the referential state primitives introduced in chapter 5 suggest that it may be possible to map the different kinds of copular constructions to the different focus articulations, provided their referential states are available. For now it is good to know that a well-defined subset of copula clauses map straightforwardly onto the constituent focus articulation; section 5.5.3 offers a more detailed account of focus domains in copula clauses.

3.2.2.2 *Explicit contrastive focus*

When a clause contains a contrastively focused noun phrase, it has a constituent-focus articulation. We can detect contrastive focus if the contrast is explicit.⁹ The examples in (30) represent different ways in which explicit contrastive focus can be expressed.

- (30) a. But there is rich compensation in Barbara Jefford's magnificent Volumnia: why has this superb actress been given **only two roles** by the RSC in 30 years? [BNC, A8s:23]
- b. The sounds came nearer; dragging, crawling sounds, as if **not one but several creatures** were struggling across the floor. [BNC, G1L:2192]
- c. And many more believed because of his word; and they said to the woman, Now we believe, **not because of thy speaking**: for we have heard for ourselves, and know that this is indeed the Saviour of the world.
[erv-new-1881:302-5]
- d. Generous Spirits will always have a Concern for the Benefit and Credit of their Country: And how far the Honour and Interest of Great Britain are concern'd in the Cultivating of Our Language, I presume not to say; only, That a neighbouring Nation has taken Care of **Theirs**, and found their Accounts in't. [brightland-1711:22-3]

First, explicit contrast is possible within one constituent. The use of a focus particle like “but” or “only”, as in (30a), is a clear signal of contrastive focus, because “only” explicitly selects one value from a larger set of values. The use of two alternatives contrasted by “but” within one NP constituent, as in (30b), is another form of explicit contrast. In some cases the use of the negator “not”, as in (30c), is an explicit negation of one alternative, which implies the existence of other alternatives to which the one alternative is contrasted.

Second, explicit contrast is possible with reference to a constituent in the preceding context, as for example in (30d). The pronoun “theirs”, which refers to “their language”, contrasts explicitly with “our language” in the previous sentence.

Third, explicit contrast can also occur by contrasting one constituent with another one that is located in the *following* context. Such could be argued to be the case in (30d) too, where “our language” explicitly contrasts with “theirs” in the following context.¹⁰

3.2.2.3 *Emphatic prominence*

A third kind of constituent-focus clauses that can be detected are those containing a constituent with “emphatic prominence”. This term is used by Callows (1974) for constituents that are marked so as “to express strong feelings about an item or to indicate that what follows is unexpected”. The sentences in (31) provide different examples of emphatic prominence.¹¹

- (31) a. We were **right in the middle** of an arc of gunfire and there were search lights into the sky trying to pick out aircraft. (BBC, 2009)
- b. Other food colourings, **particularly** the synthetic ones, have been known to cause allergic dermatitis, mainly in food workers exposed to large amounts. [BNC BMI:617]
- c. **The same** Honest John who once described his Treaty negotiation as Game, Set and Match and who now prays at night the French *will* reject it so all the blame can be piled elsewhere. [BNC, CH1:2146]
- d. **One and the same** practice may be performed by a nurturant or a hostile mother, may occur within an easygoing or a rigidly authoritarian home, or may take place against a background of love or of hate. [BNC, EEK:768]
- e. All that will, of course, now change, with the government’s decision to allow the supermarket giants in. But the move is **not without** opposition. (BBC, 2011)
- f. **We ourselves** are to some extent part of this problem, but we do at least live and work in the village full time. [BNC, A7D:2291]
- g. He has never sought to evade his responsibility for the appalling consequences of his errors...**not one word of excuse** came from Mr Hemingway. [BNC, A7W:766]

Emphatic prominence is marked by using adverbs like *right*, as in (31a), which do not really add to the referential meaning of the constituent, but do make it more prominent. An apposition such as the one in (31b) is another method that effectively yields highlighting of a constituent. The lexeme *the same* in (31c) and the expression *one and the same* in (31d) highlight the noun phrase they are part of. Positive negations like *not without*, as in (31e), are another way to express emphasis on a constituent (see section 9.4). A language may also have a set of emphatic pronouns, or, as in (31f) for English, use pronouns that otherwise have the function of a bound anaphor for the purpose of highlighting when these pronouns appear in an undominated position. Another highlighting construction is (31g), where the

unmarked negation *no* receives emphatic prominence by expanding it to *not one word*.

Levinsohn (2009) reports that phonological features such as pitch, heavy stress and vowel lengthening can also be used to convey emphatic prominence (see also Selting, 1994 for prosodic features co-occurring with emphatic speech style). But since this dissertation focuses on *written* communication, prosodic means are not taken into consideration here.

In sum, we have seen several types of constituent focus that involve explicit linguistic means (word order, adverbs, focus particles, negation etc). We will look back to these indicators of constituent focus in chapter 9, where the experiments are described which are aimed to find the changes in word orders used for constituent focus in English.

3.2.3 Thetic sentences

So called “**thetic sentences**” have a focus domain that includes the subject and the predicate. They are used to introduce a new entity (through the subject) or event (the predicate + subject) into the discourse. Such sentences can occur at the beginning of a narration, as in example (32a), or at a point in the narration where a new entity enters the scene, as in examples (32b,c).

- (32) a. It is evident by Experience, that **there are several Arts and Sciences**, which can not be learn'd in any great Perfection, without the Knowledge of Latin, or Greek, or other Antient Languages. [anon-1711:5]
- b. **There was a very picturesque little watering-place** where the boats used to fill their casks. Near this on one occasion, tempted by the beautiful water and the warm air, I stripped and plunged in. [fayrer-1900:234-6]
- c. The jib was pressing her so heavily that I determined to take it off and work under the reefed mainsail alone. **There were two men in the boat**, who had on pea-jackets and heavy sea-boots. I told them what I was going to do, and ordered one man to roll the jib round the stay, a common practice in a Bermudian boat. [fayrer-1900:366-9]

The main goal of thetic sentences is to introduce a new participant or event into the discourse, and the sentence in (32a) is an example of how this is achieved at the very beginning of a story.¹² This sentence is the first one from an essay on education, introducing the subject *several arts and sciences*, which is taken up in subsequent sentences.

Example (32b) is at a point in an autobiography where a new episode starts. The author introduces the topical element of this episode, which is the *watering-place*, anchoring it in information that has, to some degree, already been established, by referring to *the boats* (these have been mentioned in the prior discourse).

Another example of a thetic sentence is (32c), which introduces *two men* as new participants, which are subsequently referred to as *them*. This example illustrates how thetic sentences do not necessarily have to consist of constituents that are *all* completely *new* to the discourse scene (or to the mental model in the addressee's mind). The *boat* is already known, for instance. In fact, any thetic sentence is uttered

against the background of a situation, time and/or location, which is either not specified but understood, or overtly specified. Scene settings or “points of departure”, as we will call them, can occur with any of the three focus articulation types, and will be discussed more fully in section 3.3.2. Important for the discussion onthetic sentences is that the stage setting elements are not to be confused with topics or foci. They are *outside* of the focus domain.

One of the determining characteristics ofthetic sentences is that they contain a subject that is usually new to the mental model of the addressee (the notion of “new” will be explored further in chapter 5).¹³ The newness of the subject is not the only factor; there are at least two more considerations to be made, since the main feature ofthetic sentences is that their focus domain spans the whole core of a clause. The first consideration is that the predicate (with its internal arguments) must be (relatively) new too, and the second point is that constituent focus overridesthetic focus. So when the subject provides the value for an open proposition that has just been raised, there is constituent focus, and there can be nothetic articulation.

- (33) a. “Who would want to listen to you?”
 b. “**An educated man** will read my books!”

An illustration is hard to find, but (33) should serve as an example for both principles that have just been mentioned. The subject of (33b), *an educated man*, is completely new, but it provides the value for the variable raised in (33a) “an *x* who listens to you”, which means that (33b) has constituent focus on the subject and is not of athetic articulation. The second motivation for not recognizingthetic articulation here is that the focus domain does not seem to include the predicate, since the event *read* in (33b) really can be inferred from *listen* in (33a), and *my books* in (33b) is anchored to the speaker, so not completely new too.

In sum, we can say that athetic articulation can only be there if we have evidence that the predicate is part of the focus domain, that the subject is new, and that the subject is not providing the value for a variable that has just been raised.

3.2.4 Focus domain generalizations

The three focus articulations or focus structures discussed above are based on a threefold distinction in focus domain size: one constituent (constituent focus), the VP (topic-comment articulation) or the subject with the predicate (thetic sentences). Role and reference grammar generalizes from a fixed number of three focus domains to what it calls the “Actual focus domain” (van Valin, 2005). VanValin argues for less restriction on the size of the focus domain, giving an example where he distinguishes two constituents in one sentence that, as he says, each have one focus domain (34).

- (34) a. Bill gave [_{DO} the BOOK] [_{PP} to MARY]

VanValin states that *the book* has contrastive focus, whereas *to Mary* has complete focus. Even though one may disagree on the particular names of the focus types here (this could be a corrective answer, containing two contrastive focus domains, in

response to an enquiry like: “Did Bill give the chapter to John?”), it is clear that there may be more than one focus domain in any one clause. VanValin’s approach does not deny the existence of the three focus articulation types as given in (24), but it says that the number of focus articulations may not be fixed to “3”: there may be other domains of highlighting that, for example, involve more constituents.

3.3 Interactions with focus articulations

There are several phenomena that work in parallel with the focus articulations. First of all there is the “Principle of Natural Information Flow”, which states that established information tends to precede unestablished information. Then, while every clause has one of the three focus articulations given in (24), it may also contain a “point of departure” or “frame setter”, as has been hinted on in the examples above. Topic-comment clauses may sometimes have one particular constituent that is highlighted more than others, even though the focus domain spans the whole predicate. We will refer to this as the “dominant focal element”. The form of a clause (as visible in as word order, suffixes, overencoding or the use of particles) may furthermore be influenced by discourse-related constraints, which serve to divide the text into smaller units. Some of these may signal local cohesion, while others may signal smaller or larger episode boundaries.

What all of these phenomena have in common, is that they can occur in parallel with one or more of the introduced focus articulation types. Since the overall purpose of this dissertation is to further an understanding of how the expression of *focus* has changed in English, we need to clearly discern the interactional phenomena mentioned above, so that we can see if, in addition to any changes in the way the focus articulations are expressed, the expression of these phenomena has also changed.

3.3.1 The principle of natural information flow

Many languages in the world tend to order non-verbal constituents according to the “Principle of natural information flow” (Comrie, 1989, Firbas, 1964, Kaiser and Trueswell, 2004).¹⁴ This principle basically says that “established” information precedes non-established or less established information. Whether some piece of information, such as a participant or a location or time, is established or not depends on whether it has been mentioned in the discourse before or perhaps is evident from the extralinguistic situation surrounding the communication or can be taken for granted as shared knowledge between the speaker and the addressee.

Constituents (mainly arguments and adjuncts) will only satisfy the Principle of Natural Information Flow if the *syntax* of a language allows them to. Languages which mark noun phrases for their role in the clause morphologically, such as Russian, Turkic etc, will obviously allow more reordering of constituents, and the Principle of Natural Information Flow plays an important role in this reordering.

Present-day English is usually regarded as having a rigid SVO word order, but there are still several situations where the syntax allows for alternatives, such as the relative position of the direct object and the indirect one, as in (35a-b).

- (35) a. John gave the knife to **a boy**.
 b. John gave the boy **a knife**.

In the example in (35a), the indirect object *a boy* is less established than the direct object *the knife*, as can be seen from the articles. The example in (35b) has it the opposite way: the indirect object *the boy* is more established than the direct object *a knife*. The principle of natural information flow is operating in both examples: the more established information precedes the less established information.¹⁵

The principle of natural information flow is also at work in the presentational construction of (36a), where the least established information is *a handsome prince*, and comes completely clause-finally.

- (36) a. Once upon a time there was **a handsome prince**.
 b. George and I were to be victims, I was to be taken to *the top floor* and George to the third floor up. *The house* had already been damaged. I was to have broken my leg attempting to get from bed to the top of the stairs. I was duly bandaged by the first aid folk, and then placed in position. I waited for my rescuer. I did get a shock. Until the end of the war so very few folk had beards, and then only short ones nicely trimmed, but into the room came **a most handsome young man** with a black fuzz of over eight inches. [BNC-UK B2E:1213]

Another construction where the principle of natural information flow can be seen to work is the locative inversion in (36b). The *room* of the PP *into the room* can be inferred from *the house* and *the top floor*, which is already established information in the preceding context. Like (36a), the example in (36b) also has athetic focus articulation, introducing a new participant in the discourse scene.

3.3.2 Point of departure

Clauses with any of the three articulation types can optionally have a “point of departure” (Beneš, 1962, Levinsohn, 2000).¹⁶ This is a constituent, a phrase or subordinate clause that indicates an important change in the course of the discourse in terms of location, time, situation or referential point of view. The formal definition of a point of departure in (37) derives from Levinsohn (2000: 8).

- (37) *Point of departure*
 A point of departure is a constituent fulfilling the following conditions:
 i) It is placed at the beginning of a clause or sentence;
 ii) It expresses a change in the point of view in the discourse;
 iii) It anchors to something that is accessible to the addressee (either from the preceding linguistic context or through shared knowledge)

The “point of view” in a discourse can be compared with the position of a camera. Just as a camera can capture one and the same situation or event from a different point of view, so too can an author describe a situation or event from a different angle. A sentence like “John walked home”, for instance, can be looked at from a temporal point of view (such as: “At five o’clock, John walked home”), from a

locational point of view (“From the drug store, John walked home”) or from another circumstantial point of view (for instance: “With tears in his eyes, John walked home”).

Crucial for a point of departure is that it not only establishes a particular time or location, but that it does so in relation to the established context—the context that is already available in or for the current mental model (see chapter 2). This context can be inter-textual, in which case the point of departure anchors to something specific in the preceding text or can be inferred from some person, thing or event in the preceding context. It can also be extra-textual, anchoring in global time or known facts within the world. We will have a look at some points of departure from an existing text.

- (38) a. So I must leave here the fruitless exclaiming at my self, and go on with my Voyage. **From the Brasils**, we made directly away over the Atlantick Sea, to the Cape de bon Esperance, or as we call it, The Cape of Good Hope. [defoe-1719:162-3]
- b. The People, who by the Way are very numerous, came thronging about us, and stood gazing at us **at a Distance**; but **as we had traded freely with them, and had been kindly used**, we thought our selves in no Danger. [defoe-1719:187-9]
- c. **But when we saw the People**, we cut three Boughs out of a Tree, and stuck them up at a Distance from us. [defoe-1719:190-1]

The late Modern English example in (38a) is taken from a book written by Daniel Defoe, where he describes his travels. He has just been making a self examination, which has put the story completely off the theme line. The return to the theme line is the sentence that starts with the adverbial clause of location *from the Brasils*. This provides the *locational* point of departure for a set of propositions that continue the description of his travels.

A bit further in the story, line (38b) provides us with a *situational* point of departure in the form of the adverbial clause *as we had traded freely with them, and had been kindly used*. This point of departure comes at the point where the perspective changes from *the people* to *we*. At this point, the author inserts a *temporal* point of departure in line (38c): *when we saw the people*. This temporal point of departure is then followed by several topic-comment articulation propositions with *we* as the topic.

Adverbial points of departure are argued to occur only sentence-initially or clause-initially (Dooley and Levinsohn, 2001, Erteschik-Shir, 2007, Lambrecht, 1994: 121, 125, 129, Levinsohn, 2009, Virtanen, 2004). Adverbial phrases have other functions when they occur in non-initial positions (Levinsohn, 1992). The locational adverbial *at a distance* in (38b), for example, does specify the location where *the people* were looking from, but it does not change the perspective from which the narrative develops.

Levinsohn (2009) argues that points of departure need not necessarily be a change in situation, but they can also be a change in “referential” perspective. A story may be from the point of view of one participant, but at some point it may

continue from the point of view of another participant. Instead of just switching the topic, languages may use specific linguistic devices to signal such a referential change in perspective. An example is the use of *however* in the second position, as in (39).

- (39) a. Meantime the fire of the ships ahead, and the approach of the Ramillies and Defence, from Sir Hyde's division, which had now worked near enough to alarm the enemy, though not to injure them, silenced the remainder of the Danish line to the eastward of the Trekroner.
 b. **That battery**, however, continued its fire.
 c. This formidable work, owing to the want of the ships which had been destined to attack it, and the inadequate force of Riou's little squadron, was comparatively uninjured. [southey-1813:463-5]

The narrative in (39a) takes the perspective of the fire, commenting on how it affects the “Danish” resistance. Line (39b) changes the perspective to talk about the “battery”. This now becomes the theme, and we learn that it continued to fire (39b), and was uninjured (39c). The “battery” is not completely new—in fact it is probably still accessible in the addressee’s mental representation of the situation, since it was mentioned in line 449 (which is 14 lines before the current snippet starts).

The category of what Levinsohn (2009) calls “referential point of departure” overlaps to a large extent with what others have called “topic” in the sense of “aboutness topic”, but more specifically where it involves a *change* in the aboutness topic (Krifka, 2007). These categories overlap in the sense that if a clause has a referential point of departure, it also is the topic. But the notions differ, since clauses with a topic-comment articulation always have a “topic”, but they do not always have a referential point of departure—they may have no overtly expressed point of departure, or they may have a temporal or spatial point of departure.

What we should remember about points of departure in general is that we can expect sentence-initial or clause-initial constituents containing information that is accessible to the addressee (in the sense that it is not completely unestablished), and that these constituents come in addition to one of the three focus articulations.

3.3.3 Dominant focal element

The focus domain in topic-comment articulation is relatively broad, spanning the whole predicate. This includes the verb, any following arguments of the verb, and also any adjuncts that follow.¹⁷ The question whether particular elements within the focus domain stand out as more dominant than others has been asked by several researchers. Firbas (1964) argued that one element in the comment has a “higher degree of communicative dynamism”. Heimerdinger (1999: 167) argues that there always is one element of the predicate that is more important than the others, and that this element is the one receiving the accent. In English, that would be the rightmost element within the predicate, since English has ‘the principle of end-focus’ (Leech and Short, 1981).

Others, however, argue that stress on the final constituent in the predicate does not necessarily single out *that* particular constituent, but only indicates the right

edge of a domain (Chomsky and Halle, 1968, Gussenhoven, 1983b).¹⁸ Since stress on the right edge of a focus domain is the default way of demarcating that domain, no special significance should be added to the particular constituent that is at the right edge.

The situation is different, as Levinsohn (2009) argues, when there is a marked (or non-default) order of constituents *within* the predicate (the focus domain of the topic-comment articulation), or when, as is the case for some languages, particular particles are used within the predicate. The constituent that is singled out within the predicate by the marked order or the particle is what Levinsohn regards as the “Dominant Focal Element”.

The unmarked word order within the comment of a topic-comment articulation is the one that, first of all, (a) complies with the language’s syntax rules, and, if there is room for variation, (b) satisfies the Principle of Natural Information Flow (established information precedes unestablished information). A marked word order may appear in English, for instance, with the dative alternation (40a-c).

- (40) a. He gave the girl a book.
 b. He gave the book to a girl.
 c. He gave a book to [_{DFE} the girl].

The word order in (40a) favours a definite indirect object, and that in (40b) a definite direct object. What these first two have in common is that established information (“the girl” in 40a and “the book” in 40b) precedes unestablished information. The variant in (40c), however, has a marked word order: the unestablished information “a book” precedes the established information “the girl”. This word order results in an increased highlighting of “the girl”, which is the last constituent in the predicate.¹⁹

One reason for highlighting a constituent as the dominant focal element is to mark it as an entity that will be picked up again in a subsequent clause, as for example in (41).

- (41) a. The next day was more idly expended in despatching [_{OBJ} a flag of truce] [_{PP} **to the governor of Cronenburg Castle**], to ask whether he had received orders to fire at the British fleet; as the admiral must consider the first gun to be a declaration of war on the part of Denmark.

[southey-1813:118]

The example in (41a) illustrates how a Dominant Focal Element establishes a referent that is picked up subsequently. The verb *despatch* is followed by the indefinite direct object *a flag*, which represents addressee-new information. Next follows *the governor*, which fills in the recipient role of *despatch*. But this last constituent is relatively more established, linking back by inference to *Cronenburg castle* in line 73 of the narrative. What we have here is a marked word order, where less established precedes more established information, in violation of the Principle of Natural Information Flow. The result is not only that *the governor of Cronenburg Castle* becomes the Dominant Focal Element, but also that the subsequent sentence

takes this person as its topical subject, as is illustrated by the pronominal reference *he*.

As we consider how focus changed in the history of the English language, the occurrence of Dominant Focal Elements is something that we need to be aware of. The question is to what extent English has allowed for Dominant Focal Elements. We will see a partial answer to this as we consider two narrative texts in chapter 4, but for the rest the research into Dominant Focal Elements is outside the scope of this dissertation—we concentrate on the development of Presentational Focus and Constituent Focus.

3.4 Marked versus unmarked focus

The syntax of a language describes the linguistic strategies that are used to express grammatical functions and relations, and these strategies may include agreement, case and, where necessary word order. Also part of a language's grammar is the set of "default" or "auto-pilot" word order regularities that help language users process the input more easily. But on top of these observed regularities, languages usually allow for variation. A principle question is whether any and every variation in the *form* of linguistic expressions signals differences in semantic and/or pragmatic *meaning*. If we answer this question with "yes", then it is fruitful to contrast the unmarked (or default) linguistic form with a marked one. The unmarked form associates with an unmarked meaning. No special motivation is needed to use the unmarked form—it is the default one. A marked form is only used if one specific marked meaning is to be expressed. It is important to recognize the asymmetry between marked and unmarked. Use of a marked form is a signal for the presence of one particular meaning. But use of the *unmarked* form is not a signal for a particular meaning.

If we look at word order, for instance, and agree on the idea that SVO is the default or unmarked word order in Present-day English, then marked forms such as OSV ("in he came") and OVS ("into the room came Harry") call for explanations, since these deviate from the standard, but we would not need to "explain" why a sentence has the default SVO form.

When it comes to focus, there are a few ways in which we can apply the marked versus unmarked distinction. The first application is in the realm of the focus articulations, which associate with the focus domain sizes. The unmarked focus articulation is the topic-comment one, which is the default way of telling things. We have a topic in our mental representation, and add a piece of information about this topic. We would expect topic-comment articulation to appear wherever possible. The other two focus articulations are marked ones and call for an explanation. The constituent-focus articulation should only surface when one constituent needs to be highlighted against the background of information that is presupposed—information of which the speaker or writer assumes that it is already in the addressee's mental representation. The presentation-focus articulation associates with a particular function, namely the introduction of a new participant or a whole situation to the scene. In sum, we can talk of the "unmarked focus articulation", implying a clause

or sentence is in the topic-comment articulation, or the “marked focus articulation”, implying there is a constituent-focus or presentation-focus articulation.

Another way to look at marked versus unmarked focus is by highlighting the role of the Principle of Natural Information Flow. We could say that any focus occurring in a sentence or clause that satisfies this principle is unmarked focus, whereas any occurrence of focus that results in a constituent with established information *following* a constituent with unestablished or less established information should be labelled marked focus. This is, perhaps, not so helpful, since it implies that one particular focus articulation, the constituent-focus one, is marked in a language like English, since it usually has the focused constituent clause-initially, where it violates the principle of Natural Information Flow, whereas it is unmarked in an SOV language like Chechen, where the focused information should immediately precede the finite verb, which usually entails that the order of Natural Information Flow is left intact. It would also imply that a word order like OVS, which normally results if the object contains information that is more established than that in the subject (Ward et al., 2002), can be seen as having unmarked focus, which is counter-intuitive.

The way of looking at marked versus unmarked focus adopted in this dissertation will be that both the unmarked focus articulation (topic-comment) as well as the marked focus articulations (sentence-focus and constituent focus) can have a marked and an unmarked form. We have already seen this principle at work in the explanation of the Dominant Focal Element in section 3.3.3: the topic-comment articulation can have a marked form, where the information in the comment contradicts the Principle of Natural Information Flow, and when it does so, the function of this form is to put additional emphasis on the comment-final constituent, possibly with a view of making it ready for reference in a next clause or sentence.

The constituent-focus too can have a marked and an unmarked form, but its association with the Principle of Natural Information Flow is different. The unmarked form of constituent focus is, in fact, *against* this principle, as illustrated in (42a).

- (42) a. [Who]_{FD} did it? [The butler]_{FD} killed him.
 b. You just saw [whom]_{FD}?

English *wh* words routinely violate the principle of natural information flow in favour of the more rigid syntax rules, which require the question operator to appear clause-initially. That is why the focus domain in (42a) is clause-initial. Given the asymmetry between unmarked and marked forms, no special meaning needs to be associated with the unmarked *wh*-word order as in (42a), but the marked order in (42b) calls for a marked meaning (that is: pragmatics comes in). The marked meaning to be associated with (42b) is not in the realm of semantics, but pragmatics. This marked word order associates with strong emotions, the exact content of which depends on the context. The strong emotions could be of surprise, but other contexts may demand a strong emotion of indignation. In sum, both (42a) and (42b) are examples of constituent focus, but the first word order is unmarked constituent

focus, whereas the second word order conveys marked constituent focus, and associates with a strong emotion.

The lesson to be drawn out of the examples above for this current dissertation is that we should not only try to determine if and how the expression of the focus articulations has changed, but also if and how marked and unmarked variants within these focus articulations have changed over time.

3.5 Focus and newness

As we are trying to identify focus, we should keep in mind that focus not necessarily equates with new information. The relation between focus and newness is a subtle one.

Halliday (1967: 176) argues that “new information” receives “information focus”, and Kiss (1998) too uses the category of “information focus” to refer to constituents that are focused because they contain new information. It would seem that these authors make use of a “focus-to-new” principle as in (43).

(43) *Focus to new principle*

Assign focus to the element in a sentence that is new.

The focus articulations discussed in 3.2 lead to a slightly different perspective on the relation between newness and focus. Each focus articulation is defined by a focus domain, and the size of the focus domain depends on the informational content of the constituents. That is to say: the focus domain is defined by containing elements that are new.

However, as we have seen in the different types of constituent focus in section 3.2.2, constituents may be focused without necessarily containing completely or relatively new information. The idea, then, that the focus domain *exclusively* consists of those elements (be they noun phrases, adjuncts or verbs) which are new, is not completely correct. I argue that, as a generalisation, only a one-way relation between new and focus can be defined:

(44) *New to focus principle*

Every constituent that is referentially “new” belongs to the focus domain in the clause or sentence that contains it.

The principle in (44) is a general one that, by definition, always holds. The focus domain by definition contains at least the material in the sentence or clause that is referentially new, but it may, by the definition of focus in (21), also contain non-new material that is contrastive, unpredictable, or otherwise of high communicative interest.

The “new-to-focus-principle” defined in this chapter is a good incentive to the discussion in chapter 5 that leads to the definition of different information state categories, one of which is “new”. The combination of a solid definition of newness, the new-to-focus principle and the acknowledgement of different focus articulations will allow us to locate clauses with different focus articulations, after which we can evaluate their characteristics and, ultimately, derive quantitative and qualitative conclusions as to the way focus has changed in the history of English.

3.6 Discussion

The research described in this book aims at understanding the interaction between syntax and focus, as stated in (11), but in order to do so we need to have a clear definition of what we mean by “focus”, and that has been the aim of the chapter at hand. The discourse processing model described in chapter 2 came up with the concept of the mental model an addressee makes of the text he reads or the story he listens to. Every new line of discourse adds to this model, and the definition in (21) couches focus in the notion of that mental model.

This model helped in defining and understanding three “focus articulations”, which differ in the domain focus occurs in: (a) presentational focus has the whole core of the clause as its domain, (b) the topic-comment articulation has the predicate as its focus domain, and (c) the constituent focus articulation has just one basic constituent as its domain (section 3.2). The focus articulations interact with non-pragmatic factors, such as syntax and text organization, in order to arrive at word orders (section 3.3). Recognizing these factors helps us establish when focus is pragmatically marked (section 3.4). The information we have gathered so far allows us to do preliminary research in chapter 4 that will help us sort out which word orders or constructions can be seen as strategies for conveying particular focus articulations.

¹ The definition loosely derives from a number of existing definitions (Dooley and Levinsohn, 2001, Kiss, 1998, Loos, 2003).

² The reason why ellipsis leads to focus is that material can only be elided (left out) if it has already been established previously, and that means that whatever *is* expressed overtly is most likely to be non-established information. Such information must at least be part of the focus domain (see section 3.5 for the new-to-focus principle).

³ Gussenhoven distinguishes the following types of focus: presentational (the focus is an answer to a question), corrective, counterpresuppositional, definitional, contingency, reactivating (bring old information into the foreground), and identificational (one alternative out of a set of alternatives).

⁴ Sentence (25d) is a diversion, but it is anchored to the mainline of the narration by a temporal adjunct clause, which in itself has the topic-comment structure.

⁵ The *it*-cleft construction is treated more fully in chapters 10-12 of this dissertation.

⁶ English equative clauses are only in this word order (with subject first), since there is no other way to know what the subject is or the complement except by looking at word order. Copula clauses where the complement is not a NP can vary their word order: both *The tree is in the garden* and *In the garden is the tree* have “the garden” as subject, since “in the garden” can only be a complement.

⁷ There are more classes into which copula clauses in general could be divided: specification, predication, equation and identification (Mikkelsen, 2005).

⁸ An author of a book can play us, readers, a trick, by starting the first chapter of the book with just this sentence “The runner was a beautiful lady”. In that case the identity of the

runner has not yet been established, quite likely because the author does not want his readers to know this yet (he wants to surprise us with the lady's identity later). What happens in the reader's mind in this case is that a mental entity for "runner" is created with the information that *is* available, and we are left with the feeling that the runner somehow already is (or should be) familiar to us.

⁹ Since this dissertation is not about *spoken* English, we do not take prosodic means into account, which may be used to express contrastive focus.

¹⁰ Instead of recognizing (30d) as an example where "our language" is explicitly contrasted with a constituent in the following context, one could label "our language" as a *foil*: a constituent that is put into a particular position so that another constituent in the following context can contrast with it (Levinsohn, 2009). Whatever terminology is being used, *foils* (or constituents that contrast with a following constituent) are themselves treated as if they have contrastive focus.

¹¹ The data with the reference starting "BNC" have been extracted from the British National Corpus Online service, managed by Oxford University Computing Services on behalf of the BNC Consortium. All rights in the texts cited are reserved.

¹² Thetic sentences can have several functions within a discourse. Sasse (2006) did a typological search in Indo-European languages and distinguishes five functions: annuntiative (announcing an event, for instance in newspaper headlines), introductive (introducing a participant), interruptive (an event that interrupts the main storyline), descriptive (scene setting) and explanative (explaining an event that has just been mentioned).

¹³ The participant that is presented in the grammatical subject is either completely new to the mental model of the addressee, or the participant comes as a complete surprise at this point in the story, e.g: "*We were talking, when into the room came John.*" The participant "John" is already known from previous episodes, but he comes as a surprise into the current scene, and the unexpectedness is conveyed by using a construction (locative inversion) which is normally used to present new information.

¹⁴ Firbas' paper discusses the work of Mathesius, who was the founder of the Prague linguistic school (Mathesius, 1942). Firbas and Mathesius link the principle of natural information flow to the idea of "functional sentence perspective", which goes back to Henri Weil (1844). This principle normally puts the "theme" (established information) before the "rheme" (most informative information). Comrie's contribution is couched within a description of case marking systems, and is more in terms of degrees of definiteness (Comrie, 1989: 127-128). The term "Principle of natural information flow" has been gleaned from Comrie's work by Levinsohn (2009).

¹⁵ It is possible to change the word orders, arriving at a constellation that violates the principle of natural information flow. This leads to "marked" focus, as will be discussed in section 3.4.

¹⁶ The term "point of departure" first comes from Weil (1844) as 'le point du depart'. This notion compares with, but is not necessarily completely the same as "scene-setting" or "topic" (Lambrecht, 1994: 118) or the notion "theme" in the Prague School.

¹⁷ The focus domain does not include *right-dislocated* elements.

¹⁸ Chomsky and Halle (1968) introduced the Nuclear Stress Rule, which, when applied cyclically, result in a nuclear accent on the rightmost constituent within a focus domain. Gussenhoven (1983a) signalled cases where this does not seem to happen, and posited the Sentence Accent Assignment Rule as an improvement.

¹⁹ All this is not to say that *any* word order alternation can *always* be explained in terms of highlighting or focus. There are obviously much more factors playing a role in word order alternations, such as constituent weight (the number of lexical elements), participant status (the size of a participant's coreferential chain), idioms etc.

This chapter describes my text-charting approach to the study of the changes in the expression of focus in the history of English. A key observation concerning these changes is that Old English allowed the expression of focus in *two* positions in the clause: the clause-initial position was mainly used for constituent focus (see 3.2.2) and the clause-final position was mainly used for presentational focus (see 3.2.3). We will have a look at the changes in both types of focus.

In present-day English, constituent focus is often expressed in clause-final position. The difference with Old English is illustrated by the following OE examples of constituent focus and their present-day English translations (where the subject is bolded and the verb forms are underlined):¹

- (45) a. Ða axode he hine hwæthis nama wære. [coeuphr:159-160]
 then asked he him what his name was
 Ða cwæð he, Smaragdus ic eom geciged.
 then said he Smaragdus I am called
'He asked him what his name was, and the other one answered: "I am called Smaragdus."'
- b. He is leo geciged, of Iudan mægðe. [cocathom1:4912]
 he is lion called of Juda's province
'He is called the Lion of the province Judah.'
- c. Efne sceal mæden geeacnian on hyre innode & oncennan sunu
 behold will virgin conceive in her inside and bare son
 & his nama bið geciged Emmanuhel, [cocathom1:2365-2366]
 & his name will.be called Immanuel
 þæt is gereht on urum gedeode: God is mid us.
 that is explained in our language God is with us
'Behold, a virgin will conceive and bare a son, whose name will be called "Emmanuel," which means "God is with us" in our language.'

Present-day English allows the expression of the type of constituent focus as in (45a) almost exclusively in clause-final position, witness the translations in (46a-c) that follow the word order patterns of the possibilities illustrated in (45a-c):

- (46) a. ??Smaragdus I am called.
 b. *I am Smaragdus called.
 c. I am called **Smaragdus**.

This chapter lays the groundwork for the analysis of the changes by which the three possible positions associated with constituent focus in OE illustrated in (45) were reduced to one: the clause-final position, as illustrated in (46c); a more detailed account follows in chapter 9. The changes in the constituent focus word order are closely linked to major structural changes that took place in the history of English, which I briefly mention here (and in more detail in section 0). Old English had a

version of what is known in the literature as the “verb-second constraint”, which can be exemplified by (47).

- (47) a. Bearn, [for hwilcum þingum] come þu hider? [coeuphr:147]
 child for what matter came you here
 ‘Child, for what cause have you come?’
- b. On Ispanianlande þære Speoniscan leode wæs [coaelive:7814]
 in Spain’s land of the Spanish people was
se halga martir þe hatte Uincentius to menn geboren.
 the holy martyr that called Vincent to mankind born
 ‘In the Hispanian land of the Spanish people the holy martyr called
 Vincent was born to mankind.’
- c. Æfter þisum wordum he eode on ðone weg þe him getæht wæs
 after these words he went on the way that to him pointed was
 oð ðæt he becom to þære ceastre geate. [coapollo:222]
 until that he came to the city’s gate
 ‘After these words, he went on the way that was pointed out to him, until
 he came to the city gate.’

The kind of verb-second in OE divides into two types: obligatory inversion, as in (47a), and pragmatically driven inversion, as in (47b-c).² The clause-initial position in the two types of V2 is category-neutral, and can serve a variety of pragmatic functions: it can be a point of departure (see 3.3.2), a discourse-linker, but it can also host a *focused* constituent, as in (45a), and it is this last feature of the V2 system that is most important for this book.

The alternation between (47b) and (47c) also shows that there are different subject positions: (i) a subject immediately before the finite verb, and (ii) one following the finite verb. These two subject positions have been correlated with their information state (see for instance van Kemenade and Los, 2006a): subjects containing non-established information, such as the NP in (47b), occur after the finite verb, while subjects containing established information, such as the pronoun *he* in (47c) occur before the finite verb.

There is a third subject position in OE: the clause-final position. This position, which is also known as the “late-subject” position, following Warner (2007), is exemplified in (48).

- (48) a. Fæder her is cumen **aneunuchus of cinges hirede**. [coeuphr:142]
 father here has come a eunuch of the king’s household
 ‘Father, a eunuch from the king’s household has arrived.’
- b. Ongemang þissum, com ham **Pafnuntius** [coeuphr:88]
 in.the.midst of.this came home Paphnutius
 ‘In the midst of this, Paphnutius came home.’

The late subject position is primarily used with unaccusative verbs (see van van Kemenade, 1997, Warner, 2007), and it is, as I will show, primarily used for the second kind of focus I will be concentrating on: presentational focus (the introduction or reintroduction of a major participant in a narrative; see section 3.2.3 and chapter 8).

What this discussion shows is that there is a close interrelation between word order as defined by syntax, and focus strategies. This chapter will identify the syntactic changes that formed the backdrop for changes in the expression of constituent and presentational focus.

Before I go on to a more detailed discussion of these changes, I first outline a model for the interaction between three major factors that can have an effect on word order: syntax, focus (information structure), and text (the location within a text can make a difference, as we have seen in (47b-c)).

4.1 A model for word order variations

I am going to assume a model of how the three factors identified in the introduction to this chapter interact to account for the variation in word order found in texts. *Syntax*, roughly speaking, makes use of different linguistic strategies, including word order, to express grammatical functions and reduce the processing load for the speakers by defining “standard” word orders (see the definition I use in section 1.1). *Information structure* makes use of word order to define focus domains (which translate into focus articulations), and also involves the Principle of Natural Information Flow as discussed in chapter 3. *Text-structure*, in the sense of the position of a clause with respect to major or minor paragraphs, can also correlate with particular word orders (I will exemplify this in section 4.1.1).

Since the overall aim of this study is to investigate the interaction between syntax and focus, I will adopt the working hypothesis that the three factors are independent from one another, in the sense that all kinds of combinations of syntax, focus articulations and text-positions can co-occur.³ Roughly speaking, the three factors span a three-dimensional space where each factor is represented by one axis.⁴

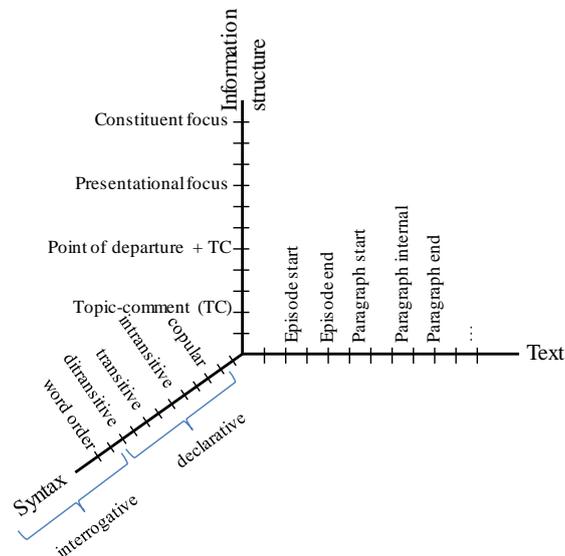


Figure 3 Visualization of the three-dimensional syntax-focus-text space

If the three factors are viewed as three orthogonal axes, then what does a *point* in the space spanned by these axes stand for? A point in this space is specified by the combination of a particular syntactic specification, a particular focus articulation and a particular point in the text structure. The “value” of this point in the 3D-space is the linguistic realization (in terms of word order, morphology and so on) for the combination of syntax, focus and text-structure in a particular language. It is these values that change over time as the English language develops from OE to LmodE.

The different focus articulations have been treated extensively in chapter 3, and I will now proceed with a fuller account of word order as it correlates with text structure, and then show how I intend to model word orders in this study.

4.1.1 Text-structure and word order

The position within a text can correlate with particular word orders. We have seen this to some extent in the examples (49a,b), but I would like to treat the influence of text-structure on word order in more detail in this section, although this factor will figure prominently too in the treatment of the OE narrative (4.6.4) and the LmodE one (4.7.4). Let us consider a small part of a fictional narrative from LmodE:

- (50) a. About half an Hour afterwards they came all up in a Body a-stern of us, and pretty near us, so near that we could easily discern what they were, tho' we could not tell their Design. **And** I easily found they were some of my old Friends, the same Sort of Savages that I had been used to engage with; **and in a little Time more** they row'd a little farther out to Sea, 'till they came directly Broad-side with us, **and then** row'd down strait upon us, 'till they came so near, that they could hear us speak.
- b. Upon this I order'd all my Men to keep close, lest they should shoot any more Arrows, **and** made all our Guns ready; **but** being so near as to be within hearing, I made Friday go out upon the Deck, **and** call out aloud to them in his Language to know what they meant, which accordingly he did; whether they understood him or not; that I knew not.
- c. But as soon as he had call'd to them, six of them, who were in the foremost or nighest Boat to us, turn'd their Canoes from us, **and** stooping down, shew'd us their naked Backsides, just as if in English, saving your Presence, they had bid us kiss. [defoe-1719:2-11]

This part of the larger narrative roughly divides into three paragraphs, which can be seen from the use of a clause-initial adverbial phrase in (50a,b) and a clause-initial adverbial clause in (50c). Roughly spoken, the PP-S-V_{fin} word order in LmodE correlates with a paragraph-partitioning function. The conjunctions *and* and *but* are positioned at the start of clauses so as to provide cohesion within the three paragraphs. And the combination of a conjunction and a time adverbial (twice in 50a) seems to indicate the borders of smaller developmental units within a larger paragraph. There is more to say about the different strategies that either serve cohesion or text-partitioning, and I intend to do so later, within the framework of the discussion on the word orders found in the OE text (4.6) and LmodE text (4.7).

4.1.2 Modelling word orders: the slot-structure model

I intend to model the word order patterns occurring in the different English time periods by determining a “slot-structure” for the language in a particular time period. This slot-structure gives a sequence of slots that host constituents with a particular grammatical function (such as “subject” or “object”), with a particular information content (such as “established” information) or a combination of the two (such as “established arguments”). The aim of the slot-structure is to facilitate the “default” word order patterns of a language as well as deviations from these patterns in a theory neutral way.

The slot-structure approach has its roots in Longacre & Levinsohn (1978), but closely follows the latest trends (Dooley and Levinsohn, 2001, Levinsohn, 2009), and resembles the word order structure that is used, for instance, by Role and Reference Grammar (RRG; see: van Valin, 2005). Slots of sentences divide into four basic “slot areas”, as in (51), each of which can be subdivided into smaller slots where this is helpful for the analysis of the particular language that is being researched:

(51) *Basic slot areas*

Area 1: Sentence introducers like connectives

Area 2: Pre-nuclear constituents (the pre-core slot in RRG)

Area 3: The nuclear predication (the core in RRG terms)

Area 4: Post-nuclear adjuncts and right-dislocated constituents
(the right periphery in RRG)

The slot-structures that I find useful for Old English and late Modern English are based on the analysis of the narratives in section 4.6 and 4.7. I provide them at this point in the chapter, since I would like to refer to elements of these slot structures in the next section where we look at word order phenomena in the history of English. The slot-structure I use for Old English is provided in (52), and an example of sentences using the structure is given in Table 2.

(52) *Slot-structure for Old English*

PreCore

Intro: Conjunctions, disjunctions, logical connectors
like *forþam* “because”

PreAP: Preverbal adverbial phrases, PPs or *þa* ‘then’

PreSbj: Preverbal position filled by **subjects** as well as by RefPoD

Core

Vb1: Usual place for the finite verb (alternative is Vb2)

Sbj: Core-internal subject position

Est: Established arguments

Nest: Non-established arguments

AP: Adverbials, PPs

Vb2: Usual place for the non-finite verb (sometimes hosts finite verb)

PostCore: Anything that is clearly extraposed past the core-end, late subjects

Table 2 Old English slot-structure

#	Intro	PreCore PreAP	PreSbj	Core					PostCore	
				Vb1	Sbj	Est	Nest	AP		Vb2
45		Smaragdus <i>Smaragdus</i>	ic <i>I</i>	eom <i>am</i>					geciged <i>called</i>	
45			He <i>he</i>	is <i>is</i>		leo <i>lion</i>			geciged <i>called</i>	of Iudan mægige <i>from the</i> province Judah
47		On Ispanian lande þære Speoniscan leode <i>in the Hispanian land</i> <i>of the Spanish people</i>		wæs <i>was</i>	se halga martir þe hatte Uincenius <i>The holy martyr</i> <i>called Vincent</i>		to menn <i>to mankind</i>		geboren <i>born</i>	
47		Æfter þisum wordum <i>after these words</i>	he <i>he</i>	eode <i>went</i>			on done weg þe him geteht wæs <i>on the way that was</i> <i>pointed out to him</i>			
48	Fæder <i>Father</i>	her <i>here</i>		is <i>has</i>					cumen <i>come</i>	an eunuchus of cinges hirede <i>a eunuch of the</i> king's household

The three different OE subject positions discussed in the preamble to this chapter are visible in the slot structure: (i) the “PreSbj” slot is used in (45a,b) and in (47c), (ii) the core-internal “Sbj” slot is taken up by the non-established subject of (47b), and (iii) the “PostCore” slot is a multi-purpose one, able to host material such as the PP in (45b), but also the late-subject in (48a).

The slot-structure that will be used for late Modern English is provided in (53) and exemplified in Table 3, by using lines from the “reeve-1777” text (see the further discussion in section 4.7).

(53) *Slot-structure for late Modern English*

PreCore

Con: Conjunctions, disjunctions

PreC: Pre-core position for points of departure (PP, AP)

Core

Sbj: Core-internal subject position

Vb1: Usual place for the finite verb (alternative is Vb2)

Mid: Adverb, subject or negation between finite and non-finite verb

Vb2: Usual place for the non-finite verb (sometimes hosts finite verb)

Arg: Any argument of the main verb

AP: Adverbials, PPs

PostCore: Anything that is clearly extraposed past the core-end

The picture in LmodE looks strikingly different from that in OE. Only one subject position is now left, and that is the “Sbj” slot (the previous “PreSbj” one) that marks the start of the core area. A postverbal logically new subject such as *a battle* in line #77 may appear in the “Mid” slot, but only where the “Sbj” slot is filled by the expletive *there*. The “PreCore” hosts points of departure, such as in line #10, and the PostCore can be used for focus, such as in line #68.

The rationale for choosing these particular slot structures for OE and LmodE will be provided in sections 4.6 and 4.7, which discuss the two narrative texts. It should be noted here that the slot-structure for OE in particular, and that for LmodE more remotely, closely resembles the topological fields that are part of the “topological field model” used for the description of German (Drach, 1937). The German ForeField matches the slot-structure’s PreCore, the German Left Bracket is the same as our “Vb1” slot, the German MiddleField is divided up in the slots Sbj-Est-Nest-AP in our approach, the Right Bracket for German matches the slot-structure’s “Vb2” slot, and the PostField is the same as our “PostCore”.

The difference between the slot structures for OE in (52) and LmodE in (53) illustrate the structural differences between the two language variants: where the OE has a PreCore area that is clearly demarcated by the presence of the “Vb1” slot, hosting the finite verb, such clear demarcation is no longer present in LmodE. The PostCore area also seems to be less clear in LmodE. The two dedicated slots for the subject in OE (the PreSbj and the core-internal Sbj slot) have become one slot in LmodE; it is still possible for subjects to occur in the “Mid” slot in LmodE, but this is no longer a *dedicated* slot for the subject. It is these kinds of changes in the word order *structure* that make the LmodE language so different from the OE one.

Table 3 Late Modern English slot-structure

#	Con	PreCore	Core				Vb2	Arg	AP	PostCore
			Sbj	Vb1	Mid	Arg				
10		after the death of his prince	he	entered				into the services of the Greek emperor		
11	and		—	distinguished		his courage		against the encroachments of the Saracenes		
68			he	survived			his father		but a short time	
77	and	so	there	was	a battle	fought				

4.2 Old English syntax and focus

The OE examples (48a) and (48b) and their PDE translations in the start of this chapter illustrate some of the key changes in word order patterns that have taken place in the English language, but these need to be substantiated by in more detail. This section fills that gap by providing an overview of selected word order phenomena and word order changes in the history of the language. Where appropriate, I will show how these phenomena could be seen against the background of the slot-structure models for OE and LmodE in (52) and (53), and what position these word orders take in the syntax-pragmatics-text structure space. The word order changes described here allow me to formulate important principles at work in the interaction between syntax and focus.

Section 1.2.2 showed a glimpse of subject-auxiliary inversion, and since this word order phenomenon is so clearly related to V2, and the decline of the V2 system in English is probably one of the largest causes for the changes in presentational focus and constituent focus that are going to come up later on in this book, we will have a close look at it.

4.2.1 Syntactic triggers of V2

Roughly speaking, subject-auxiliary inversion is the process where the subject is placed *after* the finite verb and the first position in the clause is taken up by another constituent. PDE uses an auxiliary as the finite verb in subject-auxiliary inversion, as we have seen in the examples in (5), which I repeat here for convenience:

- (5) a. Who **did** you rob for this? [BNC HTY:160]
 b. **Does** the pattern seem satisfactory in the longer term? [BNC K8Y:808]
 c. In no way **did** she wish ill health on the woman. [BNC JXS:3195]
 d. Not a tear **did** she shed. [BNC EFP:35]

The triggers of subject-auxiliary inversion in PDE are questions and negations: the presence of a *wh*-constituent (5a), a negated PP (5c) or a negated NP (5d).⁵ Questions and negations have triggered subject-auxiliary inversion from OE onwards, and they still do so. We could call the motivation for this kind of subject-auxiliary inversion *syntactic*, since it automatically happens with constituents belonging to the grammatical category of “*wh*-question constituent” and “negated constituent”. This kind of subject-auxiliary inversion, then, is the linguistic realization of a particular value along the “syntax” axis of the three-dimensional model introduced in section 4.1.

This particular trigger for the syntactic subject-auxiliary inversion has not changed since OE, but OE was able to perform subject-auxiliary inversion without actually inserting an auxiliary; starting with a [Sbj-V_{fin}... XP ...] clause, the inverted one would have the constituent order [XP-V_{fin}-Sbj ...], as exemplified in (54a).

- (54) a. Bearn [for hwilcum þingum] **come** þu hider? [coeuphr:147]
 child for what matter came you here
 ‘Child, for what cause have you come?’
 b. [Which lord Lovel] **does** your honour enquire after? [reeve-1777:63]

If we skip the vocative *bearn* ‘child’ in the OE example (54a), then the *wh*-question constituent ‘for what cause’ comes clause-initially, and so does the constituent ‘which lord Lovel’ in the LmodE example (54b). The slot-structure approach analyses the OE example (54a) as: *Child* [_{PreCore} *for what cause*] [_{Core} [_{Vb1} *came*] [_{Sbj} *you*] [_{Est} *here*]] (the vocative *bearn* ‘child’ is kept outside the structure). In formal terms, the analysis is: [_{CP} [_{PP} *for what cause*] [_C *came*] [_{IP} *you* [_I *t_v*] [_{VP} *t_S* *here* *t_{PP}*]]] (where traces due to movement are marked with *t*): the *wh*-constituent is in SpecCP, and the finite verb moves out of the VP, through IP into the head C⁰ of the CP.

The difference between OE and LmodE in a formal analysis is that OE still allows the finite lexical verb to move from its base position as VP head through the IP head into the CP head (V-to-I-to-C movement), but LmodE no longer allows this. Only auxiliaries are allowed to appear in second position in these instances (which is interpreted as I-to-C movement in generative terms), and when no auxiliary or modal is associated with the TAM form of the verb, *do*-support is called upon, as in (54b). Roberts (1985) and Kroch (1989), on the basis of data from Ellegård (1953) demonstrate that “V-to-I” movement for lexical verbs was lost by the middle of the 16th century (in the context of questions and negation), whereas “I-to-C” movement has remained.

4.2.2 Pa-initial as V2 trigger

Another trigger for subject-auxiliary inversion is the temporal adverb *Pa* ‘then’. When this word occurs clause-initially in OE, then the finite verb must appear in second position, so in this sense it works the same way as the syntactic V2 trigger explained in the previous section. We have seen the *Pa*-initial at work in example (48a) in the beginning of this chapter, which I repeat here for convenience.

- (48) a. *Pa axode he hine hwæth his nama wære.* [coeuphr:159-160]
 then asked he him what his name was
Pa cwæð he, Smaragdus ic eom geciged.
 then said he Smaragdus I am called
 ‘He asked him what his name was, and the other one answered: “I am called Smaragdus.”’

Both clauses in (48a) start out with *Pa* ‘then’, which is followed by the finite verb *axode* ‘asked’ or *cwæð* ‘said’, and then the subject (in both cases a pronoun) follows. There are several authors who mention the fact that subject-finite-verb inversion occurs in Old English for *pa*-initial (as well as *bonne*-initial) clauses (Enkvist, 1986, van Kemenade, 2002, van Kemenade and Los, 2006a). Two explanations are put forward: a syntactic one and a text-structural one. The syntactic explanation offered by Los (2012) is that *pa* ‘then’ is a clausal marker of modality when it occurs clause-initially. This explains why *pa*-initial clauses behave exactly the same as negation-initial and *wh*-initial ones. A different explanation is offered by Enkvist (1986), who sees the clause-initial *pa* as marking the start of a major section, and being connected with a “lively narrative style”. If this last explanation is followed, then the *pa*-initial clauses are examples of the linguistic realization of a value along the “text” axis of the three-dimensional space introduced in section 4.1.

The *þa*-initial trigger for subject-auxiliary inversion decreases in the history of the English language. The discussion on T-initial clauses of the LmodE text in section 4.7.5.2 shows that there are still instances available where this pattern occurs, but it has by and large disappeared from PDE.

4.2.3 Pragmatic triggers of V2/V3

Throughout the history of English, other clause-initial PPs, NPs and Adverbs have also given rise to subject-auxiliary inversion, and the motivation for this inversion is argued to be pragmatic; related either to focus demarcation or topic/background demarcation (and Los, 2012, see for example van van Kemenade and Los, 2006a). Some examples that could be explained under the heading of “focus demarcation”, which means that the first constituent is the domain for the constituent focus, are these:

- (55) a. Sunnendei wes **ure drihten** iboren of þe halie [cmlamb1:273]
 sunday was our lord born from the holy
 Maiden Marie for ure hele.
 maiden Mary for our bliss
 ‘Our Lord was born for our bliss from the holy virgin Mary on a Sunday.’
- b. (I have appeased him, if a degraded Chief can possibly be appeased, but it will be thirteen days, days of resentment and discontent, before my recantation can reach him.) [johnson-1775:547]
 Many a dirk will **imagination**, during that interval fix in my heart.
- c. (Melody is the most intractable quality.) [bain-1878:94-95]
 Of this alone can **little or no idea** be imparted by translations.

Example (55a) emphasizes the temporal NP *Sunnendei* ‘Sunday’ as being the day on which the Lord (that is Jesus Christ) was born (the following context stresses that it was a ‘Sunday’ when the Lord was baptized in the river Jordan, so ‘Sunday’ is both thematic and focused). Example (55b) puts emphasis on *many a dirk*; not only through the choice of word (which includes the quantifier *many*), but also through its clause-initial position. The example in (55c) is typical too: a PP is positioned clause-initially due to the presence of the focus adverb *alone* (this adverb excludes alternatives, thereby emphasizing the one option given in the PP). With the loss of V2 in the history of English, this clause-initial position for focus disappears too, as we will see in sections 4.3 and 4.4.

The function of the first constituent does not seem to be restricted to focus; the clear demarcation of a domain provided by the finite verb can also be used to separate established from non-established information. This is where the V2 organisation can alternate with a V3 one, as van Kemenade’s (1987) argues: there is a distinction between pronominal and nominal subjects when the first constituent is something other than a *wh*-question, a negation or the temporal adverb *þa/þonne*. The examples in (56), where the finite verb is underlined and the subject is bolded, illustrate this.

- (56) a. Ongemang þisum sende **Eufrosina** anne cniht, [coeuphr:93]
 in.the.midst of.this sent Euphrosyne one servant
 þone þe heo getreowost wiste.
 who that she most.faithful knew
 ‘Meanwhile Euphrosyne sent a servant (one whom she knew to be very faithful).’
- b. On Ispanianlande þære Speoniscan leode wæs [coaelive:7814]
 in Spain’s land of.the Spanish people was
se halga martir þe hatte Uincentius to menn geboren.
 the.holy martyr that called Vincent to mankind born
 ‘In the Hispanian land of the Spanish people the holy martyr called Vincent was born to mankind.’
- c. Æfter þisum wordum he eode on ðone weg þe him getæht wæs
 after these words he went on the way that to.him pointed was
 oð ðæt he becom to þære ceastre geate. [coapollo:222]
 until that he came to the city’s gate
 ‘After these words, he went on the way that was pointed out to him, until he came to the city gate.’

All three examples start with a temporal PP, but (56a,b) have the subject follow the finite verb, whereas (56c) has it precede the finite verb. Van Kemenade (1987) argues that the generalization for this variation is that subjects following the finite verb, as in (56a,b), are lexical NPs, whereas those preceding the finite verb, as in (56c), are pronouns. Los (2012) as well as Hinterhölzl & van Kemenade (2012) interpret the asymmetry slightly differently, arguing that the finite verb in both instances forms the demarcation between a clause-initial more discourse-linked part (*Ongemang þisum* ‘in the midst of this’ in (56a) and *Æfter þisum wordum he* ‘after these words he’ in (56c)), and the new information that follows the finite verb. This is an area of ongoing research, where the referentially enriched English texts, as discussed in chapters 5-7, may play a key role. The example in (56b), for instance, illustrates that the clause-initial part does not necessarily have to be discourse-linked, but contains more “established” information: the mentioning of the Spanish land and people makes use of world knowledge that is readily available. An example with a little bit more context that illustrates the PreCore (see 1.2.1) as containing discourse-linked material is (57a):

- (57) a. (An Antiochia þære ceastre wæs sum cyningc Antiochus gehaten:)
 æfter þæs cyninges naman **wæs seo ceaster** Antiochia geciged.
 after that king’s name was this city Antioch called
 ‘(In the city of Antioch there was a king named Antiochus,)
 from whom the city itself took the name Antioch.’ [coapollo:3-4]

The clause-initial PP *æfter þæs cyninges naman* ‘by the king’s name’ links back directly to the end of the preceding clause *Antiochus gehaten* ‘called Antiochus’, and it does not seem to be contrastively focused at all. The situation may be more complicated, however, since the topic *ceaster* ‘city’ is maintained over the two sentences, and since it *follows* upon the finite verb, it would have to be interpreted as “new”, which is at odds with the fact that it is topical. Examples with the XP-S-V

versus XP-V-S alternation abound in Old English (see Los, 2012 and references therein), but the alternation is reported to decrease and (almost) completely be lost in Present-day English.⁶

A slightly different view on PreCore areas that contain more than one element can be offered against the background of the division between focus articulations and points of departure provided in chapter 3. I would like to turn to the PP-Sbj-V_{finite} word order, following the examples in (58), where the PreField contains two constituents.

- (58) a. *Æfter þisum wordum he eode on ðone weg þe him getæht wæs,*
 after these words he went on the way that him shown was
oð ðæt he becom to þare ceastre geate. [coapollo:222]
 until that he came to of.the city's gate
'After these words, he went on the way that had been pointed to him, until he arrived at the city's gate.'
- b. *On ðissere egeslican reownesse Apollonius geferan* [coapollo:191]
 in this terrible tempest Apollonius' companions
ealle forwurdon to deaðe,
 all became to death
 (and Apollonius ana becom mid sunde to Pentapolim þam ciriniscan lande).
'In this terrible tempest the companions of Apollonius all perished (and only Apollonius managed to escape by swimming to Pentapolis which is in the Cyrenian country).'
- c. *(Eala þu sæ Neptune, manna bereafigend and unscæddigra beswicend, þu eart wælreowra þonne Antiochus se cyngc.)*
 For minum þingum þu geheolde þas wælreownesse
 on my case you reserved this cruelty
þæt ic þurh ðe gewurde wædla and þearfa, [coapollo:197]
 that I through you became poor and needy
and þæt se wælreowesta cyngc me þy eað fordon mihte.
 and that the cruel king me the.easier destroy might
'(O thou Neptune of the sea, bereaver of men, and deceiver of the innocent! thou art more cruel than Antiochus the king)
On my account have you reserved this cruelty, that I through you might become poor and needy, and that the cruel king might the more easily destroy me.'

The example in (58a) has a PP that provides a point in time 'after these words', which is why I would like to interpret it as a point of departure (see 3.3.2) that is followed by a topic-comment articulation; such an analysis determines the position of this word order in the syntax-pragmatics-text structure space posited at the start of this chapter. Formal approaches such as van Kemenade (2000) and also Los (2009) analyze the initial PP in the Spec-CP, and the subject pronoun in something like the Spec-AgrSP or Spec-FP, which is a projection between the CP and the IP that is meant to host "topical" subjects. The slot-structure approach assigns the PP first constituent as well as the subject pronoun to smaller slots within the PreCore, as we have seen in 4.1.2.

The example in (58b) casts doubt on an interpretation of the PP-Sbj-V_{finite} construction that assigns the subject a topical role. The clause starts out with a (temporal) point of departure again, as in (58a), but the subject is not pronominal, nor is it topical. One analysis would be to say that the finite verb has failed to move and is in the “Vb2” slot. But the reason for such failure to move cannot be the same as that of late subject constructions: where late subjects are completely clause-final (they occupy the PostCore slot), the subject in (58b) is not (it is in the Core area). An alternative analysis would be that the subject is a “foil”: it is placed in a position preceding the finite verb so that it has the same word order as the constituent “only Apollonius” in the next sentence, which contrasts with it (see Levinsohn, 2009 for an extensive discussion on “foils”). Examples like these suggest that there need not be a one-to-one reversible mapping between a syntactic construction (the PP-Sbj-V_{finite} one) and a particular function (that of a topical subject).

The example in (58c) seems to have the same word order as that in (58a) and (58b), but there is a slight difference: the clause-initial PP should probably be considered to be an argument of the verb *geheoldan* ‘reserve’ (to reserve something for someone), which is reason to believe that the PP is highlighted here: positioning a referentially unestablished argument from the Core (the MiddleField) to the PreField goes against the Principle of Natural Information Flow, and constitutes a sign that can be picked up immediately by an addressee. The question is whether the highlighting of this constituent is a manifestation of constituent focus, in which case the remainder of the clause should be regarded as backgrounded, or a manifestation of a dominant focal element within a topic-comment structure. The latter option is less likely in this case, since the VP *geheolde þas wælreownesse* ‘reserved this cruelty’ is established information; it is a rewording of the end of the previous clause. The structure of an example like (58c), then, looks much like the ones in (58a,b), since both have the word order PP-Sbj-V_{finite}, but they take up a different position in the syntax-pragmatics-text structure space: one that differs on the syntax axis (since the PP is an argument of the verb) and on the pragmatics axis (example (58c) must be understood as constituent focus instead of the unmarked topic-comment one).

4.2.4 Adverbs as topic-domain dividers

There have been several approaches arguing that fixed-position adverbs demarcate the dividing line between topical and focal information (van Kemenade, 2002, van Kemenade and Los, 2006a). The preceding section has shown that the temporal adverb *þa* or *þonne* in clause-initial position not only triggers Subject-auxiliary inversion, but also signals the start of a major section in a text. This same adverb (again *þa* or *þonne*) can also occur in clause-internal position, and when it does so, Kemenade and Los (2006a) argue that it functions to divide given from new information: the material that precedes the temporal adverb is given (established information), whereas the material following is new (non-established). It is not completely clear where the position of the temporal adverb would have to be in terms of a generative approach (somewhere between the CP and the IP, one would

say); the text-charting approach discussed later in this chapter would put the temporal adverb together with the established material in a slot somewhere in the PreField. Whatever interpretation is taken, it is clear that the word order with an internally placed *þa* or *þonne* serves to signal something associated with the topichood of the first constituent after which it occurs. The analysis of the Old English text discussed later in this chapter (see section 4.6.4.1) suggests that this particular word order has a function on the pragmatics and text-structural axis: it indicates a topic-comment articulation, where the topic is shifted from a preceding one to the current one (this is the “referential point of departure” discussed in section 3.3.2).

4.2.5 Late subjects

There is a construction with the subject appearing in the PostField, as noted by Warner (2007), and illustrated here with an example from the OE text discussed later in this chapter:

- (59) a. Fæder her is cumen **aneunuchus of cinges hirede.** [coeuφr:142]
 father here has come a eunuch of the king's household
'Father, a eunuch from the king's household has arrived.'

One analysis for the construction is that the subject is extraposed (rightward movement) here. The fact that the verb in this construction is almost invariably an unaccusative supports an additional analysis in which the subject is situated in its original position, since subjects of unaccusatives start as objects of the verb (see Warner (2007), following Burzio's (1986) unaccusative generalization).

In terms of the slot-structure model, it is clear that the subject appears after the “Vb1” and the “Vb2” slots in the PostField. Whatever analysis is taken, the function of the late subjects seems to be pragmatically motivated: it conveys presentational focus (the introduction of a new participant in subject position). Apparently this pragmatic motivation combines with a text-structural function (late subjects are usually combined with an initial constituent functioning as a point of departure) as well as with a particular syntactic constellation (late subject constructions typically involve the verb “be” or one of a selected few unaccusative “presentative” verbs), so that this word order fills a well-defined part in the syntax-pragmatics-text structure space.

4.3 Syntactic changes

We have seen that the first constituent in OE, as demarcated by the finite verb in second position, was used for several different purposes: syntactic (section 4.2.1), text-structural (4.2.2) and information-structural (4.2.3). The word order “XP-V_{fin}-Sbj...V_{nonfinite}”, then, can be regarded as a value in the three-dimensional word order space that occurs in several different locations: the first location has a “*wh* question” value or a “negated constituent” on the syntax axis; the initial *þa* or *þonne* clauses have a “major episode start” value on the text structural axis; the information-

structural category either has an “episode-internal cohesion” value on the text structural axis or a “constituent-focus” value on the focus axis.

The decline of subject-auxiliary inversion for non-*wh* clauses has been touched upon in the introduction in Figure 1, and the resulting picture is repeated here because of the important consequences it has had on the changes in focus realizations. Important for the interpretation of the trend depicted in Figure 4 is that the subject position in all instances is the core-internal “Sbj” slot—the slot occurring between the finite and the non-finite verb (see also the algorithm in (8)).

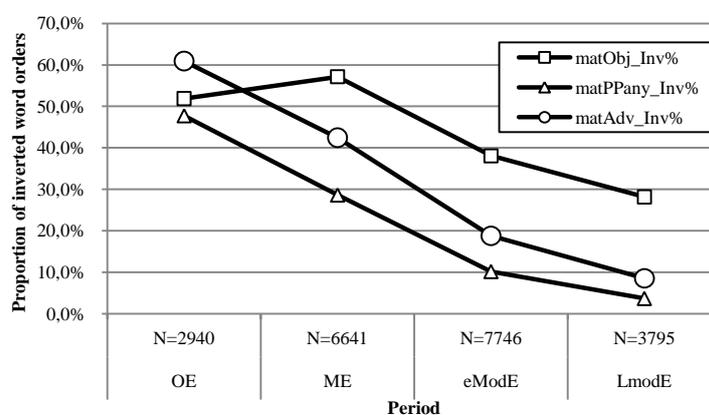


Figure 4 The decline of subject-auxiliary inversion in main clauses from OE (Old English) until LmodE (late Modern English)

What this figure shows is that the multi-functional clause-initial position available with subject-auxiliary inversion has gradually decreased: first for clause-initial PPs, then for adverbs (which includes the temporal adverb *þa* ‘then’), and to a lesser extent for object NPs.

There is a major difference between OE as we have seen it so far and Present-day German, but the combination of pragmatic, syntactic, and text-structural motivations for constituents to appear in the OE PreField begs for a comparison: Speyer (2010) argues that there is a ranked constraint hierarchy determining what is put in the PreField: if *dann* or a scene-setting adverbial is present, it is first of all put in the PreField; if that is not the case, but there is a contrastive constituent, then that appears in the PreField; otherwise the default option is to put the topic in the PreField. Since German allows no more than one constituent (with very few exceptions) in the PreField, the situation there is less complicated, but it seems that the same kinds of constituents competing for a place in the German PreField also compete for a place in the OE PreField. I will leave this as a matter of future research.

The main point in this section on subject-auxiliary inversion is that the change in this word order pattern had a considerable effect on the English language: a clearly demarcated position usable for information structure purposes disappears. This fact

is one of the key triggers for the rise of alternative structures taking over, as we will see in chapters 9-12.

4.4 Changes in the expression of focus

The word order phenomena presented in this section, taken from the existing literature, take up particular areas in the syntax-pragmatics-text structure space. Some of these areas are well defined and remain the same in history (the *wh*-question placement and the negated constituents), but others are less easy to define (such as the area taken up by the PP-Sbj- V_{fin} word order and other word orders where the finite verb could be interpreted as having failed to move forward, resulting in a more subclause-like word order) or their significance changes over time.

We saw in 4.2.3 that subject-auxiliary inversion provided a clearly demarcated domain (the PreCore) in OE for the expression of constituent focus. But we also saw in Figure 4 that subject-auxiliary inversion decreases over time. I summarize the impact this had on constituent focus and on presentational focus in the following ways:

- (60) *Subject-auxiliary inversion impacts constituent focus*
 The decrease in subject-auxiliary inversion means the loss of a strategy to mark constituent focus: demarcate an area (the PreCore area) for the focused constituent where it occurs against the Principle of Natural Information Flow.
- (61) *Subject-auxiliary inversion impacts presentational focus*
 The changing syntax of English makes subjects increasingly occur before the finite verb, which means that the late-subject construction decreases as a strategy to mark presentational focus by demarcation and placement that is against the canonical word order (but in accordance with the Principle of Natural Information Flow).

The principles at work are: (a) demarcation of an area where one particular constituent can naturally “stand out”, (b) mark by going against the Principle of Natural Information Flow and (c) mark by going against canonical word order. We will see in the remainder of this book that English retains these three principles for the expression of constituent focus and presentational focus, but where one construction disappears (the subject-auxiliary inversion, which is the consequence of OE’s V2 nature), other constructions take over.

The constructions that appear depend on the focus articulation we are studying. Chapter 8 will show that presentational focus remains to be expressed in the clause-final (PostCore) position until two important things happen: (a) the pressure of V2 loss for the verb to occur before the finite verb reaches a critical level, and (b) an alternative construction arises. This alternative construction, as will be seen later in this study, is a *there* expletive construction, which takes care of the subject-before-the-finite-verb pressure by positioning a grammatical (but semantically empty)

subject where it is expected to occur, and it takes care of the pressure inherent in presentational focus to have the subject appear as late in the clause as possible.

Chapters 9-12 will show that when the first position is jeopardized for the expression of constituent focus, an existing construction, the *it*-cleft, is hijacked for this type of focus. This is possible, because the *it*-cleft contains (a) a clearly demarcated area for the focused constituent, (b) the focused constituent as such occurs after the main clause's finite verb, and (c) the location of the focused constituent precedes the remainder of the clause. These are exactly the characteristics we saw the V2's first position had, so the *it*-cleft is a worthy alternative. Chapters 9-12 will tell the story more fully, but it may be beneficial to know where this study is heading to as we review the OE and LmodE texts in sections 4.6 and 4.7. I start presenting the investigations of individual texts by introducing the text-charting approach.

4.5 The text-charting approach

The approach taken here to determine the structure of Old English, Middle English and late Modern English is a charting method that has its roots in Longacre & Levinsohn (1978), but closely follows the latest trends (Dooley and Levinsohn, 2001, Levinsohn, 2009). Dooley and Levinsohn (2001: 43) describe charting as a “*visual display of a text in such a way as to make features of interest apparent (by lining them up, for instance)*.” What we are interested in first and foremost is getting an idea of the “default” word order in narrative texts from different periods in English, and then seeing where, how and why deviations from these word orders appear (deviations may also be the appearance of constituents in slots that “normally” hold other material).

4.5.1 Choosing texts to chart

Since the text charting approach is labour-intensive, it will only be done for a limited number of texts; we will take an Old English one and a Late Modern English one as our main texts, giving a detailed analysis of the features found there. These analyses will be supplemented by less detailed ones that “fill the gap” in time. The detailed analyses of the texts in this chapter will give us insight into the word order structure (through the slot-structure) of English, which will help us steer the quantitatively driven corpus research methods in subsequent chapters into the right direction; it will enable us to fine-tune our number-crunching corpus searches.

The basis on which the two texts discussed in this chapter have been chosen is not arbitrary. If the amount of variation that can be attributed to other factors than pragmatics is to be as small as possible, we should concentrate on the least complicated genre: narrative. This genre is simple in text organization, consisting of episodes that are usually organized in a straightforward chronological way, and texts from this genre tend to have one or two main protagonists as well as a limited number of other participants, so that we can observe what happens when attention shifts from one to the other. The criteria which the texts for the in-depth analysis of word order have to satisfy are the following:

- (62) *Text choice criteria*
- a. The text has to be a 3rd person account (there should be no 1st person narrator)
 - b. The text should be part of the parsed corpus
 - c. The beginning of the text in the corpus should be the beginning of the story as a whole
 - d. The genre of the text should be that of narration (which includes: fiction, biography and history)
 - e. The text may not be a translation from another language

The choice in (62a) for 3rd person over 1st person potentially allows us to see more variation in the way participants are referred to: the 1st person is, as it were, always immediately available, only requiring pronouns like “I” or “we”, while 3rd person narratives need to use other NP types, depending on the amount of disambiguation that is needed and the amount of text that intervenes between references to the person. The requirement in (62b) that the text should be part of the parsed corpus enables us to take the syntax of the clauses we encounter into account in a unified manner. We should, as stated in (62c), ideally have a text that starts at the beginning of a story, since that point in the text provides clear situations for scene settings and character introductions, potentially allowing us to see presentational focus at work. The last requirement in (62e) is that the text may not be a translation, since ideally we do not want to end up analyzing a text, finding different word orders, and realizing that (part of) the variation is caused by the fact that a translator copied word orders from the source language to the target one.

The number of texts that comply with these criteria is not very large. The two main texts that have been selected are a biography of Saint Euphrosyne as an Old English narrative, and “The champion of virtue” as a late Modern English narrative. Both texts are 3rd person accounts with a clearly identified main character through most part of the story, as well as several secondary characters. The fact that these texts have a clear main characters should make the narrative more cohesive, while the presence of secondary characters allows us to see switches from the main character to secondary ones and back again, potentially allowing us to observe “referential points of departure” at work (see 3.3.2).

4.5.2 Text-charting as a technique

The charting of a text can be illustrated by seeing how the example sentences provided in (3) of the introduction (see section 1.2.1) are to be represented in a chart (and see also the OE slot-structure examples in Table 2):

Table 4 Charted representation of the sentences in (3a-c)

#	Intro	PreCore		Core							PostCore
		PreAP	PreSbj	Vb1	Sbj	Est	Nest	AP	Vb2		
25	þa <i>then</i>			wurdon <i>were</i>	hire yldran <i>her parents</i>			swidlice <i>exceedingly</i>	geblissode <i>made-joyful</i>	þurh hi <i>through her</i>	
88		Ongemang þissum <i>Meanwhile</i>		com <i>came</i>	[Postposed]		ham <i>home</i>			Pafnuntius <i>Paphnutius</i>	
98	þa <i>then</i>		se cniht <i>the servant</i>	bæd <i>asked</i>		hine <i>him</i>	[IndSp]			with him to Eufrosinan <i>with him to Eufrosyne</i>	
				þæt <i>that</i>	he <i>he</i>		[Postp]		come <i>would come</i>		

What a chart like the one in Table 4 yields is a division of a language's "standard" in columns or "slots": each slot is mainly reserved for one particular kind of constituent. The chart above, for instance, normally has the finite verb in the "Vb1" slot, but there are situations where the finite verb should be charted in the "Vb2" slot, especially in clauses with one verb form that is preceded by one or more arguments (arguments are normally part of the Core). Slightly more accommodation in terms of dedicated slot positions is available for the subject: the subject very often occurs in the core-internal "Sbj" slot, but there is an additional "PreSbj" slot available for situations where such a constellation is more in line with the word order that is observed. The main idea of charting is that the columns capture the generalizations, so that constituents appearing in "odd" places can be dealt with in terms of exceptions, signalling word order phenomena motivated by text-structural or pragmatic reasons. Line #88 in the chart above, for instance, is an example of such a deviation: the subject does not appear in the "Sbj" slot, but in the slot marked "PostCore" (several examples of "late" occurring subjects are needed to determine that the most fitting column for such subjects is the "PostCore" slot; late subjects have been touched upon in 4.2.5).

The result of charting a text yields a slot-structure for the language that has been charted, and such a slot-structure represents a particular language, since it conveys the word order of that language from which pragmatically or text-structurally inspired deviations can be derived most economically (see for instance Clark, 2012). For the English language, which we are investigating in this book, we regard each stage of the language as a separate language, and we therefore would expect each stage to yield a different slot-structure. It is with this language-characterization potential of the slot structure in mind that we will take a brief detour into an attempt to derive a slot-structure from a text automatically.

4.5.3 Automatically charted texts

The hypothesis that a slot-structure is representative of a language's structure can be made more plausible by demonstrating how a slot-structure is arrived at in a relatively unbiased way. The algorithm in (63) can be used to chart a text and derive its slot-structure (this procedure is available in the program "Cesax" that is discussed in chapter 6):

- (63) *The charting process used for texts from the English parsed corpora (Cesax)*
- a. Pass 1: Assign each constituent in the text to a slot in linear order, but:
 - i. Reserve the first slot for connectives (conjunctions)
 - ii. Combine appositives into the preceding slot
 - iii. Combine consecutive adjuncts into one slot
 - b. Determine the most likely slot for the finite verb by frequency
 - c. Pass 2: Consider each clause with a finite verb
 - i. Shift the finite verb rightwards to its slot
 - ii. Shift all following constituents accordingly
 - d. Determine the most likely slot for the non-finite verbs by frequency
 - e. Pass 3: As pass 2, but now for non-finite verbs
 - f. Determine the most likely slot for the subject by frequency
 - g. Pass 4: Consider each clause with a subject
 - i. Shift the subject through empty slots as far right as possible to its own slot

Application of the procedure above leads to a chart for a text where the constituents are mapped into columns in such a way, that (a) connectives are in the first slot, (b) there are fixed slots for the finite and the non-finite verb, and (c) there are one or two slots where the majority of the subjects occur.⁷ The variation that we can expect for the different stages of the English language, representative of the kind of word order variation we are interested in, reduces to:

- (64) *Parameters of slot-structure variation*
- a. The total number of slots needed
 - b. The location of the finite-verb slot
 - c. The location of the non-finite verb slot
 (“b”+“c” combine as:
 the number of slots between the finite and non-finite verb)
 - d. The location of the preferred subject slot
 - e. A possible secondary location for the subject

The slot structure we come up with can be regarded as representing the structural backbone of a language variant, but charting does not stop here. After a chart has been successfully constructed for a text, we need to indicate which constituents have ended up in “odd” slots, and what the possible reasons for these deviations are. When we find, for instance, that the “normal” slot for the direct object (we will assume that this is the slot that is most frequently used for it) is the one following the finite verb in an SVO language, and we stumble upon a direct object occurring before the subject (so that the word order OSV results), then we add a marking like “[Preposed O]” in the slot where the object would normally have been. The markings we use may vary, and include: “Preposed”, “Postposed”, “Left dislocated”, “Right dislocated”.⁸

If the slot-structure as derived by charting a text is representative of the language used in that text, then we should be able to see language change at work in changes in the slot structure. In order to get an initial idea of the changes this kind of approach would show for English, I have implemented the algorithm described in

(63), and run it for a selected number of texts. The results of automatically charting texts in terms of the positions (and existence) of key slots are shown in Table 5.

Table 5 Change in the position of three crucial slots of automatically charted texts

Text	Period	Year	PreSbj	Vb1	CoreSbj	Vb2
coeuphr	O1-4		3	4	5	8
coapollo	O3		3	4	5	7
cmmarga	M1		3	4	5	7
meli	M3		3	4	5	7
malory	M4		3	4	5	6
kempe	M4		3	4	-	6
roper	E1		3	4	5	7
deloney	E2		3	4	5	6
armin	E2		3	4	5	6
perrot	E2		4	4	-	7
burnetroc	E3		3	4	-	6
behn	E3		3	4	-	6
defoe	B1	1719	3	4	-	7
reeve	B1	1777	3	4	-	6
long	B3	1866	3	4	-	6
fayrer	B3	1900	3	4	-	6
benison	B3	1908	3	4	-	6

What this table shows is: (a) there is very little to no change in the size of the PreCore (three slots remain enough); (b) the CoreSbj slot disappears somewhere between late ME and eModE; (c) the number of slots between Vb1 and Vb2 gradually decreases from 3 in OE to 1 in LmodE.

To sum up the section on the text-charting approach we can conclude that text-charting, provided it is done in a systematic way, allows one to derive a generalized structure that fits word order variations found in a language, so that deviations from the ‘standard’ pattern can be readily identified. The next two sections use the text-charting approach in an OE and a LmodE text, allowing us to see some changes in the expression of presentational and constituent focus.

4.6 Old English narrative

The basis for determining the different word orders in Old English is formed by a text that describes the life of “Saint Euphrosyne” (abbreviated as “coeuphr”), which is taken from the YCOE (Taylor et al., 2003). The YCOE compilers categorize it as a text that has *not* been translated from Latin. The fact that it probably is an untranslated narrative with a 3rd person main protagonist (the woman Euphrosyne herself) and several other important persons (her father and the abbot of the minster she ends up in) make it a text that is very suitable for our purposes (see 4.5.1). We will be able to see switches in *time* and *location* due to its narrative character, and switches between the main three protagonists due to the text’s particular content. Since it is a saint’s *life* story, there is a good stretch of presumably relatively

unmarked topic-comment clauses to be expected. And we will be able to see presentational focus at work on the moments where key participants are introduced into the story.

The observations that are being reported here about the Euphrosyne text are based on an 88 line sample of the text that has been “charted” in the sense that the important ingredients of all main clauses have been divided over the OE slot structure that is discussed in detail in section 4.6.2.

4.6.1 Narrative text

This section provides the vernacular text of the first 100 sentences of the “CoEuphr” together with a relatively literal gloss supplied by Skeat (1835-1912), and it is divided into major paragraphs in accordance with the findings described in section 4.6.4. The line numbering follows the YCOE version of this story.⁹

[1.2] III ID FEBRUARI: NATALE
SANCTE EUFRASIÆ VIRGINIS.

[1.2] FEBRUARY 11. ST. EUFRASIA
(OR EUPHROSYNE), VIRGIN.

[1.3] SVM WER WÆS ON
ALEXANDRIA MÆGÐE PAFNVN-
TIVS genemned, se wæs eallum
mannum leof and wurd, and Godes
beboda geornlice healdende, [1.4] and he
þa genam him gemeccan efenbyrde his
cynne; [1.5] seo wæs mid eallum
wurdfullum þeawum gefylled, [1.6] ac
heo wæs unwæstmære. [5.7] Þa wæs hire
wer þearle gedrefed forþam him nan
bearn næs gemæne, þæt æfter his
fordside to his æhtum fenge, [5.8] and heo
þa dæghwamlice hire speda þearfendum
dælde. [8.9] And gelomlice heo cyrcan
sohte, [8.10] and mid halsungum God
wæs biddende þæt he him sum bearn
forgeafe, swiþost forþam heo geseah
hire weres sarignysse. [10.11] And he sylf
eac ferde geond manige stowa, gif he
weninga hwilcne Godes man findan
mihte þæt his gewilnunga gefultumian
mihte. [12.12] Þa æt nyhstan becom he to
sumum mynstre; [12.13] þæs mynstres
fæder wæs swyðe mære beforan Gode.
[13.14] And he þa micelne dæl feos þider
ingesealde, [13.15] and miccle
þeodrædene nam to þam abbode, and to
þam gebroðran; [13.16] and þa æfter

[1.3] There was a certain man in the
province of Alexandria named Paphnu-
tius, who was beloved and honoured by
all men, and who diligently kept God’s
commandments. [1.4] He then took
himself a spouse of birth equal to his
own rank; [1.5] she was filled with all
honourable virtues, [1.6] but she was
barren. [5.7] Then was her husband
exceedingly afflicted, because there was
no child between them who should
succeed to his possessions after his
death. [5.8] She therefore daily distributed
her wealth among the poor, [8.9] and
frequently sought churches, [8.10] and
with supplications besought God that He
would give them a child, especially
because she saw her husband’s sorrow.
[10.11] He himself likewise travelled
through many places, (to see) if
perchance he might find some man of
God who might assist his desires. [12.12]
Then at last he came to a minster. [12.13]
The father of this minster was very
powerful before God. [13.14] So he paid
in a great sum of money, [13.15] and made
much friendship with the abbot and with
the brothers. [13.16] And then after a good

micelre tide cyððe he þam abbode his gewilnunge. [16.17] Se abbod þa him efnargode, [16.18] and bæd God geornlice þæt he þam þegne forgeafe bearnes wæstm. [18.19] Þa gehyrde God heora begra bene, [18.20] and forgeaf him ane dohtor. [19.21] Mid þy Pafnuntius geseah þæs abbodes mæran drohtnunge, he seldan of þam mynstre gewat; [19.22] eac swylce he gelædde his wif into þam mynstre, to þam þæt heo onfenge þæs abbodes bletsunge, and þæra gebroþra.

[22.23] Þa þæt cild wæs seofon wintre, þa letan hi hi fullian, [22.24] and nemdon hi Eufrosina. [23.25] Þa wurdon hire ylðran swiðlice geblissode þurh hi, forþam heo wæs Gode andfencge, and wlytig on ansyne. [25.26] And mid þy heo wæs twelf wintre, þa gewat hire modor. [26.27] Se fæder þa gelærde þæt mæden mid halgum gewirtum, and godcundum rædingum, and mid eallum woruldlicum wisdom; [26.28] and hio þa lare to þam deoplice undernam, þæt hire fæder þæs micclum wundrode. [29.29] Þa asprang hire hlisa and wisdom and gelærednys, geond ealle þa ceastre, forþam heo wæs on þeawum gefrætwod, [29.30] and manige wurdon atihhte þæt hi gyrndan hire to rihtan gesynscipe, [29.31] and hit to hire fæder spræcon; [29.32] ac he symle ongen cwæð, Gewurþe Godes willa. [33.33] Þa æt nyxtan com him an þegen to, se wæs weligra and wurþra þonne ealle þa oþre, and hire to him gyrnde. [35.34] Þa onfeng se fæder his wedd, [35.35] and hi him behet. [36.36] Þa æfter micelre tide þa heo eahtatynewyntre wæs, þa genam se feder hi mid him to þære stowe, þe he gewunlice to sohte, [36.37] and mycelne dæl feo þider insealde. [38.38] And cwæð to þam abbode, Ic hæbbe broht hider þone wæstm þinra gebeda, mine dohtor, þæt þu hire sylle þine bletsunge, forþam ic wille hi were syllan. [41.39] Ða het se

while he told his desire to the abbot. [16.17] So the abbot compassionated him, [16.18] and prayed God earnestly that He would give the nobleman the fruit of a child. [18.19] Then God heard the prayer of them both, [18.20] and gave them a daughter. [19.21] When Paphnutius had seen the abbot's great service, he seldom left the minster. [19.22] Likewise he brought his wile into the, minster, to the intent that she might receive the abbot's blessing, and that of the brethren.

[22.23] When the child was seven years old, then they had her baptized, [22.24] and named her Euphrosyne. [23.25] Then her parents rejoiced exceedingly on her account, because she was acceptable to God, and fair in countenance. [25.26] When she was twelve years old, her mother died. [26.27] Then the father instructed the maiden in holy writings and godly readings, and in all worldly wisdom. [26.28] She so deeply received the lore (=teaching) that her father greatly wondered thereat. [29.29] Then sprang her fame and wisdom and learning throughout all the town, because she was adorned with virtues, [29.30] and many were attracted so that they desired her in honourable marriage, [29.31] and spake of it to her father. [29.32] But he always answered: "God's will be done." [33.33] Then at last came to him a noble who was wealthier and worthier than all the others, and desired her for himself. [35.34] Then the father received his pledge, [35.35] and promised her to him. [36.36] Then after a great while, when she was eighteen years old, her father took her with him to the place where he usually went, [36.37] and paid in there a great sum of money, [38.38] and said to the abbot; 'I have brought hither the fruit of thy prayers, my daughter, that thou mayest give her thy blessing,

abbod hi lædan to spræchuse, [41.40] and lange hwile wið hi spræc [41.41] and lærde hi clænnysse and geþyld, and Godes ege hæbban. [43.42] And heo þa wunode þær seofon dagas, [43.43] and geornlice hlyste þæra broðra sanges, [43.44] and heora drohtnunga beheold, [43.45] and þæs ealles swiþe wundrigende cwæð, Eadige synd þas weras þe on þisse worulde syndon englum gelice, and þurh þæt begitad þæt ece lif. [47.46] And heo weard bihydig be þissum. [47.47] Þa þy drittan dæge cwæð Pafnuntius to þam abbode, Gang fæder þæt þin þeowen ðe mæge gegretan, and þine bletsunge onfon, forþam we willað ham faran.

[50.48] Þa se abbod com, þa feoll heo to his fotum [50.49] and cwæð, Fæder gebide for me þæt God mine sawle him sylfum gestreone. [52.50] Þa aþenode se abbod his hand, [52.51] and hi gebletsode [52.52] and cwæð, Drihten God, þu þe oncneowe Adam ær he gesceapen wære, gemedema ðe þæt þu gymenne hæbbe þisse þinre þeowenne, and þæt heo sy dælnimende þæs heofonlican rices. [56.53] Hi þa æfter þissum wordum ham ferdon. [56.54] Wæs his gewuna Pafnuntius þæt swa oft swa him ænig munuc to com, þonne lædde he hine into his huse, and bæd þæt he his dohtor gebletsode. [58.55] Þa gelamp hit embe geares ryne, þæt hit wæs þæs abbodes hadingdæg. [60.56] Þa sende anne broðor to Pafnuntie, [60.57] and laþode hine to þære symbelnyse.

[61.58] Þa se munuc to his healle com, þa ne funde he hine æt ham. [62.59] Midþy þa Eufrosina þone munuc þær wiste, þa gecigde heo hine to hire [62.60] and cwæð, Sege me broþor for þære soðan lufan hu fela is eower on þam

because I wish to give her to a husband.' [41.39] Then the abbot bade take her to the parlour, [41.40] and spake a long while with her, [41.41] and taught her purity and patience, and to have the fear of God; [43.42] and she abode there seven days, [43.43] and devoutly listened to the brothers' canticles, [43.44] and beheld their conversation; [43.45] and greatly wondering at all this said, "Blessed are these men who in this world are like unto the angels, and by such means shall obtain everlasting life." [47.46] And she became observant of this. [47.47] Then on the third day said Paphnutius to the abbot, "Come, father, that thy handmaid may salute thee, and receive thy blessing; because we desire to go home."

[50.48] When the abbot came, she fell at his feet, [50.49] and said, "Father, pray for me, that God may beget my soul unto Himself." [52.50] Then the abbot extended his hand [52.51] and blessed her, [52.52] and said: "Lord God, Thou who knewest Adam ere he was created, vouchsafe to have care of this Thine handmaid, and that she may be a partaker of the heavenly kingdom." [56.53] So after these words they returned home. [56.54] It was Paphnutius' custom that, as often as any monk came to him, he brought him into his house, and prayed that he would bless his daughter. [58.55] Then it befell, in about the course of a year, that it was the abbot's ordination-day. [60.56] Then he sent a brother to Paphnutius, [60.57] and invited him to the solemnity.

[61.58] When the monk came to his hall, he did not find him at home; [62.59] but when Euphrosyne knew the monk was there, she called him to her [62.60] and said: "Tell me, brother, for true charity, how many are there of you in

mynstre. [64.61] Þa cwæð he, þreo hund muneca and twa and fiftig. [65.62] Heo þa git axode [65.63] and cwæð, gif hwilc þider in bugan wile, wile eower abbod hine underfon? [67.64] Gea cwæð he, [67.65] ac mid eallum \$gefean \$he \$hine \$underfehð swidor for þære drihtenlican stefne þe þus cwæð, [67.66] þone þe me tocynd, ne drife ic hine fram me. [69.67] Singað ge ealle cwæð heo on anre cyrcan, [69.68] and fæstað ge ealle gelice? [70.69] Se broþor cwæð, Ealle we singað gemænelice ætgædere, [70.70] ac ure æghwilec fæst be þam þe him to anhagað, þæt ure nan ne beo wiþerræde wiþ þa halgan drohtnunga, ac wilsumlice do þæt he do.

[73.71] Ða heo þa ealle heora drohtnunga asmead hæfde, þa cwæð heo, Ic wolde gecyrran to þyllicre drohtnunga [73.72] ac ic onsitte þæt ic beo minum fæder ungehyrsum, se for his idlum welum me wile to were geþeodan. [76.73] Se broþor cwæð, Eala swustor, ne geþafa ðu þæt ænig man þinne lichaman besmite; [76.74] ne ne syle þu þinne wlite to ænigum hospe, [76.75] ac bewedde þe sylfe Criste, se þe mæg for þisum gewitenlicum þingum syllan þæt heofonlice rice. [80.76] Ac far nu to mynstre digellice, [80.77] and alege þine woruldlican gegyrlan, [80.78] and gegyre þe mid munucreafe; [80.79] þonne miht þu swa yþest ætberstan. [82.80] Þa gelicode hire þeos spræc, [82.81] and heo þa \$to \$him cwæð, Ac hwa mæg me beofesian. [84.82] \$Sodlice ic nolde þæt hit þa \$dydon \$þe \$nænne geleafan nabbað to Gode. [85.83] \$Se \$broþor \$hyre \$to \$cwæþ, Loca nu, [85.84] þin fæder sceal mid me to mynstre, [85.85] and biþ þær þry dagas oððe feower; [85.86] þonne send þu ða hwile æfter sumum ure gebroþrum; [85.87] ælc wile bliþelice cuman to ðe. [88.88] Ongemang þissum,

the minster?" [64.61] Then said he: "three hundred monks and two and fifty." [65.62] Then again she asked [65.63] and said, "If any one desire to turn in thither, will your abbot receive him?" [67.64] "Yea," said he, [67.65] "but with all (joy will he receive him), the rather for the Lord's voice who thus said: [67.66] 'him that cometh to Me, I will not drive him from Me.'" [69.67] "Sing ye all," said she, "in one church, [69.68] and fast ye all alike?" [70.69] The brother said, "We all sing in common together [70.70] but each of us fasteth according as he hath leisure, so that none of us be contrary to holy living, but do willingly that which he doeth."

[73.71] When she had enquired about all their manner of living, then said she (to the brother), "I would turn to such a life, [73.72] but I fear to be disobedient to my father, who for his vain (and transitory) riches desireth to join me to a husband." [76.73] The brother said (to her), "Sister! suffer thou not that any man defile thy body, [76.74] neither give thou thy beauty to any shame. [76.75] But wed thyself to Christ, who for these transitory things can give thee the heavenly kingdom. [80.76] But go now to a minster secretly, [80.77] and lay aside thy worldly garments [80.78] and clothe thyself with the monastic habit. [80.79] Thus mayest thou most easily escape." [82.80] Then this speech pleased her, [82.81] and she thereupon said (to him): "But who may shear me? [84.82] (Verily), I would not that any one should (do it who) hath not faith in God." [85.83] (The brother said to her): "Look now, [85.84] thy father is going with me to the minster, [85.85] and will be there three days or four. [85.86] Then send thou meanwhile after one of our brothers. [85.87] Any one will blithely come to

com ham Pafnuntius, [88.89] and swa he þone munuc geseah, þa axode he hine to hwi he come. [90.90] Þa sæde he him þæt hit wære þæs abbodes hadungdæg, and he to him cuman sceolde mid him to his bletsunga. [92.91] Pafnuntius þa weard geblissod swiðe, [92.92] and \$sona mid him \$þam \$broþor ferde to mynstre. [93.93] Ongemang þisum sende Eufrosina anne cniht \$þone \$þe \$heo \$getreowost \$wiste [93.94] \$him \$cwæð. \$far \$to \$þeodosies \$mynstre [93.95] \$and \$gang \$into \$þære \$cyrca. [93.96] \$and swa hwilcne munuc swa \$þu \$finde innan cyrcan, bring hine to me. [97.97] Þa \$lamp \$hit þurh Godes mildheortnyse, gemette he an þara muneca widutan þam mynstre. [98.98] Þa se cniht bæd hine þæt he come mid him to Eufrosinan.

[99.99] Þa he to hire com, þa grette heo hine [99.100] and cwæð, Gebide for me.

thee. [88.88] In the midst of this came home Paphnutius, [88.89] and as soon as he saw the monk, he asked him why he had come. [90.90] Then he told him that it was the abbot's ordination-day, and that he was to come to him with him to receive his benediction [92.91] Then Paphnutius was greatly rejoiced, [92.92] and (at once) went with him (the brother) to the minster. [93.93] Meanwhile Euphrosyne sent a servant (one whom she knew to be very faithful, [93.94] and said to him, "Go to Theodosius' minster, [93.95] and go into the church, [93.96] and whatsoever monk (thou shalt find) in the church, bring him to me." [97.97] Then (it happened), by God's mercy, (that) he met one of the monks outside the minster; [98.98] and then the servant prayed him to come with him to Euphrosyne.

[99.99] When he came to her, she saluted him, [99.100] and said: "pray for me."

4.6.2 Word orders motivated by syntax or text organization

The word orders that can readily be attributed to syntactic or text-organizational differences in Old English, as found in the Euphrosyne text, are summarized in Table 6.

Table 6 Word orders in Old English motivated by syntax or text-organization

Variation cause	Name	Word order	Function
Syntactic	Default	S	V _f ... Neutral
	Subclause		C S ... V _f Subclause
	V-initial		V _f ... Mood
Text	Ref PoD	S <i>þa</i> (AP)	V _f ... Reference change
	T-initial	<i>þa</i> (AP) (<i>and</i> V _f 0 ...)	V _f S ... Time change
	AP-initial	PP	V _f S _{lex} Development
		PP S _{pro} (O)	V _f Start
	T-correlated	[_{IP} <i>þa</i> S V _f ...] <i>þa</i>	V _f S ... Time change
	AP-correlated	[_{IP} PP S ...] <i>þa</i>	V _f S ... Time change
	Logical	[_{CP} L (S) V _f ...]	V _f ... Reason, purpose
	Conjunct	<i>and</i> (AP)	- (S) ... V _f Cohesion

Each of the word orders above warrants further discussion and will be illustrated by one or more examples. The kinds of syntactic variation in Table 6 that lead to different word orders are the “Subclause” pattern that is characterized by an initial complementizer and the “V-initial pattern” that is used for sentences starting with a finite verb. The “Subclause” pattern is used in subordinate complement clauses, while the V-initial pattern can find its motivation in the use of a different mood (imperative and interrogative mood). The “Default” pattern distinguished in Table 6 should be seen as a word order pattern that is found in the Euphrosyne text, but for which there is no apparent syntactic, text-organizational or pragmatic motivation (assuming the topic-comment articulation is unmarked in the sense that it requires no motivation).

Text-organizational reasons for varying word orders can be related to partitioning a text in larger or smaller units (“T-initial”, “AP-initial”, “T/AP-correlated”), to expressing cohesion or continuity (“Conjunct”), to expressing a change in participant point of view (“Ref PoD”) and to logical structures (“Logical”).

The syntactic and discourse reasons for variation in OE word order are treated in sections 4.6.3 and 4.6.4, while section 4.6.5 distinguishes word order variations that are focus-motivated. The word order variation should be seen against the background of the slotted clause structure that can be derived from the charting of the Euphrosyne text as shown in Table 7 (see 4.1.2 for the introduction of the slot structure).

Table 7 Division of Old English into slots

#	Intro	PreCore		Core							PostCore
		PreAP	PreSbj	Vb1	Sbj	Est	Nest	AP	Vb2		
25	þa <i>then</i>			wurdon <i>were</i>	hire yldran <i>her parents</i>			swidlice <i>exceedingly</i>	geblissode <i>made-joyful</i>	þurh hi <i>through her</i>	
88		Ongemang þissum <i>Meanwhile</i>		com <i>came</i>	[Postposed]		ham <i>home</i>			Pafnuntius <i>Paphnutius</i>	
98	þa <i>then</i>		se cniht <i>the servant</i>	bæd <i>asked</i>		hine <i>him</i>	[IndSp]			with him to Eufrosinan <i>with him to Euphrosyne</i>	
				þæt <i>that</i>	he <i>he</i>		[Postp]		come <i>would come</i>		

The slots that result from the charting process (see 4.5) as applied to the OE text are part of the larger PreCore-Core-PostCore division, and are the following:

- (65) *Names and functions of the OE slots*
- PreCore
 - Intro: Conjunctions, disjunctions, logical connectors
like *forþam* “because”
 - PreAP: Preverbal adverbial phrases, PPs or *þa* ‘then’
 - PreSbj: Preverbal position filled by **subjects** as well as by RefPoD
 - Core
 - Vb1: Usual place for the finite verb (alternative is Vb2)
 - Sbj: Core-internal subject position
 - Est: Established arguments
 - Nest: Not-established arguments
 - AP: Adverbials
 - Vb2: Usual place for the non-finite verb (sometimes hosts finite verb)
 - PostCore: Anything that is clearly extraposed past the core-end

Important for the remainder of this book is that the OE structure as laid out in (65) has **two** slots available for subjects. There is the dedicated subject slot *inside* the Core proper, and this slot is denoted as “Sbj”. But the slot marked as “PreSbj” (one of the *two* PreCore slots) can also host subjects, although it is more general purpose: it also hosts referential points of departure like *he þa* ‘he, then’. We will see later in the discussion on the late Modern English narrative that the dedicated “Sbj” slot inside the OE’s Core disappears over time. But more on that in section 4.7.

4.6.3 Syntactic variation

Variation in syntax can lead to a difference in word order—especially for languages like English where syntax (in the sense of grammatical functions and relations) partly needs to be expressed by word order. According to the model of the three axes discussed earlier in this chapter, syntactically motivated word order variation is not *necessarily* dependent on or correlated with a particular point on the text-organization axis or the pragmatic axis. The kind of *syntactic* variation that influences word order in Old English we find in the Euphrosyne text is: tense (periphrastic tenses involve the placement of two components), argument structure, mood (declarative versus interrogative) and subordination (main clause versus subordinate clause).¹⁰

4.6.3.1 Default

The “default” word order is the word order used in main clauses where the text-organizational component is unmarked (the word order does not function to signal the start or the end of a section, nor does it generally occur as a continuation marker), and the pragmatic axis is unmarked (the unmarked topic-comment focus articulation is used), and the syntax is relatively simple. It is difficult to speak of “unmarked” syntax, since every clause needs syntax to express its argument structure, tense, mood and aspect. Nevertheless, one could argue that any of these

parameters could have a more marked or less marked value. The declarative mood can be regarded as least marked among moods, and the simple present and past tenses are also probably least marked with respect to the other tenses. What I will refer to as the “default” word order, then, is the order that emerges when the values on the syntax, text and focus axis are least marked. The pattern found in the Euphrosyne text that seems to be least biased in these terms, is [S – V_{finite} ...], which conforms to Baker (2003: ch. 12). Two examples of this default word order in the Euphrosyne text are given in (66).

- (66) a. (Ða æt nyhstan becom he to sumum mynstre.) [coeuphr:12-13]
 þæs mynstres fæder wæs swyðe mære beforan Gode.
 the minster's father was very powerful before God
 '(Then at last he came to a minster.)
 The minster's father was very powerful before God.'
- b. Ic wolde gecyrran to þyllicre drohtnunga. [coeuphr:71c]
 I wanted belong to such living
 'I would like to belong to such kind of living.'

Both (66a) and (66b) are least marked as far as the focus axis is concerned: they are both topic-comment clauses, where the subject represents established information, and the verb phrase contains the non-established information that should be added to the addressee's mental model. The subject *þæs mynstres fæder* “the minster's father (abbot)” in line (66a) is not new—it can be inferred directly from the *mynstra* ‘minster’ in the preceding line, since the mention of a minster causes a link to be made to long-term working memory, where the prototypical minster has a “father”.

The subject in (66b) is a first person pronoun, so within the story itself and for the addressee (the reader) of the story, it is a prime candidate for the topic part of a topic-comment clause. The predicate, which expresses Euphrosyne's desire to become a monk, is a new development at this point of the story.¹¹ The result, then, is a topic-comment word order that is not marked by constituent reorderings or by material that shows continuation within an episode (that is: cohesion) or a breakpoint of an episode.

The examples make clear that the [S – V_{finite} ...] pattern is one that shows up when there are no particular reasons to use word order as a signal for pragmatics (both examples use the pragmatically least marked topic-comment word order) and text organization (none of the two examples seem to signal the start of a paragraph or the continuation of one). The syntax of (66a) does not seem to require a particular order between the main constituents S, V_{fin} and AP either (where the AP is *swyðe mære* ‘very powerful’), since any order would do to convey the grammatical relations.¹² This observation is remarkable in a sense, because it implies that the “PreSbj” slot in (65) is the more basic host for the subject, whereas when one reads through the text, one gets the impression of a wide variation in word order. Indeed, of the first 50 sentences in the Euphrosyne text 24 have the subject in the PreSbj slot, whereas 19 have it in the core-internal “Sbj” slot. However, this seemingly diverse picture is deceptive: all of the 6 neutral main clauses (those that do not belong to a pattern in Table 6 having a text-organizational, a pragmatic or a

syntactic motivation) have their subject in the PreSbj slot, and the core-internal “Sbj” slot is all but reserved for the T-initial and PP/T-correlative word orders discussed in sections 4.6.4.2, 4.6.4.4 and 4.6.4.5.¹³ The fact that the “PreSbj” slot is more or less the default subject position in OE is also implied by Fischer et al (2000: 49), although they do not state this in so many words.

4.6.3.2 Subordinate clauses

Subclauses can be divided into relative clauses, complement clauses and adverbial clauses. We will not treat relative clauses in this chapter, and the discussion of adverbial clauses is taken up later, in sections 4.6.4.5 and 4.6.4.6. Complement subclauses in Old English have a standard verb-final pattern of [C S ... V_i].¹⁴ The subordinating conjunction C seems to occupy the position that would otherwise be available for a finite verb near the start of the clause (as in the “default” pattern as well as all other patterns we have seen so far).¹⁵ Two examples of the subclause pattern are given in (67).

- (67) a. Eala swustor, ne geþafa ðu þæt ænig man þinlichaman besmite.
 dear sister not suffer you that any man your body defile
 ‘My dear sister, do not allow any man to defile your body.’ [73]
- b. Fæder gebide for me þæt God mine sawle
 father pray for me that God my soul
 him sylfum gestreone. [49]
 him self would.get
 ‘Father, pray for me, that God would get my soul for Himself.’

Both examples (67a) and (67b) follow the pattern where the subordinator *þæt* ‘that’ is immediately followed by a subject, then followed by verbal arguments, and finally by the finite verb.

4.6.3.3 V-initial

The verb-initial pattern is the default pattern for clauses in the imperative mood (47b, 49b, 60b, 74a) and in interrogative mood (line 63c and line 67b in the text, the latter of which is copied in example 68a below). As such, it results from variation along the “syntax” axis and does not seem to depend on variation across the text-organizational or pragmatic axes.

Be that as it may, there are a few occurrences of the V-initial pattern that are found in declarative mood sentences, which means that their V-initial word order cannot be attributed to a variation in mood. The charted part of Euphrosyne has one such occurrence, which is shown in (68b).

- (68) a. **Singad** ge ealle on anre cyrcan? [coeuphr:67]
 sing you all in one church
 ‘Do you all sing in one church?’

- b. **Wæs** his gewuna Pafnuntius [coeuphr:54]
 was his custom of.Pafnuntius
 (þæt swa oft swa him ænig munuc to com, þonne lædde he hine into his
 huse, and bæd þæt he his dohtor gebletsode).
'It was Paphnutius' custom
(that, as often as any monk came to him, he brought him into his house,
and asked that he would bless his daughter).'

The clause in (68b) very much is a non-standard one. It starts off with the finite verb *wæs* 'was', which is then followed by the predicate (not the subject) *his gewuna* 'his custom'. The identity of the *his* 'his' is supplied in a parenthetical way, after which the subclause starts. The main clause has no formal subject, witness the fact that the Present-day English back translation has to use a dummy subject *it* to properly translate the Old English.

Although it is difficult to generalize the function of a construction from one example, this particular V-initial instance serves to mark a breakpoint in the narrative that is not connected to the timeline, but concerns a piece of background information. This particular information is needed to understand the subsequent narrative, which talks about a monk visiting the house of Pafnuntius—a key development in the story as a whole, since it is this visit that helps Euphrosyne decide to opt for the monastic life.

Los (2000) follows Enkvist (1986) and others in arguing that the verb-initial main clauses are typical of "lively narrative style", and that they are used to "introduce a new episode" in the discourse; an episode that does not necessarily retain the same theme (this is in opposition to the function of the T-initial clauses, as discussed in section 4.6.4.2). Our observation on the discourse function of the verb-initial word order, then, coincides with that of Los.

4.6.4 Discourse variation

Authors use linguistic clues, sometimes even variation in word order, in order to divide the text into larger episodes or smaller (developmental) units. There are two basically different functions of text organizational word orders. The first indicates a breakpoint in the text—either the end of a section or the start of a new section. The second indicates cohesion, expressing that the current sentence and the preceding sentence form a tight unit. Old English has several breakpoint-indicating word orders (RefPoD, T-initial, T-correlated, AP-correlated, and, as we have seen in 4.6.3.3, part of the V-initial word orders) and one cohesive word order (which is referred to as "Conjunct" and discussed in 4.6.4.7).

4.6.4.1 Referential point of departure

A narrative is usually told from the perspective of one of the participants in the story, and that perspective can change as the story unfolds. Changes in perspective are very obvious when the narrative is first person, but even third person narratives zoom in on one participant at a time, and tend to change this perspective when a new scene is being set.

Clauses with any of the three articulation types defined in section 3.2 can optionally have a “point of departure” (Beneš, 1962, Levinsohn, 2000).¹⁶ This is a constituent, a phrase or subordinate clause that indicates a change in the course of the discourse in terms of location, time, situation or referential point of view. Not every change in the terms just mentioned necessarily is a change in referential point of view. An author may decide to keep the main attention on one particular person or on one particular location, even while a change in time takes place, by placing the time constituent in a non-obtrusive location.

As a logical extension of the “point of departure”, Levinsohn (2009) coined the change in referential perspective (that is: which participant is the thematic one in a paragraph or small episode) that can take place in a narrative a “referential point of departure”. The notion of “point of departure” comes very close to that of “theme” and “topic”, as explained in 3.3.2.

It appears that Old English has a particular construction for conveying a referential point of departure, namely the [S *þa* (AP) V_f ...] construction:

- (69) a. (Ða wæs hire wer þearle gedrefed forþam him nan bearn næs gemæne, þæt æfter his forðside to his æhtum fenge,) and heo **þa** dæghwamlice hire speda þearfendum dælde. [7-8]
and she then daily her wealth to-the-poor shared
‘(Then was her husband exceedingly afflicted, because there was no child between them who should succeed to his possessions after his death ;) and she therefore daily distributed her wealth among the poor.’
- b. (Ða het se abbod hi lædan to spræchuse, and lange hwile wid hi spræc and lærde hi clænnysse and gepyld, and Godes ege hæbban.) [39-42]
And heo **þa** wunode þær seofon dagas,
and she then lived there seven days
and geornlice hlyste þæra broðra sanges,
and devoutly listend the brothers’ songs
‘(Then the abbot bade take her to the parlour, and spake a long while with her, and taught her purity and patience, and to have the fear of God;) and she abode there seven days, and devoutly listened to the brothers’ songs.’

The example in (69a) is equal to line 7 in the chart of the narrative, which speaks from the perspective of *hire wer* ‘her husband’. The referential perspective changes in line 8, where the narrator zooms in on Paphnutius’ wife (whose name we don’t get to hear), and she remains the perspective in clauses 9 and 10 too. Example (69b) shows how the referential perspective changes from *se abbod* ‘the abbot’ in lines 39-40 to that of *heo* ‘she’ (Euphrosyne) in lines 41-42. The distinguishing mark of referential perspective changes is the *þa* particle (usually encoded syntactically as a time adverbial) following upon a clause-initial subject, which may optionally be preceded by the conjunction *and* ‘and’.¹⁷

Van Kemenade’s (2009) stance on the *þa* particle in second (or even further) position is that *þa* (and *þonne*) “...take discourse-linked material on their left”. I go one step further in combining my observations from the Euphrosyne text with the framework of Levinsohn (2009): the *þa* particle in second position signals a

referential point of departure. The noun phrase preceding the *þa* particle has to refer to an already established participant in the story, a participant who is available as topic from the immediately preceding context.

When it comes to the charting slot structure that is proposed in Table 7 and in (65), the clauses with a referential point of departure form an interesting challenge. The subject and the temporal adverb *þa* ‘then’ combine into one and the same “PreSbj” slot: sentence 42 shows that the Sbj + *þa* can be preceded by the conjunction *and* (which means that no overlap with that slot is possible), while sentences like 8, 62 and 81 show that an AP can intervene between the Sbj + *þa* complex and the “Vb1” slot.

Referential points of departure are not restricted to the temporal adverb *þa* in Old English, but may include other temporal adverb phrases, witness the following example:

- (70) a. (Ða þæs on mergen com Pafnuntius to þære ceastre, and þa æfter Godes willan eode he into cyrcan.)
 Eufrosina **betwux þysum** becom to þam mynstre [coeuphr:138-140]
 Euphrosyne between this came to that minster
 þe hire fæder to sohte.
 that her father to visited
 ‘(The morning afterwards Paphnutius came to the city, and then, according to God’s will, he went to the church.)
 Meanwhile Euphrosyne arrived at the minster that her father visited.’

The context has “Pafnuntius” (the father of Euphrosyne) as main topic, telling us that he is visiting a church. Then the camera zooms in on Euphrosyne, looking what *she* is doing, and *she* becomes the referential point of departure. This shift in topic is signalled nicely by the adverbial phrase *betwux þysum* ‘meanwhile’, which is stashed between the subject *Eufrosina* and the finite verb *become* ‘came’.

The temporal adverbial *phrase* as a means to signal a referential point of departure is but a minority feature in OE: the Euphrosyne text has only one occurrence of it, against 14 occurrences of the temporal adverb *þa* as referential point of departure. But we will see in the Late Modern English text (4.7.4.1) that the tables are turned: the LmodE variant of *þa* is used much less as referential point of departure than the temporal adverbial phrase.

4.6.4.2 T-initial

The clauses in Table 6 that I have labelled “T-initial” are ones that start with the temporal adverb *þa* ‘then’. They have already been briefly mentioned in the context of the V-initial clauses discussed in section 4.6.3.3, since those V-initial clauses that have no syntactic motivation for the finite verb to start the clause seem to fulfil a text organization function like the T-initial clause. Two examples of T-initial clauses are given in (71).

- (71) a. **Þa** æt nyhstan becom he to sumum mynstre. [coeuphr:12]
 then at last came he to one minster
 ‘Then at last he came to a minster.’

- b. **Þa** gehyrde God heorabegrabene, [coeuaphr:19-20]
 then heard God them both prayer
and forgeaf him ane dohtor.
 and gave them one daughter
 ‘Then God heard the prayer of them both, and gave them a daughter.’

The first example (71a) is the standard one, complying with the [*þa* (AP) V_f S ...] word order. It seems that the initial *þa* makes it almost impossible for a subject to appear before the finite verb (but see an exception in (3c), charted in Table 4) and the generalizations for pronominal subjects observed by van Kemenade (1987), and the discussions in section 4.2.3). The example in (71b) shows that any *and*-initial clause *following* upon a T-initial one follows the word order of the T-initial clause, and has an elided subject. The *and*-initial clauses following upon T-initial ones, then, differ from the other *and*-initial clauses (which are treated as “conjunct clause” in section 4.6.4.7). The first ones want to have the finite verb closely following upon the *and*, whereas the last ones (the Conjunct clauses) want to have the finite verb as close to the end of the clause as possible (similar to the subclause structure, which will be discussed in section 4.6.3.2).

An interesting observation that can be made about T-initial clauses is that the V2 word order they trigger (the fact that they have to be immediately followed by the finite verb—unless an adverb or adverbial phrase intervenes, as in 71a) otherwise only occurs in *ne*-initial (negated) main clauses and in *wh*-question main clauses (see the earlier discussion in 4.2.3). Van Kemenade and Los (2006a) understand the clause-initial *þa* (as well as the variant *þonne*) as occupying the [Spec,CP] position (and the immediately following finite verb is then in the C-head position), and analyze them as discourse operators, which is why they, on a par with a *wh*-operator or a negative operator, trigger movement of the finite verb to the C-head position. The common denominator between the *wh*-operator, the negation operator and *þa* as discourse operator is that of “clause-typing”. The particular discourse function signalled by the T-adverbs is, according to van Kemenade and Los (and also: Los, 2000, 2006a), that of “discourse continuity”: they signal a new episode in a text, but one with more thematic continuity than signalled by V-initial clauses.

As for the position the T-initial clauses take with respect to the charting slot model illustrated in Table 7 and in (65), we can say that the clause-initial *þa* ‘then’ occupies the “PreAP” slot, which it shares with PPs and other adjuncts.

The T-initial clauses are pragmatically unbiased: they do not necessarily belong to one particular focus articulation and they do not signal some kind of marked focus. They do have a particular function in the *discourse* structure, however, something that has already been noticed by Enkvist & Warvik (1987). T-initial clauses are one of the clause-types, along with T-correlated and AP-correlated ones which are discussed in the following sections, that signal larger episodes within a narrative. The T-initial, T-correlated and AP-correlated clause types together partition the first part of the Euphrosyne story into episodes as shown in Table 8.

The table lists the lines in the text that make up the episode, then the clause type, and then the function of this particular episode in the narrative as a whole. The episodes signalled by T-initial clauses are not necessarily very large—they can be

but one sentence long, as in line 25 and 33 (although these sentences do contain several clauses). What they signal is the point of view of the author: he sees a distinguishable new development in the story that either takes place at one particular time (e.g. the candidate asks for her hand in lines 33a-c) or addresses one common theme (e.g. the seeking for help in lines 7-11).¹⁸

Table 8 Narrative divisioning by special clause-types

Lines	Clause type	Function
1-6	(story start)	Introduce Pafnuntius and his wife
7-11	T-initial	They seek to find help for their barrenness
12-15	T-initial	Pafnuntius finds the minster that will play a key role in the narrative
16-18	T-initial	Pafnuntius shares his need with the abbot, who prays
19-22	T-initial	Pafnuntius strengthens ties with the minster when the prayers are answered
23-24	T-correlated	The child receives her name: Euphrosyne
25	T-initial	The parents are blessed because of young Euphrosyne
26-28	AP-correlated	12-year Euphrosyne is taught by her father
29-32	T-initial	Famous Euphrosyne is sought by young men for marriage
33	T-initial	One suitable candidate asks for her hand
34-35	T-initial	Her father promises her to this candidate
36-38	T-initial	Father + Euphrosyne visit the minster to get a blessing for the coming marriage
39-46	T-initial	Euphrosyne spends time in the minster and adopts their way of life

4.6.4.3 AP-initial

Main clauses starting with an adverbial *phrase*, a PP, are rare in the Euphrosyne text: only four occurrences in the 352 sentences of the whole text. While this makes a treatment of them representative of OE difficult, the occurrences that are found in the text, as listed in (72) and (73), do contain food for thought.

- (72) a. **Ongemang** þissum, com ham Pafnuntius [coephr:88]
 in.the.midst of.this came home Paphnutius
'In the midst of this, Paphnutius came home.'
- b. **Ongemang** þisum sende Eufrosina anne cniht, [coephr:93]
 in.the.midst of.this sent Euphrosyne one servant
 þone þe heo getreowost wiste.
 who that she most.faithful knew
'Meanwhile Euphrosyne sent a servant (one whom she knew to be very faithful).'
- (73) a. Ða cwædon hi, **to** niht we hi gesawon. [coephr:187]
 then said they tonight we her saw
"Then they said: 'Tonight we saw her'"
- b. Ða sædon sume, **be**weningasum man hi beswac. [coephr:196]
 then said some perchance some man her deceived
"Then some said: 'Perhaps a man has deceived her.'"

The examples in (72a,b) both start with the adverbial phrase *ongemang þissum* ‘in the midst of this’, which provides both a clear temporal starting point, and links to the preceding discourse through the demonstrative “this”. It is therefore a clear “point of departure” as described in section 3.3.2. The syntax of these first two examples coincides with that of the T-initial clauses: the initial time adverbial is immediately followed by the finite verb, after which the subject comes.

The charting of Euphrosyne has positioned the clause-initial time adverbials into the “PreAP” slot, which is part of the slots preceding the core-proper (see Table 7). There are, in fact, two possible positions for the time adverbials in the PreCore: they can precede the “PreSbj” slot or combine into it. The situation where they combine into the “PreSbj” slot has been analyzed as signalling a referential point of departure in section 4.6.4.1. Those where the PP precedes the “PreSbj” slot (in which case the PP is in the “PreAP” slot) have been discussed in section 4.2.3, and I have argued that a distinction may have to be made between those PPs that are an argument of the lexical verb, in which case they are likely to have constituent focus, and those that are not (as the ones in (72) above). The reason for this distinction in PP types is that non-argument PPs are likely to be less restricted as far as their clause-internal position is concerned than are argument PPs. The latter are expected to fill a position in the Core of the slot-structure, and their occurrence anywhere else is a strong signal to the addressee that they fulfil a different function; one that is, as I argue, related to pragmatics.

The AP-initial constructions as such, then, do not necessarily associate with any particular focus articulation; they are pragmatically unbiased in that respect. Fuller descriptions of word orders with initial AP may connect with particular values on the pragmatics axis (calling to mind the three dimensional syntax-pragmatics-text structure space posited in the introduction to this chapter). The AP- $V_{\text{fin}} \dots V_{\text{non-final}} \dots S$ construction as in (72a), for instance, seems to consistently signal Presentational Focus. It does seem clear, though, that AP-initial clauses where the AP is *temporal* have the same text-organization function as the T-initial pattern (indicating the start of smaller developmental units). A correlation between the function of these two constructions is logical, given the semantic similarity between the adverb *þa* ‘then’, and adverbial phrases like *Ongemang þissum* ‘meanwhile’: both provide a temporal point of departure that is linked with the immediately preceding sentence.

A slightly different category is formed by the examples in (73a,b): they are at the beginning of a direct-speech interaction, and they have a word order that is different from the T-initial clauses. In fact, the word order they have (PP-S-O- V_{fin}) does not come close to any of the main-clause word orders in Table 6; it looks more like the subordinate clause word order with the finite verb in final position. Fischer et al. (2000: 49) observe that personal pronoun subjects (such as *we* in 73a) tend to precede the (moved) finite verb in OE. But this leaves (73b), with a non-pronominal (and non-specific) *sum man* ‘someone’ unexplained. If we acknowledge the general tendency of both subject and object pronouns to precede the finite verb, then we come as close to an explanation of the word orders in (73a,b) as we can within the framework of this current dissertation.

What is most important to note at this point is the fact that the few sentence-initial adverbial phrases of time and location we find in OE serve as points of departure; a function that they keep having throughout the development of English, as we will see in section 4.7, where we explore an LmodE narrative.

4.6.4.4 T-correlated

The T-correlated clauses in Table 6 are very much a characteristic of OE, and serve to introduce a (from the point of view of the author) significant temporal point of departure, much like the T-initial clauses, as explained in section 4.6.4.2. They mark even larger divisions than those of the T-initial ones. An example is line 23 of the story:

- (74) a. **Þa** þæt cild wæs seofon wintre, **þa** letan hi hi fullian,
 then that child was seven winters then let they her baptize
 ‘When the child was seven years old, they had her baptized.’ [coeuphr:23-24]
- b. and [**þa** hi þa þær hi nahwær ne fundon],
 and then they then there her nowhere not found
 hi weopon hi swylce hio dead wære. [coeuphr:200]
 they bewept her as.if she dead were
 ‘And when they did not find her anywhere, they bewept her as if she was dead.’

The word order pattern for T-correlated clauses is [[_{IP} þa S V_f ...] þa V_f S ...]. This order is similar to that of T-initial clauses, but where T-initial clauses allow for a clause-initial *þa*-AdvP word order (an adverb or adverbial phrase follows the *þa* particle), the T-correlated clauses in fact have the AdvP-*þa* word order, if the first subordinated clause containing the *þa* is correctly labelled as AdvP and regarded as adverbial clause (the first *þa*-clause could also be analyzed as a left dislocation, in which case the second *þa* functions like a resumptive pronoun).

The initial *þa*-clause in T-correlated clauses in Euphrosyne can have a word order pattern [*þa*-S-V_{fin}...], which is reminiscent of the main clause word order pattern with the finite verb in “Vb1”, but also a word order pattern [*þa*-S-...-V_{fin}], which belongs more to the subclause’s pattern with the finite verb in “Vb2”.¹⁹ More research is needed to find out whether there is a significance in use between the two in terms of values on the text-axis or pragmatics-axis. Van Kemenade and Los (2006a) explain the failure of the finite verb to occur in the second position in examples like (74a) from a formal grammar point of view. They note that the first *þa* functions as a subordinating conjunction, which occurs as CP-head, and thereby blocks the finite verb from moving to the CP-head position where it would normally occur as part of the verb-second syntax of Old English. But this does not explain the position of the finite verb *wæs* ‘was’ in (74a). I leave this matter for future research.

What about the position of the T-correlated sentences in the slot approach shown in (65)? It seems that the temporal adverbial clause headed by the first *þa* is in the PreAP slot, while the second *þa* should be put in the PreSbj slot, since it alternates with the subject, witness the example in (74b).

4.6.4.5 AP-correlated

The clause type that is coined AP-correlated in Table 6 seems to function like the T-correlated one discussed in section 4.6.4.4, but there are only a few occurrences in the Euphrosyne text. I have labelled this word order “AP-correlated”, since its main characteristic is a clause-initial adjunct clause where a preposition governs a finite clause (an IP). The AP-correlated clauses mainly seem to have the word order pattern [[_{PP} P [_{IP} S ...]] *þa* V_f S ...], and the word order in the IP within the PP has a restriction that is not found in the first *þa*-clause in the T-correlated ones: the subject must follow the PP introduction (this is consistent in all the occurrences in Euphrosyne). The finite verb, however, does not necessarily need to follow directly upon the subject within the subordinate PP (this is the same as in the T-correlated clauses). Two examples of AP-correlated clauses are shown in (75).

- (75) a. And **mid** þy [heo wæs twelf wintre], **þa** gewat hire modor.
 and with that she was twelve winters then died her mother
 ‘When she was twelve year, her mother died.’ [coeu-phr:26]
- b. **Mid** þy [**þa** Eufrosina þone munuc þær wiste], [coeu-phr:59]
 with that then Euphrosyne the monk there knew
þa gecigde heo hine to hire
 then called she him to her
 ‘When Euphrosyne knew the monk was there, she called him to her.’

The AP-correlated clauses that start with *mid* ‘with’, like their T-correlated counterparts, serve to mark a significant change in *time* in the story, one that is a temporal point of departure for subsequent clauses. The temporal point of departure is established by the situation or event described in the IP subclause within the PP. The size of the episode indicated by AP-correlated clauses is large: the episode starting in (75a) spans lines 26-35, and the one in (75b) spans lines 59-79 (which is as far as the chart has been made).

I would like to treat sentence-initial adverbial clauses with other prepositions than *mid* ‘with’ in this section as well; alternative prepositions may be *gif* ‘if’ and *swa* ‘like/as’.²⁰ Sentences that have initial adverbial clauses with these prepositions are exemplified in (76).

- (76) a. **Gif** ic nu fare to fæmnena mynstre, **þonne** [coeu-phr:129]
 if I now go to women’s convent then
 secð min fæder me þær and me þær findað.
 seeks my father me there and me there finds
 ‘If I go now to a women’s convent, then my father will seek me there and find me.’
- b. and **swa** he þone munuc geseah, **þa** axode he hine [coeu-phr:89]
 and as he that monk saw then asked he him
 to hwi he come.
 for what he came
 ‘And as soon as he saw the monk, he asked him why he had come.’

Note that the sentence-initial *gif* ‘if’ clause in (76a) and the *swa* ‘as’ clause in (76b) behave exactly like the *mid* ‘with’ clauses in (75a,b), in the sense that the word order within the subordinate PP is subject-initial, and the main clause word order is PP-

þa/þonne-V_{fin}-S. Due to the semantics of *gif* ‘if’ and *swa* ‘as’, it is to be expected that the *gif* ‘if’ clauses have more of a logic-division function, while the *swa* ‘as’ clauses have more of a time-division function.

The behaviour of the AP-correlated sentences with respect to the slot approach in (65) is exactly the same as that of the T-correlated sentences: the adverbial clause headed by the preposition combines in the PreAP slot.

What is most important to take away from this section is that the AP-correlated sentences seem to be the predecessors of the sentence-initial adverbial *clauses* that appear in later stages of English (see section 4.7.4.2, where they are labelled “AP-initial”); all of these serve to indicate the start of larger episodes within a text.

4.6.4.6 Logical

Under the heading of “Logical” clauses I combine subordinate adverbial clauses of purpose and reason. Such adverbial clauses start with a logical conjunction like *forþam* ‘because’, and they basically follow the pattern of [L (S) V_f ...], where “L” denotes the subordinating adverbial. Just as the subordinate clauses that are part of the T-correlated and AP-correlated patterns, the Logical pattern looks much like the default one, since the subject (unless it is elided) immediately precedes the finite verb, while it deviates from the standard “Subclause” pattern described in section 4.6.3.2. The “Logical” pattern has the finite verb follow as soon as possible after the subject, whereas the “Subclause” pattern has the finite verb as much to the end of the clause as possible.²¹ The logical pattern serves to provide cohesion: a tight logical link within the narrative, and (77) has some examples from the Euphrosyne where the logical pattern serves this function.

(77) a. (þa wurdon hire ylðran swiðlice geblissode þurh hi,) [coeuþr:25]

forþam heo wæs Gode andfencege
because she was to-God acceptable
'(Her parents were blessed greatly because of her)
since she found favour with God.'

b. (þa asprang hire hlisa and wisdom and gelærednys, geond ealle þa ceastre,) [coeuþr:29]

forþam heo wæs on þeawumgefrætwod,
because she was with virtues adorned
'(Then sprang her fame and wisdom and learning throughout all the town,)
because she was adorned with virtues.'

The example in (77a) shows how the clause-initial logical conjunction *forþam* ‘because’ is followed immediately by the subject, this is followed by the finite verb, and then the remainder of the clause follows. The example in (77b) shows how the verbal arguments follow the initial finite verb, and the clause is then “closed off” by a verbal past participle, just as happens in the default pattern.

Notice however that in all of these instances the subordinate “logical” clause follows the main clause under which it is hierarchically kept. This is the majority

pattern, with only one of the ten occurrences in the 352 sentence-large Euphrosyne text showing a different clause-ordering:

- (78) a. Ða **forþam** se sylfe Smaragdus wæs wlitig on ansyne, [coeph:173]
 then because that same Smaragdus was beautiful in appearance
 swa oft swa ða broðra comon to cyrcan, þonne besende
 as often as those brothers came to church then sent
 se awyrgeda gast mænigfealde geþohtas on heoramod.
 that accursed spirit manifold thoughts into their minds
*'Then, because the same Smaragdus was beautiful in countenance, as
 often as the brothers came to church, the accursed spirit sent manifold
 thoughts into their minds.'*

The word order in (78a) is no exception to the rule that logical subordinate clauses come at the end of their main clauses, because in the current example the sentence-initial logical clause is embedded into the structure of a T-correlated clause.

It is fair to conclude that logical subordinate clauses in OE appear at the end of their main clause hosts, so that they do not have the text-organizing strength of “regular” points of departure, which, by definition, occur main-clause initially (see section 3.3.2). Logical subclauses *do* serve to organize the flow of the text, and to provide tight local cohesion.

4.6.4.7 *Conjunct*

Conjunct clauses are those that start with a conjunction like *and* ‘and’ or *ac* ‘but’. Many of these clauses have an elided subject, which is a clear signal of these clauses’ main function, that of providing tight cohesion.²² We will restrict ourselves here to those conjunct clauses that come with an overt subject. They do not necessarily belong to a separate category, but can usually be grouped together with one of the word orders that have been reviewed so far (most of them allow for the addition of a clause-initial conjunction).

What remains, then, is a group of main clauses that start with a conjunction, that do not have an elided subject and that do not belong to any of the previously discussed types. They can be referred to as coordinate clauses, and one of the proponents to treat them separately is Mitchell (1985: 1685, 1753). This class of clauses clearly fulfils a cohesive function at the discourse level, tying clauses together in an additive way (through the conjunction “and”) or adversatively (through “but”). It is, however, notoriously difficult to determine the pragmatically unmarked word order of conjunct clauses. The part of the Euphrosyne story considered for this chapter suggests a word order pattern of [*and* (AP) - (S) ... V_f...]. This pattern deviates from the default word order, which has the finite verb follow as soon after the subject as possible (it looks more like the Core-internal order with the finite verb in the Vb2 slot). Conjunct clauses of this type follow the subclause pattern (except that they start off with a conjunction instead of a subordinator), having the finite verb occupy the final position of the core, the “Vb2” slot, which is the position that is occupied by the non-finite verb (such as a past

participle) in main clauses that contain an auxiliary.²³ Two examples of conjunct clauses are in (79).

- (79) a. And gelomlice heo cyrcan **sohte**, [coeuphr:9]
 and frequently she churches sought
 ‘She visited churches frequently.’
- b. and hio þa lare to þam deoplice**undernam**,
 and she the teaching to that depth took.in
 þæt hire fæder þæs micclum wundrode. [coeuphr:28]
 that her father of.that greatly wondered
 ‘She took in the teaching to such extent, that her father wondered greatly.’
- c. Ealle we singað gemænelice ætgædere,
 all we sing common together
 ac ure æghwilc**fæst** be þam þe him to anhagað. [coeuphr:70]
 but of-us each fasts by that that him to pleases
 ‘All of us sing together, but each of us fasts according to his inspiration’
- d. Hlaford, ic hæbbe Cristenne fæder, and soðne Godes þeow,
 lord I have Christian father and true God’s servant
 and he hæfd mycclæ æhta, [coeuphr:104-6]
 all he has many possessions
 and his mæcca min modor **is** of þyssum life **gewiten**.
 all his consort my mother is of this life departed
 ‘Sir, I have a Christian father who is a servant of God, and he has many possessions. And his consort, my mother, is departed from this life.’

Example (79a) has a conjunct pattern with an adverbial intervening between the initial *and* and the subject *heo* ‘she’, while examples (79b-d) show the subject immediately following the conjunction (*ac* ‘but’ in this case) since there is no clause-level adverbial. While examples (79a) and (79b) follow the verb-final pattern (the one where the finite verb is in the “Vb2” slot), since both the subject and the direct object precede the verb, this nevertheless is the minority pattern for conjunct clauses in the Euphrosyne text (and, indeed, for Old English in general). The example in (79c) may be slightly misleading: this has the word order pattern [*and* S V_fPP], but the PP is quite likely extraposed due to its length. The word order [*and* S V_f...V_n] as in (79d), where the subject is in the PreCore, occurs more frequently in the Euphrosyne text, which is why it has been posited as the most unmarked (the default) pattern in Table 6. The example shows that the subject is in the PreCore, since there is an explicit Middle Field, demarcated by the finite verb *is* ‘is’ (signalling the Vb1 slot) an argument *of þyssum life* ‘from this life’ (signalling the Core), and the non-finite verb *gewiten* ‘departed’ (a signal of the Vb2 slot).

An explanation for the word order difference is difficult to obtain. Fischer et al. (2000) state that a large number of conjunct clauses do not follow the verb-second pattern (that is: one first constituent followed by the finite verb in the second position), but instead have a verb-final order (where both the finite and, if present, non-finite verb appear in the Vb2 slot, with verbal arguments preceding it). Bech (1999) studied the phenomenon and found that only a small number of conjunct clauses are verb-final, but a majority of the verb-final (main) clauses are conjunct ones. It seems likely that, given the option of two different word orders (finite verb

in the Vb1 versus the Vb2 slot), different authors and different times have led to different functions (in terms of values on the text and pragmatics axes) associated with the conjunct clauses.

The scope of this chapter is too limited to investigate the conjunct clauses more deeply, since we are only using a chart of a part of one text to find rough indications of word order variation that are pragmatically motivated as opposed to stemming from syntactic or discourse organization considerations. From this limited perspective, we can only say that the conjunct clauses serve a cohesive function, since they tightly bind clauses together by signalling through the use of the conjunction. Conjunct clauses form the “glue” of the so-called development units (Dooley and Levinsohn, 2001, Levinsohn, 2009), whereas the T-initial, T-coordinated and PP-coordinated clause types overtly mark the breaks between development units.

4.6.5 Focus in Old English

In the previous sections 4.6.3 and 4.6.4, we have seen the word order patterns in Old English that can be attributed to variations in text organization (e.g. indicating smaller or larger episode boundaries, or indicating cohesion) or in syntactic function (e.g. indicating argument structure, tense, mood, aspect or subordination). All of these word order patterns could be regarded as *pragmatically* unbiased in the sense that they do not necessarily signal a particular type of focus.

Without abandoning the hypothesis of a three-dimensional space projected by the axes of syntax, pragmatics and text-organization (see the start of this chapter), there are some word order patterns that seem to lead to a single *point* in this space rather than to a *plane* or a *line*. The T-correlated (4.6.4.4) and AP-correlated (4.6.4.5) patterns, for instance, follow one particular syntactic construction (thereby fixing the value on the “syntax” axis), and seem to associate quite naturally with an initial point of departure that is followed by a topic-comment articulation (thereby fixing the value on the text-organizational and the pragmatic axes): the initial adverbial clause is a natural locus of the point of departure, and this adverbial clause tends to include a participant who then appears as topic in the subject position of what follows. Other word order patterns may have preferences for one or more articulations too, which means that parts of the three-dimensional space that has been hypothesized are empty.

What we will do in this section is look in the Euphrosyne text for deviations from the pragmatically unbiased (but syntactically or text-organizationally motivated) word order patterns as laid down in Table 6, and see where these word order patterns are used to signal a particular focus domain. The kind of focus we are looking for differs per articulation type, as shown in Table 9, but our main goal is to find focus on *constituents*.

Table 9 Focus types per focus articulation

Articulation	Focus domain	Focus types
Thetic articulation	Whole clause	Subject focus Proposition focus
Topic-comment	Predicate + adjuncts	Dominant focal element (DFE – see 3.3.3)
Constituent focus	One constituent	Contrastive focus and emphatic prominence Open proposition element

As we examine the word order deviations in the Euphrosyne text for those that have a pragmatic motivation, we will keep Table 9 in mind. What we also need to keep in mind is the slot-structure of the OE clause that has been derived from the Euphrosyne text and is shown in (65) and in Table 7. The slot structure, which models OE word order, has several slots allocated to the “core”. We need to clearly define the borders of the “core” of a sentence, so that we know what is inside or outside it, since positioning a constituent *outside* the core looks like a particularly marked way to signal something like focus. The structure of the clause in terms of PreCore, Core and PostCore is one that directly relates to the charting of the narrative, as can be seen from Table 7.

Sentences in Germanic languages in general, as stated in the introduction to this chapter, can be roughly divided in parts according to the topological field model: (a) the prefield, (b) the left bracket (usually containing the finite verb), (c) the middlefield, (d) the right bracket (usually containing a non-finite verb), and (e) the postfield. Such a division is the ideal or prototypical situation, and as we have seen in section 4.6.2-4.6.4, we usually only see parts of these elements. Nevertheless, the information from the preceding sections helps us establish the rules that allow us to determine where the core (consisting of elements (b)-(d)) ends. A summary of the core-end-rules is in (80).

(80) *Core-end rules*

- a. Whenever there is a finite auxiliary and a past participle, then the **past participle** marks the end of the core.
- b. The finite verb in a **complement clause** marks the end of the core.
- c. The late occurring finite verb in a **conjunct** clause marks the end of the core.

The rule in (80a) is a ground rule for the core structure of Old English (as well as other Germanic languages), which says that the structure of the core is normally indicated by a finite verb, arguments and then a past participle.

The subclause rule in (80b) depends on the established unmarked word order of complement subclauses for Old English. The subordinator *þæt* ‘that’ has taken the position where the finite verb would normally be, which is the start of the core: the “Vb1” slot in terms of (65). Old English subclauses are pragmatically neutrally closed off by the finite verb.²⁴

Part of the conjunct clauses, to which (80c) makes reference, follow the subclause structure, which is why they have a similar core-end rule as that for subordinate complement clauses. We should be aware, however, of the fact that conjunct clauses appear in two types: one where the verb is early, and one where the verb is late. It is only the latter type that helps us determine the core-end position.

4.6.5.1 Split constituents

By splitting a noun phrase into two parts an author can satisfy two opposing demands: (a) the Principle of Natural Information Flow (see section 3.3.1), and (b) syntax. What these two demands have to do with split constituents will become clear as we consider the examples in (81), which showthetic focus articulations where the (unestablished) subjects are introduced using split constituents.

- (81) a. **Svm wer** wæs on Alexandriamægde **Pafnuntivs genemned**,
 one man was in Alexandria province Pafnuntius called
se wæs eallum mannum leof and wurd, [coeuphr:3]
 that was to-all to-men loved and valued
'There was a certain man in the province of Alexandria named Paphnutius, who was beloved and honoured of all men.'
- b. **Pa** æt nyxtan com him **anþegen** to, [coeuphr:33]
 then at last came him a noble to
se wæs weligra and wurþra þonne ealle þa oþre,
 that was wealthier and worthier than all the others
 and hire to him gyrnde.
 that her to him desired
'Then at last came to him a noble who was wealthier and worthier than all the others, and desired her for himself.'

Example (81a) is the start of the story about Euphrosyne, where the first major protagonist, the father of Euphrosyne, is being introduced. Since the father is completely new to the story, he should be put at the end of the clause in order to comply with the Principle of Natural Information Flow. But OE syntax requires the first constituent to be filled, and there is a preference for the subject to take this position in S+V+PP constructions. There is no topical subject available yet, since this is the start of the story. The solution chosen by the author is that of a split constituent: put the main part of the non-established subject in the default subject position (clause-initial), but put the remainder, the apposition to the subject, which gives away the name of the person, in a sentence-final position. This ordering has the advantage that the second clause, a *se*-clause, is able to pick up the participant very easily. The *se*-clause can either be interpreted as a separate main clause that starts with a *se* demonstrative, or as a relative clause. The former interpretation is possible, since there is no complementizer, and there is no other sign of the *se*-clause being a subclause. The latter interpretation is possible too, if the stand-alone *se* is regarded as a relative pronoun.

The introduction of the most serious competitor for the hand of Euphrosyne (the one that gets approval from her father) is shown in (81b). If we regard the *se* clause in this example as an extraposed relative clause, then we have a split constituent

with *an þegen* ‘a nobleman’ as its head NP. Even if one regards the *se*-clause as an independent main clause, on a par with the situation in (81a), the introduction of the relatively important “nobleman” in the story still is accompanied by a split constituent: the PP *him to* ‘to him’ is split into two parts by the insertion of the subject *an þegen* ‘a nobleman’.²⁵ There are only three occurrences of such a split PP construction in the story of Euphrosyne, and two of them (including the one in 81b) are used in a thetic focus articulation. The second one is shown in (82).²⁶

- (82) a. Wæs his gewuna Pafnuntius [coeph:54]
 was his custom of-Paphnuntius
 þæt swa oft swa him **ænig munuc** to com,
 that as often as him any monk to came
 þonne lædde he **hine** into his huse,
 then led he him into his house
 and bæd þæt **he** his dohtor gebletsode.
 and asked that he his daughter bless
‘It was Paphnutius’ custom that, as often as any monk came to him, he brought him into his house, and prayed that he would bless his daughter.’

The subject *ænig munuc* ‘any monk’ in (82a) has been inserted into the PP *him to* ‘to him’, and introduces a new participant to the scene—even though this participant does not refer to any individual. The example is one of thetic articulation (the focus domain spans the whole clause, including the action “come” and the subject), where the new subject is introduced in an emphatic way.²⁷ The author seems to use this construction to highlight the attitude of Paphnutius: whatever kind of monk would visit, Paphnutius would ask him to bless his daughter. It also highlights an important twist in the story, since it is exactly through this love-inspired attitude of asking monks to bless his daughter, that one day a monk inspires his daughter to adopt the monastic life herself, which then leads to a long-term traumatic experience for Paphnutius.

The split-constituent method when used to convey clauses with a thetic articulation either keep the constituents as close as possible within the slots they normally occur in, as per (65), or they use the “PostCore” slot for the information that is to be highlighted most.

4.6.5.2 Apposition and focus

The introduction of new participants is quite often accompanied by apposition: characteristics of the hearer-new referent are added in an appositive phrase or clause. We have seen some apposition already in the previous section on split constituents, notably in example (81a), where the referent is syntactically introduced in the subject of the clause, and an appositive NP at the end of the clause adds information that makes it easier for the reader to establish the identity of the person that has been introduced. The first mention of the referent (the subject NP *some man*) causes a mental entity to be created, while the sentence-final appositive NP results in characteristics to this mental entity being added in the addressee’s mental model of the situation. If apposition then is a feature of hearer-newness, and if a hearer-new syntactic subject is associated with presentational focus (as suggested in

section 3.2.3), then it should come as no surprise that a syntactic subject that is accompanied by an appositive NP is highly likely to point to presentational focus, as illustrated in (83), which are taken from the larger part of the Euphrosyne narrative, outside the charted sample.

- (83) a. **Agapitus min lareow** me rehte be þe hu swyðe
 Agapitus my master me related about you how sorely
 þu gedrefed eart, æfter þire dehter, and hu þu þæs abbodes
 you afflicted are about your daughter and how you of.the abbot's
 fultumes bæde, and his broþra. [coeuphr:257]
 aid requested and his of.brother
'Agapitus my master has told me about you, how greatly you are afflicted about your daughter, and how you have asked the aid of the abbot and his brothers.'
- b. **Pafnuntius þa witodlice, hire fæder,** þa he ham com
 Paphnutius then truly her father when he home came
 ofestliceode inn to þam bure þe his dohtor inne [coeuphr:184]
 quickly went in to that room that his daughter in
 gewunode beon.
 living was
'But when he came home, her father Paphnutius very quickly went into the room where his daughter usually was.'

The situation of (83a) is as follows: Euphrosyne is at the end of her life, and her father has come to her without knowing who she is. He tells her that he is still so much in grief and pain over the disappearance of his daughter, and she tries to comfort and encourage him, linking him to her own spiritual teacher, Agapitus. Since it has not yet been established that Euphrosyne and her father have this common acquaintance, she is introducing him in (83a) by mentioning Agapitus as the syntactic subject, and tagging him with the apposition “my master”. The whole sentence is *thetic*, involving information that Euphrosyne’s father should realize is now shared information between them, but as is the case in presentational focus, the most important piece of information is located in the subject of the clause.

The situation in (83b) is a bit earlier in the narrative: this is where Euphrosyne’s father Paphnutius finds out that his daughter is missing. Here we have a hearer-old subject “Paphnutius” about whom information is given: isn’t this a typical topic-comment articulation? Nevertheless, the syntactic subject is accompanied by apposition, so one might wonder whether apposition really is such a clear sign of the hearer-new status. There are two things going on here, I believe. It is true that the status of Paphnutius is hearer-old, but he has been out of the scene for some time, and, perhaps more importantly, there is an important change in his geographical whereabouts: he apparently has left the minster his daughter just entered (line 138), and now comes back to his own house, where, for all he knows, his daughter should be. The second important thing to note is the fact that a two AdvPs intervene between the subject Paphnutius and the apposition “her father”. We have seen in section 4.6.4.1 that the [S *þa* (AP) V_f ...] construction is typical of a referential point of departure. Indeed, one could say that the (83b) is a sentence that

is marked by having a switch in topic (a switch from Euphrosyne to Paphnutius) as well as a switch in time, which is indicated by the temporal point of departure *when he came home*.

In sum, apposition quite likely associates with the introduction of a hearer-new referent, in which case it may signal presentational focus (without distorting anything in the slot model of the sentence, as shown in (65)), or with the appearance of a referent that is new to a particular geographical scene, where it may either indicate presentational focus or a topic-comment articulation (as it does in 83b). This last use of apposition does seem to “cram” the slots in the pre-core, which, according to the model in (65), only have a “PreAP” and a “PreSbj” one. It seems that the PreSbj slot contains both the subject + *þa* complex (sign of the referential point of departure), the adverb *witodlice* ‘truly’, the apposition of the subject *hire fæder* ‘her father’, as well as the temporal adverbial clause *þa he ham com* ‘when he came home’. Slot-cramming, apparently, is a clear-enough signal to the reader, while it keeps the general word order structure (the slot-division) intact.

4.6.5.3 Unestablished information as DFE

The notion of “Dominant Focal Element” or DFE has been defined in section 3.3.3 as the constituent that ends up in a right-shifted position in the predicate of a clause that has the topic-comment articulation; it is the constituent that is *marked* as more informative or prominent than others within a focus domain that consists of more than one constituent (Heimerdinger, 1999, Levinsohn, 2009).²⁸ The shifting to the right may be either (a) contrary to the Principle of Natural Information Flow established information tends to precede unestablished information (see Comrie, 1989, Firbas, 1964), or (b) contrary to the unmarked (neutral) order of constituents. Dominant Focal Elements can occur *inside* the core and *outside* it. We concentrate on DFEs occurring *after* the core in this section, and look at core-internal DFEs in section 4.6.5.4.

There are two principally different situations in which post-core DFEs can occur, and in both situations these occur in sentences with a topic-comment articulation. The first situation, which is treated in this current section, is when the DFE contains *non established* information that should have been positioned in the core, but is moved out of it for highlighting, as exemplified in (84).

- (84) a. Ic hæbbe broht hider **þone wæstm þinra gebeda, mine dohtor**. [38]
 I have brought here the fruit your of-prayer my daughter
‘I have brought you the fruit of your prayers: my daughter.’
- b. (Se abbod þa him efnsargode, and bæd God geornlice)
 þæt he þam þegne forgeafe **bearnas wæstm**. [17-18]
 that he the nobleman would-give child’s fruit
‘So the abbot compassionated him, and prayed God earnestly that He would give the nobleman the fruit of a child.’
- c. and heo þa dæghwamlice hire speda **þearfendum** dælde. [8]
 and she then daily her food to.the.needy shared
‘and she daily distributed her wealth among the poor.’

The example in (84a) to a great extent follows the “default” order that is least marked from the point of view of syntax and discourse, as explained in section 4.6.3.1, following the pattern of [S (AP) V_f ...]. However, the past participle *broht* clearly marks the right edge of the core, so that the constituents following it are outside the core in the “PostCore” slot according to (65). This marked location signals highlighting, and the reason for this highlighting is clear in the immediately following context: this daughter is the topic of conversation in the following two lines 38c and 38d (see the chart of the text at <http://erwinkomen.ruhosting.nl/phd>).

We should keep in mind, at this point, that the highlighting caused by having a constituent in the PostCore slot is *not* automatically one of “constituent focus”. The case in (84a) does not have the “fruit” contrasting with anything mentioned previously or afterwards, nor is this “fruit” the value of a variable that has been instantiated earlier in the text. Instead, (84a) has a topic-comment articulation, and the PostCore slot only functions to host a DFE: one constituent of the whole predicate that receives additional highlighting, although the domain of focus spans the *whole* of the predicate.

The example in (84b) is a topic-comment articulation in a subclause, where *God* is the topic, and the predicate “give the nobleman a child” demarcates the focus domain. The pragmatically neutral word order for a subclause would have the finite verb *forgeafe* ‘would give’ demarcate the right border of the core. The direct object *bearnes wæstm*, ‘the fruit of a child’, occurs in the PostCore slot (see the model in (65)), which means that it is a DFE—despite the fact that its late occurrence is completely in line with the Principle of Natural Information Flow.

Old English has a choice in positioning non-established verbal arguments before or after the “Vb2” slot, the slot where we would expect to find the non-finite verb. The placement before the finite verb as in (84c) is the pragmatically most neutral one, since it satisfies the Principle of Natural Information Flow where syntax does not need word order to express grammatical relations. The direct object *hire speda* ‘her food’ precedes the indirect object *pearfendum* ‘to the needy’, and both of them occur before the finite verb *dælde* ‘shared’, which indicates the end of the Core.²⁹

DFEs are not restricted to the topic-comment articulation, but may also occur as part of presentational focus: this last articulation too has a focus domain that is larger than one constituent (the subject plus the verb phrase), so that one of the elements can be made to stand out more than the other(s) by an author. An example of a presentational focus DFE is given in (85), with the introduction of a “eunuch”.

- (85) a. Fæder her is cumen **aneunuchus of cinges hirede**. [coeuphr:142]
 father here has come a eunuch of the.king’s household
 ‘Father, a eunuch from the king’s household has arrived.’

The sentence in (85a) starts out with a vocative and then a locative adverbial *her* ‘here’, which functions as a point of departure. There is presentational focus, since the focus domain contains both the verb phrase *is cumen* ‘has come’ and the subject *an eunuchus of cinges hirede* ‘a eunuch from the king’s household’. The start and end of the core are indicated by the finite verb *is* ‘has’ (in the “Vb1” slot) and the past participle *cumen* ‘come’ (in the “Vb2” slot), so that it is clear that the subject

has been positioned *outside* the core in the “PostCore” slot, which is a marked position. The eunuch receives a bit more prominence within the focus domain, which is a general characteristic of “presentational focus” as opposed to the more generic “thetic articulation” (see section 3.2.3).

4.6.5.4 Established information as DFE

The second position where DFEs can occur is the “PostCore” slot (see 65). Constituents positioned there are recognizable as DFEs when they contain information that is relatively *more established* than other parts of the predicate, thereby violating the Principle of Natural Information Flow, as for example (86).

- (86) a. *Ʒa wurdon hire ylðran swiðlice geblissode þurh hi* [coeuþr:25]
 then were her parents exceedingly gladdened through her
‘Then her parents rejoiced exceedingly on her account.’

The post-core DFE example in (86) is a T-initial clause, indicating the start of a new episode, but otherwise following the topic-comment articulation, continuing the “parents” topic from lines 23 and 24. The post-core constituent is the PP *þurh hi* ‘on her account’, which is part of the focus domain, but its placement is contrary to the Principle of Natural Information Flow, since the reference to *hi* ‘her’ is more established than the reference to the blessing. There seems to be a good reason for putting the reference *hi* ‘her’, which refers to Euphrosyne, in this prominent “PostCore” slot, because she becomes the topic of the next clauses (lines 25b, 25c, 26a and 26b). Having established information in a PostCore slot, then, appears to be one method of introducing topic-shift.

4.6.5.5 Adverbial DFEs

The DFEs need not be restricted to verbal arguments or PPs, but can also come in the form of adverbs, where it is possibly harder to speak of “established” versus “unestablished” information. If we keep in mind the core-end rules laid down in (80), then any constituent following a past participle ought to be considered as a DFE, as illustrated in (87).

- (87) a. *Pafnuntius þa weard geblissod swiðe.* [coeuþr:91]
 Paphnutius then became gladdened greatly
‘Paphnutius became very glad.’
- b. *Ʒa hira brydguma gehyrde þæt heo losad wæs,* [coeuþr:193]
 then her bridegroom heard that she lost was
þa weard he swiðe gedrefed.
 then became he greatly troubled
‘When her bridegroom heard that she was lost, he became greatly troubled.’

The adverb *swiðe* ‘greatly’ in (87a) as well as in (87b) modifies the core-final past participles, which are in the “Vb2” slot (see 65). But in (87a) this adverb occurs after the participle (in the “PostCore” slot), whereas in (87b) it occurs before the participle (in the core-internal “AP” slot). Since adverbs have a relatively normal

position *inside* the core, the fact that they occur *outside* it in the PostCore slot is enough to mark them as a DFE (provided there is a larger focus domain, which is the case here). The adverb *swiðe* ‘greatly’ in (87a) is a DFE, and the fact that it is highlighted makes sense within the context following line 91 of the story: it is Paphnutius’ gladness that brings him to a prompt visit to the abbot’s ordination, providing an ideal occasion for his daughter Euphrosyne to get someone to cut her hair, and then make her way to the minster as a monk in disguise.

The amount of troubling that happened to the bereaved bridegroom in (87b), however, is probably not too significant. We briefly read that he has people looking for Euphrosyne, but then he completely vanishes from the story, clearly indicating the minor role he plays in the eyes of the author.

The adverbial DFEs consolidate the picture we are gradually seeing about the interaction between focus and syntax: word order slots *can* be used for highlighting purposes, but there is no direct syntax-to-focus correlate, since the fact that a constituent is in the PostCore slot does not automatically and always translate into it expressing one particular syntactic function or being part of one particular focus articulation; we have seen that it can be a DFE within the predicate domain of a topic-comment articulation, but also one within the larger domain of thetic articulation (leading to a presentational focus reading).

4.6.5.6 Preposing

The term “preposing” is used for constructions that have the direct or indirect object appear before the grammatical subject, resulting in an OSV word order. Present-day English preposing usually has the function of highlighting the preposed object, which must somehow be linked to the preceding context, or at least provide a value for an open proposition in the preceding context (see Birner and Ward, 1998), but this is not necessarily the case in Old English. Two preposing examples from the Euphrosyne text are these:

- (88) a. (Hi þa ealle wucan fæstan, and on heora gebedum þurhwunodon,)

ac him **nan swutelung** ne com swa him gewunelic wæs

but to.them no revelation not came as to.them usual was

þonne hi hwæs bædon. [coeuphr:225-227]

then they something prayed

'(All the week they fasted and continued to pray.)

But they did not receive any revelation as would have been customary

when they prayed for something.'
- b. Þa axode he hine hwæthis nama wære. [coeuphr:159-160]

then asked he him what his name was

Þa cwæð he, Smaragdus ic eom geciged.

then said he Smaragdus I am called

'He asked him what his name was, and the other one answered: "I am

called Smaragdus."'

The syntactic subject *nan swutelung* ‘no revelation’ is bolded in (88a), and it is preceded by the indirect object *him* ‘to them’, while the finite verb *com* ‘came’ follows. The focus articulation here is the topic-comment one, where *him*

‘them/they’ continues the topic of the preceding two clauses, while the predicate constitutes the focus domain: *nan swutelung ne com* ‘no revelation came’. Even though the surface word-order is that of a preposing construction (object-subject-finite verb), the actual “preposing” that has taken place is the rearrangement within the predicate: the discourse-new subject *nan swutelung* ‘no revelation’ has shifted from its core-internal “Sbj” slot (where it would be more in line with the Principle of Natural Information Flow) to the preverbal “PreAP” slot (in terms of the slot division in (65)).³⁰ I interpret this as signalling that it is a Dominant Focal Element (see section 3.3.3 as well as the discussion in 4.6.5.3).

The example in (88b) is preposing of the complement *Smaragdus* before the subject *ic* ‘I’ (with nice fillers of the Vb1 slot in the form of the finite verb *eom* ‘am’ and for the Vb2 slot in the form of the non-finite verb *geciged* ‘called’). This preposing clearly is a clear example of constituent focus, since the value provided by the NP *Smaragdus* fills in the open variable created by the *wh*-question in the preceding sentence.

4.6.5.7 The *it*-cleft

There is one construction that is absent in the first part of the Euphrosyne text that has been charted for this chapter, but it is present in a later part, and because of its significance in the last part of this dissertation, we will have a look at it here: it is the *it*-cleft construction (89).

- (89) a. (Wa me, hwa sceal mine ylde afrefrian, to hwam sceal ic gan þæt me fultumige? Min sar is getwyfyld.)
 Nu hitis **for eahta and þryttidan gearan** [coeuphr:283]
 now it is for eight and thirty years
 þæt min dohtor me losode,
 that my daughter to.me was.lost
 (and me nan swutelung ne com, þeh ic his geornlice gyrnde).
 ‘(Woe to me! Who will comfort me in my old age, and to whom shall I go that will help me? My sorrow is doubled.)
 It is now thirty eight years ago that I have lost my daughter,
 (and no revelation has come to me, though I have earnestly longed for it).’

The construction in (89a) is the predecessor of the Present-day English *it*-cleft, and in that sense it should be the prime candidate for a linguistic device that expresses constituent focus. However, if we look at (89a) in its context, it is hard to call it a constituent focusing device at all, because (a) the constituent that is set apart (*for 38 years*) is not an argument of the verb *losode* but a temporal adjunct, and (b) this temporal adjunct is not really contrasted with anything in the preceding or following context, so there is no *contrastive* focus, nor does the temporal phrase supply a value for an open proposition (it is not the answer to a question that has just been raised).

The cleft construction in C matches quite well on the word order structure of OE shown in (65): the subject *hit* occupies the “PreSbj” slot, the verb *is* the “Vb1” slot, and the time adverb *for eahta and þryttidan gearan* ‘for thirty eight years’ fits the

core-internal “AP” slot. The subordinate clause started by *þæt* ‘that’ maps straightforward into the word order model, just like other subordinate clauses.

It is hard to say from this one sample how much highlighting (that is: constituent focus) actually is going on here, which is one of the reasons we will have to look to more examples of this kind of construction in closer detail (see chapters 9-12).

4.7 Late Modern English narrative

The basis for determining the word order patterns in Late Modern English is formed by the first 100 sentences of the text “reeve-1777”, which is taken from the PPCMBE (Kroch et al., 2010). The text is a fiction narrative entitled “The champion of virtue: a Gothic story”, and it is written by Clara Reeve.³¹ It satisfies the narrative conditions stated in (62) since it is not a translation from another language and is a clear 3rd person narrative. The main character is a Christian knight called sir Philip Harclay, who returns from fighting for his country under king Henry V, sets out to visit one of his old friends, finds he has died, and then wants to take care of this friend’s affairs. Just as with the Euphrosyne story, this kind of text mostly contains topic-comment clauses (85%), and provides ample changes in time and location, as well as the introduction of characters and events. It shows presentational focus at work and illustrates word order variations due to syntactic, discourse and pragmatic reasons, as we will see.

4.7.1 Narrative text

What follows here are the first 100 sentences of the “reeve-1777” text, divided into major paragraphs in accordance with the findings described in section 4.7.4. The line numbering is copied from the PPCMBE version of this story.³²

[1.3] The Champion of Virtue.

[1.4] A Gothic Story.

[1.6] In the minority of Henry the Sixth, king of England, who also was crowned king of France, when the renowned John duke of Bedford was regent of France, and Humphrey the good duke of Gloucester was protector of England; a worthy knight, called sir Philip Harclay, returned from his travels, to England, his native country. [1.7] - He had served under the glorious king Henry the Fifth with distinguishing valour, [1.8] had acquired an honourable fame, [1.9] and was no less esteemed for christian virtues than for deeds of chivalry. [1.10] After the death of his prince, he entered into the service of the Greek emperor, [1.11] and distinguished his courage against the encroachments of the Saracens. [1.12] In a battle there, he took prisoner a certain gentleman, by name M. Zadisky, of Greek extraction, but brought up by a Saracen officer, [1.13] this man he converted to the christian faith, after which he bound him to himself by the ties of friendship and gratitude, and he resolved to continue with his benefactor. [2.14] After thirty years travel and warlike service, he determined to return to his native land, and to spend the remainder of his life in peace, and by devoting himself to works of piety and charity, prepare for a better state hereafter.

[2.15] This noble knight had in his early youth contracted a strict friendship with the only son of the lord Lovel, a gentleman of eminent virtues and accomplishments. [2.16] During sir Philip's residence in foreign countries, he had frequently written to his friend, [2.17] and had for a time received answers, [2.18] the last informed him of the death of the old lord Lovel, and the marriage of the young one; [2.19] but from that time he heard no more from him. [2.20] Sir Philip imputed it not to neglect or forgetfulness, but to the difficulties of intercourse, common at that time to all travellers and adventurers.

[2.21] - When he was returning home, he resolved, after looking into his family affairs, to visit the castle of Lovel, and enquire into the situation of his friend. [2.22] - He landed in Kent, attended by his Greek friend and two faithful servants, one of which was maimed by the wounds he had received in the defence of his master.

[2.23] - Sir Philip went to his family seat in Yorkshire, [2.24] he found his mother and sister were dead, and his estates sequestered in the hands of commissioners appointed by the protector. [2.25] - He was obliged to prove the reality of his claim, and the identity of his person, by the testimony of some of the old servants of his family after which every thing was restored to him. [2.26] He took possession of his own house, [2.27] established his household, [2.28] settled the old servants in their former stations, [3.29] and placed those he brought home in the upper offices of his family. [3.30] He left his friend to superintend his domestic affairs, [3.31] and attended by only one of his old servants, he set out for the castle of Lovel, in the west of England. [3.32] - They travelled by easy journeys, [3.33] but towards the evening of the second day, the servant was so ill and fatigued he could go no further, [3.34] he stopped at an inn where he grew worse every hour, [3.35] and the next day expired. [3.36] Sir Philip was under great concern for the loss of his servant, and some for himself, being alone in a strange place; [3.37] however he took courage, [3.38] ordered his servant's funeral, [3.39] attended it himself, [3.40] and having shed a tear of humanity over his grave, proceeded alone on his journey.

[3.41] As he drew near the estate of his friend, he began to enquire of every one he met, whether the lord Lovel resided at the seat of his ancestors; [3.42] he was answered by one, he did not know,- by another he could not tell,- by a third, that he never heard of such a person. [3.43] Sir Philip thought it strange that a man of lord Lovel's consequence should be unknown in his own neighbourhood, and where his ancestors had usually resided. [3.44] - He ruminated on the uncertainty of human happiness; [3.45] this world, said he, has nothing for a wise man to depend upon, [3.46] I have lost all my relations, and most of my friends; [3.47] and I am uncertain whether any are remaining. [3.48] - I will however be thankful for the blessings that are spared to me, [3.49] and I will endeavour to replace those that I have lost. [3.50] - if my friend lives he shall share my fortune while I live, [3.51] and his children shall have the reversion of it; [3.52] and I will share his comforts in return. [4.53] - But perhaps my friend may have met with troubles that have made him disgusted with the world. [4.54] Perhaps he has buried his amiable wife, or his promising children, and tired of public life, [4.55] he is retired into a monastery, [4.56] - at least I will know what all this silence means.

[4.57] When he came within a mile of the castle of Lovel, he stopped at a cottage, [4.58] and asked for a draught of water, [4.59] a peasant, master of the house brought it, [4.60] and asked if his honour would alight and take a moments refreshment. [4.61] - Sir Philip accepted his offer, being resolved to make farther enquiry before he approached the castle. [4.62] - He asked the same questions of him, that he had before to others, [4.63] which lord Lovel, said the man, does your honour enquire after? [4.64] the man whom I knew, was called Arthur, said sir Philip, [4.65] ay, said the peasant, he was the only surviving son of Richard, lord Lovel as I think? [4.66] - very true friend, he was so. [4.67] - alas sir, said the man, he is dead! [4.68] he survived his father but a short time. [4.69] - dead say you. [4.70] - how long since? [4.71] - about fifteen years to the best of my remembrance. [4.72] - sir Philip sighed deeply- [4.73] alas, said he, what do we by living long, but survive all our friends! [4.74] - but pray tell me how he died. [4.75] - I will sir to the best of my knowledge. [4.76] \$An \$'t please your honour, I heard say, that he attended the king when he went against the Welch rebels, and he left his lady big with child; [4.77] and so there was a battle fought, [4.78] and the king got the better of the rebels,- [4.79] there came first a report that none of the officers were killed, [5.80] but a few days after there came a messenger with an account very different, that several were wounded, and that the lord Lovel was slain, which sad news overset us all with sorrow, [5.81] for he was a noble gentleman, a bountiful master, and the delight of all the neighbourhood. [5.82] - He was indeed, said sir Philip, all that is amiable and good, [5.83] he was my dear and noble friend, [5.84] and I am inconsolable for his loss. [5.85] - but the unfortunate lady, what became of her? [5.86] why \$an \$'t please your honour, they said she died of grief for the loss of her husband, [5.87] but her death was kept private for a time, [5.88] and we did not know it for certain till some weeks afterwards- [5.89] The will of heaven be obeyed, said sir Philip, [5.90] but who succeeded to the title and estate? [5.91] - the next heir, said the peasant, a kinsman of the deceased, sir Walter Lovel by name. [5.92] I have seen him, said sir Philip, formerly, [5.93] but where was he when these events happened? [5.94] at the castle of Lovel, sir, [5.95] he came there on a visit to the lady, [5.96] and waited there to receive my lord, at his return from Wales; [5.97] when the news of his death arrived, sir Walter did every thing in his power to comfort her, [5.98] and some said he was to marry her, [5.99] but she refused to be comforted, [5.100] and took it so to heart that she died.

4.7.2 Pragmatically neutral word orders

The word orders that are *pragmatically* neutral in this late Modern English narrative are summarized in Table 10. The variation in the different word orders listed in this table is related to syntactic and text-organizational matters.

Table 10 Pragmatically neutral word orders in late Modern English

Variation cause	Name	Word order	Function
Syntactic	Default	(Conj) S V _f ...	Neutral
	Subclause	C (AP) S V _f ...	Complement
	V-initial	S=0 V _f ... V _{nonf} ...	Imperative mood Adverbial clause
Text	Ref PoD	S then V _f ...	Reference change
	AP-initial	[_{CP} ...] S V _f	Major section
		[_{AdvPP} ...] S V _f	Minor section
	Logical	[_{CP} L (S) V _f ...] V _f ...	Reason, purpose
	Conjunct	Conj CP (S) ... V _f ... (Conj) S=0 V _f ...	Cohesive step Cohesion

Each of the word orders above warrants further discussion and will be illustrated by examples in the following sections. While the word orders above are neutral from a *pragmatic* point of view, they are not neutral from other points of view. Just as in the case of the Old English narrative discussed in section 4.4, we will look at three of the main causes for word order variation: syntax, text-organization and pragmatics.

The word order variation should be seen against the background of the slotted clause structure that can be derived from the charting of the Reeve text as shown in Table 11 (see for comparison the OE one in Table 7).

Table 11 Division of late Modern English into slots

#	Con	PreCore	Core					PostCore
			Sbj	Vb1	Mid	Vb2	Arg	
10		after death of his prince	he	entered				into the services of the Greek emperor
11	and		—	distinguished		his courage		against the encroachments of the Saracenes
16		During sir P's residence in foreign countries	he	had	frequently	written		to his friend

The slots that have, on the basis of examining the first 100 sentences in the Reeve text, been chosen for LmodE are part of the larger PreCore-Core-PostCore division, and are the following:

(90) *Names and functions of the LmodE slots*

PreCore

Con: Conjunctions, disjunctions

PreC: Pre-core position for points of departure (PP, AP)

Core

Sbj: Core-internal subject position

Vb1: Usual place for the finite verb (alternative is Vb2)

Mid: Adverb, subject or negation between finite and non-finite verb

Vb2: Usual place for the non-finite verb (sometimes hosts finite verb)

Arg: Any argument of the main verb

AP: Adverbials, PPs

PostCore: Anything that is clearly extraposed past the core-end

There are several things that have changed from the OE slot division in (65) to the LmodE division in (90): the two subject slots in OE (the “PreC” slot and the “Sbj” slot in (65)) have combined into one “Sbj” slot in LmodE; the core-start is no longer signalled by the finite verb “Vb1” slot, but by the “Sbj” slot; some of the slots that were between “Vb1” and “Vb2” in OE have moved to the right, so that the core-end signal is much less clear in LmodE.

4.7.3 Syntactic variation in Modern English

Syntactic reasons to vary word order in the late Modern English “Reeve” text include the expression of argument structure, subordination (complementation and adjunction), tense, aspect and mood (declarative versus imperative and interrogative).

4.7.3.1 Default word order and complementation

The LmodE text teaches us that there is very little difference between the syntax of main and subordinate clauses: both have the S-V_f word order (witness the main clause in 91a, and the subclause in 91b), both optionally allow an adjunct (in the form of an AP clause) to *precede* the subject (91c-d) and both optionally allow an adjunct (in the form of an adverb) to *follow* the subject (and precede the finite verb), as in (91e-f). The only real difference is the fact that subordinate clauses have a subordinating conjunction such as *that*, *why* or *when*.

- (91) a. **Sir Philip** went to his family seat in Yorkshire. [reeve-1777:23]
 b. I heard say that **he** attended the king. [reeve-1777:76]
 c. During sir Philip’s residence in foreign countries, **he** had frequently written to his friend. [reeve-1777:16]
 d. I am thinking that tho’ young Edmund wants not my assistance at present, **he** may hereafter stand in need of my friendship. [reeve-1777:436]
 e. **He** pathetically lamented the loss of all his friends. [reeve-1777:331]
 f. ..., and sir Philip thought **he** still followed him. [reeve-1777:221]

The canonical word order for main clauses is SV, as in (91a), where a clause-initial subject is followed by a finite verb. This same word order also appears in (91b),

which is a subordinate clause. Time adverbials such as *frequently* and *hereafter* may occur immediately after the finite verb in what is left of the “Middle Field”, as in (91c) for a main clause and in (91d) for a subordinate clause. When there is no discernable Middle Field, as is the case when there only is a lexical verb and no auxiliary, then the default position for adverbials seems to be between the subject and the finite verb, as in (91e) for a main clause and (91f) for a subordinate clause. There are variations in the word order of time and place adverbials, as we will see in section 4.7.4 of this chapter, but some of these can be attributed to text-organization (they serve to indicate boundaries of developmental units in the narrative) and others result in pragmatic differences (they are highlighted when they appear clause-finally, for instance).

As in the discussion of the Old English text, syntactic variation due to negation will not be treated here, since this requires separate attention that is beyond the scope of this chapter (and this dissertation), where we only take a preliminary and cursory look at word order variation in late Modern English, in order to know whether the variation that we discuss in this and subsequent chapters is to be attributed to pragmatic factors or not.

4.7.3.2 V-initial

The Reeve text shows several different syntactic motivations for clauses to begin with a verb. Main clauses in imperative mood (92a) and polar questions in interrogative mood (92b) start with a verb (if we do not take conjunctions into consideration), both of which compare well with what was already present in Old English, except that polar questions now need *do*-support, which is not needed in Old English, since the *lexical* verb moves up front, as in (68b).

- (92) a. But pray tell me how he died. [reeve-1777:74]
 b. And does the present lord Lovel reside at the castle? [reeve-1777:101]
 c. He landed in Kent, attended by his Greek friend and two faithful servants. [reeve-1777:22]

One noteworthy syntactically motivated V-initial pattern in late Modern English is the one inside a subordinate adverbial clause like *attended by his Greed friend* in (92c). The V-initial character results from elision of the subject in the subordinate clause. This particular pattern is important enough to take note off, because it is one of the pattern(s) that has come to replace the Old English T-correlated and AP-correlated ones (see Table 6). Instances of correlation such as (74a), repeated here below for convenience, where the subject of the two clauses differs (“the child” versus “they”), occur as adverbial clauses with their own subject and finite verb in PDE. But as soon as the subject of the subordinate clause coincides with that of the main clause, a rendering in PDE is in place that uses a participle form of the verb and elides the subject, as in (92c) (subject is “he”), and as in (75b) (with the subject “Euphrosyne” expressed as “Eufrosina” and “heo”), repeated below.

- (74) a. **Þa** þæt cild wæs seofon wintre, **þa** letan hi hi fullian,
 then that child was seven winters then let they her baptize
 ‘When the child was seven years old, they had her baptized.’

- (75) b. Mid þy [þa Eufrosina þone munuc þær wiste],
 with that then Euphrosyne the monk there knew
 þa gecigde heo hine to hire
 then called she him to her
 ‘Knowing the monk was there, Euphrosyne called him to her.’

The V-initial adverbial clause pattern exemplified in (92c) and (75b) is part of the discourse-motivated AP-initial pattern that will be discussed to a fuller extent in section 4.7.4.2.

4.7.4 Discourse variation in Modern English

Just like Old English texts, Modern English texts too contain linguistic clues that are used to divide the text into larger episodes or smaller developmental units, and these clues are not necessarily associated with a difference in pragmatic meaning. The particular linguistic devices that are used in Modern English differ to a considerable degree from those used in Old English: the Old English correlative structures have completely disappeared over time, the temporal adverb *then* has only to a limited degree taken up the referential point of departure demarcating function of its Old English predecessor *þa* (when used in second position, as in section 4.6.4.1), and two different types of AP-initial clauses have entered the scene. We will have a look at each of these structures in the following subsections.

4.7.4.1 Referential point of departure

When the referential theme, the main person that is being spoken about, switches from one to another, this often signals the start of a different episode or developmental unit, and we have adopted the term “referential point of departure” in section 3.3.2 to describe the linguistic device (usually some kind of fronting) that a language uses to signal this change. What we have seen for Old English is that it uses the subject plus the temporal adverb *þa* ‘then’ in second position to signal such a referential point of departure, but the LmodE equivalent *then* is not used for this feature as much anymore. The first 100 sentences of Reeve don’t even make use of this device at all, but there are occurrences of it later on in the text, such as those shown in (93).

- (93) a. Old Wyatt: “John, do you run back and acquaint my Lord of it.”
 Philip: “Not so,” said Sir Philip; “it is now almost dark.”
 John: “’Tis no matter,” said John, “I can go it blindfold.”
 Sir Philip; **then** gave him a message to the Baron in his own name,
 acquainting him that he; would pay his; respects to him in the morning.
 John flew back the second time, and soon returned with new
 commendations from the Baron, and that he would expect him; on the
 morrow. [reeve-1777:201]

- b. The baron agreed with him in opinion, that a man was of much more service to the world who continued in it, than one who retired from it, and gave his fortunes to the church, whose servants did not always make the best use of it.
 Sir Philip **then** turned the conversation, and congratulated the baron on his hopeful family. He praised their persons and address, and warmly applauded the care he bestowed on their education. [reeve-1777:332-336]
- c. He (=the Baron) listened with pleasure to the honest approbation of a worthy heart, and enjoyed the true happiness of a parent.
 Sir Philip **then** made further enquiry concerning Edmund, whose appearance had struck him with an impression in his favour. [reeve-1777:337-339]

The first instance of “then” that could be regarded as referential point of departure device is in (93a), where it serves to transition away from the direct-speech exchanges between Old Wyatt (a peasant who has initially invited Sir Philip to stay with him), Sir Philip and John, the son of Old Wyatt, who returns from an errand. The “then” sentence transitions into a more descriptive paragraph of the narrative, which is built around the referent “Sir Philip”: even though the next sentence has “John” as subject, the use of *him* late in this next sentence refers to the paragraph’s overall referential theme, which is “Sir Philip”.

A section that is later in the story, as repeated in (93b), first has the “baron” as theme, but switches to “Sir Philip” using the temporal adverb “then” in second position. This “Sir Philip” is retained in the following sentence. The following context is shown in (93c), and it starts with an ambiguous “He”, which in a later edition of the book has been corrected to “The baron” (Reeve, 1967). The switch back to “Sir Philip” is then again made using the “then” adverb in second position, but this is the last time this device is being used to indicate a change in referential point of departure in the whole story, which consists of 751 sentences. Other uses of “then” in the story show that “then” in second position does not *have* to signal a change in referent anymore, as in (94).

- (94) a. He stopped at the place where his good servant was buried, and caused masses to be said for the repose of his soul, went home by easy journeys, without meeting any thing remarkable by the way. His family rejoiced at his return, he settled his new servant in attendance upon his person, he **then** looked round his neighbourhood for objects of his charity. When he saw merit in distress, it was his delight to raise and support it. [reeve-1777:482-488]

The example above in (94a) shows “then” in second position in a main clause, but it is not used to signal a change in referential point of departure. On the contrary, the referent stays fixed to “he” throughout the whole episode. The use of “then” in the sentence it occurs in merely helps order the events a bit more efficiently, indicating that “he” looked round his neighbourhood only *after* he had settled his new servant.

From the point of view of the slot-structure, the referential point of departure, which consists of the subject followed by *then*, resides in the “Sbj” slot (see the

overview of the slots in (90)): the slot immediately preceding the finite verb slot “Vb1”.

To sum up, it seems clear that the use of “then” in second position as a referential point of departure switching device has decreased considerably by the LmodE period, which raises the question what (if any) device has taken over. I’m afraid that the answer to this question will have to wait for future research, since the selection of the Reeve text that has been used does not offer the kind of switch in referential point of departure that is needed to figure out what linguistic device is used for the switch.

4.7.4.2 AP-initial

The first 100 sentences of the Reeve show that there are two types of AP-initial sentences (those with a structure AP-S-V_{fin}, where the “AP” comes in the pre core “PreC” slot in (90)), each fulfilling their own role in the organization of the text in units: (a) sentence initial adverbial *clauses*, and (b) sentence initial adverbial *phrases*. The distinction between them becomes evident when we look at their number of occurrences within the Reeve text sample: there are 4 adverbial *clauses* versus 20 adverbial *phrases* sentence-initially. The adverbial clauses are mainly being used to indicate the start of a larger episode, as is evident from sentences [1.6], [2.21], [3.41] and [4.57] in the narrative in section 4.7.1, while the adverbial phrases serve as starting points for smaller developmental units within these episodes.

- (95) a. **When** he was returning home, he resolved, after looking into his family affairs, to visit the castle of Lovel, and enquire into the situation of his friend. [reeve-1777:21]
 b. **After** the death of his prince, he entered into the service of the Greek emperor. [reeve-1777:10]

The AP-initial *clause* in (95a) comes at the start of an episode that runs from line [3.41] until [4.56] (fifteen sentences), while the AP-initial adverbial *phrase* in (95b) only demarcates the start of a development unit running from [1.10]-[1.11] (two sentences).

The observations about the adverbial clauses are in line with the findings of Ford (1993), who noted that initial adverbial clauses do “text-organizing work” (p.17). Initial *temporal* adverbial clauses in particular “provide temporal backgrounds for accounts, to encode new time frames ...” (Ford, 1993: 41). Ford’s description of the function of these clauses coincides with the observations that can be made from the Reeve text sample: sentence-initial temporal adverbial clauses serve as points of departure for larger discourse units (that is: episodes). The sentence-initial adverbial clauses have taken over from the “T-correlated” and “AP-correlated” ones in OE (see sections 4.6.4.4 and 4.6.4.5), clause-types that have not survived the OE period.

As for sentence-initial adverbial *phrases*, Virtanen (1992) notes that they are “crucial signals of text-strategic continuity and indicators of textual shifts, as well as points of departure for the textual unity they introduce”. The “continuity” provided by sentence-initial adverbial phrases becomes apparent when we view them as a kind of chain that form the (temporal) backbone of the text. If we look at the story’s

first episode, which consists of sentences [1.6]-[2.14], then the adverbial phrases are: *in the minority of Henry the Sixth* (1.6), *after the death of his prince* (1.10), *in a battle there* (1.12) and *after thirty years travel and warlike service* (2.14). These points in time indeed form the chain of events described in the episode. Like good points of departure they either link back into generally known history (the reign of *Henry the Sixth*), provide a development in time themselves (the *death of his prince*, and *thirty years travel and service*) or provide a point in time that is anchored to events or places mentioned in the preceding developmental unit (*in a battle there* links to the battle field that has just been mentioned).

The function of the sentence-initial adverbial phrases has, if we compare the findings for LmodE with those of OE in section 0, remained the same throughout time: they serve as points of departure for smaller developmental units. Their number, however, seems to have increased, if we compare the LmodE and OE texts with one another. The reason for this seems to be that they have taken over from the T-initial sentences, which have all but disappeared.

4.7.4.3 Logical

The “Logical” clauses combine subordinate adverbial clauses of purpose and reason, just as noted in the parallel section 4.6.4.6 on OE logical clauses, and they start with logical conjunctions such as “because”, “since” and “therefore”, which must be assigned a position in the “Con” slot of the LmodE model in (90):

- (96) a. That shall be as your honour pleases, **since** you will condescend to stay here. [reeve-1777:194]
 b. I hope no offence; the only reason of my sending was, **because** I am both unable and unworthy to entertain your honour. [reeve-1777:185-186]

The subordinate logical clauses, such as the *since*-clause in (96a) and the *because*-clause in (96b), generally appear at the *end* of the main clause they are contained in. This main-clause-final position (they must occupy the “PostCore” slot in (90), since nothing can follow them), as well as the fact that they are less concerned with the timeline but more with the logical structure of the text, both indicate that they do *not* serve as points of departure in the sense discussed in section 3.3.2. Their behaviour does not seem to differ from that of their counterparts in OE.

4.7.4.4 Conjunct

In the Old English “Euphrosyne” text, main clauses starting with a conjunction *and* had a clearly distinct word order, and they fulfilled a clear cohesive function in that they tightly knitted the clauses of one developmental unit together. By the time of late Modern English, the conjunctions of the conjunct clauses still occur in the initial slot as per the model in (90), but they no longer have a completely distinct word order pattern (witness the SV word order in 97b); the only way in which they differ from other main clauses is that *and*-initial clauses more frequently co-occur with subject-elision.

- (97) a. I have lost all my relations, and most of my friends;
 b. **and** I am uncertain whether any are remaining. [reeve-1777:46-47]

The section of the Reeve text in 4.7.1 has 15 instances of main-clause subject-elision, 10 of which are in *and*-initial clauses. Four of the main-clauses that lack the initial conjunction *and* are part of a series of main clauses, the last of which has the *and* conjunction, as exemplified in (98).

- (98) a. He_i took possession of his own house, [reeve-1777:26-29]
 b. 0_i established his household,
 c. 0_i settled the old servants in their former stations,
 d. **and** 0_i placed those he brought home in the upper offices of his family.

The first sentence (98a) of the tightly joined micro-unit has an overt subject pronoun *he* (referring to the main character, Sir Philip). The subject is elided in all three following main clauses (98b-d), but only the last one (98d) contains the conjunct *and*.

It is clear, then, that conjunct-clauses, even though they have taken over the word order of that of regular main clauses (as in 97b above), still serve to provide cohesion inside a developmental unit.

4.7.5 Focus in Modern English

We have established basic word orders in late Modern English that differ due to their syntactic function (they indicate argument structure, tense, mood, aspect or subordination) or their discourse-organization function (e.g. indicating the start of developmental units or episodes, or indicating cohesion within a developmental unit). All of these word order patterns could be regarded as pragmatically neutral in the sense that they do not necessarily signal a particular type of focus.

This section finds focus constructions by using two different methods: (a) we look in the Reeve text for deviations from the basic word order patterns as laid down in Table 10, and see where these word order patterns are used to convey focus, and (b) we look at clear cases of presentational focus or constituent focus in the Reeve text, and see by what constructions or word orders these are accompanied.

4.7.5.1 Expletive constructions

Absent from the OE text, but available in the Reeve text are main clauses that use the expletive *there*. Expletives are placeholders: the pronoun *there* syntactically functions as subject in the place of a sentence's "logical" subject, especially if the latter provides completely new information. Let us have a look at the role of the expletive *there* in this Reeve excerpt:

- (99) a. “but pray tell me how he_i died.”
 b. “I will sir to the best of my knowledge. An't please your honour, I heard say, that he_i attended the king_k when he_k went against the Welch rebels, and he_i left his lady big with child;
 c. and so **there** was a battle fought, and the king_k got the better of the rebels,-
 d. **there** came first a report that none of the officers were killed,
 e. but a few days after **there** came a messenger with an account very different, that several were wounded, and that the lord Lovel_i was slain, which sad news overset us all with sorrow.” [reeve-1777:74-80]

The exchange in (99a,b) speaks about Lord Lovel (referred to by the pronoun *he_i*), but then in (99c) a new situation is presented, and the expletive *there* is used to signal that there is discontinuity in topic; we are in a diversion to a battle led by “the king” in which Lord Lovel only played a minor role.³³ One motivation for the use of the expletive construction here is that the *battle* is discourse and hearer-new, and that LmodE does not allow such new elements to appear as syntactic subjects before the finite verb if it can avoid it. This hypothesis is confirmed by (99d), where the discourse and hearer-new subject *a report* is introduced by an expletive *there* construction. That the presentational focus articulation can be accompanied by a point of departure is shown in line (99e), which again introduces a new event with (relatively) new participants, but set apart from the previous events by a sentence-initial adverbial phrase.

A second motivation for using an expletive construction may be found in its ability to explicitly signal underspecification in the temporal or spatial setting of an event that is being reported; they set up a “new stage” that is only partly specified (Bolinger, 1977, Erteschik-Shir, 2007, Roos, 2012). A slightly different angle is presented by Biber et al. (1999), who see existential *there* as being used to “focus on the existence or occurrence of something”. The use of *there* in line (99c) could be seen as an intentional effort on the part of the writer to leave the time when this battle was fought and the place where it was fought unspecified (since it is irrelevant to the point the author wants to make anyway). And while the sequential nature of the events reported in (99c,d,e) is guaranteed by the use of the tenses, the location where these events happened are intentionally kept unspecified by the expletive. The use of *there*, then, can satisfy a number of desires that are related with syntax (canonical subject position), pragmatics (presentational focus) and text-organization (underspecification of the point of departure).

The question arises how the elements of the *there* sentences expressing presentational focus map onto the charting slots as proposed in (90). What happens in the case of (99c) is that the existential *there* occupies the “Sbj” slot, which signals the start of the core, as depicted in Table 12.

Table 12 Division of late Modern English “there” clauses into slots

#	Con	PreCore	Core						PostCore
			Sbj	Vb1	Mid	Vb2	Arg	AP	
77	and	so	there	was	a battle	fought			
79			there	came	first		a report		[79b]
80	but	a few days after	there	came			a messenger	with an account very different	[80b]

All the *there* constituents from (99c-e) are placed in the “Sbj” slot, while the “logical” subjects (those that receive a role from the main verb) appear either in the “Mid” slot (99c) or in the “Arg” slot (99d-e), depending on whether the “Vb2” slot has to be filled or not. The focus domain of *there* clauses comprises all the slots starting from “Vb1” and moving rightwards. But there is no one slot “reserved” for the most important element of the presentational focus, the (logical) subject: it can be in two different slots.

The *there* construction, then, is one linguistic device that LmodE uses to introduce clauses withthetic articulation, but here too we see that there is no exact syntax to focus mapping, and, as we will see in the next section, there is no exact “presentational focus”-to-syntax mapping too, since presentational focus may be expressed by different means.

4.7.5.2 T-initial

The Reeve text has one instance of a *then*-initial sentence of the structure *then*-V_{fin}-S (just as in OE, see section 4.6.4.2), but it does not occur in the small sample of the first 100 sentences; it occurs towards the end of the story in line 623, and is repeated with some preceding and following context in (100).

- (100) a. Upon this the cabal drew back, and mr. Wenlock protested that he meant no more than to mortify his pride, and make him know his proper station.
 b. Soon after sir Robert withdrew, and they resumed their deliberations.
 c. **Then spoke** Thomas Hewson: “There is a party to be sent out tomorrow night, to intercept a convoy of provisions for the relief of Rouen.”

[reeve-1777:619-624]

What we have in the narrative stretch in (100a-c) is three small developmental units: (100a) is set out by the adverbial phrase “Upon this”, (100b) starts with “Soon after”, and (100c) begins with “Then”. The function of the sentence-initial “then” is, at first glance, similar to that of the pragmatically neutral sentence-initial PPs (with structure PP-S-V_{fin}): it is to start a smaller developmental unit, and the events described in the unit start at a time that follows on the (reference) time of the preceding sentence (conform the findings of Thompson, 1999). Such is the function the sentence-initial *þa*, the predecessor of *then*, already fulfilled in OE. However, where the OE T-initial construction was pragmatically neutral, it seems that the LmodE is no longer: it indicates presentational focus. While presentational focus uses the syntactic subject position to introduce a *new* referent, the status of “Thomas

Hewson” deserves a bit more investigation, since he is not entirely new at this point in the story. Thomas Hewson is first mentioned in line 576 of the story, but he is not actively present until he takes turn to speak in line 623, which is the line above in (100c). Since it is only at that time that he really “enters” the mental model of the addressee, (100c) can safely be regarded as an example of presentational focus.

The mapping of the constituents in the clause in (100c) onto the LmodE slot model in (90) must be such that the adverb *then* occupies the “Sbj” slot, since the subject proper appears in the “Mid” slot. This is a remnant feature of OE, where the “PreC” slot preceding “Vb1” in (65) was used partly by the temporal adverb *þa* ‘then’, and partly by the subject.

Under the heading of T-initial clauses I would also like to regard sentences of the structure AP[time]-V_{fin}-S, which look a lot like locative inversion (but locative inversion has adverbial phrases of *location* rather than of time; see Bresnan (1994) and Salzmann (2004)). There is one such clause in the larger part of the Reeve text:

- (101) a. The whole cabal of his enemies consulted together in what manner they should vent their resentment against him, and it was agreed that they should treat him with indifference and neglect, till they should arrive in France, and when there, they should contrive to render his courage suspected, and by putting him upon some desperate enterprize, rid themselves of him for ever.
- b. **About this time** died the great duke of Bedford, to the irreparable loss of the English nation.
- c. he was succeeded by Richard Plantagenet, duke of York, as regent of France, of which great part had revolted to Charles the dauphin.
- [reeve-1777:584-587]

The preamble (101a) to the T-initial clause in (101b) talks about Edmund (using pronouns like “his” and “him”) versus his enemies. A totally new development starts in (101b) and is centered around “the great duke of Bedford”, who is referred to again in (101c) using the pronoun “he”. The new development in (101b) is accompanied by a point of departure in the form of the temporal adverbial phrase *about this time*, which establishes the time frame to that of the events described in (101a).

The AP[time]-V_{fin}-S construction, then, is of the same focus articulation as that of the *then*-V_{fin}-S construction exemplified in (100): presentational focus. Both introduce a new referent into the mental model of the addressee by using a syntactic subject that occurs after the finite verb (presumably in the “Mid” slot), and both start with a temporal point of departure (which seems to occupy the empty “Sbj” slot).

4.7.5.3 Apposition and focus

The majority pattern in LmodE is for the subject to precede the finite verb, with a notable exception formed by the expletive *there* construction discussed in the previous section, which provides a strategy to syntactically abide by the rule (the empty syntactic subject *there* appears before the finite verb), while at the same time place the discourse and hearer-new logical subject (the person or thing that has the

role of the lexical verb's agent) after the verb, in compliance with the Principle of Natural Information Flow.

Another strategy that somehow appeases the introduction of a discourse and hearer-new syntactic subject, is to put the subject before the finite verb, in its default "Sbj" slot, but have it followed by an appositive clause (which adds into the "Sbj" slot). There is one example of such a construction in the sample of the Reeve text associated with presentational focus, and there are several associated with constituent focus.

- (102) a. In the minority of Henry the Sixth, king of England, who also was crowned king of France, when the renowned John duke of Bedford was regent of France, and Humphrey the good duke of Gloucester was protector of England; **a worthy knight**, called sir Philip Harclay, returned from his travels, to England, his native country. [reeve-1777:6]
- b. (When he came within a mile of the castle of Lovel, he stopped at a cottage, and asked for a draught of water.)
A peasant, master of the house brought it, and asked if his honour would alight and take a moments refreshment. [reeve-1777:57-70]

The first example of the appositive strategy is (102a), which is the very first sentence of the whole narrative. The discourse and hearer-new participant *a worthy knight* is introduced as syntactic subject preceding the finite verb *returned*, but intervening between these two is the appositive clause *called sir Philip Harclay*.

An example associated with constituent focus is (102b), where the discourse and hearer-new referent *a peasant* is introduced sentence-initially, before the finite verb *brought*, but its occurrence is appeased by the appositive NP *master of the house*. This is an example of constituent focus, since the predicate *brought it* closely relates to *asked for ... water* in the preceding sentence, and the NP subject *a peasant* can be seen as filling in the variable generated by the implicit addressee in the preceding sentence.³⁴

A strategy for constituent focus that is often accompanied by apposition too is the "one-constituent answer": the answer to a *who*, *what*, *where* or *when* question is not a full sentence, but only a single constituent (an NP or a PP), which resolves the matter of new information appearing before or after the finite verb vacuously, since a single constituent cannot be assigned a slot in the model of (90):

- (103) a. "And who is he, said the knight?"
"One Edmund Twyford, the son of a cottager in our village."
 [reeve-1777:276-7]
- b. "But who succeeded to the title and estate?"
"The next heir, said the peasant, a kinsman of the deceased, sir Walter Lovel by name." [reeve-1777:90-91]

The answer to the *who* question in (103a) consists of just one NP *one Edmund Twyford*, but this NP has an appositive one *the son of a cottager in our village*. The answer to the *who* question in (103b) likewise only has one NP *the next heir*, but is then followed by two appositions: *a kinsman of the deceased*, and *sir Walter Lovel*

by name. This strategy effectively bypasses the tricky matter of the syntax wanting to place syntactic subjects before the finite verb, whereas the Principle of Natural Information Flow would want such completely new participants to occur later in the sentence.

4.7.5.4 Preposing

The term “preposing” is used for constructions that have the direct object (or more generally another XP) appear before the grammatical subject, and its function can be that of (contrastive) topic or of focus in Present-day English (Birner and Ward, 1998). There is one instance of preposing in the Reeve text sample, and it is shown in (104).

- (104) a. After the death of his prince, he_i entered into the service of the Greek emperor, and distinguished his_i courage against the encroachments of the Saracens. In a battle there, he_i took prisoner a certain gentleman_j, by name M. Zadisky, of Greek extraction, but brought up by a Saracen officer.
 b. **This man**_i he_i converted to the christian faith, after which he_i bound him_j to himself_i by the tyes of friendship and gratitude, and he_j resolved to continue with his_j benefactor_i. [reeve-1777:10-13]

The narrative has Lovel as its main topic, and he is referred to by the personal pronouns (indicated by the index “i”). The end of (104a) introduces “a certain gentleman”, adding an apposition to facilitate the addressee processing this new referent in his mental model of the situation. The next clause, (104b), still retains Lovel as pronominal subject, but uses a preposed demonstrative object NP *this man* to refer to the just introduced “Zadisky”.

The division of the constituents in (104b) in terms of the LmodE slotting model in (90) is not too complicated: the preposed object *this man* occurs in the “PreC” slot, the subject *he* in the “Sbj” slot and so on.

Birner and Ward’s (1998) observations about the possible functions for object preposing in PDE (the preposed object is either a contrastive topic or it has focus) do not seem to work in (104b): there is no implicit or explicit alternative person to whom the prisoner Zadisky is to be contrasted, nor is there a salient open proposition in (104a) for which *this man* in (104b) provides the value. Indeed, reading *this man* with any kind of contrast seems far-fetched: it would imply that there were numerous other people taken prisoner by Lovel, but only one of them (Zadisky) he converted to the Christian faith. Since no mention at all is being made of other prisoners, this is highly unlikely.

What we have, then, probably is a smooth and natural transition from the just introduced new referent “Zadisky”, who would be expected to be topical after his lengthy appositional introduction, back to the episode’s topic Lovel (the pronoun *he*_i). This transition is smooth and natural, because it complies with the Principle of Natural Information Flow: the referent that is still available in the “cache” of the addressee’s memory comes first, and only then is followed by the referent that is in second position in terms of saliency. Such a structure could be seen as a carry-over from OE, where the first constituent is widely used to provide a pragmatically

neutral link to the immediately preceding context (Los, 2012). The preposing found in Reeve, then, is not associated with any particular focus articulation.

4.7.5.5 Established information as DFE

The Principle of Natural Information Flow would have less established information follow upon the relatively more established information, and this can be particularly visible in a predicate that consists of multiple components in English (see 3.3.3). The sample of the Reeve text does not contain a situation where the order is changed within the predicate, violating the Principle of Natural Information Flow, but the larger Reeve text does, and one of these situations is shown in (105).

- (105) a. “And how came sir Walter to leave the seat of his ancestors?”
 b. “Why sir he married his sister to this said lord, and so he sold the castle to **him**.” [reeve-1777:105-107]

The predicate of the topic-comment articulation sentence in question is *sold the castle to him* in (105b). Both the direct object *the castle* and the prepositional object *him* are established information, but the *more* established information, the participant that is topmost in the mind of the speaker, is *him*, and it is this relatively more established information that is positioned *after* the less established *castle* (which was last mentioned in line 101). The fact that the natural word order is changed marks *him* as the Dominant Focal Element within the focus domain; it is the constituent that is most newsworthy or surprising at this point. Table 13 gives the division of the sentence’s constituents into the slots according the model of (90):

Table 13 A late Modern English dominant focal element

#	Con	PreCore	Core						PostCore
			Sbj	Vb1	Mid	Vb2	Arg	AP	
95			he _i	married			his _i sister	to this said lord _k	
96	and	so	he _i	sold			the castle	to him _k	

It is clear from the slot division in Table 13 that the constituents fit their structural slots well, and that there is no deviant word order that signals highlighting. The fact that *to him* is a DFE purely stems from the violation made to the Principle of Natural Information Flow. Had the writer written (105b’) “*and so he sold him the castle*”, there would not have been any DFE: the syntax would be satisfied, and so would the natural information flow. We see again that there is no one-to-one mapping between syntax and focus: one and the same syntactic construction (a ditransitive verb with its arguments) can be realized either with a pragmatically neutral word order, as in (105b’), or with a word order that contains a DFE, as in (105b).

Old English allowed unestablished information to become the DFE when it was moved out of the core of the clause, but this situation is less clear in Late Modern English, where the end of the core is not so clearly visible. The Reeve text sample does not offer examples of DFEs that have unestablished information.

4.7.5.6 The *it*-cleft

Just as we saw for the OE text, the 100 sentence sample of LmodE text does not contain an example of the *it*-cleft construction either, in spite of the fact that it is one of the devices that we would have expected to meet increasingly in LmodE to express constituent focus, just as it does in PDE. There is, however, one occurrence of an *it*-cleft in the larger part of the Reeve text:

- (106) a. During his sleep, many strange and incoherent dreams arose to his imagination. He thought he received a message from his friend lord Lovel, to come to him at the castle, that he stood at the gate and received him, that he strove to embrace him, but could not, but that he spoke to this effect.
 b. “Though I have been dead these fifteen years, I still command here, and none can come here without my permission,
 c. know that it is **I** that invite,
 d. and bid you welcome, the hopes of my house rest upon you.”

[reeve-1777:213-218]

The context in (106a) is the main character of the Reeve story, lord Philip, having a dream in which someone speaks to him (106b-d). The identity of the person speaking in the dream is left a bit implicit, though it can be deduced from the “15 years” reference, and the statement that the person “still” commands “here”, implying that it is someone in command of the castle 15 years ago. The speaker is asserting his authority in (106c) with an *it*-cleft construction, stating that *he* is the person who invites the dreamer, lord Lovel, which contrasts with any other potential inviters.

The mapping of this construction to the slot-structure is not too difficult: the first part of the *it*-cleft is a straight-forward copula clause with *it* in the “Sbj” slot, *is* in the “Vb1” slot, and *I* in the “Arg” slot. The second part is a subordinate clause, starting with the complementizer *that* in the “PreC” slot, after which the predicate follows (which in this case simply is *invite*). Seen from the perspective of information flow, the copula clause could be regarded as one where more established information “*I*” follows rather than precedes the less established *it*. This is, however, not entirely clear, since it is hard to speak of a “more” or “less” established state of a pronoun, if *it*, in fact, does not need any establishment at all—it simply cannot refer to anything.

The fact that this construction occurs only once in this whole text makes it difficult to make generalizations about it, which is one of the reasons we will look at it in more detail in chapters 9-12.

4.8 Discussion

I started out in this chapter by distinguishing three factors that can contribute to the use of different word orders: syntax, information structure and text-organization. In order to work towards an honest answer to the research question in (11), which is about the *relation* between syntax and focus, I adopted the working hypothesis that these three factors are independent, and I also presented the slot-structure approach

as a theory neutral method to chart the variation in word order we find in the history of English. After discussing several relevant results on Old English syntax and focus (0), I highlighted the decrease of subject-auxiliary inversion, one of the syntactic changes that took place in English and that is relevant for the change in focus (4.3). I then gave a brief preview on the changes in the expression of focus that play a role in this study (4.4). With these fundamental issues settled, I started introducing the text-charting approach used in the remainder of this chapter (4.5). Automatic charting of several selected texts from different periods gave another viewpoint into the changing word order patterns in English: (a) the reduction of the number of positions available in the “Core” area, and (b) the related disappearance of a dedicated slot for subjects in the “Core” area.

The remainder of this chapter on narrative text word orders is an attempt to identify pragmatically marked word orders and patterns in Old English and Late Modern English, by using an in-depth analysis of one narrative from each of the time periods. The OE text of Saint Euphrosyne reveals several word orders and devices used for text organization, some of which apparently remained constant throughout the further development of English, witness their occurrence in the LmodE text “The champion of virtue”: a temporal adverb like “then” in second position has remained a signal for a referential point of departure, although LmodE seems to use temporal prepositional phrases for this purpose more frequently; the start of larger episodes is signalled by T-correlated (4.6.4.4) and AP-correlated (4.6.4.5) constructions in OE, and taken over by sentence-initial adverbial clauses (4.7.4.2) in LmodE; the start of smaller developmental units is marked mainly by T-initial (4.6.4.2) constructions in OE, but the emerging AP-initial ones in OE (0) take on this role completely when we reach LmodE (4.7.4.2); cohesion within developmental units is signalled by Conjunct clauses (4.6.4.7) in OE, and though the syntax of these clauses changes, they retain this function in LmodE (4.7.4.4).

The relation between syntax and word order has changed fundamentally from OE to LmodE, a matter that is visible in the changes in the slot models: where OE could host the subject in *two* different slots, LmodE almost exclusively has *one* slot left for it; the core-start is marked by the “Vb1” slot in OE, but by the “Sbj” slot in LmodE, which means that the subject can appear almost nowhere else; where OE had a clear core-end marking in the Vb2 slot, such clarity has greatly decreased in LmodE (it is sometimes hard to know where the core ends); the complementizer, the functional element introducing a subordinate (complement) clause, resides in the Vb1 slot in OE, but has moved leftward into the “PreC” slot by LmodE.

When it comes to focus constructions, it is the introduction of hearer-new subjects as part of presentational focus that appears to be most challenging, calling for creative solutions. OE uses split constituents (4.6.5.1) as one device that allows fulfilment of two demands: (a) have one part of the NP as syntactic subject appear before the finite verb (which seems to be the canonical position, even in OE), and (b) position the other part of the NP as close to the end as possible, where new information is expected according to the Principle of Natural Information Flow. When clauses can be started with a good point of departure, it is possible to use the “PostCore” slot for presentational focus (4.6.5.3). The reverse, however, is not true:

that a constituent appears in the PostCore slot is *not* always an unambiguous indication of it having either presentational or constituent focus (it may just be a DFE, which is part of a larger focus domain; see 4.1.5.3 for examples). LmodE has switched to the strategy of expletives (4.7.5.1) to indicate presentational focus, which allows to (a) have a syntactic subject (the expletive pronoun *there*) appear before the finite verb, (b) have the NP of the logical subject, which contains the new information, follow the finite verb, (c) explicitly signal that the point of departure is unspecified for time and/or place. The postverbal subject of the presentational focus may end up either in the “Mid” slot or in the “Arg” slot, which means that the syntactic strategy for this construction is not completely fixed.

A feature that slightly overlaps with split constituents, and that is also associated with the introduction of hearer-new participants is apposition: this is used to a limited extent in the introduction of new subjects in OE (4.6.5.2), but its use has increase over time, so that we find it much more frequent in LmodE (4.7.5.3). Since apposition is associated with hearer-new participants in general, it can accompany presentational focus, where we have a new subject, but also constituent focus. LmodE uses single-NP-constituent constructions with appositives as answers to *wh* questions, so as a method to convey constituent focus.

There are two articulations that have a focus domain spanning multiple constituents: topic comment, where the domain equals the predicate, and presentational focus, where the domain consists of the subject and the predicate. One of the constituents within the topic-comment or the presentational focus can receive special highlighting and function as the Dominant Focal Element. Both OE as well as LmodE allow relatively more established information to be postposed after relatively less established information within the predicate of a topic-comment articulation, thereby overruling the Principle of Natural Information Flow, which is a mark to the addressee that the constituent in question is a DFE. The OE text also shows instances of DFEs with relatively *less* established information being situated past the core of the clause (either as part of a topic-comment articulation or as part of presentational focus), but similar DFEs are not noted in the LmodE text.

The two texts hesitantly show the changes in the *it*-cleft construction that will be dealt with in-depth later on (see chapters 10-12). The *it*-cleft construction as such is already present in OE (section 4.6.5.7), but does not seem to function as a constituent focusing device yet. The one example from LmodE (section 4.7.5.6) shows that it *does* fulfil that function by then, but that is as much as we can say about it from these two texts.

In fact, the observations on the *it*-cleft are indicative of what the single-text-comparison approach is giving us: we end up only having a few examples, we are able to see how these function in the wider context of a text, but we are probably not seeing the wider picture of the language in its transition stages. If we want to get such a broader picture of what takes place, we need to look at much more data, but, given the constraints on time, this means that we cannot look with as much detail as done in this chapter.

The approach of the remainder of this dissertation, then, will be a two-step one: (a) find ways to recognize sentences with presentational focus and constituent focus,

and (b) see how the means of conveying these two focus articulations have changed over time. The approach we will take to recognize focus articulations is to (i) annotate constituents for (relative) newness (chapters 5-6), (ii) develop a method to assign constituents to focus domains based on this new annotation and on syntax (chapter 7), and then (iii) implement this method for presentational focus (chapter 8) as well as for constituent focus (chapter 9).

¹ The examples in (45b) is constituent focus, since *leo* ‘lion’ contrasts with *lamb* ‘lamb’ in the immediately preceding clause, that runs like this: *He is lamb gehaten for þære unscæddignysse lambes gecyndes. & wæs unscyldig for ure alysednysse. his fæder liflic on sægednys. on lambes wisan geoffrod* ‘He is called lamb because of the innocent nature of a lamb, and he was, though innocent, for the benefit of our redemption by His Father sacrificed as a physical offering, like a lamb.’

² The inversion that occurs after clause-initial *þa* ‘then’ behaves as the syntactic inversion type. It will be discussed later in 4.2.2 since it is mainly used pragmatically for text-structuring purposes.

³ A key assumption in this 3D-approach is the definition of syntax in section 1.1 of chapter 1, which assigns a more confined role to syntax than Chomsky (1957) does.

⁴ Assuming only *three* axes is a simplification; syntax, for instance, may itself be thought of as having more axes: one for the mood, one for the tense, one for aspect, one for verb frame and so on. Van Kemenade & Westergaard (2012), for instance, provide a detailed analysis of changes in Middle English, noting that unaccusative verbs behave different than unergative ones. There are other influences on word order that are beyond the scope of this book, such as semantics (word order to help express scope), lexis (including fixed expressions) and constituent weight.

⁵ A polar question as in (5b) also leads to subject-auxiliary inversion, but I leave it out of the discussion here, since there is no XP triggering this inversion.

⁶ The pragmatic function of XVS constructions, where the S is clause-final, is now expressed by locative inversion constructions. Such constructions do not show verb movement but lead to V_{finite}-Subject order because the subject stays low in the structure (see the discussion on “late subjects” in section 4.2.5, and chapter 8 on presentational focus).

⁷ Subjects that occur before the finite verb stay there, while subjects occurring after the finite verb may end up into the “majority” subject slot if that slot has been determined to appear after the finite verb.

⁸ The applications of charting do not even stop here. Clark (2012), for instance, follows Dooley & Levinsohn’s (2001) method of participant tracking in charted texts to determine what strategies different languages use to keep track of and switch between participants.

⁹ The chart of the text is available on the author’s website:
<http://erwinkomen.ruhosting.nl/phd>.

¹⁰ Overlap between syntax and pragmatics is possible, such as when interrogative mood is used instead of declarative mood as a topic-setting device, or when subordination is used to signal backgrounding. Some research (Tomlin, 1985) sees a clear correlation between main clause and foregrounding on the one hand and subordinate clause and backgrounding on the other hand. Other research does not arrive at such a strong division (Thompson, 1987).

¹¹ The story has Euphrosyne shaving her hair and redressing in order to enter a male minster as a monk.

¹² Indeed, any of the permutations between [S, V_{fin}, AP] allows an addressee to figure out what the subject, finite verb and predicate are. OE only uses two of these permutations: the [S V_{fin} AP] one and the [AP V_{fin} S] one. This latter would convey presentational focus. The reason the other permutations are not used is the verb-second rule that is part of OE word order formation, requiring the finite verb to be in the second position of the main clause.

¹³ The six main clauses with their subjects are: 3^a (subject is *sum wer* ‘some man’), 3^b (subject is *se* ‘that one’), 5^a (subject *seo* ‘that one’), 13^a (subject *þæs mynstres fæder* ‘father of the minster’), 33^b (subject *se* ‘that one’), 38^b (*ic* ‘I’).

¹⁴ Fischer et al. (2000: 62) report one more subclause word order, where the verb immediately follows the subject, but this order results from negation, which I leave out of the current discussion.

¹⁵ Generalists would argue that the conjunction occupies the C⁰ position, which explains the reason for an initial verb to be impossible, since initial verbs would otherwise be found in the C⁰ position.

¹⁶ The term “point of departure” first comes from Weil (1844), and compares with, but is not necessarily the same as “scene-setting” or “topic” (Lambrecht, 1994: 118).

¹⁷ The subject-*þa* order is not always preceded by a conjunction *and*, witness lines 53, 62.

¹⁸ The functions of the episodes as shown in Table 8 read like a plot synopsis, which is probably the result of my own summarizations of the episodes. Brinton (1990) investigated the function of the Middle English discourse particle *gan*, and found that the actions correlated with this particle *do* in fact correlate to a plot synopsis. More research would be needed to see if the actions that correlate with the T-initial, T-correlated and PP-correlated clauses have the same effect.

¹⁹ The whole of Euphrosyne has 5 instances of a T-correlated clause where the initial subclause has [*þa*-S-V_{fin}...] word order (23, 48, 58, 193, 336), while there are 6 instances where the initial subclause has [*þa*-S-O/PP-...V_{fin}] order (71, 99, 221, 229, 245, 313), which is more what we would expect in a subclause. It is obvious that this matter needs much more investigation, but since it is clearly outside the scope of this current study, I leave it for future research.

²⁰ The *gif* ‘if’ and *swa* ‘like/as’ words are treated as prepositions in the parsed English corpora, and these prepositions take a clause as complement.

²¹ A generative description for this pattern could, perhaps, be the following. The subordinator (be it a complex adverbial like *forþam* ‘because’ or a simpler like *gif* ‘if’) occupies the specifier of the CP, and there is an invisible subordinating complementizer C⁰. The finite verb is still attracted to occur as high in the hierarchy as possible: it would like to go to C⁰, but this being blocked, it stays in I⁰. The subject is in Spec,IP.

²² Conjunct clauses with an elided subject are for instance the following lines in the first part of the Euphrosyne text: 3c, 10a, 15a, 18a, 20a, 24a, 25c.

²³ The question why conjunct clauses of this type pattern after subordinate clauses is difficult to answer. The verb-final pattern can, in generative terms, be said to arise due to the failure of the finite verb moving to the head of the CP (in complement clauses it is the *that* complement that moves to the CP head). However, there is no obvious reason (at least in a generative account) why conjunctions like *and* or *but* would block verb movement by occurring as specifier or head of the CP.

²⁴ This situation changes gradually towards Present-day English, where subclauses (like main clause) have the unmarked SVO word order. One of the pressures for this change may have come from the fact that objects usually contain new information, the significance of which is indicated by moving them out of the core into the PostCore slot as dominant focal elements. When this becomes norm rather than exception other methods had to be sought to put an additional nuance of emphasis on a new-information object when it is part of the predicate (but see the lack of such new-information DFE examples for late Modern English in 4.7.5).

²⁵ There is no reason to argue for a separate placement of the pronoun *him* near to the finite verb, since there are enough occasions (see for example lines 28b, 36a, 47c in the chart) where a less established subject (in the form of a lexical NP) follows the finite verb, and is only then followed by a direct or indirect object pronoun.

²⁶ The third occasion of a split PP is in (i):

- (i) (Ða gearn Agapitus pyder, and he Smaragdum forðferendne geseah,
and Pafnuntium samcwicne on eorðan licgan.)
Ða wearp he him **wæter** on. [coeuphr:316]
then threw he him.DAT water on
'(Then Agapitus_i run there, and saw that Smaragdus was dying, while
Paphnutius_j was half alive, lying on the ground.)
Then he_i threw **water** onto him_j.'

The focus articulation of (i) is topic-comment, with *the* 'Agapitos' being the topic, and the new information is that Agapitos throws water onto Paphnutius. The splitting of the PP *on him* 'onto him' seems to be motivated by: (a) the principle of natural information flow (have the established constituents *he* and *him* early on), and (b) the problem that a word order like *Then threw he onto him water* is infelicitous.

²⁷ The emphasis is, strictly speaking, on the word *any* in the subject, but English does not allow the focus domain to be smaller than a syntactic constituent, so that the whole subject NP constitutes the domain.

²⁸ On a par with his predecessors, Heimerdinger regards *any* constituent that occurs at the right edge of the predicate as a "Dominant Focal Element". Levinsohn (2009) deviates from Heimerdinger, regarding it unlikely that a re-shuffling of constituents that is in accordance with the Principle of Natural Information Flow is to be interpreted as highlighting, and I agree with him.

²⁹ There is no *overt* indicator of the core start, but the combination *heo þa dæghwamllice* 'she, then, daily' must be placed before the core, so that the objects and the finite verb are part of the core proper.

³⁰ An alternative view would be to say that the finite verb *ne com* 'did not come' is in the "Vb2" slot. The object *him* would then be in the Core-internal slot CoreArgEst for established arguments, and the subject *nan swutelung* in slot CoreArgNest for the non-established arguments. The addressee would then note the movement of the subject from the CoreSbj slot to the CoreArgNest slot, and this core-internal movement would then be perceived as a signal that the subject is a DFE.

³¹ The term "Gothic" refers to a fiction genre.

³² The chart of the text is available on the author's website:
<http://erwinkomen.ruhosting.nl/phd>.

³³ This particular sentence from the Reeve text could be regarded as an example of the so-called "transitive expletive construction", where an expletive *there* is used in combination

with an otherwise transitive lexical verb. I hesitate to actually label it as such, since the transitive expletive constructions noted in the literature all have an overt subject as well as object, but here the subject has been left out, since the lexical verb has been put in the passive (comparable to a Present-day Dutch rendering *er werd een oorlog gevoerd waarin de koning de overhand op de rebellen behaalde*), probably to avoid going into details about the details of the war, since these are irrelevant to the discussion here. Transitive expletive constructions are reported to have lived a short life, disappearing by the end of early Modern English (Links, 2010).

³⁴ This analysis assumes that the verb phrase *brought it* as a whole is so predictable from the context (it is the logical result of someone *asking for it*), that it is not part of the focus domain.

In the search for how focus has changed in the English language, we divided clauses in three focus articulations (chapter 3), and we looked at the realization of two of these (presentational focus and constituent focus) in an Old English and late Modern English text (chapter 4). But it was noted at the end of this last chapter, that a detailed examination of individual texts may not give us the generalizations we are looking for. With this in mind, the current chapter is the start of a corpus approach: (i) we enrich existing corpora with the minimal amount of information needed to derive the focus domain in each clause (chapters 5-6), (ii) we use corpus queries to look for instances of the focus articulations we are interested in (chapters 7-9), and (iii) we draw generalizations from the results we get (chapter 13).

This chapter is the first of the two chapters needed for step (i): we define a set of referential state primitives (labels that serve as a basis for informational categories) with which we can label individual constituents. The labels for the referential states, *combined* with the syntactic information that is included in the parsed English corpora, serve as the basis to determine focus domains, focus articulations and focus types. Once we have labelled several texts with these referential state primitives (by a method described in chapter 6), we are ready to proceed with step (ii). The actual corpus research into English presentational and constituent focus starts in chapters 8 and 9.

5.1 Criteria for referential state primitives

The introduction to this chapter argues for a relation between focus and referential state: such referential state indicators should, if combined with syntactic information, help determine the focus domain. Crucial in this hypothesis is the “referential state”, which should be understood as the way a constituent’s information relates to the mental model (see chapter 2). I would like to illustrate this notion of “referential state” by briefly looking at the noun phrases in (107).

- (107) (The syntactic study of cleft structures is widely assumed to have originated in Jespersen’s work on English, ...)
 [_{NP} **It**] seems to be [_{NP} **a lesser known fact**], however, that
 [_{NP} **counterparts of** [_{NP} *it*-clefts in [_{NP} **Romance languages**]]] had already
 been identified long before [_{NP} **Jespersen’s descriptions**] first appeared in
 [_{NP} **print**]. (Dufter, 2009: 83)

The NP *a lesser known fact* is relatively new information for the reader’s mental model, while the pronoun *it* does not refer back to an antecedent, but only functions as placeholder for the main-clause-final constituent. The first NP in the subordinate clause, *counterparts of it-clefts in Romance languages*, is as such new to the text, but builds on the phrase “cleft structures” that is mentioned in the preceding context.

This large first NP consists of several smaller ones, and the smallest of them, *Romance languages*, could hardly be labelled as “new” to the reader—here the author assumes that this entity is already known to the reader from wider world knowledge: it is already present in the reader’s long term memory (see 2.3.2). The NP *Jespersen’s descriptions* as a whole is likewise new to the text, but is linked to “Jespersen’s work” in the preceding context.

Given the different types of information the NPs represent in the example, in particular the different ways these types of information interact with a reader’s mental model, we can conclude that simple notions such as “new” and “given” are not precise enough to express the constituent’s referential state variation that can be observed. The main goal of this section on “Referential state”, then, is to look for a concise and sufficient set of referential states, which serve as primitive building blocks that we can use to derive “higher order” information status notions, such as “(aboutness) topic” and “focus (domain)”.

The need to know the referential states of constituents ties in with the overall aims of this study, which concentrates on the relation between syntax rules and focus rules. It is focus rules where information states will be shown to play an important role. (I use the term “rules” here to refer to the regularities observed in the language as used by native speakers.)

The taxonomy of referential states we arrive at in this chapter should be a set of “primitives” by the criteria stated in (108).

- (108) a. The referential states do not overlap with syntax.
 b. The referential states do not overlap with one another.
 c. The referential states are sufficient in the sense that combining constituents according to their syntax and referential states allows all relevant information state distinctions to be made.¹

The criteria above are important to keep in mind once we start to evaluate existing information state taxonomies. The criterion that referential states should not overlap with syntax in (108a) excludes from the realm of “primitives” taxonomies that define the cognitive status of a referent depending on the form of the noun phrase with which they are referred to, since the noun phrase form is syntactic information. The non-overlapping criterion in (108b) is needed for any set to be called true primitives, and the criterion in (108c) can also be thought of as a general property of primitives: one uses primitives (e.g. atoms) as building blocks to arrive at larger and more meaningful structures (e.g. molecules). This last criterion is also intended to keep the number of referential states to a minimum: only those that are needed to make relevant distinction on the information structure level should be accepted.

The set of concise yet sufficient referential state primitives we arrive at can serve as labels with which we enrich existing syntactically parsed corpora, a process that is described in detail in chapter 6. It is the combination of syntax and referential states that combine into the recognition of focus domains, which, in turn, allow us to *quantify* changes in the relation between syntax and focus rules.

The next section evaluates existing taxonomies of information state categories, paving the way for the introduction of the set of referential state primitives defined in section 5.3.

5.2 Existing taxonomies

As our aims are very practical – how to enrich the largest number of texts in a consistent manner, with the smallest expenditure of time and effort –, we will only review existing information state taxonomies here that, like us, are trying to identify a set of primitives. This means we will not consider taxonomies that seek to establish a large and fine-grained set of information states, such as Riester et al. (2010). Riester et al. combine referential information with semantic and other information, whereas we aim for a set of primitives based on referential status alone. Neither will we discuss taxonomies that annotate constituents with information status like “focus” or “topic” (Götze et al., 2007) – this information presupposes too much, and the risk of low interrater agreement is too high. Our hope is that categories such as “focus” and “topic” can be *derived* from the combination of referential and syntactic information. Our practical purpose is to look for a set of *primitives*, which are, in essence, undervivable (see the criteria in 108).

5.2.1 A taxonomy of given and new

One of the first to propose a set of information state descriptors that is more fine-grained than “given” versus “new” was Prince (1981). She came up with the taxonomy of information states for written communication shown in Figure 5.

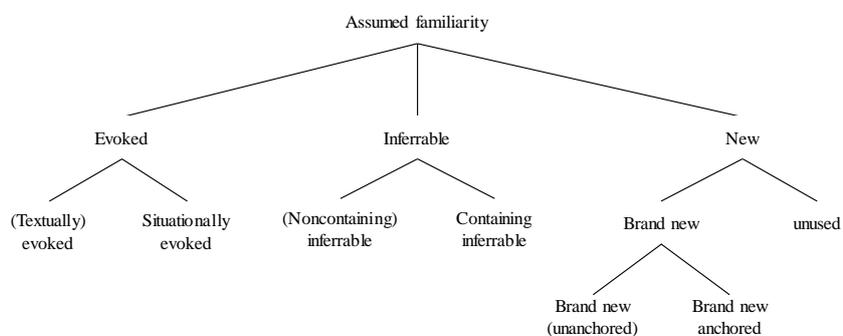


Figure 5 Prince's (1981) taxonomy of given-new information

At the first level Prince distinguishes *Evoked*, *Inferable* and *New*. A referent is called *Evoked* if it already is present in the mind of the hearer. Prince distinguishes between *textually evoked* items, which are those that have occurred in the text before arriving at the current point, and *situationally evoked* items, which are present in the extra-textual context. An example of a textually evoked item would be the pronoun *she* in (109g), which refers back to *my mother* in (109f). A situationally evoked item would be a reference of the author to himself in a text, such as the use of the 1st person singular pronoun “I” in (109a).

- (109) a. [_{NP} **I**] am the second son of [_{NP} **a family of eight**], - six sons and two daughters, -
 b. and was born on December 6, 1824, at [_{NP} **Plymouth**], where [_{NP} **my**] father and mother were on [_{NP} **a visit**] after one of [_{NP} **his voyages to India**].
 c. My father was one of three sons of Captain J. Fayrer:
 d. [_{NP} **the eldest**] was the Rev. Joseph Fayrer, rector of St Teath, Cornwall;
 e. the third, Edward, a midshipman in [_{NP} **the navy**], was drowned when H. M. S. Defence foundered, with all hands, in a gale of [_{NP} **wind**] in the Baltic in 1811.
 f. My mother was [_{NP} the only daughter of a Lancashire gentleman named Wilkinson]:
 g. [_{NP} **she**] was descended on [_{NP} **the female side**] from John Copeland, who took David, King of Scots, prisoner at [_{NP} **the battle of Neville's Cross**].
[fayrer-1900:7-13]

The *Inferrable* referents are not literally available in the preceding context, but can somehow be derived from it. An example of inferrable items would be *the eldest* in (109b) as well as *the third* in (109e), both pointing to *three sons of Captain J. Fayrer* in (109c). The referents here (the two different sons) are not identical to their antecedent (the captain), but they stand in a clear relation to one another—they are all part of the larger set of *sons* of their father.

A subtype of inferrable referents are the *containing inferrables*, which are referents that can be inferred from information within the same NP they occur in. The NP *the battle of Neville's Cross* in (109g) is an example, since the head noun *battle* is one of the possible inferences that can be made from the location *Neville Cross*.

The class of *new* referents refers to those entities which are not referred to in the preceding discourse context, and which cannot be inferred from the preceding context. This class is divided into subclasses. Prince calls a referent *unused*, if it is not in the text as such, but the author assumes it is known to the reader. Examples would be names of places (such as Plymouth and India in (109b)) and people (such as John Copeland in (109g)), but also references to generally known entities such as *the sun* or *the moon*.

Brand new referents are those that are not available in the prior context, nor does the author assume his readers to know them. The class of *brand new unanchored* referents are those that are indisputably new, such as *a family of eight* in (109a), *a visit* in (109b) and *a gale of wind* in (109e).

The mention of *his voyages to India* in (109b) is new in the text, but it is “anchored” to established information through the possessive pronoun *his*, which points to the *father* mentioned earlier in the sentence. Prince labels this as *brand new anchored* information.

Prince (2002) demonstrates how her taxonomy can be applied to annotate noun phrases. She clarifies her categories by introducing the distinction between

discourse-old and discourse-new, where the latter divides into hearer-old and hearer-new, as in Figure 2.

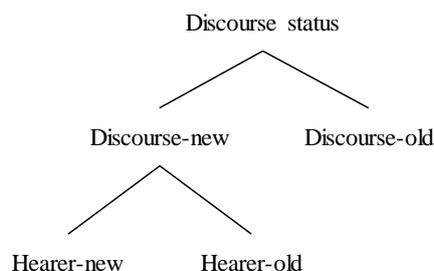


Figure 6 Information states based on discourse and hearer

From the point of view of the practical criteria stated in (108), some problems should be noted. The referent type of *wind* in line (109e) could, by Prince's system, be identified as *unused* or *brand new unanchored*. This, we would argue, misses a generalisation that cannot easily be derived from the syntax of the NP *wind*. The observation is that an NP like *wind* is inert to a referencing system. It cannot refer back to something in prior discourse or to any specific referent in the extralinguistic world, for that matter. It can also not serve as antecedent for following items—nothing in the following text can refer back to *wind*. We will come back to referentially “Inert” entities in section 5.3.4.

Although Prince's taxonomy is well-thought out and well-founded, it contains redundancies which we would like to exclude from our set, given our criteria in (108a-c). The category of “containing inferrables” is derivable from syntax: definite NPs that contain a postmodification can be classified as containing inferrables. Another redundancy is in the categories *brand new anchored* and *brand new unanchored*. The difference between them lies in the presence of an anchor within an NP. Such an anchor can be deduced from syntax (it is a constituent within the NP), and from its information state (which is *textually evoked*), and hence should be excluded from our set of primitives.

5.2.2 The topic acceptability scale

As a potential alternative to Prince's taxonomy we could consider Lambrecht's (1994: 165) *topic acceptability scale*. This scale is a condensed version of Prince's (1981) taxonomy, containing only 5 instead of Prince's 7 categories, and is illustrated in Figure 7.

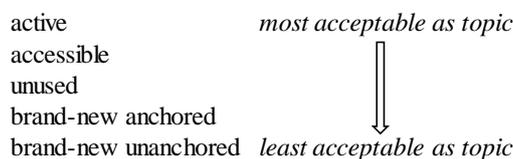


Figure 7 The topic acceptability scale (Lambrecht, 1994)

Lambrecht's scale contains a set of information statuses that condense those suggested by Prince in a clear way. His status of *active* is a subset of Prince's status *textually evoked*. A referent is "active" if it is textually evoked and its antecedent is nearby. Its status of *accessible* can be subdivided into *textually accessible* (those are *textually evoked* entities with more distant antecedents), *situationally accessible* (*situationally evoked* in Prince's taxonomy), and *inferentially accessible* (the *inferrables* in Prince's taxonomy).

Lambrecht's reduced set of categories retains the difference between *brand new anchored* and *unanchored*, which we argued earlier to be redundant for our purposes. Note also that Lambrecht's set splits up Prince's *textually evoked* into *active* and *accessible* on the basis of the distance to their antecedents as well as the existence or absence of intervening references to other entities in the text. The set of categories we are looking for would be more in line with Prince here: we would simply assign the category of *textually evoked* to an NP, and rely on information from other levels (syntax) to derive the activation state.

The purpose of Lambrecht's topic acceptability scale does not, in fact, coincide with our purposes, since it is: "measuring the degree of pragmatic well-formedness of a sentence containing a topic expression". From that perspective, activation states are important because entities that are active make the best topics. Such topics refer to items that have already been introduced and are in fact being talked about in the immediately preceding context. *Accessible* topics are also available, but in the wider context. We will consider other information state taxonomies based on activation state in the next section.

5.2.3 The givenness hierarchy

Following Chafe's (1976) seminal paper that launched a cognitive theory on how to distinguish degrees of givenness, Yule (1981) started extending Chafe's original set of "new" versus "given" by dividing "given" into "current non-new" and "displaced non-new"—labels that refer to entities that have been mentioned once before ("current") and more than once before ("displaced"). He found a correlation between the form of the referring expression (e.g. indefinite NP, definite NP, pronoun) and the level of givenness of the referent.

This three-level correlation was extended to an "accessibility marking hierarchy" with some fifteen levels by Ariel (1999), in which the various expressions of the NP were ranged on an accessibility scale. This is shown in Figure 8.

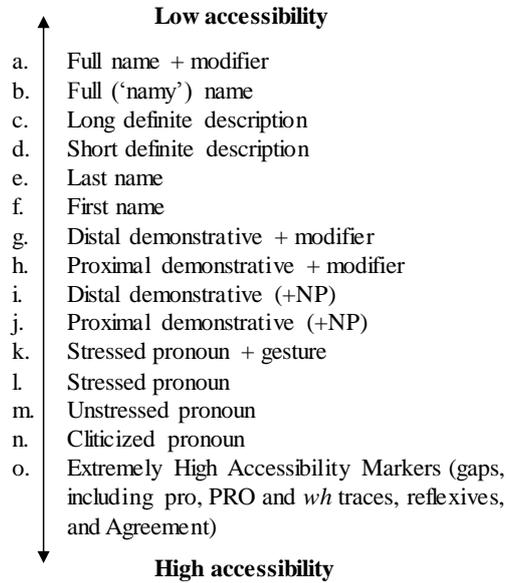


Figure 8 Ariel's accessibility marking scale (Ariel, 1999)

While Ariel posited this hierarchy in terms of noun phrase types, she did not name or distinguish all the information states that should correspond to these noun phrase types. It were Gundel, Hedberg and Zacharsky (1993) who introduced a theory that distinguishes six "cognitive statuses", each of which corresponds to a subset of noun phrase types, as illustrated in Figure 9.

In focus	<i>it</i>	
Activated	<i>that, this, this N</i>	
Familiar	<i>that N</i>	
Uniquely identifiable	<i>the N</i>	
Referential	<i>indefinite this N</i>	
Type identifiable	<i>a N</i>	

Figure 9 The givenness hierarchy (Gundel et al., 1993)

By selecting a particular kind of referring expression, the speaker signals that the referent has a particular cognitive status. Note that *activated* also entails that all statuses that are lower on the hierarchy are met. The hierarchy entails that as soon as an entity is found to be, for example, *activated*, it also has the status *familiar* and lower statuses. But the reverse is not necessarily true—an entity with status *activated* is not necessarily also *in focus*.

The relationship between accessibility status and NP type helps us to match syntactic expression with cognitive status, which in turn allows us to avoid postulating redundant categories in our set of referential state primitives.

The highest cognitive status is that of *in focus*. This label has nothing to do with the term “focus” as it is used in information structure. A referent has the status of *in focus* if it is the current center of attention. Such a center of attention roughly corresponds to what is usually identified as “topic”. A pronominal subject such as *he* in (110b) is one example, as is an ellipited subject (a zero pronominal), as in *finished the sloop* in (110b).

- (110) a. I obliged him to set up the sloop which I had brought with me from England, as I have said, for the use of my colony, in order to send the refreshments I intended to my plantation. [defoe-1719:93]
 b. Accordingly he got hands, and finished the sloop in a very few days. [defoe-1719:94]

The status of *activated* is assigned to a referent that is available in short-term memory.² This includes, for example, the *uniquely identifiable* expressions defined below, but it also includes the communication participants (references to the author, such as “I”, and to the addressee, such as “you” or “we”).

One step further along the hierarchy, an item has the cognitive status of *familiar* if the author is not only able to identify the referent from the expression, but if this referent is already available in the memory of the addressee. The pronoun *he* and the expression *the sloop* in (110b) both refer to entities that have already been mentioned in (110a).

An item is *uniquely identifiable* if it not only has a unique referent, but the addressee is able to identify the referent on the basis of the linguistic expression alone. The identification may be made from the preceding linguistic context, from the extra-linguistic context, or from world knowledge. A noun phrase such as *England*, but also the complex noun phrase *the sloop which I had brought with me from England* in (110a) are examples of expressions with a cognitive status of *uniquely identifiable*.

An item is *referential* if the addressee is not only able to think of an example of the object described, but if one *particular* referent is intended to which the speaker is going to refer. Gundel et al. use the example of *this dog* in the sentence “This dog kept me awake”. The reader understands that *this dog* refers to a particular dog, but he is unable to determine the identity of that dog. The expression *my colony* in (110a) is another example, since it refers to a particular “colony”, but the identity of “my colony” cannot be *uniquely* determined, since the expression itself does not give enough information to identify the colony the author is writing about. By using a *referential* expression the author introduces or maintains a specific referent, which can function as the theme for the following sentences, and which can be referred to.

An item is *type identifiable* if the addressee (a hearer or reader) is able to think of a specific example of the object that is described by the expression used. The expression *hands* in (110b) is an example of an item with this cognitive status. Gundel et al (1993) classify all indefinite NPs in English as type identifiable.

Note that the identification of this set of cognitive statuses relies on information from the syntax in that cognitive status correlates with the form of the NP (pronoun,

definite/indefinite NP, etc.). This violates the criterion specified in (108a), which means that we are not able to use the givenness hierarchy as a set of referential state primitives. We need a minimal set of information status categories that is independent of the *forms* of the referring expressions and may serve as the basis from which the cognitive statuses in the *givenness hierarchy* can be derived. The precise mapping from this set of primitives to this *givenness hierarchy* in all likelihood depends on the language-specific forms of referring expressions.

In our search for a set of referential state primitives, we now turn away from approaches where constituents only receive an information state category to those where constituents also receive a link to their antecedent—if they have one.

5.2.4 Coreference resolution

Our own project group in Nijmegen initially set out to perform *coreference resolution* on the parsed English corpora manually, and make use of a limited set of *coreference types*, which are type labels that are only used for noun phrases that have an antecedent. The process of coreference resolution evolved as a generalization of the specific task of pronoun anaphor resolution in computational linguistics (Hobbs, 1978, Soon et al., 2001). We will turn to computational linguistic approaches later in chapter 6, when we will discuss *how* referential state annotation should be added to the texts, but for now the mere concept of coreference resolution is of importance. The task of coreference resolution is to find the correct antecedent for each and every noun phrase in a text—provided the NP has an antecedent.

The initial efforts of the Nijmegen group were based on work with Cesac, a coreference editor for syntactically annotated corpora (Komen, 2009a). The program allows making a link from a noun phrase to an antecedent manually (see Figure 10).³

Cesac only allows adding information status categories for noun phrases that have an antecedent. The program allows specifying the *type* of coreference relation, which comes close to the referential state category we are looking for. The categories it discerns are in (111).

- | | |
|----------------|--|
| (111) Identity | - The referent of the constituent is identical to that of its antecedent |
| CrossSpeech | - As “Identity”, but crossing a direct or indirect speech boundary |
| BoundAnaphor | - The constituent is an anaphor bound to the antecedent within the clause (e.g. a reflexive) |
| Subset | - The constituent is a subset of the larger category of the antecedent |
| PartOfWhole | - The constituent is a part of the larger whole in the antecedent |
| Cataphoric | - The antecedent follows instead of precedes the constituent |
| Inferred | - The antecedent is related to the constituent in another way |

expansion of tags with the category KIND and the addition of “non-specific” categories, as shown in (113).

(113) *Specific tags*

- OLD - Items available in prior discourse⁵
 - ACC-sit - Accessible from the *situation*
 - ACC-inf - Accessible from *inference* to items in prior discourse
 - ACC-gen - Accessible from *general* world knowledge
 - NEW - Other items
 - KIND - Generic noun phrases denoting kinds
- Non-specific tags*
- NONSPEC - A new instantiation of a not actually existing referent
 - NONSPEC-old - A reference to a previously mentioned non-specific referent
 - NONSPEC-inf - A non-specific referent inferred from another NP
 - QUANT - Quantifier noun phrases (like “all people”)

The first small set of annotation labels comes very close to what is useful as a bare minimum set of information status categories, except that the categories of ACC-sit and ACC-gen could be collapsed into a single category and there is no category for “inert” NPs.⁶

The PROIEL annotation manual reports that the category KIND is sometimes close to other categories: an NP like *death* should normally be tagged as KIND, but an NP like *his death* should receive a more specific label; an NP like *eternal life* would normally fit the category KIND, but should receive the ACC-gen tag; constituents that should normally be labelled as KIND, but which “pick up” a previous one are to be labelled as OLD (PROIEL, 2011). The extension to the larger tagset seems to be unnecessary from the point of view of a “bare minimal” approach we are looking for as advocated by requirement (108c). Some of the added categories (QUANT, NONSPEC, NONSPEC-old, NONSPEC-inf) can be derived from the syntactic environment of the constituent or by checking its antecedent (see section 5.4). This violates our criteria in (108a-b). The category of KIND does not seem to be derivable from syntax, but it overlaps with NEW, OLD and ACC-gen, and it is unclear at this point whether making a distinction between discourse-new entities that refer to individuals and those that refer to kinds is needed within the framework of looking for focus domains.

In sum, the initial PROIEL approach comes very close to what we are looking for: a set of referential state *primitives*.⁷ The finer differentiations made by the expanded set of PROIEL categories do not seem to match the criteria in (108), as we will see in sections 5.4.3 and 5.4.4.

5.3 The Pentaset as referential state primitives

The discourse processing model described in chapter 2, and the knowledge of existing taxonomies from section 5.2 provide enough background to define a small set of referential state primitives satisfying the criteria in (108). The set of five referential state primitives that I argue for and which we will refer to as the “Pentaset”, is shown in Figure 11.

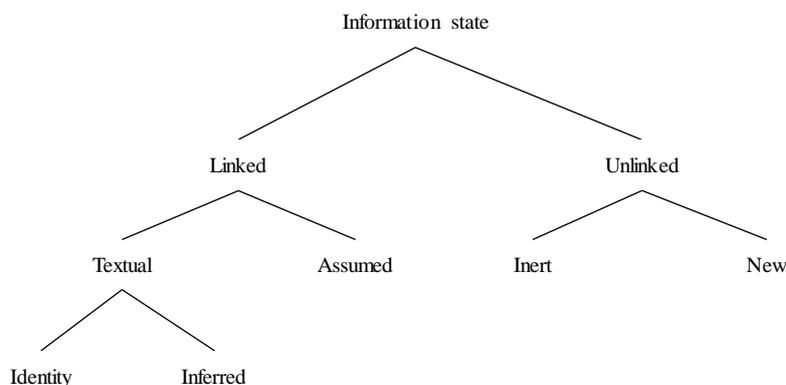


Figure 11 The referential state primitives in the Pentaset

The first distinction that the Pentaset makes is between noun phrases *with* and *without* antecedents. Those *with* antecedents (either in the discourse, or in the real world) can be regarded as “Linked”, while those without are “Unlinked”, as shown in Figure 11. The Pentaset recognizes that antecedents can be inter-textual or extra-textual. Those with antecedents in the text are separated into those for whom the mental entity of the constituent and of its antecedent are identical (category IDENTITY) and those for whom this is not the case (category INFERRED). Noun phrases with an extra-textual antecedent, which leads to mental entities having a link to long-term memory, are labelled with the category ASSUMED. The noun phrases without antecedents divide into two groups depending on their ability to be referred to in subsequent clauses. Those that cannot be referred to are called INERT, while those that can be referred to are labelled NEW.

We now turn to the formal definitions of the Pentaset categories, all of which are grounded in the mental model (see 2.3.2), while we build the definitions on the basic notion of “mental entity” as formally defined in (19) of section 2.3.1.

5.3.1 Identity

The first category that we turn to is that of “Identity”, where a noun phrase refers to something that is already available in a reader’s mental model:

(114) **Identity**

A constituent NP_i with mental entity $MEnt(NP_i)$ has the referential status “Identity” if there is an NP_j with $j < i$, such that $MEnt(NP_j) = MEnt(NP_i)$.

The definition of the referential state primitive “Identity”, as in (114), is straightforward: the entity referred to by the constituent must be exactly the same as an entity that is already available in the current situation model of the reader. Some examples of Identity relations are given in (115).

- (115) a. It has sometimes been assigned to B. C. 100, but in **that year** Glauca was praetor. [long-1866:525-526]
 b. Antonius also had opportunities of improving himself during his quaestorship in Asia B. C. 113, and again when **he** had the province of Cilicia B. C. 103. [long-1866:373-374]

No new mental entity needs to be created for the NP *that year* in (115a), since it can be associated directly with the one that has already been created for *B.C.100* in the current situation model. The referential state of *that year* can therefore be assigned the value “Identity”.

Example (115b) has the pronoun *he*, which refers to the person of Antonius. Since the pronoun *he* directly associates with the mental entity of the NP “Antonius”, which is already available in the current situation model, the referential state of *he* receives the value “Identity”.

5.3.2 Inferred

The relation of “Inferred” is, perhaps, the most difficult relation to define of all the relations in the Pentaset. Prince (1981: 236) states that “a discourse entity is inferable if the speaker assumes the hearer can infer it, via logical—or, more commonly, plausible—reasoning, from [other] discourse entities”. Inference according to this definition, then, assumes there are two entities (which are called “discourse entities” by Prince). One entity has already been processed by the addressee and has led to the creation of a mental entity in the model of the current discourse that is built in the mind of the reader. The situation by which the *first* mental entity was created evoked a link to a model of some kind (the word *restaurant* for instance evokes the model of a restaurant, available in long term memory), having particular “slots” that could be filled in (such as *waiter*, *table*, *bill* and so on when the word *restaurant* occurs in a discourse). It is these “slots” that are ready to contain mental entities standing in an “inference” relation to the first mental entity.⁸

The process of inferring the existence of one referent from that of the other is described by Prince as “logical reasoning”. An accurate definition needs to be as specific as possible about the “logical reasoning” process that is used for inference, and what follows is my own attempt to formalize this process. Given a sentence such as: *I got onto the bus, but the driver was ill*, the mental entity created for *driver* stands in an inference relation with that of *bus*. The process of inference can be illustrated by (116).

- (116) a. There exists a set *B* of which *x* (denoting ‘the bus’) is one element
 b. There exists a set *D* of which *y* (denoting ‘the driver’) is one element
 c. Find an appropriate class-member relation: buses have drivers
 d. Inference: *Have*(*x*, *y*)

We start in (116a) by finding a set to which the constituent *the bus* belongs: the set of *buses*. The next step (116b) introduces a new entity, *the driver*, for which we also find a set: *drivers*. The step (116c) describes the search in predefined class-member

relationships. There are several relationships that are not applicable, such as: busses have wheels, busses have windows, busses have timetables etc. But we find one relation that matches the situation: busses have drivers. Step (116d) is the inference: if there is a bus, and all busses “Have” some particular “entity”, then our bus also must have this “entity”, and so this “entity” must exist, as implied in (116c).

The crucial restriction in the process of inference is in the kind of relations we allow for in a step like (116c). The set relation of “Have” is one, but another possible relation would be “Subset”. Consider a sentence like: *She came running down the steps and she fell down four* (Prince, 1981 ex. 28i). The numeral *four* denotes a subset of all the *steps*.

I argue that we can restrict the category of “Inferred” enough by only allowing for *direct* relations instead of *indirect* ones. Examples of *direct* relations are: Have(x,y), Subset(x,y) and Member(x,y).⁹ *Indirect* relations are linked direct relations, such as for instance: Have(x,y) + Subset(y,z), where y provides the link between the two.

An example with an indirect relation, which, we argue, does not lead to an inference, is in a sentence such as: *I got into the bus, but the uniform was wrinkled*. The *uniform* cannot be *directly* inferred from *bus*, but there does exist an *indirect* relation: busses have drivers and drivers have uniforms.

(117) **Inferred**

A constituent NP_i with mental entity $MEnt(NP_i)$ has the referential status “Inferred” if

- (i) there is no NP_j with $j < i$, such that $MEnt(NP_j) = MEnt(NP_i)$, but
- (ii) there is an NP_k with $k < i$, such that:
 - a. $MEnt(NP_i) \in S_x$
 - b. $MEnt(NP_k) \in S_y$
 - c. there exists a *direct set relation* between set S_x and S_y .

The definition of the referential state category “Inferred” is given in (117). There are two noun phrases, and as the reader reads a text, it has made a mental entity in the current situation model for each of them. The referents of these mental entities, however, are not the same, so that an Identity relation does not exist between them. Condition (ii.c) of the definition states that a simple (direct) set relation must exist between the set of which the referent of XP is a member and the set of which the referent of YP is a member.

We can see how this definition works out in practice from the examples in (118).

- (118) a. Accordingly runaway slaves stayed there, and were of course maintained by the guardians of the temple, until the masters came to reasonable terms with the slaves and confirmed **the agreement** by a solemn oath, which no master was ever known to have violated. [long-1866:38-39]
- b. The fear of the deities of the place secured the performance of the oath; for divine vengeance soon followed an act of perjury. **Some perjurers** had been deprived of their sight on the spot. [long-1866:40-42]

Example (118a) contains two constituents, and the second one stands in an *Inferred* relation to the first one. As the reader processes the text above, he creates a mental entity in the situation model for *reasonable terms* (it gets referential state “New”), and the set of *terms* is evoked from long term memory. Reading on, the reader gets to *the agreement*, creates a mental entity in the situation model, and this entity evokes the set of *agreements* from long term memory. Having evoked this set, the mind of the reader checks if there is a direct relation with any of the sets belonging to the mental entities that are in the situation model, and he sees that there is the relation *Have*(*agreements*, *terms*). The reader then makes a link from *the agreement* to *reasonable terms*. The Pentaset classifies this link as “Inferred”.

The last sentence in example (118b) has *some perjurers*, the referent of which is part of the whole set of *perjurers*. This set stands in direct relation to the set of *acts of perjury*, of which *an act of perjury* is a member. The relation between these two sets is “Perform”: all perjurers perform acts of perjury.

5.3.3 Assumed

An author (or speaker) may assume that the addressee is able to link a particular concept with an entity that is already available in the addressee’s long term memory (that is, in the location of the mind where the brain stores “general knowledge”, or in the location where it stores knowledge related to the discourse situation; see the mental model in section 2.3.2). If this is the case, then the addressee creates a mental entity in the situation model, and links it to the entity in long-term memory. The linguistic expression the author uses to refer to the concept gets the referential state “Assumed” assigned to it. The formal definition is in (119).

(119) Assumed

- A constituent NP_i with mental entity $MEnt(NP_i)$ is “Assumed” if
- (a) there is no NP_j with $j < i$, such that $MEnt(NP_j) = MEnt(NP_i)$,
 - (b) nor such that $MEnt(NP_j)$ can be inferred from $MEnt(NP_i)$, but
 - (c) there exists an NP_{LTM} (in long-term memory),
such that $MEnt(NP_{LTM}) = MEnt(NP_i)$.

The definition of “Assumed” in (119) assumes that all world knowledge (which includes for instance sun, moon, stars) as well as situational knowledge (which includes the speaker, the hearer, the book that is being written, etc) is stored in the addressee’s long-term memory. The entities referred to in long-term memory are not by default part of the situation model, but only when they are evoked (which means that a link is established). Certain entities from this world are available in the minds of both interlocutors, and they are conscious of this fact. So when a speaker or author uses a linguistic expression XP that has not been used previously in the text, he may safely assume that the concept denoted by this expression is already available in the mind of his addressee.

- (120) a. The leader was Athenion, a **Cilician** born, and the bailiff of two rich brothers. [long-1866:117]
b. He was a man of courage and could read **the stars**. [long-1866:118-119]

- c. No reason is given by Diodorus for Tryphon leaving the east side of the island and establishing himself in the west, but **we** may conjecture that as he had failed before Morgantia, and there were on the east side of Sicily the large cities Messana, Catana, Syracuse, and others, the new king did not feel quite safe there. [long-1866:159-160]

The reference to *Cilicia* in (120a) can be made by the author, because he assumes that the readers of his book are familiar to the general geography of the world in that time. It is for that reason that a simple name suffices, instead of a postmodified noun phrase like “Cilicia, a region in X”. It is also reasonable to assume that people on earth are familiar with the existence of *stars*, so that the expression *the stars* in (120b) can receive the referential category of “Assumed”. This explains why it is a definite expression, though there is no prior mention of stars in the text. The pronoun *we* in (120c), finally, receives the referential category “Assumed” because it refers to a set of people that belong to the “world” of the written material, which consists of the author and the readers.

5.3.4 Inert

An important differentiation made by the Pentaset (see Figure 11) is between constituents that are “linked” and those that are “unlinked” in the following sense: for the “linked” expressions the addressee should already have a mental representation (be that in the current mental model or in epistemic memory), while this is not the case for the unlinked ones. The unlinked expressions are further divided in two groups: the “inert” ones and the “new” ones. The first group of expressions are, in a sense, “inert” to the referential system: they have no antecedent, and no entity in a subsequent clause can take them as antecedent (they function more like attributes of other entities). The formal definition of this referential category is given in (121).

(121) Inert

A constituent NP_i with mental entity $MEnt(NP_i)$ is “Inert” if

- (a) there is no NP_j with $j < i$, such that $MEnt(NP_j) = MEnt(NP_i)$,
- (b) nor such that $MEnt(NP_j)$ can be inferred from $MEnt(NP_i)$, and
- (c) it is **not** possible that there exists an NP_k with $k > i$, such that $MEnt(NP_k) = MEnt(NP_i)$.

The definition of “Inert” in (121) does not only require the absence of preceding constituents that have been marked with “Identity”, but also the absence of preceding constituents marked as “Inferred”. The reason for this is that we, by the requirements stated in (108), do not want to have overlapping referential state categories. In terms of the situation model, the reader meets a noun phrase, looks for a “match” to link it to among the available mental entities in the situation model, when it finds none, it looks for a match in long term memory and then for a possible inference with one of the sets evoked by mental entities available in the model. When none of these candidate links lead to a match, the reader checks if the entity is one that can be referred to later or not. If this is not the case, then the mental entity

created in the short term memory is not propagated to the situation model, and disperses.¹⁰

In some sense the referential category of “Inert” is like the one of “New”: it concerns the introduction of a referent that has not in any way been referred to previously in the text. There is, however, one important difference: the mental entity which is created for an Inert XP is completely inaccessible in subsequent clauses. Let us have a look at some examples:

- (122) a. But Tryphon, suspecting that Athenion would take some opportunity to attack him, put his general in **prison**.
 b. Triocala, which Tryphon chose for his royal residence, was naturally **a strong place**. It was so called, as people said, but perhaps they did not say true, because it possessed three good things, abundance of excellent water, a territory rich in wine, oil, and grain, and perfect security, for it was **a large impregnable rock**. [long-1866:166-167]
 c. Ann is **a teacher**. ?The teacher caught a bus. (Johnson-Laird, 1983: 383)
 d. But **there was another Apartment** in the House where the Prince or King, or whatever he was, and several other were. [defoe-1719:373]

The NP *prison* in (122a) does not refer to one particular prison building, but only to the concept of being confined. As such, it is new to the text. There is no expression in the preceding text that exactly the same entity. The subsequent sentence does not make mention of a prison, nor is it able to. Suppose the first sentence would be followed by a sentence like: “The general spent seven years in that place”. Such a follow-up would be impossible, because the location referred to by *that place* must be one particular location, while *prison* in the preceding clause only denotes the concept of being confined in a non-specified location, without singling out one particular entity.

The noun phrases *a strong place* and *a large impregnable rock* in example (122b) describes a quality, an attribute of “Triocala”, the location Tryphon chose for his royal residence. They are new to the discourse, which is reflected in the choice of expression: the noun phrases start with an indefinite article. While the attribute *strong* is added to the location Triocala, no mental entity for *a strong place* is retained in the mental model.¹¹ This is clear from the inability of the pronoun *it* in the subsequent clause to refer back to *a strong place*. Instead, it refers to the location Triocala.

Johnson-Laird (1983: 383), building on work from Stenning (1977, 1978), recognizes the fact that an indefinite expression like *a teacher* in (122c) does not lead to the introduction of a unique entity in the mental model (see also the discussion on equative clauses in 3.2.2.1). It is precisely for this reason that the second sentence in (122c) is not licit: when the noun phrase *the teacher* is used, the expectation is that there already exists a unique token in the current mental model to which *the teacher* can refer.

Expletive pronouns like *there* in (122d) are grammatically the subject of a sentence, but they do not link back to any tangible entity in or outside the mental model and they cannot be referred back to either: they are inert.¹²

We have seen three types of expressions that are very prone to receive the referential category of “Inert”: bare nouns within a prepositional phrase (e.g. *prison*), attributive indefinite noun phrases in the complement position of an equative clause (e.g. *a strong place*) and expletive subject pronouns. We leave the question of which (if any) other situations allow for or require “Inert” expressions to further research, since this question, intriguing as it may be, is outside the scope of this current research.

5.3.5 New

Constituents that receive the referential category of “New” are like the “Inert”, “Inferred” and “Assumed” categories, in that they do not have an antecedent within the text (or utterance). What distinguishes “New” and “Inert” ones from constituents labelled “Assumed” is that the former don’t have an extra-linguistic antecedent, one that is already available in the mind of the interlocutor. Referentially “New” constituents distinguish themselves from “Inert” ones in their ability to be referred to in subsequent clauses. “Inert” ones cannot be referred to later on (since no mental entity is created for them in the situation model), but “New” ones can. Referentially “New” constituents are able to establish a topic: a mental entity that is kept in the mental model for reference in subsequent clauses or sentences. Since both “Inert” and “New” constituents do not have an antecedent, the formal definition of “New” in (123) is much like that of “Inert”.

(123) New

A constituent NP_i with mental entity $MEnt(NP_i)$ is “New” if

- (a) there is no NP_j with $j < i$, such that $MEnt(NP_j) = MEnt(NP_i)$,
- (b) nor such that $MEnt(NP_j)$ can be inferred from $MEnt(NP_i)$, but
- (c) it **is** possible that there exists an NP_k with $k > i$,
such that $MEnt(NP_k) = MEnt(NP_i)$.

The definition in (123) says that a constituent receives the referential state category of “New” when it denotes an entity, and there is no constituent in the preceding context of the discourse that either refers to exactly the same entity, or that links to another entity by an inference. Unlike constituents with the referential state “Inert”, the ones labelled “New” create a mental entity in the situation model to which constituents in a following clause can link (either through an “Identity” or an “Inferred” relation). It should be emphasized, though, that constituents labelled as “New” are not necessarily referred to later on in a text. English texts seem to have a large number of noun phrases that *do* lead to the creation of a mental entity in the situation model, but to which no further reference is made. The examples in (124) should illustrate the category “New”.

- (124) a. **A man named Silus** had given evidence against Piso the client of Crassus: it was hearsay evidence, which the Romans allowed, but they did not overvalue it. Crassus in his cross-examination of Silus said to him: It is possible, Silus, that the man from whom you say that you heard this said it in a passion. [long-1866:408-410]
- b. He had also erected spacious and lofty buildings on the shores of **the salt lagoon named the Lucrine Lake**, for the purpose of breeding oysters. But the lagoon was public property and let to a Publicanus or public contractor, named Considius, who complained of Orata's encroachments on the lagoon, and brought an action against him. [long-1866:477-478]

The indefinite noun phrase *A man named Silus* in (124a) creates a mental entity in the situation model of the addressee that can be referred to in subsequent clauses. A few clauses after the one where the referent is established, it is picked up again by the proper name *Silus*, and then later on in that sentence by the pronoun *him*.

The noun phrase *the salt lagoon named the Lucrine Lake*, even though it carries a definite article, is the first reference to this lagoon. The information that is needed to establish the unique referent of *lagoon* is contained within the postmodification *named the Lucrine Lake*. With the referent thus uniquely established, the next clauses can simply refer back to it by the noun phrase *the lagoon*.

5.4 Is the Pentaset sufficient?

We have reviewed existing taxonomies as candidates for referential state primitives, and since none of them proved to be in line with the requirements stated in (108), I have proposed the “Pentaset” in the previous section. The question arises whether this small set of states is really sufficient in the sense of criterion (108c): are we able to use the Pentaset such that combining constituents according to their syntax and “Pentaset” state allows all relevant information state distinctions to be made? Provided we do not only label noun phrases for referential state category, but also store their antecedents (if they have one), all other information that is needed to determine the “information structure” of sentences can, as I argue, be derived by combining the syntax and referential states of the constituents and their antecedents. The main goal of this current section and the next section 5.5 is to make this hypothesis plausible. The strategy taken in this section is to show that when syntactic information is combined with the information in the Pentaset primitives, most of the information state categories available in the existing taxonomies (see 5.2) can be derived, whereas section 5.5 serves to give an idea of how the Pentaset, combined with syntactic information, relates to higher order information structure notions such as topic and focus.

5.4.1 Pentaset categories versus alternatives

The overview in Table 14 compares the Pentaset with our own set (Cesac) as well as with other information status category sets proposed in the literature (see section 2): Prince (1981, 1992); Lambrecht (1994), the Proiel set as described in Haug (2009), and Gundel, Hedberg and Zacharsky (1993), here abbreviated as GHZ.

Table 14 Comparison of information status category sets

Pentaset	Cesac	Proiel	Prince	GHZ	Lambrecht
Identity	Identity CrossSpeech BoundAnaphor Cataphoric	Old Old-inactive	Evoked textually	In focus Activated Familiar Uniquely Idt Referential	Active Accessible
Inferred	PartOfWhole Subset Inferred	Acc-inf	Inferrable containing non-containing	-	Accessible
Assumed	-	Acc-sit Acc-gen	Unused Evoked situationally	-	Unused
New	-	New	Brand-new anchored unanchored	Type Idt.	Brand-new anchored unanchored
Inert	-	-	-	-	-

The Pentaset, which is in the first column, is indeed the most concise set, as compared to the other sets. Each of the categories in the Pentaset is represented by more than one category in at least one of the other information status category sets.

5.4.2 Deriving other categories from the Pentaset

The question needs to be answered, however, whether the Pentaset, concise as it is, does not “throw away” information in the sense that it is too generic, and does not allow to make distinctions that are (perhaps implicitly) deemed to be significant by the sets of information state categories in other taxonomies mentioned in section 5.2. Table 15 serves to answer this question to some extent: it illustrates the relation between the referential state primitives of the Pentaset and information state categories of the other sets. The first column has the Pentaset’s referential state, and the last column has the corresponding information state in another set, based on additional criteria that are stated in the second column.

The criteria needed to determine the information state of other sets can be broken up into several categories. The *antecedent distance* of constituents that are marked as “Identity” and “Inferred” in the Pentaset is used by Lambrecht, Proiel, Cesac and GHZ. The *noun phrase type* is used by GHZ to distinguish several categories, which is not surprising, given the fact that the whole idea of the Givenness Hierarchy is to establish a relation between the form of a referring expression and its cognitive status. The presence or absence of immediate *children* from a particular *noun phrase category* is used by Prince, Lambrecht and GHZ. The existence of a *speech boundary* between the source and the antecedent is used by Cesac.

Table 15 Deriving other information status categories from the Pentaset

Pentaset	Criterion	Set	Information state
Identity	none	Prince	Textually evoked
	antecedent relatively far away	Lambrecht	Accessible
	antecedent in immediate context	Lambrecht	Active
	antecedent > 15	Proiel	Old-inactive ¹³
	antecedent < 15	Proiel	Old
	antecedent follows	Cesac	Cataphoric
	antecedent Refl.Pro	Cesac	BoundAnaphor
	cross speech boundary	Cesac	CrossSpeech
	antecedent <= 1	GHZ	In Focus
	antecedent >1, NPtype: <i>dem, dem+N</i>	GHZ	Activated
	antecedent >1, NPtype: <i>dem+N</i>	GHZ	Familiar
antecedent >1, NPtype not: <i>dem, dem+N</i>	GHZ	Uniquely Identifiable	
Inferred	none	Prince	Noncontaining inferrable
	none	Proiel	Acc-inf
	Source is part of antecedent whole	Cesac	PartOfWhole
	Source is subset of antecedent	Cesac	Subset
	Source is <i>not</i> part of whole or subset	Cesac	Inferred
	antecedent relatively far away	Lambrecht	Accessible
	antecedent in immediate context	Lambrecht	Active
	antecedent >1, NPtype: <i>dem, dem+N</i>	GHZ	Activated
	antecedent >1, NPtype: <i>dem+N</i>	GHZ	Familiar
antecedent >1, NPtype not: <i>dem, dem+N</i>	GHZ	Uniquely Identifiable	
Assumed	1 st or 2 nd person	Prince	Situationally evoked
	1 st or 2 nd person	Proiel	Acc-sit
	3 rd person	Prince Lambrecht	Unused
	3 rd person	Proiel	Acc-gen
New	none	Proiel	New
	No Identity/Inferred/Postmodifying child	Prince Lambrecht	Brand-new unanchored
	One Identity/Inferred child	Prince Lambrecht	Brand-new anchored
	One Postmodifying child	Prince	Containing inferrable
	One Postmodifying child	GHZ	Referential
	No postmodifying child	GHZ	Type Identifiable

Almost all of the criteria mentioned above can actually be derived from the syntactic information in the parsed English corpora and from the coreference information supplied by the Pentaset. One type of criterion has not been mentioned yet: the semantic relation between the source and the antecedent. Cesac uses that to determine whether the information state is PartOfWhole, Subset or Inferred. This particular distinction cannot be derived from the available syntactic information and the referential information supplied by the Pentaset, but I argue that these distinctions are not necessary according to the criteria in (108), since it is hard to see how the finer distinctions in the semantics of the “Inferred” category would make a

difference in the information state notions we are looking for. Including the area of semantics in our quest for the relationship between syntax and information structure would only serve to complicate further a picture which is quite complex as it is already. Further research should show if the finer semantic distinctions are necessary for information structure research purposes.

5.4.3 Generics

In our search to answer the question whether the five referential categories proposed by the Pentaset are sufficient to derive relevant information structure distinctions, as stated in the requirement of (108c), we need to be sure that our set of primitives is not too small, so that we miss distinctions that need to be made due to their relevance for information structure. There are two distinctions made by other taxonomies, which the Pentaset does not make, and I would like to zoom in on them. This section treats the first one: generics.

Gundel et al (1993) take as their lowest ranked cognitive status “Type Identifiable”, which they define as “The addressee is able to access a representation of the type of object described by the expression”. They give as an example the NP *a dog* as in (125a). No specific dog can be pointed at, but the addressee is able to retrieve a mental picture of the characteristics of a “dog”. While it is true that no specific dog can be pointed at, the dog is, in fact, a specific one: the one dog that was there outside last night, and was barking. The Pentaset way of dealing with *a dog* in (125a) is straightforward: it gets assigned the referential status of “New”, since a new mental entity is created in the situation model of the addressee.

(125) a. I couldn’t sleep last night.

A **dog** (next door) kept me awake. (Gundel et al)

b. **Prophets** wear sandals. (Proiel)

c. Isaiah, Elijah, ... wear sandals. (Proiel)

d. If **man** lands on the moon, it will be a great step forwards. (Proiel)

Haug et al (2009) argue for a cognitive status that goes one step “lower”, as it were, than Gundel’s “Type Identifiable”; they introduce the category of KIND. This new category is used for “generic referents such as ‘the lion’ in ‘The lion has a mane’” (PROIEL, 2011: 4). The Proiel coding manual continues to exemplify this category by comparing (125b) with (125c): the former has the NP *Prophets*, which cannot be replaced by a list of specific prophets in the latter, and is therefore labelled as “KIND”. The manual gives another example in the form of the NP *man* in (125d): this NP does not refer to one particular man, but to the whole of mankind, and is therefore to be labelled as “KIND”.

In the Pentaset approach a kind-referring expression such as *Prophets* in (125b), which points to a set rather to one individual, is treated just like any other NP: if it occurs for the first time in a text and does not link back to a previously mentioned entity, it receives the referential category “New”. I argue that they can be treated as entities from a referential point of view: it is possible to link back to sets in much the same way that individuals can be referred back to. A follow-up sentence on (125b),

for instance, could be: “They also wear leather belts”, where “they” refers back to “prophets” in the previous sentence.

A follow-up on (125d) cannot be done in the form of a pronoun. However, picking up the set “man” can effectively be done by repeated use of “man”: “*If man lands on the moon, man could also land on mars. But man is not unlimited in his abilities.*” The repeated use of *man* repeatedly refers to the same mental entity, which is a representation of the whole of mankind. The kind of environment, the conditional clause, inside which the word *man* is found is a special kind, and it is the topic of the next section.

5.4.4 Referential islands

In our review of distinctions that are made by other taxonomies but not by the Pentaset we now come to the second one: referents that are created in what could be called “referential islands”, which are opaque contexts such as negation, quantification and modality.¹⁴ The Proiel (2011) tagset (see section 5.2.5) reserves a special tag, called “NonSpecific”, for NPs that lead to the creation of mental entities in these referential islands, since they seem to resist being referred to outside of the opaque island contexts. The example given is repeated here:

- (126) a. No one lights a lamp_i and hides it_i. (Proiel ex. 2, adapted from Lk 8:16^a)
 b. *Jesus continued to speak about it_i.
 c. *Jesus continued to speak about this lamp_i.

The Proiel guidelines (2011) explain that, in the case of (126a), it “does not make sense to use a pronoun *it* to refer to the lamp which no one lights [outside of negation, e.g. in the next sentence].” Indeed, a follow-up sentence that does not retain the opaque context, such as (126b), cannot pick the *lamp* from (126a) up by using a pronoun. Picking up the *lamp* by using a definite NP such as *this lamp* in (126c) is equally impossible.

The idea of referential islands goes back to Karttunen (1969, 2003), who introduces the class of, what he calls, “short term referents”, which are the entities tagged in Proiel as “NonSpecific” that only exist as long as an opaque context is continued. Karttunen’s examples are:

- (127) a. You must write a letter_i to your parents and mail the letter_i right away.
 *They are expecting the letter_i. [Karttunen ex. (25a)]
 b. John wants to catch a fish_i and eat it_i for supper.
 *Do you see the fish_i over there? [Karttunen ex. (25b)]
 c. I don’t believe that Mary had a baby_i and named her_i Sue.
 *The baby_i has mumps. [Karttunen ex. (25c)]
 d. You must write a letter_i to your parents. It_i has to be sent by airmail. The letter_i must get there by tomorrow. [Karttunen ex. (26)]
 e. Mary wants to marry a rich man_i. He_i must be a banker. [Karttunen ex. (27)]
 f. Mary wants to marry a rich man_i. He_i lives in New York.

- g. My wife just phoned, and told me she wants to buy me a new shirt_i for my birthday. Which reminds me: I still need to do some shoppings and invite friends. Anyway, as for this new shirt_i, I would want it_i to be blue—navy blue.

The verbs *must*, *want* and *believe* in (127a-c) create a context such that when a new mental entity is created (*letter*, *fish* and *baby*), the entity can only be referred back to when the context is maintained. Follow-up sentences in (127a-c) that do *not* maintain the opaque context do not allow referring back to the entities created in the opaque contexts. Example (127d) shows that it is possible to maintain a referent created in an opaque context for more than one sentence, provided that the sentences also contain an opaque context (which is facilitated by the verbs *has to* and *must* in this example). Example (127e) illustrates a referential island (set up by the verb *want*) inside which a generic entity is created, one that does not refer to an individual, but to a set (all men who are rich). The fact that the ensuing sentence, provided that it continues the opaque context, is able to maintain reference to the generic entity thus created confirms the conclusion from the previous section that set references do not differ from individual references in terms of creation and maintenance. An alternative continuation of the sentence in (127f) illustrates one of the points Karttunen makes: the *rich man* can be specific or non-specific. The difference between (127e) and (127f) is that when the *rich man* refers to a set (that is: non-specific), the back reference to it must continue the opaque context, whereas when the *rich man* is one particular individual (the specific reading), the next reference to him should not be in a continuation of the opaque context. Example (127g) illustrates that the *new shirt*, which is created in an opaque context (due to the modality connected with the verb *want*), can still be referred back to after the referential island is left: but only if the opaque environment of the referential island is copied.

As the Proiel (2011) guidelines rightly state, opaque contexts do not only include modality (the theme of Karttunen's sentences), but also negation, quantification, and other modalities, of which my own examples in (128) testify.

- (128) a. John didn't read a book_i last night.
 i. *He only looked at it_i
 ii. *He only looked at the book_i.
 b. All books yearn for a reader_k.
 i. *She_k loves them too.
 ii. *The reader_k is in the library.
 c. If a student asks a question_m, you should be happy.
 i. *Its_m form is irrelevant.

The sentence negation in line (128a) creates a context, and when a new entity such as "a book" is instantiated *within* this context it seems this cannot be picked up again in the next sentence—not by a pronoun like "it", nor by a definite NP like "the book". The quantifier *all* in (128b) creates a similar context, and it seems to be impossible to refer back to a new entity such as *a reader* that is instantiated in that

context: not by *she*, nor by *the reader*. The example in (128c) sets a conditional mood context, and it seems likewise problematic to refer back to an entity like *a question* instantiated in it.

While the existence of referential islands is beyond doubt, it is sometimes possible to pick up a referent outside the opaque context, by using non-default stress or by using different types of NPs, witness the examples in (129).

- (129) a. John didn't **read** a book_i last night. He **repaired** one_i.
 He took it_i to the library later.
 b. All books yearn for a reader_k. This reader_k is in the library.
 c. If a student asks a question_m, you should answer it_m.
 The form of the question_m is irrelevant.

The example (129a) uses stress on “read” and “repaired” to arrive at a reading where “one” in the second sentence links back to “a book”. This kind of reference, however, is not of the “Identity” type—it is “Inferred”: the initial “a book” creates a mental entity that happens to be a set, and the noun phrase “one” refers back to one element from this set. The quantifier context created in (128b) allows for an escape too, witness the example in (129b), where “this reader” refers to exactly the same set as the one instantiated in the situation model by “a reader”. This illustrates that it is sometimes possible to refer back to an element in an otherwise closed context with an “Identity” link. The conditional context in (128c) allows for a similar escape hatch: the pronoun *it* refers exactly to the same set of questions that are represented by the mental entity created in the situation model for *a question*.

What we see above, is that non-specific or generic entities that enter the situation model as new instantiations under certain contexts can be very hard to refer back to, but it is not always impossible. Nevertheless, even in the “retrievable” cases of (129) there is an odd characteristic related to the referential islands: if entities are not referred to soon enough, they really are beyond reach.

- (130) a. John didn't **read** a book_i last night. He went to the cinema instead.
 i. *When he came home, he **repaired** one_i.
 ii. When he came home, he **repaired** a book_k.
 b. All books yearn for a reader_k and really want him_k to sit down with them.
 They do that every day of the year, but there is hardly anybody taking notice of this.
 i. *This reader_k is in the library.
 c. If a student asks a question_m, you should be happy. Go home and tell your family. And if they are not around, then write them an email.
 i. ?The form of the question_m is irrelevant.

The “book” introduced in the negation context in (130a) really is beyond recovery when we have a sentence like *He went to the cinema instead* follow it, and only then try to refer back to it. The attempt to use “one” to at least refer back with “Inferred” to “a book” fails, because “the cinema” now has become the best antecedent of “one”. The second attempt in (130a.ii) is to repeat the generic NP itself, “a book”, and this one fails for different reasons: the mention of “a book” in the not-negated

context causes the reader to look in his situation for a mental entity that is compatible, and when none is found, a new mental entity is created. This entity does not really seem to link back to the mental entity representing the set of books that was created in the negation context in (130a). Even an “Inferred” link seems impossible, and it seems, in fact, that the mental entity representing the set of books has completely *left* the situation model. It is here that we have the first example of a mental entity with a “restricted life” rather than a “text-long life” or a “clause-long life”.

The situation in (130b) is similar to the one above: the mental entity created for the set of readers triggered by “a reader” has disappeared by the time we reach (130b.i) and cannot be recovered again. Notice that the “reader” does have a life within the limited quantifier context, since it is picked up by the pronoun “him” in the first sentence. The example in (130c) is a bit harder: it may still be possible to use “the question” in (130c.i) despite the intervening two sentences.

To sum up what we have found on referential islands: entities that receive the referential status “New” in contexts like negation, quantification and modality often cannot be referred back to outside these contexts, and even when referring back to them from outside the opaque context is possible, this should be done in the immediately following clause, or else the mental entities seem to have disappeared. This brings us back to the matter of the criterion on referential state primitives stated in (108c), which says that the set of primitives should be “sufficient” so as to derive relevant information state distinctions. Since it remains to be shown whether entities created in referential islands behave differently in terms of information structure, we will err on the safe side if we make sure that the situations such as the ones illustrated in (128)-(130) are recognizable, and ways to do this include the following:

(131) *Recognizing contexts leading to short-life referents*

- a. Use a separate referential category for entities created in contexts that might lead to short-life referents.
- b. Do not add referential categories, capitalizing on the combination of syntax and the Pentaset categories to recognize entities created in contexts that could lead to short-life referents.

Option (131a) seems to be the solution chosen by the Proiel project, while I argue that for the information structure research done with the English parsed corpora, the solution described in (131b) is sufficient: we do not add categories to the Pentaset, but instead rely on the syntactic and referential category labels to help us discern suspicious contexts. All that is needed to make this solution plausible is to show that the situations described in (128) can be discerned automatically, and I would like to illustrate that this is possible by showing a “suspicious context” from one of the English texts we are working with:¹⁵

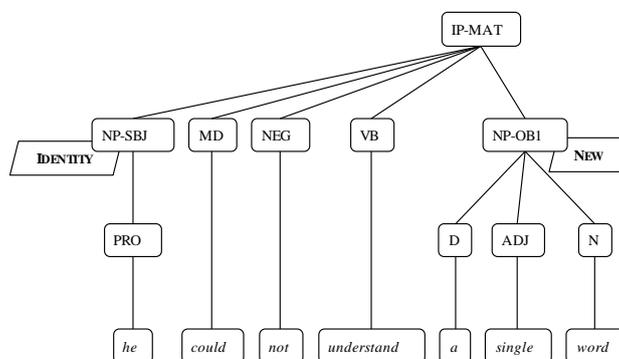


Figure 12 Suspicious context in a text from the parsed English corpora

What is shown in Figure 12 is a graphical representation of a sentence from the corpora, comparable to (128a), illustrating that each word has a word category assigned (the word “he” has the label “PRO”, signalling that it is a pronoun, for instance), that each constituent has a label telling its major category (such as “NP” and “VB”) as well as a function (such as “SBJ” and “OB1”) where applicable. The hierarchical relation between the words and constituents is visible from the tree drawing, and is actually encoded in the parsed corpora. What is not included in the “standard” parsed corpora are the referential category labels “Identity” and “New” for the subject and the object in the example above. Chapter 6 describes how these are going to be added, so as to arrive at “enriched” texts.

If we suppose that we have such an “enriched” text, so that it includes both the syntactic information as well as the referential categories, then it is quite obvious that the suspicious context depicted in Figure 12 can be easily recognized: it involves an NP, which is a child node of a main clause (labelled “IP-MAT”), which has referential category “New”, and which is positioned somewhere after a sentence negator “NEG”, and this sentence negator is a child node of the main clause too. Writing an algorithm to recognize this and similar kinds of situations is not difficult at all, which shows that the solution offered in (131b) is good enough—at least for texts that have been parsed syntactically, and referentially.

Sentences with a quantifier context, such as (128b), can be recognized automatically too, but now the distinguishing factor is not the presence of a sentence negator, but the presence of a quantifier (which is marked by the label “Q” in the parsed corpora) as part of the subject. A conditional context, such as (128c), can be recognized by the presence of the conditional “if” (which is labelled as a preposition “P” that is followed by an adverbial clause marked “CP-ADV”). The kinds of modalities discussed by Karttunen, see examples (127a-f), all involve using a modal verb, and since the modal verbs are a limited set, these contexts can be recognized by checking for the presence of an “MD” verb from this set.

5.4.5 Conclusions

The previous sections show different ways of looking at the sufficiency of the Pentaset, which is a necessary condition for a good-enough set of referential categories, as formulated in (108c). We first looked at the relation between the categories of the Pentaset and the alternative sets that were discussed in 5.2, and found that the Pentaset is the most concise one. We took this comparison one step further in 5.4.2, where we saw that most of the information structure categories that are used by the other taxonomies can be derived by combining Pentaset categories with syntactic information. There were a few categories that could *not* be derived by the Pentaset (such as the diversification of “Inferred” into “PartWhole” and “Subset”), but I have argued that the underivable further diversifications are not significant from the point of view of information structure. We turned to the question whether generic noun phrases need to be treated differently from the others or not in section 5.4.3, and we concluded that it is enough to label generics as “Inert” in those contexts where they cannot be referred to later anymore (they are more attributive in those situations), and to label them as “New” in other contexts: they do lead to the creation of a mental entity in the situation model of a reader, but this mental entity is a *set* rather than one particular item from a set. It is generally not possible to derive the generic character of a noun phrase from the Pentaset category and the syntax, but I have argued that it is very unlikely that this difference is necessary for information structure purposes, so that we are still satisfying the requirements in (108) when we do not label generics separately as KIND or something similar. The quest for the sufficiency of the Pentaset finished in section 5.4.4, where we looked at referential islands: opaque contexts that often do not allow mental entities created in them to be referred back to. We have seen that the contexts in which this happens are determinable in “enriched parsed texts”: syntactically parsed texts that are enriched with the referential categories of the Pentaset. Since these contexts are automatically recognizable, there seems to be no need to introduce another referential category label for the noun phrases occurring in these contexts.

What these sections have shown, then, is that it is very likely that the Pentaset offers sufficient differentiation when it comes to alternative information state categories, such as those offered by the taxonomies we looked at. What remains to be shown, though, is whether the Pentaset categories (in combination with the syntactic information) are sufficient to derive the “higher order” notions used in information structure research: those of topic and focus.

5.5 Deriving topic and focus

The previous section showed the relationship between the Pentaset of referential state categories and the information states as defined by the taxonomies discussed in section 5.2. What we now turn to is a more experimental chapter, where we will see how the Pentaset, combined with syntactic information, relates to higher order information structure notions such as topic and focus. All this serves to underscore the hypothesis that sees focus domains as derivable from syntactic and referential

state information, which lies at the basis of the approaches to automatically look for presentational focus in chapter 8 and constituent focus in chapter 9.

We will look at two attempts that are concerned with deriving an approximation of the notion of “topic” (topic guessing algorithms and centering theory), and we then turn to a specific example of matching one construction (the copula clause) to the automatic determination of focus domains by making use of syntactic and referential state information.

5.5.1 Topic guessing

The most unmarked of the three focus articulations adopted in chapter 3 is that of topic-comment, and since this articulation not only defines the size of the focus domain (which is the predicate, the verb with its internal arguments), but also uses the information structural notion of “topic”, it would be good to see whether using syntactic and Pentaset information allows one to retrieve topics. The notion of “(aboutness) topic” can, if we loosely follow Reinhart’s (1981) definition, be summarized as “the entity that the utterance is about” (which is much in line with: Givón, 1983, Krifka, 2007, Neeleman et al., 2009). Vallduví (1990) argues that topics function as index cards in the mind of the addressee, specifying where new information should be stored. The topic-comment articulation, then, is a sentence with a topic, an entity that is already established in the mental model, about which the “comment” provides new information.

Eckhoff and Haug (2011) have for some years been working on an algorithm to guess what the topic of a sentence is. They report a 90% agreement between the outcome of their algorithm and that of human intuition. The rough structure of that algorithm is this:

- (132) *Algorithm to identify the aboutness topic (Eckhoff and Haug, 2011)*
- a. Is this a main clause but not a presentation construction?
 - b. Get topic candidates: main clause verb arguments that are linked to the preceding context.
 - c. Rank the candidates according to parameters:
 - i. information status
 - ii. animacy
 - iii. morphosyntactic realization
 - iv. saliency
 - v. syntactic relation
 - vi. word order
 - vii. antecedent properties

Their algorithm starts by checking whether a particular clause is a main clause, and if so, whether it is not a presentation construction (step a). If the clause is accepted, then step (132b) looks at all the arguments available for the main clause verb, and if they have an antecedent (which can be recognized by checking that their information status is OLD, as per the Proiel tagset in section 5.2.5), then they are kept separate for the next step, (132c), which ranks the candidates according to seven parameters. Almost all of these parameters are derivable from syntactic and referential state

information; the exception is “animacy” (c.ii), which is only partly encoded in the syntax (only third person singular pronouns differentiate for gender). What Eckhoff and Haug’s topic guessing algorithm illustrates is that the combination of (morpho)syntactic and referential information of the constituents under review and of their antecedents is by and large enough to give an extremely good guess of the notion “aboutness topic”.

Another attempt at guessing topics was implemented in the “Cesac” program that is briefly discussed in Komen (2009a). The syntactically annotated texts that were enriched with referential states discussed in section 5.2.4 could be automatically converted into a table where each row contained a main clause with a guess for the topic in that clause—provided Cesac had detected it as a topic-comment clause.

- (133) *Algorithm to determine the topic of a topic-comment clause (Komen, 2009a)*
- a. If this clause is declarative mood, continue with step (b)
 - b. Determine the number of NPs that have an antecedent:
 - zero: stop → this is not a topic-comment clause
 - one: stop → we **found** the topic!
 - else: order all the NPs in [empty > Dem > Pro > Dem+NP > other NP]
 - c. Determine the amount of NPs on the highest level:
 - one: stop → we **found** the topic!
 - else: continue with step (d)
 - d. Get the NP ranked highest in [subject > object > PP object]¹⁶

For each main clause that is found in the text, step (133a) determines if it is in declarative mood (this is information available from the syntactic encoding of the text). The next step (133b) checks all the constituents of the clause, and if they have the syntactic category of a Noun Phrase, and their referential state is such that they link back directly or through an anchor to the preceding context, then they enter a collection. The size of this collection determines how the algorithm proceeds: if the collection is empty, there is no topic candidate, which means that this cannot be a topic-comment clause, and if there is one topic candidate, then this must be the topic. If there are more candidates, then they are ordered according to the syntactic category of the NP, resembling Gundel’s (1993) givenness hierarchy (see 5.2.3). Step (133c) checks how many NPs in the collection have a syntactic category that is highest. If one is highest of all on this scale, we found the topic, but if this is not the case, then there is one more tie-breaker: step (133d) checks if one of the topmost NPs is a subject. If this is the case, then we found the topic; if not, then the algorithm is not able to determine the topic.

5.5.2 Centering theory

Centering theory aims at finding a topic in each sentence in a narrative, in order to detect topic continuity and various kinds of topic shifts (Grosz et al., 1995). Centering theory proper does not speak of “topic”, but seeks to determine what the “attention states” of entities in a clause are, which of them is the current and the following “center of attention”. Having found $C_f(U_n)$ a set of “forward looking” centers in sentence n , and having determined $C_b(U_{n+1})$ the “backward looking

center” in sentence $n+1$ (usually chosen from the forward looking centers in the previous sentence), it then determines “transition types” (“continuation”, “retaining” or “shifting”), the value of which depends on whether the center of attention is retained or shifts.¹⁷ Crucial for us to understand at this point is the way in which $C_b(U_n)$ is determined. This process runs along the following lines:

- (134) *Determining the center of attention in sentence n* (derived from: Grosz et al., 1995)
- a. Construct $C_f(U_{n-1})$: the set of forward looking centers in sentence $n-1$
 - i. Add all the referring expressions in sentence $n-1$
 - ii. Rank them according to criteria of category, syntax and so on
 - b. $C_b(U_n)$ becomes the highest ranked entity in $C_f(U_{n-1})$

The set of forward looking centers is filled with all the referring expressions in a sentence, which are then ranked by several criteria (we will come to that), and then the backward looking center of the next sentence picks the best candidate (the most salient one) from among the forward looking centers of the previous sentence. The criteria that are being used to rank the forward looking center entities and to choose the backward looking center are:

- (135) *Criteria for ranking the forward center and determining the backward looking center* (derived from: Grosz et al., 1995)
- a. Rank according to linguistic expression: Pronoun > Noun phrase
 - b. Rank according to grammatical role: Subject > Object > Other
 - c. If a constituent in $C_f(U_{n-1})$ is realized by a pronoun in sentence n , then $C_b(U_n)$ must be a pronoun (“Rule 1”)
 - d. The $C_b(U_n)$ is the entity that also exists in $C_f(U_{n-1})$ and is highest ranked in it

The factors used above to determine the ranking of the forward looking center are the form of the linguistic expression and the grammatical role, both of which are already part and parcel of the syntactically parsed English corpora. However, practical implementations of the centering theory added more criteria in order to have a more realistic ranking in the forward looking centers.

Beaver (2004) describes a constraint based implementation of centering, and proposes to rename the “backward looking center” into “topic”; the backward looking center is the “the most significant discourse entity under discussion in both the current and previous sentences”. Beaver redefinition of backward looking center into topic states:

- (136) *OT centering’s definition of “topic”* (Beaver, 2004)
- The *topic* of a sentence is the entity referred to in both the current and the previous sentence, such that the relevant referring expression in the previous sentence was minimally oblique.
- If there is no such entity, the topic is undefined.

The term “minimally oblique” points to the criteria that are used to rank the referring expressions in the previous sentence. The criteria used in OT centering, as far as they are important to determine what is “minimally oblique” are listed here:

- (137) *Constraints to determine what is minimally oblique in OT centering*
- a. PRO-TOP: The topic is pronominalized.
 - b. FAM-DEF: Each definite NP is familiar (the referent of the NP is familiar and no new information about the referent is supplied by the definite).
 - c. SUBJECT: The topic is in subject position

Without going into details about the way these constraints are used in the “COT” algorithm (the algorithm proposed by Beaver (2004), in which he uses the constraints above and several others to determine what the “best fit” in terms of coreference resolution is for a whole sentence), the constraints listed in (137) do not only need information about the syntax, such as the ranking criteria in (135), but they also need to have referential information: the FAM-DEF constraint needs to be able to evaluate whether noun phrases have antecedents, and what their antecedents are. An implementation of centering for German by Strube and Hahn (1999) also proposed that the “information status” of the entities is needed to help determine their ranking in the set of forward looking centers.

The point of this section has been to show that a successful theory such as centering determines its “topic” by taking into account exactly those features that become available in syntactically annotated texts that are enriched with referential information. The ranking of the topic candidate constituents is based on: the categories of the noun phrases (e.g. “pronoun”, “demonstrative”), the referential states of the noun phrases (basically whether the referent of a noun phrase is “familiar” or not), and the grammatical role played by the noun phrases (such as “subject” and “object”).¹⁸

5.5.3 Deriving focus domains

I have argued in previous sections that the information provided by the syntactically parsed corpora, when enriched with the referential state categories from the Pentaset, provides sufficient material to determine “higher order” notions within the information structure research such as topic and focus. The previous two sections on topic guessing and centering have zoomed in on ways to derive “topic”, and this section concentrates on ways to derive focus. Specifically, this section offers a case study on how focus domains (and consequently focus articulations) can be determined on the basis of syntactic and referential state information. The case study concentrates on one particular construction, the copula clause, and the results are promising enough to increase the likelihood that the Pentaset enrichments are, indeed, sufficient when it comes to determining focus articulations.

5.5.3.1 Copula clauses in general

The general strategy behind the answer is that the combination of syntax with referential information should allow one to determine the focus *domains*—at least to some extent. What we will do here is look at one particular syntactic construction, the copula clause, and see how we can combine syntactic and referential information to provide a mapping between this construction and a focus domain division (see also section 3.2.2.1). The definition of copula clauses we will use here is quite a

generic one: XP + *be* + YP (all sentences that consist of two constituents and a form of the verb “to be”). We will restrict the possible values of XP and YP to noun phrases, prepositional phrases and APs (excluding clausal XPs) for this particular case study.

If we take into account the different possible syntactic categories of XP and YP, and combine that with the possible different referential category values, we end up with quite a lot of combinations, but if we group several of these together, we get Table 16. The process of checking the possible focus articulations for each row in the table has been done mostly with texts that have been enriched with Pentaset information, and the word orders shown in the table are the *surface* word orders as found in these texts.

Table 16 Types of “XP *be* YP” copula clauses depending on the referential and syntactic categories of their components (surface word orders)

#	XP		YP		Focus domain	Articulation
	Syntax	Pentaset	Syntax	Pentaset		
a	NP	Identity	AP	-	predicate	TC
b	NP	Identity	PP	Assumed	predicate	TC
c	NP	Identity	NP	Inert	predicate	TC
d	NP	Identity	NP	New	complement	CF
e	NP	New*	AP	-	core	PF
f	NP	New	NP	Inert	core	PF
g	NP	New*	NP	New	core	Thetic
h	AP	-	NP	New	core	PF
j	NP	Inert	AP	-	predicate	TC
k	NP	Inferred	AP	-	predicate	TC
l	NP	Inferred	NP	Inert	predicate	TC
m	NP	Inferred	NP	New	complement	CF
n	NP	Assumed	AP	-	predicate	TC
o	NP	Assumed	NP	Inert	predicate	TC
p	NP	Assumed	NP	New	complement	CF
q	PP	Assumed	NP	New	complement	CF

Each line in Table 16 represents one possible combination of syntactic and referential categories for XP and YP, which is then followed by the focus domain belonging to this representation.¹⁹ Examples for each of the combinations above are presented here, where the subject in each sentence is depicted in bold-face, and the focus domain is indicated by square brackets:

- (138) a. In autumn and winter **the corn** [was bruised]. [fleming-1886:377]
 b. **The driver of that car** [is from Finland].
 c. **Diodorus** [was a native of Sicily]. [long-1866:9]
 d. A stiff clay produces a coarse barley; a light chalk a light grain; and a loamy land a full, plump grain; [fleming-1886:49]
these are [only a few examples of many which might be quoted].

- e. In very wet years, and especially when lands have been flooded,
[**parasitic diseases of plants** are most common]. [flaming-1886:58]
- f. In the next year [**Marius** was consul]. [long-1866:257]
- g. [**The first teacher of Crassus** was L. Caelius Antipater the historian].
[long-1866:338]
- h. [Next in importance to food and water in stable-kept horses is **grooming**].
[flaming-1886:472]
- i. In this time of year, [**it** is cold].
- j. What is the weather_i in Siberia? In the winter, **it**_i [is cold].
- k. In good upland hay the flowering heads of the grasses should be plentiful.
Meadow hay [is long]. [flaming-1886:128]
- l. Grasses are divided into natural and artificial.
The former [are true grasses]. [flaming-1886:110]
- m. There was also a rising of the slaves in the west part of the island, about
Segeste and Lilybaeum Marsala, and other neighbouring parts.
The leader was [Athenion, a Cilician born, and the bailiff of two rich
brothers]. [long-1866:117]
- n. **The world** [is beautiful].
- o. **The earth** [is a planet].
- p. **This book** is [the answer to your problems].
- q. [Under the table] is **a good place to hide**.

We start with the XP of the copula clause being a noun phrase with referential category “Identity”, leading to examples (138a-d). The first two of these, (138a,b), where the YP is an AP or a PP, are examples of *predicational* copula clauses (Akmajian, 1979). (Remember that points of departure like “in autumn and winter” are not part of the core, and do not co-determine the focus articulation.) The third one, (138c), seems to be more specificational, but since the YP has a referential category of “Inert”, the whole of the copula clause still is predicational, having a topic-comment articulation. Of those starting with an “Identity” subject only the fourth one, example (138d), has a constituent focus articulation, and would be called “specificational” or “identificational” by researchers like Akmajian (1979) and Mikkelsen (2005).

Next we turn to examples (138e-g), which illustrate the situation where the first XP of the copula clause is a noun phrase with referential category “New”. These situations generally lead to a focus domain spanning the whole of the core, which is thethetic focus articulation, but some are more clearly presentational focus (marked “PF”; these situations are clearly used to introduce a new participant). There is one exception (which is why some situations are marked “New*”): the referentially “New” subject NP may *not* be one that generates a variable, such as a free relative; we will come back to that category later in this section.

Example (138h) offers a clear case of presentational focus, where we have the copula clause word order AP *be* NP, and where the NP has referential category “New”. The focus domain is the whole core, but it is clear that this construction serves to introduce a new participant (in this case the “participant” is a generic noun

grooming, but as we have seen in section 5.4.3, generics can be treated like other NPs when it comes to information structure).

If the first XP has a subject NP with referential state “Inert”, like in (138i), the question is whether the subject is part of the focus domain or not. If we do *not* count it as part of the focus domain, we arrive at a topic-comment articulation, but then the referentially “Inert” subject would be the “topic”. This is not really possible, so we have to conclude that in this kind of situation the focus domain spans the whole core, resulting in “thetic” articulation. We should be aware, though, for a seemingly similar but fundamentally different construction like the one in (138j): the pronoun *it* is no longer “Inert”, but refers back to “weather” in the previous sentence, so that it has referential category “Identity”. This kind of constellation is what we have in (138a) too, and the focus articulation is like there: topic-comment.

The situations where the first XP is a subject NP with referential state “Inferred” closely follow the pattern of those with referential state “Identity”: when the YP is an AP, as in (138k), we get the topic-comment articulation, which we also get when the YP is a noun phrase, with referential category “Inert”, as in (138k), while there is constituent focus when the noun phrase has referential category “New”, as in (138m).

When the first XP is a subject noun phrase with referential state “Assumed”, the pattern matches that of “Identity” and “Inferred” subject noun phrases: with an AP as YP, as in (138n), the topic-comment articulation results. The same happens when the YP is a noun phrase with referential category “Inert” in (138o). When we have an “Assumed” subject like “This book” in (138p), and the complement is completely new, then the focus domain only comprises the complement, and we have constituent focus, comparable to (138d) and (138m).

I have not come across examples of PP-initial copula clauses in the enriched English texts, but there is the often-cited example of “Under the table is a good place to hide”, where the PP is argued to function as subject (for instance Faarlund, 1990: 112). More data from the historical corpora would be needed to classify these kinds of examples as well as those where there is a clausal subject

5.5.3.2 Copula clauses and variable creating expressions

There is a syntactically distinct category of XP (subject) or YP (complement) noun phrases that needs separate attention, and that is the category of those that create a variable. Examples of variable creating noun phrases are free relatives, such as “what I wanted to say” in (139a-c), which are described, for instance by Bresnan & Grimshaw (1978). While all free relatives lead to the *creation* of a variable, the *resolution* of them only occurs in specific contexts. Whenever the context in which a free relative occurs leads to the resolution of the variable, there is constituent focus.²⁰ Consider the following examples of copula clauses with a free relative as example of a variable creating NP (again with subjects bolded and the focus domains demarcated by square brackets):

- (139) a. [**What I wanted to say** is good].
 b. **What I wanted to say** is [a few words].
 c. [**Just a few words**] is what I wanted to say.
 d. [**What you see** is what you get].
 e. **What I wanted to say** is [this]: “Linguistics is great”. (*resolution*)
 f. “People are great”. [**That**] is what I wanted to say. (*resolution*)
 g. Is that the house? [**The kitchen**] is what I wanted to see. (*resolution*)

When the variable created by the free relative is not resolved, such as in (139a-d), then the copula clauses satisfy the mapping described in Table 16: (139a-c) are of type “e” and lead to presentational focus, while (139d) is of type “g” and leads to athetic articulation. In the other examples (139e-g) the variable resolution takes place within the copula clause, so that they are examples of constituent focus. Example (139e) is of type “g” (according to Table 16), since it has a referential “New” subject NP and a referential “New” complement NP. This last examples does not map onto the default “thetic” articulation, since the constituent-highlighting achieved by the fact that the complement “fills in” the variable created by the subject overrules: there is constituent focus articulation. The examples (139f,g) have the free relative as complement, illustrating that the focus domain now is the subject, which supplies the value of the variable that is created by the free relative. Example (139f) compares to type “d” of Table 16 (the subject has referential category “Identity”), and example (139g) compares to type “m” of Table 16 (where the subject has referential category “Inferred”).

Since copula clauses with free relatives occur in two flavours (those that lead to variable resolution and those that do not), determining the focus domain requires a step for which referential state information is needed: we need to determine the state of the NP (be it the subject or the complement) in the copula clause that is *not* the free relative. If the referential state of that NP shows that it either creates a mental entity (the state would be “New”) or links to an existing one (the stat can be “Assumed”, “Inferred” or “Identity”), then we have a variable resolution situation, and, consequently, constituent focus. If this is not the case, then the focus articulation can be determined in the “normal” way as described in section 5.5.3.1, where we can accept the free relative NP as having referential state “New”. In sum, the focus domains of the situations in (139) can all be determined programmatically.

The kind of variable-creating noun phrases that lead to situations exemplified in (139) is not limited to free relatives. There are at least four categories of variable-creating noun phrases that can be discerned:

- (140) *Variable-creating noun phrases*
- a. Free relative
 - b. Restricted relative clause with a generic head
 - c. Definite noun phrase with a verbal head noun
 - d. A pronoun with referential category “Identity” pointing back to a variable-creating noun phrase in the preceding context

Examples of free relatives have been provided in (139), but the category (140b) of restrictive relative clauses with a generic head achieve the same effect. The free

relative “what I wanted to say” is equivalent to “the **thing** I wanted to say”; “who I saw yesterday” is equivalent to “the **person** I saw yesterday”; “where I went to last night” is equivalent to “the **place** I went to last night”. The generic head nouns do not provide a specific enough value for the resolution of the variable, created by the relative clause, to be reached. The identification of copula clauses with a restrictive relative clause that has a generic head noun at first glance seems to involve two steps in the parsed English corpora: one would have to identify that the noun phrase in the copula clause (a) has a relative clause, and (b) has a generic head noun. The identification of generic head nouns is the challenge here, but this does not seem to be an undoable task, since the number of generic head nouns is probably restricted.²¹ However, the challenge of identifying generic heads can be circumvented, since we can generalize that *any* copula clause that has two noun phrase, one of which contains a *restrictive* relative clause, must always have a constituent focus articulation: the complement (or subject) will always provide the more detailed value for the variable created in the relative clause of the subject (or complement). The fact that a relative clause is *restrictive* already implies that the head noun is more generic, and needs restriction to reach identification. Since restrictive relative clauses are marked as such in the English parsed corpora, variable creating noun phrases of the type in (140b) can be recognized programmatically.

We have already seen the category of (140c) exemplified in chapter 3, section 3.2.2.1, in the form of “the murderer”. This is a definite noun phrase with a head noun that is derived from a verb with the agentive suffix *-er*.²² Such a noun phrase really is a shortcut to type (140b) “the person who killed (Mr. X)” and ultimately to type (140a) “who killed (Mr. X)”, which means that it too is a variable creating expression, and leads to constituent focus in copula clauses like “The murderer is John” and “John is the thinker of the family”. Not only verbal nouns like “murderer”, “killer”, “sleeper”, “walker” count, but nominalized past participles like “deceased” (“the deceased is John”) act the same way. The identification of variable-creating expressions of these kinds of verbal nominalizations requires *morphological* information about the nouns we encounter in the texts we search. Such information should be regarded as belonging to the realm of “syntax”, so that we can still argue that the combination of syntactic and referential information is sufficient to determine the focus articulation of the kinds of copula clauses containing type (140c) subjects or complements.²³

The final category we need to address is that of (140d): pronouns that have referential category “Identity”, but that link back to a variable creating definite noun phrase of type (140c). Examples of such a situation are these:

- (141) a. I saw the *murderer* on television yesterday. **It** is [John].
 b. A: “I know John and James. Do you know who the *murderer* is?”
 B: “Yes, **it** is [John].”
-

Both examples in (141) have a variable creating definite noun phrase “the murderer” in one clause, which is referred back to by a pronoun “it” in the ensuing copula

clause. The first copula clause is of the type “d” in Table 16 (a subject with “Identity” category and a complement that is referentially “New”), but the second one does not occur in Table 16: it has both subject and complement with a referential category of “Identity”. The focus articulation of the former type coincides with that in Table 16, so that no additional measures are needed to recognize it programmatically. The focus articulation of the latter type does not need additional measures either: whenever the XP and YP constituents in a copula clause are both noun phrases with an “Identity” referential category, either the first provides a value for the variable in the referential chain of the second or the second for the first, as can be seen from the following example:

- (142) a. How moche rather our mother holy chyrche which is the spouse of christ,
hath an heed of her owne; that is to saye the pope.
-
- b. And yet neuerthelesse [**chryst Iesu hyr housbande**] is her heed.
[fisher-e1-h:134]

The noun phrase *Christ Jesus her husband* in (142b) has referential status “Identity”, and links back to *Christ* in (142a), while *her head* in (142b) links back with “Identity” to *a head of her own* in (142a). While this is a situation of two “Identity” noun phrases in a copula clause, in this situation the first noun phrase provides the value for the variable that was created in the referential chain of the second noun phrase. The creation of the variable in (142a) does not result from the use of an agentive noun (such as “murderer” in (141)), but starts with the indefinite noun phrase *a head of her own*, which evokes the question who this head is. A first possibility for the value is offered by the end of (142a): *the pope*, but then (142b) offers another value for this variable *Christ Jesus her husband*.²⁴ In order to be able to determine the focus domain for the IdentityNP-IdentityNP type copula clauses in (140d), then, we need to know which of the two noun phrases links back to a variable-creating noun phrase in the preceding context. This shows the necessity of being able to “follow” the chain; to look back at the syntactic and referential situation of an antecedent noun phrase. The way by which the parsed English texts will be enriched described in chapter 6 and the methods that are proposed to search in these texts (chapter 7) make it possible to annotate the location of antecedents and to “follow” antecedent chains.

The exercise on matching one syntactic construction (the “XP *be* YP” copula clause) onto all possible focus articulations by making use of the available syntactic and referential information has worked out quite well, which increases the plausibility that syntax and referential categories in general determine the focus structure of a clause. This matter needs more verification in future research, where perhaps other constructions could be reviewed in a manner like the one used here, but for now it seems reasonable to say that the Pentaset satisfies the sufficiency condition in (108c), and we can go ahead and enrich the existing English parsed corpora with this kind of referential category encoding.

5.6 Discussion

This chapter is the first step in the corpus approach of looking for changes in English focus: we endeavour to enrich existing corpora with the minimal amount of information needed to automatically determine the focus articulations. The first step laid in this chapter involves a thorough definition of the kind of annotation we want to enrich the existing syntactically parsed corpora with. Having reviewed several candidates—theories that define cognitive states or information states of referring expressions—a minimal set of five “referential state” primitives has emerged: the Pentaset.

This chapter has used several different perspectives to show that this “Pentaset” can indeed be regarded as a set of primitives: the Pentaset categories are more concise than other taxonomies (5.4.1), the different cognitive and information states used by other taxonomies can be derived from the Pentaset (5.4.2), it is possible to calculate several measures for the notion of “topic” by combining Pentaset with syntactic information (5.5.1 and 5.5.2), and it seems to be possible to map syntactic constructions to focus domains by making use of the Pentaset categories (5.5.3). We have also looked at generics as well as entities created in referential islands (5.4.3 and 5.4.4), and we have concluded that it does not seem likely these categories need to be added to the set of referential primitives.

With a clearly defined set of referential states in place, the next chapters show how we can semi-automatically add referential state information to the existing parsed texts (6), and how we can search the enriched texts for combinations of syntactic and referential information (7), forming the prelude to the actual corpus research described in chapter 8-9.

¹ Section 5.4.2 shows to some extent how combining syntax, semantics and referential states leads to finer-grained taxonomies of information state categories, which underscores the point of view in this chapter that information structure is compositional (since these latter categories are definable in terms of syntax, coreference and referential state categories).

² Gundel et al state that they “make only minimal assumptions ... about the representation of referents in long- and short-term memory”. So when they state that referents with a particular status are e.g. in short-term memory, then this is part of the model they posit, and not necessarily demonstrated by experiments.

³ There are some tools available (such as MMAX) that facilitate manual annotation of coreference links, but none of the available ones were completely “ready to go”, so that some adaptation would have been necessary anyway to use them for the tasks we were planning to do (Müller and Strube, 2001, 2006). MMAX, for instance, takes as a starting point unparsed text, whereas we start from syntactically parsed text.

⁴ The interrater agreement of the OE text “Apollonius” resulted in values for Cohen’s kappa ranging from .198 (slight agreement) to .629 (substantial agreement).

⁵ One version distinguished information state OLD from OLD-inactive, where the former has antecedents within a frame of 15 preceding sentences and the latter antecedents that are further away.

⁶ Right now the “inert” constituents do not receive any tag (personal communication).

⁷ The Pentaset and the initial PROIEL tagset both differentiate between “OLD” (Pentaset “Identity”), “ACC-inf” (“Inferred”), “NEW” (“New”). PROIEL divides the Pentaset’s “Assumed” into “ACC-sit” and “ACC-gen”, and it does not have an equivalent for the Pentaset’s “Inert”.

⁸ Note that this kind of reasoning disfavours “optional” inferences: the kind of slots that *could* be there, but that do not necessarily belong to the standard model of a situation. The mention of *restaurant* evokes certain slots that really belong to a restaurant (although this may differ between cultures and in time), such as *table*, *waiter* and *bill*. An optional slot may be a *playground*: many restaurants have them, but not all, and they are not evoked in a standard way when “restaurant” is mentioned (well, they are if a particular restaurant such as “MacDonalds” is mentioned).

⁹ The sets of the first and second noun phrases may also be identical, witness the following example from the English student learner’s database created by van Vuuren (2012):

- (i) The knight was brought up when England still fought **a lot of battles**.
- (ii) In contrast, his son was taught how to live life at the court, for, due to **fewer battles**, courtlife became more important.

The noun phrase *a lot of battles* in line (i) belongs to the set of “battles”, while the second noun phrase *fewer battles* in line (ii) belongs to the same set. The referential state of the second noun phrase will be labelled as “Inferred”, since the first mention *a lot of battles* evokes the larger set of *battles*, of which the second noun phrase is another subset.

¹⁰ Noun phrases with the category “Inert” are the linguistic equivalent of short-lived particles like positrons: they are destined not to survive in time, but do leave their “impression” (attributive character) in the world around.

¹¹ The observant reader may note that “Triocala ... was naturally a strong place” is followed by the clause “It was so called”, where “so” somehow relates back to “a strong place”. However, referential categories are, for the moment, restricted to noun phrases, which “so” clearly is not. Even if we were to extend referential categories to be attached to adverbs too, the antecedent of “so” should probably *not* be a mental entity made for “a strong place”, but only the attribute “strong”, and it is not clear that attributes have a kind of “life of their own” within the situation model; I would say they can only exist as attachments to mental entities.

¹² The word *there* is not treated as a place adverb in English sentences like (122d), since it has lost its ability to refer to a particular place. The parsed corpora treat these instances of *there* as expletive subjects.

¹³ The OLD-inactive information state is described as a “subtag” in the Proiel’s annotation guidelines and should be used for antecedents that are further away than a measure that is to be determined experimentally, and is currently set to “13 sentences” (PROIEL, 2011).

¹⁴ Thanks to Ans van Kemenade for coining this vivid term.

¹⁵ This particular example is taken from [long-1866:364].

¹⁶ Practice with a number of texts from different time periods has not come up with a situation where this last step in the algorithm is inconclusive: I have not come across a situation where more than one NP is left at the start of step (d), while none of them is the subject.

¹⁷ Instead of looking at the realm of the “sentence”, centering restricts itself to what Grosz et al call the “utterance” (from which the “U” derives), which can be compared to what we would call the “finite clause”.

¹⁸ The core constraints proposed by Beaver (2004) and their relation to syntactic and referential information are the following:

Constraint	Meaning	Relation to syntax / referential category
AGREE	Anaphoric expressions agree with antecedents in gender and number	Each referring NP is enriched with a link to its antecedent
DISJOINT	Co-arguments of a predicate are disjoint (<i>principle B</i> effect)	When NPs do not have the same referent, they are not on the same coreference chain
PROTOPIC	The topic is pronominalized	The syntactic category of each NP is in the syntactic encoding of the text
FAMDEF	Each definite NP is “familiar”: (a) referent is mentioned in the discourse before (b) the NP does not provide new information about the referent	An NP is “familiar” if it has the referential state of “Identity” (see section 5.4.1)
COHERE	The topic of the current sentence is the topic of the previous one	The referential link of the topic NP points to the topic NP of the previous sentence
ALIGN	The topic is in subject position	The grammatical category of each NP can be derived straightforwardly from the syntactic annotation

¹⁹ Not all logically possible combinations (that would be $3*5*3*5=225$) are presented in the table, since not all of them are possible, and not all of them have been identified in the enriched texts available. The examples have been found by using a query of the kind discussed in chapter 7. The query used for the copula construction can be reviewed in appendix 14.2.1.

²⁰ Huddleston and Pullum (2002) distinguish between two types of free relatives: “fused relatives”, which are true NPs and occur in variable resolution contexts, versus clausal free relatives, which do not occur in variable resolution contexts. I would like to keep apart the free relatives as such (all of which *create* a variable) and the context in which they occur (some contexts lead to variable resolution; others do not).

²¹ One would have to make a list of all generic head nouns, and then label these nouns with a feature like “generic head” in the parsed English corpora.

²² An agentive suffix is a derivational suffix that transforms a verb into a noun that identifies the agent performing the action described by the verb. The agentive suffix can also be in the form *-or* as in *actor*, *surveyor*. The process of forming variable creating NPs from verbs through derivational morphology (140c) is less flexible than that of using free relatives (140b) or restrictive relative clauses (140c).

²³ From a practical point of view, however, we do have a bit of a problem: the parsed English corpora do not provide the (derivational) morphological information we need to have in order to determine whether a head noun has an agentive suffix or is a nominalization of a past participle. This purely *practical* lack of information does not conflict with the *theoretical* claim that syntax + referential category is sufficient to determine the focus domains in copular clauses.

²⁴ The context is that of a written-out sermon, where the preacher argues in favour of the Catholic Church with the pope as head and against the teachings of Martin Luther.

Part III

Methodology

In order to answer the major research question (11) on how the interaction between syntax and focus changed in English over time, we need to be able to *quantify* changes that took place in the expression of focus, and we are in the middle of an attempt to do that by automatically determining focus domains (see chapter 3). This method only works if we have more than the available syntactic information in the parsed texts: we need to have referential information of each referring expression, and this information boils down to: (a) the referential state, and (b) a link to a possible antecedent.

Chapter 5 thoroughly derived a minimal set of referential states, the “Pentaset”, and this chapter shows how the existing corpora can be annotated with the Pentaset. This forms the onset for the next chapter, where we will see how the newly enriched texts can be searched for the changes in focus we are looking for.

6.1 How to add referential state primitives

The task we have to accomplish in this chapter is finding a method to add referential information to the parsed texts, and this information consists of two elements: each noun phrase needs to have a label with its referential state (taken from the Pentaset), and if a noun phrase has an antecedent, it needs to have a pointer to that antecedent.

The process of finding out which constituent refers back to which other constituent is known from computational linguistics as “coreference resolution”. Coreference resolution, as well as the more limited pronominal anaphor resolution, have a history of algorithms, which differ in their effectiveness. Hobbs’ algorithm, for instance, attempts to find the correct antecedents for 3rd person anaphoric pronouns, and reports an accuracy of 88%, provided that perfect syntactical and morphological information are present (Hobbs, 1978). The Resolution of Anaphora Procedure (RAP) provided by Lappin and Leass (1994), assuming the data have been parsed through a full syntactic parser and a morphological analyzer, report 86% accuracy. More recent algorithms are stochastically oriented, they don’t need their data to be parsed syntactically or morphologically before hand, and they reach an accuracy that approaches 80% (Kehler et al., 2004, Soon et al., 2001). Other recent algorithms combine statistics with linguistic information; the “kernel-based method” starts from scratch (raw text), derives a syntactic structure using existing tools, and then uses the “syntactic tree kernels” stochastically in the coreference resolution step (Versley et al., 2008). Many of the automatic coreference resolution approaches are limited to resolving only a subset of noun phrases: Hoste (2005: 173), for instance, only looks at “coreferential information for pronominal, proper noun and common noun coreferences”. What these (and similar) approaches have in common is their overall aim: resolve coreference as much as possible automatically, in order to serve as a component of larger systems with particular purposes like text

summarisation, term extraction or text categorisation (Mitkov et al., 2007). One application of this is identifying the particular piece of information a user is looking for, and then providing him with a link to it (i.e. internet searching).

We need to resolve coreference for a different purpose. What we want to know for each sentence in a text is how the old, new, prominent and/or topical information is ordered, and how this interacts with syntax. The resulting picture should ultimately help us understand what the meaning is of a particular word order or construction, including the relative importance and the topicality of the constituents involved. None of the existing automatic coreference resolution methods finds the correct coreference information for all the constituents in a text. Their results may be 75-80% or more correct, but it is not clear whether and how the remaining 20-25% incorrectly labelled constituents might mislead us when we try to answer the form-meaning puzzle we are interested in. That is why our overall aim is to get near perfect coreference resolution.

There are, in principle, several solutions to overcome the problem of false-positives mentioned above, each with their limitations. One solution might be to supply all the references manually from the start. We have tried this approach, but found it too labour intensive, and too prone to errors. Another solution would be to use an existing automatic algorithm anyway, and check all the references it found manually. This would require a checking process where *all* the references that have been made are suspicious—again a huge and labour intensive task. A third solution, which is the approach advocated in this dissertation, is to opt for a semi-automatic process consisting of the following two main steps:

1. The computer resolves as much as possible automatically.
2. The computer asks the user's input for situations it recognizes as suspicious.

This approach should be less labour intensive than the previous two approaches, since the suspicious coreference situations are automatically selected by the computer and presented to the user. There are more advantages to a semi-automatic approach, but these should be regarded as side effects. Such an approach forces us to specify the factors contributing to the coreference resolution, and it forces us to define suspicious situations. The result is that we gain insight into the coreference specifics of the language we are working on. The Cesax algorithm I propose opts for a constraint-based automatic part of the coreference resolution.¹ Such an approach allows one to easily change the relative contribution of the different factors, as these will vary with the different periods of English, and are likely to be different for other languages.

Existing constraint-based coreference resolution algorithms have been taking centering theory as a starting point (Beaver, 2004, Gegg-Harrison and Byron, 2006, Grosz et al., 1995). What they typically do is provide the harmonically best aligned set of referents for all noun phrases in one clause. Beaver's (2004) COT algorithm, for instance, which is based on the centering approach, tries to find a best match for all noun phrases in one sentence² at a time.³ However, what our semi-automatic coreference resolution wants is an evaluation of the coreference situation one

constituent at a time. If we would work clause by clause, then suspicious situations would require the user to manually resolve the coreference of all the clause's constituents at the same time, which we regard to be undesirable. This is one of the main reasons why we need to develop our own coreference resolution algorithm.

Another reason why we need to divert from existing constraint-based algorithms has to do with the kind of constraints we want to use. Since we want to use the results of the coreference resolution to say something about information ordering, which includes a notion such as topic, we should not use constraints that already include a notion of "topic" in them. That is the reason why we don't use constraints such as PRO-TOP and COHERE from centering. Instead, we need to use more primitive constraints, on which these higher level constraints are probably based.

The approach advocated in this dissertation, then, works constituent by constituent, and uses a set of hierarchically ordered constraints, which are derived from the morphological and syntactical information we can glean from the existing corpora. The approach recognizes suspicious situations and asks the user to solve a particular coreference situation when it recognizes its own inability to do so correctly.

Section 6.2 introduces the data we are working with and zooms in on the coreference resolution task that needs to be done. Section 6.3 gives a short overview of the coreference resolution algorithm proposed, and then focuses on the individual parts of this algorithm in subsequent subsections. This section includes a discussion of the constraints that are being used and the suspicious situations it recognizes. Section 6.4 presents a case study of our algorithm: the annotation of a chapter from an 18-th century history book. Conclusions and a discussion for further research are then presented in section 6.5.

6.2 The data and the task

The syntactically annotated English texts we are dealing with come in a labelled bracketing format, which use a tagset defined for the Penn-Helsinki-York corpora (Kroch and Taylor, 2000). This tagset is larger than the one used for the Wall-Street Journal corpus (Marcus et al., 1993). Example (143) contains a sentence from a text in this format. The Treebank format provides a hierarchy of a sentence's syntax using two kinds of nodes: (a) parent nodes labelled with a syntactical category, which contain one or more children, and (b) end nodes labelled with a word category, containing one word in the vernacular. An example of the first kind of node is the subject of the main clause, which is labelled NP-SBJ, and which contains two child nodes: a PRO\$ and a N. An example of an end node is the PRO\$ node, which contains the vernacular word *my*.

(143) Penn-Helsinki-York Treebank example

```

(IP-MAT (CONJ But)
  (NP-SBJ (PROS my) (N Partner))
  (IP-PPL (VAG remembering)
    (. ,))
  (CP-THT (C that)
    (IP-SUB
      (IP-SUB-1
        (NP-SBJ (PRO I))
        (HVD had)
        (VBN given)
        (NP-OBJ (NUM 500) (NS Moidores))
        (PP (P to) (NP (D the) (N Prior)
          (PP (P of) (NP (D the) (N Monastery)
            (PP (P of) (NP (D the) (NPRS Augustines))))))))))
      (CODE <$$font>)
      (. ,))
      (CONJP (CONJ and)
        (IP-SUB=1
          (NP-OBJ (NUM 272)) (PP (P to) (NP (D the) (ADJ Poor))))))
      (. ,))
      (VBD went)
      (PP (P to) (NP (D the) (N Monastery)))
      (. ,))
    )
  )

```

The syntactical hierarchy of clauses is encoded through the labelled bracketing system. The syntactic system that is used assumes a relatively flat structure, leaving open how the constituents within one clause are hierarchically structured with respect to one another. The most important parts of syntactic information encoded in the labels of the Treebank data are the following:

- Basic syntactic constituent types such as IP, NP, CP, PP etc.
- Optionally, some functional information is added, such as: SBJ (subject) and OBJ (object) for noun phrases, and MAT (matrix/main clause) or SUB (sub-clause) for inflectional phrases (IPs).
- The word category for the end nodes, e.g.: P (preposition), D (determiner) etc.
- Number marking for nouns and proper nouns. Singular nouns have the label N, whereas plural ones have NS. Likewise, singular proper names have the label NPR, while plural ones have NPRS.

Having explained the basics of the Treebank format we have to deal with, we will now look at the task of the coreference algorithm. What the algorithm wants to do is add features related to coreferentiality to particular constituent nodes. The Penn-Helsinki Treebank format only offers limited opportunities to fulfil this task. The Treebank format allows adding features at the level of the label. For example, a label like NP-SBJ-PGN:3ns-CREF:IDT could be constructed, which would then be interpreted as an NP that functions as a subject (the SBJ part), which is 3rd person neuter singular (3ns), and which refers back with an “Identity” link (the “CREF” equals “IDT”) to an antecedent which has the identifying code of “1245”. It would be possible to query texts containing these kinds of labels with an engine like “CorpusSearch” (Randall et al., 2005). Consider the query to look for *subject* noun

phrases with a *referential status* of “Identity” that are 3rd person neuter singular in (144).

(144) *CorpusSearch2 query example*

```
1 query: (IP-MAT iDoms NP-SBJ*) AND
2 (NP-SBJ* HasLabel *CREF:IDT*) AND
3 (NP-SBJ* HasLabel *PGN:3ms*)
```

One complication in the approach above is that of encoding and querying referential pointers. Each constituent can not only take features, but also a reference to another constituent. In order to facilitate such a referential system, we would need to equip each constituent with a unique identifier, such as NP-SBJ-ID:101. The constituent with this identifier could then be referred to from another constituent by using its identification number, for example: NP-OB1-ID:124-REF:101. Querying these kinds of references using *CorpusSearch*, however, will become a difficult task.

I am arguing that an *xml* format, for instance one like the Text Encoding Initiative (TEI-P5), is a better choice for holding the referentially enriched parsed English texts than the labelled bracketing format is (Sperberg-McQueen and Burnard, 2009). The two formats are comparable in that they both allow nesting of constituents using nodes, and they both allow adding features and cross-referencing in principle, but *xml* has several practical advantages. Since *xml* has become a standard format, a wide variety of technologies are available to process *xml* coded texts, whereas the options for the labelled bracketing format as it is used in the available English treebanks are limited to one program (*CorpusSearch*), which, according to its website, is not being developed any further. The labelled bracketing as it exists, however, would need to be extended so that it could facilitate the features and referencing that come with the referential status enrichment. Processing of such an extended labelled bracketing format would then require the development of new software. All this is not necessary when *xml* is chosen, since tools (such as the Xpath and Xquery languages) are available to query *xml* encoded texts, and several different *xml* standards exist that are meant to be used for searchable texts.

All the *xml* formats allow adding features and referencing either through attributes or special daughter nodes, so that there seems to be no necessity to create a new standard. There is a two-way distinction between formats using stand-off *xml* and nested *xml*, and this difference is discussed further in chapter 7. The implementation of TEI-P5 described in this dissertation uses the TEI-P5’s embedded tree tag set. The main elements of the tags that are used to encode the hierarchical phrase structure of the parsed corpora are shown in (145).

(145) *Main elements of the TEI-P5 embedded tree tag set*

- <eLeaf> An end-node containing one text element (a word or punctuation mark).
- <eTree> A hierarchical element from the phrase-structure. An <eTree> may contain one <eLeaf> or one or more <eTree> elements.
- <forest> This typically is a line in a text, and may contain one or more <eTree> elements.

`<forestGrp>` A collection of `<forest>` elements constituting one text groups together in a `<forestGrp>` element.

Each text contains one `<forestGrp>` tag, which has one `<forest>` child for each line in the text. These `<forest>` tags contain a hierarchical structure of `<eTree>` elements, which indicate the phrase structure of this sentence, as for example in Figure 13.



Figure 13 Conversion from (a) psd format (right) to (b) psdx format (left)

The actual words and punctuation marks of the sentence are found in the `<eLeaf>` elements. The `<forest>`, `<eTree>` and `<eLeaf>` tags themselves contain a limited number of attributes in order to facilitate querying them. Each `<forest>` element, for instance, contains identifiers `Location` and `TextId` which contain information similar to that in the ID label of the labelled bracketing format. Each `<eTree>` contains a numerical identifier attribute, which serves to facilitate the coreference information we want to add. The `psdx` format allows an unlimited number of features, divided into feature sets, to be added, as illustrated in (146).

(146) *Coreference information in the psdx format*

```
<fs type="coref">
  <f name="RefType" value="Identity" />
  <f name="IPdist" value="20" />
</fs>
<ref target="321">
<fs type="NP">
  <f name="GrRole" value="Oblique" />
  <f name="PGN" value="3s" />
  <f name="NPtype" value="QuantNP" />
</fs>
```

The tags used for feature sets `<fs>`, features `<f>` and references `<ref>` are all taken from the TEI-P5 tag set. The example in (146) contains two feature sets: one for the coreference information, and one for noun phrase information. The coreference information notes the type of link to the antecedent in the feature called `RefType`, and it contains a distance measure to the antecedent in `IPdist`. The `<ref>` tag gives us value of the antecedent's `Id` field.

I would like to finish this section on the choice of the format for the data (the texts) we are working with by summing up several advantages of using the *xml* implementation of the TEI-P5 tagset I have chosen to encode the coreferentially enriched and syntactically parsed corpora of English:

(147) *Advantages of using xml for storing parsed corpora*

- a. Texts encoded in *xml* can be effectively searched with *existing standard* tools like Xpath and Xquery, which have not been developed specifically for the purpose of searching *xml* encoded corpora, but come as a kind of bonus.
- b. The TEI implementation we use (the P5 one) allows constituents to be equipped with reference information as well as an expandable set of features.

The importance of the advantage mentioned in (147a) cannot be stressed enough: once we are able to use standard “off-the-shelf” tools and apply them to corpus research, we save ourselves a lot of work. Not we but others are developing the much-needed research tools, and possibly even improving on the query language we can use. To name but a few windows of opportunity that open up for free once we have our texts available in *xml*: (a) we can visualize our texts using *xslt* (a method of transforming *xml* into *html*), (b) we can manually tweak our texts using a wealth of freely available *xml* editors, (c) we can perform simple searches using the *xpath* standard, (d) we can define complex queries using the *xquery* standard, and (e) we can make use of programming tools available for *xml* structured datasets.

Using a standard like the TEI-P5, mentioned in (147b), has the advantage that it already defines a hierarchical mapping of the structure of sentences with their constituents and phrases into an *xml* tagset, and it has a definition of an expandable set of features (as visualized above in 146).

6.3 The coreference resolution algorithm

The Cesax coreference resolution algorithm proposed in this section builds, as explained in the introduction, on existing algorithms, although its overall approach is totally new. It builds on COT in the sense that it, like COT, makes use of a set of hierarchically ordered constraints to evaluate possible solutions. But it diverts from COT, since the constraints it uses are different, it is not restricted to the 1-clause frame associated with centering, and since it does *not* consider the most optimal solution for *all* noun phrases in a clause. It is similar to the Hobb's algorithm in that it treats every noun phrase individually. However, as we will see, the way it “steps” through the noun phrases diverges principally from that algorithm.

This section only presents the *Cesax algorithm*, without focusing on its actual implementation, although an implementation is available and can be freely installed.⁴ The *Cesax algorithm*, then, proceeds in stages, as shown in Figure 14.

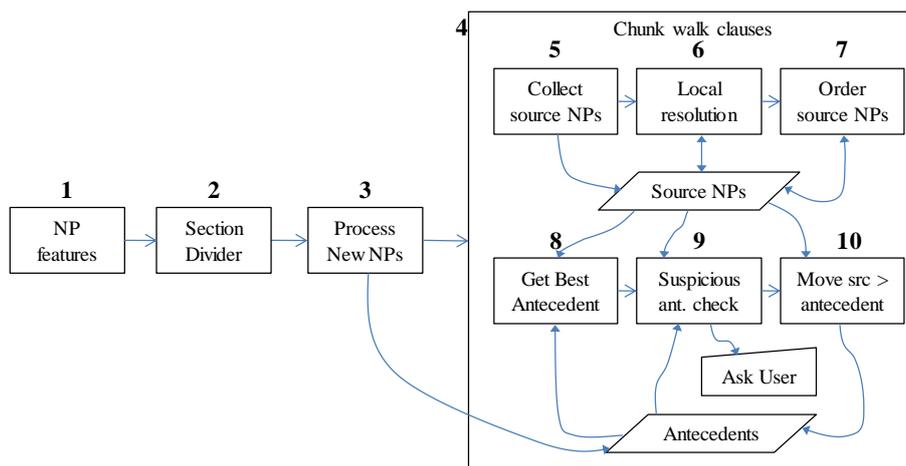


Figure 14 The semi-automatic coreference resolution algorithm

The first two stages are pre-processing. Stage 1 derives NP features like grammatical role and NP type. Stage 2 tries to divide a text up into smaller sections, so that processing is faster, and there is no risk of crossing section lines with the coreference resolution process. Stage 3 is an important step that tries to identify discourse new noun phrases in one complete sentence. The main processing occurs in stage 4, where the clauses of a sentence are parsed in a particular order. Stage 5 collects the noun phrases in the clause that need coreference resolution. Stage 6 tries to resolve any local coreferencing, such as those involving reflexive pronouns. The remaining noun phrases are ordered in stage 7, depending on their grammatical role and NP type. Each noun phrase is now taken in preferential order from the resulting collection of source NPs, and stage 8 looks for the best antecedent (if any) from the collection of potential antecedents available so far. Stage 9 checks whether the source-antecedent match found in stage seven is suspicious. If so, the user is asked to confirm or improve the resolution. The final stage 10 takes the resolved source noun phrase from the collection of sources and moves it into the potential antecedent collection, so that subsequent noun phrases can refer back to it.

Each stage in the algorithm deserves more detailed attention, which is why we will look at these stages in subsequent sections.

6.3.1 Gathering NP features

The core of the coreference resolution algorithm in stage seven hinges on the availability of some basic information for each NP. Like other pronoun and coreference resolution algorithms, the *Cesax algorithm* gathers this information

before the actual resolution begins. In particular, Cesax gathers the following information in stage one:

- NP type.
- Grammatical role.
- Person, gender and number.

Noun phrase types include for example: pronoun, definite NP, demonstrative, proper noun etc. The NP types can be determined by looking at the labels of the noun phrase's children. A proper name, for instance, will be a constituent labelled "NP", having one or more children labelled "NPR", which is the tag used to denote proper names.

Typical grammatical roles are subject, argument and object of a prepositional phrase. One additional role used by Cesax is the "possessive determiner". This is a noun phrase whose first child is, for instance, a genitive case of a proper name, such as in (148).

(148) (NP (NPR\$ Stephen's) (NS books))

Getting correct and detailed person/gender/number information for each noun phrase is a major factor that promotes correct coreference resolution.⁵ Number information can be gleaned from the labels of the NP's children: if the head noun is marked as N, we have a singular noun phrase, and if it is marked as NS, we have a plural one. The grammatical gender can, unfortunately, only be determined for pronouns (and for Old English: for demonstratives). The pronominal information in itself provides us with only partial gender resolution, since some pronouns are used for more than one gender. The plural pronouns, as well as the first and second person pronouns, don't distinguish gender at all. The grammatical person (first, second or third) follows straightforwardly from the pronominal paradigm. Non-pronominal noun phrases are all 3rd person.

6.3.2 Divide the text into sections

There are two reasons for wanting to divide a (larger) text into smaller sections. The main motivation has to do with the coreference resolution process. Anaphoric references tend to find their references within a well-defined section of a text, and by dividing the text in appropriate sections, such as chapters, sermons, stories or letters, unwarranted referential links are automatically excluded. The second reason is more algorithm-internal. The speed of the algorithm partly depends on the size of the antecedent's collection, and since Cesax allows antecedent candidates from a user-definable amount of preceding sentences, that collection grows sentence by sentence.⁶ Some noun phrases are deleted when they are processed, but only when Cesax is sure that they cannot be referred to again by subsequent constituents.⁷

The section division module uses two different clues to determine where section breaks occur. Both clues are part of extra textual information provided in nodes with the label "CODE". Old English texts are broken up into sections depending on the

Toronto text numbers, while Middle English and Modern English texts are broken into sections whenever a <heading> marker is found.

6.3.3 Identify discourse new noun phrases in the current section

Having divided the text into sections, the algorithm walks through each section sentence by sentence. The first pass through a sentence looks at every noun phrase and determines whether it is discourse new. If so, the noun phrase is put into the antecedent's collection straight away. Collecting potential antecedents for a whole sentence provides the algorithm with some alternatives to deal with sentence-internal cataphoric coreference situations. By the time the individual noun phrases within the sentence are processed in step 8, this current step 3 has already made some antecedents available, and within the domain of the sentence they can be cataphoric.

The big question for this step is how we can determine whether a noun phrase is discourse new. We certainly cannot get a full 100% of the discourse new noun phrases without the user's input, but we can check for several situations, as do some other algorithms too (Vieira, 1999). We can, for the moment, distinguish the situations defined in (149).

(149) *Discourse-new noun phrases*

- i) Definite noun phrases with restrictive postmodification. We assume that the postmodification used in noun phrases like *the bicycle of my mother* and *the car my mother drives* indicate that the noun phrase is discourse-new.
- ii) Indefinite noun phrases. E.g.: *a nice umbrella*.
- iii) Quantificational phrases. E.g.: *several boys, all people etc.*

There are other situations where the noun phrase potentially is discourse-new too, but those need to be checked with the user (which happens in stage 9 of the algorithm). Such situations include: anchored noun phrases (e.g.: *my daughter*), free relatives headed by a demonstrative pronoun (e.g.: *those living overseas*), and *wh*-headed free relatives (e.g.: *what I do*).

6.3.4 Process the clauses of each sentence in chunk order

The Cesax algorithm works its way through the sentence one noun phrase at a time. It takes the noun phrase types and their grammatical roles into account explicitly. That is why the algorithm has to deal with the main and sub-clauses of a sentence in a particular order, such that clauses containing potential antecedents are dealt with before clauses containing noun phrases that refer to these antecedents. The order of treating sub and main clauses is not breadth-first, since we would then miss out on example (150b), which requires the sub-clause to be parsed first. The order also is not depth-first, since that would not allow us to find the correct antecedent in (150a,e,f), which is located in the main clause. Depth-first would tempt the algorithm to make links in (150c,d), where these should not be made.

- (150) a. When he_{i/j} got into the boat, Peter_i sat down.
 b. When Peter_i got into the boat, he_{i/j} sat down.
 c. He_i sat down, after Peter_{*i/j} had come.
 d. He_{*i/j} walked, when Peter_i had laughed, into the room.
 e. Peter_i sat down, after he_i had gotten into the boat.
 f. In came, though she_i wasn't feeling too well, Mary_i.

This is why Cesax opts for a two-stage procedure. The first stage, discussed in section 6.3.4, recognizes and processes all unambiguously discourse-new NPs of a sentence (implemented as a `<forest>` element). The second stage recursively walks the main and sub-clauses of the sentence using the “ChunkWalk” algorithm. This algorithm starts processing the clauses (labelled as IP) in the sentence in a breadth-first, depth-last order. Each potential coreference source (an `<eTree>` element labelled as NP or as PRO\$—a possessive pronoun) is added to the source collection when it is encountered. The items in the source collection are processed in preferential order (see 6.3.7) as soon as either (a) a new IP is encountered, or (b) the last element of the current IP has been added to the source collection.

6.3.5 Collect the source NPs

Step four in the algorithm takes all the noun phrases in the clause that is currently reviewed, and adds them to the collection of source constituents. This step does *not* make exceptions to any of the noun phrases in the clause: they do not necessarily have to be daughters of a main or subordinate clause; their NP type is unimportant; they are *all* collected.

6.3.6 Perform local coreference resolution

Step five in the Cesax algorithm visits all the noun phrases in the clause, and sees if coreference can be resolved in a local manner. Local resolution applies to the three distinct situations listed in (151).

- (151) *Local resolution situations*
- a. Reflexive pronouns.
 - b. Appositives.
 - c. Certain bare nouns that are inert to coreferencing.

Reflexive pronouns need to be linked to the “nearest” subject higher up in the hierarchy. Appositives are linked to the nearest preceding noun phrase. Both source noun phrases are to be deleted from the collection of sources after their coreference has been resolved, because none of them can serve as an antecedent for other noun phrases.

The third category warranting local resolution consists of certain types of bare nouns. The bare noun complement of a copula clause, e.g. “professor” in (152a), cannot be referred to and cannot itself refer to an antecedent. This is because this bare noun does not refer to a person, but functions as a class label. The bare noun “professor” denotes the class of individuals having the quality of being a professor.

The noun phrase “professor” can thus be regarded as “inert” to the coreference resolution process.

The same situation seems to hold for bare noun complements of prepositions, witness example (152b). The noun phrase “bed” does not refer to one particular bed, but opens a class of items characterized by being beds.

- (152) a. John is professor.
 b. We went to bed. [Boswell 1776]

Inert noun phrases receive the label “inert”, are taken out of the source collection, and don’t get moved into the antecedent collection.

6.3.7 Determine the order of treating source NPs

The remaining source collection contains noun phrases for which the coreference needs to be figured out. Some may contain totally new information, some assumed information, and some may actually refer back to an antecedent. Step six of the Cesax algorithm calculates a preferential order by which the source noun phrases need to be reviewed as to their coreferential status. The idea is that the “best” source candidate will refer back to the most “salient” antecedent. What is “best” for source candidates is determined by two linguistic hierarchies. The first hierarchy, illustrated in 157, looks at the noun phrase type. Pronouns are more likely to refer back than definite noun phrases and so forth. The second hierarchy, illustrated in 156, looks at grammatical roles. Subjects are more likely to refer back than objects and so on. The preferential order takes these two properties into account. The noun phrase type figures as the first factor, and if there are more constituents with the same noun phrase type, then the order is determined by the grammatical role.

6.3.8 Get the best antecedent for each source NP

All the work done in steps one to seven can be described as preliminary. It all leads up to step eight, which aims to find the best possible antecedent for the source noun phrase currently under scrutiny. The Cesax algorithm sides with the COT algorithm in using a hierarchical set of constraints to determine what constitutes the best antecedent from the set of potential antecedents. The constraints used in Cesax are listed in Table 17, in the hierarchical order currently used for them.

This section offers an explanation for each individual constraint, and then shows the basic way by which constraint evaluation can take place.

Table 17 Constraints used to determine the best antecedent

Constraint	Description
AgrGenderNumber	One violation when gender/number of source disagree with gender/number of antecedent.
Disjoint	One violation when source and antecedent are in the same IP
EqualHead	One violation when the source head noun does not agree with any of the head nouns in the chain of the target
NoCataphor	One violation for an antecedent that is following the source instead of preceding it.
NoClause	One violation for an antecedent that is a clause (IP).
AgrClause	One violation mark when a source does not have PGN 3s/3ns, yet does agree with an antecedent IP.
NoCrossAgrPerson	One violation when there is agreement in person at a cross speech boundary.
NearDem	One violation for an antecedent that already has a coreference, unless the antecedent NP also contains a near demonstrative.
AgrPerson	One violation when the source has a different person than the antecedent.
IPdist	One violation for every IP between source and antecedent
GrRoleDst	The number of the NP's grammatical role on this scale: Sbj > PossDet;Arg > PPObj > other
NPtypeDst	The number of the antecedent NP's type on this scale: Zero > Pro > Proper > DefNP;AnchoredNP > DemNP > Other
NoCrossEqSubject	One violation when source and antecedent are both subject and cross a speech boundary. Or: one violation when the source's IP is imperative, the source itself is an argument and the antecedent is a subject.

If the constraint evaluation results in two (or more) “best” antecedents which are evaluated equally well, in that they have the same violations, then the user needs to resolve this ambiguity. Likewise, if the antecedent suggested by the constraint ranking algorithm is relatively far away, then again the user needs to be consulted. And even if we are left with one best antecedent, some more checking is needed in step 8, described in section 6.3.9.

Since the constraints used in step eight form the heart of the algorithm, they require a detailed description. Instead of the constraint AGREE, which is hierarchically the topmost constraint used in the COT algorithm, Cesax has two separate ones: AGRGENDERNUMBER and AGRPERSON. This is because person agreement depends on whether a coreference relation crosses a speech boundary. If it does so, then the constraint NOCROSSAGRPERSON has to overrule AGRPERSON, since in many cases there should *not* be agreement in person across a speech boundary. For instance, second person pronoun “you” should point to third person common name “Peter” in example (153). Likewise, first person “me” points to third person “John”.

- (153) a. John_k asked Peter_m:
b. “Will you_m help me_k?”

The DISJOINT constraint derives straight from the COT algorithm (Beaver, 2004). Cesax uses the Treebank inherited syntactic structure to see whether the source noun

phrase and the potential antecedent are part of the same IP or the same NP. If that is so, then this combination gets one violation mark for the DISJOINT constraint.

The EQUALHEAD constraint is based upon the literature on definite noun phrase coreference resolution (Soon et al., 2001, Vieira, 1999). If a source noun phrase contains a head noun, and a potential antecedent also contains a head noun, then there is a violation if these head nouns don't match, unless the source's head noun matches one of the other head nouns in the coreference chain of the potential antecedent. An example where there *is* a match is given in (154). The source noun phrase "the market" can have "the market in the center of town" as antecedent without a violation of EQUALHEAD. Notice that this constraint is also conformed to vacuously, if either the source or the potential antecedent does *not* contain a head noun. The constraint EQUALHEAD is a special one within Cesax, since it facilitates a slight form of progressive learning. As more coreference resolutions are being made, more combinations of which head noun can, in principle, refer back to which other head noun become available. These combinations are taken into account whenever EQUALHEAD is evaluated.

- (154) a. John went to the market in the center of town.
 b. He bought shoes at the market.

The constraint NOCATAPHOR has not been taken from other algorithms. It is based on the observation that coreferencing favours anaphoric over cataphoric references. Since all constraints are violable, cataphoric coreference relations are still possible though.

Another constraint peculiar to Cesax is the NOCLAUSE constraint. Cesax allows coreference relations to be made from a noun phrase to an IP (a clause). However, this is less common than references from noun phrases to noun phrases, hence the NOCLAUSE constraint. When there is a potential clausal antecedent to a noun phrase, the constraint AGRCLAUSE adds one more restriction. The source noun phrase should at least be third person singular, and when gender is specified for the source noun phrase, it should be neuter. This discourages "he/she/they/you" from referring back to a clause, and favours "it/that".

Like other coreference resolution algorithms that are not based on centering, Cesax too has a constraint that progressively disfavours distant antecedents (Bouma, 2003, Soon et al., 2001, Vieira, 1999). The constraint IPDIST measures the amount of clause boundaries between the source noun phrase and the potential antecedent.

The NEARDEM constraint tries to capture the intricate, and seemingly language dependent, behaviour of the near demonstratives like *this*, *these*. A language like Dutch use them to facilitate topic switch, a language like Chechen uses them to point to a major participant on the discourse level. Present Day English seems to use the near demonstrative to refer back to a secondary participant that now comes more into the picture, but only when it is plural, as illustrated by (155). The main topical participants are referred to by "they" in all three sentences. But the "few Indian houses" (and by implication: the people living in them) become a minor participant that needs to be referred to in (155c). This is done by the near demonstrative "these", signalling a shift in topic. This topic shift is confirmed by (155c-d), where

the inhabitants of the Indian village are referred to by non-subject pronouns “their” and “them”.

- (155) a. When they_i came to the few Indian houses_k which they_i thought had been the town, they_i were under a great disappointment.
 b. They_i consulted what to do, and were some time before they_i could resolve.
 c. For if they_i fell upon these_k, they_i must cut all their_k throats.
 d. It was ten to one but some of them_k might escape. [Defoe 212.324-330]

The Cesax algorithm does not make use of COT constraints PROTOP, FAMDEF and COHERE. These constraints determine for instance that a pronoun in one sentence is likely to refer back to the topic of the previous sentence. Instead of working with the notion “topic”, Cesax takes into account two linguistic hierarchies. The constraint GRROLEDST recognizes that it is more likely for an antecedent to refer back to a subject than to an object. The constraint assigns a number of violation marks depending on the position of the scale in (156), which is partly based on observed preferences, and partly on the accessibility hierarchy (Keenan and Comrie, 1977). Since the most salient source noun phrases are processed first by the Cesax algorithms, the potential antecedents also need to be lined up.

- (156) *Grammatical role scale*
 Subject > Argument; Possessive > PP-object > other

The constraint NPTYPEDST is partly based on observed preferences and partly based on the NP types for English, as these correlate with the givenness hierarchy (Gundel et al., 1993). The givenness hierarchy as such consists of cognitive states such as “In Focus” and “Activated” (see 5.2.3), but for any specific language it roughly translates into a hierarchy of NP types. The saliency of a potential antecedent decreases depending on the kind of noun phrase, as shown in (157).

- (157) *Givenness hierarchy translated into NP types for English*
 Zero > Pronoun > Proper name > Definite NP; Anchored NP >
 Demonstrative NP

Both the GRROLEDST as well as the NPTYPEDST constraints are ones that probably require language specific fine-tuning.

One final constraint is NOCROSSEQSUBJECT, which looks at situations where the coreference relation crosses a direct speech boundary—where the source is direct speech and the antecedent narrative, or vice versa. The possible coreference relations are quite restricted in a cross speech situation—in particular with a subject as antecedent. As illustrated by (158a), for instance, it is less likely that the subject of the direct speech “you” would coincide with that of the indirect speech “John”. Likewise, as illustrated by (158b), there is a tendency to leave the subject of the containing a speech introducer (here: “John”) implicit in the direct speech fragment. Both constraints are soft ones, and they are currently only decisive if all other constraints of the Cesax algorithm have failed to come up with a good candidate.

- (158) a. John_k told Peter_m: “You_m have come to know me_k as a colleague”.
 b. John_k told Peter_m: “You_m must come”.

Although it is arguable whether the way in which constraints are evaluated is part of an algorithm or not, it may be instructive to show the principle of constraint evaluation that has been chosen in our implementation of the Cesax algorithm. The algorithm in (159) shows the steps that are taken in calculating what the best antecedent is, given one particular noun phrase for which we are trying to resolve the coreference.

(159) *An algorithm to calculate the antecedent of one noun phrase*

- Step 1** For each *constraint* in the set of constraints, ordered by their evaluation level...
- Step 2** For each *candidate* in the collection of potential antecedents: calculate the weight of this constraint and add it to the total weight for this candidate.
- Step 3** If only 1 candidate is left with the minimum weight, then exit and return this candidate. Otherwise go to the next constraint.

The constraints are evaluated one-by-one, starting with the top-level one. All candidates in the current collection of potential antecedents are evaluated for this particular constraint, and the total evaluation number for each candidate is adapted accordingly. The next step in the algorithm checks how many candidates are left with the minimum total evaluation number. If there is only one left, then this is the best candidate we can come up with, and which we then have to check against suspicious situations, as explained in the next section. If more than one candidate with minimal weight is left, we go and evaluate the candidates against the next constraint. In this step it would not even be necessary to actually calculate evaluations for all candidates—only those with the minimum evaluation number would need the evaluation of an additional constraint. If the evaluation algorithm has gone through all constraints for all possible antecedents, and found that there is still more than one candidate with the smallest evaluation number, then no further check for suspicious situations is needed—the user can immediately be consulted for his input in resolving the conflict.

6.3.9 Check for suspicious coreference solutions

Suppose step seven in the algorithm has come up with one best antecedent for the source noun phrase currently being evaluated. Step eight of the overall algorithm shown in Figure 14 checks this source-antecedent pair against a set of known “suspicious” situations, which are found and defined manually. If the solution belongs to one of these suspicious situations, then the user needs to confirm or modify the solution found by the algorithm. Table 18 presents the most important suspicious situations currently used by the Cesax algorithm. Some of the situations partially overlap with the constraints from Table 17.

The situation GENDERNUMBERDISAGREEMENT coincide with the AGRGENDERNUMBER constraint. The CROSSSPEECH situation is triggered when

either the NOCROSSAGRPERSON or NOCROSSEQSUBJECT has a violation. The EQUALHEAD situation completely matches the constraint with the same name.

Table 18 Suspicious situations

Situation	Description
GenderNumberDisagreement	When gender/number of source disagrees with the gender/number of the antecedent that was found.
CrossSpeech	The suggested coreference link crosses a speech boundary, and (a) there is agreement in person, or (b) the link goes from subject to subject.
EqualHead	The source head noun does not agree with any of the head nouns in the chain of the target, and some additional conditions are met. E.g: source and destination are proper nouns.
AgrGender	When the source has a specific gender, then the antecedent should agree with it.
Ambiguity	One violation for every full NP in the IP of the target
NTypeSrc	Certain sources are unlikely to function as source.
Disjoint	Source and antecedent are in the same syntactic domain.
CloseVicinity	When more than one candidate is in close vicinity, while these candidates differ only marginally in evaluation.

The AMBIGUITY situation counts the total number of full noun phrases (as opposed to pronouns and demonstratives) in the clause inside which the potential antecedent resides. The idea is based on the observation that it is hard for an algorithm to determine to which full NP a source noun phrase refers if there are two or more full noun phrases in the same clause containing the potential antecedent. Ambiguity is illustrated in (160) and (161).

- (160) a. [_{NP} The parents]_k brought [_{NP} their children]_m to [_{NP} the station]_n.
 b. As the train left, they_k waved them_m goodbye.
- (161) a. John took [_{NP} a book]_k from [_{NP} the shelf]_m.
 b. Mary looked at it_k.

The resolution of (160b) would need more context in order to know whether the subject “they” of (160b) refers to the parents or the children. The antecedent of “it” in (161b) should be “a book” from (161a), but this is not something we could expect a coreference resolution algorithm to be able to know.

The AGRGENDER situation does not derive from a constraint used in step seven. It requires the user to give his judgment whenever the best antecedent that has been found for the source noun phrase under consideration is less specific in gender. For instance, when a source noun phrase is masculine, as for example the pronoun “he” in (162b), and the best antecedent’s gender is not given, as for example “the king” in (162a), then the user’s input is needed.

- (162) a. The king stood before the army.
 b. He looked intently at his men.

The NPTYPESRC situation makes sure that coreference relations having for instance an indefinite noun phrase as source are double checked with a user. It is very

unlikely, for instance, that expressions like “many people”, “two children”, “words of wisdom” have an antecedent.

The DISJOINT situation matches the constraint with the same name used in step seven. Other suspicious situations may come up, as Cesax processes more and more texts. When they are identified, they can easily be added to the list of suspicious situations. The algorithm does not give a priority to such situations—as soon as at least one suspicious situation is met, it halts and asks for the user’s input.

6.3.10 Move the NP from the source to the antecedent collection

The last step of the Cesax algorithm is to move the source noun phrase from the source collection to the antecedent collection. Whenever such a move is made, some additional checking is done to see whether the *antecedent* should actually be kept in the antecedents’ collection. If the source noun phrase refers back to a pronominal antecedent, then this antecedent may be safely removed from the antecedents’ collection. Any further references should not be allowed to link back to this pronominal antecedent; they should target the source noun phrase that is now moved into the antecedent’s collection instead.

Note that the above only holds for pronominal antecedents. As soon as the antecedent is a noun phrase with a more substantial head (e.g. a nominal head), it could potentially be referred to by further noun phrases.

6.4 Case study: a history book from 1866

This section describes a case study in which the semi-automatic coreference resolution algorithm is applied to a single text, in order to test its effectiveness. The text for this case study is the text named “long-1866” taken from the PPCMBE corpus (Kroch et al., 2010). The text consists of three chapters from a history book entitled “The decline of the Roman empire” written by George Long. It contains 3083 noun phrases, and these have all been annotated for coreferentiality using Cesax.

About 54% of the noun phrases were processed automatically by the algorithm, while the user was consulted for the remaining 46% of the cases. The user agreed with about 40% of the suggestions made, choosing other options for the remaining 60% of the situations where consultation was deemed necessary by Cesax. About 5% of the automatically processed coreference resolutions were found to be erroneous, so that the total success rate of the algorithm (the number of correctly automatically resolved coreference situations and the number of correctly made suggestions) totals to 70%.⁸

The evaluation of the coreference resolution algorithm given here lacks some of the metrics that have been proposed and evaluated recently in the field of automatic coreference resolution (see Recasens et al., 2010 for an overview of the different metrics). The task of semi-automatic coreference resolution for the purpose of linguistic research differs in key aspects from the NLP task of automatic coreference resolution, so that a comparison between the two systems may not be very helpful.

Nevertheless, future work on Cesax should seek to provide these metrics, and investigate which kinds of comparisons between the systems are helpful and needed.

Table 19 shows which types of coreference relations were established in the Long text, which largely coincide with the bare minimum Pentaset introduced in section 5.3.⁹ The majority of cases (close to 50%) were IDENTITY relations, and more than a third (40%) were completely discourse NEW. The reference type ASSUMED deals with noun phrases that point to knowledge shared between the speaker and the hearer. The NEWVAR reference type is peculiar to the Treebank format used—it refers to variables introduced for instance by *wh* clauses.

Table 19 Reference types used in the case study

Reference type	Frequency	Count
Assumed	3,1%	97
CrossSpeech	1,7%	53
Identity	37,6%	1158
Inert	8,6%	265
Inferred	1,8%	55
New	41,5%	1280
NewVar	5,7%	175

One more piece of information that can be gleaned from the case study concerns the kind of constraints that proved to be crucial in deciding which antecedent formed the best fit for the source noun phrases. A constraint is “crucial” when after its application only one antecedent candidate remains. Table 20 shows those cases in which the various constraints, in order of increasing hierarchical level, given in the column marked “Level”, proved to be crucial. The top four crucial constraints are: the number of clauses there are between the source and the antecedent (IPdist), the grammatical role of the antecedent (whether it is subject, object, P-complement and so forth), whether the noun phrase heads—if present—match up (EqualHead), and the NP type (e.g. pronoun, definite NP etc) of the antecedent (NPtypeDst).

Table 20 Crucial constraints in the “Long” text

Constraint	Frequency	Count	Level
(none)	7,6%	235	-
NPtypeDst	1,8%	55	15
GrRoleDst	16,1%	496	14
IPdist	57,8%	1783	13
AgrPerson	0,5%	15	11
NearDem	0,1%	4	10
NoCrossAgrPerson	0,2%	6	9
NoClause	0,5%	16	5
NoCataphor	0,1%	3	4
EqualHead	12,1%	374	3
AgrGenderNumber	0,4%	11	1

One final piece of information concerns the question how far the established anaphoric links go back. It appears, not surprisingly, that most of the anaphoric links are either in the same clause or in the immediately preceding clause. A diminishing number of links appears in subsequently preceding clauses, up to a distance of more than 200. What this shows is that it does make sense to let the algorithm supply potential antecedents that come from further back—otherwise the user would have to resort to completely manual annotation.

6.5 Discussion

The challenge taken up in this chapter has been to find a way to add the referential state primitives derived in chapter 5 to the existing syntactically parsed corpora. Section 6.1 describes the strategy we are taking, which is a semi-automatic approach: make an algorithm that looks for the antecedent of each noun phrase and determines its referential state. As much as possible is resolved automatically, but suspicious situations are recognized, and the user is asked to select the correct antecedent in such situations, or to label the constituent with one of the “unlinked” states (the states “New” or “Inert”). The algorithm takes as starting point treebank texts from the parsed English corpora (6.2); these have already been annotated syntactically. The original labelled bracketing is first transformed into an *xml* format that conforms to the TEI-P5 standard. This standard allows features—such as the NP type, and coreference information—to be added to constituent nodes using the `<fs>` tags. The approach of the Cesax algorithm (6.3) builds on existing ones. Cesax handles the coreference resolution process constituent by constituent like the more traditional Hobbs algorithm, and it uses a hierarchical evaluation of constraints to arrive at the most plausible antecedent, like the newer COT algorithm does. The biggest novelty of the Cesax algorithm is the semi-automatic approach, which boils down to recognizing suspicious coreference solutions that need to be verified with the user. Unlike the COT algorithm, Cesax is not based on centering. The constraints used by Cesax take into account hierarchies such as the noun phrase type (157) and grammatical role scale (156) in order to determine the most likely antecedents for a

given noun phrase. A case study of the algorithm as applied to one text (6.4) shows that 54% of the coreference situations were resolved automatically, while the user had to be consulted in the remainder of the cases. The user agreed with approximately 40% of the suggestions made by the algorithm in the situations that were not resolved automatically, and 95% of the 54% automatically made coreference links were correct, bringing the overall success rate of the algorithm to about 70%. The figure of 70% probably comes across as rather low from the point of view of computational linguists, but where fully automatic algorithms end up without knowing where the “mistakes” are, the current algorithm reaches the accuracy that is required for the kind of linguistic research described in this book.¹⁰ It should also be taken into account that computational linguistic methods often do *not* provide a full coreference resolution of *all* the noun phrases available in a text, since they only focus on the noun phrases that are linked in the text. This leaves out making a distinction between three important referential states: “Assumed”, “New” and “Inert”.

Future work can focus on fine-tuning of the constraints depending on the text period, critical evaluation and possibly extension of the constraints, and fine-tuning of the suspicious situations. The more we are able to differentiate *really* suspicious situations from correct coreference resolutions, the better the accuracy of the result will be. Future work should also provide more heuristics for the performance of the semi-automatic algorithm: the standard metrics used for automatic coreference resolution (see section 6.4) should be calculated for the automatically resolved part of Cesax, and the interrater agreement should be calculated for the machine-guided manual resolution part of the algorithm.

The current tool that implements the Cesax algorithm is a stand-alone computer program called “Cesax” (Komen, 2011b). Future work should at least provide facilities for collaboration, such as through the realization of an internet repository of texts and through a related system of double-checking enriched texts. Future work might also involve providing a web-based service for the coreference resolution similar to those provided by tools such as “brat” (Stenetorp et al., 2012), “BART” (Versley et al., 2008) and “MMAX2” (Müller and Strube, 2006); both web-based as well as stand-alone approaches have their advantages, and should therefore be provided for the users.

With a method to enrich the existing syntactically parsed corpora, the next challenge is to find ways to query the enriched texts for a combination of syntactic and referential information. Chapter 7 focuses on that, and when we are done there, we are ready for the actual corpus research described in chapters 8 and 9.

¹ The term “Cesax” originally is the name of the xml-version of Cesac (coreference editor for syntactically annotated corpora), but has subsequently come to denote the coreference algorithm it uses.

² Beaver’s article mainly shows simple clauses, but his examples (24) and (25) contain some slightly more complex ones.

³ COT tries to find the best coreference resolution for all noun phrases in one clause at the same time. It does this by evaluating all possible connections with the preceding clause against a set of hierarchically ordered constraints. This evaluation process resembles OT.

⁴ An implementation (in the .Net version of visual basic) is available on <http://erwinkomen.ruhosting.nl/software/Cesax>.

⁵ A coreference resolution algorithm would, at this point, ideally make use of a dictionary look-up to find the information needed. Our implementation of the algorithm, however, cannot do this, since no suitably annotated dictionaries of the different stages of English exist.

⁶ The amount of time the algorithm takes after it has resolved the coreference of one noun phrase and before it needs the user’s input on the coreference of the next noun phrase is, obviously, data dependent, but still below one second in the texts we have been working with.

⁷ Noun phrases that are deleted from the collection of potential antecedents once they have functioned as antecedent, are, for instance, pronouns. But full noun phrases are never taken out of the collection.

⁸ The 70% consists of 54% automatically resolved coreference times 95% success rate, to which are added the 40% correctly made suggestions of the 46% part of the text that could not be handled automatically: $0,95 * 0,54 + 0,40 * (1 - 0,54) = 0,70$.

⁹ The category “assumed” refers to information that is assumed to be shared knowledge between the author and the reader, “identity” relations point to the same referent, as does “cross speech”, but then across a direct speech boundary. An “inert” noun phrase does not refer to something, and cannot be referred to. An “inferred” relation is a bridging expression such as part-of-whole. A “new” label indicates a totally discourse-new constituent, while a “newvar” points to the introduction of a new variable, e.g. in a *wh*-clause.

¹⁰ A 100% success rate is probably never possible in any algorithm, but Cesax approaches this by allowing the user to review all the automatically made links, which are the ones where errors could creep in.

The major research question as formulated in (11) is how the interaction between syntax and focus changed in English over time, and we are almost ready to actually tackle this question. The strategy we have formulated is that we (i) enrich existing corpora with coreference information, assigning referential states to each and every noun phrase (see chapters 5-6), and then (ii) query the resulting texts by taking note of syntactic and referential information in order to distinguish focus domains (the whole core, the predicate, or just one constituent), which equal the three focus articulations defined in chapter 3:thetic articulation, topic-comment articulation and constituent focus.

The chapter at hand focuses on the question of *how* we are to query the syntactically parsed texts that have been enriched with the referential information. The search “engine” we are going to need to be able to actually *quantify* the changes in the expression of English focus and its relationship with syntax have to fulfil a few requirements, which are formulated in (163).

- (163) *Requirements on a search engine usable in quantifying focus change*
- a. Detect syntactic environments that could signal a particular focus articulation
 - b. Detect the referential state of constituents
 - c. Detect (a) and (b) for antecedents of constituents

The necessity of the requirements stated in (163) will only become fully clear in chapters 8-9, but since the choice of the search engine, discussed in this chapter, hinges on them, I will briefly review these requirements with the help of a few sentences from “The three musketeers” in (164). The requirement in (163a) states that a search engine that is going to be of any use for our purposes needs to be able to detect “syntactic environments”. An example of a syntactic environment that links to presentational focus, for instance, is the constituent order of Locative-FiniteVerb-Subject, such as “At the door stood two horses” in (164d). The horses mentioned in this sentence are completely new participants in the story, and line (164d) serves to introduce them as new topic, which are taken up in the narrative’s usual topic-comment articulation in line (164e).

- (164) a. This began to be annoying.
b. All these successive accidents were perhaps the result of chance; but they might be the fruits of a plot.
c. Athos and d'Artagnan went out, while Planchet was sent to inquire if there were not three horses for sale in the neighborhood.
d. At the door stood two horses, fresh, strong, and fully equipped.
e. These would just have suited them. (Dumas, 1878: 154)

The example illustrates that if we want to look for presentational focus, we must not only be able to detect a particular syntactic environment (the constituent order of PP-V-S), but we must also be able to evaluate the referential state of a constituent: if the subject is referentially new, then we are bound to have presentational focus. This latter requirement illustrates why (163b) is a necessary condition for a search engine.

The second example I would briefly like to focus on is (164b). We know that there is constituent focus on “the fruits of a plot”, since this constituent provides an answer to the question “What is the source of these accidents?”, and also because the noun phrase “the fruits of a plot” contrasts with the noun phrase “the result of chance”. If we wanted to define a search that is able to detect the kind of constituent focus present in (164b), then it would need to detect that (a) the sentence consists of two equative clauses of the form Subject-Be-Complement, (b) the subject “they” of the second clause refers to the same entity as the subject “all these successive accidents” of the first clause, and (c) the complement in the second clause is referentially new. This illustrates requirement (163c), since we need to know the syntactic status of the antecedent of “they” in order to fully detect sentences with constituent order of the type illustrated by (164b).

These examples are sufficient to illustrate that the three requirements stated in (163) must be met by a corpus engine that is able to serve us in our quest for focus changes in English. This chapter reviews existing query languages and engines against the background of the three requirements, showing that none of the existing ones are able to meet the requirements completely. The chapter then presents a solution in the form of the program “CorpusStudio” that uses the query language “Xquery” with built-in extensions.

Readers who are not at all interested in the technical details of query languages could skip this chapter, because the searches discussed in chapters 8-9 will be described as much as possible in plain language, so that no in-depth knowledge of Xquery is required to follow the discussion. Readers who are familiar with query languages may find it sufficient to read about the additional functionality provided by CorpusStudio in section 7.3, and are advised to read through the corpus research example in section 7.4.

7.1 Choosing a text format and a query language

Corpus research in general can nowadays often be done using web-based tools, such as for instance the search interface provided by Mark Davies for the British National Corpus (Davies, 2004-2012). Tools like this facilitate formulating a query in a simple language, but they are limited to searches on the level of individual words and part-of-speech labels.

There are special programs allowing off-line syntactic searches in historical English texts, but these programs are often command-line oriented, and researchers often directly invoke queries using the Window command prompt. There are several major problems associated with this approach: it is error-prone, since the user can easily overlook one step in a series of queries that need to be done in a particular

order, the approach leads to unreplicable results, since the researcher may forget the exact series of queries that he used or the particular input files he used, and the results of the approach are irretrievable, since there is no record of the input files, queries and query order used to obtain a particular set of results.¹

One off-line program that is able to deal with the parsed English corpora is “CorpusSearch2” (Randall et al., 2005), a program that allows querying the *treebank* format (sometimes referred to as “labelled bracketing”) in which the texts are provided. The texts that are enriched with the “Cesax” program (a stand-alone program implementing the algorithm described in chapter 6) can only be queried with CorpusSearch2 if they are exported from Cesax into a treebank format. However, CorpusSearch2 is not able to query one important part of the information that is present in the enriched texts: it cannot access *antecedents* of constituents (the noun phrase in the preceding context that has the same referent as the constituent under question has), as discussed in section 6.2. I have shown above in example (164) that some of the queries trying to detect focus articulations make use of these antecedents, which is why the query language we choose needs to be able to access that bit of information. The problems with the treebank format have already been touched upon in section 6.2, where I concluded that it is better to use the *xml* format instead, but in the context of this chapter I would like to give one more example (165) that illustrates the problems that the labelled bracketing format combined with the CorpusSearch2 engine experiences in accessing antecedents.

(165) *Treebank data with coreferential information*

```
(IP-MAT
  (NP-SBJ-ID:101 (PROS-ID:102-REF:80 my) (N Partner))
  (VBD went)
  (PP (P to) (NP-ID:103 (D the) (N Monastery)))
  (. .))
(IP-MAT
  (NP-SBJ-ID:104 (PRO I))
  (VBD met)
  (NP-OBJ-ID:105-CREF:IDT-REF:101 (PRO him))
  (ADV there)
  (. .))
```

The direct object *him* in the second sentence has an antecedent *my partner* in the previous sentence with an ID of number ‘101’. If we have a query that is looking at the situation in the second sentence, and this query needs to have access to the antecedent of the direct object (it may, for example, want to know whether the antecedent refers back to something itself), then the query language CorpusSearch2 would have to facilitate two things: (a) have a command that allows stripping the antecedent’s ID from the source node’s label NP-OBJ-ID:105-CREF:IDT-REF:101, and (b) access the antecedent that has the number ‘101’ (even though the query is already processing a line in which it is not available). The first task is not possible with the current version of CorpusSearch2, since this version does not offer a command to obtain a particular part of a label; it is only able to look for the *presence* of (part of) a label through the function HASLABEL. The second task is not possible

in principle: once CorpusSearch2 has processed one sentence, it continues with the next one and has no access to the previous sentence anymore.

The problems appearing in the example with the treebank data are related to two different matters: the encoding of the data (treebank format) and the search engine (which is sentence oriented). As I have demonstrated in section 6.2, and as appears from the treebank example in (165), the data can much better be encoded in the *xml* format, since that format is not only able to keep the hierarchical structure of a sentence's syntax (which is what the treebank format can do too), but it is more suitable to contain node identification and other feature information at the different levels in the hierarchy, and it can easily contain cross-references between the constituents it encodes. The second problem, that a search engine is oriented to processing a text sentence-by-sentence, is a separate one: it is not related to the way in which a text is encoded. While I am not aware of query engines considering the data of a treebank on the level of the text as a whole, this is not impossible in principle. And as for *xml* oriented query engines: some of these process data in chunks (just as the CorpusSearch2 engine for labelled bracketing treebank data), and some query engines process the data on the level of a whole text.

There are several query languages around that have been designed to query texts annotated in *xml*, such as: TigerSearch (Brants et al., 2002), Tgrep2 (Rohde, 2005), Annis (Zeldes et al., 2009) and DtSearch (Kloosterman, 2007), to name but a few.² Almost none of the existing query *languages* are able to access constituents' antecedents just like that, since coreferentially enriched syntactically parsed texts are a recent development. One exception is the Xquery implementation used to search through the Alpino *xml* treebank (Bouma and Kloosterman, 2007). This implementation has a user-defined function `resolve-index`, repeated in (166), which is able to access the antecedent of a node.

(166) *Alpino Xquery function to access an antecedent* (Bouma and Kloosterman, 2007)

```

1  declare function alpino:resolve-index($constituent as element(node))
2      as element(node)
3  { if ( $constituent[@index and not(@pos or @cat)] )
4    then $constituent/ancestor::alpino_ds/
5          descendant::node
6          [@index = $constituent/@index and (@pos or @cat)]
7    else $constituent
8  }
```

What the function in (166) does is check if the node `$constituent` contains an attribute `@index`, which, in the Alpino *xml* implementation, is a pointer to an antecedent, and then retrieve this antecedent node, but this node has to be part of the tag `<alpino_ds>`, which is the Alpino equivalent of a sentence. This is how standard Xquery can be used to access antecedents within one sentence in Alpino-*xml* texts.

What is needed to search the referentially enriched corpora goes one step further: we need to be able to retrieve antecedents at the level of one whole text. We do want to make use of the Xquery language (Boag et al., 2010), since it offers us several important advantages: the language allows searches on *xml* coded information that is hierarchically oriented (which is the case for the Alpino *xml* treebank and also for

the *psdx* output of Cesax), it is an open standard (which means that other people are continuously improving it), it allows extension through user-definable functions, and it provides access to the constituents' antecedents *in principle*. I am emphasizing "in principle" here, because there is one problem that needs to be tackled. The Alpino method of retrieving an antecedent as shown in (166) only gets antecedents within one sentence.³ Additional measures need to be taken to access antecedents within a text as a whole. What follows is an account of the necessary extensions to Xquery, such that it is able to be used for our focus-oriented questions.

7.2 Accessing constituents' antecedents

So far we have stated that we will query texts using the Xquery language in order to find the focus types we are looking for. Some of the situations that are indications of constituent or presentational focus types require us to find antecedents of constituents across the level of one sentence. The Xquery language does not prohibit this in principle, so that one option would be that we process one whole text at-a-time and then access the antecedents through the means that are built into Xquery: through the axes.⁴ This requires us to load a whole text into memory, and process the query functions we have on that text. If we do this, then we will be able to access the antecedents of constituents that are anywhere in this text. There is a practical limitation in that the Saxon implementation of Xquery we use has memory limitations, which is why this option is not a workable solution (Saxon, 2009).

Another option is to process a text sentence-by-sentence, and then access the antecedents through built-in extension functions. This seems to be a good alternative, but we need to be able to make an extension function that is capable of: (a) accessing text-level antecedents, while (b) the text is still being processed in a sentence-by-sentence order. The method described in (167) provides a solution to this problem.

(167) *Accessing a text-level antecedent from within sentence-by-sentence processing*

- a. Load the *xml* text into the wrapper as an "xml document".
- b. Process each sentence (which is implemented as a `<forest>` in a *psdx* text) in this document using Xquery.
- c. Add a user function in the wrapper that retrieves the antecedent's *xml* code from the "xml document" that has been loaded.

The solution starts by loading the entire *xml* document into the wrapper (167a). This is necessary anyway for sentence-by-sentence processing, because the *xml* format of the texts that have been enriched in Cesax, the *psdx* format, contains one text as a whole. The *psdx* texts are internally divided in `<forest>` parts (one `<forest>` roughly corresponds to one sentence). Step (167b) is where the Saxon implementation of Xquery processes one `<forest>` node a time (this alleviates potential memory problems). The step described in (167c) is the crucial one in the solution to retrieve antecedents: we make use of a user function, which is written in the programming language of the wrapper, and which has access to the whole loaded "xml document".⁵

We see, then, that the wrapper program we build and use around the Xquery implementation is of vital importance to the focus research we are doing. It is the wrapper that should provide additional functions, comparable to the `alpino:resolve-index()` function described in (166), but now operating at text-level.

7.3 CorpusStudio: a wrapper around Xquery

We have seen that the desire to be able to access antecedents influences the choice of the query language and the specifications of the wrapper program. But there are several more reasons why having a wrapper program around an Xquery engine is advantageous.

- (168) *Advantages of having a “wrapper” around an Xquery engine*
- a. It is a windows-oriented wrapper for researchers who are not used to work with command-line interfaces.
 - b. All the queries that belong to one corpus research project are kept in one place.
 - c. The wrapper provides a table-oriented output with numerical results of our queries.
 - d. The wrapper shows additional context for each output of a query.
 - e. The wrapper can contain functions that allow access beyond the current sentence.

Since the advantages mentioned in (168) constitute the motivation for deviating from available tools and using our own development, I would briefly like to explain the importance of the points that are being made. Researchers in linguistics, even those involved in corpus linguistics, do not necessarily have the advanced computer skills that are found with those working in the field of computational linguistics (168a). Researchers are probably keener to work in a what-you-see-is-what-you-get environment than in a command-line environment. A windows-oriented wrapper—be it in the form of a stand-alone program or in the form of a web-based application—would therefore be much more suited for such researchers (as well as the students they work with, probably).

A totally different reason for having a wrapper around the Xquery engine is (168b), which states that a wrapper program may be made in such a way, that it keeps all the queries belonging to one particular corpus research project together. These queries (as well as the corpus research project itself) can be supplied with meta-information, so that retrievability of corpus research projects that have been done in the past increases, and research projects can be archives in a way that allows them to be retrieved successfully at a later stage.

Queries in corpus research projects often work like filters: the output of a query is a compilation of the sentences found in the input that satisfy the conditions stated in the query. The advantage of having a wrapper, according to the reason in (168c), would be that the wrapper program is able to give a table of the number of sentences fulfilling the conditions in a query, and subdivide this table over time-periods or text genres. The wrapper program would even be able to accompany a query's output

with a definable context of x preceding and y following sentences, as stated in (168d).

The program “**CorpusStudio**”, which I wrote, is a wrapper around Xquery (as well as around CorpusSearch2 oriented projects) which not only allows one to process texts sentence-by-sentence, and add additional Xquery functions that have access to the text as a whole, but also provides for the functionality listed in (168a-d). It is a stand-alone program, which has the advantage that a user is not dependent upon the availability of internet access. There are drawbacks to stand-alone solutions too, such as the fact that such solutions are usually limited to one or two platforms (CorpusStudio is only available for the Windows platform). Future work should therefore seek to make available a web-based implementation that provides CorpusStudio’s functionality in a platform independent way. Both the stand-alone as well as the web-based approaches should include a way to share and improve user-defined Xquery functions and they should stimulate collaboration in projects. Corpus research assignments for courses too might benefit from a web-based approach.

Be that as it may, CorpusStudio is a user-friendly environment to query the existing parsed English corpora as well as the coreferenced English corpora discussed in chapter 6. A full description of the program can be found in the user’s manual (Komen, 2009b), and section 14.1 of the appendix provides a short introduction to the main relevant functions of the program. What we will do here is show the most important functions that have been added to the wrapper, implementing the functionality stated in (168e).

The program CorpusStudio not only provides a wrapper around the software that is used to perform queries, it also adds several built-in Xquery functions that are either useful or essential. The useful functions are built-in shortcuts that come in handy for the work with syntactically annotated texts, but that could be rewritten as user-defined functions. The essential functions provide for functionality that could otherwise not be encoded by user-defined functions.

7.3.1 Antecedents and coreferential chains

Chapter 6 has explained a method to add “referential” information to the already available syntactically parsed English texts: each noun phrase receives a referential category (from the set of categories that has been defined in chapter 4), and if the noun phrase relates to a particular noun phrase occurring earlier (or later) in the text, then a pointer to that constituent is added in the *xml* code of the text. As has been argued in section 7.2, accessing noun phrase antecedents that do not occur in the same sentence as the noun phrases referring to them requires additional functions from the wrapper. The two most important functions that provide this functionality are in (169).

(169) *CorpusStudio Xquery basic antecedent functions*

ru:ant(\$ndThis) – Return the antecedent of node \$ndThis.

ru:chnext(\$ndThis) – Get the first node that has \$ndThis as antecedent.

The function `ru:ant()` and `ru:chnext()` provide the most elementary information that is needed to work with antecedents. If we have a noun phrase and we want to access the constituent it refers to (provided it refers to something), then we get this constituent by using `ru:ant()`, even if the constituent is located in another sentence in the text we are working with. The function `ru:chnext()` works the other way around: if we have a noun phrase, and we want to know which other noun phrase refers to our noun phrase, the function `ru:chnext()` gives us the *first* constituent that has ours as antecedent. The referentially enriched texts are such, that each constituent can have no more than one antecedent, but there may be more constituents having one and the same antecedent. This is the case, for instance, if there is a parenthetical constituent, as in (170).

- (170) a. In a battle there, he_i took prisoner [a certain **gentleman**, [by name M. **Zadisky**]_j, of Greek extraction, but brought up by a Saracen officer]_j,
 b. [**this man**]_j he_i converted to the christian faith, after which he_i bound him_j to himself_i, by the ties of friendship and gratitude, and he_j resolved to continue with his_j benefactor_i. [reeve-1777:12-13]

The first line of this story (170a) introduces a new person by using a generic expression *a certain gentleman*, which is then postmodified by a number of different characteristics that serve to make clear who this person is (they help to link the mental representation of this person with elements stored in long-term memory). The first postmodification *by name M. Zadisky* is encoded by the creators of the corpus as a parenthetical NP. Since this parenthetical NP refers to exactly the same physical person as *a certain gentleman*, the referentially enriched version of this text has a link of type “Identity” from the parenthetical NP to it. However, the policy of creating referentially enriched texts is to *not* have parenthetical NPs as main elements in a coreferential chain (the chain of noun phrases referring one to the other with a link of type “Identity”: they co-refer to the same participant). This is why the next NP that refers to our newly introduced person, the NP *this man* in line (170b), refers back to the constituent *a certain gentleman* instead of the parenthetical *by name M. Zadisky*. As a result, there are two constituents having *a certain gentleman* as antecedent: (1) the parenthetical NP *by name M. Zadisky* and (2) the next NP on the coreferential chain *this man*.

All texts in general, but narratives in particular contain references to participants that form coreferential chains, and investigating these chains can help us answer questions such as: “How are major participants (those with long chains) encoded, and how minor ones?” and: “Which syntactic constructions are used to start new major or minor participants?” The coreferential chains consist of what I would call a “main line”, which is the chain that runs from the last mention of a participant back to its first mention via a continuous series of backwards references of type “Identity”. There can, however, be some small branches to this main line, for instance in the form of appositive noun phrases (such as *by name M. Zadisky* in (170a) above). In order to distinguish constituents that are linked with “Identity” and are on the “main” line of a coreferential chain from those that are not, CorpusStudio comes with two additional functions, as shown in (171).

(171) *CorpusStudio Xquery advanced antecedent functions*

ru:antidt(\$ndThis) – Return the antecedent of node \$ndThis, provided the link to this antecedent has referential status “Identity”.

ru:chnextidt(\$ndThis) – Return the first node that has \$ndThis as antecedent, and that has a link to it with referential status “Identity”.

The function `ru:antidt()` gives an antecedent of the node that is passed on as argument, provided that the link to this antecedent is of type “Identity”. If the function were applied to *this man* in (170b), it would return the constituent *a certain gentleman*. This function could have been derived from `ru:ant()` in Xquery, since it really is a shortcut to taking the resulting node of the `ru:ant()` function, and testing whether this node has the value “Identity” for the feature “RefType”. The function `ru:chnextidt()`, however, provides functionality that cannot be obtained in an alternative way. It checks constituents that have the argument passed on in `ru:chnextidt()` as antecedent, and returns the one that is nearest and has a link of type “Identity”. In example (170), the result of applying `ru:chnextidt()` to the constituent *this man* would be the pronoun *him* in the same sentence.

If one would want to traverse the constituents that form a coreferential chain step-by-step, then the function `ru:antidt()` allows us to do this in a “backward” way (starting further on in a text and going back to an antecedent, the antecedent of that antecedent and so on), whereas the function `ru:chnextidt()` gives us the possibility to do this in a “forward” way.⁶

There is one more function that makes working with coreferential chains easier, and that is the function `ru:chlen()`, as shown in (172).

(172) *CorpusStudio Xquery coreferential chain functions*

ru:chlen(\$ndThis, ‘following’) – Return the number of constituents in the coreferential chain that runs from the end of the chain until reaching \$ndThis.

ru:chlen(\$ndThis, ‘preceding’) – Return the number of constituents in the coreferential chain that starts with \$ndThis, and then runs from antecedent to antecedent.

The function `ru:chlen()` gives the length of the following or preceding coreferential chain, and this is something we may find useful as we are investigating the relation between syntax and information structure. If we were to apply `ru:chlen(..., ‘preceding’)` to the constituent *this man* in (170b), it would return a length of ‘2’, since the preceding coreferential chain consists of (1) *this man* and (2) *a certain gentleman* (the parenthetical NP *Zadisky* is not part of the backward-running coreferential chain of *this man*). If we apply `ru:chlen(..., ‘following’)` to the constituent *this man* in (170b), then we would get a length of ‘3’. The coreferential chain that follows consists of (1) *this man*, (2) *he* and (3) *his*.

7.3.2 Preceding and following sentences

Suppose we want to retrieve the answer to a question—which is a practical example that is actually needed in section 9.10. The question will usually be in one sentence,

which is part of one `<forest>` in the *psdx* file, and if there is an answer to this question (so that it is not a rhetorical question), then it is very likely to be found in the immediately following sentence—hence in the following `<forest>`. There is a problem in accessing the following sentence, since *CorpusStudio* processes one `<forest>` a time, which means that the normal Xquery commands that are used only have access to the elements in this one single `<forest>`, and not in the following one. So accessing the next sentence by through a path specification like `$ndThis/ancestor::forest/next-sibling::forest[1]` does not work, because the context available to the Xquery processor only holds one `<forest>`, and not any preceding or following ones. This is where the built-in function `ru:line()` comes in, as defined in (173).

(173) *CorpusStudio Xquery function to access preceding and following lines*

ru:line(\$intNumber) – Return the `<forest>` element that is `$intNumber` away. If the number `$intNumber` is negative, then return the `<forest>` element that precedes `$intNumber` lines earlier.

The function `ru:line()` takes a number as argument. If we want to have the immediately following sentence, then we can specify `ru:line(1)`. The immediately preceding sentence can be accessed likewise, but with a negative number: `ru:line(-1)`.

7.3.3 Matching strings

The parsed English corpora contain labelled constituents in a wide variety. Instead of just one label for a noun phrase, NP, there are labels like NP-SBJ (subject NP), NP-VOC (vocative NP), NP-RSP (resumptive NP) to name but a few. If we, for instance, want to get all different kinds of object NPs, we would need to look for constituents with label NP-OB1, NP-OB2, NP-PRD. There are two features that make working with groups of labels easier: (a) the Xquery facility to have variables with a global scope, and (b) the built-in *CorpusStudio* function `ru:matches()`. This function is defined in (174), and an example of using globally defined variables with this function is given in (175)

(174) *CorpusStudio Xquery function to match strings*

ru:matches(\$strIn, \$strPattern) – Check if the `$strIn` matches one of the patterns that are defined in `$strPattern`.

(175) *Matching a label with a globally defined variable*

```
1  $define $_object := 'NP-OB*|NP-PRD*';
2  ...
3  for $search in //eTree[ru:matches(@Label, $_object)]
4  ...
```

The example definition in line 1 of (175) assigns the value `NP-OB*|NP-PRD*` to the global variable `$_object`. (These global variables must be defined in a definitions file.) What we want to say with this definition is that objects can be any constituents whose labels start with NP-OB (so NP-OB1 is okay, and so is NP-OB2), as well as those whose labels start with NP-PRD. The part of the query that makes use of this

variable is in line 3 of (175). This line starts a `for`-expression that looks for all `<eTree>` nodes where the attribute string (which is in `@Label`) matches at least one of the expressions separated by vertical bars in the global variable `$_object`.

7.3.4 Returning output

The usual purpose of a query is to find constituents that satisfy particular criteria. A query such as “subS+V+O” in line #1 of Table 48, for instance, finds subclauses with a subject, a direct object and a finite verb. The code of such a query can look like the example in (176).⁷

(176) *Query* matS+V+O

```

1  <TEI>
2  {
3    for $search in //eTree[ru:matches(@Label, $_subIP)]
4
5    (: Get the subject, but exclude some non-subject types :)
6    let $sbj := tb:SomeChildNo($search, $_subject, $_nosubject)
7
8    (: Get an object, excluding non-object types :)
9    let $obj := tb:SomeChildNo($search, $_object, $_noobject)
10
11   (: Get the finite verb :)
12   let $vb  := tb:SomeChild($search, $_finiteverb)
13
14   (: All three constituents must exist :)
15   where ( exists($sbj)      and
16           exists($obj)      and
17           exists($vb)
18         )
19   return ru:back($search)
20 }
21 </TEI>

```

The query looks in each subclause (line 3), and has a match if it finds a subject (line 6), an object (line 9) and a finite verb (line 12). Line 19 concludes the query by returning the value provided the built-in function `ru:back()`. Essentially the only important thing done by `ru:back()` is to pass on the numerical identifier of the subclause constituent that was found to the wrapper, the program CorpusStudio (each constituent is kept as an `<eTree>` node, and each such node has an attribute `@Id` with a unique numerical identifier for that node). As the queries are being executed, the wrapper program does two things for each matching constituent it finds. It uses the numerical identifiers to locate the whole sentence as well as the user-defined number of preceding and following context sentences, and appends this information to appropriate *html* output files, which will later be shown in the output (see section 14.1.4 in the appendix). The second thing done by the wrapper program is that it stores the resulting constituent’s identifier in an *xml* output file that can be referred to later.

The output of one query, then, is *not* a collection of `<forest>` nodes containing the matching `<eTree>` nodes, feeding into the next query in line (see Table 48 for an example of a collection of queries that need to be executed). The reason for this is the query execution order as specified in (345): since all necessary queries are executed one after another on one `<forest>` node, the fact that there has been a match on a particular query or not is enough to determine whether the next query

should be executed, and if execution of the next query is needed, then the available `<forest>` node can be passed on to that query without the need for the previous query to return it as output. If a `<forest>` node has been “ousted” by one query, it is not fed into subsequent queries, which prevents unnecessary processing time.

7.3.5 Returning numbers

The result of executing one or more queries is a table (such as Table 49) with the number of occurrences of the phenomenon (such as subclauses with a direct object preceding the finite verb) we were looking for, divided over time periods (see section 14.1.4 in the appendix). But what if we are not looking for the *number of hits* of a particular phenomenon, but for some other numerical measure? We might, for instance, want to know the average size of preposed direct objects in English, and compare these with the average size of direct objects in their canonical position. Xquery allows us to count the number of words or constituents, and the built-in function `ru:avg()`, as shown in (177), can then be used to keep track of the averages.

(177) *CorpusStudio Xquery function to prepare numerical results*

`ru:avg($intNum, $strType)` – Include `$intNum` in the calculation of the averages for `$strType`.

`ru:ard($ndThis, $strType)` – Include the distance from `$ndThis` to its antecedent in the calculation of the average referential distances for `$strType`.

`ru:out($strCsv)` – Append the semicolon-separated list of values supplied in `$strCsv` to one global output file of this corpus research project.

A more specific numerical measure is the referential distance as introduced by Givón (1983). This is the distance between a constituent and its antecedent, and it can be measured in terms of number of words, number of clauses or number of sentences. The built-in function `ru:ard()` takes a constituent node as argument, and keeps track of the average referential distances.

The program *CorpusStudio* shows the results of `ru:avg()` and `ru:ard()` by adding a table with one row for each situation we have specified. The columns subdivide the results over the different time-periods that are specified, and each average is also supplied with the number of constituents evaluated and the standard deviation. A query that illustrates the use of the `ru:ard()` function is in (178).

```
(178) Query anyDem
1 <TEI>
2 {
3   for $search in //eTree[ru:matches(@Label, $__anyp)]
4     (: Check the Nptype feature of this NP :)
5     let $type := ru:feature($search, 'Nptype')
6
7     (: Store referential distances, depending on the type :)
8     let $bHasArd := if ($type = 'Dem') then
9       ru:ard($search, 'Dem')
10    else
11      false()
12
13   where ( $bHasArd )
14
15   (: Return the independent demonstrative we found :)
16   return ru:back($search)
17 }
18 </TEI>
```

Line 3 of query `anyDem` considers all the noun phrases in a text, and then gets the value for the feature “Nptype” in line 5 (this feature has been added to the parsed texts in preparation of coreference resolution with Cesax, as explained in section 6.3.1). If this value is “Dem”, then we have an independent demonstrative pronoun, and in that case line 9 calls the `ru:ard()` function with the noun phrase node as argument and the type specified as “Dem”.

Table 21 The result of using the `ru:ard()` function

Type	OE			ME			eModE			LmodE		
	Count	ARD	Sdev									
Dem	71	2,58	4,03	56	2,80	4,01	46	2,74	3,42	73	2,64	3,00

The result of applying the query in (178) to those texts that have been referentially enriched is given to us by CorpusStudio in the form of Table 21. The results are divided over the four main periods, and each of these periods has the average referential distance (ARD), the number of occurrences taken into consideration (Count) and the standard deviation (Sdev).⁸

Greater flexibility than `ru:avg()` and `ru:ard()` is offered by the output function `ru:out()`, as shown in (177). This function allows one to pass on one or more numerical or text results per hit to a “csv” file, which is a semicolon-separated file that can be read by a program like Microsoft Excel. Each line is automatically supplied with the locational details of the hit: the name of the text, the period abbreviation of the text, and the line number of the text in which the result occurs.

Concluding section 7.3, we can say that the program CorpusStudio is not just a wrapper around Xquery, but adds additional functionality (in the form of the built-in functions that handle access to antecedents, coreferential chains and numerical results) to the corpus researcher who is interested in investigating the relationship between syntax and information structure.

7.4 Querying coreferenced corpora

We have seen that CorpusStudio allows considerable flexibility in defining functions, queries and query execution order. This section shows how we can put all of that to work in order to query coreferenced corpora. We do this with a task that serves to evaluate two things: (a) the value of the referentially enriched corpora, and (b) CorpusStudio’s facilities to help relate syntax with information structure.

Suppose we have a set of coreferenced texts and we set ourselves a task that combines syntactic information with referential status. Our task will be to look at main clauses that contain: (a) a subject, (b) a finite verb, and (c) a prepositional phrase. What we want to know is whether the proportion of PPs containing “New” information has changed significantly over time. Are PPs used more to express new information or not?

We will use a query for this task that employs CorpusStudio’s feature to “subcategorize” the output: divide the results of one query over a definable subset of categories. The query, `matS+V+PP`, retrieves all the main clauses with the correct content: a subject, a finite verb and a PP that contains at least one NP. It then determines the referential status of this last NP, and divides the results accordingly.

```
(179) Query matS+V+PP
1  for $adjunct in //eTree[ru:matches(@Label, $_anypp)]
2  (: Get the usual [search] value: the parent matrix IP :)
3  let $search := $adjunct/parent::eTree[ru:matches(@Label, $_matrixIP)]
4
5  (: Find the subject of this IP and the finite verb :)
6  let $subj := tb:SomeChildNo($search, $_subject, $_nosubject)
7  let $vb := tb:SomeChild($search, $_finiteverb)
8
9  (: Get the (first) NP object of the PP, and its reftype :)
10 let $obj := tb:PPobjectOrNP($adjunct)
11 let $ref := ru:feature($obj, 'RefType')
12 let $cat := if (ru:matches($ref, 'New|Inferred|Assumed')) then 'new'
13             else if (ru:matches($ref, 'Identity'))         then 'old'
14             else 'other'
15
16 where ( exists($subj) and
17         exists($vb) and
18         exists($obj)
19       )
20 return ru:back($adjunct, '', $cat)
```

Line (179.1) of the query starts by selecting prepositional phrases, which are characterized by having an `<eTree>` element whose `@Label` attribute matches one of those defined by the variable `$_anypp`.⁹ The prepositional phrase that is selected is assigned to the variable `$adjunct`.¹⁰ We would like to limit our search to main clauses with their direct child-constituents, which is why line (179.3) checks if the parent constituent of `$adjunct` is a constituent with a main clause label as defined in the variable `$_matrixIP`.¹¹ Having obtained the main clause variable `$search`, we can now look for the subject `$subj` in line (179.5) and the finite verb `$vb` in line (179.7).

Line (179.10) obtains a variable `$obj`, which contains the noun phrase governed by the PP node `$adjunct` that we are currently treating. The noun phrase `$obj` is the main object of our attention in this query, and we want to know the referential

status of this NP. We retrieve its referential status in line (179.11) through the built-in `ru:feature` function (see the online [manual](#)), which gives us the value of the grandchild `<f>` feature node with feature name `RefType`. Lines (179.12-14) derive the value of the subcategorization variable `$cat`, which can be *new*, *old* or *other*, depending on the particular referential category found for the noun phrase `$obj`.

The `where` clause in lines (179.16-18) makes sure that we only proceed if we have actually found a main clause (in `$search`) that contains a subject (in `$subj`), a verb (in `$vb`) and PP object (in `$obj`), without any specification as to the order in which these occur.

The last line of the query in (179.20) uses the built-in `ru:back` function, which makes sure that, if all the conditions have been met, we return a `<forest>` node as a result (this typically is one line in the text we are processing, see 6.2). These returned `<forest>` nodes are then used by CorpusStudio to count the number of results and to show the user where the results are located. The call to `ru:back` in line (179.20) also contains the subcategorization variable `$cat`, which makes CorpusStudio not only give us a row in the summary table with the number of PPs that meet all the conditions of (179), but it will make three additional rows, which give us the number of referentially *new*, *old* and *other* PPs.

Several lines in the query make use of functions such as `tb:SomeChild` and `tb:SOMEChildNo`—these are defined in the “Definitions” section of the CorpusStudio project. We will briefly consider one function to see how this feature of Xquery works.

```
(180) Function tb:SomeChild
1  (: -----
2     Name : tb:SomeChild
3     Goal : Return the first child of [$this] having a label like $strLabel
4     History:
5     24-02-2010      ERK      Created
6     ----- :)
7  declare function tb:SomeChild($this as node()*, $strLabel as xs:string?)
8     as node()?
9  {
10     (: Get ALL the children of me :)
11     let $all := $this/child::eTree
12     (: Select those that have the indicated label :)
13     let $ok := $all[ru:matches(@Label, $strLabel)]
14     return
15     if (empty($ok))
16         then ()
17         else $ok[1]
18 }
```

The function `tb:SomeChild` as shown in (180) starts with a `declare` line where the input arguments and the output type are defined. Line (180.10) gets all the direct `<eTree>` children of the input node `$this`, and line (180.12) selects those of the children that have a `@Label` attribute like the `$strLabel` argument supplied by the calling function. The function finishes in lines (180.13-16) by returning either “nothing” if we have not found a child fulfilling the conditions, or else the first child that fulfils the conditions.

When the query in (179) is executed, we get the number of PPs that are *new*, *old* and *other* according to the referential status division made in (179.12-14). Table 22

gives the numerical results, as divided over all the subperiods where the enriched corpus texts are from.

Table 22 Prepositional phrases in main clauses found by query (179)

Result	O3	O14	M1	M2	M3	M4	E1	E2	E3	B1	B3
matS+V+PPnew	74	51	50	31	43	43	137	53	49	108	319
matS+V+PPold	62	48	25	38	13	20	67	37	16	52	79
Texts in this period	1	2	1	1	1	2	2	1	2	2	3
D[corp]	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

The number of occurrences is limited, but if we combine the results into the four main periods (Old English, Middle English, early Modern English and late Modern English), then we get a good idea of the development. Figure 15 shows the result of combining the subperiods into larger periods (O3 and O14 are both part of OE).¹²

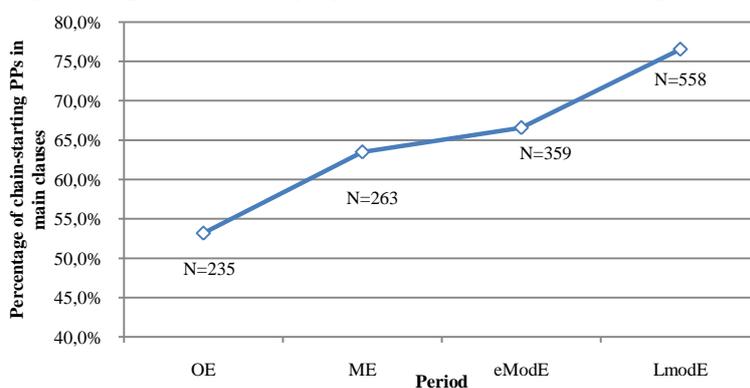


Figure 15 Chain-starting PPs in main clauses

What we see graphically in Figure 15, and quantitatively in Table 22, is that the PPs in main clauses are increasingly *new* by the definition in (179.12). The question arises what kind of newness this is. The referential statuses that form the category *new* as in (179.12) are: “New”, “Inferred” and “Assumed”. These are the referential statuses a constituent has that can potentially start off a coreferential chain. Those with referential status “New” are new to the addressee as well as to the discourse. Those with status “Inferred” infer a new participant from an existing one, and those with category “Assumed” refer to an addressee-known entity. NPs in all three categories can be referred to subsequently, and are therefore the constituents that can lie at the basis of coreferential chains.

There are, as usually is the case in corpus research, several questions coming up from the discussion so far. If the PPs from query (179) and Figure 15 have such a “wide” definition of newness, we would like to know whether PPs that are *new* in a stricter sense behave. A follow-up experiment, a variation to the query in (179), selects PPs in main clauses *and* subordinate ones, and calculates the percentage of *new* PPs according to two definitions: (a) those that start off a coreferential chain

(defined as in 179.12), and (b) those that are new in a strict sense: they have referential category “New”, and do not even have an anchor.¹³ The results of this experiment are shown in Figure 16.

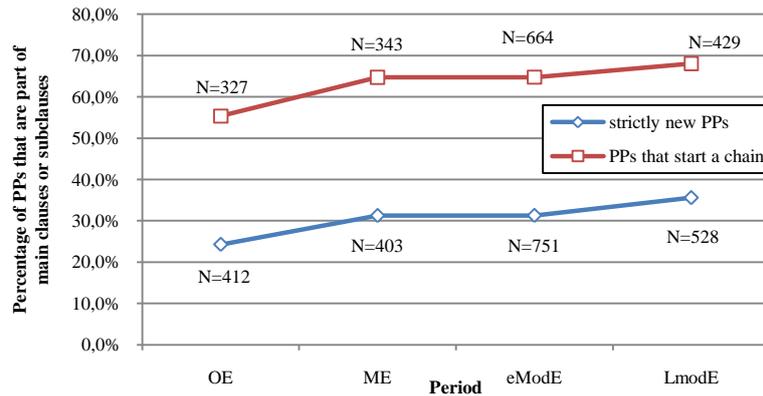


Figure 16 New and chain-starting PPs found in main clauses and subclauses

What we can conclude from the tendencies in Figure 16 is that even strictly new PPs gradually increase from just over 20% in OE to almost 40% in LmodE.¹⁴ This means that PPs are increasingly being used as a vehicle to contain unestablished information instead of established information. One more observation is that the picture for *all* finite clauses as in Figure 16 does not greatly differ from the picture we obtained for just the *main* clauses as in Figure 15.

We would now like to know what *kind* of coreferential chains are started by PPs, and this is where the Xquery facilities of CorpusStudio can be put to an even fuller use. We have two questions about the nature of these chains. The first question concerns the length distribution of the chains being started by PPs. The experiment that is needed to get the distribution of the lengths of the chains formed by PPs in main and subordinate clauses is shown fully in (181).

(181) *Query finS+V+PPchain*

```

1 for $adjunct in //eTree[ru:matches(@Label, $_anyp)]
2   (: Get the usual [search] value: the parent matrix IP :)
3   let $search := $adjunct/parent::eTree[ru:matches(@Label, $_finiteIP)]
4
5   (: Find the subject of this IP and the finite verb :)
6   let $subj := tb:SomeChildNo($search, $_subject, $_nosubject)
7   let $vb := tb:SomeChild($search, $_finiteverb)
8
9   (: Get the (first) NP object of the PP, and its reftype :)
10  let $obj := tb:PPobjectOrNP($adjunct)
11  let $ref := ru:feature($obj, 'RefType')
12
13  (: Filter out the Inert and NewVar ones :)
14  let $ok := ru:matches($ref, 'New|Inferred|Assumed')
15
16  (: Get the distribution of the chainlength :)
17  let $distri := ru:distri(ru:chlen($obj, 'following'), 'finNewPP')
18
19  where ( exists($subj) and
20         exists($vb) and
21         exists($obj) and
22         $ok and
23         $distri
24       )
25  return ru:back($adjunct)

```

Line (181.14) in the query makes sure we only get PPs that can potentially start off a chain. The distribution of the chain is then taken care of by two built-in functions in line (181.17). The function `ru:chlen` obtains the length of the chain starting at the PP's noun phrase. This function “walks” the coreferential chain in order to find the chain length. The `ru:distri` function is one of the built-in statistical functions. It keeps track of the chain lengths and, after running the query through all the texts, gives a logarithmically scaled distribution of these lengths. The results of this query are in Table 23.

Table 23 Length distribution of chains started out by main clause and subclause PPs

length range	OE	ME	eModE	LmodE
1	89,0%	83,5%	80,0%	88,4%
2	6,1%	8,7%	11,7%	6,2%
3-4	3,3%	4,5%	5,4%	4,0%
5-8	0,0%	1,6%	2,0%	1,0%
9-16	1,7%	1,6%	0,7%	0,5%
17-32	0,0%	0,0%	0,2%	0,0%

The distribution of the lengths of the chains as shown in Table 23 tells us that there are no major changes going on. So, even though the PPs increasingly are being used to start off chains of participants, the distribution of the lengths in these chains does not change dramatically. The numbers in OE and LmodE are quite comparable, in fact.

The last question we would like to be answered also concerns the difference in chains started by PPs. We want to know whether there is a change in the number of such chain-starting PPs that contain at least one subject constituent. The presence of a subject constituent on a chain is a rough indication that the chain belongs to a

participant of some importance, since it is typically the subject that can function as agent of an action.¹⁵

The way to measure the presence of a subject on a coreferential chain is to use an Xquery function that “walks” the chain: it transitions from one constituent to the next by using a built-in function like `ru:chnext`. As it does so, it checks if the constituent it ends up in is a subject or not. Walking a chain in this way can be done by using a “recursive” Xquery function: one that keeps invoking itself until specified conditions are met. While using such functions is quite technical, the fact that Xquery allows one to do so is very practical for our purposes, and it demonstrates nicely how we can make use of the coreferential chains that have been derived through the texts we have enriched in Cesax.

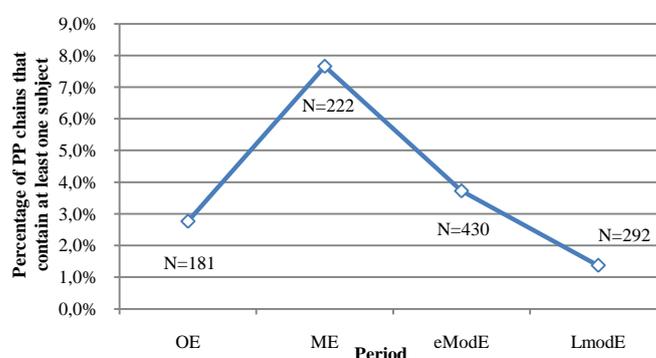


Figure 17 PP-initiated chains with at least one subject

The results in Figure 17 show that the percentage of PP-initiated chains with at least one subject (which is an indication of a relatively substantial participant in a story) is small overall, ranging from 3% in OE to 7% in ME.¹⁶ The largest change is, in fact, the transition from OE to ME, and after that the percentage gradually decreases into LmodE to reach a level that only marginally differs from OE.

We may conclude, then, that PPs are gradually being used more often to point to strictly new information, they are also gradually being used more often to start off coreferential chains, but the length distribution of these chains does not change dramatically, and their use to point to relatively more important participants remains marginal.

The examples in this section illustrate how CorpusStudio is able to combine syntactic and referential information to yield results in the area of diachronic information structure research. What CorpusStudio needs in order to do this, is combine two pieces of information: (a) the syntactic information that already is available in the parsed English corpora, and (b) the referentially enrichments to these parsed corpora. This, then, illustrates that the combination of Cesax and CorpusStudio allow us to find answers to research questions that are involved in the interaction between syntax and information structure.

7.5 Discussion

In order to investigate the overall research question on how the interaction between syntax and focus changed in English over time (as formulated in 11), we need to have corpora that contain syntactic as well as referential information, since these basic building blocks allow us to determine the focus domains. The computer program Cesax (described in chapter 6) allows enriching existing parsed English with coreferential information, which tells us which constituent has which other constituent as antecedent, and it also tells us what the referential state (in terms of the Pentaset) of each noun phrase is. The program does this in a semi-automatic way, saving us a lot of manual labour. Once syntactically parsed texts are enriched with Cesax, this yields texts in the *psdx* format (an *xml* format where the hierarchical structure of labelled bracketing from the Treebanks has been replaced by a hierarchical structure of embedded tree tags), which contain both syntactic as well as referential information.

Section 7.1 of this current chapter compared existing query languages in their applicability for the information structure related corpus searches we will do in chapters 8 and 9. While the existing CorpusSearch2 engine would be able to work with an adaptation of the bracketed labelling treebank format that contains the referential state labels for each noun phrase, it would not be suitable for algorithms that require access to the antecedents of constituents (7.2). This reason, as well as the major advantages in using an *xml* kind of format (see the arguments in (147) of section 6.2) are enough justification to use the open standard “Xquery” language as search engine. The computer program “CorpusStudio” that is described in section 7.3 provides a user-friendly interface to define corpus research projects that make use of the Xquery language (although it also facilitates querying the existing treebank files using the CorpusSearch2 engine). CorpusStudio has several built-in Xquery functions that allow easy access to antecedents, even to the extent of following coreferential chains downwards or upwards.

A limitation of CorpusStudio is the fact that the program right now basically is a single-platform (Windows) stand-alone program.

Section 7.4 demonstrates CorpusStudio’s capabilities in the area of information structure research by a case study on the development of discourse-new prepositional phrases. The results of the case study are twofold. In terms of the development of the prepositional phrases the case study shows that PPs are gradually being used more often to point to strictly new information, and they are gradually being used more often to start off coreferential chains, but the length distribution of these chains does not change dramatically, and their use to point to relatively more important participants remains marginal. In terms of this current chapter’s goal, the case study effectively demonstrates that the combination of texts enriched with Cesax and then queried with CorpusStudio is capable of handling the kind of information structure related questions we are looking for within the overall framework of this study. This means that we are now ready for the real corpus research: the quest for presentational focus changes in chapter 8 and the search for changes in the expression of constituent focus in chapter 9.

¹ Using a command file (also called “batch file”) that holds the calls to particular queries in the desired order only partly alleviates these problems, since the command file does not hold the text of the queries, which are kept in separate query files. And command files themselves are meant to be used under a command-prompt (or shell), so that these, again, require advanced computer skills from researchers working in linguistics.

² Some of these query languages are aimed at searching the Negra (Tiger) *xml* format, which is a stand-off format: a format where nodes are stored in lists, and the hierarchy is resolved by cross-list indexing. The Alpino aimed tools (the Xpath implementation “DtSearch” and Alpino’s Xquery tools) are aimed at *xml* formats that are ordered hierarchically. But conversion between the stand-off and hierarchically ordered *xml* formats is always possible.

³ The Xquery language does not prohibit accessing earlier (or later) sentences from any constituent in principle. The main reason Alpino is not able to do this is that it stores texts per sentence instead of as a whole. One of the reasons this may have been done is the fact that Xquery processing can easily run into memory problems when larger texts are processed as a whole.

⁴ It would be possible to access the information *n* sentences back through an Xpath axis definition like `ancestor::forest/preceding-sibling::forest[n]`, provided all the `<forest>` nodes (the sentences) are loaded in memory.

⁵ This access is not through Xquery functions, but either through less memory-intensive Xpath functions or, if that is not an option in the wrapper we use, by programmatically “walking” an *xml* document node-by-node.

⁶ The notions “backward” and “forward” are not necessarily equal to “anaphoric” and “cataphoric”: if we traverse a chain “backward”, the surface position of the antecedent to the noun phrase we are currently visiting will usually be located in the preceding part of the text, in which case we have an anaphoric reference, but it may also be situated *after* the constituent we are currently visiting, in which case there is a cataphoric reference.

⁷ The actual query will be more complicated, if one, for instance, wants to exclude empty subjects, and exclude interrogative sentences, to name but a few realistic criteria.

⁸ The main point being made here, which is that CorpusStudio *allows* the researcher to investigate the average referential distance. I realize that the standard deviation of the numbers given here is actually quite high—so high that the averages can no longer be called significant. The reason for this is that the referential character of demonstrative pronouns differs greatly, depending on things like (a) whether this is a near demonstrative (like “this”) or remote one (like “that”), and (b) what the grammatical role of the demonstrative is. The simple example shown here has lumped together *all* demonstrative pronouns, without taking these and other factors into account.

⁹ This variable is defined in the “Definitions” section of the corpus project as a shortcut for nodes with the label `PP` as well as those with the label `PP-*`. The latter ones are PPs with a further (often functional) specification, and include, for instance, `PP-LFD` (a PP that occurs in a left-dislocated position).

¹⁰ As a matter of convention, we use the `$_` prefix for globally defined variables and the simple `$` prefix for variables that are defined inside the Xquery function where they are being used.

¹¹ The variable `$_matrixIP` is a shortcut for nodes with a label like `IP-MAT`, `IP-MAT-SPE` etc.

¹² The sub period “O14” means that we have an Old English manuscript from the 4th (final) subperiod of OE, but the original could have been from any time within OE, starting with O1.

¹³ An example of an anchored NP is *his voyages to India* in (109b), which as an NP is referentially “New”, but links to an existing participant through the “anchor” pronoun *his*. Anchored NPs are not as new as unanchored ones.

¹⁴ D[corp] is 100%. Fisher’s exact test shows for “PPs that start a new chain”: the change from OE to ME is significant ($p < 0,05$), as is the change from OE to LmodE ($p < 0,05$), but the changes from ME to eModE and from eModE to LmodE are not significant. Fisher’s test for “Strictly new PPs” gives similar results. See for details the appendix, section 14.3.1.

¹⁵ A more advanced study would have to take into account the *kind* of action (mirrored in the kind of verb) that the participant belonging to the PP-started chain takes. While the measure we take here is, therefore, but a very rough estimate, it is nevertheless important, since it illustrates the capabilities of intelligently “walking” the coreferential chains that CorpusStudio supports.

¹⁶ The only significant difference according to Fisher’s exact test ($p < 0,05$) is: OE-ME ($p = 0,0475$). See for details the appendix, section 14.3.314.3.1.

Part IV

Results

The corpus research described in this and subsequent chapters builds on the groundwork that has been provided in the preceding chapters. We started by recognizing that clauses can be divided into three basically different focus articulations, depending on the focus domain: constituent focus (the domain is one constituent), topic comment articulation (the domain is the predicate) and thethetic focus articulation (the domain is the subject + predicate). I have alluded to a unidirectional relation between “newness” and “focus” in section 3.5: any constituent that is new within the mental model that the addressee creates as the discourse develops is extremely likely to be part of the focus domain. I have subsequently shown that syntactically parsed corpora can be enriched with referential information (chapters 5-6), which I claim forms the basis of information structure notions, and that this information is accessible for queries combining syntax and information structure (chapter 7).

The chapter at hand combines the consolidated information in an elegant way: I am going to quantify changes that have taken place in English in the expression of presentational focus, which involves those thetic articulation constructions where the *subject* is the most informative part. Bailey, who worked extensively on the thetic articulation in ancient Greek, writes about this articulation:

- (182) “I use the term ‘thetic’ for a sentence that serves primarily to introduce an entity or state of affairs into the discourse (what is also called ‘presentational’ function) and I assume that theticity is prototypically expressed cross-linguistically by ‘sentence-focus’ constructions (i.e. where **the subject is in some way marked as non-topical**).”
(Bailey, 2009 - emphasis mine)

Crucial to my method for finding instances of presentational focus is Bailey’s observation that presentational focus can be recognized by a “subject that is in some way marked as non-topical”, which is in line with Lambrecht (1994). The approach we take here is to look for “new” subjects, which are, as by the line of thought expressed above, extremely likely to be part of the focus domain, and, consequently, an indicator of presentational focus.

The impact statement (61) in the introductory part of chapter 4 gives an idea of what we are going to find: the changes in English syntax lead to increasing placement of the subject before the finite verb, as visible from the decrease in subject-auxiliary inversion (see section 4.3, Figure 4), but this jeopardizes the “late-subject” construction (see section 4.2.5), which has been the construction par excellence for the expression of presentational focus. Recognizing this syntactic change, I nevertheless posit the following hypothesis:

(183) *Presentational focus hypothesis*

The position for presentational focus in written English will remain to be after the finite verb, despite the loss of V2.

What I am arguing is that the position where presentational focus occurs remains to be after the finite-verb; either in the PostCore area as in OE, or in the Core area. A number of forces conspire to achieve this. First, there is the placement of the subject in the PostCore area where it deviates from the SV word order, which forms a clear signal of focus. Second, there is the Principle of Natural Information Flow (see section 3.3.1) which stipulates that the non-established newly introduced participant be as far to the end of the clause as possible. And third, there is the aim of presentational focus: introduce (or reintroduce) a participant, and then make a comment about it. This last requirement is best met if the first and second mention of the participant are as close to one another as possible, which means that the first mention should be close to the end of the clause in which it is introduced.

But how is the hypothesis in (183) met where the syntax of English changes? The main impact of the loss of V2 on strategies of presentational focus was that all subjects became preverbal. Recall from the slot-structure (52) in chapter 4 that OE subjects could occur in the PreCore, the Core and the PostCore areas, whereas the slot-structure (53) for LmodE has only retained a dedicated subject position in the PreCore area. The loss in subject positions jeopardized the late-subject constructions—which has reduced to locative and severally well defined other inversions (Birner and Ward, 1998). This chapter will show that there is another construction coming up in ME, one with the expletive pronoun *there*, which ultimately takes on the function of presentational focus. The reason for this, as will become clear in 8.4.3, is that this construction satisfies the forces conspiring together that are mentioned above.

Before we start looking for presentational focus, we will have a closer look at the notion of newness (in section 8.1) where we will also consider the limitations of this approach. We will then look at the texts we are using and the algorithm that helps us find instances of presentational focus (8.2), and then continue with several experiments (8.4 until 8.6).

8.1 Newness and presentational focus

The question what “new” information is came up in chapter 5, where the referential state primitives were introduced in relation to the mental models discussed in chapter 2. Instead of making a binary distinction between “established” information and “unestablished” information, and instead of making gradual distinctions between information that is “less established” and “more established”, the Pentaset of referential state primitives (chapter 5) recognizes five states: “Identity”, “Inferred”, “Assumed”, “Inert” and “New”. As the label “primitive” suggests, these referential states are building blocks from which we can derive the information state categories we need. The question now is what kind of “new” information state we need to have in order to find the “new” subjects that are part of clauses with a presentational focus articulation.

The approach we take is to detect presentational focus by finding “new” subjects, where “new” is defined in different ways: (i) referentially “New” subjects (section 8.4), (ii) unanchored referentially “New” subjects (section 8.5), and (iii) referentially “Identity” subjects with a relatively distant antecedent (section 8.6).

There is at least one category of presentational focus that we will not be able to capture with our approach of looking for new subjects, since there are situations where the subject is “most informative” (see Bailey’s definition in 182) even though it is not “new”. A subject can be most informative without being new if it represents a participant who appears at a location where his physical appearance was not expected. An example of such an opportunity that our approach will miss is given in (184).

- (184) a. Ongemang þissum, com ham Pafnuntius, [coeuφr:88]
 In.the.midst of.this came home Paphnutius.
‘In the midst of this, Paphnutius came home.’

This example is taken from the “Euphrosyne” text discussed in chapter 4. A little bit of context is necessary to understand that this is indeed an example of presentational focus. Paphnutius is the father of the main character, the woman called Euphrosyne. She has, just before we get to the sentence above, secretly been making enquiries into the possibility of entering the monastic life, something which she fears to be against her father’s wish. She had taken this opportunity, while her father was away on one of his trips. It is at that point that the sentence in (184) informs us that her father, Paphnutius, comes home. His arrival was not something Euphrosyne had been hoping for or expecting. So the appearance of Paphnutius onto the scene is unexpected and surprising.

This is enough information about the context, and we can now continue by looking at the rationale for analysing this sentence as an example of presentational focus. The crucial question to ask is: “Where is the focus domain?” We can exclude *ongemang þissum* ‘in the midst of this’ from the focus domain, since this is clearly a temporal point of departure (see chapter 3, section 3.3.2 on points of departure). Of the remainder *com ham* ‘came home’ is a predicate that clearly gives the reader some new information, especially in relation to the main character Paphnutius. And even though the subject *Pafnuntius* refers to a person who can be regarded as “established” information, Paphnutius is the most informative part of the sentence at this point, because previous clauses (a) had Euphrosyne as topic and (b) were in a physical situation excluding Paphnutius. The fact that *he*, of all people, enters this physical situation is the most informative one, so we must, at the very least, include the subject Paphnutius in the focus domain. This leaves us with a focus domain that contains the predicate as well as the subject, so that the clause, by definition, must be categorized as having athetic focus articulation, and more specifically, a clause that has presentational focus, since the subject holds the most important piece of information (be it in relation to the predicate and, indeed, the addressee’s mental model of the physical situation).

Examples like (184) above will not be found by the approach taken in this chapter, since they crucially make use of the way in which characters enter and

leave scenes. This information is not available to us or derivable from the syntactic and the referential information with which the texts have been annotated and would require a detailed pragmatic analysis of each text.

Nevertheless, we *can* have a look at all the situations where “new” characters come into a text and see how these are handled. We can take note of the word orders and constructions used to deal with such kind of presentational focus. We can then compare our findings from different time-periods and see what diachronic trends we observe and relate these trends to diachronic changes in English syntax and information structure.

If we happen to be so lucky that we find particularly “exclusive” word orders or constructions used for presentational focus of “new” characters, then we can try to take the matter one step further. We can do the reverse of what we have been doing so far. We can look for these “exclusive” word order patterns, and check if we find instances where the subject is not really “new” or unestablished, and if these are instances of presentational focus of the kind illustrated by example (184), where the subject *is* the most informative part of the clause on contextual grounds. The late-subject position may be a good candidate for this.

8.2 Looking for presentational focus

The corpus-based investigation into presentational focus described in this chapter builds on the enriched syntactically annotated English corpora (see chapter 6). Relevant constituents have been enriched with the referential state primitives defined in the Pentaset (see chapter 4).

The restriction we have in this approach is the very size of the corpus we are working with. Only a limited number of texts have been “cesaxed” (that is: supplied with referential annotation using the Cesax computer program). Continuing efforts are on the way to extend the size of the referentially enriched corpus, but the status of the corpus, at the time of writing, is shown in Table 24.

Table 24 Texts that have been enriched with referential information

Text file	Name	Words	Period
CoApollo	Apollonius of Tyre	6545	OE (950-1050)
CoVinceB	Saint Vincent	728	OE (1050-1150)
CoEuphr	Euphrosyne	3658	OE (850-1150)
CmSawles.m1	Sawles Warde	4111	ME (1150-1250)
CmKentse.m2	Kentish Sermons	3534	ME (1250-1350)
CmHorses.m3	Horses	5902	ME (1350-1420)
CmReynar.m4	Reynard the fox	8850	ME (1420-1500)
CmCapser.m4	Capgrave's sermons	1569	ME (1420-1500)
Fabyan-e1-h	Fabyan's chronicles	5478	eModE (1516)
Fisher-e1-h	Fisher's sermons	4853	eModE (1521)
Perrot-e2-h	Perrot biography	4831	eModE (1592-1603)
Behn-e3-p1	Oroonoko	5475	eModE (1668-1688)
Jpinney-e3-p1	Letter from Pinney	881	eModE (1685-1686)
Brightland-1711	Brightland	1341	LmodE (1711)
Defoe-1719	Defoe	9378	LmodE (1719)
Skeavington-184x	Skeavington	9132	LmodE (184x)
Long-1866	Long	8851	LmodE (1866)
Fleming-1866	Fleming	9038	LmodE (1886)

Clauses with presentational focus having a new subject are found by querying the available texts through the help of the CorpusStudio program (see chapter 7). The corpus research project that is used for this purpose has an algorithm along the lines of (185).¹

(185) *Algorithm to get presentational focus clauses*

Step 1: Consider each NP in the text, and check if it satisfies conditions:

Condition a: grammatical role is "Subject"

Condition b: the NP is child of a main clause or subclause

Condition c: the clause is not an interrogative one

Condition d:

Approach i: referential status is "New"

Approach ii: referential status is "New" + NP has no anchor

Approach iii: referential status is "Identity" + distance > threshold

Step 2: Let *cat* be the word order type of this clause

Step 3: Let *len* be the length of the chain started by this NP

Step 4: Output:

Subcategorize on *len*

Provide *cat*

Show the clause of which the NP is part

The procedure outlined in (185) checks each NP in step #1 for the conditions, modulo the approach taken. Approach (i) simply checks the referential status, which has to be "New". Approach (ii) checks if the NP is "unanchored new" according to the definition in (193), and approach (iii) checks if the NP has a status of "Identity", but contains an antecedent that is further away than the threshold we derive

experimentally. It also checks to make sure the NP is really part of a main clause or subclause (instead of, for instance, a non-finite participial clause), and makes sure the clause is not part of a question (since questions throw in unexpected complications in terms of word order and referential statuses).

Step #2 of the algorithm stores the word order type of the clause, so that this is available as part of the output. The word order types recognized are the ones that determine the position of the new subject with respect to key elements of the clause: its start, its end, the position of the finite verb, and, if available, the position of a non-finite verb form such as a past participle. The word order types recognized by the Xquery implementation of the algorithm in (185) are shown in Table 25.

Table 25 Word order categories for subjects

Category	Word order
Initial	S ... V _{finite}
PreV	... S ... V _{finite}
VS	... V _{finite} S
Mid	... V _{finite} ... S ... V _{non-finite}
PostVnonf	... V _{finite} ... V _{non-finite} ... S
PostVf	... V _{finite} ... S

Step #3 in the algorithm in (185) determines the length of the coreferential chain that starts off at the NP currently being scrutinized. This information can be gained from the enriched texts, because every noun phrase stores a link to its antecedent—if it has one. The algorithm in (185) needs the length of the coreferential chain to “subcategorize” the output in step #4 on the basis of different chainlength classes. The Xquery implementation of the algorithm subdivides four categories of coreferential chain lengths, as shown in Table 26.

Table 26 Coreferential chain length categories

Category	Coreferential chain lengths
Zero	0
Small	1
Medium	2-5
Large	6 and higher

Coreferential chains of length “zero” occur when a participant is introduced in subject position, but there is no noun phrase in the subsequent clause or discourse that refers back to it. This is, as we will see in the experiments, the most common situation. The category of “large” is determined on the basis of the experiments described in the next sections. It appears that any text has a small number of participants that have a relatively large coreferential chain.

8.3 Subject positions

The subject is syntactically the key element in presentational focus, which is why this section reviews the different possible subject positions. The subject can appear

in the PreCore area, the Core, or in the PostCore; but how are the subject positions identified in Table 25 related to these three possibilities? Consider the following examples of subject positions:

- (186) a. **This noble knight** had in his early youth contracted a strict friendship with the only son of Lord Lovel. [reeve-1777:15]
- b. But from that time **he** heard no more from him. [reeve-1777:18]
- c. Trending away on either side of the port was **a bold rocky coast, varied here and there with shingly and sandy beaches.** [fayrer-1900:54]
- d. Nor should **a Horse** be rejected on account of a large belly. [skeavington-184x:69]
- e. Fæder her is cumen **aneunuchus of cinges hirede.** [coeuphr:142]
 father here has come a eunuch of the.king’s household
 ‘Father, a eunuch from the king’s household has arrived.’
- f. Vpon the v. day played togyder **an Henauder and a Squyre called Iohn Stewarde** whiche daye also the Englysshe man wan the worshyp. [fabyan-e1-h:180]

Subjects in the “PreCore” area can be “Initial”, as in (186a), as well as “PreV”, as in (186b); in the latter case the subject is before the finite verb (the Vb1 slot), but another constituent precedes it. Example (186c) illustrates the “VS” word order according to Table 25, but it is not clear if the subject is part of the Core area or the PostCore area, since there is no clear Core-end signal such as a non-finite verb. Example (186d), illustrating the “Mid” word order, has the Vb1 slot filled with the modal *should* and the Vb2 slot with the non-finite verb forms *be rejected*, so that it is clear the subject in-between is in the Core area. Also clear is the “PostVnonF” example in (186e): the subject is in the PostCore area, since it follows on the past participle *cumen* ‘come’, which fills the Vb2 slot. The last example (186f) illustrates the “PostVF” word order, where the subject is completely clause-final, a constituent intervenes between the finite verb *played* and the subject, but it is often not completely clear whether the subject is in the Core or the PostCore area.

In sum, it is easy to know whether a subject is in the “PreCore”, but a decision whether it is in the Core area or the PostCore area can only be taken if the Vb2 slot is filled. One way to decrease the ambiguity would be to recognize which kinds of non-Verb constituents can be regarded as “alternative” fillers of the Vb2 slot, but this is a matter of research beyond the scope of the current study.

8.4 Presentational focus with “New” subjects

The first experiment conducted on quantifying the changes in presentational focus uses approach (i) from the algorithm in (185): it checks for all clauses that have a referentially “New” subject. Section 8.4.5 will show that this is not a sufficient condition for the recognition of presentational focus, but I will use it as a first approximation before finetuning the search algorithm in section 8.5. The query that looks for new subjects is performed on the referentially enriched subset of the parsed English corpora (see Table 24). We will look at the outcome of this experiment from different angles, taking into account differences in the position

where new subjects are found and differences in the lengths of the chains that are started off by the new subjects.

8.4.1 Subject chain length differences

The clauses found by the algorithm in (185) can, regardless of the clausal position of the focused subject, be subdivided on the basis of the length of the chain started by the “New” subjects, as shown in Figure 18.

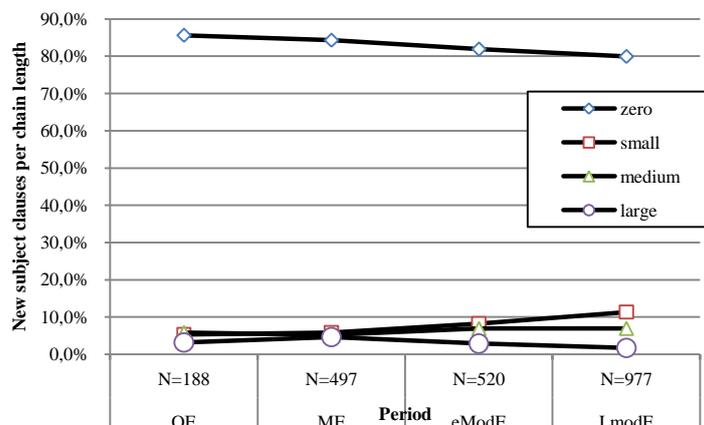


Figure 18 New subject presentational focus per chainlength category

By far the majority of clauses (80-85%) have a subject that has no chain at all: no further reference is being made to the newly introduced entity. There is a rise in the number of small-chain-subject presentational focus clauses (small chains have just two constituents), and this is mostly at the cost of a relative decrease in “New” subjects that start a relatively “large” chain. The differences between the behaviour of presentational focus depending on the length of the chains which the focused subjects start is not that huge, and this is in line with what we would have expected. Since the types of texts ought to be equivalent between the time-periods that we perform our experiments on, there should be no big changes: the same kinds of stories should roughly yield the same number of references to each individual participant, which translates in the expectation that chain length distribution stays equal.²

The small changes that we do observe may be attributed to the changing role of the subject in English: where in Old English the subject is used to keep track of a protagonist in a story and clause-initial adverbials provide cohesive linking, Present-day English uses the subject for both functions (Los, 2012). The net result is a decrease in subject elision and an increase in inanimate subjects, which combine into an increase in new subjects that have little or no chain attached to them whatsoever, which is exactly what we see in Figure 18. An example of a short-chain subject is given in (187).

- (187) a. **The Sight of their poor mangled Comrade** so enrag'd 'em, as before, that they swore to one another they would be reveng'd; [defoe-1719:260]
 b. *þa they* saw their poor mangled comraded, *þa they* got enraged, as before, and swore to one another they would be revenged. [OE alternative]

The inanimate subject *the sight of their poor mangled comrade* in (187a) is referentially “New”, which is why it was found by our algorithm. However, the subject clearly provides a link with the preceding discourse by referring to *their poor mangled comrade*. Even though we do not have an Old English equivalent of this text, if it existed, such an equivalent could well have maintained *they* as topic, while the current subject would have been expressed as a temporal point of departure, arriving at the T-correlated structure in (187b).

8.4.2 Subject position differences

Another way to look at the results of the experiments described by the algorithm in (185) is to see if there are differences in the position of the new subject over time, dividing clauses into the word order categories that are defined in Table 25. This revision results in Figure 19, which shows the developments of the presentational focus word orders irrespective of the lengths of the chains that start out from the newly introduced participant.

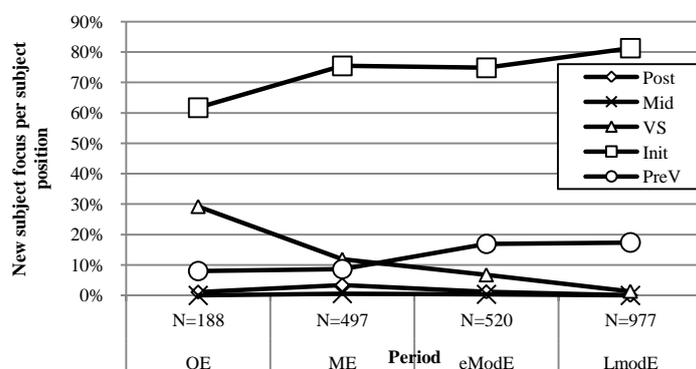


Figure 19 New subject presentational focus per clause type (see Table 25)

There is a clear and gradual increase of the “Init” and “PreV” word order types (indicative of a subject in the PreCore area) to contain presentational focus constructions, and this is at the cost of the “VS” word order; but this later order is ambiguous between Core and PostCore (see 8.3).³ These developments are what we would expect: the English word order in general changes from a kind of V2 (see 1.2.1) in main clauses to SVO (as described in chapter 4). Those instances where V2 would accept a subject *following* the finite verb (be that in the Core or in the PostCore area) decrease as the syntax becomes more SVO like. Apparently these instances include the referentially “New” subjects.

What if we were to combine chain length and position in our search for the behaviour of main clauses with referential new subjects? When we look at the data

retrieved with the corpus research described in section 8.2, and we zoom in on the referentially new subjects that start off a coreferential chain of medium to large size (see Table 26), then the division of clauses with respect to the position of the subject becomes as in Figure 20:

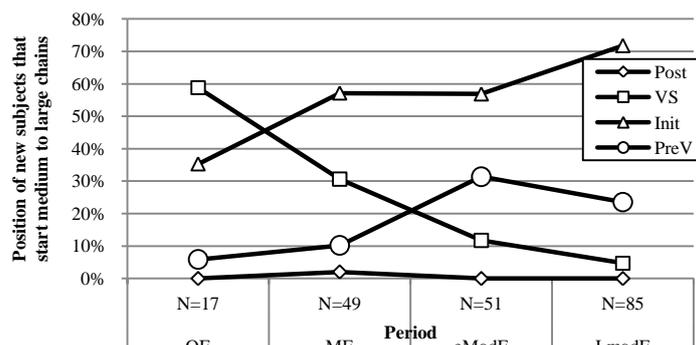


Figure 20 New subject presentational focus for medium and large subject chains

The most significant difference between the results for medium and large chain starting new subjects in Figure 20 versus all new subjects in Figure 19 is the steeper decline of the “VS” word order (where the subject position is ambiguous between Core and PostCore): its decline starts in OE with 30% for all new subjects, but with an almost 60% for those subjects that have a larger chain.⁴ Care must be taken, however, since the significance of the results has decreased: the division of positions for OE in Figure 19 is based on 188 occurrences, whereas it is based on only 17 occurrences in Figure 20. Nevertheless, it is clear that presentational focus constructions change for participants that have a longer lifespan in a text—these are the participants that may be regarded as more “protagonist” like, since narrative protagonists tend to be referred to more than once in a story. The kind of changes we see for these more pronounced participants is illustrated in (188).

- (188) a. Ða færinga com **Arcestrates**, ealre þære þeode cyningc, [coapollo:233]
 then suddenly came Arcestrates all that people’s king
 mid micelre mænio his manna
 with great company his of.men
 ‘Then suddenly came Arcestrates, king of all that people, with a great company of his men.’
- b. (Yet, nevertheless, Sir John Perrott_k wanted noe Adversarys, whatsoever he_k attempted or performed.)
 For presently, upon his_k Returne from Sea, **one Thomas Wyriott, a Justice, and a headie Man**, did preferre a Petition, with Artickles, agaynst Sir John Perrott_k unto the Queene;

The example shown in (188a) is quite a typical Old English introduction of a new participant in subject position, especially in combination with an auxiliary or an unaccusative verb. The type of the clause, according to the Old English clause types defined in Table 6, is T-initial. Such a clause is used at a moderate-sized new

development in a story. The new participant is introduced by an NP with an apposition: it not only lists the name of the person, but some additional characteristics as well. Such an NP type is a quite common one for new participants wherever they are introduced in a story. The apposition serves to link the new person to things that may be assumed to exist in one’s mind already (in this case the concepts of “king” in general and the reference to the established information “that people” specifically).

The approach taken in (188b) to convey presentational focus is different. The clause has a temporal point of departure *upon his returne from sea*, after which the new subject comes and then the finite verb *did*. The signal that the subject is new is not given through word order, but through the built-up of the subject NP: the use of the indefinite article *one* (meaning ‘a certain’) and the use of the appositive construction. This last strategy is the same for the LmodE example as for the OE example. The difference between the two may be partly due to the verb used in the sentence introducing the participant: an unaccusative verb of motion in Old English, versus a transitive verb (*prefer* ‘present’) in the LmodE example. However, an author has the option of choosing the kind of verb with which a participant is introduced; if a transitive verb is needed to convey an action taken by a new participant, two clauses may be used: one that contains a lexically light verb (such as an auxiliary) to introduce the participant, and the second that conveys the action. Such a strategy for (188b) could have resulted in: *When Perrot returned from sea, there was a certain Thomas Wyrriott, a just and stubborn man, and this man presented with a petition that included article against Perrot to the queen.*

8.4.3 Two strategies for postverbal new subjects

Since there is a clear division between clauses where the subject occurs before and those where the subject occurs after the finite verb, we need to do some additional work: we need to determine the development of postverbal subjects in English *in general*, and compare this more general trend with the development of referentially new subjects. A corpus research project that looks at the position of subjects with respect to the finite verb yields the results depicted in Figure 21.⁵

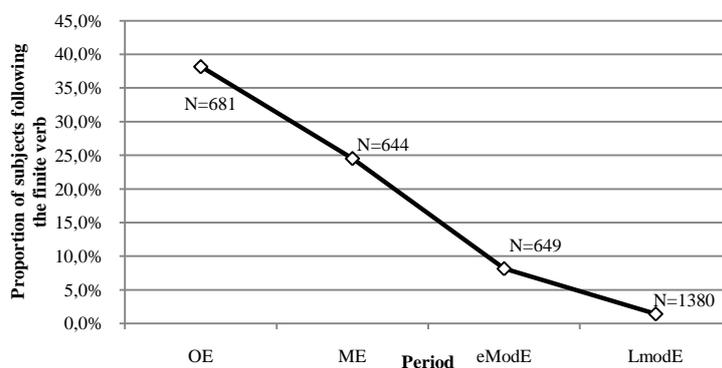


Figure 21 The decline of subjects occurring after the finite verb in main clauses

What we see is that there is a steady and almost linear decline of subjects occurring after the finite verb in the main clause from almost 40% in OE to some 2% in LmodE.⁶ This is what we would have expected, given the general tendency of English to become more of a rigidly structured SVO language, as forced by the loss of V2, described in chapter 4. Postverbal subjects are still possible, but only in well defined exceptional cases, and in the other situations there is an alternative strategy that has taken over: make use of a syntactic subject (an expletive pronoun like *there* or *it*) that is semantically empty, so that the logical subject has to receive a different syntactic status (Biber et al., 1999: 942-956, Bolinger, 1977, Hartmann, 2008, Roos, 2012). The two strategies co-occur in late Modern English, as shown in (189).

- (189) a. But **there was another Apartment** in the House where the Prince or King, or whatever he was, and several other were. [defoe-1719:373]
 b. And so **there was a battle fought**, ... [reeve-1777:76]
 c. The first Object we met with, was the Ruins of a Hut or House, or rather the Ashes for the House was consumed; and just before it, plain now to be seen by the Light of the Fire, **lay four Men and three Women kill'd**. [defoe-1719:418-419]

The expletive strategy is used in (189a), where the logical subject *another apartment* is referentially new, and occurs after the finite verb, while it syntactically is a complement in a copula clause with *there* as syntactic subject. It is not completely clear from (189a) what the position of the logical subject is in terms of the slot-structure, but example (189b) (repeated from (99c) in chapter 4) indicates that it might be the “Mid” slot (the slot immediately following the non-finite verb Vb1 slot). The strategy with a syntactic subject that is completely new is used in (189c). The additional effect of the expletive strategy is that it explicitly signals an underspecification in time or place (see sections 4.7.5.1 and 4.8). This could be one of the motivations for using an expletive in (189a): if the sentence would have been rephrased as “*In the house was another apartment where the Prince and several others were*”, then the clause-initial adverbial *in the house* would be seen as point of departure (or frame-setting), leading the reader to expect the ensuing discourse to continue to speak about other matters relevant to the location *in the house*. But this is not the case: the discourse continues with “the prince and several others” as well as “the house” as major participants.⁷

If we combine the decrease in referentially new syntactic subjects (as in Figure 19) with the decreasing occurrence of syntactic subjects in general (as shown in Figure 21), the question arises what the correlation is between the postverbal subject position and the presentational focus articulation: are postverbal subjects always referentially new, or do they always indicatethetic focus? A follow-up experiment looks into just that by (a) detecting postverbal subjects and (b) expletive subjects with a logical subject occurring postverbally.⁸ The results of that experiment are shown in Figure 22.

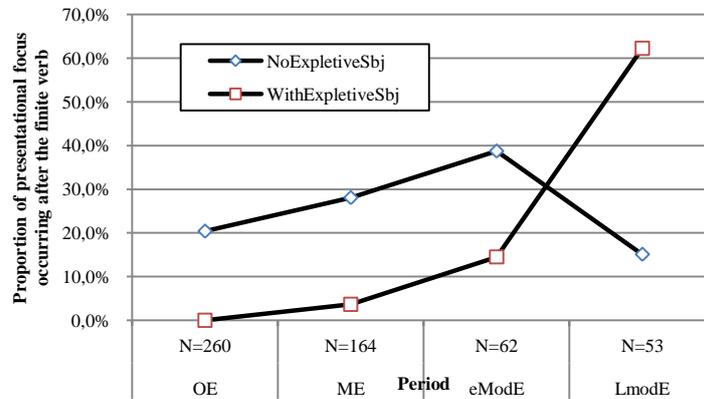


Figure 22 Postverbal presentational focus with syntactic subjects versus expletives

The baseline used in the measurements for Figure 22 (which are the numbers depicted by “N=” in the figure) consists of all the main clauses that either have a postverbal subject of any kind or have an expletive subject accompanied by a postverbally occurring “logical” subject, which syntactically comes across as a complement. The line in Figure 22 labelled “NoExpletiveSbj” gives the proportion of clauses from the baseline with a postverbal subject that start a new chain: these are the instances of presentational focus that make use of the “postverbal syntactic subject” strategy. The line labelled “WithExpletiveSbj” gives the proportion of clauses that use the “expletive subject with postverbal complement” strategy for presentational focus: they have an expletive as syntactic subject, and a logical subject occurring after the finite verb in the main clause.

What we see is that the postverbal syntactic subject strategy steadily rises from 20% in OE to around 40% in eModE, but after that it sharply falls to 15% in LmodE.⁹ This fall coincides with a sharp rise of the expletive strategy from 15% in eModE to over 60% in LmodE.¹⁰ It is by the time of LmodE that the expletive strategy, exemplified in (189a), which slowly gained momentum in the previous periods, takes over as the default strategy for presentational focus.

These observations are in line with Lambrecht (2010), who compared the “subject-focus mappings” of French and English, and found that subject focus (which is what we refer to as presentational focus) in English very often occurs in the canonical position before the finite verb, whereas French prefers other constructions, leading to new subjects to appear in the postverbal position. While Lambrecht finds French to ban syntactic subjects to be focused at all (which is why focused constituents that are logically subjects come out syntactically as complements in cleft constructions in French), it seems that the constraint on focused subjects in English is restricted to their position: focused subjects may decreasingly occur after the finite verb.

8.4.4 The other postverbal subjects

There is one matter I would like to follow up on with regards to the results shown in Figure 21 and Figure 22: what is the story behind the referentially linked subjects occurring after the finite verb? If we look again at the postverbal subjects, which, according to Figure 21, sharply decrease over time as the syntax of English changes, we can measure the proportion of postverbal subjects that link back to what the addressee has already available in his mental model; they are *not* new, but have a referential category of “Assumed”, “Inferred” or “Identity”. If we look at that proportion, we get Figure 23.

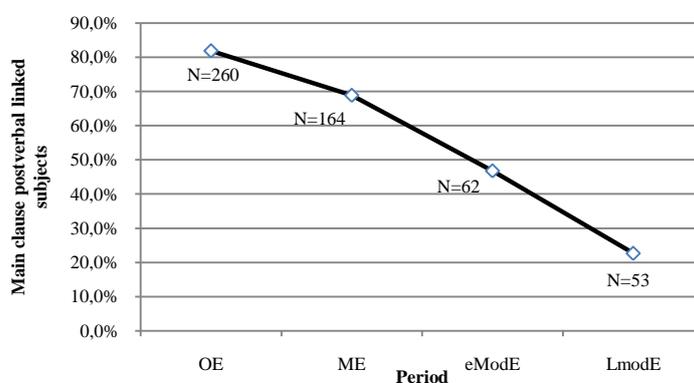


Figure 23 Main clause subjects that occur after the finite verb and that are linked

What we see is the flipside of Figure 22: a sharp decrease in postverbal subjects that are linked.¹¹ By late Modern English there still are over 20% of linked subjects occurring postverbally, and it would be good to know what kinds of subjects *do* occur postverbally, but are *not* referentially new, so that they are quite likely *not* associated with presentational focus. An inspection of the data reveals that most of them actually are instances of negation-initiated subject-auxiliary inversions: the occurrence of negators like *nor* and *neither* triggers the finite verb to occur immediately after such a negator, so that the subject, whether it is referentially linked or not, necessarily follows the finite verb.

- (190) a. nor did I ever design they should drown him. [defoe-1719:53]
 b. Nor would that do. [defoe-1719:45]
 c. and away went he. [defoe-1719:401]

Apart from the subject-auxiliary inversion types shown in (190a,b), there also is one instance of a locative inversion, as shown in (190c). As a follow-up on Figure 22, we look at the “late-subjects”, which may be regarded to be those occurring in the “PostVf” and “PostVnonF” positions (see Table 25). The number of positively recognizable late-subjects is quite low, but Table 27 shows how many of these late subjects are “linked” and “unlinked”.

Table 27 Late subjects that are linked and unlinked

	OE	ME	eModE	LmodE
Linked	11	8	5	0
Unlinked	12	11	12	0
Total	23	19	17	0

Conclusions from Table 27 are speculative due to the low numbers, but if the trend we see here reflects reality, then the OE period does not seem to distinguish late subjects according to their referential category: they can be established or unestablished. This changes gradually towards eModE, by which time the number of non-established late subjects is much higher than the established ones. By LmodE the *there* expletive construction has taken over, so that no more clearly identifiable late subjects are found.

To recapitulate the picture evolving from Figure 22 and Table 27, by the late Middle English period the situation has become as it is in Present-day English: postverbal subjects in general are only allowed in a grammaticalized versions of subject-auxiliary inversion (as triggered by *wh*-constituents and negation) and by locative inversion (Birner and Ward, 1998). Presentational focus that makes use of referentially new subjects either makes use of the expletive strategy, or puts the new subjects in the canonical preverbal position, employing other means (such as apposition) to signal the addressee that the subject is new, and to avoid leading the reader into a topic-comment articulation reading, where a presentational focus one is intended.

There is one more issue that remains to be investigated in future research, and that is the question whether there is a correlation between the kind of verbs used in new-subject clauses and the position where the new subject occurs (before or after the finite verb). Lozano and Mendikoetxea (2010) find three factors promoting the use of a VS construction for present-day English L1 and L2 speakers: (a) the verb is unaccusative, (b) the subject is heavy, and (c) the subject is focused (in the sense of being “non-presupposed, new or (relatively) unfamiliar information”). A feature indicating the type of verb has not yet been added to the parsed English corpora this study makes use of, but once that is done, the verb type can also be taken into account.

8.4.5 Preverbal new subjects

In addition to concentrating on the postverbal subjects, we should also take a closer look at the referentially new subjects occurring *before* the finite verb, which are, according to Figure 19, quite a large group already in the Old English period. An inspection of the results reveals that the preverbal new subject by the end of the late Modern English time period consist of a few different types, which are illustrated by the examples in (191). Not all of these are, as we will see, examples of thethetic or presentational focus articulation.

- (191) a. **The digestibility of food** is an important consideration in feeding, as with some kinds more is absorbed into the system than others. With scarcely any of them is digestion complete throughout. [fleming-1886:5-6]
- b. There was here a consecrated piece of ground with a temple, to which slaves used to fly when they were badly used, and the masters could not forcibly take them away. Accordingly runaway slaves stayed there, and were of course maintained by the guardians of the temple, until the masters came to reasonable terms with the slaves and confirmed the agreement by a solemn oath, which no master was ever known to have violated. **The fear of the deities of the place** secured the performance of the oath. [long-1866:36-40]
- c. The river Alba is not mentioned, I believe, by any other writer, but it is very probably the river Allava in the Antonine Itinerary. If the governor crossed this river before reaching Heraclea, it must be a stream east of Heraclea, but **some geographers** have identified the Allava with a river west of Heraclea. [long-1866:70-73]
- d. The Roman governor however revoked the promise of freedom which had been made to the slaves of Morgantia, and **many of them** went over to the insurgents. [long-1866:114-5]
- e. **What is called “cellulose”** is usually fairly well digested. [fleming-1886:18]
- f. It would appear that only a certain amount of each substance can be digested from a given quantity of food, and **rest or work** will not cause an animal to digest more, though it may happen that two animals of the same breed will digest different quantities of the same food. [fleming-1886:21-22]

Examples (191a,b) illustrate the first group, which is that of the postmodified definites. The postmodified definite subject *the digestibility of food* in (191a) is the very first line of a chapter in a text about horses, so that its referential status as “new” is quite clear. Even though it is what Prince (1981) would call a “containing inferrable”, because the head noun *digestibility* can be inferred from the noun *food* which is contained within the noun phrase as a whole, the source of the inference *food* is new too, so that the subject is new in every respect. The second example with a postmodified referentially new subject (191b) slightly differs, because here we have the postmodification *the deities of the place* infer from *the temple* in the preceding context. This “anchor” reduces the newness of the subject. This example also differs in the newness of the predicate: *secure the performance of the oath* links back to *confirm ... by a solemn oath* in the previous clause. This clause is not athetic one at all: it is an example of constituent focus. The clause gives an answer to the question “What made no one violate the oath?” The answer to that question is: the fact that people feared the deities that were worshiped at the temple. All this is to show that we have to be very careful with referentially new subjects that are anchored (which is why we exclude them in the experiments in section 8.5).

Another group of referentially new subjects is formed by quantified noun phrases, as for example *some geographers* in (191c) and *many of them* in (191d). Bailey (2009: 134) excludes such sentences from his in-depth research on the thetic articulation in ancient Greek, and I will do so too in section 8.5, because sentences

with quantified subjects come closer to having a topic-comment reading: a comment is being made about a particular subset of a larger group. This larger group consists of “geographers” in (191c) and to “them” in (191d). Both of these larger groups are anchors in the sense that they link back to already established participants—either with inference (*geographers* can be inferred from *writers*) or directly (*them* refers to *the slaves of Morgantia*).

A different group of referentially new subjects occurs in *wh*-cleft clauses, as for instance the subject *what is called “cellulose”* in (191e). The general rule for *wh*-clefts, as we will see in more detail in chapter 9, section 9.11, is that the referentially new status of the free relative subject does not count so much in determining its focus articulation as does the fact that the noun phrase complement provides the value for the open proposition generated by the free relative subject. This is why we will have to exclude *wh*-clefts from the clauses that carry presentational focus.

A final group of referentially new subjects consists of those that are part of a clause that contains a sentence negation, as in (191f). The example makes it quite clear that clauses of this type should not have been captured as having presentational focus: they have constituent focus. The subject (in this case *rest or work*) represents one (or more) categories for which the predicate (here: *cause an animal to digest more*) does *not* hold, which very strongly implies contrast between the subject and one or more alternatives (even though these alternatives may not have been stated explicitly). Constituent negation has already been introduced as a sign of constituent focus in section 3.2.2.2.

In sum, there are several situations where clauses have referentially new subjects, but there is no presentational focus: (a) anchored subjects, (b) quantified subjects, (c) *wh*-clefts and (d) clauses with sentence negation. These facts do *not* form a challenge to Bailey’s statement on the recognition ofthetic sentences in (182), or to Lambrecht’s (1994: 144) observations onthetic sentences, since both Bailey and Lambrecht carefully describe the demands imposed on presentational focus by the subject. Bailey says the subject must be “non-topical”, and if it is anchored in any way, then it links to established information, so is a likely candidate as a topic. Lambrecht says that the new subject must not be linked “either to an already established topic or to some presupposed proposition”.

8.4.6 Constituent focus versus presentational focus

The previous section ended with the observation that there are several situations where a clause has a referentially new subject, yet does not contain presentational focus. I suggest one more test to see if new subjects, even though they are not to be disregarded on any of the grounds stated above, could possibly turn out to be part of constituent focus instead of presentational focus.

Since constituent focus is extensively treated in chapter 9, I will give a preview here of the conclusion: there are two clearly recognizable constituent focus situations that we can automatically check for: (i) the presence of a focus particle or emphatic adverb in a noun phrase, and (ii) local contrast within one noun phrase. Both these features are clear diagnostics of the presence of constituent focus, in the

sense that *if* a constituent has any of these features, *then* it is very likely to have constituent focus. I am not claiming the reverse: it is not true that constituent focus is always indicated by the presence of a focus particle or by local contrast. And I hasten to add that there are other indicators of constituent focus, as will become clear in chapter 9. Nevertheless, if we find referentially new subjects within main clauses and check to see whether they contain a focus particle, an emphatic adverb or local contrast, then we will get some idea of the possibility for constituent focus to “override” presentational focus.

A slightly adapted version of the corpus research queries that have been used for the data earlier in this chapter, which makes use of the diagnostics above, looks for constituent focus on new subjects, and finds 2 possible occurrences in a total of 785 main clauses (with such a referentially new subject). These two occurrences are shown in (192).

- (192) a. Ða ða þis geban þus geset wæs, þa wæron mid gitsunge
 then when this proclamation thus set was then were with avarice
beswicene na þæt an his find ac eac swilce his frind.
 seduced NEG that one his foe butalso such his friend
 (and him æfter foran and hine geond ealle eorðan sohton ge on dunlandum
 ge on wudalandum ge on diglum stowum, ac he ne weard nahwar funden.)
 [coapollo:114-119]

‘When this proclamation was made in this way, not only his foes but also his friends were seduced by avarice, (who went after him and sought him over all the earth, as well in downlands as woodlands, and in obscure places, but he was nowhere found).’

- b. (They wish to do as the rich do: they would enjoy before they have laboured; and so kicking against the law by which **society** exists, they bring ruin on themselves and often on others.)

Thus **even the wealthiest and most fortunate of our modern societies consist** of one set of men, who have laboured for their own good and that of their country, and of another set, who will not labour, but are mean enough to live on those who have done the work. [long-1866:509-512]

The example in (192a) is from Old English, and has an instance where the postverbal new subject contains local contrast: *not only his foes but also his friends*. However, the postverbal subject is no example of constituent focus, as we see from the ensuing context: it starts the chain of a new referent (his foes and friends) who, as we learn, start looking for him (Apollonius, the main character of the narrative) and do not find him. The fact that the constituent is internally contrastive may be one of the factors that contribute to its clause-final position (its canonical position according to section 0 would have been immediately following upon the finite verb *wæron* ‘were’). The other factor driving this subject to the clause-final position is undoubtedly the fact that it is syntactically “heavy”; several researchers have observed that heavy NPs tend to shift rightward (Ross, 1967). The charting framework that has been used to interpret the OE narrative texts in section 4.4 would label the clause-final subject as a DFE, a dominant focal element, because it is in a

marked focus position (the unmarked one being the canonical position for subjects within the main clause of a T-correlated sentence). This is perhaps as much as we can conclude from this example: it is presentational focus (since it clearly introduces a new participant that becomes topical in the next clause), but the subject is slightly more marked than other instances, due to the presence of local contrast within the subject NP.

The example in (192b) came out as a possible candidate for constituent focus instead of presentational focus because, even though the subject is referentially new, the subject NP contains an emphatic adverb: *even*. The referential status of the subject, however, is not entirely new, since the NP contains an anchor in the form of *our modern societies*, which can be inferred from the generic *society* in the previous clause. There is a comparison between “society” in general in the preceding clause and one particular kind of society (the wealthiest and most fortunate one today) in the sentence we are considering. Such contrast points to contrastive focus: this is not an example of presentational focus with a completely new subject at all.

In sum, we see that the data in the referentially enriched texts of the parsed English corpora agree that main clauses with referentially new subjects can safely be regarded as instances of presentational focus. One potential counterexample, a subject with the referential status of “New”, contains a constituent inside it that links with the preceding context, so that its status is not as new as could be. This is one more reason, on top of those mentioned at the end of section 8.4.5, why we now turn to look at the behaviour of *unanchored* new constituents (see 193).

8.5 Presentational focus with unanchored “New” subjects

The previous sections have touched upon a kind of subjects that are referentially new but do not (or not always) seem to be indicative of presentational focus: unanchored new ones (Prince, 1981). The term “anchor” has been used several times now in relation to referentially new constituents, and (193) gives a more formal definition of “unanchored new” constituents, basing it on the Pentaset (chapter 5).

(193) *Unanchored new*

A constituent is “unanchored new” if it has the referential state “New” and does not contain a descendant constituent with a referential state of “Identity”, “Inferred” or “Assumed”.

The definition of “unanchored new” constituents states that a constituent should not only be referentially “New”, but it may also not contain a link to already established information by means of an anchor, where an anchor is part of a constituent that has a referential state of “Identity”, “Inferred” or “Assumed”. The existence of different types of anchors is exemplified in (194).

- (194) a. [_{NP} **his** trousers]
 b. [_{NP} the **Lord’s** voice]
 c. Jane walked into the kitchen and looked at the [_{NP} the door of the **refrigerator**].

Even though the NP *his trousers* in (194a) may have the referential state of “New”, it links to an already established participant through the possessive pronoun *his*, which, being a pronoun, will have referential state “Identity”. The NP *the Lord’s voice* in (194b) may, again, be new as a whole, but it contains an anchor in the form of *the Lord*, which has the referential state of “Assumed”. The NP *the door of the refrigerator* in (194c) can very likely be “New”, but it contains an anchor, since the *refrigerator* has a referential state of “Inferred”: it can be inferred from *kitchen*, since kitchens tend to have refrigerators.

An experiment that looks for the position of unanchored new subjects with respect to the finite verb in main clauses, and that excludes quantified subjects, results in Figure 24.¹²

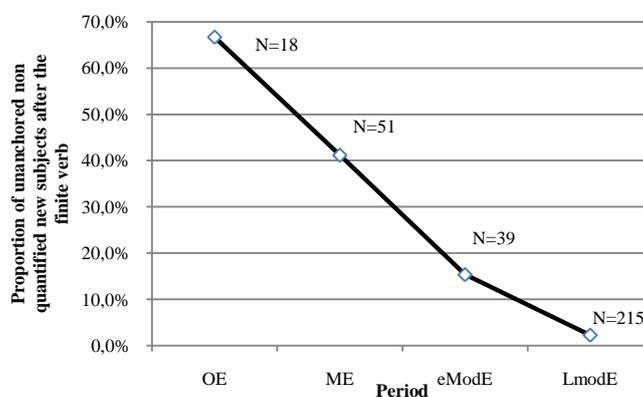


Figure 24 Unanchored non-quantified subjects occurring after the finite verb in main clauses

The number of times that subjects satisfying the strict conditions we have stated above occur is quite low in OE, ME and eModE but the trend nevertheless clearly coincides with what we have seen in earlier experiments: the postverbal position loses its ability to host syntactic subjects that are new.¹³ The only remaining exceptions in LmodE are those we have mentioned earlier: locative inversion (of which the enriched texts happen to have very few examples) and negation-motivated subject-auxiliary inversion (see the examples in 190).

8.6 Presentational focus with reintroduced subjects

There is one final, more speculative approach, that could in principle allow us to capture presentational focus: look at the behaviour of linked subjects with distant antecedents. Such subjects refer to participants who are re-introduced into a narrative, and the start of this chapter stated that re-introduction can be one form of presentational focus. We only know for sure that a participant is being reintroduced when we are aware of the scene that has been built up in the mental model of an addressee, where a particular participant has been absent for a while, and then enters that scene again (which is what happens in example (184) above). But in practice,

we may be able to capture a subset of these instances if we look out for “subject reintroduction” of a participant, as defined in (195).

(195) *Subject reintroduction*

A constituent is a subject reintroduction if it satisfies the following conditions:

- (a) it is a subject,
- (b) it has the referential state of “Identity”, and
- (c) the distance to its antecedent is larger than a definable constant.

An experiment that uses the value of “50” as the minimal distance for subject reintroduction yields very few results, as shown in Table 28, with some examples in (196).¹⁴

Table 28 Reintroduction of subjects after an absence of more than 50 clauses

	OE	ME	eModE	LmodE
Reintro PostVfSbj	2	3	0	0
Reintro PreVfSbj	5	11	14	15

- (196) a. The two most distinguished orators of this time were L. Licinius Crassus and **M. Antonius**, both of whom have often been mentioned. Crassus, who came forward as a speaker when he was a very young man vol. i., p. 320, was Quaestor probably in B. C. 109, and in Asia, where he devoted himself still further to oratorical studies under Metrodorus of Scepsis, a rhetorician of the Academy, of whom Crassus had a high opinion. [long-1866:320-321]
- b. On his return from Asia Crassus went through Macedonia to Athens, where he carefully read with Charmadas the Gorgias of Plato, in which dialogue he most admired that Plato while ridiculing orators showed himself to be the greatest of orators. He heard other philosophers and rhetoricians at Athens, and he would have stayed longer, if he had not been vexed because the Athenians would not repeat for his pleasure the mysteries, which had been celebrated two days before the arrival of Crassus at Athens. **M. Antonius** used to read Greek authors as well as Latin in his retirement at Misenum, for he had little time at Rome. [long-1866:354-7]
- c. **I** could not satisfie my self, however, without venturing on Shore once more, to try if I could learn any Thing of him or them. [defoe-1719:265]

The LmodE example in (196b) has *Marcus Antonius* as reintroduced subject: he has not been mentioned for 94 clauses. This occurrence should be seen against the background of (196a), where the author introduces two orators: *Crassus* and *Antonius*. He first speaks a number of sentences about *Crassus*, and then switches over to *Antonius*. What we have is the reintroduction of a participant in the subject position, but it is difficult to read anything but a topic-comment articulation in the text: there is no indication whatsoever that *Antonius* is treated as a completely new or surprising character at this point.

The example in (196c) is representative of a number of instances from Table 28: it illustrates a reference to the first person (either “I” or “we”) after a period of absence. It is clear that such instances are not examples of reintroduction at all: they are typical topic-comment constructions, and should be disregarded.

This is where the attempt to look for reintroduction of subjects has to stop: there simply is too little data available to make any further refinements and come up with results that have enough significance to work with. Future work on a greater number of enriched texts should attempt to look further into presentational focus resulting from the reintroduction of participants.

8.7 Discussion

In search for an answer to the research question (11) about the relation between syntax and focus, this chapter has looked at a subset of thetic focus articulation clauses. Thetic focus in general is defined as having a focus domain that covers the predicate as well as the subject, and by combining this definition with the “Focus to new” principle (see (44) in section 3.5), which says that new constituents must be part of the focus domain, we arrive at one of the hypothesis that underlies this chapter: main clauses with new subjects are indicators of presentational focus. This is an important hypothesis, because it allows verification through the texts from the parsed English corpora that have been enriched with referential information (section 8.2), which was the goal of the corpus approach we started to implement from chapter 5 onwards.

The first presentational focus experiments looked for subjects that have received the referential category of “new”, and find that the coreferential chains that start out from such subjects differ only marginally with respect to their distribution in length: the proportion of short chains increases slightly. This effect can be attributed to the changing role of the subject in English: where the clause-initial position in Old English could be used for local linking, Present-day English much more needs to use the subject for that purpose.

The position of referentially new subjects with respect to the finite verb changes over time. The most notable change is for new subjects that occur after the finite verb (which is in the PostCore slot in the model adopted in 1.2.1 and in chapter 4): their proportion changes from 30% in OE to almost zero in LmodE (and the effect is even greater for chains of participants that are referred to more than twice). This decline should be seen against the background of the overall decline for subjects (be they new or old) to occur postverbally: their proportion decreases from 38% in OE to 1,4% in LmodE. One reason for this decline is to be attributed to the change in English core structure: where OE has a [Vb1 ... Vb2] core where the subject can quite naturally appear after the finite verb (the Vb1 slot), LmodE has redefined the core as [S ... Vb1 x Vb2 O], having done away with the dedicated Core-internal position for the subject (see the discussion in chapter 4). These syntactic changes can be related to the loss of V2. When we compare the number of postverbal subjects that are new with the overall number of postverbal subjects, there is an increase from OE to eModE, followed by a sharp decrease in LmodE. This trend

combines with the dramatic rise of the expletive strategy in LmodE as in Figure 22: take a referentially inert expletive pronoun as syntactic subject, and place the “logical” subject after the finite verb. The expletive strategy has the additional effect of explicitly underspecifying the point of departure of a clause.

Further scrutiny of the data from this first experiment reveals that by LmodE the postverbal subjects that are *not* new result either from locative inversion or from negation-triggered subject-auxiliary inversion. A look at the data also reveals the nature of some of the new subjects occurring in the PreCore area: these consist of (a) postmodified noun phrases, (b) quantified noun phrases, (c) free relatives and (d) subjects under the scope of a sentence negator. Subjects in the first two of these groups often contain an anchor, which makes their status less “new”, and makes them less likely candidates for the presentational focus articulation (unless they are reintroduced). The last two groups (c) and (d) are often indicative of constituent focus.

Section 8.5 describes the results of a follow-up experiment where we distinguish subjects that are new according to a finer definition, which excludes anchored and quantified constituents. The finer definition leads to a considerable reduction of the overall results we get, but what surfaces is an even more pronounced decline of the proportion of postverbal new subjects than we saw in the earlier experiments. Section 8.6 describes an attempt to locate instances of presentational focus where the subject is not completely new, but it is a reintroduction of a participant after it has been away for a number of clauses. The results we see here are so few, that they are too insignificant to lead to any conclusions: a larger amount of referentially enriched texts is needed to follow up this line of research.

Returning to the research question in (11) about the relation between syntax and focus, a line of cause-and-effect surfaces: the loss of V2 leads to an increasing pressure on the subject to occur in the PreCore area, a reduction of subject-auxiliary inversion to grammaticalized contexts (*wh*-questions and negation), and a reduction of late-subject construction to locative inversions. The hypothesis in (183), however, which states that presentational focus in English attempts to retain its clause-final position is borne out: when the *there* expletive subject pronoun appears in English, the use of the expletive strategy takes over as a presentational focus strategy, and the effect of this is that the logical subject continues to appear as late in the clause as possible.

¹ The actual Xquery code of the queries used in this corpus research project is provided in appendix 14.2.2.

² D[corp] is 100%. Fisher’s exact test indicates that the changes from OE to ME, to eModE and to LmodE are all insignificant ($p > 0,05$). Even the change from OE to LmodE is not significant ($p = 0,0693$). More data would be needed to see if these trends remain insignificant. See for details the appendix, section 14.3.4.

³ D[corp] is 100%. The change in “Init” from OE to ME and from eModE to LmodE are significant, but the change from ME to eModE is not according to the two-sided Fisher’s exact test ($p < 0,05$). For the “PreV” line only the change from ME to eModE is significant ($p < 0,05$). As for the “VS” line: all the changes are significant according to the two-sided Fisher’s exact test ($p < 0,05$). See for details the appendix, section 14.3.5.

⁴ D[corp] is 100%. The two-sided Fisher’s exact test indicates that for the “VS” line the changes from OE to ME and from ME to eModE are significant ($p < 0,05$). The change from eModE to LmodE is not significant anymore, nor can any of the changes on the “Init” line be regarded as significant according to this test. See for details the appendix, section 14.3.6.

⁵ The corpus research project “SbjPosition” uses a query “matSbjPos”, and this query finds all main clauses (those that are not appositive and that are not the second part of a main clause in one sentence) with a finite verb and an overt subject. It subcategorizes on the different possible positions of the subject, and the picture in Figure 21 shows the proportion of clauses where the subject occurs in one of the positions after the finite verb. A differentiation into the subject positions identified in Table 25 is this:

		OE	ME	eModE	LmodE
PostVnonf	Vf_V_Sbj	1%	1%	1%	0%
PostVf	Vf_Sbj	2%	2%	1%	0%
Mid	Vf_Sbj_V	4%	10%	3%	1%
VS	VfSbj	31%	12%	2%	1%
Initial	Sbj_Vf	41%	49%	48%	67%
PreV	Y_Sbj_Vf	21%	27%	43%	32%
	N=	681	644	649	1380

⁶ D[corp] is 100%. The transitions from OE to ME, from ME to eModE and from eModE to LmodE are all highly significant according to the two-sided Fisher’s exact test ($p < 0,01$). See for details the appendix, section 14.3.7.

⁷ The context of the story, including the preceding and following line, is this:

[372] In short, most of the Indians who were in the open Part of the House, were killed or hurt with the Grenado, except two or three more who press'd to the Door, which the Boatswain and two more kept with their Bayonets in the Muzzles of their Pieces, and dispatch'd all who came that Way. [373] But there was another Apartment in the House where the Prince or King, or whatever he was, and several other were, [374] and these they kept in till the House, which was by this time all of a light Flame, fell in upon them, and they were smother'd or burnt together.

⁸ The code for the query looking for expletive sentences is provided in appendix 14.2.2.

⁹ The non-expletive postverbal strategy has a D[corp] of 100%. Fisher’s two-tailed exact test shows that, in fact, *none* of the changes from OE to ME, from ME to eModE and from eModE to LmodE are significant ($p < 0,05$). More data is needed to get a clearer picture of what happens to the use of postverbal subjects for presentational focus.

¹⁰ The expletive strategy has a D[corp] of 56% (since it is absent in OE, for instance), but all of the transitions from OE to ME, from ME to eModE and from eModE to LmodE *are*

significant according to Fisher's two-tailed exact test ($p < 0,05$). See for details the appendix, section 14.3.8.

¹¹ D[corp] is 100%, and Fisher's two-tailed exact test indicates that all between-period transitions are significant ($p < 0,05$). See for details the appendix, section 14.3.9.

¹² It would have been nice to look for subjects occurring after the Vb2 slot, the non-finite verbs, but numbers are really too limited. The code for the query looking for sentences with unanchored new subjects in presentational focus is provided in appendix 14.2.2.

¹³ D[corp] is 100%, and Fisher's two-tailed exact test indicates that the transition from OE to ME is not significant ($p < 0,05$), but the transitions from ME to eModE is, and so is the one from eModE to LmodE. See for details the appendix, section 14.3.10.

¹⁴ None of the transitions between periods are significant according to Fisher's two-tailed test ($p < 0,05$). See for details the appendix, section 14.3.11.

After chapters 5-7 laid the foundations for a corpus based investigation of the major research question in (11), which asks what we can learn about the relation between syntax and focus, the previous chapter looked at the presentational focus articulation, mainly in its use to introduce a new participant into a narrative. This focus articulation reveals a strong push to comply with the Principle of Natural Information Flow (more established information precedes less established information), so that the new participant is typically found in thethetic focus articulation as a referentially new subject that occurs clause-finally—that is: in the PostCore. The change from the late-subject construction to the *there* expletive construction for presentational focus is attributable to a change in English syntax: the increasing demand for the subject to occur before the finite verb, which is a consequence of the loss of V2.

We will see in this and subsequent chapters that different aspects of the loss of V2 are the driving forces behind changes in the constituent-focus articulation (see 3.2.2). The aim of this focus articulation is to single out one constituent as the highlighted or focused one, while the remainder of the clause is then to be understood as backgrounded and often presupposed. What we want to know in light of the research question in (11) is: (a) in what way has constituent focus changed over time, and (b) what does this tell us about the interaction between syntax and focus? We are going to look for an answer to this question in light of the statement (60) from section 4.4, which describes the impact of the changing syntax on the expression of constituent focus: a decrease in subject-auxiliary inversion jeopardizes the possibility to use the PreCore area for constituent focus. This is why I posit the following hypothesis:

(197) *Constituent focus hypothesis*

The position for constituent focus in written English shifts from the PreCore area to the PostCore area as a result of the loss of V2.

In section 4.2.3 of chapter 4, we have seen that subject-auxiliary inversion provided the PreCore area as a locus where constituent focus took place in English. What I argue is that there are two principles behind the choice of the PreCore area as the locus for constituent focus, and that, while the locus of constituent focus changes, these two principles are retained in English.

(198) *Constituent focus demarcation principle*

The focused constituent preferably occurs in an area of the sentence where it has a clear left and right boundary.

- (199) *Constituent focus placement principle*
 The focused constituent preferably occurs where it violates the Principle of Natural Information Flow.

The “Constituent focus demarcation principle” in (198) explains that the PreCore area in OE is a logical option for constituent focus, since it has a clearly defined left boundary (the start of the clause) and right boundary (the finite verb). The “Constituent focus placement principle” in (199) can be met by the PreCore area in OE too, since constituent focus often involves relatively less established constituents, and clause-initial placement yields a violation of the Principle of Natural Information Flow. Such placement can be a clear signal for focus, but care has to be taken. The constituent to be highlighted may, in principle, have almost any referential *state* (as defined in chapter 4): it may have the information state category “new”, “assumed”, “inferred” or “identity” (see the observations made by Krifka, 2007: 29, Lambrecht, 1994: 209). This is why the interaction between the form by which constituent focus is expressed on the one hand, and the Principle of Natural Information Flow on the other, may differ in written communication. Constituent-focus on a referentially new object can be signalled by placing it clause-initially, as done in OE, where it violates the Principle of Natural Information Flow. But if the constituent is referentially linked, there may not be a violation of the Principle of Natural Information Flow, so that the position is not a signal.

What are the options for retaining the demarcation principle (198) and the placement principle (199) when English syntax changes? The loss of V2 means an increasing occurrence of the subject before the finite verb, leading to a loss in a clearly demarcated PreCore area, and an increasing occurrence *after* the finite verb of non-subject constituents—including focused constituents. But this violates the placement principle (199). We will see in this and subsequent chapters that there are constructions even after the loss of V2 that retain the principles in (198) and (199), and that these constructions are increasingly used for constituent focus.

While I argue for constituent focus to change in accordance with the syntactic changes, as described hypothesis in (197), the approach in this chapter is to keep all options open. This is one of the reasons why we are mainly going to look for constituent focus by locating examples that express constituent focus as indicated by other features than word order. A second reason is that a typological study on focus in the languages of Europe by Miller (2006) concludes that (constituent) focus may associate with a particular position in the clause, but that word order is very unlikely to be the *only* distinguishing feature of such focus. In sum, we will look for constituent focus marking features, and then see what constructions or word orders correlate with constituent focus in different time-periods.

As for the kinds of constituent focus indicators, there is one obvious indicator that needs to be mentioned, and that is intonation. Intonation allows singling out a constituent or part of a constituent in an acoustic (tonal) way, which signals unequivocally to the hearer that this constituent should be regarded as having special emphasis (Gussenhoven, 2007, Halliday, 1967). Depending on the particular tonal contour, but also depending on the preceding (or following) context the hearer can

figure out what the heightened prominence signals (the focus may signal contrast or correction, for instance). But intonation as such is not always a good diagnostic for constituent focus, since some languages use intonation to demarcate the right edge of the focus domain in the topic-comment articulation, and other languages apparently do not use intonation to express focus at all (Kügler and Skopeteas, 2006). The constituent at the right edge receives an intonational peak, but is not necessarily to be understood as having *constituent* focus.

There are other indicators of constituent focus, which are not dependant on intonation, and not on the position within the clause. Constituent focus in English tends to be accompanied indicators having the following characteristics:

(200) *Constituent focus indicator characteristics*

- a. An open proposition for which the constituent to be focused provides the value.
- b. An explicit indication of contrast.
- c. An explicit indication of emphatic prominence.

Obvious indicators that make use of the characteristic in (200a) are the different cleft constructions (9.11) and answers to *wh* questions (9.10), since all of these contain variable-creating mechanisms.

Table 29 Possible constituent focus diagnostics

Diagnostic	Description	Treated in
Adverbs, particles	Adverbs like “only” and “indeed”, where they are part of an NP or a PP	9.2
Negation	The negation of one NP or PP constituent implies contrast with another one	9.3
Positive negation	Positive negation is a kind of special emphasis on the NP or PP involved	9.4
Local contrast	NPs (or PPs) of the type “not ... but ...” contain explicit contrast that is confined to one constituent	9.5
Emphatic pronouns	Reflexive pronouns, in combination with the normal set of pronouns	9.6
Apposition	Mentioning of different characteristics of a participant by apposition	9.7
Split constituents	Constituents that belong together, but occur in different positions in the sentence (including extraposed relative clauses)	9.8
Left dislocation	The position of the resumptive NP or PP within contrastive left dislocation	9.9
<i>wh</i> answers	Answers to constituent <i>wh</i> questions like “what” and “who”	9.10
Cleft constructions	Three types: <i>wh</i> -clefts, reversed <i>wh</i> -clefts and <i>it</i> -clefts	9.11

Explicit contrast, the characteristic mentioned in (200b), is associated with particular focus adverbs like “only” (9.2), negation (9.3), local contrast (9.5) and contrastive left dislocation (9.9). The last characteristic is the explicit indication of emphatic prominence (200c), and we can expect to find this with certain adverbs (9.2), with

emphatic pronouns (9.6), and with positive negation constructions like “not without” (9.4). There are two more potential indicators that we will look at, since these are features that we have come across in the treatment of the two narratives in chapter 4: apposition (9.7) and split constituents (9.8). The diagnostics we will be reviewing are listed in Table 29.

Some of the diagnostics we are reviewing here have been mentioned in section 3.2.2 of chapter 3, where the different focus articulations were introduced, but some appear here for the first time. As we review the diagnostics, we will find that not all of them are a valid diagnostic of constituent focus at all, some can be used, but only with additional stipulations, and some are unequivocal diagnostics.

9.1 Looking for constituent focus in the main clause

The experiments later on in this chapter detect noun phrases (and sometimes prepositional phrases) that comply with the diagnostics in Table 29. The experiments then determine what the *position* is of that constituent with respect to several major landmarks of the main clause (subclauses are *not* taken into account): the beginning of the clause, the end, the position of the finite verb, and, if present, the position of the non-finite verb (e.g. a past participle as *seen* in *he had never seen anything like it*). Where we find only few results, we will only consider the rough division of “preverbal” versus “postverbal”, which tell us whether a focused constituent precedes the finite verb or follows it. Where enough information is available, we distinguish the five positions in Table 30, where “XP” denotes the NP or PP whose position we determine, the bracketed constituents are optional, and the other constituents are obligatory.¹

Table 30 Word order categories for main clause constituents

Category	Word order
Initial	XP (y) V _{finite}
PreVf	YP XP (y) V _{finite}
ImmPostVf	(x) V _{finite} XP
Mid	(x) V _{finite} (y) XP (z) V _{non-finite}
PostVnonf	(x) V _{finite} (y) V _{non-finite} (z) XP
PostVf	(x) V _{finite} (y) YP (z) XP

The five positions above are mutually exclusive, and help us make the kind of word order differentiations that are interesting for Old English, which, like West-Germanic languages, generally divides clauses in a Prefield (everything that precedes the “Vb1” slot, which normally hosts the finite verb; see the slot-division in (65), section 4.6.2), Middlefield (whatever is between the “Vb1” slot and the “Vb2” slot; this last slot normally hosts the non-finite verb) and the Postfield (all that follows the “Vb2” slot). The “PreVf” position above corresponds to the Prefield (of which “Initial” is the clause-initial part), the “Mid” to the Middlefield, and the “PostVnonf” to the Postfield. The remaining two positions above, the “ImmPostVf”

and “PostVf” ones, do not allow a direct link with the Middlefield or the Postfield, since they lack a clear indication of the right border of the Middlefield.

9.2 Adverbs as diagnostics for constituent focus

Adverbs can be used in many different positions, and some of the current research explores their influence on information structure when they occur sentence-initially (Los and Dreschler, 2012, Virtanen, 2004). There is a limited set of adverbs or particles that can be used to indicate that one particular noun phrase or prepositional phrase is focused, which is why we explore them as constituent focus indicators in this section. Consider the examples in (201) that serve to illustrate this diagnostic.

- (201) a. But there is rich compensation in Barbara Jefford's magnificent Volumnia: why has this superb actress been given **only two roles** by the RSC in 30 years? [BNC, A8S:23]
- b. Having described the job the next step is to identify what kind of person will fit it. Sometimes called a candidate specification, it states the essential attributes that you require and also **the merely desirable ones**. [BNC, AYJ:108]

The word *only* can function as a focus adverb (or particle) that is positioned within an NP or PP, and that modifies it.² If it does then there is a very high probability that we are dealing with constituent focus. The constituent *only two roles* in (201a) is part of a *why* sentence, and the fact that this sentence (rhetorically) asks a *why* question about an event implies that the reader should already be familiar with the event as such. Assuming, then, that “Volumnia has been given two roles by the RSC in 30 years” is familiar to the addressee, the highlighting of the NP *two roles* by the addition of the focus adverb *only* is indeed an indication of constituent focus.

The word *merely* can, in the same way as *only*, also be an indication of (contrastive) constituent focus, provided it is part of one NP or PP. This is indeed the case in the example (201b), where *merely* is part of *the merely desirable ones*. There is explicit contrast between *the essential attributes* and *the merely desirable ones*, so that we can be sure of the fact that we are dealing with constituent focus here. This is another example of a combination of diagnostics that are being used to mark a particular type of focus: (a) explicit contrast between two NP constituents, (b) the use of a particle (the adverb *merely*), and (c) the end focus position (the Principle of Natural Information Flow would have the newest information, which usually is the most informative and most important bit, last, and here we have a reversal of this principle: the most important information is last, even though it is not necessarily the newest bit of information).

9.2.1 Adverbs for focus and emphasis

The way we can do quantitative corpus research with adverbs as diagnostics for constituent focus is the following. We use the Cesax program to add an “adverb type” feature to each of the adverbs that is being used to modify an NP or a PP. We divide the adverbs in such a way that there is one category with all the adverbs used to signal *contrastive* focus. We then make a corpus research project within the

program CorpusStudio where we locate all NPs and PPs modified by an adverb of the category *contrastive*, and note the position of these constituents as defined in Table 30.

There are two types of adverbs that are important for recognizing constituent focus. The first type of adverb signals contrast, and the second type emphatic prominence. Table 31 shows which adverbs, including most of the spelling variants, that have been assigned to these two categories, divided over the four major time periods.

Table 31 Adverbs for focus and emphasis found in the parsed English corpora

Period	Focus adverbs	Emphasis adverbs
OE	ana, elles	
ME	but, only	euer, rygt, riht, singularly, specially
eModE	alone, alonly, aloane, but, eune, yet, only, onelye, onlye, oonly, onely, onelie, onlie, oonelie, meerely, meerly, merely, singular, solely	chiefly, chieflie, chiefe, chiefly, cheifely, clene, clean, deep, diametrally, directly, dyrectlie, directlie, directly, especially, espetially, especyall, espeshallie, especiallie, espetially, especially, especiallye, esspetiallye, even, euyn, eene, e'ne, ee'ne, eev'n, euen, euin, een, evyn, evyne, e'en, ever, exactly, excellent, ful, full, flat, faste, imediately, iust, iuste, just, juste, fast, marveilous, meanly, oft, particularly, particularly, perfectly, quite, right, ryghte, ryght, righte, rygth, shortly, sodainely, sound, soone, specialli, speciallie, specyallye, speciallye, specyally, spetiall, strait, streighte, streight, strayght, straight, straitte, streght, straught, streyght, utterly, vast, very, verie, verye, welle,
LmodE		altogether, directly, exactly, quite, strait

The division of the adverbs above into the category “Focus adverb” and “Emphasis adverb” is based on the context in which these occur. A computer-generated example illustrating the use of each of the adverbs can be found at the author’s website: <http://erwinkomen.ruhosting.nl/phd/adverbs.htm>.

9.2.2 Determining the position of constituents with a focus adverb

The algorithm that locates the noun phrases and prepositional phrases modified by an adverb for focus or emphasis is implemented in the program CorpusStudio in the Xquery language. The steps that the algorithm takes are in (202).

(202) *Algorithm to get adverb related constituent focus*

Step 1: Consider each Adverb in the text, if it fulfils the conditions:

Condition a: the adverb type is “Contrastive” or “Emphatic”

Condition b: the adverb is at the top-level of an NP or PP

Condition c: this NP or PP is part of a clause (an IP)

Condition d: this clause has an overt subject and a finite verb

Step 2: Let *cat* be the position of the NP or PP within the IP (as per Table 30)

Step 3: Output:

Subcategorize on *cat*

The algorithm starts by selecting adverbs that, per condition (202.1.a), have been labelled with an adverb type of “Contrastive” or “Emphatic”. The adverb that meets this condition is then further scrutinized by checking if it occurs at the top level of an NP or PP (condition 202.1.b), if this NP or PP is part of a finite clause (an IP; condition 202.1.c), and if this finite clause has an overt subject and a finite verb (condition 202.1.d). We are only interested in the latter type of clauses, since these provide clear landmarks (clause start, clause end, finite verb) that help us determine the position of the adverb-marked constituent. Step 2 determines the position of the constituent we found within the clause, and we use this position to subcategorize the results in the output.

9.2.3 Results for the position of constituents with a focus adverb

When the algorithm above, encoded in Xquery, is run on the parsed English corpora that have been enriched with adverb type information, we get a detailed overview of the position of NPs (including subjects) and PPs modified by a contrastive or emphatic adverb. In order to get a more general overview of what is going on, we will first have a look at the general position of the constituent with respect to the finite verb (preverbal versus postverbal), as in Figure 25 (a more detailed look into the slot-positions defined in Table 30 follows).³

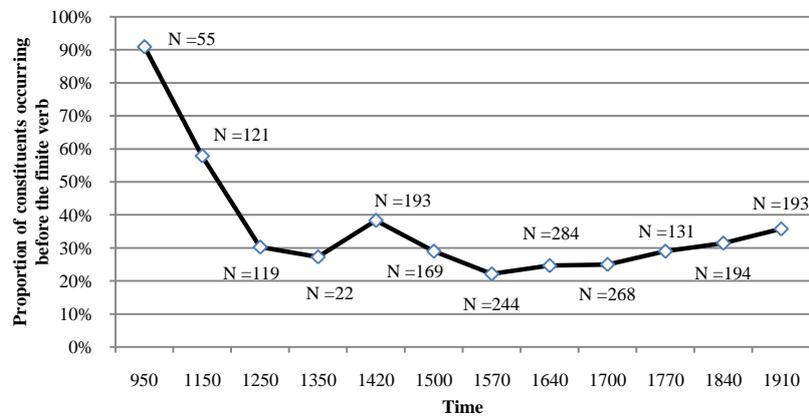


Figure 25 The proportion of NPs and PPs modified by a focus adverb occurring before the finite verb in main clauses

The general picture conveyed by Figure 25 gives us insight into the changes going on in English, although the number of occurrences (the *N* in the figure) is rather low for the OE period (the years marked as 950 and 1150).⁴ We see that OE starts with adverb-marked constituent focus occurring overwhelmingly before the finite verb, but that this picture has radically changed by the start of ME (the year marked 1250). It is by that time that the majority of adverb-marked constituent focus occurs *after* the finite verb. The trend to express constituent focus only after the finite verb continues until the year 1500, the end of the ME period. This period is followed by a

steady increase in the proportion of constituent focus from 15% at the start of eModE to 35% at the end of LmodE.

The corpus research project that collects the data along the lines of the algorithm in (202) subcategorizes on the position of the focused constituent with respect to the clause start, clause end and the finite verb. More details about the behaviour of the adverb-marked constituent focus in terms of position can, therefore, be collected if we make a finer distinction in terms of position. Table 32 shows the more detailed division in position, at the cost of less detail in the time period (see Table 30 for the proper definition of the word order categories used here).

Table 32 Positional distribution of adverb-marked constituent focus

				OE	ME	eModE	LmodE
PostVnonf	Vf	V	X	5%	13%	18%	18%
PostVf	Vf	X		18%	29%	25%	28%
Mid	Vf	X	V	1%	0%	0%	0%
ImmPostVf	Vf	X		9%	25%	33%	21%
Initial	X	Vf		59%	25%	17%	22%
PreVf	Y	X	Vf	10%	8%	7%	10%
			N	176	503	796	518

What we learn from Table 32 is that the preferred adverb-marked constituent focus position in OE really is the clause-initial one—the other positions only marginally contribute. The next thing we see is that from ME onwards there is a consistent preference in positions: (a) first the immediately postverbal or the clause-final position, and then (b) the completely clause-initial positions. Some examples from these periods should help us understand more clearly what is going on.

- (203) a. (Witodlice þa þa se halga wer Benedictus eallunga forlet to leornienne þa boccræftas, þa geteohhode he to secenne westenstowa.)
 & **his fostormodor ana** him fyligde,
 and his nurse only him followed
 forþam þe heo hine swiðe geornlice lufode. [cogregdh: 989-990]
 because that she him quite tenderly loved
 ‘(Truly when the holy man Benedict left everything to acquire learning, he prepared himself to seek a lonely place,) and only his nurse followed him because she loved him quite tenderly’
- b. ‘(When Benedict abandoned his studies to go into solitude,) he was accompanied **only by his nurse**, who loved him dearly.’ (Zimmerman and Avery, 1980)

The subject *his fostormodor ana* ‘only his nurse’ in example (203a) occurs in the clause-initial position (if we skip over the conjunction *and*). This contrasts with the more recent Present-day English translation provided in (203b), where a passive is employed, so that the agent of the main verb (*fyligan* ‘follow’ in OE, and ‘accompany’ in PDE) occurs in a clause-final position. The clause-final position is not only the place where we can expect DFEs (dominant focal elements; see section

3.3.3), but in this current situation it also provides for a more natural connection with the information about “his nurse” neatly stored in a relative clause, where it is readily interpreted as backgrounded material. We now leave OE and take a look at two examples from the ME period in (204).

- (204) a. And sir Lyonell waked whyles he slepte. ... and in the meanewhylye
 And sir Lyonell waked while he slept ... and in the meantime
 com there three knyghtes rydyng, ... and there followed hem
 came there three knyghts riding ... and there followed them
 three **but one knyght**. And when sir Lyonell hym sawe, he thought
 three only one knight and when sir Lyonell him saw, he thought
 he sawe never so grete a knyght ... [cmmalory: 2430-2434]
 he saw never so great a knight
*‘Sir Lionell kept watch while he slept. ... In the meanwhile they were
 approached by three knights on horseback... These three were followed by
 only one knight. When sir Lionell saw him, he thought that he had never
 seen such a great knight...’*
- b. (The hond of God is myghty in confessioun, for therby God foryeveth thee
 thy synnes,)
 for **he allone** hath the power. [cmctpars:1530-1532]
 for he alone has the power
*‘The hand of God is mighty in confession, because that is the means
 through which God forgives your sins, since only He has that power.’*

We have seen from Figure 25 that the preference for adverb-marked constituent focus is strongest by the end of the ME period, which is around 1500. Example (204a) is from this period, illustrating the point through the clause-final positioning of *but one knyght* ‘only one knight’. The ME strategy is to put the subject clause-finally, and it does so by using an expletive subject *there*.

There is a minority of instances where ME texts have adverb-marked constituent focus on the first constituent in a clause, and (204b) is one example of these instances. The reason why the clause with constituent focus is clause-initial may have nothing to do with the fact that it is *focused*, but more with the fact that it is the grammatical *subject*. The English language has an increasing tendency throughout for subjects to occur before the finite verb (in the PreCore area), irrespective of whether they are focused or not, and the ME period already sees an increased pressure in having subjects precede the finite verb. Example (204a) has a syntactic subject, the expletive *there*, precede the finite verb. The pressure for a subject to occur before the finite verb can be illustrated by Figure 26, which shows the percentage of subjects occurring *after* a finite form of the verb “have” (the lexical verb “have” as in (204b) is preferred by excluding clauses with participles, such as ‘he has seen her’).⁵

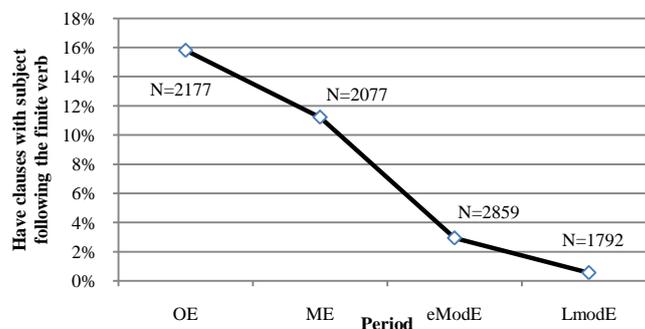


Figure 26 Percentage of “have” clauses with postverbal subject

What Figure 26 shows, then, is an example of the tendency for one lexical verb (the verb “have” in its simple transitive sense) to increasingly disallow postverbal subjects, irrespective of their pragmatic status. If we extrapolate the picture we get by looking at the behaviour of “have”, we realize that the preverbal position in English increasingly loses its power to signal that a constituent is focused; the alternatives are to add a focus adverb to a clause-initial constituent, or to position a focused constituent clause-finally. These alternatives already become visible in ME, but are most clearly visible in LmodE, witness the examples in (205).

- (205) a. ... they were also allowed to distribute private charity – for **the French only** understand or understood then the combination of public & private charity –, [nightingale-189x:307]
 b. Maize, beans, or peas, with bran and cut hay, formed the basis of the usual food allowance. The oats and linseed were used **only for sick or delicate-feeding horses**. [Fleming-1886:373-374]

The Modern English period gradually sees a trend where the proportion of adverb-marked constituent focus increases to occur before the finite verb, as in (205a), where *the French only* should be understood as one noun phrase. Despite this increase, the proportion of adverb-marked constituent focus occurring *after* the finite verb (as in the ME example) remains the majority, as illustrated by (205b).

To summarize what we learn from adverb-marked constituent focus: OE uses the first constituent for such focus, ME mainly uses the clause-final one, and LmodE still uses the position after the finite verb in the majority of cases.

9.3 Negation as diagnostic for constituent focus

The negation of one NP or PP can be seen as a form of explicit contrast, as we have seen in section 3.2.2.2, and contrast is one of the constituent focus indicators mentioned in (200). The reason why negation relates to contrast is that if we negate, and therefore exclude, one entity, we presuppose that there must be another entity for which the proposition that is being evaluated *does* hold (see Los, 2012 for the relation between negation and focus in the clause-initial position). Consider the examples in (206).

- (206) a. And many more believed because of his word; and they said to the woman, Now we believe, **not because of thy speaking**: for we have heard for ourselves, and know that this is indeed the Saviour of the world.
[erv-new-1881:302-5]
- b. Redemption was **not a cheap process**. It wasn't just something you did lightly, you had to weigh it all up and consider the cost of it.
[BNC KNA:106]
- c. The Sight of their poor mangled Comrade so enrag'd 'em, as before, that they swore to one another they would be reveng'd, and that **not an Indian who came into their Hands** should have Quarter. [defoe-1719:360]

The negated constituent in example (206a) stipulates the reason, *because of thy speaking*, which is not the basis for the fact that a group of people started believing that Jesus is the Saviour of the world. The fact that the group *started* believing in Jesus is stated explicitly as *Now we believe*, and there must have been a reason for their belief. The current context gives this reason explicitly as *we have heard for ourselves*. But even if the “positive” reason would not have been given, there would still have been the presupposition that such a positive reason exists. The second example (206b) illustrates the same point. The negation of *a cheap process* signals the existence of the opposite: that redemption is an expensive process. So the occurrence of a negator inside the NP signals contrast, and explicit contrast between constituents is one kind of constituent focus.

When *subject* noun phrases are negated, the diagnostic may fail to indicate constituent focus, but this depends on the referential status of the subject. If we have a referentially new subject, such as the subject *not an Indian* in example (206c), then this is already a clear indication ofthetic focus—that the focus domain spans a whole clause. And this is indeed what we find for (206c).

The last examples shows that we are able to use negation of NPs or PPs as an indicator of constituent focus in a quantitative research only to a limited extent: negation of a constituent by itself is not sufficient for recognition of the focus articulation. To establish the focus articulation we may also need to know (a) the grammatical role that the negated constituent fulfils, and (b) the referential status of the constituent. This is how the recognition process goes: if the negated constituent is *not* a subject, then it is very likely that we have constituent focus, but if the constituent *is* a subject, then we need to consider its referential status. If it is referentially “new”, then we probably have athetic focus articulation, but if it is not, then the picture becomes much more complicated: we need additional information about the syntax of the clause and the referential statuses of its constituents to determine the focus articulation, and it is not clear at all whether this can be done automatically. Of the two pieces of information we need, the grammatical role of the negated constituent and its referential status, the latter one is only known in the coreferentially enriched part of the parsed corpora, so that the value of negated constituents as a diagnostic for automatic constituent focus recognition is limited. There are two different approaches, then, that we could take: (a) look at the placement of negated PPs and negated non-subject NPs in all the syntactically

parsed corpora, or (b) restrict the search to the referentially enriched texts, so that we can also include those *subject* NPs that are not referentially new. Both approaches have drawbacks: the former compels us to look at a subset of all the data, inherently skewing the results, and the latter is very likely to include too little data to be of any real significance. It is for these reasons that none of the two approaches have actually been implemented as part of this current study.

9.4 Positive negation as diagnostic for constituent focus

Positive negation is the positive meaning of an NP or PP resulting from a combination of a grammatical negator and a word within the NP or PP that lexically has a negative meaning. This kind of negation is quite often a means of emphatically highlighting, as we have seen in section 3.2.2.3, and emphatic prominence is one form of constituent focus according to (200). Consider the examples in (207).

- (207) a. All that will, of course, now change, with the government's decision to allow the supermarket giants in. But the move is **not without opposition**. (BBC, 2011)
- b. The performance was **not without mishap**. He did lose the lines on more than one occasion and thrashed around helplessly through pauses that seemed eternal, until the A.S.M.'s quiet voice in his ear managed to get him back on to the right track. [BNC H92:1898]

Example (207a) shows how positive negation within the PP *with opposition* results in *not without opposition*, which is a double negation that can be understood as emphatically saying *with a lot of opposition*.⁶ Example (207b) similarly emphasizes that there was *a lot of mishap* in the performance, a fact that is further substantiated by the next sentence, which states some of the things that went wrong during the performance.

Finding situations of positive negation through an automated corpus research is difficult, since one part of the positive negations, the noun with the inherently negated meaning, is determined lexically, and it may not be possible to recognize such nouns in the texts from all the different English time periods automatically. An NP or PP constituent with positive negation consists of two crucial elements: (a) at most one overt negator (this is the negator *not* in Present-day English), and (b) another word in the NP or PP that inherently contains negation. The examples above have the word *without*, but NPs with positive negation can have words like *unnecessary*, *unintended* in combinations like *not an unnecessary precaution* and *not an unintended consequence*. Words like *unintended* are not explicitly marked with a “negative” feature in the parsed corpora, which makes them more difficult to recognize, especially since there are other words starting with *un* that do not have a negative feature, such as: *unification*, *unity* etc. In sum, positive negation is a valid diagnostic for highlighting, but it is not investigated in this dissertation with an automated corpus search.

9.5 Local contrast as diagnostics for constituent focus

There are some situations where an NP or PP bears contrast within its own constituent, and since contrast is one of the constituent focus indicators according to (200), it is worthwhile to look into them as a diagnostic. When a constituent has local contrast, one entity within the constituent is given preference over an other one, so that such constituents can be recognized by the presence of a negation (which shows the denial of one option) as well as a conjunction like “but” (which introduces the preferred option).

- (208) a. The sounds came nearer; dragging, crawling sounds, as if **not one but several creatures** were struggling across the floor. [BNC, G1L:2192]
 b. Democracy and unlimited government may be connected. However, it is **not democracy but unlimited government** that is objectionable. [BNC, EAJ:455]
 c. The two tests were explained in that case by the Lord Chancellor... who commented that **not the law but our mode of life** has changed over the years. [BNC, HXW:324]
 d. The person using the system provides the expertise necessary for the making of the work and is, for copyright purposes, the author of the work. That expertise may be applied directly or indirectly; for example, a person writing a report may draft it out on paper and then hand it to a word processor operator who enters it into the computer. In these circumstances, the author is **not the operator but the person writing the report**. [BNC, HXD:358]

The subject *not one but several creatures* in (208a) provides explicit contrast between “one creature” and “several creatures”. This is a clear sign of constituent focus, and this focus type takes precedence over the other focus articulations. The subject is “new”, since it introduces the entity “several creatures” into the mental model of the addressee, but the predicate is not new—the “struggling across the floor” can be inferred from the “sounds” and the “dragging” mentioned in the immediately preceding clause.

The *it*-cleft construction in (208b) contains a clefted constituent *not democracy but unlimited government* (which both have been mentioned just before), which contrasts “democracy” with “unlimited government”. We will focus on *it*-clefts in subsequent chapters, but notice that we have a combination of two strategies that are used to express constituent focus here: (a) a locally contrastive constituent, and (b) occurring in a construction (the *it*-cleft) that is often used for constituent focus.

The subject *not the law but our mode of life* in example (208c) contains explicit contrast between two entities, and is a clear indication of constituent focus, especially since the clause’s predicate *has changed over the years* contains information that is clearly assumed to be established in the addressee’s mental model of the situation.

The complement *not the operator but the person writing the report* of the equative clause in (208d) is a locally contrastive constituent, and it too expresses constituent focus. This clause has an established subject (“the author” is a class

description referring back to “the author of the work” with “Identity”), and it has an established complement (both “the operator” and “the person writing the report” have been mentioned in the previous clause). Within this context, the complement provides a *wh*-constituent answer to the constituent question: “Who is the author of such kind of work?” One possible answer is denied (that is: “not the operator”) and one other possible answer is confirmed (that is: “the person writing the report”).

9.5.1 Finding local contrast

An automated corpus search should be capable of finding several clear instances of local contrast, since the key elements of the contrast, an adversative conjunction like “but” and a negator, are both identifiable as elements of a noun phrase from the syntactic encoding of the parsed English corpora. The algorithm that locates the noun phrases with local contrast is described in (209).

(209) *Algorithm to detect local contrast*

Step 1: Consider each NP in the text, and check if it satisfies the following conditions:

Condition a: the NP contains a negator

Condition b: the NP contains a contrastive conjunction

Condition c: the NP is part of a main clause or complement clause

Condition d: this clause has an overt subject and a finite verb

Step 2: Let *cat* be the position of the NP within the IP (as per Table 30)

Step 3: Output:

Subcategorize on *cat*

The algorithm starts by selecting noun phrases, and checks if they satisfy the four conditions. Conditions *a* and *b* are used to see if the noun phrase has local contrast, and condition *c* and *d* check the clause of which the noun phrase is part: we want this to be a finite clause with an overt subject and a finite verb. Step 2 determines the position of the NP we found within the clause, and we use this position to subcategorize the results in the output.

9.5.2 An experiment with local contrast

When the algorithm described in (209) is executed on all of the four parsed English corpora, we do not get too many results, since this particular method of conveying constituent focus does not occur very often. This is why the results from subperiods are grouped into the four larger periods, and the six-fold word order division from Table 30 is collapsed in a two-fold one, as shown in Figure 27, while a full breakdown into the slot-structure positions is provided in Table 33.⁷

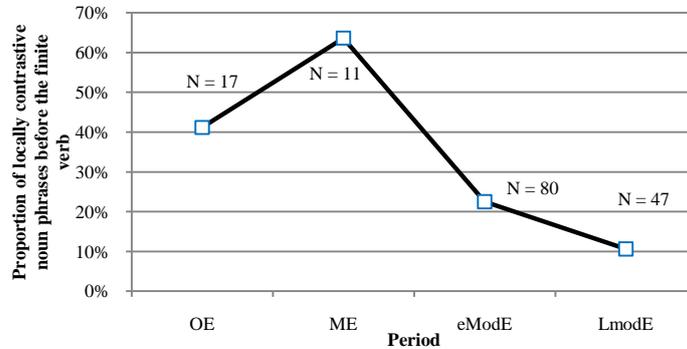


Figure 27 Percentage of local contrast noun phrases before the finite verb

Table 33 Positional distribution of local-contrast marked constituent focus

				OE	ME	eModE	LmodE	
PostVnonf	Vf	V	X	12%	0%	20%	6%	
PostVf	Vf	X		24%	18%	18%	17%	
Mid	Vf	X	V	0%	0%	0%	0%	
ImmPostVf	Vf	X		24%	18%	40%	66%	
Initial	X	Vf		24%	55%	11%	4%	
PreVf	Y	X	Vf	18%	9%	11%	6%	
				N	17	11	80	47

There is a decline in the proportion of locally contrastive noun phrases occurring before the finite verb—at least from ME until LmodE.⁸ Striking is the absence of locally contrastive constituent focus in the “Mid” area in all the periods. The results from the OE and ME period, however, may not be significant enough, since the total number of occurrences, 17 and 11 respectively, is rather low. The declining trend in constituent focus occurring before the verb as shown by this experiment confirms the results obtained for the adverb-modulated constituent focus experiments in section 9.2.

The fact that the OE and ME period show fewer results may stem from the difference in expressing local contrast (that is: the kind of NPs we are looking for in this section) for these periods. The OE and ME periods make more use of split constituents than does PDE. Another difference has to do with the way negation is expressed (Fischer et al., 2000, van van Kemenade, 1999, 2000, 2011). The examples in (210) should illustrate this

- (210) a. se hælend cwæð þis gehyrende, **Nys halum** læces
 the saviour said this to.the.listeners not.is whole.DAT of.doctor
 nan þearf ac seocum. [cowsgosp.o3:520]
 no need but sick.DAT
 ‘It is not the healthy who need a doctor, but the sick.’

- b. Ða cwæþ he, **ne** underfod **ealle menn** þis word
 then said he not understand all men this word/message
ac þam þe hyt geseald ys. [cowsgosp.03:1264]
 but those who it given is
'He said: "Not everyone understands this word, but only those whom it has been given."'
- c. Na us, Drihten, **na us, ac þinum naman** sele
 not us Lord not us but your to.name give
 þu wulder. [cobenrul:41]
 your glory
'Not to us, Lord, not to us, but to your name goes all the glory.'

The constituent “not the healthy but the sick” would have been the locally contrastive one in example (210a), but not so in the OE syntax. The two contrastive parts of the constituent have been split up over two noun phrases: (a) *halum* ‘whole ones’ and (b) *ac secum* ‘but sick ones’. What is more: the negator has contracted with the verb *be* into a negated variant *nys* ‘not is’. A similar picture obtains for the constituents *ealle menn* ‘everyone’ and *ac þam þe hyt geseald ys* ‘but those to whom it has been given’ in example (210b). These constituents together form a locally contrastive NP, but the second part has been dislocated to the end of the clause. Both examples lack a constituent negator as part of the first NP, which apparently is a feature that only gradually increases in English (van van Kemenade, 2011). Constituent negation, even without a concordant sentence negation, is already possible in OE, witness the example in (210c), where the negator *na* is a constituent modifier within the dative-case constituent *na us ac þinum naman* ‘not to us but to your name’.⁹

9.6 Emphatic pronouns as diagnostics for constituent focus

When reflexive pronouns occur as a modification of a noun phrase, they serve, as mentioned in 3.2.2.3, to put emphatic prominence on the entity referred to by the noun phrase, and emphatic prominence is one of the possible indicators of constituent focus according to (200). Very often the kind of emphatic prominence reached with emphatic pronouns is a clear indication of constituent focus, witness the examples in (211).

- (211) a. In A.D. 313 the Emperor Constantine issued the Edict of Milan which gave to Christians the right to practise their religion openly on an equal basis with other religions. In A.D. 325 **the Emperor himself** professed Christianity, which then became the official religion of the Roman Empire. [BNC HWB:932]
- b. “Can you work out where he was likely to have been put in?” “Not without knowing how long he was in the water. “Maybe we’ll get some idea of that from the medical report. Let’s think a bit more about **the man himself**. The doctor put him in the middle or late fifties, though, being a doctor, he hedged a bit by saying he might be anywhere between forty-five and sixty.” [BNC H0D:301]

The subject *the emperor himself* in (211a) is highlighted through the use of the reflexive pronoun *himself*. This is an example of constituent focus: we know from the previous sentence that there are people who “practise their religion (Christianity) openly”, and the current sentence assigns the thus established predicate of “professing Christianity” to one particular person, namely the emperor.

The highlighted constituent *the man himself* in (211b) is not a subject, but a PP object. The “man” being referred to is a person who has been found dead, and detectives are trying to figure out how this happened. In this situation we are dealing with a topic-comment articulation, where the topic is “we” and the comment is “think more about the man”. The highlighted constituent, then, is not an example of constituent focus. It is, however, an example of a DFE (dominant focal element—see 3.3.3), since it comes in a position within the clause where the Principle of Natural Information Flow is violated. The referent of “the man” is established in the mental model of the addressee, while “a bit more” is not, yet the more established follows upon the less established constituent, and this does not occur out of syntactic necessity.¹⁰

What we see from the examples above, then, is that emphatic prominence through emphatic pronouns is not a sufficient diagnostic for the recognition of constituent focus in and by itself. Detecting constituents with emphatic pronouns does get us constituents that are somehow highlighted, but we need to take additional measures to distinguish DFEs, which occur inside the comment part of the topic-comment articulation, from real constituent focus situations. Since this is an additional burden for the automated corpus research that will give us quantitative results, and this is not a trivial one, it is, for the current study, better to keep the emphatic pronouns out of the study that focuses on clear diagnostics of constituent focus.

9.7 Apposition as diagnostics for constituent focus

A diagnostic of constituent focus that has been mentioned in section 3.2.2.3 as well as in conjunction with the two texts investigated in chapter 4 is that of apposition: that of describing a participant by one or more additional noun phrases on the same syntactic level. The fact that one participant is described by more than one noun phrase is an indication that the entity referred to is probably new to the mental model of the addressee, and needs additional attributes in order to be linked to previous knowledge, or to be established as uniquely identifiable. Several examples of apposition from Present-day English are given in (212).

- (212) a. Other food colourings, **particularly the synthetic ones**, have been known to cause allergic dermatitis, mainly in food workers exposed to large amounts. [BNC BMI:617]
- b. A review of all government policy affecting the environment was announced yesterday by Chris Patten, **the Secretary of State for the Environment**, in a Conservative Party conference speech that flagged a shift in the Government's ideological stance. The review will lead to a “green” White Paper, planned for publication next summer and likely to provide the kernel of the Tories' next general election manifesto. [BNC A53:91]

The subject *other food colourings* in (212a) is supplied with an apposition *particularly the synthetic ones*, which identifies it clearly as a referentially new entity. The question is whether this in itself is enough to make the jump to such a constituent belonging to the constituent focus articulation. The example in (212a) seems to deny this: the subject is referentially new, but so is the predicate, so that we are dealing with a thetic focus articulation.

The person *Chris Patten* in the passive by-phrase in (212b) is apparently new to the addressee (or at least not well known enough), so that a clarifying apposition is added stating his function. It seems, however, that the syntactic subject of the sentence is new too, and in fact, the sentence is an example of thetic focus, introducing a *review* which is picked up on in the immediately following sentence.

In sum, we may conclude that apposition is a good diagnostic for referential newness, but not of constituent focus. This conforms to what has been said about the asymmetrical relation between newness and focus in section 3.5: constituents that are referentially new are part of the focus domain, but the focus domain is not necessarily restricted to one (or more) referentially new constituents.

9.8 Split constituents as diagnostics for constituent focus

The OE Euphrosyne text treated in section 4.4 showed several interesting cases of split constituents, and all of them seemed to have resulted from the desire to satisfy several constraints at the same time: (a) put syntactic information where it is needed (such as subjects before the verb), and (b) the Principle of Natural Information Flow: put less established information before more established. Splitting of constituents in two parts is a feature that was probably more in use in Old English, but still happens to some extent today—especially with extraposed relative clauses. In fact, we probably need to distinguish extraposed relative clauses from other split constituents. Two examples of extraposed relative clauses are shown in (213).

- (213) a. This old man would fix his eyes upon Edmund, whenever he could do it without observation—sometimes he would sigh deeply, and **a tear** would start from his eye, **which** he strove to conceal from observation. One day Edmund surprized him in this tender emotion, as he was wiping his eyes with the back of his hand. [reeve-1777:541-3]
- b. The baron agreed with him in opinion, that **a man** was of much more service to the world **who** continued in it, than one who retired from it, and gave his fortunes to the church, whose servants did not always make the best use of it. [reeve-1777:332]

The subject *a tear* in example (213a), which is from a late Modern English text, is only the first part of a constituent that includes a postmodifying relative clause, but this relative clause is extraposed—it appears at the end of the sentence. The question is why the author has chosen to use a split constituent, an extraposed relative clause, and whether this signals the use of a particular focus structure. We can start addressing the question by looking at the referentiality of the noun phrases in the clause. The subject *a tear* is referentially new, although one could argue that it is inferable from *his eyes* in the preceding sentence. Be that as it may, the prepositional object *his eye* in the current clause is referentially not new (it has an “Identity” link to *his eye* in the preceding clause), yet the PP *from his eye* comes clause-finally. What we see is that the Principle of Natural Information Flow has been overruled in this example, and the author would not have been *required* to select this order.¹¹ As for the question what the focus structure of the clause is, we have to decide whether the focus domain only includes the subject *a tear*, in which case we have constituent focus, or includes the verb phrase *would start from his eye*. The last choice seems the most fitting one: the focus domain includes the verb phrase, so that we have *thetic* focus here. The thetic focus is used for presentational focus: it introduces the significant participant *a tear* (and all that is associated with it). It is this participant that the old man tries to “conceal”, but Edmund nevertheless sees the tear when the old man “wipes his eyes”. In sum, the presence of an appositive relative clause does *not* seem to be an indication that the information contained in it is backgrounded (in fact, it is of vital importance for the ensuing storyline), and extraposition as such cannot be linked to constituent focus in this case.

The subject *a man* in example (213b) has an extraposed *restrictive* relative clause. It could, at first glance, be more easily identified as being part of constituent focus. We can argue that the subject *a man* is referentially new, but it is part of a type of open-value proposition *X is of service to the world*, where the value of the variable *X* is supplied by the whole noun phrase *a man who continues in the world*. We know that the proposition reflects presupposed information, since the text explicitly says that *the baron agreed with him in opinion*, and it is the baron’s opinion that is being repeated here. However, if we ask ourselves what is being contrastively focused with what, then it is not the constituent *a man ... who continued in it* versus *one who retired from it*. The contrast really only is between *continued* and *retired*, and it is not clear at all what the focus domain is. We could even argue that *a man* is not referentially new at all, but is an instance of the

prototypical “man” which is “Assumed” to be readily available in the addressee’s mind.

All this is to say that extraposed relative clauses do not always provide us with a clear link to constituent focus. But does the same hold for other split constituents—the ones that we find in Old English, for instance? Consider the example in (214).

- (214) a. Ða æt nyxtan com **him** an þegen **to**, [coeuφr:33]
 then at last comes to.him a nobleman to
 (se wæs weligra and wurþra þonne ealle þa oþre, and hire to him gyrnde.)
 ‘Then at last came to him a noble who was wealthier and worthier than all the others, and desired her for himself.’
- b. (Heo þa þone wiflican gegyrlan hire ofdyde, and hi gescrydde mid werlicum, and on æfentid gewat of hire healle, and nam mid hire fiftig mancsas, and þa niht hi gehyde on digelre stowe.)
 Ða **þæs** on **mergen** com Paphnutius to þære ceastre, [coeuφr:133-138]
 then of.the on morning comes Paphnutius to the city
 ‘(Then she put off her womanly garb, and clothed herself with a man’s and in the eventide departed from her hall, and took with her fifty mancuses, and that night she hid in a secret place.)
 Then afterward, in the morning, Paphnutius came to the city.’

The PP *to him* in (214a) is split into two parts in such a way that the order is reversed (*to him* becomes *him to*) and *an þegen* ‘a nobleman’ now is captured between the two parts. But is there any relation to constituent focus? This does not seem to be so here. What we have is presentational focus (part of the thetic articulation) on the referentially new subject *an þegen* ‘a nobleman’. This new participant is placed, conform the Principle of Natural Information Flow, as far as possible to the end of the clause, and then picked up as new topic in the next clause by the demonstrative *se*.¹²

The constituent *on þæs mergen* ‘in the morning’ in (214b) is also split in two parts, so that the demonstrative *þæs* occurs before the PP *on mergen* ‘in morning’. The reason for this splitting may be found in contrast: the “morning” in this line is contrasted with the *niht* “night” in the previous line. Nevertheless, even this example does not provide a situation with constituent focus, since the contrasted constituent is part of a point of departure, and not of an argument or adjunct in the main part of the clause.

The above examples show that there is no correlation between the occurrence of split constituents (extraposed relative clauses or others) and constituent focus on the head of the split constituent. This is why no quantitative experiments are done to locate the split constituents.

9.9 Contrastive left dislocation

The reason to look at “contrastive left dislocation” as a possible structure related to constituent focus is that this structure contains a constituent that is contrastive, and contrast has been identified as one of the possible characteristics of a constituent that is part of a constituent focus articulation (see 200). Left dislocation in general is

related to the different kinds of split constituents that have been discussed in the previous section, but it is not quite the same. The most notable difference between the two phenomena is that split constituents consist of two parts that together form one constituent, whereas this is not the case with left dislocation. A left-dislocated constituent, such as the constituent *weather at sea*, *weather on the mountains* in (215a), is complete as it stands, and not split out in parts. It is positioned before the body of a sentence, and is referred to from within the sentence with a resumptive, such as the object pronoun *it* in (215a). The presence of such a resumptive is one of the main characteristics of left dislocation.

- (215) a. Weather at sea, weather on the mountains, he could foretell **it** always.
[meredith-1895:473]
- b. Under-Secretary for Foreign Affairs at the age of forty—**that**'s good enough for anyone, I should think.
[wilde-1895:80]
- c. On the whole, however, you are not to take gloomy views for there is nothing to mourn at, to despair at: a serious cheerfulness; **that** is the right mood in this as in all cases.
[carlyle-1835:372-374]

Left dislocations come in different kinds, and the one in (215a) is called a “Hanging Topic Left Dislocation” (de Vries, 2007, Prince, 1984). The type that could possibly have a link to constituent focus is called “Contrastive Left Dislocation” (CLD), and (215b) serves as an example for this type. The NP *that* is used to resume the left dislocated constituent *Under-secretary for foreign affairs* is a demonstrative pronoun, and this demonstrative receives contrastive stress. The same goes for the noun phrase *a serious cheerfulness* in (215c), which is resumed with the demonstrative pronoun *that*. Cornish (1999) argues that there is a strong tendency for stressed demonstrative pronouns to be contrastive, and so does de Vries (2007). I argue that demonstrative pronouns as such need not necessarily have a contrastive interpretation: what they do is identify one particular constituent, but identification is only one ingredient of contrast (the other element being the explicit or strongly implicit presence of alternatives). Stoop (2011) and Veeninga et al (2011) argue that CLD in Dutch is related to mark the shift to a different topic, so not to constituent focus. Nevertheless, when demonstratives function as resumptives for a left dislocated constituent in English, they seem to co-occur with a contrastive interpretation, which is why I will have a closer look at their behaviour and see if the resumptive demonstrative pronouns can serve as a diagnostic for constituent focus.

9.9.1 Finding CLD resumptives

The search for resumptives of contrastive left dislocation constructions can be done with parsed English corpora, since the annotators have added functional labels which help us locate left dislocated constituents (the label extension “LFD”) and resumptives (extension “RSP”). The recognition of demonstrative pronouns is something that has already been done as part of adding the “NPtype” feature to noun phrases (see section 6.2). The algorithm that locates the resumptives of left dislocations is described in (216).¹³

(216) *Algorithm to find demonstrative resumptives of left dislocation*

Step 1: Consider each constituent in the text, and check if it satisfies these conditions:

Condition a: the constituent is marked as left-dislocated

Condition b: the resumptive NP in the same clause has Nptype ‘Dem’

Condition c: this clause has an overt subject and a finite verb

Step 2: Let *cat* be the position of the NP within the IP (as per Table 30)

Step 3: Output:

Subcategorize on *cat*

The algorithm starts by selecting left-dislocated constituents: all those that have “LFD” in their label (these may be clauses or phrases). It then checks for crucial elements in the clause associated with the left-dislocated constituent. Condition *b* locates the resumptive noun phrase and checks if its Nptype feature is ‘Dem’, and condition *c* checks whether the clause of which the noun phrase is part has an overt subject and a finite verb. Step 2 determines the position of the NP we found within the clause, and we use this position to subcategorize the results in the output.

9.9.2 An experiment with CLD resumptives

When we apply the algorithm to find demonstrative pronouns that function as resumptives for contrastive left dislocated constituents to all the four parsed English corpora, we get the results as summarized in Table 34 (since the positional variation in terms of Table 30 only involves ImmPostVf and PreVf, these two categories are abbreviated as “preceding V_{fin} ” and “following V_{fin} ”).

Table 34 *The position of CLD resumptive demonstrative pronouns*

	OE	ME	eModE	LmodE
Subject precedes V_{fin}	65%	82%	43%	73%
Subject follows V_{fin}	9%	1%	0%	0%
Object precedes V_{fin}	22%	17%	58%	27%
Object follows V_{fin}	4%	0%	0%	0%
<i>N</i>	645	163	40	26

The results in the OE period are as we would have expected: the clause-initial position hosts about 87% of the contrastive resumptive NPs.¹⁴ The absence of a divergence from this trend towards LmodE is interesting in itself: it could indicate that the clause-initial position *retains* its ability to host contrastive constituents, which contradicts the results we have seen so far, where we have been observing a shift of the contrastive position to the end of the clause. However, the decline in the occurrence of CLD is so drastic, that by the LmodE period there are only 26 occurrences to choose from, which means that the significance of the results in LmodE has decreased. An example of an outcome that militates against our expectations is given in (217).

- (217) a. Howbeit he that sent me is true; and the things which I heard from him, **these** speak I unto the world. [erv-new-1881:692]
 b. Now since all Languages are naturally equal to us, therefore the first Language we hear, **that** we shall first understand. [anon-1711:1735]
 c. Being so near as to be within hearing, I made Friday go out upon the Deck, and call out aloud to them in his Language to know what they meant, which accordingly he did; whether they understood him or not; **that** I knew not. [defoe-1719:8-9]

The result in (217a) is a clear LmodE example of a resumptive object demonstrative pronoun in clause-initial position, but it is from a New Testament translation, and could therefore have been influenced by the source language or by the translator's desire to use an "elevated" liturgical register (which is likely to contain archaisms). The results in (217b,c), however, are not from translations and nevertheless do show the demonstrative pronoun with its constituent focus in the clause-initial position.

It is difficult to answer the question *why* the demonstrative pronoun resumptives of left dislocations are in the position we find them. One answer to this question could be that there are conflicting forces at work with opposing demands on the position of the constituent that has contrastive focus. One such tension could be the desire to have a referential demonstrative pronoun occur as near as possible to its textual antecedent, in order to not jeopardize the demonstrative's antecedent identification. If that constraint overrules the Modern English constraint of having constituent focus in a clause-final position, then this would explain the reason for the uniform placement of demonstrative resumptives in a clause-initial position.

An alternative answer to the question *why* the demonstrative pronoun resumptives occur in the position they do can be that the necessary (contrastive) focus interpretation is the result of combining the identificational properties of demonstrative pronouns with the particularities of the pre-subject position, so that only the combination of these two would lead to constituent focus. It is striking, in this context, that the loss of demonstratives from the first position in OE seems to have been one of the major causes that led to the loss in the abilities of the first position to host contrastive focus in later stages of English (Los, 2012).

9.10 Constituent answers as diagnostics for constituent focus

A standard case of constituent focus occurs where an NP constitutes the answer to a *who* or *what* question: such constituent supply the value for a variable that is created in the question (see to 200a). A constituent question like "Who did the dishes?" establishes an open proposition where the variable (the agent of the dish-washing in this case) is quite probably to be found in the next clause, the answer to this question. An NP constituent answer can only be expected to be given in answer to certain *wh* questions: *who*, *what*, *where* and *when*. Question words like *why* and *how* do not give the kind of constituent answers we are looking for, since a reason (the answer to *why*) or a means (an answer to *how*) are usually expressed as clauses, so that the focus domain of the answer is not restricted to one constituent. We also need to take distinguish between different kinds of answers that are given to *what*

questions, since some of these have a constituent focus articulation (where an NP constituent provides the answer to the *what* question), whereas others have a sentence focus articulation. If one asks “What happened”, for instance, the focus domain of the answer is very likely to span the whole clause, which is thetic focus instead of the constituent focus we are looking for.

The parsed corpora of English have labelled the question words in such a way, that we are able to locate them, but the question is whether we would be able to automatically (programmatically) capture the NP or PP constituents that provide the answers to the questions. A cursory look at the possible answers to *who* questions in (218) shows us that our task is not going to be accomplished automatically.

- (218) a. “And **who** was he?” inquired Mr Pickwick. “Vy, that’s just the wery point as nobody never know’d,” replied Sam. [dickens-1837:88-89]
- b. “But yet there is one who is thought to exceed them all, though he is the son of a poor labourer.”
 “And **who** is he,” said the knight?
 “**One Edmund Twyford**, the son of a cottager in our village. He is to be sure as fine a youth as ever the sun shone upon, and of so sweet a disposition that nobody envies his good fortune.” [reeve-1777:275-278]
- c. “You know I don’t love to hear you talk about Politics; they belong to us, and Petticoats should not meddle: but come, **Who** is the Man?”
 “Marry!” said she, “you may find him out yourself, if you please.” [fielding-1749:153-156]

The answer to the *who* question (218a) is lacking, because the person who is supposed to answer the question simply doesn’t do it. The answer to (218b) is given in the next clause, which, as is quite common in answers to *who* questions in PDE, only consists of the NP that supplies the variable for the open proposition. The question in (218c) does lead to a response, but this response is not the answer. If we were to simply take the first NP in the response as the answer, we would be quite led astray. We have to conclude, then, that the idea of using answers to *wh* constituent questions as a diagnostic for constituent focus is not something that can be dealt with through a corpus research algorithm.

9.11 Clefts as diagnostics for constituent focus

The question to what extent *it*-clefts function as a diagnostic for constituent focus (which first came up in chapter 4, section 4.7.5.6) will be addressed in chapters 10-12, since it needs much more attention and the answer can offer us insight into interchange between syntax and information structure. What we look at in this section is the relation between *wh*-clefts (sometimes referred to as “pseudo-clefts”) and constituent focus. The reason we look into these kinds of clefts is that they contain a free relative, and such a relative generates a variable, which is one of the things that relates to constituent focus according to (200a). Examples of a *wh*-cleft and a reversed *wh*-cleft are given in (219).

- (221) a. **What I have often asked myself** is how other linguists manage to keep abreast with the rapid developments in the different fields of linguistics while still finding time to go on writing articles themselves. One colleague who has proved to be able to do this and who I have the honour to introduce to you tonight is Mr. ... (Declerck, 1984: 257)

Declerck (1984) already noted that *wh*-clefts can occur as “discourse openers”, but only if they are of the “informative presupposition” type, as the one in (221a), which is a clear indication that the free relative NP subject does not represent ‘established’ information. Hedberg (2007) studied the use of clefts in spoken English, and she found several instances of what she labelled “informative” free relatives—ones that represent discourse-new information, although the number of free relatives in *wh*-clefts she assigned the status of “topical” (hence: established information) was much larger.

All this is to say that there is no implicational relationship between a free relative NP and its referential status, so that we cannot derive the focus articulation of the different kinds of *wh*-cleft from the information status of these free relative NPs.

9.11.2 Constituent focus and *wh*-clefts

What can we say about the relationship between *wh*-clefts and constituent focus? We have seen that the information status of a free relative NP does not derive straightforwardly from the fact that it is a free relative, so we cannot make a generalisation about the referential status of the free relative subject in a *wh*-cleft. But there is something else we can say with confidence: the free relative expresses an open proposition, and the NP complement in a *wh*-cleft provides the value for this variable. If we take, for example, the *wh*-cleft construction used in (220b-c), we have the open proposition “John can make *x*”, and we have the construction assign the value *a painting* to this variable *x*. This, then, is almost a prototypical situation of constituent focus—not on the free relative NP subject, but on the complement NP.

Having established the relationship between *wh*-clefts and constituent focus, we should now ask ourselves if *wh*-clefts can serve as a diagnostic for constituent focus in our search for a change in the *position* of focused constituents with respect to the word order in the sentence. The answer to that should be “no”. We cannot use *wh*-clefts in this sense, because the position of the focused constituent, the complement NP, is “fixed” by the definition of the *wh*-cleft itself: the complement must always follow the finite form of the verb “be”, otherwise we don’t have a *wh*-cleft but something else.

9.11.3 Constituent focus and reversed *wh*-clefts

Reversed *wh*-clefts are similar to *wh*-clefts in the sense that both contain a free relative NP. The difference is that this free relative NP is the subject in a *wh*-cleft, whereas it is the complement in a reversed *wh*-cleft. Both the *wh*-cleft and the reversed *wh*-cleft are equative constructions, and such constructions can in principle be specificational or predicational (see section 3.2.2.1), and the referential status of the subject and the complement can differ. We have seen in chapter 3 that only

specificational equative clauses with a subject that is referentially newer than the complement have a constituent focus structure. The reversed *wh*-clefts are no exception to this rule: only specificational ones have a constituent focus structure. While this is something we need to keep in mind, reversed *wh*-clefts seem to associate with constituent focus very often, witness the examples in (222).

- (222) a. I've done my best. I thought that was **what I was being paid for**.
[BNC HD7:2590]
- b. “Oh, and Elsa, if anyone asks you what nationality you are, say you're Swiss.” “Why? I don't want to say I'm Swiss. I'm proud of being German.” “Be guided by me, my dear girl. If you wish to keep your job, Swiss is **what you need to be**.”
[BNC HTG:275]
- c. It's good to see you out and about. Fresh air is **what you need**—that and time will see you through these early discomforts. [BNC H82:177]

The reversed *wh*-cleft in (222a) has the subject *that*, which has the referential status of “Identity”, since it links back to the whole first sentence. The referential status of the free relative *what I was being paid for* is “New”, since the purpose (implied head noun) for which the person is being paid is stated here for the first time (the presupposition that there is a reason for which “I” am being paid is internal to the free relative, and bears no relation to the referential status of the whole free relative NP). However, even though the subject in the equative clause represents established information and the complement is new, the focus articulation is not a topic-comment one, but a constituent focus one. We are not so much dealing with a topical “that” and a comment being made about this topic. The fact that the free relative contains an open proposition (I am being paid for *x*) and that the value for this proposition is provided by the subject “that” is more important here: we have constituent focus on the subject “that”.

The same reasoning goes for the reversed *wh*-cleft construction in (222b). Even though we have an established subject “Swiss” and a complement free relative *what you need to be* (which links back with “Identity” to *nationality* in the preceding discourse), the presence of the open proposition “you need to be of nationality *x*” overrules all other matters, so that we end up with constituent focus on the subject “Swiss”. This is confirmed by the observation that there is explicit contrast: the nationality “Swiss” is compared to “German”.

The reversed *wh*-cleft in (222c) has a referentially new subject *fresh air*, and the fact that there is something “needed” by the protagonist may be presupposed in the free relative, but is referentially new in the context. Overruling the referentiality concerns, however, is the presence of the open proposition “you need *x*” and the fact that “fresh air” provides a value for this *x*, so that here again we have constituent focus on the subject.

9.11.4 The development of *wh*-clefts

We have seen that *wh*-clefts and reversed *wh*-clefts are diagnostics for constituent focus, since the NP complement in a *wh*-cleft and the NP subject in the reversed *wh*-cleft provide the value for the open variable created by the free relative. But we also

know that the definition of the *wh*-cleft and the reversed *wh*-cleft specify a particular word order for these constructions, so that each construction in its own is not usable as a diagnostic for the question we set out to answer in this chapter, which is how the position of constituent focus changes over time.

Nevertheless, if what we have seen in previous sections of this chapter is true, namely that, as I have stated in the hypothesis (197), the preferred position for constituent focus shifts from clause-initial to clause-final, then we would expect a development where either reversed *wh*-clefts start to appear earlier than *wh*-clefts, or a development where the former appear more frequent than the latter, since such trends retain the principles in (198) and (199). This is a hypothesis we can verify with a corpus research project, and the results of such an attempt are shown in Table 35.¹⁵ The corpus project involves a search of all the four parsed English corpora, and the free relatives are detected by making use of the labels provided by the researchers who have created these corpora. The clauses in which we look for the different *wh*-clefts may be main clauses or complement clauses. We have excluded question sentences as well as sentences that do not have an overt subject.

Table 35 Occurrence of *wh*-clefts versus reversed *wh*-clefts

	OE	ME	eModE	LmodE
<i>wh</i> -cleft	0	7	14	33
reversed <i>wh</i> -cleft	2	40	14	64

The results of the corpus experiment seem to show two things confirming the hypotheses above: reversed *wh*-clefts appear slightly earlier than *wh*-clefts, and they occur more frequently (except for the eModE period). However, the total number of results is rather low, and the significance of the results is therefore rather low too.¹⁶

When we take a closer look at the results, it becomes clear that large part of the reversed *wh*-clefts consists of clefts with a demonstrative pronoun (*this*, *that*, *these*) as subject, but also that there are a number of reversed *wh*-clefts that may not have constituent focus at all. Some of the disputable ones are shown in (223).

- (223) a. “What was that?” [boethri-1785:264-266]
 “The end, added she, of all things; for the end of all things is **what they pursue**.
- b. Messieurs, you are today **what you were yesterday**. [carlyle-1837:148]
- c. After this introductory preface, the three chums informed Mr Pickwick in a breath, that money was, in the Fleet, just **what money was out of it**; that it would instantly procure him almost anything he desired.
 [dickens-1837:479]
- d. You know, my dear Tom, how much I admire your proficiency in the New School of breeding. You are, **what I call, one of the highest finish'd fellows, of the present day**. [colman-1805:105-106]

The example in (223a) is a situation where the subject of the reversed *wh*-cleft *the end of all things* is highly topical, since it is a verbal repetition from the immediately preceding clause. This makes it rather difficult to decide between a reading where

we have a topic-comment structure, with *what they pursue* representing the most informative part of the utterance, and one where we have a constituent-focus reading, where *the end of all things* is emphasized as representing the crucial value for the open proposition “they pursue something”.

A slightly different problem occurs in (223b,c), where we have overt contrast, but not with the subject of the reversed *wh*-clefts, but with an adjunct. Example (223b) has contrast between “today” and “yesterday”, and (223c) has contrast between “in the fleet” and “out of it”. These observations prohibit a reading where we have constituent focus on the subject.

The problem in (223d) is yet of another kind. We have a well established pronominal topic “you” as subject, and then we have a free relative NP that contains a modifier “one of the highest”, which seems to make the equative clause into a kind of predicational one. It certainly has no specificational reading, so that it is hard to agree on constituent focus on the subject “you”.

To conclude, then, the *wh*-clefts help us little to nothing; partly because they occur so rarely, and partly because the relationship between *wh*-clefts and constituent focus does not always turn out to be what we had expected it to be. We should, therefore, leave the *wh*-clefts out of the discussion concerning the preferred position for constituent focus.

9.12 Discussion

After the chapters 5-7 paved the way for a corpus based research into the development of focus, and chapter 8 did just that for presentational focus, the chapter at hand has concentrated on constituent focus, and the way it changed in English. The development of the way this focus articulation is expressed can be correlated with the changes in English syntax, as has been stated in the hypothesis (197) in the beginning of this chapter: the loss of V2 forces constituent focus from the PreCore to the PostCore area. The corpus research in this chapter aimed at verifying this hypothesis by finding and verifying non word-order related diagnostics to reveal the preferred position of focused constituents, although we realize with Miller (2006), that position is very unlikely to be the *only* landmark of constituent focus.

The diagnostics reviewed in this chapter have been chosen based on the likelihood that they are indicative of constituent focus, but not necessarily fixed to a particular word order. Several of the diagnostics proved to be not so helpful. The fact that a constituent (an NP or PP) is negated (9.3) indicates that it is part of the focus domain, and there is a link with constituent focus if the constituent is not a subject. If it is a subject, then we need to know its referential status: negated new subjects point tothetic focus, while it is only negated established subjects that associate with constituent focus. But if a diagnostic for constituent focus such as negation associates with the syntactic function of subject, then we are very likely to get skewed results, since subjects increasingly appear before the finite verb in English anyway, unrelated to them being focused or not. The diagnostic of positive negation (9.4) and that of emphatic pronouns (9.6) are not necessarily indicators of

constituent focus—they only tell us that an NP or PP is emphatically prominent. Apposition (9.7) does not work as a diagnostic for constituent focus either—but it is a clear signal that a constituent is referentially new. Split constituents (9.8) sometimes coincide with constituent focus, but not always; they are more a sign of a strategy that several constraints are satisfied in parallel. The demonstrative pronoun used as resumptive for contrastive left dislocation (9.9) seems to indicate constituent focus, but the results we obtain deviate from the other findings, since they show a uniform tendency for the position of constituent focus to be clause-initial. The hypothesis that they indicate constituent focus may have to be revised, or, alternatively, there may be an overruling constraint at work here, which wants to minimize the distance between a demonstrative pronoun and its antecedent. Answers to constituent questions (9.10) like “who”, “where”, “when” can relate to constituent focus, but they are in practice so unpredictable, that we cannot automatically look for the constituent that answers the question: sometimes there is no such constituent at all. The different types of *wh*-clefts (9.11) do not always associate with constituent focus, and they occur too infrequently to be helpful in shedding light on our research question.

What we end up with in this chapter are two clear diagnostics of constituent focus: the presence of contrastive adverbs in an NP or PP (9.2) and overt local contrast within an NP (9.5). Both of these diagnostics illustrate an answer to the research question in (11): the loss of V2 leads to a change in the preferred position for constituent focus from the clause-initial (PreCore) one in OE to the clause-final (PostCore) one in LmodE, although the end of ME (around 1500) differs significantly. The fact that the “new” position for constituent focus is the PostCore one is confirmed by the absence of constituent focus in the “Mid” area (the Core area between the Vb1 and Vb2 slots). The problem with the PostCore area in LmodE, however, is that it is not such a clearly demarcated area (see the demarcation principle (198)), nor does it provide for the focused constituent to *precede* the rest of the clause (see the placement principle (199), but this is contingent upon the referential status of the focused constituent). This is where one potential candidate comes in that we have not discussed in this chapter: the *it*-cleft construction. This construction *does* satisfy the demarcation principle (since the clefted constituent is demarcated as the complement in a copular construction) as well as the placement principle (since the clefted constituent precedes the remainder of the clause), and it is the topic of chapters 10 to 12. Implications of the findings on presentational and constituent focus will be discussed in chapter 13.

¹ The five positions chosen here are reminiscent of those chosen for the subject position in section 8.2 where we dealt with presentational focus; see Table 25.

² The particle *only* can also occur in a position outside of the NP or PP it modifies, but it has a slightly different meaning then (see for instance Hendriks, 2004 and references therein).

³ The query looking for focus adverbs is supplied in appendix 14.2.3.

⁴ The years mark the *end* of the period from which the texts have been taken. So the point “1150” contains texts from the period between 950 and 1150. The D[corp] is approximately 45% (so more than half the texts do not contain noun phrases with the contrastive or emphatic adverbs we are looking for). The transitions from O1-2 to O3-4 and then to M1 are all significant according to Fisher’s double-sided exact test ($p < 0,05$), but all the other transitions between periods are not.

⁵ D[corp] is 86%, and all the period-transitions are significant according to Fisher’s two-tailed exact test ($p < 0,05$). The code for the query looking for *have* sentences is provided in appendix 14.3.12.

The percentages for all subject positions (see Table 30) occurring in any time period is as follows (the verb *have* does not occur with a subject in the “Mid” position):

	OE	ME	eModE	LmodE
PostVnonF	0,0%	0,3%	0,0%	0,0%
PostVf	0,6%	0,2%	0,1%	0,0%
ImmPostVf	15,2%	10,7%	2,8%	0,6%
Initial	62,1%	57,1%	60,9%	70,3%
PreVf	22,1%	31,6%	36,2%	29,1%

⁶ A neutral variant would, in my opinion, be “But there was some opposition to the move.”

⁷ The query that looks for local contrast is provided in appendix 14.2.4.

⁸ D[corp] is 16% (see 1.4.3). The transitions from OE to ME and from eModE to LmodE are not significant according to Fisher’s two-tailed exact test ($p < 0,05$), but the transition from ME to eModE is ($p = 0,0085$), even with the limited amount of data available. See for details the appendix, section 14.3.13.

⁹ A thorough investigation into the development of constituent focus in English is outside the scope of this thesis.

¹⁰ It would have been possible to say “Let us think about the man a bit more”. See the discussion on the dative alternation in section 3.3.3.

¹¹ If the author had not wanted to put constituent focus on *a tear*, he could have done so in various ways. One way would have been to not single out one particular tear at all: “his eye would start to run”. If one particular tear has to be introduced, so that it can be picked up in a relative clause, the author could have said: “... and his eye would produce a tear, which he strove to conceal from observation”.

¹² One can interpret *se* ‘this’ either as demonstrative, in which case the next clause is an independent main clause, or as a relative pronoun, in which case the clause is a relative clause.

¹³ The code for the query looking for CLD resumptives is provided in appendix 14.2.5.

¹⁴ D[corp] is 19% (see 1.4.3). The period-transitions in the “Subject precedes finite verb” line are significant according to the two-tailed Fisher’s exact test ($p < 0,05$). The period transitions in the “Object precedes finite verb” line are significant too, except for the OE to ME transition. The transitions in the other two lines (subject or object following the finite verb) are not significant except for the OE to ME transitions. See for details the appendix, section 14.3.14.

¹⁵ The code for the query looking for *wh*-clefts is provided in appendix 14.2.6.

¹⁶ D[corp] is 6% for the *wh*-clefts and 10% for the reversed *wh*-clefts.

The huge changes in the treatment of constituent focus presented in chapter 9 bring us back to the main research question in (11) how syntax and information structure interact. If Old English had a privileged clause-initial (or “PreCore”) position for constituent focus, and it lost this position in Middle English, then this raises the question what alternative strategy the language started to use, in order to express constituent focus.¹ The intuitive answer to this is that the cleft construction may have filled the gap, since *it*-clefts are often seen as focusing constructions par excellence; Lambrecht (1994: 70-71), for instance, sees a non-focus related use of the *it*-cleft as a “conventionalized pragmatic accommodation.

With the question on the inherent constituent focus function of the *it*-cleft, the following three chapters zoom in on this construction, looking at it from a synchronic and a diachronic point of view. The chapter at hand looks at *it*-cleft constructions in general, in order to provide the groundwork for the language-specific treatment of them in chapters 11 and 12. Section 10.1 lays the foundation by giving a clear definition of *it*-clefts. The second part of this chapter, section 10.2, discusses the *function* of *it*-clefts, which, as mentioned above, seems to have been pinpointed as that of “focusing”. Hasselgård (2004), on the other hand, shows that *it*-clefts in Scandinavian languages function as a thematizing device and are used to organize segments of a text (in terms of the 3D model proposed in 4.1 the *it*-cleft is a linguistic realization of particular values on the “text-structure” axis). It is this function of *it*-clefts that I will identify as the predominant one for present-day Chechen (chapter 11) and the initial one for Old English (chapter 12).

10.1 Defining clefts

Before we look at the numerical evidence on clefts in present-day Chechen and in the history of English, we need to be able to decide what a cleft is. A definition of clefts should be based on the *form* of the construction and its components alone, and it should not include references to its *function*—otherwise we would not be able to objectively note its function in a particular stage of a language.

I will argue specifically (in section 10.1.2) that time adjunct clefts are to be regarded as proper *it*-clefts (contra Ball, 1991), in order to pave the way towards the treatment of clefts in Chechen and English.

As we consider several constructions that should not be regarded as proper *it*-clefts, a small set of objective criteria emerges by which we can say whether any construction is an *it*-cleft or not.

should be captured in our definition, which will take account of English as well as cross-linguistic variation.

- (227) a. This is **a serious problem** we have here.
 b. Those are **my biscuits** you're eating. (Ward et al., 2002: 1420)
 c. Ða wæs **þy æfterangeare**, cwom sum monn [cobede:1152]
 then was the next year came some man
 in Nordanhymbra mægde.
 in Northumbrian's country
 'Then the next year a man came to the country Northumbria.'
 d. It could have been **Darwin himself** who introduced Dr Benjamin Bynoe,
 the Beagle's surgeon, to Gould. [BNC HRB:277]
 e. It was **in September 1990** that The Royal London Hospital, Whitechapel,
 celebrated its 250th anniversary of continuous service to the community.
 [BNC AOX:843]

The pronoun *it* can be replaced by a demonstrative pronoun as in (227a,b), or even completely left out as in the Old English cleft in (227c). Contrary to Jespersen's summary in (225), but in line with the definition in (226), the form of *to be* can vary in tense, mood and aspect, as in (227d). Also in line with (226), the clefted constituent does not necessarily have to be a noun phrase, but could be a prepositional phrase, as for example (227e). Lambrecht (2001) proposes a more elaborated definition:

- (228) "A cleft construction is a complex sentence structure consisting of a matrix clause headed by a copula and a relative or relative-like clause whose relativized argument is co-indexed with the predicative argument of the copula. Taken together, the matrix and the relative express a logically simple proposition, which can also be expressed in the form of a single clause without a change in truth conditions." (Lambrecht, 2001: 484)

Lambrecht does not stipulate the presence or form of a pronoun like *it*, which is justified by the variation in the data from Present-day English (227a-b) and Russian (229a), where a demonstrative pronoun is used, and from Old English (227b), which is subject-less.

Russian has been argued to have clefts (Gundel, 1977, Kimmelman, 2009), but Lambrecht's definition seems to exclude them, by demanding that a cleft construction should be "headed by a copula". Russian clefts do not use a copula in the present tense, as shown in (229a), but the absence of the copula in the present tense is not restricted to Russian *clefts*, it is a characteristic of any copula construction (NP *be* XP) in that language. If languages like Russian, which do not use an overt copula verb in certain situations, are recognized as having clefts, then the cleft definition should not require the presence of a copula verb, but might better build on the presence of a copula construction. I will leave the discussion on copula constructions to section 10.1.3, and the question whether the Russian constructions are to be considered as proper *it*-clefts is taken up again in section 10.1.8.

- (229) a. Eto **ja** kupil produkty segodnya.
 this I bought groceries today
 ‘I’m the one who did the groceries today.’

The recent work of Calude (2008) gives a definition of clefts as in (230). This definition describes one particular kind of English *it*-clefts, as indicated by the use of the word “typically”.

- (230) “IT-clefts are focusing constructions, in which typically a simple sentence (though complex sentences can also be involved) is ‘cleaved’ such that the pronominal *it* appears in initial/subject position, followed by the copula *be*, the clefted constituent which expresses the highlighted or focused element, and finally, the cleft clause, modifying the clefted constituent.”
 (Calude, 2008)

This particular kind of cleft has a set word order (*it + be + clefted constituent + cleft clause*) and a set function (namely “focusing”). The word order of *it*-clefts, however, should not be stipulated, since there are other reasons why this may differ. Whenever an *it*-cleft has a *wh*-word in the clefted constituent, it is this constituent that must be fronted instead of the pronoun *it*, as in (231a). There are clefts like (231b), which have an alternative word order that may have been influenced by information structure or discourse reasons.

- (231) a. **What sort of a brooch** was it that you lost, Mrs Cheveley? [wilde-1895:600]
 b. **Dear little William**, Vicky's eldest boy, a sweet, darling, promising child, on whom my own darling doted, and who has that misfortune with his poor little left arm, it is, who is come for sea bathing and change of air.
 [victoria-186x:558]

Since *it*-clefts may have non-typical word orders, as in (231a,b), and the “focusing” function of clefts is called into question by examples such as (227e), Calude’s definition cannot serve as a general one. In fact, none of the definitions given above is generic enough to include the data in (227), (229) and (231).

The definition of the *it*-cleft should not only be inclusive enough, but it should also be able to exclude constructions that look like *it*-clefts, but are not. This is why we turn our attention to the more disputable cleft-like constructions, and then later in section 10.1.6 return to the matter of finding a proper definition for the *it*-cleft.

10.1.2 The status of adjunct *it*-clefts

The term “Adjunct clefts” refers to *it*-cleft constructions where the clefted constituent does not get a role in the cleft clause assigned by the verb, since it is an adjunct. The main argument for accepting adjunct clefts as genuine *it*-clefts is relatively straightforward. Two main components of an *it*-cleft are the cleft clause, which is a relative clause, and the clefted constituent, which is the relativized element of this relative clause. The difference between adjunct and argument *it*-clefts is in the relation between this relativized constituent and the cleft’s relative clause: the relativized constituent is either an adjunct within the relative clause or it is an argument of it. If adjunct relative clauses are acceptable, then so are adjunct *it*-

clefts. Relative clauses that have an adjunct “gap”, such as the ones in (232), certainly have been recognized as genuine relative clauses (Hukari and Levine, 1995, Schachter, 1973: 27).

- (232) a. The time [(when) he leaves] is coming near.
 b. The inconvenience remained until the middle of the 18th century [when the Parliament of Great Britain agreed to adopt the Gregorian or “New Style” calendar]. (Doherty, 2006)

Given the existence of relative clauses with an adjunct “gap”, and provided the form of the cleft clause is also accepted as a relative clause, there seems to be little reason not to accept adjunct clefts as genuine *it*-clefts.

Ball rejects time adjunct clefts in Old English, arguing that if there is “no perceptible gap in the complement, and because there is a non-cleft analysis available, there is no motivation for a cleft analysis” (Ball, 1991: 612).

Lambrecht (1994) seems to argue against adjunct clefts, since his definition (see 228) says that the “relativized argument is co-indexed with the predicative argument of the copula”, which I interpret to mean that it should be possible to trace back the clefted constituent to an *argument* position inside the cleft clause. He does not explicitly say that this *position* in the clause should be an obligatory (that is: argument) one or may be an optional (that is: adjunct) one. His point of view is in line with that of Akmajian (1979: 163), who states that “clefted sentences” should contain a variable that is specified by the “post-copular item”. In other words: not only should it be possible to trace back the clefted constituent to a position inside the cleft clause, but it must be a “variable” there, which it can only be if it is an *argument* of the main verb in the cleft clause. Neither Lambrecht, nor Akmajian provide arguments for their position in excluding adjunct clefts from the realm of genuine *it*-clefts.

Accepting adjunct clefts as genuine *it*-clefts, Jespersen (1949) included examples of clefted adjuncts such as a reason adjunct (“*it was because he was ill that he did not come*”) and a time adverbial (“*It was yesterday that he died*”). A well-known English grammar book states that adverbials of time and place may be used as clefted constituent (Quirk et al., 1985: 951). The inclusion of adjuncts as clefted constituents continues with Gundel (1977), and Prince (1978) introduces the category of “informative presupposition” clefts, most of which have “thematic scene-setting adverbials” as clefted constituents. Prince recognizes a construction like (233) as an informative presupposition cleft. The temporal PP *in this year* does not have a role assigned by the verb *accede* in the cleft clause *Yekuno Amlak ... acceded to the ... throne*.

- (233) “It was in this year that Yekuno Amlak, a local chieftain in the Ambasel area, acceded to the so-called Solomonic throne.” [Example #45 in Prince]

Declerck (1983) too accepts *it*-clefts with a time adverbial, noting that some of the time adjunct clefts accept a sentence-level adverbial (“*Today it is 5 years ago that John died*”). This has important consequences, since the presence of such an adverbial makes clefts undecleftable (“**John died 5 years ago today*”), which

means that “deleftability” cannot be used as an *it*-cleft diagnostic. (I will introduce other diagnostics in section 10.1.8.)

González-Cruz (2003) accepts *optional* positions in the cleft clause, which amounts to saying that *it*-clefts do not obligatorily have an *argument* position in the cleft clause. Hasselgård (2004) goes even one step further. Not only does she accept adjunct clefts as legitimate *it*-cleft constructions, but she sees adjunct clefts as having the “basic” function of clefts, namely that of “thematizing”: the introduction of a theme that spans a paragraph or larger discourse section. We will return to the *function* of clefts later in 10.2, but it is important for the main line in this dissertation to note the link between the adjunct status of time adverbials and the function of thematizing.

Calude’s definition in (230) sees *it*-clefts as focusing devices, which excludes most of the adjunct clefts. Patten (2010) notes that “NP-focus *it*-clefts and non-NP-*it*-clefts are instances of a single construction” (p.263), so that she is clearly in favour of including adverbials as clefted constituents.

I conclude by claiming that adjunct clefts are acceptable as *it*-clefts on the basis of the existence of adjunct relative clauses. Even though there are scholars who have come up with cleft definitions that exclude adjunct clefts, they have either failed to provide arguments for this, or they define the function of the cleft in its definition: when a definition of clefts takes as its starting point that clefts are focusing constructions, most adjunct clefts are automatically excluded. As I have argued in the beginning of this chapter, a definition of the *it*-cleft should be based on the *form* of the construction and its components alone, and it should not include references to its *function*.

10.1.3 Specification and predication

It-clefts have generally been regarded as prototypically having a specificational reading: the clefted constituent provides the value of a variable established in the cleft clause. The *it*-cleft in (224), repeated here for convenience, is a typical example of the specificational function of the cleft. The cleft clause establishes a variable *x* for which the proposition holds that “*x* makes a woman into a nurse”. The clefted constituent then supplies the value for *x*: *x* = “not technical training only”. The specificational semantics of *it*-clefts makes them ideally suited to function as answers to *wh*-questions (e.g. *Who killed John? It is the butler who killed him.*)

(224) It is **not technical training only** which makes a woman into a nurse.
[nightingale-189x:120]

However, several researchers have argued that *it*-clefts are constructions built on an copula main clause, those of the type *NP + be + XP* (Hedberg, 1990, Patten, 2010). Hedberg notes that equative constructions, which are a subset of copula constructions, can be specificational, as in (234a), or predicational, as in (234b).

(234) a. That woman is mayor of Cambridge. (both: Hedberg, 1990)
b. The only girl who helps us on Friday is Mary Gray.

I side with Hedberg (1990) and Patten (2010), who argue that, since cleft constructions in some ways are an extension of equative constructions, they can be expected to be specificational as well as predicational. Nevertheless, predicational *it*-clefts have long been subject to discussion (Declerck, 1983, Jespersen, 1927, Prince, 1978), and since not every predicational copula clause is suitable as a basis for an *it*-cleft, we too need to look at predicational constructions. The construction in (235a) is an example of a genuine *it*-cleft that is predicational.

- (235) a. It is **a long lane** that has no turning. (Jespersen, 1927: 89)
 b. A lane that has no turning is long.
 c. A long lane has no turning.

This construction can be paraphrased as in (235c), but also as the predicative one in (235b). Declerck (1983) rejects (235a) as a genuine *it*-cleft construction, arguing that the cleft clause is not a “restrictive” relative clause to “a long lane”. I agree that the semantics of the construction are closer to the predicative reformulation in (235b), since it is a predicative construction. The type of relative clause is not a decisive factor. Much more important is the syntax of the construction in (235a), which is such that the clefted constituent *a long lane* has a subject role assigned by the main verb *has* in the cleft clause. On the basis of this formal criterion, the construction should be accepted as an *it*-cleft.

Ball (1991) as well as Hedberg (1990) argue that an *it*-cleft can be interpreted predicationally when the clefted constituent, much like in (235a) above, consists of an indefinite determiner, and adjective and a noun, such as the ones in (236a,b). The indefinite determiner typically functions to denote classes, and an adjective can be used as predicator. I agree that these *it*-clefts have a predicational interpretation, but disagree that this is a deciding factor to exclude them. I regard them as genuine *it*-clefts, since the clefted constituents have a role (that of direct object in 236 a and b) in the cleft clauses.

- (236) a. It’s **a nice dress** you are wearing. (taken from Hedberg, 1990: Ch3:19)
 b. It was **a simple and uneventful life** that Schubert lived.

In sum, *it*-clefts can have a specificational or a predicational semantics, but the deciding factor for a construction to be called an *it*-cleft is, as I argue, a syntactic one: the clefted constituent has to have a role inside the cleft clause.³

10.1.4 Complements versus clefts

Constructions such as (237a) have to be disregarded as *it*-clefts, since the clefted constituent does not leave a gap (either argument or adjunct) *in* the cleft clause, but only relates to the cleft clause as a *whole*. Such constructions can be rewritten as in (237b). They have the pronoun *it* serving as a place-holder for an extraposed subject, which itself is a subordinate clause. Some refer to this construction as an “extraposition” (Ward et al., 2002).⁴ The occurrence of extraposition is not a factor by which these constructions distinguish themselves from genuine *it*-cleft, because the cleft clause itself is regarded as an extraposed one by some researchers (Patten, 2010). I will refer to it as a “complement” construction, since the difference between

a cleft and this construction boils down to the difference between a relative clause and a complement clause. A relative clause, such as (237c), has a gap (in this case a direct object gap) that coindexes with the head noun, whereas a complement clause, such as (237d), does not—the clause rather describes the content of the head noun.

- (237) a. It is **not good** [that they quarrel all day]. (complement construction)
 b. [_{Su} That they quarrel all day] is not good. (canonical form)
 c. It is **not a good example** [that he gave].
 d. It is **not a good example** [that they quarrel all day].
 e. It was **good** that he looked when I saw him last (Delahunty, 1984)

Variant (237d) does turn the clefted constituent as in (237a) into a noun phrase, but this does not change the nature of the construction. I base my decision that (237d) is not a genuine *it*-cleft on the fact that *a good example* does not coindex with an argument or adjunct gap *inside* the cleft clause, just as *good* does not do so in (237a).

The “clefted constituent” in the complement construction (237a) is an AP, but this is not something that distinguishes it from genuine clefts either. It is possible to have APs as clefted constituents, but only when they leave a gap in the cleft clause, as in (237e).

Since the complement constructions are more varied than the examples in (237) suggest, we review several more of them in (238).

- (238) a. Those who are under the impression that British forestry is a dead or dying industry have no idea as to the amount of business done in forest trees, and it is **a pity** that the Chancellor of the Exchequer and the officials of the Board of Agriculture are not better informed as to what is being done in this respect. [weathers-1913:251-2]
 b. It may **well** be that the world shall never be able to say with any certainty whether it was wise or foolish. [trollope-1882:292]
 c. Is it **so** then, that Men have no proper and genuine Good planted within them, but that they must be forced to go abroad to seek it? [boethpr-e3-p1:465]
 d. Thus it is found that oats, and beans or peas, and maize and oats, are more beneficial than either of these grains given singly, and a variation in their relative proportion, at intervals, is also strongly recommended. So it **is** that in the diet scale of large studs we often find two or three kinds of grain in the ration, in addition to the hay and straw, roots and grass. [fleming-1886:353]
 e. It is **not** that the earth has any particular attraction towards bodies which fall to it, but, that all these bodies possess an attraction, every one towards the other. [faraday-1859:137]

The constructions in (238a,b) have the *form* of a cleft (that is: *it + be + XP + RC*), but, like (237a) they are complement constructions. The clefted constituents *a pity* and *well* do not have a role in the cleft clause, but link to the clause as a whole.

There is a similar problem with the construction in (238c). If we understand *so* to be a cataphoric reference to the cleft clause as a whole, then the constituent *so* does not have a co-indexed counterpart *within* the cleft clause.⁵

The construction in (238d), where *so* functions as a discourse adverb, is an example of a frequently occurring type, which does not seem to have a clefted constituent at all. The main clause in this construction could be rephrased as *it is true that*, which makes it comparable to the negated version in (238e), which can then be rephrased as *it is not true that*. Again, these constructions are complement ones, since the clefted constituents *true* and *not true* relate to the cleft clause as a whole, and do not have a position inside it.

In sum, the deciding factor to discern the complement cleft-look-alike from genuine *it*-clefts again is the syntactic restriction: the clefted constituent has to coindex with a gap—argument or adjunct—inside the cleft clause.

10.1.5 Referential status of the pronoun

Examples (239a,b) have the same construction in the second sentence, but depending on the preceding sentence the first one (239a) is not a cleft (just as the example from the parsed English corpora in 239e), while the second one (239b) is.

- (239) a. There was someone at the door yesterday. It was **my neighbour** who had a package for me.
 b. Was that the mailman? It was **my neighbour** who had a package for me.
 c. Was that the mailman? Who had a package for me was **my neighbour**.
 d. Was that the mailman? The person who had a package for me was **my neighbour**.
 e. The gratification I yesterday received, greatly improved my opinion of this place. It is **a city**, indeed, where a reflecting mind can scarcely fail of being kept constantly awake. [montefiore-1836:79-80]

The essential difference between the construction is syntactic in nature: (239a) consists syntactically of a subject *it*, main verb *was*, and an NP complement *my neighbour who had a package for me*. The construction in (239b) has the same subject *it* and main verb *was*, but the NP complement is only *my neighbour*; the complement and the relative clause together do *not* form a constituent. The relative clause *who had a package for me* syntactically associates more with the subject *it*, and one could say that *it* in the example here is a place-holder for the extraposed subject *who had a package for me*, so that (239b) is equivalent to (239c). More generally, as has been argued for instance by Patten (2010), the subject pronoun *it* and the relative clause *who had a package for me* form a discontinuous definite constituent, and rephrasing should be done by first replacing the pronoun *it* with a generic but definite head noun (e.g. *the thing, the person, the time*) that is modified by the relative clause. So (239d) is an even better rephrasing of (239b).

There are at least three features that distinguish genuine *it*-clefts, such as (239b), from their counterfeits, such as (239a). The first feature is the difference in syntax explained above: the cleft clause either associates with the pronoun *it* or with the clefted constituent. The second feature is the anaphoricity of the pronoun *it*. In a genuine *it*-cleft, the pronoun *it* is not anaphoric, but in the counterfeit construction, it is. In the counterfeit example (239b) the *it* refers back to *someone* in the preceding sentence. The third feature is the actual focus domain.⁶ The genuine *it*-cleft restricts

the actual focus domain to the clefted constituent (that is, to *my neighbour* in 239b), whereas the counterfeit one broadens it to the whole noun phrase complement, including the relative clause (that is, to *my neighbour who had a package for me* in 239a).

The three features above are not unrelated. Restriction of the actual focus domain to the clefted constituent can only be achieved if the pronoun *it* is not anaphoric, and if the cleft clause is not syntactically part of the clefted constituent. As soon as *it* is anaphoric, the other two features (the syntactic unity of the relative clause with the clefted constituent, and the widening of the focus domain) naturally follow.

The focus domain restriction only works one way. A narrow focus on the clefted constituent yields a cleft reading, but a wider focus does *not* necessarily yield a cleft counterfeit. That narrow focus on the clefted constituent has the effect of unambiguously providing a cleft reading can be seen by forcing the focus domain to be that of the clefted constituent, for instance by adding a focus particle as in (240a), or using a negation as in (240b).⁷

- (240) a. There was someone at the door yesterday. ??It was **only my neighbour** who had a package for me.
 b. There was someone at the door yesterday. ??It wasn't **my neighbour** who had a package for me.

In none of these two examples is the second sentence a logical continuation of the first one, and I argue that this is due to the way by which *it* seeks an antecedent. Pronouns most naturally associate with a constituent that is (a) preceding, and (b) nearby. There are two candidate antecedents for *it*: (i) *someone* from the preceding sentence, and (ii) [*the person (=it)*] *who had a package for me* which is following in the same sentence, and which is not syntactically part of the clefted constituent. Apparently the “proximity” constraint is hierarchically more important than the “precedence” one, so that the antecedent of *it* becomes the cleft clause *who had a package for me*. Once the antecedent of *it* has been established, the value for the variable introduced by *someone* is not set in the immediately following discourse, which is contrary to expectations.

Adjunct *it*-clefts, as discussed in section 10.1.2, generally have focus on the cleft clause, and not on the clefted constituent. Since the focus domain of such clefts is restricted to the cleft clause, there can never be the ambiguity whether the clefted constituent should be included in the focus domain or not.

Whatever definition of the *it*-cleft one assumes, it should either have the prohibition against anaphoric *it* pronouns, or it should have the syntactic restriction. The syntactic restriction may seem closer related to the *form* of the *it*-cleft (see the discussion at the beginning of 10.1) than the prohibition against anaphoric *it* pronouns. Since the syntactic restriction leads to the anaphoricity prohibition and vice versa, I take the liberty to use the anaphoricity prohibition in the definition, which will, as we will see in section 10.1.7, lead to a relatively easy diagnostic.

10.1.6 Towards a definition

Based on the preceding discussion, this section formulates a first approximation of a definition of the *it*-cleft. The goal for this definition is to be as universal and concrete as possible. There are at least two reasons why the definition we arrive at should be universal. The first reason is that we want to use the definition for *it*-clefts within a range of different stages of English, and each stage should be considered a language in its own right. The second reason is that we would like to compare the observations made for English with those for other languages like German and Swedish, but also for Chechen.

The definition of the *it*-cleft we come up with should be as concrete as possible: the decision whether a construction is an *it*-cleft or not should be made on explicit criteria, in order to allow a maximally precise quantitative and qualitative comparison of clefts diachronically and synchronically.

A truly universal definition of the *it*-cleft should be explicit about the obligatoriness of its components, and, if necessary about the order of the components. The four components introduced in (224) are listed in (241).

- (241) 1. Cleft pronoun
 2. Copula
 3. Clefted constituent
 4. Cleft clause

The obligatoriness of the first two components listed in (241), the cleft pronoun (e.g. *it*) and the copula (a form of *be*), is complicated. The start of chapter 10 saw the *it*-less Chechen cleft in (229b), and typological research has revealed that there are more languages that have *it*-clefts, but do not use a cleft pronoun like *it* (Harries-Delisle, 1978). Old English sometimes leaves the cleft's subject unexpressed too, as illustrated in (242). The main clause in this example is *gefyrn is* '[it] is a long time ago'. In this situation the clefted constituent is syntactically a complement, and the main clause does not have an overt subject at all (see the discussion on expletive *pro*-drop in Hulk & van Kemenade (1993) and in Haerberli (2002)).

- (242) Eala, gefyrn is þæt ðurh deofol fela þinga misfor. [cowulf:1157]
 alas long-ago is that through devil many things misformed
 'Many things have been malformed through the devil a long time ago.'

Languages may have different variants of a pronoun like "it" to use in a cleft. English may use the demonstratives *that* or *this* in clefts which otherwise function the same as *it*-clefts, such as the ones in (227a,b), while a language like Russian only uses a demonstrative for its *it*-clefts, as in (229a). Demonstrative pronoun cleft subjects, like their personal pronoun counterparts, may never be anaphoric, as per the discussion in 10.1.5, since if they were, the cleft clause would be a restrictive relative clause under the clefted constituent, and the construction would cease to be a genuine cleft.

As for the obligatoriness of the second component in (241), a copula (a form of the verb *be* in English), it is Russian we should look at (Gundel, 1977). The example in (229a) does not contain an overt copula, but this is because the language as such

does not use an overt copula in the present tense. What is vital, following the discussion in 10.1.1 on Russian, is not so much the presence of a copula verb, but the presence of a copula *construction*. The form of such a construction can vary from language to language. Russian does not use an overt copula verb in a copula construction in the present tense, but it does in the past. English always requires the presence of a verb in a copula construction, irrespective of the tense that is used.

The variety in pronoun use and in the presence of *be* for *it*-clefts can best be captured by defining *it*-clefts as constructions that have a copula construction as their main component. If a language requires the presence of a subject in a copula construction, then the language has to have an overt subject in an *it*-cleft. If the language requires the presence of the verb *be* in some form, then it has to be present in such a form in the *it*-cleft too. The particular tense, mood or aspect of *be* can vary, as we have seen in examples (227a,c), and by basing the cleft definition on the copula clause, we put the burden of stating what is and what is not required in terms of tense, mood and aspect to the definition of the copula construction, where it should be.

It goes without saying that the third and fourth component listed in (241), the clefted constituent and the cleft clause, are obligatory elements of a cleft construction.

As for word order within an *it*-cleft construction, this should not be a stipulation for inclusion or exclusion of clefts, since language specific factors or information ordering factors may determine the particular word order of the cleft's constituents. The definition of a cleft should therefore not stipulate any word order.

To recapture the universality requirements of *it*-clefts, I conclude that (a) a good definition of the *it*-cleft construction requires the main structure of a cleft to be that of a copula clause, and (b) does not stipulate a particular word order.

Sections 10.1.2-10.1.5 revealed that there are several objective requirements that distinguish a genuine *it*-cleft from other constructions. These requirements boil down to the two stated in (243). The requirement (243a) states that the clefted constituent should have an argument or adjunct role in the cleft clause. This makes it clear that adjunct clefts are allowed (see 10.1.2), and that predicational clefts with the correct semantics are allowed too (see 10.1.3), but complement constructions are not to be regarded as *it*-clefts (see 10.1.4). The requirement (243b) helps to objectively disambiguate *it*-clefts from cleft look-alikes that have exactly the same surface form but differ with respect to the constituent that the cleft clause associates with (see 10.1.5).

(243) *It-cleft requirements*

- a. The clefted constituent should have an argument or adjunct role in the cleft clause.
- b. If the cleft has an overt subject (e.g. a pronoun like *it*), it may not be anaphoric.

The definition of the *it*-cleft in (244) builds on the existing one in Lambrecht (2001), which I have introduced in section 10.1.1, and on the insights offered by Hedberg (1990) and Patten (2010). It furthermore incorporates the two universality

requirements derived above, and it contains the two objective *it*-cleft requirements in (243), which derive from the comparison between cleft and alternative constructions in sections 10.1.2-10.1.5.

(244) *Definition of an it-cleft*

An *it*-cleft construction is a complex sentence structure consisting of (a) a copula matrix clause whose subject, if overtly expressed, is semantically empty and non-anaphoric, and (b) a relative clause whose relativized argument or adjunct is coindexed with the predicative argument of the matrix clause.

The first universality requirement—the obligatoriness of the components—is met by the fact that the definition states that the basic building block of an *it*-cleft is a “copula matrix clause”. The second universality requirement states that word order does not matter, and this is met, because the definition does not stipulate any word order.

The requirement in (243a) is met since the definition explicitly states that there may be a relativized *argument* or *adjunct* that is coindexed with the predicative argument of the main clause (that is: the complement of the equative matrix clause). The requirement in (243b) is met by stating that if there is a subject, it may not be anaphoric. The definition also states that if there is a subject it has to be “semantically empty”. This means that an *it*-cleft may not have a *lexical* subject; only a neuter personal pronoun like *it* can be used, or the more generic demonstrative pronouns like *this* and *that*.⁸

The next sections derive three clear-cut and concrete diagnostics based on the definition in (244), which are then shown to help discern clefts from non-cleft constructions.

10.1.7 Cleft diagnostics

There are a few diagnostics that have been used to see if a construction is an *it*-cleft. One diagnostic is that of “decleftability”, as for instance formulated in Lambrecht’s (1994) cleft definition in (228): “... the matrix and the relative express a logically simple proposition, which can also be expressed in the form of a single clause without a change in truth conditions”. However, we have seen in section 10.1.2 that cleft constructions allow a sentence-level adverb to be present, and when they do so, they cannot be felicitously declefted. It is for this reason that we cannot use decleftability as a fair *it*-cleft diagnostic.

Calude (2008) provides a test to see if a particular construction is a genuine *it*-cleft or a complement construction (which he calls an “extraposition”). His test says that if the pronoun *it* can be replaced by the cleft clause without any further changes, the construction is not an *it*-cleft. However, this test is not always able to capture non-cleft complement constructions, witness example (238e), which is repeated here for convenience.

- (238) e. It is **not** that the earth has any particular attraction towards bodies which fall to it, but, that all these bodies possess an attraction, every one towards the other. [faraday-1859:137]
- f. *That the earth has any particular attraction towards bodies which fall to it is not.

The construction in (238e) was identified as a non-cleft complement in 10.1.4, since “not” modifies the relative clause as a whole instead of being coindexed with an adjunct or argument gap in it. Nevertheless, the pronoun *it* cannot be simply replaced by the relative clause, as per the unacceptability of (238f).

The definition of the *it*-cleft in (244) serves as a starting point to formulate several diagnostic tests that can be used to check if a given construction is a cleft or not. The following three diagnostics are necessary and sufficient for any construction to be called a cleft, as I will show.

- (245) *Cleft structure*
The clause containing a cleft construction must consist of a copula construction and a “cleft clause”: a subordinate clause that has the form of a relative clause.
- (246) *Cleft pronoun*
The subject of the clause containing a cleft construction can be a pronoun or it can be empty, but it may never be anaphoric.
- (247) *Cleft coindexing*
The relativized argument or adjunct of the cleft clause must coindex with the clefted constituent.

The *Cleft structure* diagnostic in (245) ensures that the global structure of the construction is in place. It also makes sure that cleft look-alikes with different verbs are rejected, such as *it happened in 1994 that I met this lady*. The diagnostic does not dictate the *form* of a copula construction. Present-day English has it as: subject + *be* + complement. But other languages may differ in how they express copula constructions.

The *Cleft pronoun* diagnostic in (246) requires that the cleft’s subject is non-anaphoric. This ensures that copula clauses with a complex complement, where the relative clause in the cleft is a restrictive relative of the predicative argument, are excluded (see section 10.1.5). An entirely other matter are *cataphoric* personal or demonstrative pronouns. There is no restriction on them.

The *Cleft coindexing* requirement in (247) ensures that the clefted constituent has a role inside the cleft clause—either as argument or as adjunct. It rejects look-alike constructions such as *it is well that you have come to me* and *it should not be that I have to introduce you*. Such constructions certainly are close to clefts, but by the definition in (244) they miss the vital link between the clefted constituent and the gap in the cleft clause.

10.1.8 Testing the cleft diagnostics

We have come up with a definition of the *it*-cleft and several accompanying diagnostics. This section briefly shows how these can be used, by verifying examples from English and a few other languages. We start by testing the diagnostics on the English examples in (248).

- (248) a. It **is the butler** who did it.
 b. It **should have been the butler** who did it.
 c. Do you really want it **to be the butler** who did it?
 d. It is **definitely expected** that you make your own coffee.
 e. Do you know what_i I found in my bag?
 It_i is **the necklace** that you had lost.
 f. **How many years** is it that you have studied Russian?

The examples in examples (248a-c) can be accepted without much of a problem, since they fulfil all three diagnostics: they have the structure of a cleft, the pronoun *it* is not anaphoric, and the clefted constituents have a role in the cleft clause. They do have different forms of *be* in the main clause, but neither the definition in (244), nor the diagnostic in (245) require a particular form of *be*, as long as the construction is a copula one.

Example (248d) has, from a cursory glance, the appearance of a cleft, since it consists of *it* + *be* + XP + RC (just as the first *it*-cleft example 224). Nevertheless, it is excluded on the basis of the *Cleft coindexing* diagnostic (247). The clefted constituent *definitely expected* is not coindexed with the relativized constituent. In fact, there is no relativized constituent, there only is a complement clause *that you make your own coffee*.

The second part in (248e) also has the outward appearance of a cleft, but should be rejected—this time on the basis of the *Cleft pronoun* diagnostic. This diagnostic says that the subject must be “non-anaphoric”. In the example this is not the case, because *it* refers back to the constituent *what I found in my bag* in the preceding sentence. The definition also states that the relative clause may not be a restrictive one of the predicative argument, but *that you had lost* is in fact a restrictive relative clause to *the necklace*.

Example (248f) is fully acceptable *it*-cleft by the definition in (244). The main clause is a copula construction *it is X years*, the subject *it* is not anaphoric, and the relativized constituent *how many years* coindexes with a temporal adjunct position in the cleft clause *you have studied Russian [for how many years]*.

Having established the robustness of the cleft definition in (244) with respect to English, we should now see if the definition is universal enough, and the diagnostics restrictive enough to exclude false clefts from other languages, while including *it*-clefts.

- (249) Okno razbil Vasja? — Njet, eto Petja razbil okno. (Kimmelman, 2009)
 window broke Vasja no that Petja broke window
 ‘Did Vasja break the window? No it is Petja who broke the window.’

- (250) Cwajtta sho du cuo hoqu shkoliehw buolx bo. [p86-00034.5]
 eleven year is she this at.school work does
 ‘She has worked at this school for eleven years.’

Example (249) is a Russian *it*-cleft. The main clause *eto Petja* ‘that [is] Petja’ is a copula construction in Russian (even though it lacks an explicit form of *be*), so that the first part of the *Cleft Structure* diagnostic in (245) is complied with. The second part of this diagnostic, however, requires the presence of a relative clause, and this is not immediately confirmed by the data. The construction fares well on the other diagnostics. It has a non-anaphoric pronoun, complying with (246), and the clefted constituent *Petja* fulfils a subject role in the cleft clause, complying with (247). In sum, the acceptability of the Russian *it*-cleft depends on evidence for cleft clauses like *razbil okno* to be a relative clause.⁹

The Chechen example in (250) too is accepted by the definition of the *it*-cleft. Even though the main clause *cwajtta sho du* ‘[it] is eleven years’ does not contain an *it*-like pronoun, there still is an empty subject—in this case the subject is left unexpressed, which vacuously meets the *Cleft pronoun* diagnostic. The main clause is acceptable as a copula construction in Chechen, for instance as one that answers the question *Hoqu sholiehw cuo buolx binarg maca sho du?* ‘How many years has she worked in this school?’ The *Cleft coindexing* diagnostic passes too, since the clefted constituent coindexes with an adjunct time constituent of the relative clause *cuo hoqu shkoliehw buolx bo* ‘that she works at this school’.

With the definition of the *it*-cleft’s form firmly in place, we can now concentrate on its function.

10.2 The function of clefts

There are several different suggestions for the function of *it*-clefts. Some researchers see them as having an obligatory or optional (disambiguating) role at the local level (as related to the syntax of the sentence and the local information structure rules), while others recognize them as having a function at discourse level. And, in fact, both could be true at the same time. This section looks at some of the hypotheses and observations of other researchers, and tries to differentiate the functions of clefts at the local level from those that relate to a discourse level.

10.2.1 Obligatory clefts

There are languages where clefts represent a strategy to convey a particular meaning, which cannot be expressed otherwise. This is usually the result of conflicting rules in a language at the “local” level—the level that relates to the clause and its immediately preceding or following context.

Lambrecht (1994) reports extensively on French as having two such potentially conflicting rules, and I will only briefly repeat his arguments here as an illustration. Syntax requires French word order to be SV, while phonology requires that focus is marked by a pitch-accent, in compliance with the “focus-prominence” principle (Truckenbrodt, 1995). It also requires that a pitch-accent is assigned to the right edge of a phonological phrase. Focus on objects or adjuncts can be expressed by

making sure the constituents are at the right edge of a phonological phrase, but this does not work for subjects. The normal strategy to focus the subject of a sentence like (251), therefore, is to use an *être* cleft construction, such as in (252).¹⁰ The cleft provides a way to demote the grammatical status of the logical subject *ma voiture* to that of a complement, while at the same time placing it in a right-aligned IP position, where it can receive a pitch-accent in a natural way, which is then interpreted as the focus.

- (251) Ma voiture est en panne. (Lambrecht, 1994: 22ex. 1.3')
 my car is in breakdown
 'My car has broken down.'
- (252) C'est ma VOITURE qui est en panne. (Lambrecht, 1994: 223 ex. 5.11)
 it is my car that is in breakdown
 'My CAR has broken down.'

Present-day English does not have the same conflict between syntactic and phonological rules as French. Since Old English is only available in written documents, we do not know enough about its prosody (intonation and stress), so we are not able to say anything about a similar conflict in that language.

Another area where using clefts can be a strategy to resolve a conflict is that of negation. Komen (2010) shows that Chechen needs to resort to *wh*-clefts to express sentence negation when a sentence contains an element triggering negative concord.

- (253) a. Cwa a ciga *(ca) vyedu.
 No one there NEG go
 'No one goes there.'
- b. Sobien *(ca) vyedu ciga.
 I except NEG go there
 'Only I go there.'
- c. Sociga ca vyedu.
 I there NEG go
 'I don't go there.'
- d. Sobien vaac ciga ca vyedurg.
 I except am.not there NEG going.one
 'Only I am not going there.'

The appearance of a negative expression like *cwa a* 'no one' requires the presence of a sentence negator *ca* 'NEG', as shown in (253a). The same negative concord effect is reached by the word *bien* 'only', witness (253b). The negator *ca* 'NEG' can only occur once within a clause, and if it occurs, it either functions to express negative concord (253a,b) or as a sentence negator (253c). The combination of sentence negation and negative concord requires two sentence negators, which can only be done in a *wh*-cleft such as in (253d), which is a biclausal construction.

Lambrecht (2001) shows that English is required to use a cleft construction when a combination of sentence and constituent negation needs to be expressed.

- (254) a. *I do not like no chocolate.
 b. It isn't chocolate I don't like.

Given a context where one is forced to admit that there are several things one does not like, such as mustard, raw fish etc, a speaker may want to say that there is an exception to this list of disliked items. A monoclausal construction such as (254a) does not work, because English too, just like Chechen, does not allow two negators in the context of one clause. Since an *it*-cleft is a biclausal construction, it allows one negator to occur in each of its clauses, so that the double negation is expressible as in (254b).¹¹

10.2.2 Clefts for focus

Several scholars have pursued the idea that the function of *it*-clefts is related to “focus”. I will argue in chapter 12 that one of the two main functions of clefts is associated with focus, but only with one particular kind: constituent focus.

With this in mind, I would like to review how others have seen the relation between *it*-clefts and focus. Jespersen, who was the one coining the term “cleft”, states this idea as follows:

- (255) “A cleaving of a sentence by means of *it is* (often followed by a relative pronoun or connective) serves to single out one particular element of the sentence and very often, by directing attention to it and bringing it, as it were, into focus, to mark a contrast.” (Jespersen and Haislund, 1949: 147)

Quirk’s English grammar (Quirk et al., 1985: 951-953) identifies the main function of the cleft as that of “focus” on the clefted constituent, comparable to the function of adverbs like *too* and *only*. While Jespersen and Quirk concentrate on the cleft’s ability to express *constituent focus*, the definition in (226) suggests that the primary focus of the cleft is on the element that introduces *new information*, be it the clefted constituent or some part of the cleft clause. Declerck (1983) states that his predecessors generally consider the function of clefts to be that of introducing a clefted constituent which combines new information focus and constituent focus.¹²

Very much in line with Jespersen, Kiss (1998) proposes that English *it*-clefts behave like the Hungarian preverbal position, in that the clefted constituent contains exhaustive identification—a particular kind of focus. Identificational focus results, for example, by answering the *what* question in (256a) with *a hat*. But this answer does not exclude the possibility that Mary also picked something else for herself. Exhaustive identification, on the other hand, results in (256b), which does exclude the possibility of Mary picking something else for herself. The clefted constituent now contains an exhaustive list of all the elements for which the predicate *Mary picked x for herself* holds.

- (256) a. A: What did Mary pick for herself?
 B: She picked a hat for herself.
 b. It was **a hat** that Mary picked for herself. (Kiss, 1998: 249 ex. 8a)

There are a number of problems with associating exhaustive identification with the clefted constituent in English. Wedgwood, Pethő and Cann (2006), helped by observations from Horn (1981), argue against Kiss’s exhaustivity diagnostics. The first diagnostic is the incompatibility of a universal quantifier (like *every*) to appear

in the clefted constituent, but Wedgwood has a counter example, which I have included in (257a). Exhaustivity also does not work for “not only” *it*-clefts such as *It was not only John who started to sing*.¹³

- (257) a. It’s **every child** that got frightened, not just the girls. (Wedgwood et al., 2006)
 b. It wasn’t **only John** who started to sing.

While Jespersen seems to hint at *constituent* focus here, specifically mentioning “contrast”, other researchers have looked into the function of cleft constructions in relation to the information states of its components: the clefted constituent and the cleft clause. The generalization that *it*-clefts present *new information* in the clefted constituent and *given information* in the cleft clause does not hold, since, as noted by many, certain *it*-clefts have old information in the clefted constituent, and new information in the cleft clause, while other *it*-clefts contain new information in the clefted constituent as well as in the cleft clause, and are used to introduce a chapter or even a whole book. The *it*-clefts with new information in the cleft clause have long ago been noticed, and have been labelled as “informative-presupposition” clefts and “comment-clause” clefts (Hedberg, 1990, Prince, 1978).

Declerck (1983) took Prince’s idea of dividing clefts on the basis of the information states of their components one step further, and he came up with three different basic types. His first type, called “contrastive clefts”, are distinguished by having a cleft clause that contains given information in the sense that it “pursues the thematic line of the stretch of discourse in which it is couched” (Declerck, 1984: 264). The information status of the *clefted constituent* does not matter for this type of cleft—it can link back to the preceding context (in which case it represents a ‘continuous’ topic) or not (then it is a ‘discontinuous’ topic). The contrastive reading results from the stress on the clefted constituent, and the stress on this first part of the construction results from the givenness of the second part, the cleft clause.

The second type distinguished by Declerck is the “unstressed-anaphoric-focus cleft”, which is characterized by new information in the cleft clause, and given/anaphoric information in the clefted constituent. The stress in such clefts is on the cleft clause instead of on the (given) clefted constituent. The last type Declerck distinguishes is the “discontinuous cleft”, which has a new clefted constituent as well as a new cleft clause. Such clefts can open a discourse section.

Another attempt at determining the function of clefts is given by Delin (1990). She argues that it is not the information status of the clefted constituent or of the cleft clause that should be new in a cleft, but the *relation* between the cleft clause and the variable being instantiated in the cleft clause, which is very much in line with Lambrecht’s (1994) account of focus. Delin does not tell how this account works for adjunct clefts.

A comparison with Present-day German may be in order at this point. Ahlemeyer & Kohlhof (1999) looked at how English *it*-clefts are translated into German. They found that, although German is able to use *it*-clefts, it tries to avoid them when translating English. The reason for this, as they argue, is that, while English clefts are used to mark focus unambiguously, German is able to do this with less marked methods—by means of word order and particles.

- (258) a. It is **these properties** that make them attractive as anticancer agents.
- b. Gerade diese Eigenschaften lassen sie als Wirkstoffe gegen Krebs
 precisely these characteristics let them as agents against cancer
 vielversprechend erscheinen. (Ahlemeyer and Kholhof, 1999: ex. 17)
 highly.promising appear
 ‘Precisely these properties let them appear highly promising as agents
 against cancer.’
- c. The fact that one can get away with this is one of the beauties of molecular
 biology, and it is **this beauty** that we are celebrating here.
 (Ahlemeyer and Kholhof, 1999: ex. 16)
- d. ... und diese Schönheit wollen wir hier zelebrieren.
 and this beauty want we here to-celebrate

Ahlemeyer and Kholhof do not regard a proper translation of the English *it*-cleft in example (258a) to be a German *it*-cleft, but would rather use a focus particle (here *gerade* ‘precisely’), and have the constituent in the first position (the German “Vorfeld”), as in (258b). Even without using a focus particle, placement in the first position can be used in German, as illustrated in (258c), which is rendered as (258d) in German.

The observation that word order and particles in one language can achieve the same effect as *it*-clefts do in other languages is an interesting one, and something to keep in mind, since English started out as a Germanic language, and so Old English might, at least to some extent, be comparable to Present-day Germanic languages. It is especially interesting to see that a non-contrastive constituent focus device such as the focus adverb ‘precisely’, which we saw at work in chapter 9, correlates with the use of an *it*-cleft in Present-day English, which we will come to in chapter 12.

10.2.3 Clefts as an avoidance strategy

Lambrecht (2001) introduces an extensive theory on the analysis of cleft constructions in general, and argues that the function of clefts is related to focus in the following way:

- (259) “Cleft constructions are **focus-marking devices** used to prevent unintended predicate-focus construal of a proposition. Clefts serve to mark as focal an argument that might otherwise be construed as nonfocal, or as nonfocal a predicate that might otherwise be construed as focal, or both.”
 (Lambrecht, 2001)

So in Lambrecht’s view the function of clefts very much depends on what the pragmatic interpretation of a constituent would have been, if it had not been clefted. The idea of linking unmarked form with unmarked meaning and a marked form with a marked meaning is intuitively attractive, and has been successfully applied in several different areas, most notably in bidirectional OT modelling (Blutner et al., 2006). Whether Lambrecht’s idea holds for the *it*-clefts as they started to appear in the English language remains to be shown. Comparative research between English on the one hand and Swedish and Norwegian on the other hand shows that the Scandinavian languages have a strong tendency to use clefts as a strategy to keep

referentially “new” information out of the syntactic subject position (Gundel, 2002, Hasselgård, 2004, Johansson, 2001). Example (260a), which is from a Norwegian novel, is used by Gundel to show this point. The logical subject *Sofies farmor*, ‘Sophie’s grandmother’, has moved out of the *main* clause’s subject position into the cleft clause. Gundel argues that the reason for this may be the fact that the ‘grandmother’ is new information, hence needs to be stressed, but Norwegian tends to keep focal material out of the main clause subject position. This tendency is, according to Gundel, confirmed by the “strong preference against indefinite subjects”.

- (260) a. (Etter hvert som Sofie tenkte over at hun var til, kom hun også til å tenke på at hun ikke skulle være her bestandig. Jeg er i verden nå, tenkte hun. Men en dag er jeg borte vekk. Var det noe liv etter døden? Også dette spørsmålet var nok katten helt uvitende om.) (Gundel, 2002: ex. 19)
 Det var **ikke så lenge** siden Sofies farmor døde.
 that was NEG so long since Sophie’s grandmother died
 ‘(Later, when Sophie thought about her being here, she realized that she would not be here always. “I am in this world now”, she thought, “but one day I’ll be gone.” Was there life after death? This was another question the cat was probably quite unaware of.)
 It wasn’t LONG ago that Sophie’s GRANDMOTHER had died.

While acknowledging clefts as an avoidance strategy, the same authors also note that the main function of clefts in Swedish and Norwegian seems to be that of “thematizing”, which is defined as the discourse function of dividing the text into thematically organized segments. This function will be discussed in section 10.2.5.

Hasselgård (2004) introduces one important situation where *it*-clefts are used as an avoidance strategy in English. This happens when the clefted constituent is introduced with *not until*, such as in (261a).

- (261) a. However it wasn’t **until his fourth album** that the instrument’s capabilities were more fully explored. (Hasselgård, 2004: ex. 12, 12a)
 b. Not until his fourth album were the instrument’s capabilities more fully explored.

The problem with a non-clefted variant, such as (261b), is that this needs subject-auxiliary inversion due to the negation in the clefted constituent. Hasselgård concludes that the *it*-cleft may be a way of using a marked construction (the *it*-cleft) in order to avoid one that is even more marked (subject-auxiliary inversion).

I would argue for an alternative explanation of Hasselgård’s observations, which is related to the decline of V2. Subject-auxiliary inversion is a typical V2 phenomenon, and it is the decline of V2 in the history of English that makes (261b) an option that is more marked than the *it*-cleft in (261a).

It seems fair to conclude that “avoidance strategy” is at least one of the functions for which *it*-clefts are used. The idea of a hierarchy between marked constructions such as the cleft, which retains unmarked word order, and the subject-auxiliary

inversion, which introduced a marked word order, certainly is one that deserves further attention, which it will receive in chapter 13.¹⁴

10.2.4 Clefts to introduce presupposition

One of the first to introduce the term “presupposition” in relation to the cleft construction was Chomsky (1971), who argued that clefts can be divided into a *focus* and a *presupposition*. As Schachter (1973) explains, a presupposition is “a proposition that must be true in order for the (current) sentence to have a truth value”. Gundel (1977) divides the parts of the cleft in “topic” and “comment”, where the topic is the given or presupposed information, and the comment is the new information. Prince (1978) noted that not all *it*-clefts were of the “Stressed Focus” type, those introducing new information in the clefted constituent (or in the relation between the clefted constituent and the cleft clause). She described the type of “Informative Presupposition” clefts which distinguish themselves by having a cleft clause, the part of the cleft that until then had been labelled the “presupposition”, that contains new information.

Prince argued that the information in the cleft clause is encoded as presupposed in the sense that it is a non-negotiable fact. She posited the idea that speakers might be tempted to use this property of the cleft construction in order to introduce new information in such a way that the reader or hearer naturally accommodates it as a fact—i.e. as a rhetorical device. One of the functions of such a construction, then, is to “mark a piece of information as a fact, known to some people, although not yet known to the intended hearer” (Prince, 1978: 899-900). It is this function, in the opinion of Prince, that makes Informative Presupposition clefts suitable for use in historical narrative, since the author distances himself implicitly from the truth of the information packaged in the cleft clause.

Informative Presupposition clefts can be more persuasive, when they state an opinion as a fact in the cleft clause, as the one in (262a), or they can be more factual, as the one in (262b).

- (262) a. It is through these conquests that *the peasantry became absorbed into a single form of dependent lord-tenant relationship*. (Prince, 1978: example 44a)
 b. It was in this year that *Yekuna Amlak, a local chieftain in the Amba-Sel area, acceded to the so-called Solomonic throne*. (Prince, 1978: example 45)

While Prince’s reasoning is straightforward, and her examples illustrative of the point she is trying to make, she admits that presenting new information in the cleft clause as a known fact is but one of the functions of the *it*-cleft.

Patten (2010: 278-279) sees a historical development of using clefts to “state an opinion under the guise of a presupposition”.

10.2.5 Clefts as a discourse strategy

Hedberg (1988) looked into the discourse functions of different kinds of clefts, concluding from her preliminary study that, while all clefts function to separate the “topic” (i.e. the content of the clefted constituent) from the “comment” (the cleft

clause), the *it*-cleft's function is that of expressing contrast, the *wh*-cleft's function that of signalling the opening of discourse segment, and the reversed *wh*-cleft's function is that of signalling the closing of a discourse segment.

Her studies were preliminary, and based on limited data (only 12 *it*-clefts). Her dissertation (Hedberg, 1990) is based on more data (701 *it*-cleft tokens) and also seeks to identify the discourse functions of clefts. She distinguishes two basically different cleft types. The "Topic-clause" cleft is a construction that, in a sense, is 'about' the information in the cleft clause, while the "Comment-clause" cleft is one which is 'about' the information in the clefted constituent. It is only to the latter type of cleft that she ascribes a function in the discourse. Discourse-initial clefts can be used to "anchor" something in history, as in (263a), which is the first line of a background story for a TV news special report. The "Comment-clause" clefts can be used to link discourse segments, such as (263b), and they also occur in discourse-final positions, such as (263c), where they can serve to draw a final conclusion that is tied in with the preceding material.

- (263) a. It was **the death of a Chinese leader five weeks ago** that gave birth to the student movement. ...Hu Yaobang... (Hedberg, 1990: Example 88)
- b. It was **at this point** that their conversation was interrupted by Mr. Quirk. How long he might have been listening to them was not apparent; he moved softly over the grass... (Hedberg, 1990: Example 106)
- c. Nearly all the extant artifacts date from the nineteenth century. Earlier examples have decayed...From the 1800s we also have the first-hand account of native customs made by observers before white influences caused many changes.
It is **this period** which accordingly gives us the best picture of the culture and society of the northwest coast Indians. (Hedberg, 1990: Example 114)

Johansson (2002) discerns four different functions of clefts, most of which are related to discourse: contrast, topic linking, topic launching and summative. The function of "contrast" is clear from 10.2.2, and the functions of "topic launching", "topic linking" and "summative" seem to coincide with Hedberg's "discourse initial" clefts, the discourse linking ones and the discourse-final clefts, as exemplified in (263a-c).

New in Johansson's work is that he tentatively relates these four discourse functions of the *it*-clefts to the information states of the clefted constituent and the cleft clause in the way exemplified by Table 36. Evenhuis (2006) suggests separating the "Contrast" function from the discourse functions, since "Contrast" may combine with any of the functions "Topic linking", "Topic launching" and "Summative".

Table 36 Johansson's discourse functions of clefts related to information states

Discourse function	Clefted constituent	Cleft clause
Contrast	Old or New	Old
Topic linking	Old	New
Topic launching	New	New
Summative	Old	Old

We can understand Johansson's ideas about the relation between *it*-clefts, information structure and discourse better by looking at some specific examples, which are all taken from the British component of the International corpus of English (Hasselgård, 2004, ICE-GB, 2011). **Topic launching** happens in (264a), where the clefted constituent *those men and women* is discourse new, and the idea in the cleft clause that the speaker is *thinking about them* is new too. The newly introduced referents are taken up as topic in the following context by *you* and *our servicemen and women*.

- (264) a. (We must try to work out security arrangement for the future so that these terrible events are never repeated, and we shall promise you <, > bring our own forces back home just as soon as it is safe to do so.)
 It is **to those men and women serving our country in the Middle East** that my thoughts go out most tonight, and to all of their families here at home.
 (To you I know this is not a distant war. It is a close and ever present anxiety. I was privileged to meet many of our servicemen and women in the Gulf last week.) [ICE-GB S2B-030 #63-68:1:A]
- b. (C: But really what's happened with my sort of history is when I met uh did a little recording with Chandos Records uhm and the Ulster orchestra who was conducting there came up with enough money to do their first record and they got Chandos interested.)
 It was **then** that uh I fell in love with music like Hamilton Harty and a bit of Stanford.
 (And the Arn – the Arnold Bax Saga became something quite uh excellent.
 A: Well that's a day we certainly want to come back to a bit later. But if we could just for a moment concentrate on the latter years of the nineteenth century.) [ICE-GB S2B-023 #61:3:A]
- c. (I struggled terribly with them in my early teens and had no success at all.)
 It wasn't **till I was perhaps twenty-five or thirty** that I read them and enjoyed them. [ICE-GB S1A-013 #2370238:1:E]
- d. (The purpose of war is to enforce international law. It is to uphold the rights of nations to be independent and of people to live without fear.)
 It is **in that spirit** that the men and women of our forces and our allies are going to win the war. And it is **in that spirit** that we must build the peace that follows. [ICE-GB S2B-030 #103-105]

Speaker "C" in example (264b) tries to shift the topic (a function that is referred to as **Topic launching** in Table 36) from a particular period in music history to a particular kind of music, which speaker "A" recognizes, and he tries to shift back to

the topic he is interested in. Crucial for the topic shift is that the clefted constituent links back through *then* to the point in time discussed in the preceding context, while the new topic is introduced in the cleft clause as *music like HH*, which is discourse-new.

The discourse function of **Contrast** to something that has been mentioned previously requires a discourse-old cleft clause, as in (264c), where *I read them* refers to the fact that the speaker has been reading certain books in the past, something that is also implied by the preceding context of *I struggled with them in my early teens*. The clefted constituent refers to the speaker at the age of 35, which he contrasts with himself when he was in his teens.

The **Summative** function in (264d) is reached by having discourse-old information *that spirit*, which refers to the previous sentence, in the clefted constituent. Unlike the link to information structure suggested by Table 36, however, the information in the cleft clause is discourse new.

Hasselgård (2004) follows up Johansson's (2002) idea's on part of the ICE-GB that contained 51 adjunct clefts, and found that the link between information structure and discourse suggested in Table 36 is only a tendency, not a strict one. Hasselgård extends Johansson's ideas by adding the discourse function of "thematization", which she defines as "making extra clear what the theme and the rheme of a sentence are". Hasselgård's example in (265a) constitutes a complete one-line text, so that the function of the *it*-cleft cannot be one of topic-linking, topic-launching or summation, nor can it be contrast with an element in the preceding or following context. Hasselgård notes that the clefted constituent receives a kind of thematic prominence, which, in her opinion, it would not receive in a non-cleft version of (265b). A quick search on Google, however, reveals that the adjunct "With much regret" can, in fact, be used at the start of a discourse, witness example (265c), as well as in the middle of discourse, as in (265d).¹⁵

- (265) a. It is **with much regret** that I find it necessary to send you a copy of the enclosed letter which is self explanatory. (Hasselgård, 2004: ex. 11)
 b. ? *With much regret*, I find it necessary to send you a copy of the enclosed letter which is self explanatory. (Hasselgård, 2004: ex. 11a)
 c. *With much regret*, I'm putting my Birdy Elux up for auction on Ebay. (anonymous)
 d. Were I today to deliver an Inaugural Address to the people of the United States, I could not limit my comments on world affairs to one paragraph. *With much regret* I should be compelled to devote the greater part to world affairs. (Roosevelt, 1936)

Hasselgård, carefully avoiding making an actual hypothesis, plays with the idea that the "basic function" of *it*-clefts is "thematization", and that the other functions (contrast, topic-launching, topic-transition and summative) derive from it.

I agree with the conclusions of Hedberg, Johansson, Evenhuis and Hasselgård to the point that at least adjunct clefts are used in discourse functions. Hasselgård's ICE-GB study shows that 44 of the 51 adjunct *it*-clefts introduce new information in the cleft clause, which is an ideal configuration to either launch a topic in the cleft

clause against the adjunct frame in the clefted constituent, or to transition from one topic (embedded in the clefted constituent) to a new one (in the cleft clause). A biclausal structure such as the *it*-cleft forces the reader to slow down at an important point of transition.

10.2.6 Conclusions

We have seen that *it*-clefts can be used as a *local* level strategy to express a meaning which would otherwise not be possible, such as focus on the subject in French, or negation on more than one constituent within the sentence (see section 10.2.1). The *it*-cleft can also be used at the local level as an avoidance strategy (see 10.2.3), for instance to prevent subject-auxiliary inversion to happen. One of the reasons for this may be that this inversion came to be perceived as more marked than a cleft construction at some point in time. I will argue in chapter 13 that the reason for this is the decline of V2.

Due to the inherently presuppositional character of the information in the cleft clause, *it*-clefts can also be used to introduce new information as factual. In doing so, they function as a rhetorical device.

Other functions of the cleft seem to relate more to the *discourse* level. I agree with Hasselgård and others that grouping them under the banner of “thematization” makes a lot of sense: the syntax of an *it*-cleft allows singling out virtually any kind of constituent as thematic, while the cleft clause serves to embed it further in the narration.¹⁶ The *it*-cleft is a construction that can be used to launch a topic, while it is anchored in some other (perhaps generally known) event, it can be used to make a smooth transition from one topic (expressed in the clefted constituent) to a new topic (expressed in the cleft clause), and it can serve as a summative, at the end of a stretch of discourse. All of these thematization functions can be combined with contrast between the clefted constituent and an element in the context.

Since thematization (rather than expressing contrast or constituent focus) seems to be the more basic function in Germanic languages like Swedish and Norwegian, and English started out as a Germanic language, it makes sense to hypothesize that *it*-clefts in English started out historically as thematizing devices (having the functions of topic-launching, topic-transition and summation), and only later grew into its current role as the prototypical construction to express constituent focus. This is the line of thought that will be borne out by the data discussed in chapter 12, section 12.3, but before we go there, we will make a detour to Chechen, a totally unrelated Caucasian language, and see what we can learn from it.

¹ The assumption that a language “needs” to express constituent focus may not hold everywhere, since constituent focus is not always needed to express a phenomenon that is otherwise underivable. Take for instance the constituent focus resulting from the resolution of an open variable. If “John is the murderer” answers the question “Who killed Mary?”, then the focus domain undoubtedly is the subject constituent “John”, but from a communicative point of view there is no “need” to mark this constituent linguistically—either by prosody, morphology or word order: the communication situation already gives enough clues for the addressee to understand that “John” is the value supplied for the open variable.

² In the context of *it*-clefts, I will use the term “copula clauses” to refer to those with the verb *be*. I do not take into account copula clauses with other copula verbs.

³ I use the term “role”, in view of the preceding section on Adjunct clefts, in a wide manner. A role can be an argument role or an adjunct role.

⁴ I use the term “extraposition” not necessarily to indicate movement has taken place, but out of convention.

⁵ If *so* is understood as coindexing with an adjunct “in this way” in the cleft clause, then (238c) should be accepted as a genuine *it*-cleft.

⁶ I am using the term “Actual focus domain” in a particular situation to distinguish it from the “Potential focus domain” for a particular syntactic construction. These terms have been introduced by VanValin for his Role and Reference Grammar (van Valin, 2005).

⁷ The focus particle *only*, in the sense of “exhaustively”, should not be confused with the sentence-level adverbial *only*, in the sense of “just”. The latter reading of (240a) would yield a non-cleft.

⁸ I leave the question open, whether the syntactic subject pronouns *it*, *this* or *that* of the *it*-clefts have a *cataphoric* referent. In relatively simple argument *it*-clefts like “It was John who met Mary” the relative clause *who met Mary* can easily be viewed as a (free relative) NP in itself, and serve as cataphoric referent of *it*, but this is not the case in non-argument *it*-clefts—not when the clefted constituent is an NP, like “It was March that I visited my uncle”, and certainly not when the clefted constituent is an adjunct, like “It was to help you, that I have come here.”

⁹ Gundel (1977) has an example of a slightly different Russian construction (*Eto Ivan kogo ja videl*), which contains the relative pronoun *kogo* ‘whom’, so that the relative clause status of the cleft clause is clear. But the constructions mentioned above do not have this feature.

¹⁰ This is not to say that *every* cleft in French signals marked focus on one particular constituent—that is: argument focus. Lambrecht (1994) argues that this latter kind of focus can be expressed by the *être* cleft, while sentence focus, where the whole sentence consists of new information, can be expressed by the *avoir* cleft.

¹¹ There is a strategy by which English speakers can avoid using a cleft construction, namely that of using a verb that has a negative meaning. Lambrecht uses the example of replacing *does not like* by *dislikes*. So instead of using the cleft strategy as in (254b), one can use a non-clefted clause like *I don't dislike chocolate*. However, such double negatives usually have the effect of resulting in a strong positive meaning, such as *I very much like chocolate*. The same goes for the combination *not without* in a sentence like *He is not without faults, you know*. This is understood as saying that the person we are talking about is *full* of faults.

¹² He states that the “focus” (the clefted constituent) “contains new information and is heavily stressed and contrastive”.

¹³ Wedgwood et al. (2006) show that the second diagnostic, which has to do with accent based focus, does not necessarily lead to exhaustivity either.

¹⁴ This idea seems particularly applicable to a (bidirectional) optimality theory approach, since it allows hierarchical ordering of constraints. The “cost” of subject-auxiliary inversion is, in terms of generative grammar, the “I-to-C movement”: movement of the verb, which receives its finiteness specification at the I-head, to adjoin to the C-head.

¹⁵ There are many examples of “it is with much regret” on the internet too.

¹⁶ But more research is needed to actually see if this grouping is borne out by data from different languages.

The previous chapter has shown that several researchers regard the *it*-cleft construction fundamentally as a focusing device (Jespersen and Haislund, 1949, Kiss, 1998), whereas recent work on Scandinavian languages claims their role in discourse segmentation is an even more fundamental function (Hasselgård, 2004). Scandinavian languages still use the *it*-cleft partly to express constituent focus, but I claim that there is at least one language that uses *it*-clefts only for discourse segmentation, and not for constituent focus at all. The language with this interesting property is Chechen, a North-East Caucasian language.

This claim is important within the framework of the research on focus in English described in this book: if there are languages that have *it*-clefts and don't use them for focusing, then the function of English *it*-clefts too may not right from the start have to be limited to that of focusing.

Section 11.1 briefly repeats Komen's (2007b) claim that Chechen mainly uses word order to signal focus, while a secondary, but related, device to signal focus in Chechen is the *wh*-cleft. The need for Chechen to use syntactic or morphological methods to convey focus is consolidated by a closer look at Chechen intonation in section 11.2. The conclusion there is that focused constituents in Chechen are not distinguishable from unfocused ones on the basis of intonation. Given the already available methods of word order and *wh*-clefting to convey focus in Chechen, it is remarkable that the language still has an *it*-cleft construction, as discussed in section 11.3. Closer inspection reveals that the clefted constituent in the *it*-cleft always is a time adjunct, and that the function of the cleft always is related to discourse organisation. Section 11.4 draws the logical conclusion that the *it*-cleft construction as such is not necessarily related to focus. These findings motivate me to not combine the construction with the function of focusing automatically, when I consider the history of the English *it*-cleft in chapter 12.

11.1 Focus in Chechen

The central claim of this chapter is, that Chechen has *it*-clefts, but does not use them to convey constituent focus. If this holds up, then the question stands how Chechen *does* express constituent focus. Typological research on the languages of Europe as well as some Asian languages has shown that there is a high likelihood (but not a necessary implicature) for SOV languages to have an immediately preverbal position reserved for focus (Sornicola, 2006: 380). Komen (2007b) claims that Chechen is one of the SOV languages that does have an immediately preverbal position for focus, while it also uses *wh*-clefts for constituent focus. We will briefly review the arguments made by Komen (2007b) in this section.

The arguments start with an observation about the position of *wh*-question words. Table 37 shows the actually occurring word orders found in a collection of 86 sentences containing one *wh*-question word.¹

Table 37 Word orders in sentences containing *wh*-question words

#	Word order					Number	
	Pre	Wh	Between	V _{fin}	Post		
i	X	Arg	(Neg)	V	(X)	39	(45%)
ii	X	Poss	(Neg)	V		1	(1%)
iii	X	wh	(Neg)	V	(X)	27	(31%)
iv		Arg	(Neg)	V	X	10	(12%)
v		PP	(Neg)	V	X	1	(1%)
vi		wh	(Neg)	V	X	8	(9%)

All *wh*-question word constituents (whether argument “Arg”, possessive “Poss”, postpositional phrase “PP” or independent *wh* constituent “wh”) appear immediately preceding the sentence’s finite verb (which is either the lexical verb or, if an Aux is present, the Aux, and sometimes a combination of them). The only element that can intervene between a question word and the finite verb appears to be a negator. Table 37 also shows that constituents (of diverse type, marked by “X”) can precede the *wh* constituent as well as follow the finite verb.

Since *wh*-question words often (but not necessarily—e.g. when the question is “how” or “why”) indicate the presence of constituent focus, the observation that *wh*-question words must immediately precede the finite verb leads to the hypothesis that there is an immediately preverbal focus position. In order to verify this claim, I did extensive fieldwork for my MA research (Komen, 2007b). I elicited a “paradigm” of question and correction focus. A focus paradigm is a set of different sentence types with focus on varying constituents (Büring, 2005). This particular focus paradigm distinguishes 5 different sentence types: an intransitive one and four transitive ones, which differ in the grammatical case used for the subject and in the mood of the sentence.

As part of my fieldwork, I elicited question-answer focus as well as corrective focus for each of these sentence types on the subject, the object (where applicable), the whole sentence, the verb on its own, and any adjuncts (where applicable). In order to illustrate the focus paradigm strategy, and the results achieved, (266) lists some of the elicitations for the “dative-subject transitive sentences”.²

- (266) a. Joqqa-baaba gira Denina. [Komen2007:C]³
 grandmother saw Danny.DAT
 ‘Danny saw grandmother.’
- b. Hun xilla? Denina joqqa-baaba gira. [Komen2007:C.3.C.3.i]
 what happened Danny.DAT grandmother saw
 ‘What happened? Danny saw grandmother.’

- c. Joqqa-baaba hwaanna gira? [Komen2007:C.1,C.1.i]
 grandmother who.DAT saw
 Joqqa-baaba Denina gira.
 grandmother Danny.DAT saw
 'Who saw grandmother? DANNY saw grandmother.'
- d. Suuna myettariehw joqqa-baaba gira Muusana.
 me.DAT in.thinking grandmother saw Musa.DAT
 Haan-haa, joqqa-baaba Denina gira. [Komen2007:C.5,C.5.i]
 no grandmother Danny.DAT saw
 'I thought that Musa saw grandmother? No, it was Danny that saw grandmother.'
- e. Suuna myettariehw voqqa-daada gira Denina. [Komen2007:C.6,C.6.i]
 me.DAT in.thinking grandfather saw Danny.DAT
 Haan-haa, Denina ginarg joqqa-baaba jara.
 no Danny.DAT who.saw grandmother was
 'I thought that Danny saw grandfather? No, it was grandmother who Danny saw.'

The native speaker is given the information in the basic sentence, which is the one in (266a). The subject is in the dative case, because this is required by the psych-verb 'see'.⁴

The "What happened" question in (266b) aims to elicit the unmarked word order. The word order of the answer given by the native speaker indeed coincides with what we know about Chechen: the result is SOV. The answer to the subject question 'who' in (266c) yields the OSV word order, where we know that focus is on the subject.

The focus paradigm also makes use of correction focus, as exemplified in (266d). The incorrect subject "Musa" is corrected into "Danny" in the response of the native speaker. What we have here is contrastive focus. The resulting word order is again OSV.

Sometimes focus is conveyed in the answer by a *wh*-cleft, such as in (266e). The direct object of the sentence, 'grandfather', needs to be corrected by the native speaker, and he does so by answering with the *wh*-cleft 'Who Danny saw was grandmother'. The resulting word order of this *wh*-cleft is: Subject (*Denina ginarg* 'Who Danny saw'), followed by the complement (*joqqa-baaba* 'grandmother') and then by the finite verb (*jara* 'was'). A subject-complement-verb word order is a very natural one for SOV languages. The result here is that the logical direct object of the answer (that is: *joqqa-baaba*) ends up in an immediately preverbal position (that is: *joqqa-baaba ju* 'grandmother is') rather than the possibly confusing SOV word order as in (266b), which can be used both to signal focus on the direct object as well as focus on the whole sentence.

The detailed results of the focus paradigm elicited by Komen (2007b) will not be repeated here, but the data above should be enough to illustrate the conclusion that Chechen uses the immediately preverbal position to convey focus, and that it may use a *wh*-cleft as an alternative.

We have seen in this section that Chechen uses word order, if necessary in combination with a particular construction (the *wh*-cleft), to express focus. Since

this is so, the question arises whether prosody would still be used as a—perhaps secondary—device in the expression of focus. If prosody is used as a primary device, as in English, we may find that any constituent anywhere may be focused, given we pronounce it with the correct intonation pattern. Prosody could play a role as a secondary device, such as has been suggested for French, where focus is associated with a pitch-accent, and this pitch-accent normally occurs clause-finally (Lambrecht, 1994). In such a situation, the desire to emphasize a constituent that would not normally end up clause-finally motivates alternative word orders as well as the use of a construction such as the cleft. If prosody works similarly in Chechen, we would need to carefully investigate the possibility that a combination of prosody and focus lead to the use of *it*-clefts. If, on the other hand, prosody does not play a role in Chechen focus, the picture becomes much more straightforward. The influence of prosody on focus in Chechen is addressed in the next section.

11.2 Chechen intonation

This section seeks to explore the role of prosody in the expression of focus in Chechen. We know that there are languages such as French, where the need to use cleft constructions is associated with prosodic requirements, as explained above (Lambrecht, 1994). This is why we need to find out whether Chechen has similar requirements, but we will see that it does not.

Nichols' (1997) description of Chechen phonology touches on several matters of tone and intonation, pointing out that certain clitics and suffixes have an inherent high pitch, which suggest the presence of lexical tone in Chechen, and she suggests that there are intonational domains inside which downstep occurs. The study in this section will identify these domains, but the matter of downstep is left for further research. Nichols furthermore notices a kind of tonal (or: intonational) pattern: non-final clauses may end with low or high pitch, but final clauses end with a low pitch. Komen (2007a) too argues for lexical tone on certain suffixes and clitics, while Komen (2007b) reported that narrow focus in Chechen is reserved for the constituent immediately preceding the finite verb with a intonational demarcation of the part of the clause preceding the focus. The description of the Chechen intonation system in this section loosely follows the study done by Gussenhoven and Komen (forthcoming).

What we do in this section is review the intonation grammar—the set of rules used to produce boundary tones and pitch-accents—of one particular dialect: Shali Chechen (henceforth abbreviated as SC). This is done with the help of recordings from two native speakers, which were made between 1993 and 2010. The recordings of one speaker, amounting to a total of 236 sentences, have been stored in an annotated database. The framework used for the intonation grammar is derived from Gussenhoven (2004), and the software used to investigate the audio recordings of the utterances is “Praat” (Boersma and Weenink, 2005, Boersma and Weenink, 2008).

The SC dialect distinguishes two hierarchical levels of phrases: intonational phrases (11.2.1) and accentual phrases (11.2.2). Accentual phrases may contain at

most one pitch-accent. This normally is a default pitch-accent, unless there is a function word or morpheme with lexical tone (11.2.3). Against the background of the intonation rules of these first three sections, section 11.2.4 arrives at the most important conclusion for this chapter on Chechen: SC does not have separate intonation rules for the expression of focus.

11.2.1 Intonational phrases

SC aligns finite clauses with intonational phrases, and these intonational phrases invariably start and end with a low pitch. (I will refer to these intonational phrases with the abbreviation “InP”, in order to avoid confusion with the syntactic phrase “IP”.) The utterance in (267) demonstrates this feature: it consists of two finite clauses, and each finite clause fits into exactly one intonational phrase (indicated by square brackets), while each intonational phrase consists of a number of accentual phrases (indicated by the round brackets).⁵

(267) [(Hwalx^a) (teptarsh dyesh^ush) (var^a iz^a)],
 L_aH* L_aH* L_aH* L_i
 earlier BOOKS reading was he
 [(tq^a hinc^a iza) (cwa kiex^atash dyesh^ush vu)].
 L_a H* L H_a L_a H* L_i
 but now he some LETTERS reading is
 ‘He used to be reading BOOKS, but now he is reading SOME LETTERS.’

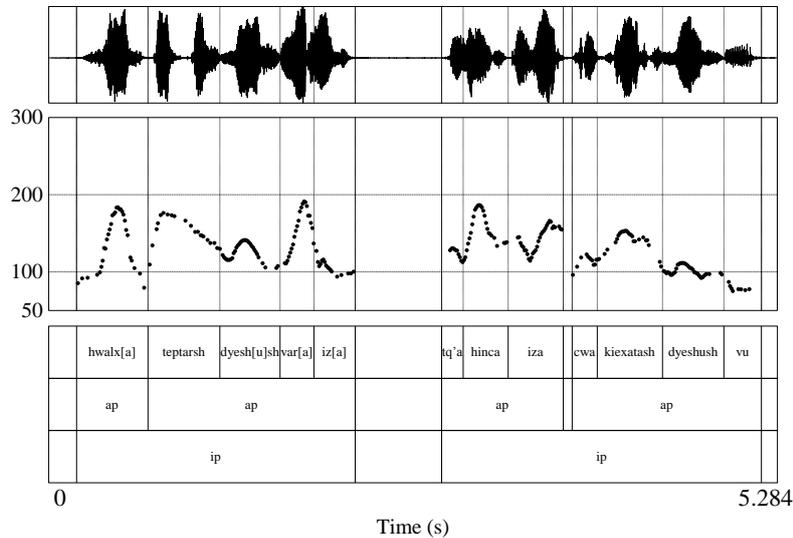


Figure 28 Pitch track and speech waveforms of an utterance of (267)

Gussenhoven (2004) shows that intonational phrases in general can be demarcated by a left and right boundary tone, and the question is whether the starting and ending low pitch observed with the intonational phrases in SC are to be seen as these boundary tones. What I argue is that the right low boundary tone belongs to the

intonational phrase pattern, and I will subsequently refer to it as L_i , where “L” denotes a low tone, and the index i indicates that it is part of the intonational phrase. The f_0 pitch track produced with Praat shows the effect of the final L_i in the first InP on *iz^a* ‘he’, and in the second InP on *vu* ‘is’.

The database that has been gathered for the intonation research shows that L_i is *not* permutable with H_i so that SC does not have a high tone at the end of an intonational phrase. In fact, there is neither a H_i to mark pre-final intonational phrases in the utterance nor is there an interrogative H_i , both of which are widespread intonational morphemes in the languages of the world (Gussenhoven, 2004: 85, 89). The SC L_i tone, then, is an obligatory marker of the end of any intonational phrase. These findings confirm Nichols’ (1997) observation on low-ending sentence-final clauses in Ingush, but in Chechen low-ending intonational phrases also occur sentence-medially, as in (267). The question where the starting low pitch of the InP comes from is next on the agenda.

11.2.2 Accentual phrases

The intonational phrases introduced in the previous chapter divide into smaller chunks, which are called “accentual phrases”, which I will abbreviate as AcP. Accentual phrases invariably start with a low pitch, and I interpret this as SC accentual phrases having a L_a left boundary tone. It is this left *ap* boundary tone that obviates the necessity to have a left L_i boundary tone, answering the question that was raised at the end of the last section.

The basic characteristics of the AcP derive from the example in (267). In the first AcP in (267), the rise of the rise-fall on *hwalx^a* is due to L_a plus a following H^* . The fall after *hwalx^a* anticipates the L_a of the next AcP, which also begins with an H^* -accented syllable, *tep*. From here, the pitch falls to *dyes^hsh*, due to initial L_a in the third AcP *var^a iz^a* ‘he was’. The third AcP has H^* on *var^a*, which is followed by L_i .⁶

The second InP illustrates a high-ending AcP. It begins with low-pitched *tq^a*, due to L_a , has H^* on *hinc^a*, while H_a on the last vowel of the topical (!) pronoun *iza* closes off the AcP. The dip between H^* and H_a in the first AcP is attributed to a L -tone that intervenes between them. Since H_a invariably occurs with a preceding L -tone and because within the InP no other sequences of identical tones can arise, given the proposed analysis, I assume that the OCP is respected within the IP (Goldsmith, 1979). Finally, the last AcP begins with a L_a on the word *cwa*, which is the number ‘one’ used as an indefinite marker, while H^* occurs on the first syllable of *kiex^atash* ‘letters’, from where the pitch slowly falls to the final L_i .

The detailed timing of the inserted L -tone that intervenes between H^* and H_a of one AcP, is not apparent either from (267) or Figure 28, because the distance between the two peaks is too short. In cases where this distance is larger, the L -tone tends to occur just before the peak on the right. Its rightward alignment is illustrated in Figure 29, a pronunciation of (268), in which two successive AcPs end in H_a . The second AcP in particular shows that the alignment of the L following on the H^* is rightmost, creating a slow fall over the stretch between the accented syllable and the

final syllable. A similar slow fall can be seen in the last AP of (267), where it is due to H* and L_i.

(268) (Shien laetta t'e) (aaxarxuochuo) (hu tosur^a).
 L_a H* L H_a L_aH* L H_a L_aH L_i
 his land onto farmer.ERG seed throw-IMPF
 "A farmer was sowing his land."

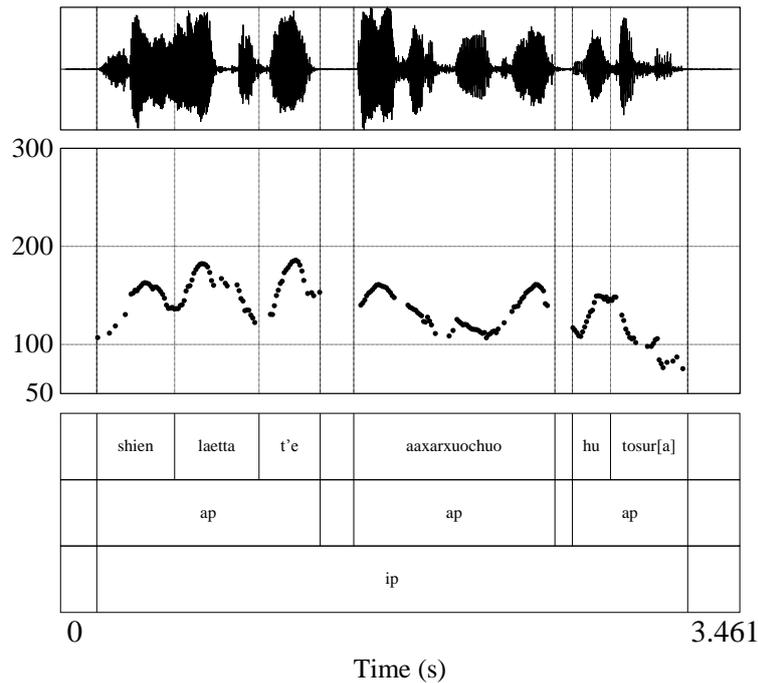


Figure 29 Pitch track and speech waveform of an utterance of (268)

In (269), I give the basic rules of the intonation grammar I arrive at for SC. Rule (269a) gives the contour of the intonational phrase, which is characterized by the obligatory right L_i tone. Rule (269b) gives the obligatory H* pitch accent in the domain of the AcP, plus the tone structure of the AcP, which begins with L_a and ends either in H_a or in no specified tone (indicated by the Ø). The last rule (269c) gives the generalization that leads to the insertion of the L-tone between H* and H_a, on the assumption that no H-tones are deleted.

(269) SC intonation rules

- a. InP: [...]L_i
- b. AcP: L_a(... H* ...) $\left\{ \begin{matrix} H_a \\ \emptyset \end{matrix} \right\}$
- c. OCP within InP

While the grammar in (269) is very simple, it correctly reflects the limited options for tonal variation that are available in SC: the choice of an AcP ending in H_a or ending in nothing. Nevertheless, there is one complication by which SC differs from most European languages, which have more complicated intonation grammars, and this is the availability of lexical tone. Once we have considered the presence and influence of lexical tone, we come back to the main question that underlies this excursion into intonation: what role does prosody play in the realization of Chechen focus?

11.2.3 Lexical tone

In the previous section we have seen that major class words will receive a pitch accent whenever they occur as the first such word in the AP. Function morphemes, i.e. functional categories like function words, clitics and affixes) come in two kinds. A number of them are unaccentable. These include personal pronouns, definite markers and preverbs. Preverbs are comparable to the particle in English phrasal verbs, like *away* in *take away*. Examples of Chechen preverbs are *dwa* ‘away’, *swa* ‘hither’, *uohwa* ‘down’, *hwala* ‘up’ etc. The other class of function morphemes comes with its own pitch accent (Komen, 2007a, Nichols, 1997). This lexical pitch accent is H*, just like the default pitch accent. Inclusion of a H*-accented function word precludes the assignment of a default pitch accent to the AcP, regardless of the position of the function word in the AcP. In (270), I list all the accented function words I have come across (there are probably more of them).

(270) *Accented function words, clitics and suffixes (not a full list)*

<i>ma</i>	negative imperative particle
= <i>a</i>	intensification clitic
= <i>a</i>	conjunction clitic
- <i>iehwa</i>	Polite imperative suffix
- <i>i/ii</i>	Polar question suffix
<i>mila</i>	Question word ‘who’
<i>masa</i>	Question word ‘how many’
<i>maca</i>	Question word ‘when’
<i>michahw</i>	Question word ‘where’
<i>hun</i>	Question word ‘what’
<i>daac</i>	negative present tense auxiliary (variants: <i>vaac</i> , <i>jaac</i> , <i>baac</i>) (when used in polar questions without question marking suffix)

Example (271) illustrates lexical tone on the negative imperative marker *ma*. The leftmost major class word, *biexk* ‘guilt’, appears without the default pitch accent. Example (272) illustrates a lexical pitch accent on the polite imperative suffix *iehwa*, which shows that suffixes can be accented in preference to their host. Verb roots such as *hwaarch* ‘wind’ are lexical categories that would normally be eligible to receive a default pitch accent. In this example *hwaarch-* does not receive the pitch accent, but the polite request suffix *-iehwa* ‘PLEASE’ does.

(271) (Biexk ma bill^alahw!) → lexical tone on negative imperative *ma*
 L_a H* L_i
 guilt NEG put-SG
 “Excuse me!”

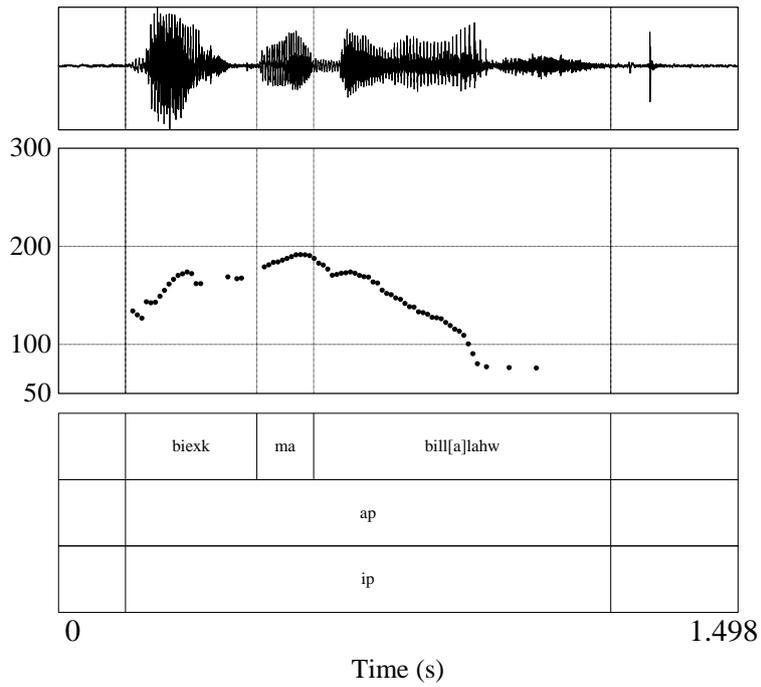


Figure 30 Pitch track and speech waveform of an utterance of (271)

- (272) (Dwaa-hwaarch-iehwa ysh suuna) → lexical tone on suffix *-iehwa*
 L_a H^* L_i
 away-wind-PLSE them for.me
 “Please wrap them up for me!”

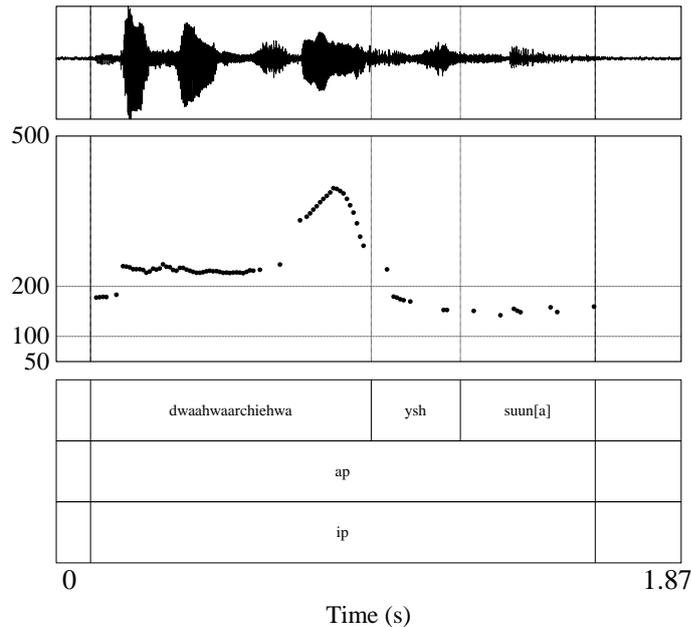


Figure 31 Pitch track and speech waveform of an utterance of (272)

The question marker clitic *-j* (realized as *-ii* when attached to a consonant) has lexical tone too, as illustrated by example (273). Default pitch assignment skips the leftmost accentable *buolx* ‘work’, which is a noun.

- (273) (Buolx biesh vu-j hwo?) → lexical tone on QM clitic *-j*
 L_a H^* L_i
 work doing are-QM you
 “Are you working?”

I have not found a lexical tone on any lexical categories. It seems probable that there are more functional categories with lexical tone, and possibly, too, the functional categories with lexical tone differ from dialect to dialect. More research is needed to validate this.

As stated in the intonation grammar (269), SC does not have a separate interrogative intonation contour. Differences in the pitch contour between declaratives and interrogatives are, however, tonally signalled by the absence of the default pitch accent and the presence of a lexical pitch accent. Frequently, these locations are different. The H^* on the polar question morpheme */-j/*, the polite imperative suffix */-iehwa/* and the clitic */=a/* occurs on or near the right edge of the word it attaches to, and this word, just as any function word with a lexical H^* , can

occur in any position with its AP. However, when default H* and lexical H* occur in the same location, there is no difference in the tonal representation. Examples (274) and (275) are a case in point. Their morpho-syntactic structures have been kept comparable by replacing the noun *hu* ‘seed’ with the wh-word *hun* ‘what’. As a result, there is no difference in the phrasal structure, with both sentences consisting of a single IP that divides into three APs. The fact that the second AP in (274) is accent type A, while that in (275) is accent type B is not significant, in view of the variability in the tonal realization of APs. In my database there are other cases of APs with a question word that are preceded by Type B APs. A final point here is that there is a phonetic difference between the realizations of H* in (274) and (275), since the pitch on the WH-word is higher. This difference may well be systematic, and if it is, it is not exclusive for interrogative sentences. While a thorough investigation needs more data, a random selection of ten IP-final APs in my database shows that there is a substantial difference between the f0 of the lexical H* and the default H*. The f0 interval between the final L_i and the peak of H* was 67 Hz in the case of a default accent and 105 Hz in the case of a lexical accent. Since default accents occurred somewhat earlier in the IP than lexical accents, the difference cannot be attributed to declination.

(274) (Shien laetta t'e) (aaxarxuochuo) (hu tesir^a).
 L_a H*L H_a L_aH* L H_a L_aH* L_i
 his land onto farmer.ERG seed threw
 “A farmer sowed his land.”

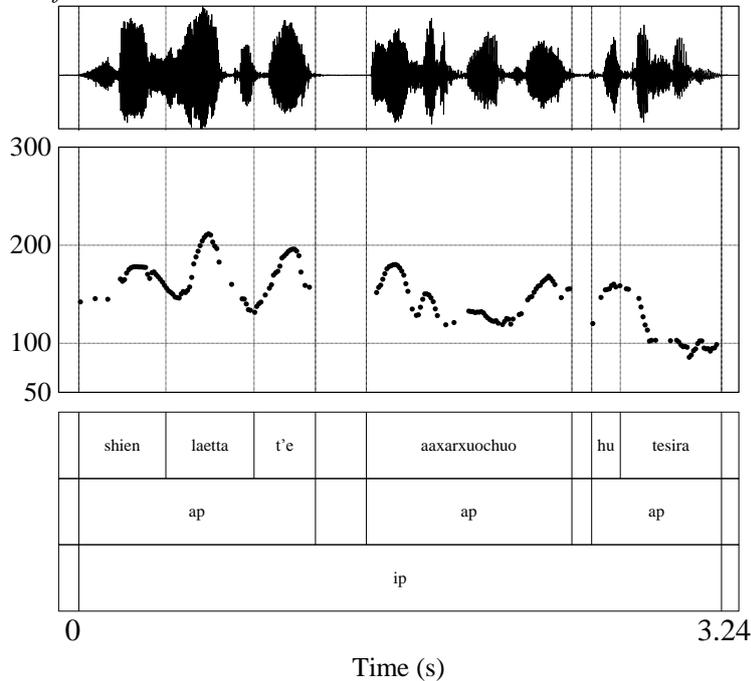


Figure 32 Pitch track and speech waveform of an utterance of (274)

- (275) (Shien laetta t'e) (aaxarxuochuo) (hun tesir^a)?
 L_a H*L H_a L_aH* L H_a L_aH* L_i
 his land onto farmer.ERG what threw
 "What did the farmer throw on his land?"

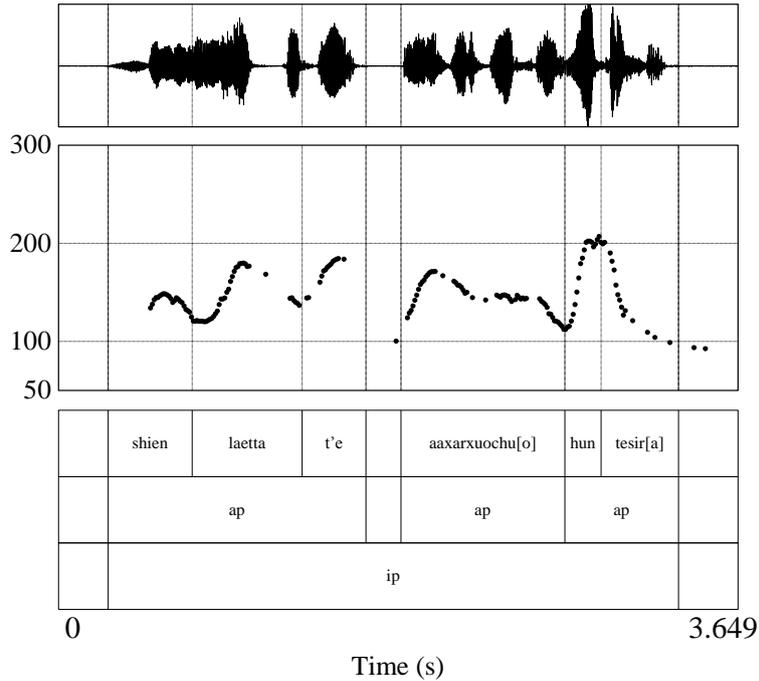


Figure 33 Pitch track and speech waveform of an utterance of (275)

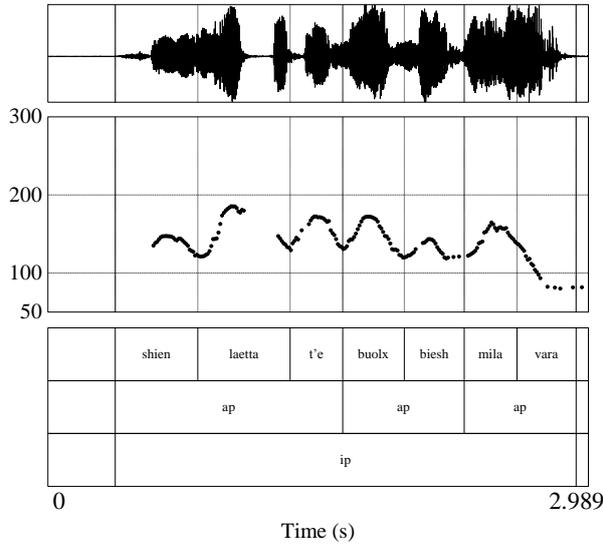
Structurally there is no difference between the intonation patterns of (274) and (275). There is an audible difference in that the last AP in the interrogative mood has lexical tone, which phonetically is a bit higher than the default H* pitch accent assigned in the last AP of the declarative mood. The amplitude difference, then, does not need to be attributed to a prosodic effect of the presence of focus within the question constituent *hun* 'what', but can be explained adequately enough by the amplitude difference between lexical tone and default H* pitch accents.

11.2.4 Intonation and focus

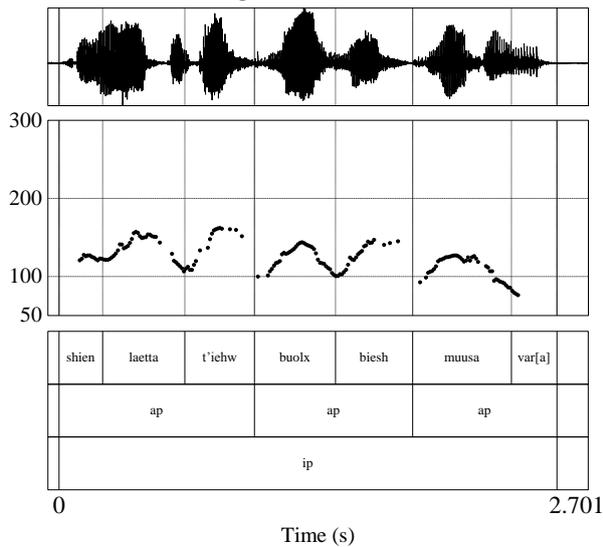
The generalization for Chechen is that a focused constituent is placed immediately before the finite verb, but there are a few exceptions (see Komen, 2007b: 29 for some of these): (a) binominative periphrastic tenses have focused *subjects* precede the tensed auxiliary, but they have focused *objects* precede the participle part of the compound tense, and (b) negation can intervene between the focused constituent and the finite verb, (c) while *wh* adverbials ('when', 'where' etc) consistently appear immediately before the finite verb, focused adverbial *phrases* do not always do so, for reasons that need further investigation.

The regular behaviour, then, for focused subjects is to appear immediately before the finite lexical verb or the tensed auxiliary, which is illustrated in (277), which has narrow focus on *Muusa* ‘Musa’ and is an answer to (276).

- (276) (Shien laetta t'e) (buolx biesh) (mila var^a)?
 L_a H^*L H_a L_aH^* L_a L_aH^* L_i
 his land onto work doing who was
 “Who was working on his land?”



- (277) (Shien laetta t'iehw) (buolx biesh) (Muusa var^a).
 L_a H^*L H_a L_aH^*L H_a L_aH^* L_i
 his land onto work doing MUSA was
 “MUSA was working on his land”



The purpose of this section has been to demonstrate that focus has no dedicated tonal feature and that focus expression takes place through word order. Specifically, the alignment constraint in (278) applies for narrow focus, which ensures that the right edge of the focus constituent ends where the finite form begins.

(278) Align (FOCUS, Right, V_{fin} , Left)

The interaction of (278) with the default syntax may lead to focus ambiguities as well as focus distinctions that are unexpected from the point of view of European languages. The paradigm collected by Komen (2007b) and the database used for this article suggest that when the focus constituent is confined to a smaller constituent *within* the NP, the position of the entire NP is determined by whether the focus constituent co-terminates with the NP, as determined by (278). Example (279) is an answer to the question “What is Musa doing today?”, while (280) answers the NP-internal question “What good thing is Musa doing today?”.

(279) [(Muusa) (cwa dika buolx biesh vu)]
 $L_a H^* L H_a L_a$ H^* L_i
 Musa some good work doing is
 ‘Musa is doing some GOOD WORK today.’

(280) [(Muusa) (cwa dika buolx biesh vu taxan^a)]
 $L_a H^* L H_a L_a$ H^* $L H_a L_i$
 Musa some good work doing is today
 ‘Musa is doing some good WORK today.’

The narrow focus for ‘some good work’ is thus structurally ambiguous with the narrow focus for ‘work’. The NP *cwa dika buolx* ‘some good work’ that represents (as in 279) or contains (as in 280) the narrowly focused constituent does *not* immediately precede the finite verb in the word orders given above, following the explanation above for the placement of focused objects in “bi-nominative” constructions (both *Muusa* and *buolx* ‘work’ are in the nominative case). While Chechen’s ergative nature is generally displayed in the morphological case of transitive verbs’ subjects, it allows for a bi-nominative construction in periphrastic tenses, but this is at the cost of a reduced word order flexibility: the direct object may only directly precede the (predicative) participle (which is the form *biesh* ‘doing’ in (279) above). This word order rule outweighs the rule for a narrowly focused constituent to immediately precede the finite (tensed) lexical verb or copula, so that the word orders as in (279) and (280) result.

What these examples show, then, is rule (278) at work: the constituent that *is* or that *contains* the focus right aligns (as close as possible allowed by overriding syntax rules) to the finite verb.

In sum, we have seen that SC does not use intonation to treat focused constituents in any different way than unfocused ones. The results for SC are, strictly speaking, not directly transferrable to the Chechen language as a whole: we would need to know the intonation grammar of all other Chechen dialects if we could make such a generalisation. However, given the results we have seen here, I expect other Chechen dialects to differ in details of the intonation grammar, such as

the number and use of the different default pitch-accents, and the particular boundary tones used for the InPs and AcPs, but I do not expect other Chechen dialects to make a principled difference between focused and unfocused constituents. More research is obviously needed to verify this claim.

11.3 Chechen *it*-clefts

We have seen that Chechen does not use intonation to signal focus, but relies on word order and a syntactic device such as the *wh*-cleft. Nevertheless, Chechen has a construction which I claim is an *it*-cleft. I make this claim, since, as we will see in section 11.3.1, the construction complies with the definition of the *it*-cleft developed in chapter 10. We will find out how we can locate these *it*-clefts (11.3.2), where they occur (11.3.3) and what they are used for (11.3.4).

11.3.1 The Chechen *it*-cleft construction

We start our research into Chechen *it*-clefts by verifying that the constructions claimed to be clefts conform to the definition developed in chapter 10. The *it*-cleft definition (244) states that an *it*-cleft first of all consists of a copula construction. An example of a copula construction is given in (281), where the bracketed (V) and (J) indicate the noun classes of the nominals, while the J is the agreeing noun-class prefix on the copula (see note 7).

- (281) Apti kuotam ju. (Full equative clause)
 Apti(V) chicken(J) J.PRS
 'Apti is a chicken.'

Copula constructions in general consist of the elements $NP_{subj} + be + XP$: there is a subject (NP_{subj}), which is combined with a complement (an XP) through the mediacy of a form of the copula verb "to be". Since Chechen is an SOV language, the normal word order for a copula construction is Subject – Complement – *be*, which is exactly what we find in example (281). This example also illustrates the noun-class agreement that is at work.⁷ The present tense form of *be* agrees in noun class with the NP complement, which is not surprising, since Chechen is a morphologically ergative language.⁸

Chechen allows for equative clauses (copula clauses with an NP complement), such as in (282), where the subject in the second sentence is implied from the first sentence: *the main weapon* is language. Such sentences are not really subject-less. The fact that there is no subject visible in the sentence signals to the reader that the subject of the first sentence continues to be operative.

- (282) Mylxa du vajn literaturan kyerta gerz? Muott bu-q. [p34-00002:117,120]
 which is our literature's main weapon language is-INT
 'What is the main weapon of our literature? It is language.'

An example of a truly subject-less copula clause is (283). The structure of the main clause is roughly: complement – *be* – when-clause. The logical subject of the sentence is the when-clause, which can be paraphrased as: *the occasion when a home is restored*. This *when*-clause cannot function as a subject grammatically, and

instead of transforming it to a grammatically acceptable subject (which would be a noun phrase), the grammatical subject may simply be left unexpressed.

- (283) Vajna massaarna doqqa sovghat dut'amuo hallakbinchu vajn
 to.us all great gift is war destroyed our
 zhimchu maxkahw cwa ghishluo, husam, c'a mettahuottiicha a.
 small in.country one building home house when.restored even
 [p86-00027:25]

'It is a great thing for all of us when even one building, one home or one house is restored in our small homeland destroyed by the war.'

This is not to say that dummy subjects are not allowed. The sentence in (284) is an example where a true dummy subject is used. The subject pronoun *iza* 'it' does not seem to have a referential function, since there is no fitting antecedent (either anaphoric or cataphoric) for it.

- (284) Iza deqq'a vaj noxchii xilarnii, vajna noxchii humadiezarnii
 It only we Chechens being-& we Chechen thing loving-&
 aella a daac. [p34-00002:35]
 said INT not.is

'It is not purely about us being Chechens and liking our Chechen ways.'

To sum up what we have seen so far: Chechen copula clauses can be with an overt subject, an elided subject or they may be without subject, and Chechen may use a dummy pronoun as subject too.

The Chechen *it*-cleft has a copula construction at its basis, and we will have a look at its form using example (285).

- (285) 3 butt xaan ju Semawashkara muhazharsh gumanitarni
 3 month time is from.Samashki refugees(B) humanitarian
 gho doocush wash bolu. [p86-00085:2]
 aid not.having living B.REL
'It is the third month that refugees from Samashki have been living without any humanitarian aid.'

- (286) Semawashkara muhazharsh gumanitarni gho doocush wash bu.
 from.Samashki refugees(B) humanitarian aid not.having living B-REL
'Refugees from Samashki have been living without any humanitarian aid.'

The *it*-cleft in (285) is based on the copula clause *3 butt xaan ju* '(it) is three months time'. This kind of copula clause is subject-less, just like the one in (283). There is no grammatical subject to the sentence, while the logical subject is the relative clause *that refugees from Samashki have been living without any humanitarian aid*.

This relative clause is a time-adjunct one, as can be seen by the validity of turning it into a full main clause in (286), which shows that the clause does not have an argument gap. The relative clause is headed by the temporal NP *3 butt xaan* 'three months time'.

In order to convince ourselves that the construction in (285) really is an *it*-cleft, we should subject it to the diagnostics defined in chapter 10, section 10.1.7.

According to the *Cleft structure* diagnostic in (245) of chapter 10, a genuine *it*-cleft should consist of a copula clause and a relative clause. The construction in

(285) complies with this. We have seen that the copula construction is *3 butt xaan ju* ‘it is three months time’, and that the relative clause is the remainder of (285).

The *Cleft pronoun* diagnostic in (246) of chapter 10 requires the subject of the copula construction to either be a pronoun or empty. Our example in (285) fulfils this condition, since it does not have a grammatical subject at all.

The third, and final, diagnostic is the *Cleft coindexing* requirement in (247), which states that the relativized argument or adjunct of the cleft’s relative clause must coindex with the clefted constituent. We have just seen that the relativized adjunct of the relative clause in (285) is ‘time’, which nicely coindexes with the clefted constituent *3 butt xaan* ‘three months time’.

We may safely conclude that the Chechen construction in (285) is an *it*-cleft that satisfies all three cleft diagnostics that are defined in section 10.1.7. The next sections will discuss how often this construction occurs in Chechen, in what circumstances, and what purposes they serve.

11.3.2 Looking for Chechen *it*-clefts

We have seen that Chechen has *it*-cleft constructions, and we have seen an example of what they look like, but we need to know two more features of them: how often do they occur, and what is their function? Chapter 12 contains a quantitative analysis of *it*-clefts in English, and if we want to compare English with Chechen in a quantitative way, we need to know how often Chechen *it*-clefts occur. A qualitative comparison of English and Chechen requires us to know what they look like in more detail (e.g. whether the clefted constituent is an argument or an adjunct within the cleft clause), and what function they fulfil.

I have chosen to find answers to the questions above by building a database of Chechen *it*-clefts. This database is built using the same corpus search software that is used for the diachronic research on English *it*-clefts described in chapter 12 (Komen, 2009c, Komen, 2011b). The research method can roughly be divided into the steps shown in (287).

(287) Chechen corpus search method

- a. Define and prepare a representative corpus of Chechen texts
- b. Define queries to look for *it*-clefts in Chechen data
- c. Combine the results of executing the queries in (b) into a database
- d. Remove non-clefts from the database formed in (c)
- e. Add features to each *it*-cleft present in the database

11.3.2.1 A corpus of Chechen texts

The first step in the search for Chechen *it*-clefts is the selection of a corpus. Computational linguists from the New Mexico State University interested in working with lesser known languages have been interested in Chechen for some time. They have developed a corpus of Chechen texts, which they subsequently have made available on the internet (Zacharsky and Cowie, 2011).

This corpus consists of two parts: a parallel and monolingual one. The parallel part of the corpus contains 324 texts from various sources, where each line of

Chechen has been provided with a free translation into English.⁹ The monolingual part of the corpus contains 624 texts without translations. The texts have been made available in untokenized plain-text format.

The texts in the corpus vary in size from 1 to approximately 800 lines, and they have been taken from the newspaper “Dajmuoxk”, and from journals like “Orga”, “Naana” and “Vajnax”. The corpus does not contain genre specification for the texts. This is why the question whether the corpus is “representative” enough of the language as a whole remains. For the moment, this is all we have available.

Since some of the texts in the parallel part of the corpus were translations from various existing English corpora (e.g. the ACE 2005 Multilingual training corpus) into Chechen, I have divided the parallel part into two sub parts: the “parOrg” part, containing 210 original Chechen texts with a translation into English, and the “parTrans” part, which contains 114 original English texts with a translation into Chechen. The corpus research, then, distinguishes between an “original Chechen” part (this is the “parOrg” part of the parallel corpus, and the whole of the monolingual corpus) and a part that is “translated from English” (the “parTrans” part).

Only a modest amount of further processing has been performed with the texts in order to prepare them for the corpus search work described in the next sections. The queries that are definable in CorpusStudio require the texts in a corpus to be available in a particular kind of *xml* format, one that is based on the Text Encoding Initiative P-5 standard (Komen, 2009c, Sperberg-McQueen and Burnard, 2009). The texts have therefore been transformed from their plain-text format into this *xml* format, they have been tokenized to facilitate subsequent searches on the word-level, and the original Cyrillic form of the Chechen has been transliterated into a Latin standard form of the language (Nichols, 2007).¹⁰

11.3.2.2 Defining queries for Chechen *it*-clefts

The corpus of Chechen texts discussed in 11.3.2.1 allows us to search sentence-by-sentence, since each sentence is stored inside a unique <forest> tag, and it allows word-by-word searches, since each word is stored in a <eLeaf> tag. Since the texts have not been syntactically parsed, we cannot look for syntactic structure. POS-tagging has only been done partially, so that we are not able to rely on part-of-speech labels in our queries.

The little we have may, however, be sufficient to give us a preliminary glance on *it*-clefts in Chechen, provided we keep in mind that there may be more around. What we may do is look for two necessary ingredients in the *it*-cleft construction, which almost always occur next to one another in Chechen. The two ingredients are: (1) the head noun of the clefted constituent, and (2) the form of *be*—the auxiliary. The particular forms we are looking for are listed in Table 38.

Table 38 Looking for Chechen *it*-clefts

<i>it</i> -cleft type	head noun	auxiliary forms
time	<i>sho</i> ‘year’, <i>sharahw</i> ‘in a year’, <i>xaan</i> ‘time’, <i>xeenahw</i> ‘in time’, <i>butt</i> ‘month’, <i>k’ira</i> ‘week’, <i>de</i> ‘day’, <i>diinahw</i> ‘on a day’, <i>k’irnahw</i> ‘in a week’, <i>sahwt</i> ‘hour’	<i>đu</i> , <i>đaac</i> , <i>đara</i> , <i>đaacara</i> , <i>đu</i> <i>j</i> , <i>đarii</i> , <i>đaacii</i> , <i>đaacarii</i> ¹¹
location	<i>mettig</i> ‘place’, <i>muoxk</i> ‘country’, <i>aara</i> ‘area’	
argument	<i>huma</i> ‘thing’, <i>stag</i> ‘person’, <i>naax</i> ‘people’	

We can locate potential time-clefts, which are *it*-clefts with a clefted constituent that is a temporal adjunct, by looking for a limited number of time-related head nouns, such as ‘year’, ‘week’, ‘month’ etc. Since Chechen is fairly head-final when it comes to noun phrases, we can rest assured that a head noun is the last element in a noun phrase.¹²

The head noun, which is the rightmost landmark of the clefted constituent, is then followed by an auxiliary. There is only a limited number of auxiliary forms we look at for our current *it*-cleft search, namely all finite forms in past and present tense, optionally with a negation suffix, and optionally with a polar question suffix attached (see the last column in Table 38).

The search for time-clefts described above also yields sentences that contain no cleft at all, such as the one in (288). This sentence has been captured by the procedure described above, since it contains a temporal head noun *xeenahw* ‘at time’, which is followed by the finite auxiliary form *ju* ‘am’. It is not a cleft, however, because there is an anaphoric subject *so* ‘I’, and there is no relative clause.¹³

(288) Cu xeenahw ju so niissa hwalxa hwyezhush. [m00125.122]
 that at.time am I straight ahead gazing
 ‘I am looking straight forward at that time.’

This shows that a process of manual selection is necessary after the procedure above has selected *it*-cleft candidates. Candidates for location adjunct clefts are extracted in the same way as time-clefts, except now the head noun we are looking for is a location. If there are any argument *it*-clefts in Chechen, then we should be able to capture at least some of them by using the head nouns in the last row of Table 38. The next section gives an example of one of the queries that have been used to select *it*-cleft candidates.

11.3.2.3 Transforming query results into a database of Chechen clefts

The procedure discussed in section 11.3.2.2 identifies almost 200 *it*-cleft candidates in the Chechen corpus, which we would like to form the basis of an *it*-cleft database. Each *it*-cleft is to be supplied with a number of features that could be helpful in answering the research questions we have: what do Chechen *it*-clefts look like, and what are they used for? The features that have been defined are listed in (289).

(289) *Chechen cleft database features*

- a. Clefted constituent
- b. Auxiliary form
- c. Auxiliary label
- d. Type

The clefted constituent in (289a) allows quick reviewing of common features in the clefted constituents. The auxiliary's form in (289b) helps identify the *it*-cleft candidate in the text. The auxiliary's label allows grouping of clefts into four categories, depending on tense (past versus present) and negation (affirmative versus negative). The “type” feature in (290d) allows for a number of different values, as shown in (290).

(290) *Chechen it-cleft types distinguished in the database*

- a. prs/pst: - present or past tense of the auxiliary
- b. neg - negated form of the auxiliary
- c. it: - overt subject pronoun available as subject
- d. rev - instead of compl-*be*-RC, the order is reversed: RC-compl-*be*
- e. q - the *it*-cleft is in question form
- f. loc - the clefted constituent is in the locative
- g. adv/when/ - an adverbial clause, when clause or inf clause is used
inf instead of a relative clause

There are several reasons why it would be desirable to get as many of the features we identified automatically—if this is possible. If a feature is determined automatically, then we will find it has the same value in similar environments. This kind of consistency is helpful, because if we should decide that a feature should get a slightly different value in some of the environments, we could adjust our automatic process accordingly.

(291) *Xquery code that finds clefts and adds features automatically*

```

1   for $search in //forest/descendant::eLeaf[@Type='Vern' and
      (ru:matches(@Text, $_aux) or ru:matches(@Text, $_auxQm))]
2   (: Get the immediately preceding <eLeaf> element :)
3   let $prec := $search/preceding::eLeaf[@Type='Vern'][1]
4
5   (: Check if this element is a time element :)
6   let $time := $prec[ru:matches(@Text, $_CheTime)]
7
8   (: Add the previous element to the time element :)
9   let $full := concat($prec/preceding::eLeaf
      [@Type='Vern'][1]/@Text, ' ', $time/@Text)
10
11  (: Note location, English gloss and Auxiliary label :)
12  let $for := $search/ancestor::forest[1]
13  let $eng := $for/div[@lang='eng']/seg
14  let $aux := tb:CheAux($search)
15
16  let $dbs := concat($full,',';,$search/@Text, ',';,$aux, ',';,$eng)
17
18  where (
19      exists($time)
20  )
21
22  (: Output for database building :)
23  return ru:back($search, $dbs)

```

The procedure shown in (291) has the double goal of identifying potential *it*-clefts, and adding some features automatically. The search starts, as explained in section 11.3.2.2, by identifying a finite auxiliary form (as defined in the variables `$_aux` and `$_auxQm`) in line #1. Such a form is located within an `<eLeaf>` element that is a descendant of the `<forest>` currently being investigated. The first `<eLeaf>` element preceding this auxiliary is stored in the variable `$prec` in line #3, while line #6 copies this `$prec` element, provided its `@Text` attribute matches with one of the temporal NP heads defined in `$_CheTime`. Line #19 and #23 make sure that whenever a `$time` element is found, the current `<forest>` element is added to the output of the query.

The intermediate lines #9-16 are used to prepare the variable `$dbs`, which contains a semi-colon separated list of 4 features calculated for this result item: the clefted constituent in `$full`, the auxiliary's form in `$search/@Text`, the auxiliary's label in `$aux` (as calculated in the function `tb:CheAux`), and, if it exists, the English translation of the Chechen sentence is passed on in `$eng`.

As explained in section 7.3.2.2, there are two more queries: one to select candidates for locative adjunct *it*-clefts, and one to select candidates for argument *it*-clefts. These queries operate in a similar way to the one in (291), but they obviously use different variables for the head nouns potentially signalling a locative adjunct clefted constituent or an argument clefted constituent.

11.3.2.4 Working with the Chechen cleft database

Execution of the three queries discussed in section 11.3.2.3 on the corpus of Chechen texts yields a database with *it*-cleft candidates. This database can be imported into the program Cesax, and then edited (Komen, 2011b).¹⁴ Figure 34 provides a glimpse of the Chechen *it*-cleft database as it is being edited.

The CorpusResults tab page in Cesax allows showing the database of results as produced by the corpus research project that has been run in CorpusStudio. Each potential *it*-cleft in the database has been reviewed manually, and the *Type* feature has been added, since it is not possible to calculate it automatically, given the status of the corpus of Chechen texts.

After the non-clefts have been removed from the database, we are left with a database that consists of 109 verified and annotated *it*-clefts in Chechen.

CESAX: Editor for syntactically annotated corpora

File Edit View Section Translation Corpus Reference Must Tools Help

General Editor Syntax Translation Report Errors CorpusResults

Corpus research project: CheCleft_Queries Database file: D:\Data Files\Corpora\CorpusStudio\Clefts\Che\CheCleft_Database_V2.xml Created: 23/02/2011 9:23:35

Additional information: Database of results created from queryline: lfm lmeCleft Leaf Analysis: Time-AuxVal,AuxLab,Type

Selected feature:

Select one result from the database

File#	Text#	Cat	Fore	P#	Select	Status
18	m00185	ps	4	C3	Duqa xaan	Verified
19	m00209	ps	1	C3	shuo xaan	Verified
20	m00209	psrev	8	C3	sov xaan	Verified
21	m00220	psrev-neg	6	C3	Kezig xaan	Verified
22	m00225	psit	1	C3	but xaan	Verified
23	m00233	ps	11	C3	shuo xaan	Verified
24	m00249	psrev	70	C3	shita-ogha shuo	Verified
25	m00256	ps	7	C3	5 sho	Verified
26	m00257	psrev	15	C3	dup-dupa xaan	Verified
27	m00263	ps	14	C3	sov xaan	Verified
28	m00272	ps	10	C3	41 shuo	Verified
29	m00272	ps	18	C3	Duqa xaan	Verified
30	m00272	ps	28	C3	25 shuo	Verified
31	m00272	ps	32	C3	itlex de	Verified
32	m00273	ps-neg	1	C3	Duqa xaan	Verified
33	m00274	ps	34	C3	geigga xaan	Verified
34	m00297	psrev	1	C3	3 shuo	Verified
35	m00298	ps	10	C3	41 shuo	Verified
36	m00298	ps	18	C3	Duqa xaan	Verified
37	m00298	ps	28	C3	25 shuo	Verified
38	m00298	ps	32	C3	itlex de	Verified

Number: 23 Text: m00233 Location: m00233.11
 File: D:\Data Files\Corpora\Chechen\lme\m00233.psd
 Period: C3 Forestid: 11 e'held: 1 Category: ps Status: ps Syntax:

[m00233] [m00233.8] Hince ambulatormi darba liehton fe'yecu, mezhienash chyraevilla ja jojna a, ishita, hveeran a, j'elwini a camgash j'ohi darvoj [m00233.9] Caarna q'ingahw joqqa g'oolle a karajo. aatta bocchu baak'a t'el'hw ziedellag [m00233.11] Taamasha a baac, tq'e qojita shuo xaan ju cuo quzahw q'ahw'ye'gu. [m00233.12] A.

User adaptable features

Pde	@
Time	shuo xaan
AuxVal	ju
AuxLab	BEP
Type	ps
Erg	[Unamazingly] it is thirty three years that he has worked here.

Notes

Initialized.

Figure 34 The chechen it-cleft database seen from Cesax

11.3.3 Discussion of the corpus findings

A quantitative comparison between Chechen and English *it*-clefts requires us to know how often these constructions occur in Chechen. The database of Chechen *it*-clefts described in section 11.3.2 has been reviewed and corrected manually, and the results stored in the database are shown in Table 39. The search yields a total of 104 *it*-clefts in the “original Chechen” part of the database, which amounts to 327 *it*-clefts per 100,000 main clauses. It is this normalized number that we will need to compare the Chechen data with the English data, which are described in chapter 12.¹⁵

The 5 *it*-clefts found in the part of the corpus that is “translated from English” amount to 73 *it*-clefts per 100,000 main clauses. I give this figure for completeness, but, due to the suspicious nature of translated texts, these clefts are to be left out of the comparison between Chechen and English.

Table 39 Results of the Chechen *it*-cleft database

Subject	Clefted constituent	Word order	Illocutionary Force	Original Chechen	Translated from English
overt	time adjunct	canonical	declarative	5 (5%)	0
(none)	time adjunct	canonical	declarative	58 (56%)	0
(none)	time adjunct	reversed	declarative	39 (38%)	3
(none)	time adjunct	canonical	question	2 (2%)	0
(none)	time adjunct	reversed	question	0 (0%)	2
(none)	locative	(any)	(any)	0 (0%)	0
(none)	argument	(any)	(any)	0 (0%)	0
<i>total</i>				<i>104</i>	<i>5</i>

A qualitative comparison between Chechen and English *it*-clefts requires us to take a closer look at the possible forms of the *it*-cleft constructions found in Chechen, and their function. We leave the discussion on their function to section 11.3.4, and first concentrate on the variation in forms.

Chechen *it*-clefts only appear to have a time phrase, a temporal adjunct, as clefted constituent: no argument or locative adjunct *it*-clefts have been found that conform to the three *it*-cleft diagnostics in 10.1.7. There is one interesting observation that needs to be made about the form of the temporal adjuncts serving as clefted constituents in Chechen. We have searched for two possible forms: NPs headed by a time noun in the nominative and in the locative case. This last type of NPs is comparable to English PPs. The noun *sho*, for instance, means ‘year’, but when it appears in the locative case, such as in *hoqu sharahw*, where *hoqu* is the inflected form of the near demonstrative, it has to be translated with a PP ‘in this year’. Interestingly, the temporal adjuncts serving as clefted constituents are all NPs in the nominative case.

There are two more observations we can make about the possible *forms* of the Chechen *it*-cleft. Although these observations show that a finer grained investigation of *it*-clefts in Chechen is needed to clarify several intriguing details that surface,

they have no influence on the main line of this chapter, which argues that Chechen has *it*-clefts, and that these *it*-clefts are not motivated by focus.

From the total of 104 original Chechen *it*-clefts, five samples have an overt pronominal subject. One of these is shown in example (292).

- (292) Hara **cwa butt** **xaan** ju Noxchiin Respublikan Q'ooman bibli'otekan
 this one month time is Chechen Republic's national library's
 bielxaxuosha de-byysa ca Iyerush q'ahwyegu. [m00225.1]
 employers day-night not regarding toil
*'It is one month since the employees of the Chechen Republic's national
 library have been working both day and night.'*

The example in (292) is the opening line of a newspaper article, which makes it clear that the pronominal subject *hara* 'this' cannot be anaphoric. Given the marginal number of *it*-clefts with an overt subject pronoun, this structure could be an innovation, in which case we should see it happen more as time passes by, or a remnant of the past, in which case we should see it relatively more often in older texts. The current study only has corpus data from one time period available, so that we cannot look at the dating of the texts. I suspect that the use of an overt subject is an innovation, since there is no *grammatical need* for copula clauses, which form the basis of the *it*-cleft, to have a subject.

The results presented in Table 39 use several more differentiations: the word order of the cleft's main clause, the illocutionary force of the sentence, and the origin of the cleft.

The main clause word order may be "canonical", in which case the relative clause occurs sentence-finally, as in (293), or it may be "reversed", in which case the relative clause is situated sentence-initially, as in (294).

- (293) (T'aehwaluonan ojla a jiesh, xaza kyg tuuxush jina ghishluo ju hara.)
Duqa xaan jara ooxa hoqu ghullaqie satesna a. [p86-00185.9]
 much time was we this to.matter hoped &
*'(It is a beautifully built building built for the future.)
 We have waited for this for a long time.'*
- (294) (Sa'iev Wumar literaturiehw kerla stag vaac,) [p86-00184.15]
 cuo noxchiin literaturiehw q'ahwyegu **tq'a sho sov xaan** ju.
 he Chechen in.literature toils 20 year more time is
*'(Umar Saiev is not a new man in Chechen literature.)
 He has been working hard in Chechen literature for over twenty years.'*

The "reversed" cleft in (294) conforms to the *it*-cleft diagnostics in section 10.1.7, since none of them prescribes a particular word order of the main clause. The question obviously arises what the function is of a reversed word order for Chechen *it*-clefts. I will leave this for further research.

11.3.4 The function of Chechen *it*-clefts

A qualitative comparison between Chechen and English *it*-clefts requires us to take a closer look at the function of Chechen *it*-clefts. Are they used as a focusing device, or as a thematization device, that is: for text organization.

Let us first consider the possibility that they are a focusing device. An argument in favour of this analysis would be that the position of the clefted constituent, the immediately preverbal one, is that of the focused constituent. This is in line with the findings of section 11.1 on focus in Chechen as a whole. However, position as such is not sufficient in this case, since Chechen is an SOV language, and the most natural position for a complement is the preverbal one anyway—focused or not. This is the same problem as that of recognizing object focus from a transitive sentence that has SOV word order: the position as such coincides with the unmarked word order, so it does not necessarily point to *constituent* focus.

If the SOV word order is not a sufficient indication of focus, then the question arises whether there are other indicators of focus-hood. There are a few standard indicators of focus-hood: (a) the presence of focus particles, and (b) the presence of a question word.

As for focus particles, the equivalent for ‘only’ (Chechen: *bien*) has not been used as indicator of focus, as far as I am aware of. It is possible that the clitic =*m* functions as focus marker in Chechen, but I am not aware of published research results in this area.¹⁶ There is no doubt that there are other focus particles in Chechen, such as for instance the particle *a*, in its use in examples (283) and (284), where it is translated as “even”.¹⁷ The corpus of 104 *it*-clefts contains 3 occurrences of *it*-clefts where the intensification particle *a* modifies the clefted constituent. One of these is shown in (295).

- (295) Tq'a ysh mella=a sixa xiica jiezash xilla jolu xaan
 but they however fast change needing been being time
 t'exjaella **shiitta-qojtasho a** du. [m00249:70]
 surpassed 12-13 year INT is
 ‘*But the time that they should have been replaced as fast as possible, has now surpassed even twelve-thirteen years.*’

The combination of a focus particle with an approximate time like ‘twelve to thirteen years’ sounds a bit awkward in English. There does seem to be some kind of constituent focus, since the time ‘twelve-to-thirteen years’ is compared with ‘as fast as possible’.

The presence of a question word to indicate focus-hood has already proven its value in section 11.1, so we can be confident to use it here too. Only 3 of the 104 *it*-clefts have a clefted constituent containing a question word, and these instances are shown in (296) and (297).

- (296) **Miel** **xaan** ju vaj karzaxdevlla? [m00300:73-74]
 how.much time is we stood.up
Ja miel **xaan** ju parghatdovla ghierta, booxush, hwiiza?!
 or how.much time is get.free to.try saying torment
 ‘*How long is it that we have stood up?*
Or how long is it that we torment ourselves, saying we try to get free?’
- (297) As horsh dyycu, hwiexado **miel** **duqa** **xaan** ju. [p34-00002:21]
 I these talk teach how much time is
 ‘*I have been talking and bringing it up for a long time.*’

The two examples in (296) do contain a question word in the clefted constituent, but it should be noted that both of them are rhetorical question. The reason they are used is not to elicit an answer, but to convey emotion. As such they do convey a form of intensification. The example in (297) is from the parallel part of the corpus, and has a “reversed-order” *it*-cleft. The question word again does not serve its role as question-elicitor, but it does convey intensification.

The corpus also contains examples like (298), which illustrate that the clefted constituent can have characteristics that are quite unlike those of constituent focus.

- (298) «Phwarmat» quollajelcha dyyna swa, **30sho gergga xaan ju**
 Phwarmat created.when since from 30 year almost time is
 so hoqu t'iehw buolx biesh volu. [p86-00064:40]
 I this on work doing am.REL
'I have been working at this since the creation of "Pharmat" – for about thirty years.'

Where constituent focus identifies and enforces one particular variant, explicitly or implicitly contrasted with alternatives (as for example Krifka, 2007), the clefted constituent in (298) contains an approximate time reference, which is an *open* set of alternatives, which makes it much *unlike* focus.

In sum, apart from the focus associated with the preverbal position, there are a limited number of examples where emphatic prominence is expressed in the clefted constituent of the Chechen *it*-cleft, but this never seems to be the main rationale for using a cleft construction.

This brings us to the second possibility for the function of *it*-clefts in Chechen mentioned above: that of text organization. If the *it*-clefts are used to indicate textual boundaries, then we expect them to occur (a) story-initially, (b) paragraph-initially and (c) story-finally. These possibilities are in line with Johansson's (2002) ideas on Norwegian, discussed in section 10.2.5, which recognizes the use of clefts for “Topic launching”, “Topic linking” and “Summative”.

In order to verify the text-organization function of *it*-clefts in Chechen, it is (at least sometimes) better to show a larger stretch of a text, so that we can better judge whether the position of the cleft coincides with a paragraph start, transition or end. This is what has been done in our first example (299), which is an article that contains a forum discussion on the use of Chechen as the principal language in elementary schools.

- (299)¹⁸ a. Kati, it is not right for us to come to this magazine's office and tell them that there is practically nothing being done and that talking is a waste of time.
 b. What is the duty of a magazine? To listen to your, my and their opinions, write them down in some way and deliver them to people. When these guys, another magazine or another newspaper raise an issue, talk about it over and over again, *then* the government can do nothing else but what it is supposed to do. But nothing happens if we stay away from discussing a problem.
 c. We have been talking for a **long time** about it. [p34-00002.29]

- d. What you said in the beginning that two to three grades in school should be in Chechen is the topic we have so far been talking about. It will happen, as long as we keep talking about it. Without giving in. If we speak about it, we should not speak about it superficially. It is not enough to speak about switching elementary classes to Chechen, when it is not understood why it needs to be switched, and it is incumbent on us to provide a foundation for that. It is not purely about us being Chechens and liking our Chechen ways. There is more to it than that.

The paragraphs in (299a-d) are the start of a reaction from one participant in the forum. He addresses the interviewer with “Kati” in paragraph (299a). Paragraph (299b) opens with a typical topic-introducer: a question. Paragraph (299d) likewise identifies a clear change of topic, which is retained as “it” throughout this last paragraph. The line in (299c) contains the Chechen *it*-cleft. It functions as a transition between the previous paragraph (299b) and the next one (299d). The link with the previous paragraph is by the pronoun “it”, which refers to the whole clause “nothing happens if we don’t discuss the problem”. The link with the next paragraph is clear too, because the start of (299d) copies the “we have been talking” element. In sum, the Chechen *it*-cleft here functions as an episode boundary marker.

There is one more *it*-cleft in this same text which we may consider, and it is shown in example (300).

- (300)¹⁹ a. Abdullah: The development of the Chechen language and literature depends mostly, as you said, on a school. The fact that the elementary school should be in the Chechen language is beyond any doubt. Not only elementary school, middle school too should be in the Chechen language. However, as of today, we shall have the financial capabilities to switch only elementary school to the Chechen language.
- b. I would like to say a few words about it, because we have been studying the problems for **a long time**. [p34-00002.253]
- c. The elementary school was switched to the Chechen language. It was at the end of the past century. I was the one who paid visits to the Regional Committee at the time of the switch.

Line (300a) starts the contribution of Abdullah, a participant in the forum. This first paragraph gives some background, and (300b) finishes this introduction by announcing that he is going to say “a few words” about this matter. The content of what he then says starts in (300c). Again we see that the *it*-cleft is in a position where it helps finish off one topic, and introduce another one.

A total of 14 *it*-clefts from the corpus (which amounts to 13%) are located at the beginning of a story or report. We can see the English translation of those that occur in the “parOrg” part of the corpus in (301).

- (301) a. It was **not long ago** that a medical insurance ZAO (closed shareholders company) called "Maks-M" opened another branch in Grozny, at Pervomaiskaya street #85. [p86-00063.2]²⁰
- b. A team of the Achkhoy-Martan financial department has hoped for **a long time** that a new building would be built. [p86-00027.2]²¹
- c. (It is **the third month** that refugees from Samashki have been living without any humanitarian aid). [p86-00085.2]²²
- d. It is **the 5th year** since the branch of the PTU #113 was opened in the village of Samashki of the Achkhoy-Martan district. [p86-00110.2]²³
- e. It has been **at least 25 years** since a literature group called "Shovda" has been working at the newspaper's group "Gums" in the city of Gudermes. [p86-00130.2]²⁴

All of the examples in (301) provide clear opening sentences for a text: they anchor a theme in a timeframe. It is interesting to see that four of the five were translated with English *it*-clefts by the native Chechens who cooperated in establishing the corpus. What the time adjuncts in the clefted constituents do is establish a link between the whole of the article and the real world. Such a link is a kind of scene-setting, and is usually not something that is developed as topic later on.

The monolingual part of the corpus contains the remaining 9 instances of *it*-clefts that start off a story or report. We have already seen one of them in (292), where it was brought up as illustration of Chechen *it*-clefts having a non-anaphoric pronominal subject. In fact, it should be noted here that four of the five *it*-clefts that use the demonstrative pronoun *hara* 'this' as subject are story-initial ones. The reason for this is probably the avoidance of ambiguity: the near demonstrative *hara* can quite easily link up with something in the previous sentence, or with the previous sentence as a whole, but this is impossible if there is no previous sentence.

What about the "Summative" function Johansson (2002) found for Swedish? Is the Chechen *it*-cleft used for that text-organizational function too? The number of times an *it*-cleft is used to *finish* a story is very limited. I have only found one example of this in the whole corpus, and this example is shown in (302).

- (302) Taamasha a baac, **tq'e qojtta sho xaan** ju cuo quzahw q'ahwyegu.
 surprise & not.is 20 13 year time is he here toils [m00233.11]
 'It is no surprise that he has been working here for thirty three years.'

The newspaper story that finishes with (302) is a small biography in praise of a doctor called Umar Astamirov, and it speaks of how good he is at his job and how well he relates to patients and people. The concluding remark about the number of years he has been working at this particular hospital is a worthy end of the biography, underlining his dedication to the work, and the hospital commitment to keep him on.

The scope of this dissertation is too limited to discuss all the remaining examples of *it*-clefts in Chechen, but what we have seen so far is that the construction is used as a story-opener (to set the scene for the rest of the story), and that it can function as a paragraph transition device in other situations. This "paragraph transitioning" function compares with Johansson's "Topic linking" one, where one

discourse topic (in the clefted constituent) is linked to a subsequent discourse topic (which is in the cleft clause). The Chechen *it*-cleft, then, is a linguistic realization in the 3D space suggested in 4.1 of particular values on the “text-structure” axis, and not of a value on the “focus” axis.

11.4 Conclusions and implications

The main claim of this chapter has been that there are languages that use *it*-clefts only for text-organization, and not for constituent focus at all, and that Chechen is one of those languages. Komen’s (2007b) work on the relation between word order and focus in Chechen showed that this language uses the immediately preverbal position for focus. Both *wh*-questions and the constituents answering those questions appear in the preverbal position, and it does not matter whether these constituents are arguments or adjuncts.

The same work also showed that *wh*-clefts can be used as an alternative to plain word order for the purpose of focusing. While the principle of using the preverbal position for focus is also operative in *wh*-clefts, these constructions make it possible to distinguish the unmarked SOV word order from constituent focus on the object.

Word order with or without the use of the *wh*-cleft construction seems to be even more important as a focusing device, given the observation that Chechen does not use intonation to convey focus on a constituent. The Chechen language breaks up sentences in accentual phrases, which vary may be up to 8 syllables long. These accentual phrases are demarcated by a left and right boundary tone, and they contain a maximum of one H* or H*L pitch-accented syllable. This syllable is the leftmost accentable one, unless the accentual phrase contains one of the function words or morphemes that have lexical tone. In that case the function word or morpheme holds an H* pitch accent. All question words have lexical tone, but focused answers to question words do not. This is why Chechen does not have a separate focus intonation pattern.

Corpus research on a set of contemporary Chechen texts from journals and newspapers reveals that Chechen does have an *it*-cleft construction, whose frequency was determined to be 317 per 100,000 sentences, but that these constructions are not used to convey constituent focus. The clefted constituents are never arguments from the cleft clause, but are always time adjuncts. The function of these *it*-clefts is that of text-organization, since they occur at the start of a story and as a link between thematically different paragraphs.

To sum up, there is at least one language that has *it*-clefts, but does not use them as constituent focus devices. This is an important observation that needs to be kept in mind as we turn to the diachronic review of *it*-clefts in the English language in chapter 12.

¹ The sentences are partly taken from the same corpus used in this dissertation, which is discussed in section 11.3.2. Another part of the sentences were taken from books and from material available on the internet. See Komen (2007b: 72) for details.

² This dissertation uses a phonemic transliteration of Chechen that is very closely related to the one used for Ingush (Nichols, 2007). The vowels and consonants roughly appear in their IPA forms, with the following exceptions. The *w* on its own represents an epiglottal stop (equivalent to Arabic “ajin”), while it represents an epiglottal fricative when it follows a voiceless consonant. The *hw* represents a pharyngeal fricative /h/, and the *gh* represents the voiced uvular fricative /ʁ/.

³ Words that are orthographically written with a hyphen in Chechen, such as *joqqa-baaba* ‘grandmother’, are glossed as a whole unit. Only when hyphens in the Chechen words indicate morpheme breaks are the glosses also broken up into morphemic units, such as with *dwaah-waarch-iehw* ‘away-wind-PLSE’.

⁴ Psych verbs are verbs of a psychological state or event. Such verbs do not have an Agentive subject, but an Experiencer one. Examples are: *see, feel, hear, like*. Psych verb subjects are in the dative case in Chechen.

⁵ Unpronounced vowels are either raised or between square brackets. The abbreviation *ip* is used for the Intonation Phrase and AP for the Accent Phrase.

⁶ It is not completely clear why the H* on *var^a* is phonetically higher than the H* on *dyes^hsh* in the previous AcP.

⁷ Chechen has 6 noun classes, which are signalled on verbs that start with a noun-class prefix. There are 4 possible prefixes (j, v, b, d), and each noun class is defined by the set of noun-class prefixes used in singular and plural. The classes “j-d” and “v-d” are used for feminine and masculine nouns, while the remaining four classes (“j-j”, “d-d”, “b-b”, “b-d”) are used for non-human nouns.

⁸ It is only when the complement is *not* an NP that the form of *be* agrees in noun class with the subject.

⁹ These free translations have been provided by native Chechen speakers “without an intermediate Russian stage” (Cowie, 2011).

¹⁰ See also footnote 2.

¹¹ The consonant *d* is a place-holder for any of the noun-class prefix consonants *v, j, d* or *b*.

¹² Notable exceptions to this rule are extraposed relative clauses and extraposed possessors, for which I refer to Komen (2008).

¹³ The clause *niissa hwalxa hwezhush* ‘looking straight ahead’ is an adverbial clause, syntactically modifying the main verb, which is the auxiliary *ju* ‘am’.

¹⁴ At the moment of writing, this program is freely available on <http://erwinkomen.ruhosting.nl>.

¹⁵ We can only fairly compare normalized numbers of occurrences between languages or language-variants, because if the corpus we have of one language is significantly larger or smaller than that of the language we are comparing it with, the difference in absolute numbers of occurrences will be mainly due to corpus size differences.

¹⁶ Molochieva (2010) discusses the function of the verbal suffix *-q*, which can indicate mirativity (the presence of unexpected information), but also “emphatic meaning” on a clause as a whole. More research would be needed to see if this suffix plays a role in “verum focus”, since it often occurs attached to the assertive or negative auxiliary. A proper investigation into

these and other phenomena in Chechen would be greatly facilitated by the presence of a parsed corpus.

¹⁷ The particle *a* can have several different functions. It can be used as focus particle (it is referred to as “intensifier” in this use), but also as negator (for instance turning *huma* ‘thing’ into *humma a* ‘nothing’), as coordinator (for instance *daada a, naana a* ‘father and mother’) and as co-subordinate clause marker (for instance *iza a dina* ‘having done that’).

¹⁸ [24] Kati, vaj hoqu zhurnalie a daexkina, vaj hoqaerga prakticheski diesh humma a daac, q’amielash dar erna du baexcha, niisa daac. [25] Zhurnalan dieqar hun du? [26] Hwuuna, suuna, hoqaarna, qiechaarna xietash dolchynga la a dyeghna, cwana kiepiehw dwaajaazdina, naaxie dwaqaachuor du-q. [27] Hoqaara, qechu zhurnaluo, gazieta shaa aj’a a aj’ina, jux-juxa diicicha, dan huma a ca xylii, t’aaqqa do wiedaluo shaa diirig. [28] Tq’a diica a ca dyycush wad ditcha, cunax humma a ca xylu. [29] Vaj i dyycush dolu **duqa xaan** ju. [30] Ahw dwaavuolalush aellarg, noxchiin mattahw shi-qo klass xila jiezash xilar, dyycush wash du vaj cq’achunna. [31] Cq’a macca a ghullaq xir du hwuuna cunax, vaj diicichahwana. [32] Q’ar ca lush. [33] Vaj dyycush xilcha, t’exula diica ca dieza. [34] Jyhwancara shkola noxchiin mattie jaaqqa jieza baxarx tye’ush daac, hunda jaaqqa jieza qietash ca xilcha, cunna bux kechbar - iza vajna t’iehw du. Iza deqq’a vaj noxchii xilarnii, vajna noxchii huma dezarnii aella a daac. [36] I doocurg qin a humnash ma du cigahw.

¹⁹ [249] Abdulla: - Noxchiin muott, literatura qi’ar duqax dolchunna, ahw ma aallara, doozadella du shkolas. [250] Jyhwancara shkola noxchiin mattahw xila jiezar - iza cwaellig cwana aaghuor shiekuonie huma daac. [251] Jyhwancara shkola hwovxa, juqq’iera shkola a xila jieza noxchiin mattahw. [252] Baq’du, vajn taxanleerachu diinahw taruonash xir jaac noxchiin mattie jyhwancara shkola jaaqqa bien. [253] Asa ocunax masiex duosh eer du, hunda aelcha i problemash ooxa tollush **dikka xaan** ju. [254] Jyhhwancara shkola noxchiin mattie jaeqqina jara. [255] I dara dwaadaxanchu bweesheran chaqqiengahw. [256] I joqqachu xeenahw obkomie liellarg so vara.

²⁰ **Duqa xaan** jaac ZAO “Maks-M” c’e jolchu medicinie straxovani jaran kompanis Syelzha-Ghaalin Pervomajski uuraman No. 85 jolchu c’iiniehw shien roghiera fili’al dwaajillina.

²¹ **Duqa xaan** jara T’iehwa-Martana rajonan fin’otdelan kollektivuo kerla ghishluo jarie satyysu.

²² **3 Butt xaan** ju Semawashkara muhazharsh gumanitarni gho doocush wash bolu.

²³ Hara **5-gha sho** du T’iehwa-Martana rajonan Semawashka jyrtahw Syelzha-Ghalin No. 113 jolchu PTU-n fili’al swajillina.

²⁴ **Laxxara 25 sho xaan** ju Gymsie ghaalin «Gums» c’e jolchu gazietan redakciehw «Shovda» c’e jolu literaturan kruzok bolxbiesh jolu.

The search for the relation between syntax and focus that kicked off this study in (11) led to a corpus research into the changes that took place in the strategies used for presentational focus (chapter 8) and constituent focus (chapter 9) in English. One of the linguistic methods to express constituent focus, the *it*-cleft, had already been identified in chapter 4, and was selected as a strategy that deserved more in-depth scrutiny. This led to the definition of the *it*-cleft in chapter 10, and as we considered the *function* of the *it*-cleft, the Scandinavian languages pointed to the possibility that constituent focus might not be the only, and not even the core function they fulfil. The previous chapter 11 confirmed this suspicion, as it showed one language, Chechen, to have an *it*-cleft exclusively functioning as a text-organization strategy and not one to signal focus.

This brings us to the current chapter, where we consider the development of the *it*-cleft in English: are the *it*-clefts in English a “focus promotion device”, and has this function evolved or was it always present? These are important questions, because of the different claims that have been made about the function of English clefts. Some would see the English *it*-cleft as a *syntactic* focusing device (Lambrecht, 2001: 472), some claim that it marks exhaustive identification (Kiss, 1998), and others recognize a class of *it*-clefts in which the clefted constituent is discourse old and not focused at all, implying that such clefts have another function (Gundel, 2002, Hedberg, 1990).

What I will show in this chapter is that the rise of the *it*-cleft in English reveals an intricate interplay between syntax and focus. The constituent focus requirements of (a) having a distinguishable domain and (b) placing the focused constituent against the natural information flow are met in OE by the PreCore position, which is clearly discernable due to the V2 nature of OE. With the loss of V2, this clearly distinguishable position fades away, and what remains is much less usable for the constituent focus requirements. This is where the *it*-cleft steps in. While its biclausal nature makes it ideally suitable for text-structuring purposes, which is what OE mainly uses it for, it is also a construction that meets the constituent focus requirements: the complement of the *it* main clause offers a clearly discernable domain for constituent focus, and the fact that the *it* clause part precedes the remainder of the clause allows structuring the information in such a way, that the focused constituent goes against the natural information flow.

12.1 Research on the history of clefts in English

The number of articles, books and chapters in grammars discussing *it*-cleft constructions in English is huge, but surprisingly little has been written about the historical development of this construction. There are several authors who have mentioned the existence of cleft constructions in Old English (e.g. Mitchell, 1985,

Visser, 1963), but, as far as I am aware, no one ventured into a comprehensive study on the rise of the cleft in English until Ball's dissertation (1991). Her extensive study is based on a corpus selected by herself.

The starting point of Ball's work is the idea that the "*it*-cleft is a periphrastic construction, whose function is to emphasize, or make prominent, the predicate" (1991: 13). So Ball already starts out with the idea that the basic function of the cleft is to emphasize. In her dissertation, Ball considers all kinds of constructions for OE and ME, while she focuses on Informative-Presupposition clefts in her subsequent article (1994), which includes data until LmodE.¹ Some of the OE constructions she looks at should not, as she argues, be regarded as clefts. She comments on the analysis of a construction like (303):

"Because there is no perceptible gap in the complement, and because there is a non-cleft analysis available, there is no motivation for a cleft analysis for the OE tokens." (Ball, 1994: 612)

- (303) a. *Da wæs æfter dissum þætte Agustinus Breotone ærcebiscop*
 then was after this that Augustine Brittain's archbishop
 gehalgade twegen biscopas. [cobede:976]
 consecrated two bishops
 '*Then after this Augustine, archbishop of Britain, consecrated two*
 bishops.'

Ball excludes constructions from being called clefts, unless the cleft clause contains a "perceptible gap" that is filled by the clefted constituent, where I understand a "perceptible gap" to refer to an argument role in the cleft clause. I disagree with Ball's stance, and, according to the definition in (244), regard constructions as clefts as soon as coindexing is possible. According to the arguments in favour of adjunct clefts discussed in section 10.1.2, coindexing between the clefted constituent and adjunct positions in the cleft clause is possible, even though such positions may not be "perceptible". This divergence in the acceptability of clefted constituents that have an adjunct role in the cleft clause has major implications for the view of the development of the *it*-cleft arrived at in this chapter.

Concerning "be" structures with a time adjunct as clefted constituent (as in 303), Ball sides with Visser (1963), who says that in such cases "to be" is not an auxiliary, but has the sense of "happen, take place". Such reasoning would indeed exclude example (303) from being an *it*-cleft construction by the main-clause diagnostic (245). The definition of the cleft in (244) takes main clause copula constructions as a starting point, which means that the presence of a copula like "be" is dependent on the way a language expresses a copula construction. If it requires a copula in such a construction, then it should also be present in the cleft.

Since Ball's starting point differs, her conclusions differ too. She does not accept constructions as genuine *it*-cleft variants until they appear in early ME. One of the first clefts to appear, as recognized by her criteria, is the early ME one in (304).

- (304) a. (Ah nis nawt lihtliche of þis meidenes mot, for gef ich sod schal seggen in
 hire ne moted na mon) [cmkathe:229-232]
 for nawt nis hit**monlich** **mot** þet ha mealed.
 for not not-is it human argument that she voices
 (ne nawt nis heo þt haued us acomen.)
*'(But this girl's argument is not to be underestimated, for, if I must tell the
 truth, nothing human speaks in her,
 because it is **no human argument** she voices,
 (nor is it she who has tamed us.)*

Ball sees a historical development which starts out with “Stressed-Focus” clefts like the one in (304), and only later (starting from late ME) starts to incorporate “Informative-Presupposition” ones. It is primarily the latter type of clefts that contain adjuncts as clefted constituents whose role in the cleft clause becomes more and more optional.

Pérez-Guerra (1998) looked into the appearance of *it*-clefts, as well as the *it*-constructions that look a lot like clefts, but do not have a coindexation between the clefted constituent and the cleft clause (see examples 238a,b,e). He coins such constructions as “EX/*it*”, and notes a rise in *it*-constructions in general. He sees this rise happening at the cost of a decrease in Right Dislocation, as in (305). I hesitate to see right dislocation as being involved in the rise of the *it*-cleft, because the structural differences are too large. The *it*-pronoun in a right dislocation example as (305) is a cataphor that corefers to the large, right dislocated, constituent *the showdown...* It seems unlikely that the equivalent *it*-cleft would have such a large constituent as clefted constituent, and it also seems unlikely that the main clause *X must come now*, which is in the foreground, would appear in a subordinate position, which is a position for backgrounding.

- (305) It must come now – the showdown between Anne Vardon and her greatest enemy. (Pérez-Guerra, 1998: Example 16)

Pérez-Guerra uses corpora starting in the late ME period, and, based on the development he finds in the EX/*it* constructions, he identifies the “spread” of the *it*-cleft as taking place in eModE.

Filppula (2009) investigates clefts in the parsed OE and ME corpora, limiting his research to those clefts that have been marked as CP-CLF by the annotators of these corpora. He observes the relatively large number of Informative-Presupposition *it*-clefts in English with a time adjunct as clefted constituent. He argues for a grammaticalization of the cleft, because he sees that the distribution of the *it*-clefts over different authors increases from OE to ME, and because the diversity in clefted constituents increases. His main point is that English clefts could perhaps, on the basis of indirect evidence, have arisen as a result of contact with Celtic.

Patten (2010), although focusing in her dissertation mainly on the rise of the *it*-cleft as a construction, seems to agree with Ball's arguments about the historical development of the *it*-cleft construction: the first *it*-clefts are Stressed-Focus ones, and the Informative-Presupposition ones are a later development. She writes:²

“I used data from the PPCME2 and the PPCEME to show that the non-NP *it*-cleft and the IP *it*-cleft have developed by extension from pre-existing *it*-cleft constructs.” (Patten, 2010: 273)

To sum up, researchers so far see the Stressed-Focus clefts as more “basic” from a historical point of view, and see the rise of Informative-Presupposition clefts as either developing from it by extension, or as a parallel, but later development. The main discrepancy between these analyses and mine is the fact that I embrace a broad definition of *it*-clefts, as described in section 10.1, which is based on relatively objective grounds. My analysis results in the inclusion of adjunct clefts, as argued for in 10.1.2, and it is these clefts that abound in Old English, as we will see in the corpus study.

12.2 Making a historical cleft database

This section describes the formation of a database consisting of *it*-clefts and all necessary characteristics for the purpose of analyzing the development of clefts. It is notoriously difficult to capture cleft constructions, and to distinguish them from cleft look-alikes, even when syntactic information is available. However, the annotators of the available parsed English corpora (see chapter 6) already identified clefts by giving the cleft clauses a separate label: CP-CLF (Marcus et al., 1993). Their work can be used as the basis for building a database of clefts, as explained in section 12.2.2. The situations where cleft constructions have not been identified as clefts by the annotators are discussed in section 12.2.3.

12.2.1 Requirements for a cleft database

In order to investigate the role of clefts in the interplay of syntax and information structure, we need a database of clefts. This database contains the text and context of all the *it*-clefts found in the parsed English corpora, and for each cleft it holds relevant syntactic features as well as the information states of the clefted components.

Table 40 Syntactic and information state features used in the English cleft database

Feature	Description	Example(s)
CleftType	Word order of the main clause	svoc, svac, svpc
CleftedCat	Syntactic category of the constituent (or gap) in the cleft clause that is co-indexed with the clefted constituent	Subject, Adjunct
CleftedType	Noun Phrase type of the clefted constituent	Bare, Dem, Pro, FullNP
CleftedCoref	Information status of the clefted constituent	New, Identity, Inferred
ClauseStatus	Information status of the cleft clause	New, Known, Inferred
FocusType	Kind of constituent focus used in this cleft.	none, Temp, Contrast

The syntactic and information state features that are stored with each cleft, as shown in Table 40, are the following. The *CleftType* provides the word order of the main components of a cleft, the *CleftedCat* a generic category for the role fulfilled by the

constituent co-indexed with the clefted constituent inside the cleft clause, the *CleftedType* reveals the clefted constituent's Nptype, the *CleftedCoref* and *ClauseStatus* give the information status of the clefted constituent and the cleft clause respectively, and the *FocusType* feature specifies the kind of constituent focus—if any—that is used in the cleft.

Since the features of the clefts in the database are the fundamental building blocks that are used in calculating the behaviour of constructions that are, on the basis of the definition in (244), identified as *it*-clefts through the development of the English language, we will look at the features in subsequent sections in more detail. The examples used in the subsections are combined in (306).

- (306) a. It may have been you who stole the cutlery.
 b. It was Jesus who had made him well. [erv-new-1881:355]
 c. I think it was decorative art-needlework she took up. [wilde-1895:708]
 d. Did you know it was the Strafford you fired into? [holmes-trial-1749:1401]
 e. It is the first time I have been quoted as an authority by an eminent outsider. [thing-187x:28]
 f. It is because your husband is himself fraudulent and dishonest that we pair so well together. [wilde-1895:777]
 g. Ða næs long to þon in þæm westenne þæt we to sumre ea cwoman.
It then was not long after that, in the wilderness, that we came to some river. [coalex:101]
 h. It is not wantonly, nor altogether wilfully, that man has so often lost his God [talbot-1901:93]
 i. What sort of a brooch was it that you lost, Mrs Cheveley? [wilde-1895:600]
 j. It was not Moses that gave you the bread out of heaven; but my Father giveth you the true bread out of heaven. [erv-new-1881:464-5]
 k. It is only in that way that large waves can be affected. [strutt-1890:442]
 l. At these lectures T. H. Huxley sat by my side, and he it was who first directed my attention to their great interest and importance. [fayrer-1900:564]
 m. Even apart from estimates of pitch, an examination of the tones of the bells of the Terling peal proves that it is only from the third and fifth tones that a tolerable diatonic scale can be constructed. [strutt-1890:197]

12.2.1.1 CleftType

The *CleftType* gives the word order of the four main components of the *it*-cleft. While the definition of the cleft in (244) does not stipulate a particular word order for a construction to be an *it*-cleft, we know that the main word order in English changed from a particular variant of V2 to one that is broadly SVO, but without V2. If we keep track of the word order that is used for *it*-clefts, we will be able to see if the *it*-cleft sticks to one particular order, or follows the general changes in English. The four main components of the *it*-cleft are the following:

(307) *The four main components of the it-cleft*

- a. The **syntactic subject**. In Present-day English this usually is the pronoun *it*. Earlier forms of English may have an empty subject. If there is an overt subject, it is identified by ‘s’.
- b. The **finite verb**. The location of the finite verb is indicated by ‘v’. The finite verb usually is *was* or *is*, since clefts are built on copula constructions. But some clefts use modals, as for instance the one in (306a). In this situation the finite verb is *may*, and so its position is marked by ‘v’.
- c. The **clefted constituent**. The letter used to identify the location of the clefted constituent varies. The ‘o’ signals an NP complement as clefted constituent, the ‘a’ is used for adverbial phrases, the ‘p’ for prepositional phrases, and the ‘i’ for clausal complements (since these contain an IP).
- d. The subordinate **cleft clause**. This is headed by a label CP-CLF, CP-THT or CP-REL in the existing annotation of the corpora. The cleft clause position is marked by ‘c’.

It is only the word order of these four main components that is marked. The positions of conjunctions and sentence-level adverbials are not taken into account, because the amount of possible CleftType values would be too high, and we would not be able to make proper generalisations.

12.2.1.2 *CleftedCat*

The *CleftedCat* is the syntactic function fulfilled by the clefted constituent in the cleft clause, if it still were overtly present in that clause. The definition of the cleft in (244) allows for argument as well as adjunct functions, and the discussion in sections 10.1.2 and 10.2.5 suggest that adjunct functions may be more prominent in early *it*-clefts. It is unclear whether all argument functions would then rise at the same time or in a particular order. The clefted category is, therefore, an essential feature to take along in the cleft database. It can have the following values:

(308) *Possible values for “CleftedCat”*

- a. **Subject**. The clefted constituent, or its co-indexed gap, functions as subject in the cleft clause. The clefted constituent in (306b) is *Jesus*. It has a co-indexed constituent in the cleft clause, namely *who*. This constituent is the syntactic subject of the cleft clause.
- b. **Object**. The clefted constituent, or its co-indexed gap, functions as direct or indirect object in the cleft clause, such as in (306c). The clefted constituent is *decorative art-needlework*, and it fulfils the role of direct object in the cleft clause *she took up decorative art-needlework*.
- c. **PPobj**. The clefted constituent, or its co-indexed gap, functions as complement of a preposition. The preposition itself is still located in the cleft clause. The *Strafford* in (306d) is the name of a ship, and the court asks the accused whether he knew he was firing at an allied ship instead of at the enemy. The *Strafford* is a complement Noun Phrase within the main clause, but it is part of the PP *into the Strafford* inside the cleft clause.

- d. **NonArgNP**. The clefted constituent, or its co-indexed gap, is a non-argument Noun Phrase inside the cleft clause. The constituent *the first time* in (306e), for instance, also functions as temporal Noun Phrase in the corresponding cleft clause. While a temporal clause may be simply expressed as a Noun Phrase when it is the complement of a copula clause, this is not usually the case when it occurs in a fuller clause. A rephrasing of the cleft clause in (306e) has the temporal clause as part of a PP, i.e: *I have been quoted as an authority by an eminent outsider for the first time*.
- e. **Adjunct**. The clefted constituent, or its co-indexed gap, is a non-argument inside the cleft clause, and it is not a Noun Phrase. It can be a clause, such as in (306f), an adverbial phrase, such as in (306g), an Adjectival Phrase, or a Prepositional Phrase.

12.2.1.3 CleftedType

The *CleftedType* looks at the syntactic category of the complement XP in the copula main clause of the *it*-cleft construction. While the definition of the *it*-cleft in (244) does not demand the complement to be of a particular syntactic category, section 10.1.3 discusses predicational versus specificational clefts, which can be distinguished partly on the basis of the clefted constituent's syntactic category. If we want to know how much of the *it*-clefts are more predicational in nature than specificational, we need to keep track of the *CleftedType* feature.

If the clefted constituent is a Noun Phrase or a Prepositional Phrase that contains an NP, the *CleftedType* gives the type of NP, which can be *AnchoredNP*, *Bare* (as in 306c), *BareWithPP*, *DefNP* (as in 306d), *Dem*, *DemNP*, *FullNP*, *IndefNP*, *NumP*, *Pro* (as in 306a), or *Proper* (as in 306b).

If the clefted constituent is not a Noun Phrase, the *CleftedType* identifies its type as follows:

(309) Non-NP values for CleftedType

- a. **AdjP**. The clefted constituent is an Adjectival Phrase, as for instance *long to bon* in example (306g).
- b. **AdvP**. The clefted constituent is an Adverbial Phrase, as for instance *wantonly* in example (306h).
- c. **IP**. The clefted constituent is a whole clause, such as *because your husband is himself fraudulent and dishonest* in example (306f).

12.2.1.4 CleftedCoref

The *CleftedCoref* feature gives the information status of the clefted constituent. This feature is important, if we want to verify the claims made in the literature that clefts are used for focus or discourse, as discussed in sections 10.2.2 and 10.2.5. Texts that have been annotated for coreference using Cesax would already have the information status of all NP constituents available, but only few texts have been annotated so far (Komen, 2012).

The database supplies several preceding and one following context line, so that the cleft can be seen in its proper context, and the information status of the clefted

constituent can be determined more precisely. The following four information states are used to differentiate the status of the clefted constituents:

(310) *Possible values for CleftedCoref*

- a. **Assumed.** The clefted constituent is known to the hearer or reader, but it is new to the current discourse. The person *Moses* in (306j), for instance, is a well-known figure to the audience, but it has not been mentioned in the preceding context of the text.
- b. **Identity.** The referent of the clefted constituent is identical to the referent of a constituent in the preceding discourse. The referent of *that way* in (306k), for instance, refers back to a preceding clause, and the referent of *he* in (306l) is the same as that of *T.H.Huxley* in the preceding context.
- c. **Inferred.** The referent of the clefted constituent is not exactly identical to the referent of a constituent in the preceding discourse, but it does relate to one, as for instance *third and fifth tones* in (306m) stand in a part-whole relation to *the tones of the bells* earlier.
- d. **New.** The referent of the clefted constituent is new in the discourse, and new to the hearer (as far as we can judge). An example could be *the first time* in (306e).

12.2.1.5 *ClauseStatus*

The information status of the cleft clause is kept in the feature *ClauseStatus*. Sections 10.2.2 and 10.2.5 discuss that the information status of the cleft clause, in combination with that of the clefted constituent, can be used to define the function of an *it*-cleft. We need to keep track of it in the database, so that we can verify those claims.

The information status of the cleft clause is something that cannot be determined automatically, but needs the annotator's judgment after careful reading of the context. In order to limit the number of possible combinations, only three possible states are allowed for: *New*, *Known*, and *Inferred*.

Cleft clauses marked with a *ClauseStatus New* are not linked to anything in the preceding discourse, whereas those marked with *Known* basically repeat a previously mentioned idea. Those marked with *Inferred* somehow link to the preceding discourse, but do not really repeat an idea mentioned there.

Marking whole clauses for information state is a careful manual process, but the clausal status needs to be taken into consideration in order to check the validity of existing historical accounts, such as the claim that stressed-focus *it*-clefts came up earlier, while informative presupposition ones followed later (see section 10.2.2).

12.2.1.6 *FocusType*

The *FocusType* feature is used to keep track of clefts that are used for constituent focus, and those that are not. This distinction is an important one, for instance to verify the claim that the more basic, and therefore historically earlier, function of *it*-clefts is that of thematization instead of focusing (see section 10.2.5), contrary to, for instance, Ball (1991). If *it*-clefts start to be used to express constituent focus only

later, then the question is whether particular kinds of constituent focus come into the picture earlier or later. This is why we do not only have to keep track of the fact that an *it*-cleft is used for constituent focus or not, but also what kind of constituent focus is used. The following Values may appear for the *FocusType*:

(311) *Possible values for FocusType*

- a. **Time.** A time-cleft is one where the clefted constituent is a temporal Noun Phrase or Prepositional Phrase.
- b. **Wh.** This type is chosen whenever the clefted constituent is or contains a question word.
- c. **Emph.** The focus type *Emph* is reserved for NP or PP with emphatic prominence. Such prominence is sometimes expressed through adverbs (e.g. *just* twenty years, *right* in the middle), sometimes through positive negation (e.g. *not without many tears shed on both sides*), and sometimes through other means (see sections 3.2.2.3 and chapter 9).
- d. **Reason.** A reason cleft links to the preceding discourse through a logical function such as *therefore*, *hence*, *because* etc.
- e. **Contrast-Adv.** The clefted constituent NP or PP is contrastive due to the presence of a focus particle such as *only* within the constituent (e.g: *only* in that way, as in 306k).
- f. **Contrast-Foll.** The clefted constituent contrasts with an element in the following context.
- g. **Contrast-Neg.** The clefted constituent is contrastive as a result of a negation within its NP or PP. The constituent *not Moses* in 306j, for instance, implies that there is at least one other person that *gave you the bread out of heaven*, and contrasts this person explicitly with Moses.
- e. **Contrast-Pre.** The clefted constituent contrasts with an element in the preceding context.
- h. **Contrast-Same.** The clefted constituent contains an explicit contrast within itself (e.g: *not the perfect but the imperfect*).
- i. **None.** These are all clefts that do not belong to any of the focus types mentioned in (a)-(i).

The *FocusType* feature is one that can be determine with relatively high objectivity, since all possible values are based on the presence or absence of particular NP types, adverbials, logical connectors etc. The *it*-clefts in the database only receive the *FocusType* of *Contrast*, for instance, when there is explicit contrast, or when a contrastive adverb is used as modifier.

12.2.2 Gathering initial data for the cleft database

Since the total number of clefts identifiable by the label CP-CLF is slightly over 700, and since much of the features that need to be gathered for each cleft are derivable from the syntactically annotated corpora, I have chosen to use the computer program CorpusStudio to not only collect all clefts, but also provide them with as much initial information as possible automatically. Such a procedure reduces the amount

of manual work, and the errors that are associated with it. This section describes what procedure has been followed, and what data have been gathered.

All subordinate clauses that were regarded as belonging to clefts have been tagged with the label CP-CLF instead of CP-THT, which refers to a complement clause, or CP-REL, which refers to a ‘normal’ relative clause. Figure 35 illustrates how a typical cleft is coded.³ This cleft is part of a main clause, identified by the label IP-MAT. The subject NP-SBJ consists of a pronoun *it*, and the clefted constituent is the complement NP-OBJ *only the successful workers*. The subordinate clause is contained within the CP-CLF. The relativizer is the *wh* pronoun *who*, which is encoded as WNP-1. The number “1” serves as coreference tag with the NP-SBJ subject trace *T*-1 inside the subordinate clause IP-SUB. The complementizer C is 0, due to the presence of the relativizer pronoun.⁴

```
(IP-MAT
  (CONJ but)
  (CONJ neither)
  (BEP is)
  (NP-SBJ (PRO it))
  (NP-OBJ (FP only) (D the) (ADJ successful) (NS workers))
  (CP-CLF
    (WNP-1 (WPRO who))
    (C 0)
    (IP-SUB
      (NP-SBJ *T*-1)
      (BEP are)
      (NP-OBJ (D the) (ADJS best) (NS performers))))
  (. .))
```

Figure 35 Coding of a typical cleft

The code in (312) would be enough to select all the main clauses and subordinate clause that contain a cleft.⁵ Line #1 looks for all constituents, which in *xml* are `<eTree>` elements, that have the `@Label` that matches the string `$_finiteIP`. This string defines the main and subclauses as `IP-MAT*|IP-SUB*`. Whenever a finite clause is found, line #3 stores the first child of this clause that has a label matching `$_anyCLF` in the variable `$cp`. Line #5-7, then, state that wherever such a `$cp` actually exists, the finite clause should be returned as a result.

(312) *Xquery code to find clefts in main and subordinate clauses*

```
1   for $search in //eTree[ru:matches(@Label, $_finiteIP)]
2
3     let $cp := tb:SomeChild($search, $_anyCLF)
4
5     where ( exists($cp) )
6
7     return ru:back($search)
```

While the procedure in (312) correctly identifies all 716 clefts encoded in the parsed English corpora, we would like to calculate and add as many features for each cleft automatically as possible. This is done in the procedure given in (313). The variable `$loc` obtains a simplified one-letter-per-constituent overview of the layout of the finite clause (e.g: svpc = subject, verb, prepositional phrase and CP).

(313) *Xquery code that finds clefts and adds features automatically*

```

1  for $search in //eTree[tb:HasLabel(@Label, $_finiteIP)]
2    let $cp := tb:SomeChild($search, $_anyCLF)
3
4    (: Get the location-layout of this whole IP :)
5    let $loc := tb:Location($search, 'detailed')
6
7    (: Get the category of the clefted constituent and its NP type :)
8    let $cat := tb:CleftedCat($cp)
9
10   (: Get the actually clefted constituent and its NP type :)
11   let $clf := tb:Clefted($cp)
12   let $npt := tb:PhraseType($clf)
13
14   (: Guess the coreference type, depending on the NP type :)
15   let $cor := tb:GuessCoref($clf)
16
17   (: Prepare the field-contents for possible database output :)
18   let $db:=concat($loc,';', $cat, ';',
19                 tb:Constituent($clf), ';', $npt, ';', $cor)
20
21   where (
22     exists($cp)
23   )
24
25   (: Output subcategorizes in [$cat] and database output in [$db] :)
26   return ru:back($search, $db, $cat)

```

The function `tb:CleftedCat` looks at the syntactic label of the first trace it finds in the cleft clause, and fills `$cat` with ‘Adjunct’, ‘Subject’, ‘Object’, ‘NonArgNP’ or ‘Other’, depending on what it finds.

The function `tb:Clefted` tries to determine what the clefted constituent is, by looking at the syntactic category of the trace in the cleft clause. These overlap for clefted adjunct or non-argument NPs, but for subject clefts there is no overlap: the clefted constituent is marked with NP-OB1 as complement in the finite clause, while its trace is marked with NP-SBJ as subject in the cleft-clause. When this case is taken into consideration, the clefted constituent can be found with relatively high certainty.

The kind of coreference relation (e.g. `Identity`, `Inferred`, `New`) that the clefted constituent has can only be guessed. The function `tb:GuessCoref` looks at the NPtype of the clefted constituent, and comes up with a suggestion for the coreference relation. Pronouns, demonstratives and traces, for instance, are likely to have an `Identity` coreference relation.

Line #18 prepares a string with the cleft’s feature values separated by semicolons. This string is made available to the CorpusStudio engine in line #25, and is used to make a database of clefts after the query has selected the clefts from the parsed English corpora.

The database that results from the query in (313) is imported into CESAX (Komen, 2011a). This program facilitates evaluating each cleft within its context, and editing the features that were automatically assigned to it. Figure 36 shows the database facilities of Cesax in action, in order to provide a better idea of how the editing of the cleft features was done.

The screenshot displays the CESAX Editor interface for editing cleft features. The main window is titled "CESAX: Editor for syntactically annotated corpora". The interface includes a menu bar (File, Edit, View, Section, Translation, Corpus, Reference, Must, Tools, Help) and a toolbar. The main workspace is divided into several panels:

- Database file:** CleftBasic_V1
- Database of results created from queryline:** InCLF. Feature values have subsequently been manually adapted.
- Selected feature:** Focus Type
- Text:** |wide-1895
- Location:** |70.861
- File:** D:\Data Files\Corpora\English\xml\Adapted\WBE\wide-1895.psd
- Period:** |B3
- Forecast:** |861
- eTread:** |18061
- Category:** Subject
- Status:** Done
- Syntax:** (IP-MAT-1 (NP-SEJ (PRO it)) (REP is) (NP-OBJ (NP (NEG not) (D the) (ADJ perfect)) (, , ,) (CONJP (CONJ but) (NP (D the) (ADJ imperfect)))) (, , ,) (CP-CLF (NP-2 (NPBO who)) (C O) (IP-SUB (NP-SEJ *TT-2) (HVP bare) (NP-OBJ (N need)) (P of) (NP (N Love)))) (, . .))

The main text area contains the following text:

[wide-1895] [70.858] We have all feet of clay, women as well as men.[70.859] but when we men love women, we love them knowing their weaknesses, their follies, their imperfections, [70.860] love them all the more, it may be, for that reason.
[70.861] It is not the perfect, but the imperfect, who have need of love.
 [70.862] It is when we are wounded by our own hands, or by the hands of others, that love should come to cure us.

Below the text is a table of user-adaptable features:

User adaptable features	Cleft Type	CleftedCat	CleftedText	CleftedType	CleftedCoref	ClassStatus	Focus Type
svoc	svoc	Subject	not the perfect, but the imperfect	DefNP	Inferred	Inferred	Contrast:Same

At the bottom, a table lists the results from the database:

Ret	Textid	Cat	fores	Para	Select	Status
609	lyell-1830	Adjunct	267	B2	Contrast:Same	Done
621	turner1-1799	Subject	83	B2	Contrast:Same	Done
649	whewell-1837	Adjunct	13	B2	Contrast:Same	Done
666	benson-1908	Subject	74	B3	Contrast:Same	Done
683	bradley-1905	Subject	89	B3	Contrast:Same	Done
728	oman-1895	Subject	405	B3	Contrast:Same	Done
765	wide-1895	Subject	400	B3	Contrast:Same	Done
771	wide-1895	Subject	861	B3	Contrast:Same	Done
8	conicoda	Adjunct	539	O14	Emph	Done
9	convinal	Subject	129	O14	Emph	Done
24	cobede	Adjunct	1704	O2	Emph	Translation
45	cobede	Adjunct	3400	O2	Emph	Translation
62	coboeth	NonAgNP	1835	O2	Emph	Done
65	cooura	Adjunct	2389	O2	Emph	Done
74	coalex	Adjunct	101	O23	Emph	Done
80	coblick	Adjunct	2737	O23	Emph	Done
81	comar3	Subject	493	O23	Emph	Done
146	cmcloud	Subject	617	M3	Emph	Translation
155	cmrtest	PParg	351	M3	Emph	Done
159	cmrtest	Subject	509	M3	Emph	Done
170	cmwycser	NonAgNP	1417	M3	Emph	Done
173	cmwycser	PParg	3875	M3	Emph	Explanation
188	cmcapstr	Subject	2101	M4	Emph	Done

Figure 36 Editing of cleft features within Cesax

The left hand side of the “CorpusResults” tab page of the program Cesax contains a listbox with information that can be used to identify and sort the *it*-clefts in the database, such as the time period abbreviation, the name of the text that is used, and the *CleftedCat* feature (see section 12.2.1.2). Once a cleft has been selected in the listbox, the right hand side of the “CorpusResults” tab page shows the context of the cleft, its syntactic make-up, and all the features belonging to it. These features can be edited, and there is room for additional notes, such as the reasons why particular feature values have been assigned.

12.2.3 Identifying additional candidates for the *it*-cleft database

We saw in section 10.1 that there is no consensus in the literature about which constructions should be called clefts, and which not. If the annotation scheme used by the creators of the parsed English corpora followed the more restrictive definition of clefts, e.g. the one in (228), or Ball’s (1994) reasoning explained in section 12.1, it is quite possible that there are some constructions that should be labelled as CP-CLF by our definition, but have not received that label in the parsed corpora. There are two categories of constructions that need to be inspected to see if they are clefts after all.

The first category involves copula clauses with a subordinate clauses that has been identified as a complement clause (CP-THT or CP-THT-x) rather than as a cleft relative clause. The second category involves copula clauses where the subordinate clause has been identified as an extraposed relative clause (coded as CP-REL-1) rather than as a cleft relative clause.

12.2.3.1 Locating additional candidates for *it*-clefts

The procedure in (314) shows how constructions that have not been labelled as clefts, but could potentially be one, can be found. Line #1 selects constituents labelled with `$_anyCLFq`, which is defined as `CP-THT|CP-THT-x|CP-REL-1`. Line #4 identifies the main or subordinate clause level that such potential clefts are part of, and line #7 looks for the subject of that clause. Line #8-10 adds an additional test for the subject: it should either be a pronoun, or empty. This gives us more than we are actually looking for; we get *all* pronoun and *all* empty subjects instead of just the *it* pronoun subjects and the expletive kind of empty subjects. This is done to allow for the wide variation in how the pronoun *it* has been written in earlier forms of English.

Line #13 looks for a form of the verb *be*, and line #14 checks the presence of any unwanted elements in the clause (that is: past participles, infinitive clauses and adjectival phrases).

(314) *Xquery code to find complement clauses that might be clefts*

```

1 for $search in //eTree[ru:matches(@Label, $_anyCLFq)]
2
3   (: Find the clause to which this cleft belongs :)
4   let $cls := $search/parent::eTree[ru:matches(@Label, $_finiteIP)][1]
5
6   (: Get the subject :)
7   let $subj := tb:SomeChildNo($cls, $_subject, $_nosubject)
8   let $subjOk := if (tb:SomeChild($subj, 'PRO*')
9                     or tb:IsStarred($subj)) then true()
10                    else false()
11
12   (: A cleft IP has a form of "BE", but no other verb or participle :)
13   let $be := tb:SomeChild($cls, $_any_BE)
14   let $van := tb:SomeChild($cls, "VAN|IP-INF|ADJP*")
15
16   where (
17     exists($cls)      and
18     not(exists($van)) and
19     exists($be)       and
20     $subjOk
21   )
22   return ru:back($cls)

```

The actual code used to find potential clauses with a cleft that has not been recognized as such adds lines like 4-18 in (313), which are used to make the result into a database. The items found in the resulting database are then evaluated line by line against the cleft diagnostics defined in section 10.1.7. Those that pass all diagnostics described in 10.1.7 are added to the database that was created with the procedure described in section 12.2.2.

12.2.3.2 *Clefts tagged as complement clauses*

With the procedure to find complement clauses that might be clefts after all as defined in (314), the corpus research project gives the results as shown in Table 41. Almost all clefts that have been tagged as complement clauses in the parsed English corpora are those where the clefted constituent has an Adjunct role within the cleft clause. The majority of them are found in the Middle English and Early Modern English periods, and most of these are Reason clefts.

Table 41 *Clefts that were mistakenly taken for complement clause constructions*

Period	Dates	Found	Added	Texts	Adjunct	Time	Reason	Contrast
OE	450-1150	574	5	4	100%	80% (4)	-	-
ME	1150-1500	254	23	14	100%	26% (6)	57% (13)	17% (4)
eModE	1500-1700	256	26	11	100%	15% (4)	65% (17)	15% (4)
LmodE	1700-1914	117	12	10	92%	0% (0)	25% (3)	67% (8)

In order to be sure that the reclassification as *it*-clefts is justified, we will have a closer look at some of the constructions that have been found. Example (315) is an Old English instance of a construction where the subordinate clause has been tagged as a complement clause. However, the clefted constituent *þa* can be understood perfectly well as serving a temporal adjunct role in the cleft clause. The temporal PP *in the tyme of Cambises* from the Middle English example in (316) can likewise serve as a temporal adjunct in the cleft clause.

- (315) (And he wæs, se ylca Tyrus, þæs ðe bec secgað, swa unhal on hys andwlitan, þæt ðæt adl, þe we hatað cancer, hym wæs on þam nebbe fram þam swyðran næsþyrle, oð hyt com to þam eage.)
 Ac hyt wæs **þa**, þæt sum man wæs farende of Iudealande,
 but it was then that some man was going from Judea land
 þæs nama wæs Nathan, ... [covinsal:5-6]
 that-GEN name was Nathan
(And he was the same Tyrus, of whom the book says he had a disease on his skin, and that he had this disease we call cancer on his face—from his right nostril until his eye.)
*'It was **then** that a certain man came from Judah, whose name was Nathan.'*
- (316) **But in the tyme of Cambises** was it y=t= the werke of god. this buyldynge went not forwarde but lettyd was it by fals accusars whyche neuer cesse in the chyrche of Cryste to lette the werkes of god as dayly experyence dooth shewe. [cmfitzja:104]
*'But it was **in the time of Cambises**, that the work of God—this building—did not go forward, but it was denounced by false accusers, which never cease to exist in the church of Christ to denounce the work of God, as is shown by daily experience.'*

Temporal clefts decrease by the early Modern English time period, but (317) is an example of a reason cleft that was not recognized as such by the corpus annotators. The reason adjunct *hence* 'for that reason' can perfectly well be thought of as being an adjunct to the cleft clause.

- (317) (For every thing which is said to be imperfect is proved to be so by the Diminution of that which is perfect.) [boethpr-e3-h:92-3]
Hence it is that if any thing in any kind be said to be imperfect, it is presently understood that in it there is also something perfect.

There are fewer unrecognized clefts from late Middle English onwards. One example of a cleft missed out is (318):

- (318) **Why** is it that men are more angry at being accused of bad reasoning than of erroneous opinions?
 (Clearly because all these faults imply an incomplete and ill-conducted cultivation of the speculative faculty, in reference to language or to reasoning.) [whewell-1837:165]

The clefted constituent *why* is a reason adjunct in the cleft clause. I have classified clefts where the clefted constituent is a *wh*-phrase as *contrastive* clefts, since *wh*-phrases have a kind of constituent focus that strongly implies the existence of a limited set of alternatives, which is often (but not always) contrasted with one particular choice in the following context. In example (318) the *wh* variable is filled in by the *because* phrase in the following sentence.

12.2.3.3 Clefts tagged as relative clauses

The difference between relative clauses modifying their head noun and relative clauses that are part of an *it*-cleft can be quite tricky, because a wider context is needed to disambiguate the two. What helps in distinguishing the two is the *Cleft pronoun* diagnostic as defined in (246). Whenever we are dealing with a “real” relative clause, the *it* pronoun is anaphoric—it refers back to an antecedent. When even that diagnostic becomes difficult, then the flipside of it can be used: if the clefted constituent and the cleft clause in the construction we are looking at form one tight meaningful unit, then it is not a real cleft.⁶

The procedure to identify potential cleft candidates that have been tagged as relative clauses follows the general procedure illustrated in section 12.2.3.1. The only difference is that the CP we start looking for in line #1 of (314) should not be the CP-THT or CP-THT-x, but it should now be the CP-REL or the CP-REL-1.

With this correction in place, the corpus research project that looks for additional *it*-cleft candidates that are annotated as relative clauses gives over 140 results, of which 26 passed the diagnostics and were added as genuine clefts to the database. The results in Table 42 show that only one fifth of the constructions found by the algorithm were indeed added as genuine *it*-clefts.

Table 42 Clefts that were mistakenly taken for simple relative clause constructions

Period	Dates	Found	Added	Texts	Adjunct	Time	Neutral	Contrast
OE	450-1150	54	2	2	0%	0%	0%	100% (2)
ME	1150-1500	27	10	6	10% (1)	0%	40% (4)	60% (6)
eModE	1500-1700	51	14	9	7% (1)	7% (1)	7% (1)	86% (12)
LmodE	1700-1914	9	0	0	-	-	-	-

A few examples suffice to illustrate constructions that were tagged as relative clauses by the corpus annotators, but have been added to the database as *it*-clefts, since they meet the criteria. The Old English example (319) has *by the Holy Spirit* as clefted constituent, which perfectly well fits as manner adjunct in the cleft clause. The pronoun *hit* ‘it’ is not anaphoric, and therefore passes diagnostic (246).

The early Modern English example (320) should be regarded as *it*-cleft, since it has a main clause structure complying with (245), the clefted constituent coindexes with the subject of the cleft clause, complying with (247), and the pronoun *it* is non-anaphoric, complying with (246).

- (319) (Ioseph, be not aferd to take Mary, þy wyfe, ynto þy kepyng.)
 hitys of þe Holy Gost þat ys qwyk yn hur, [cmmirk:2958]
 it is of the holy spirit that is pregnancy in her
 Werfor þou schalt be hur keper and norish to hur chyld.
 wherefore you will be her caretaker and nourish to her child
 ‘(Joseph, do not be afraid to take Mary, your wife, under your care.)
 It is **by the Holy Spirit** that she is pregnant, which is why you will have to
 take care of her and her child.’
- (320) So I suppose 't is **his Quality more than his Love**, has brought him into
 this Adventure. [vanbr-e3-p1:17]

The procedure described in this section has resulted in a database of *it*-clefts, which are well defined by the formal criteria in 10.1. They are the fairest representation of the *it*-clefts used in English, during the time it developed into its current form. The next section will tune in on this development.

12.3 Results from the historical cleft database

The previous sections have described how the *it*-clefts from the parsed English corpora have been identified and combined into a database. Each *it*-cleft entry in the database has a number of features associated with it, as described in section 12.2.1. This section takes these features as a basis to describe the history of the *it*-clefts, concentrating on the relation of the *it*-cleft construction to information structure. What we will see is that the *it*-cleft was mainly used for text-organization purposes in OE, and that its rise as a strategy to convey constituent focus coincides with the loss of V2 in English.

12.3.1 The number of *it*-clefts in English time periods

Clefts in general are not a very frequent occurrence, even in Present-day English. They are a marked construction that is used only in specific situations (see 10.2). Table 43 lists the absolute number of *it*-clefts found in the parsed English corpora by the procedure described in the previous sections. It also lists the number of clefts per 100,000 main clauses (denoted as *it*-cleft*).

Table 43 Number of *it*-clefts found in the parsed corpora

	O1-2	O3-4	M1-2	M3-4	E1	E2	E3	B1	B2	B3
Period	450- 950	950- 1150	1150- 1350	1350- 1500	1500- 1570	1570- 1640	1640- 1700	1700- 1770	1770- 1840	1840- 1910
Clauses	20411	96214	24398	54960	28194	34614	24944	15424	20326	17201
<i>it</i> -cleft	73	29	30	92	40	65	142	96	118	114
<i>it</i> -cleft*	358	30	123	167	142	188	569	622	581	663

Since the total number of clefts for certain periods was below a level to get much significance, the Old English sub periods O1, O2, O3 and O4 have been combined into two sub periods O12 and O34. The same was done for the Middle English periods. Figure 37 shows this general trend of *it*-clefts graphically.

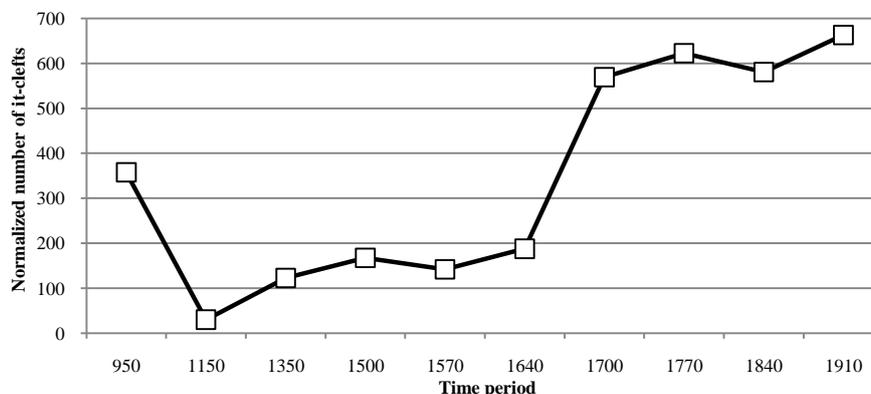


Figure 37 Number of *it*-clefts normalized per 100,000 main clauses

The numbers clearly illustrate the marginal character of the *it*-cleft in older variants of English. It is only after 1640, which is the end of the early Modern English time period, that the use of *it*-clefts increases significantly.⁷

It is also clear from Figure 37 that Old English had a relatively large number of *it*-clefts. Section 10.2.5 describes that the reason for this behaviour is the function fulfilled by *it*-clefts in the partitioning of texts. The next sections will quantify this idea in terms of the syntactic and information structural distribution of the *it*-clefts per time period.

12.3.2 Syntactic features

A number of syntactic features have been stored with each cleft, and this section follows the behaviour of clefts based on one of those features.⁸

12.3.2.1 Category of the clefted constituent

The first feature that warrants closer inspection is the syntactic category of the clefted constituent. This feature can be used to see how the percentage of clefts with an argument gap in the relative clause behaves with respect to those where there only is an adjunct “gap” (that is, where the clefted constituent has an adjunct role in the cleft clause). Figure 38 shows the make-up of the clefted constituent in terms of its syntactic category. This figure divides the time periods in the four main ones: Old English (OE), Middle English (ME), early Modern English (eModE) and late Modern English (LmodE).⁹

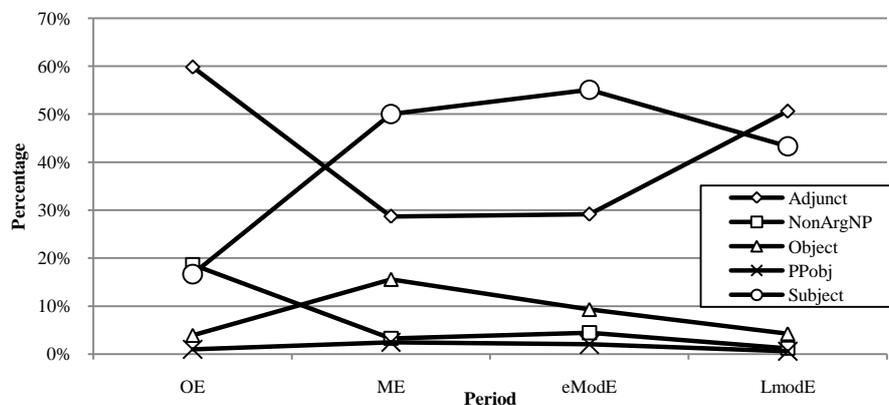


Figure 38 Syntactic category of the clefted constituent

The argument-adjunct division can be observed by looking at the lines marked “Adjunct” and “NonArgNP”. Old English starts out with a majority of Adjunct clefts (60%), and it also has a robust percentage of non-argument noun phrase clefts (20%). The number of adjunct clefts decreases to a minority of 30% in ME and eModE, while it rises again to 50% in late Modern English. The OE clefts with an adjunct gap (as well as those with a NonArgNP gap) are mainly time-clefts, as shown for example in (321). The LmodE clefts with an adjunct gap tend to function more as means and reason, as for instance (322).¹⁰

- (321) (And he wæs, se ylca Tyrus, þæs ðe bec secgað, swa unhal on hys andwlitan, þæt ðæt adl, þe we hatað cancer, hym wæs on þam nebbe fram þam swyðran næsþyrle, oð hyt com to þam eage.)
 Ac hyt wæs þa, þæt sum man wæs farende of Iudealande,
 but it was then that some man was going from Judea land
 þæs nama wæs Nathan. [covinsal:6]
 this-GEN name was Nathan
 (He was the same Tyrus of whom the book says that he had the disease on his face, which we call cancer, from the right nostril until it his eye.)
 It was **then** that a certain man was coming from the land of Judah, whose name was Nathan.

- (322) But it is **only by some means of this kind** that private ills, in such a lawless community, can be made public wrongs. [reade-1863:432]

Argument clefts—clefts with a subject or object gap in the cleft clause—are relatively rare in Old English, but they are attested, witness the “Subject” and “Object” lines in Figure 38, of which (323) is an example.¹¹

- (323) *Þa cwæð þæt wif him to þæt hitwære Swyðun,*
 then said that woman him to that it was Swithun
se ðe hine lærde mid þære halgan lare and þone
 who that them taught with their holy teaching and whom
ðe he geseah on ðære cyrcan swa fægerne. [coelive:4463]
 that he had.seen in their church so glorious
Then the woman told him (=her husband) that it was Swithun who had
instructed him with this holy teaching, and whom he had seen so glorious
in the church.

Saint “Swithun” has appeared in a dream to a bedridden man, and requested this person come to Winchester. The man doesn’t know who has appeared to him, but he relates his dream to his wife, who subsequently suggests the identity of this stranger.

12.3.2.2 Position of the clefted constituent

Another syntactic feature that could be of interest to look at is the word order of the cleft constructions that were found. Instances where an *it*-cleft contains a question word as clefted constituent are not of interest, since all of these necessarily have the question word as first constituent. The remaining word orders can be divided into those where the clefted constituent precedes the copula, and those where it follows after the copula. Figure 39 shows how the percentage of *it*-clefts where the clefted constituent *precedes* the copula changes over time.

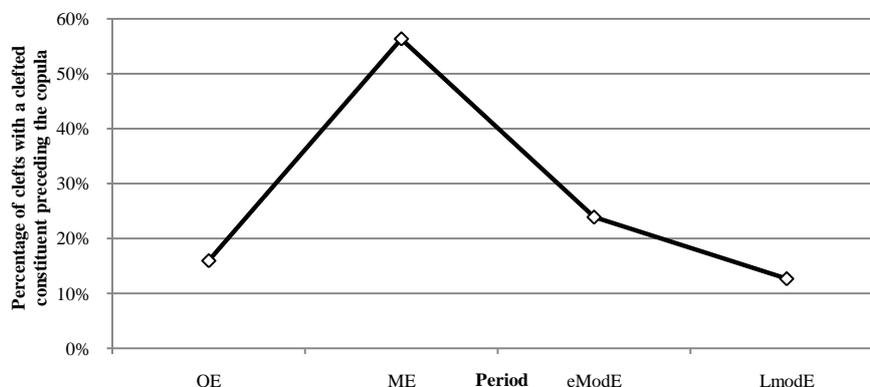


Figure 39 Clefted constituents preceding the copula

The trend towards SVO, with “S” representing the pronoun *it*, and “O” the clefted constituent, is clearly visible from ME until LmodE.¹² What is perhaps unexpected is the low percentage of *it*-clefts where the clefted constituent precedes the copula in Old English. This means that OE has a relatively large number of clefts where the clefted constituent *follows* the copula. Most clefts in OE that follow this pattern are of the type exemplified in (324).

- (324) (Æfter dissum wæs æfterfylgendre tide sum cneoh̄t in þæm mynstre in Beardan ea in longre lenctenadle hefiglice swenced.)
 Þa wæs **sumedæge**, þætte he sorgende bæd hwonne
 then was some day that he worrying asked when
 seo adl to him cwome. [cobede:1879]
 that fit to him would.come
 (SOME time after, there was a certain little boy in the said monastery,
 who had been long troubled with an ague.)
 He was **one day** anxiously expecting the hour that his fit was to come on,
 (when one of the brothers, coming in to him, said, "Shall I tell you, child,
 how you may be cured of this distemper.")¹³

This type of *it*-cleft construction starts with a temporal adverb *þa* ‘then’, and, due to the V2-character of OE, is followed by a finite form of the copula. The clefted constituent (usually a temporal NP or PP) and finally the cleft clause follows. These kinds of clefts do not have an overt pronoun *it*. The OE clefts that do have an overt *it*-pronoun, such as (323), also have the clefted constituent following the copula.

In sum, the word order of the *it*-clefts reflects the general trend in the history of English to prefer SVO, even for copula constructions, where the “O” is not a direct object, but a complement. The fact that the clefted constituent tends to follow the verb in the main clause of the *it*-cleft means that it is in a position where, according to Chapter Part IV, it receives unmarked focus.

12.3.3 Information status

The *it*-clefts in the database are annotated for information status of the clefted constituent as well as that of the cleft clause. Figure 40 shows how the information status of the clefted constituent behaves diachronically.

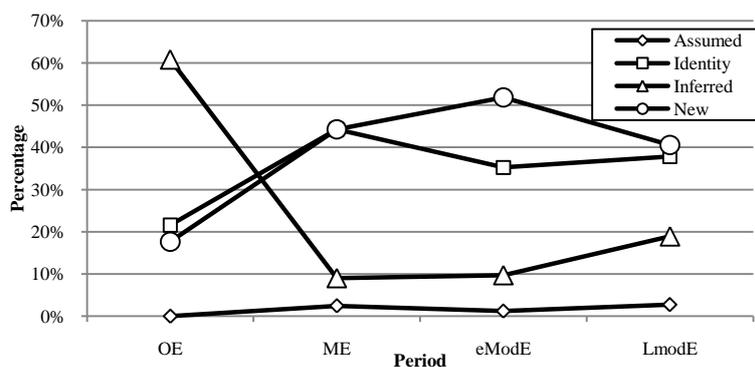


Figure 40 Information status of the clefted constituent

Old English starts with a high percentage of clefts where the clefted constituent has the status of “Inferred”.¹⁴ Those are the time-clefts, which have been annotated as “Inferred”, because the time reference builds on something in the preceding context, but cannot be identified as identical in reference with an earlier constituent. The

information status of the clefted constituent does not undergo large changes in the period between Middle English and late Modern English.

The clausal parts of the *it*-clefts too have been annotated for information state. Figure 41 shows how the information state of the cleft clause changes in time. There is a steady trend from Old English into late Modern English for referentially “New” cleft clauses to decrease.¹⁵

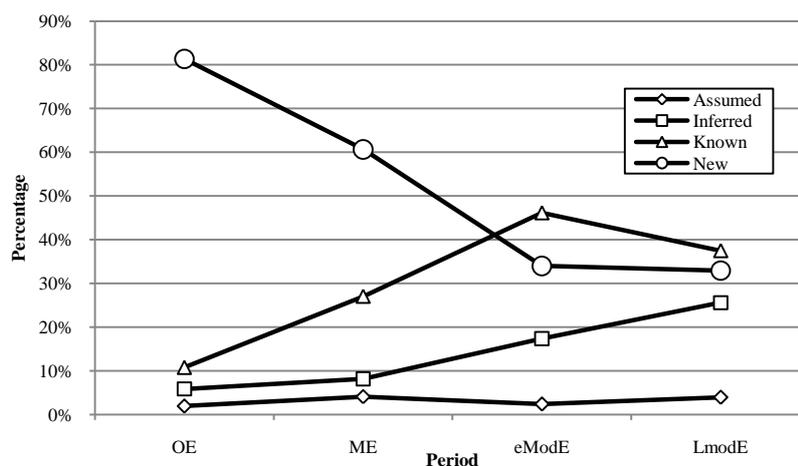


Figure 41 Information status of the cleft clause

Cleft clauses with non “New” information state increase as a result of the relative decrease of “New” ones. As explained in section 10.2.2, clefts with a referentially “New” cleft clause have been identified as a group quite early, and are generally known as “Informative Presupposition” clefts (Hedberg, 2007, Prince, 1978). An example of the latter from the cleft database is (325).

- (325) And as the winter wore on, tidings of the difficulties of transit from Balaclava to the Heights reached us, and at last the road was made. It was **in the middle of all these difficulties** that Palmerston had become Prime Minister, and that Mr. Roebuck urged on his committee.
[Trollope-1882:149-151]

The decrease of referentially “New” cleft clauses illustrates the gradual development of the prototypical Present-day English cleft, which contains a “Known” presupposition in its cleft clause. Apparently such clefts were rare (20%) in Old English, increased in Middle English (40%), and became the majority in early Modern English (65%) and late Modern English (66%).

Example (326) has a LmodE *it*-cleft where the information status of the cleft clause is “Known”, illustrating the prototypical *it*-clefts with a known presupposition (and a referentially new clefted constituent).

- (326) The wind abated a little, but the sea was terrible, the ship rolling heavily and going very slowly, for the engines were hardly working: it was **chiefly by press of canvas** we got on. [fayrer-1900:491]

Example (327) shows a typical OE *it*-cleft where the information status of the cleft clause is “New”. The information in the cleft clause that Augustine ordained two bishops is a completely new development at this point of the story.

- (327) Da wæs **æfter** **dissum** þætte Agustinus Breotone ærcebiscop
 then was after this that Augustine Britain’s archbishop
 gehalgade twegen biscopas. [cobede:976]
 ordained two bishops
 ‘After this Augustine, archbishop of Britain, ordained two bishops.’

Instead of only looking at the information status of the clefted constituent or that of the cleft clause, we could combine the two. In line with the synchronic research done by others, the information states are reduced to two values, by the following procedure:

- (328) *Information states of clefted constituent and cleft clause*
- Clefted Constituent.** The referential state of the clefted constituent, as described in 12.2.1.4, but reduced to either *Known* or *New*. A clefted constituent is *Known* if its status is IDENTITY, INFERRED, or ASSUMED, and a clefted constituent is *New* in all other situations.
 - Cleft clause.** The referential state of the cleft clause, as described in 12.2.1.5, but reduced to either *Known* or *New*. A cleft clause is *Known* if its status has been annotated as KNOWN or INFERRED, and it is *New* otherwise.

The approach of combining two values for the referential states of the clefted constituent and that of the cleft clause leads to the four categories of clefts which are shown in Table 44.

Table 44 Cleft type categories

Cleft Type	Clefted Constituent	Cleft Clause
Topic-Comment	Referential	New
Comment-Topic	New	Referential
Comment-Comment	New	New
Topic-Topic	Referential	Referential

The *it*-clefts that have a *wh* element in the clefted constituent, are almost always Comment-Topic clefts, which is why they have been excluded. The *wh* element contains the questioned, hence unknown information, and this new information is normally questioned against the background of known information in the cleft clause.¹⁶ Leaving aside the clefts with *wh* elements, the division of cleft types develops as shown in Figure 42.

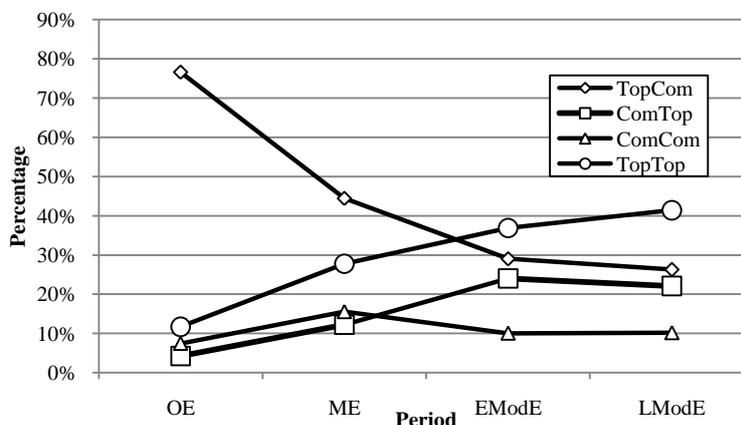


Figure 42 Combined information states of clefted constituent and cleft clause

The development of clefts with a referentially new clefted constituent (the Comment-Topic and Comment-Comment types) does not seem to be very significant, but the other two types show a steady progression.¹⁷ Both types involve a clefted constituent that is marked as “Topic”, which means that it somehow relates to the preceding context—either very specific (as in “Identity”), or by inference from something that has been mentioned, or through discourse-new, but hearer-old information (as in “Assumed”). The Topic-Comment *it*-clefts usually are those where an adjunct in the clefted constituent provides a backward link (e.g. a time link like *then* or a reason link like *therefore*), while the information in the cleft clause is new, and introduces a line of thought that is then pursued. Such clefts function as ideal text dividers (see 10.2.5). They are what others have labelled the Informative-Presupposition clefts.

The Topic-Topic clefts (together with the Comment-Topic ones) are slowly taking over from the Topic-Comment ones. These clefts are what others have labelled the Stressed-Focus ones. Such clefts link their clefted constituent to the preceding context, while the cleft clause also contains information that is already known, that is presupposed. The main characteristic of such clefts, then, is that of providing constituent focus on the clefted constituent, which is embedded in an already known context. The next section takes this observation a step further.

12.3.4 Information structure status

The study described in Los and Komen (2012) is based on a subset of the data that are now available. This current section uses the same procedure as described in Los and Komen, but now on all the data available from the cleft database. Los and Komen look at the development of the cleft construction based on three simplified features, which are derived from the full features.

(329) *Information structure states components*

- a. **Clefted Constituent.** The referential state of the clefted constituent, as described in 12.2.1.4, but reduced to two values: Referential and New. A clefted constituent is *Referential* if its status is IDENTITY, INFERRED, or ASSUMED, and a clefted constituent is *New* in all other situations.
- b. **Cleft clause.** The referential state of the cleft clause, as described in 12.2.1.5, but reduced to the values Referential and New. A cleft clause is *Referential* if its status has been annotated as KNOWN or INFERRED, and it is *New* otherwise.
- c. **Focus type.** The *Focus Type* of a cleft is based on the values described in 12.2.1.6, and it is derived manually, by evaluation of the preceding and following context. The values NEUTRAL, TIME and REASON are kept, while all the different constituent focus types (Contrast, Emphatic, Wh) are combined into EMPHATIC.

These three features are further combined into one value, the **information-structure status** (abbreviated as “IS Status”), according to the division shown in Table 45. The first three cleft types (Topic-Comment, Comment-Topic and Comment-Comment) are already known from the combined information status cleft types discussed in section 12.3.3. What is new, is that the information states of the clefted constituent and the cleft clause are only taken into consideration for clefts whose “FocusType” is not related to contrast or emphasis. As soon as an *it*-cleft has the FocusType of “Contrast” or “Emphatic”, it is assigned the IS Status “Emphatic”.

Table 45 Information Structure Status categories

IS Status	Clefted Constituent	Cleft Clause	Focus Type
Topic-Comment	Referential	New	neutral, time, reason, purpose
Comment-Topic	New	Referential	neutral
Comment-Comment	New	New	neutral
Emphatic	-	-	contrast, emphasis

The clefts that are of particular interest for our scenario of the rise of clefts are those labelled as “Emphatic”, as for example (330).

- (330) It was **only after I had been in the room for a few minutes** that I realized that everyone was staring at me.

“Emphatic” clefts are clefts that can be shown to have emphatic prominence or contrastive focus on the clefted constituent (see sections 3.2.2 and 12.2.1.6, as well as chapter 9). Emphatic prominence occurs with adverbs that add emphasis, like *chiefly*, *just*, *same*, *too* etc. The adverb *just* in (331a), for instance, does not make *twenty years* contrastive, but gives it emphatic prominence. Emphatic positive prominence can also be achieved by the combination of a negator with an inherently negative element, as in (331b), where the combination *no worse* can be rephrased as *very good*.

- (331) a. It is **just twenty years** that we had that very very happy meeting at dear Coburg, when you and dear Louise were there! [Victoria-186x:694]
 b. Alick Keith? Not from me, and Lady Temple is perfectly to be trusted; but I believe his father knew it was **for no worse reason** that I was made to exchange. [Fleming-1886:373-374]

Contrastively focused clefted constituents come in different types too, as noted in section 12.2.1.6. Some are marked by the presence of a contrastive focus adverb (such as *only*, *alone* or *but*), as in (332a). Others uniquely identify the referent for the clefted constituent, for instance when it is a pronoun, a demonstrative NP (e.g. *that Mary* in (332b)), a proper name, or an anchored NP (e.g. *my father*). Some negate a focused constituent with unique identification, as in (332c), which forces the reader to contrast it with something else. Occasionally contrastive focus is not formally marked but is clear from the context, as for example in (332d), where *another matter than for money* contrasts with *a money matter* in the preceding discourse.

- (332) a. Still, it is **but a divided attention** that we can give to the exercise. [Bain-1878:350]
 b. A Certayne man was sicke, named Lazarus of Bethania the toune of Mary and her sister Martha. It was **that Mary which annoynted Iesus with oyntment, and wyped his fete with her heere** [‘hair’], whose brother Lazarus was sicke. [Tyndnew-e1-h:985-986]
 c. And it was a bloody sacrifice not a drye sacrifice. Why then it is **not the Masse** that auaieth or profiteth for the quicke and the dead? [Latimer-e1-p:202-205]
 d. Then Throckmorton shuld say, though I know ther hath bin an vnkindnesse betwixt M. Southwell and you **for a Money matter**, wherein I trauelled to make you Friends, I doubt not, but in so honest a matter as this is, he will for the safegard of his Country joyne with you, and so you may be sure of the Lord Burgainey and his Force. Then Wyat said, it is **for another matter than for Money** that we disagree, wherein he hath handled me and others very doubly and vnneighbourly. [Throckm-e1-h:361-363]

Clefts with an emphatic clefted constituent appear, as stated above, only in the category “Emphatic”, so that the “topic-comment” and “comment-topic” categories only contain those clefts that are *not* emphatic. Figure 43 shows the division of cleft types from the full database.¹⁸

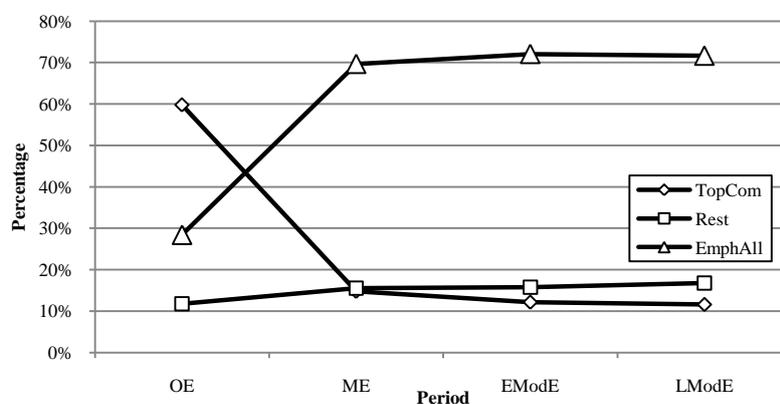


Figure 43 Information Structure Status of *it*-clefts in selected sub periods

The results stress just how different OE clefts were: they were non-emphatic Topic-Comment clefts. OE did not “need” emphatic clefts because V2 offered a position that could easily accommodate focus markers. The drastic rise of the emphatic clefts from OE to ME, as visible from Figure 43, coincides with the loss of the sentence-initial’s position (the PreCore area, which is clearly delimited by the finite verb in OE) to signal constituent focus, as shown in chapter 9.

12.3.5 Emphatic cleft types

Section 12.3.4 looked at the “Information Structure Status” of *it*-clefts, taking all “Emphatic” clefts together as one whole. This enabled us to see that the great difference between OE and ME lies in the rise of the relative amount of “Emphatic” clefts: their percentage grows from 30% to 70%. One question we should ask here is whether this rise in “Emphatic” clefts is due to a particular type of such clefts. This leads us to the question how the internal make-up of the “Emphatic” clefts in general develops throughout time.

The database of *it*-clefts holds all necessary information. The “FocusType” feature of each cleft specifies if a cleft is marked by “Emphatic Prominence”, or by “Contrast”. The FocusType then differentiates between the different kinds of contrast (Same, Pre, Foll, Adv, Neg), as defined in chapter 3.

Figure 44 shows how the division of the different types that make up the “Emphatic” *it*-clefts changes over time, while Table 46 shows their numbers (normalized per 100,000 main clauses) in the indicated periods.¹⁹ The contextually motivated contrastive types (Contrast-Pre, Contrast-Foll, Contrast-Same) have been united, because it does not seem likely that there is an external motivation for their relative make-up, and by grouping them together the essential parts of the picture stand out better.

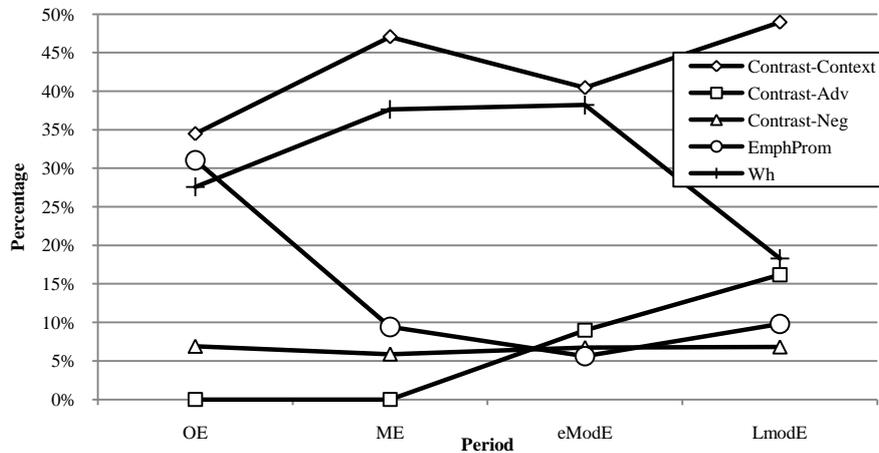


Figure 44 Division of emphatic clefts into types

The largest components of “Emphatic” clefts over time are the “Contrast-Context” ones: those where there is explicit contrast with an element in the same constituent, in the preceding context or in the following context. Their growth in absolute numbers, as shown in Table 46, is telling too. Clefts that are “Emphatic” due to the fact that the clefted constituent is a *wh*-phrase come next. Their relative decline in LmodE is, perhaps, of some significance, but, as Table 46 shows, it seems they have stabilized in terms of their absolute number of occurrence.

OE has a relatively large percentage of *it*-clefts where the clefted constituent has “Emphatic Prominence”. This percentage stabilizes from ME onwards, but, as the numbers in Table 46 show, their absolute numbers keep growing.

(333) a. *ða cwæð ic: Nu ic ongite þæt* [coboeth:2388]

then said I now I understand that
 hit nis **ecu gifu** þæt he gifð þæm yflum,
 it not.is eternal gift that he gave to.the evil.one
 ac is hwilchwugu eldcung & andbid þæs hehstan deman
 but is some delay & expectation of.the highest judge
 ‘Then I said: “Now I understand that it is not an eternal gift that he gave the evil one, but that it is some delay and expectation of the supreme judge.”’

b. *ond him ætywde ða wunda on his handum ond on his fotum*
 and to.him showed the wounds on his hands and on his feet
 ond þa gewundedan sidan, þæt hi þy sodlicor
 and the wounded side that they that truly
 ongeaton þæt hitwæs **sodlice his agen lichomadæt**
 would.understand that it was truly his own body that
 þær of deade aras. [comart3:493]
 there from death had.risen

‘He showed him the wounds on his hands and feet as well as his wounded side, so that he would really understand that it was his very own body that had risen from the death.’

- c. Since therefore ech thing seekith the good, it is playne, that is **only the good** that of all is desyred. [boethel-e2-h:235]

The OE *it*-cleft in (333a) belongs to the “Contrast-Neg” class of emphatic ones: the clefted constituent *ecu gifu* ‘eternal gift’ is negated (it is also overtly contrasted with *hwilchwugu eldcung & andbid* ‘some kind of delay and expectation’ in the following context). The cleft in (333b) is an example of an “Emphatic” cleft from OE: the clefted constituent *his lichoma* ‘his body’ has been made more emphatic by the addition of *agen* ‘own’ and *sodlice* ‘really’.

There is one development, which might seem marginal in Figure 44 at first, but should be regarded as very significant: the rise of the “Contrast-Adv” category (and one of the first ones is illustrated in 333c above). These are the *it*-clefts having a clefted constituent which has marked focus due to the presence of a *focus particle* or *contrastive adverb*. The most telling point from their behaviour is the fact that they are completely absent in OE and ME, only starting to appear in eModE. From then on they grow in both an absolute and a relative sense.

Table 46 Emphatic cleft types per 100,000 main clauses

	OE	ME	eModE	LmodE
Contrast-Context	9	50	82	217
Contrast-Adv	0	0	18	72
Contrast-Neg	2	6	14	30
Wh	7	40	77	81
EmphProm	8	10	11	43

At this point one might wonder what methods OE used to express constituent focus, given the fact that English *it*-clefts have increasingly been used for this kind of focus. The next section deals with that question.

12.4 Clefts and emphasis

The drastic rise of the emphatic clefts from OE to ME, as visible from Figure 43, coincides with the decline in emphatic, focus-marked constituents in the preverbal position from OE to ME, as visible in Figure 45. The two trends are compared in Figure 45, where the “EmphAll” line contains the percentage of *it*-clefts with overt contrast or emphatic prominence, and the “FP-Initial” line contains the percentage of NPs and PPs with focus particle that occur clause-initially in main clauses (see Figure 25 in section 9.2.3 for a combination of focus particles and other focusing adverbs). While the match is not perfect, the trends do seem to complement one another—at least until ME. From then on, preverbal focus-marked constituents increase again, while the percentage of clefts used to express Emphasis remains relatively steady.²⁰

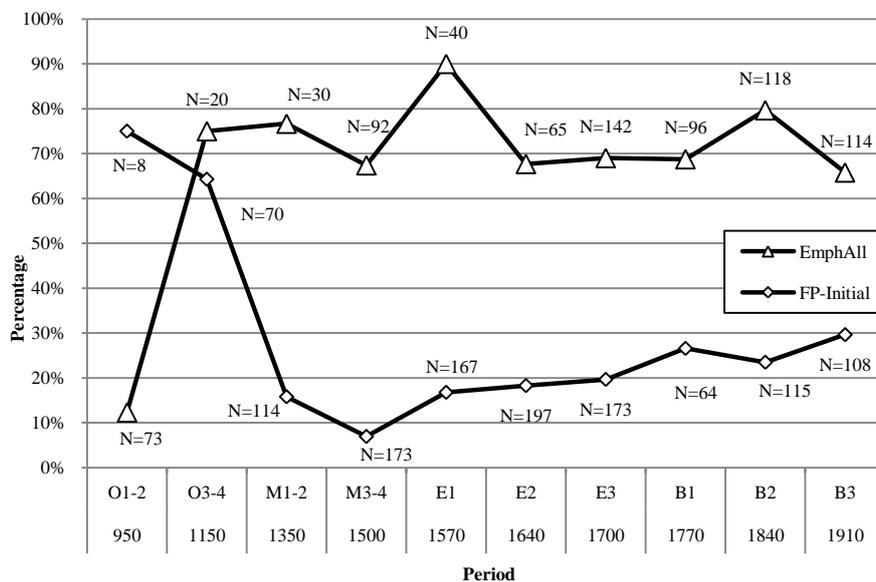


Figure 45 Emphatic *it*-clefts compared with clause-initial focus marking

As the preverbal position in English is increasingly reserved for the syntactic subject, *it*-clefts offer an alternative for the first position's loss of functionality, and a resolution for a conflict of interests: they provide a way to preserve the relative ordering of emphatic constituent and the logical main verb, now in the cleft clause, while at the same time allowing this constituent to follow the syntactic main verb (the copula) in the matrix clause.

The inspection of emphatic cleft types in section 12.3.5 revealed the rise of *it*-clefts with a focus particle in the clefted constituent, which started in eModE, witness the examples in (334a-c).²¹

- (334) a. ... that all the worlde shall to our honor and her reproch, perceiue that it was **onelye malyce, frowardnesse, or folly**, that caused her to keepe him there. [moric-e1-p1:63]
- b. (All her concern now was for his life, and therefore she hasten'd him to the camp, and with much ado prevail'd on him to go.) Nor was it **she alone** that prevailed; Aboan and Onahal both pleaded. [behn-e3-p2:20-4]
- c. It is perhaps **in this method only** we can chastise, and preserve affection, at the same time. [barclay-1743:199]

The focus particles “only” and “alone” occur in the clefted constituents from eModE, as in (334a-b), where the clefted constituent coindexes with a subject gap in the cleft clause. Clefted constituents that have an adjunct role in the cleft clause and that contain a focus particle also occur, as for example (334c).

Clause initial constituents marked with a focus particle (i.e. *only*) in Present-day English give rise to subject-auxiliary inversion, as in (335a): the word order XP – Aux – S results. This word order could be regarded as a remnant of the V2 word order in OE.²²

- (335) a. **Only with the development of factions and the growth of the party system** did it come about that monarchs found themselves confronted, in Cabinet, by Ministers presenting a united front on matters on which they had previously deliberated in the absence of the monarch. [BNC C8R:1532]
- b. It is **only with the development of more radical differentiation in the decades around the turn of the twentieth century** that it is possible to speak of a fully-fledged and optimally differentiated cultural modernity. [BNC GW4:727]
- c. Some woman-centred psychologists think, too, that **only a woman** should study female subjects, and that she should do so as much as possible, because **only she** can understand them. [BNC CMR:1388]
- d. As I could not escape from the coxcombs of the university, I surrendered myself with the best grace I could into their hands. It is **the first step only** that costs a struggle. [godwin-1805:337-338]

Present-day English offers the option between the V2 remnant XP-Aux-S word order and the *it*-cleft, such as (335b). Even though subject-auxiliary inversion is no issue when the subject itself is modified by a focus particle, as in (335c), there are nevertheless *it*-clefts with a focus particle modifying the subject of the cleft clause, as in (335d).

If we now turn to clefted constituents that have an adjunct status in the cleft clause, such as the one in (335b), we see that, even though their numbers are low, the *it*-cleft database shows an increase in their occurrence, as shown in Table 47.

Table 47 Focus-particle *it*-clefts and those of them that are adjuncts

Period	FP <i>it</i> -cleft	Adjunct role	Percentage
eModE (E1)	2	0	0%
eModE (E2)	1	0	0%
eModE (E3)	13	1	8%
LmodE (B1)	7	4	57%
LmodE (B2)	17	12	71%
LmodE (B3)	14	10	71%

Focus particle clefts as such only take off from the third eModE subperiod onwards, and as they increase in number, so does the percentage of them where the clefted constituent has adjunct status in the cleft clause. The increasing tendency to choose a cleft construction instead of the XP-Aux-S remnant V2 word order could be regarded as an indication of continued V2 decline as a whole, which is in line with the results of V2 behaviour that have been shown in the introduction, in section 1.2.2 (see also the discussion in section 0). The results there showed that subject-auxiliary inversion (an indicator of V2) in sentences starting with an adverbial

phrase, a noun phrase object or a prepositional phrase steadily decreases from 58% in OE to 5% in LmodE.

We have two developments involving adjuncts: the subject-auxiliary inversion changes (which involve adverbial phrases, as per Table 47) and the growing role of adjuncts in *it*-clefts (see Figure 45). The question rises whether there might be a connection between these two developments. In order to address this question I have constructed a corpus research project that looks for subject-finite-verb inversions involving a clause-initial adjunct that has an emphatic or contrastive adverb (or focus particle). The project looks for XP-V_{fin}-S instances (where the XP contains a focus adverb) in main clauses and compares these with the overall number of main clauses having a subject, a finite verb and an XP with focus adverb in any order. Figure 46 shows the results of this search, which indicates that subject-auxiliary inversion for constituents *with a focus particle* is decreasing in English (though the sample numbers are small).²³ This shows the percentage of clauses with word order PP-Aux-S instead of PP-S-Aux, where the PP contains a focus particle.

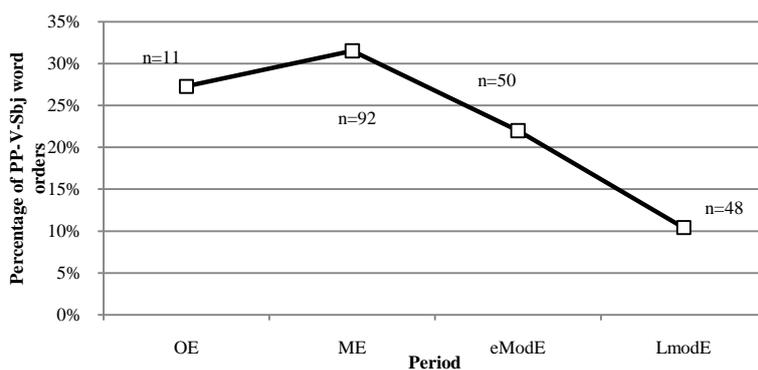


Figure 46 Subject-auxiliary inversion for clause-initial PPs with a focus adverb or particle

Languages apparently need to have well-defined ways to express contrast—be it through word order, morphology, particles or constructions. Present-day German does not seem to need *it*-clefts to express constituent focus, because it has the better, less marked option of word order (Ahlemeyer and Kholhof, 1999, Miller, 2006).

The changes in English syntax seems to have done away for the privilege of the first position (the PreCore slot) to serve as a recognizable way for constituent focus, which is one of the reasons why the already existing *it*-cleft was increasingly used for the expression of constituent focus. The contrastive reading of clefts is a natural reading, given the characteristics of the construction. The cleft's main clause is an identificational (copula) construction, and when the clefted constituent is unique, this naturally leads to specificational reading: one where the clefted constituent is seen as specifying the value of some variable. It is a short step from such a unique specification to a contrast with alternatives. This is *not* to say that a contrastive reading must always necessarily follow from a cleft—we have already seen

evidence that the historic and synchronic data do not bear this out. It is just that the characteristics of the *it*-cleft make it a nice and attractive environment for contrast.

12.5 Conclusions

The *it*-cleft has been part of the English language right from the start in Old English, but it started to grow significantly only in early Modern English, after 1640. The *it*-clefts in OE were by and large adjunct ones, which had the function of discourse partitioning (but emphatic clefts are attested already in this period). OE was a V2 language, and the thematizing function of its *it*-clefts is in line with the findings of Hasselgård (2004) and Johansson (2002) for V2 languages such as Norwegian and Swedish. The synchronic data from Chechen presented in chapter 11 showed that this is a language where *it*-clefts *only* have a thematizing function, and are not used as a focusing device. While more typological data would be needed, these findings suggest that the thematizing function of *it*-clefts may be the more fundamental one in general.

Instead of Informative-Presupposition clefts rising as an innovation after the appearance of Stressed-Focus clefts, our data show that the first category of clefts, those with informationally new cleft clauses, gradually *decreases* over time, starting from OE.

A major question in this chapter has been whether the rise of *it*-clefts in English could be related to information-structure. This has indeed been found to be the case, but the relation is not the most obvious one. It is not the information state of the clefted constituent, nor that of the cleft clause, nor, for that matter, a combination of them, that most clearly describes the rise of the cleft.

A big change took place from OE to ME, which involved a shift from using *it*-clefts mainly as a text-dividing strategy, where the clefted constituent usually was a time adjunct, to using them to express constituent focus, mainly on subjects, but also on objects. The reason for this shift, as substantiated in chapter 9, lies in the decline of the clause-initial position for constituent focus. The emphatic-clefts have become a strategy whereby constituent focus is still expressed before the logical main verb of a proposition (which is inside the relative clause), while it syntactically follows the main verb of the main clause (the copula). It is, therefore, through *constituent focus* that the cleft demonstrates the interaction between syntax and information structure from OE to ME.

From eModE onwards *it*-clefts become a strategy to avoid subject auxiliary inversion (which is a remnant of the OE verb-second system): when a constituent is marked for constituent focus by having a focus adverb or a negation, the language has the option to put this constituent clause-initially and let the finite verb follow it immediately, after which the subject comes (this is the OE verb-second option), or to put the constituent in an *it*-cleft, where the new PDE core structure of SVO is satisfied in the initial part of the cleft, and the clefted constituent nevertheless precedes the cleft clause where it logically is part of.

Objects are relatively less frequently clefted, since expressing constituent focus on objects can be done through word order, such as the OSV “preposing” word

order (Birner and Ward, 1998). The percentage of adjunct clefts rises again from eModE onwards, but the type of adjuncts differs from that used in OE. While older English mainly had temporal adjuncts, modern English has a wider variety, including reason and manner ones. Text organization is, I argue, still the main function of English *it*-clefts, but then largely in combination with contrast.

On the whole we can say that syntactic changes in English introduced a caveat in the language where the *it*-cleft's originally secondary function of accommodating constituent focus took over from its text organization function.

With a relatively clear picture of the changes in mind, the final chapter looks back and summarizes our findings about the relationship between syntax and focus, while it also looks ahead to possibilities to extend this study in the future.

¹ On "Informative-Presupposition" clefts, see section 10.2.2.

² The PPCME2 and PPCME are the parsed corpora of ME and eModE texts respectively.

³ This example is taken from [Benson-1908:236].

⁴ Earlier forms of English could have both a relativizer pronoun and a complement, so the coding here is not redundant.

⁵ Proper code would need to be enclosed in curly brackets and it would need to have `<TEI>` added in the beginning, and `</TEI>` at the end.

⁶ Some of these non-cleft constructions have the cleft clause as restrictive relative clause of the clefted constituent, whereas others have it as an appositive one.

⁷ The corpulect distribution D[corp] is 34% for all the periods in OE until LmodE together. The D[corp] values per period are: 24% (OE), 59% (ME), 23% (eModE), 77% (LmodE).

⁸ The general word order of the clefts has been annotated, but has not been worked out

⁹ For D[corp] values: see footnote 7. The significance according to Fisher's two-tailed exact test ($p < 0,05$) of the period transitions per line are as follows (see for details the appendix, section 14.3.15):

Adjunct:	OE-ME and eModE-LmodE are significant, but ME-eModE is not
NonArgNP:	OE-ME and eModE-LmodE are significant, but ME-eModE is not
Object	OE-ME and eModE-LmodE are significant, but ME-eModE is not
PPobj:	none of the transitions is significant
Subject:	OE-ME and eModE-LmodE are significant, but ME-eModE is not

¹⁰ Only "reason" clefts have been identified as a separate category in the database (by the feature called "FocusType" – see section 12.2.1.6). Most of the other adjunct clefts appear in the Contrastive and Emphatic categories, since they incorporate negators or contrastive adverbs.

¹¹ This example shows OE relative clauses as still using both the relativizer pronoun (the *se* paradigm) as well as the complementizer (which is the *ðe*).

¹² For D[corp] values: see footnote 7. All the transitions between periods are significant according to Fisher's two-tailed exact test ($p < 0,01$). See for details the appendix, section 14.3.16.

¹³ The Present-day English translation of this example is taken from a Wikisource version, which is based on several earlier translations (Jane and Sellar, 2011).

¹⁴ For D[corp] values: see footnote 7. The significance according to Fisher's two-tailed exact test ($p < 0,05$) of the period transitions per line are as follows (see for details the appendix, section 14.3.17):

Assumed: none of the transitions is significant
 Identity: only the transition from OE to ME is significant
 Inferred: the transition from ME to eModE is insignificant; the others are significant
 New: the transition from ME to eModE is insignificant; the others are significant

¹⁵ For D[corp] values: see footnote 7. The significance according to Fisher's two-tailed exact test ($p < 0,05$) of the period transitions per line are as follows (see for details the appendix, section 14.3.18):

Assumed: none of the transitions is significant
 Inferred: the transition from OE to ME is not significant; the other transitions are
 Known: all the transitions are significant
 New: the transition from eModE to LmodE is not significant; the other transitions are

¹⁶ There are *it*-clefts with a *wh* element in the clefted constituent whose cleft clause contains "New" information. One example might be (i). The author wants to make a new point in the discussion by introducing God as the bestower of blessings. But even here it could be argued that, through the position in the cleft clause, the author "assumes" this information to be available to his audience.

- (i) Who is it that diffuses blessings upon mankind and saves them from evil, but God alone, who is the guide and physician of souls? [boethri-1785:415]

¹⁷ For D[corp] values: see footnote 7. The significance according to Fisher's two-tailed exact test ($p < 0,05$) of the period transitions per line are as follows (see for details the appendix, section 14.3.19):

TopCom: the transitions from OE to ME to eModE are significant, the one to LmodE is not
 ComTop: only the transition from ME to eModE is significant
 ComCom: none of the transitions are significant
 TopTop: the transition from eModE to LmodE is not significant; the other transitions are

¹⁸ The cleft types of Comment-Topic, Comment-Comment and Topic-Topic have been gathered under the general umbrella of "Rest", so as to stress the behaviour of the Topic-Comment clefts as opposed to the Emphatic ones. For D[corp] values: see footnote 7. The significance according to Fisher's two-tailed exact test ($p < 0,05$) of the period transitions per line are as follows (see for details the appendix, section 14.3.20):

TopCom: only the transition from OE to ME is significant
 EmphAll: only the transition from OE to ME is significant
 Rest: none of the transitions are significant

¹⁹ For D[corp] values: see footnote 7. The significance according to Fisher's two-tailed exact test ($p < 0,05$) of the period transitions per line are as follows (see for details the appendix, section 14.3.21):

Contrast-Context: none of the transitions are significant
 Contrast-Adv: the transitions from ME to eModE to LmodE are significant;
 OE to ME is not
 Contrast-Neg: none of the transitions are significant
 EmphProm: only the transition from OE to ME is significant
 Wh: only the transition from eModE to LmodE is significant

²⁰ D[corp] is 49% for the FP-initial line and 75% for the "EmphAll" *it*-cleft line.

²¹ The first instance of such a cleft is from around 1500 [moreric-e1-p1:63].

²² I am assuming the kind of V2 word order for OE as hypothesized by Kemenade and others, who give accounts in terms of derivational grammar (van Kemenade, 1987, van Kemenade and Westgaard, 2012). Accounts in terms of Optimality Theory reach V2 word order by a series of hierarchical constraints (Grimshaw, 1997).

²³ D[corp] is 16%. None of the transitions are significant according to Fisher's two-tailed exact test ($p < 0,05$).

Part V

Implications

The theoretical questions that have motivated this study are how syntax interacts with focus, and what the interaction between syntax and focus tells us about grammar, the system of rules a language uses to convey what we want to say. The study described in this book contains excursions to areas such as psycholinguistics, text charting, anaphor resolution, software development and corpus research, but all these should be regarded as tools that one way or another help shed light on the main research questions, which I repeat from (11) in chapter 1: *What can we learn about the interaction between syntax and focus, when we look at the development of the English language as visible in the available syntactically parsed corpora?*

The reason to look at English is the fact that one part of the research question, the development of the syntax in English, has already been studied extensively (see chapter 4). My attempts to describe the changes in English *focus* also builds on work that has already been conducted on the expression of focus (narrow focus through *it*-clefts in particular) in Present-day English.

This final chapter returns to the main question by recapitulating the results that have been gained and then considering what the implications of these findings are for grammar in a wider perspective.

13.1 Background

The theoretical underpinnings for this study were built up in chapters 1-4, and it is against the background of the setting provided by these chapters that the results reported in this chapter should be viewed. Chapter 1 introduces the concepts that are used in this study, one of which is the view on syntax that builds on Dryer (2003): rather than seeing all word order as being determined by syntax (which would disable answering the research question about the *relation* between syntax and focus), syntax is first and foremost regarded as having the function to signal grammatical functions and relations. Word order in English is needed for this purpose to a changing degree. Languages also come with “default” word orders, which greatly alleviate the processing burden.

Chapter 2 grounds the research described by this book in a psycholinguistically oriented view of communication, whose fundamental concept is a dynamically being built up “situation model” into which “mental entities” are added and/or connected with one another. The reason to take such a situation model as a starting point is the increasing confirmation from psycholinguistics and neurolinguistics that we as humans make use of something like a situation model in our communication.

The notion of focus as explained in chapter 3 opts for one particular point of view from the many existing frameworks; I have refrained from introducing new theories or concepts in this area. It crucially involves the concept of “focus domains”, which, in turn, correlate with three main focus articulations: *thetic focus*

(the domain is the core of the sentence), topic-comment articulation (with the predicate as focus domain), and constituent focus (where one argument or adjunct constituent constitutes the focus domain). Clauses have one of these three focus domains, and can also have a “Point of departure”. The “Principle of Natural Information Flow” is recognized as one universal principle influencing word order within sentences. Deviations from canonical (or expected) word order can then lead to “Dominant Focal Elements” within a larger focus domain.

The start of chapter 4 introduces the working hypothesis that three factors playing a role in the word order of clauses (syntax, focus and text-structure) are to be regarded as independent. The word orders observed in the different stages of the English language are described in chapter 4 by making use of a text-charting approach, which determines a kind of “best-fit” slot-structure for the majority of the actual word orders found in main clauses in a text. The slot-structures that are proposed compare to the topological field model (which is used to describe German), in that they divide sentences into a PreCore, Core and PostCore area. An algorithm to come up with the best slot-structure for a given text reveals an interesting change in this slot-structure over time: (a) while the number of slots used for the PreCore stays equal over time, it is the size of the Core that gradually decreases from 3 constituents in OE to 1 constituent in LmodE (see section 4.5.3); (b) OE initially has two dedicated slots for the subject (one in the PreCore area, and one in the Core) and one alternative (the late-subject in the PostCore area), but the core-internal one disappears as a dedicated slot towards the end of the eModE period, and the late-subject all but disappears in the LmodE period (see the discussion on subject-positions in the preamble to chapter 4, and the introduction of the OE and LmodE slot-structures in section 4.1.2).

These findings correlate with the changes in English syntax that have been reported. The change that has been found most crucial for this study is the loss of V2. This loss first of all caused the reduction of three subject positions (PreCore, Core and PostCore) to one (PreCore). Since the PreCore area in the V2 system had multiple functions, which included hosting constituent focus, the loss of V2 also meant the loss of this constituent focus position (and in 13.3.2 I explain what alternative strategies for constituent focus were found). The severe reduction of the late-subject construction (the subject position in the PostCore) due to the loss of V2 meant that presentational focus was jeopardized (and in 13.3.1 I summarize how the realization of this focus articulation changed). The loss of V2, then, can be seen as the main trigger of the changes in the expression of different focus articulations in English.

13.2 Methodology

The initial studies into the pragmatic factors that influence word order differences in Old English and late Modern English narrative texts, described in chapter 4, reveal that the degree to which a constituent represents established or unestablished information partly determines whether it belongs to the focus domain of the clause or not—something which is not surprising in light of the studies by Birner and Ward

(1998), who successfully explain word order variation in Present-day English by making use of absolute and relative newness. The size of the focus domain (whether it spans one constituent, the whole predicate or even includes the subject) in turn translates into one of three focus articulations: constituent focus, topic-comment or presentational focus.

With this in mind, I have seen it as a task of primary importance to add the degree to which constituents represent established information in a maximally objective way. I have adopted a coreference resolution approach from computational linguistics, and extended it to add referential categories from a minimal set of primitives to each and every noun phrase. The set of categories are derived in chapter 5 and the extended coreferential resolution algorithm, called “Cesax”, in chapter 6. A steadily growing set of texts is now being enriched with the help of Cesax.

In order to do effective corpus research in the texts that have been enriched with the referential information, I have written a computer program “CorpusStudio”, which basically is a shell around existing search engines, and I have provided extensions aimed at working with, for instance, coreferential chains. Chapter 7 describes how this program can be used to perform corpus searches that combine syntactic information with referential information.

13.3 Focus changes

The methodology that I thus arrive at is one where existing syntactically annotated corpora are enriched with referential information and these enriched corpora are queried for particular combinations of syntax and referentiality. Chapters 8-9 use this approach to discover the strategies used in the development of the English language to express two focus articulations: presentational and constituent focus.

13.3.1 Presentational focus

One of the correlations between focus domain and syntax is clearly visible in presentational focus, as defined in section 3.2.3, and experimentally investigated in chapter 8. The focus domain in presentational focus is that of the whole core of the clause (the subject plus the verb phrase), and I have argued that unlinked new subjects are an indication of this focus type.

The experiments in chapter 8 show that there are, historically speaking, two grammatical constructions possible where presentational focus occurs: the referentially new syntactic subject can occur in the PreCore area or in the PostCore (the area between the Vb1 and Vb2 slots is excluded). If we exclude the use of expletives, then the word order used for presentational focus involving participants that are major (in the sense that they are at the start of a medium to larger sized coreferential chain) changes enormously in the course of the development of English: the proportion of presentational focus with a PostCore subject changes from 36% in Old English to a mere 4% in late Modern English, whereas the majority of presentational focus (95%) by that time occurs with the new subject in the PreCore position. What we see here may just be the result of the general tendency of

English to have subjects occur almost exclusively in a preverbal position. We should realize that there are two conflicting constraints at work here: (a) due to the loss of V2, syntax increasingly demands the subject to occur preverbally (the subject replaces the finite verb as the marker of the start of the “core” slots), and (b) the principle of natural information flow demands referentially new constituents to occur as far to the right of the clause as possible. Section 8.4 in general, and Figure 20 in particular illustrate this conflict of interests, showing that, over time, syntax is at the winning hand.

It is unclear which of the syntactic theories in general (minimalism, government and binding, role and reference grammar, optimality theory) are able to capture such a relatively low-level change over time. At each point in time over the last 1000 years there is a certain proportion of clauses going one way (allowing new subjects to occur postverbally) and another proportion going the other way (having new subjects appear preverbally). So it is *not* the case that at every point in time there was a clear “winner” in this conflict.

Returning once more to the specifics of the change in presentational focus, we can conclude that the English language did develop a strategy to accommodate for the two opposing demands of (a) having unestablished information appear late in the clause, and (b) demanding the syntactic subject to occur before the finite verb (due to the loss of V2, which meant a loss of alternative subject positions). The strategy that evolves is that of using an expletive, as illustrated in the Present-day English rendering of (81a), which is repeated here for convenience: an expletive provides for a core-start signalling syntactic subject (the expletive pronoun) in the position before the finite verb, while still allowing the unestablished information in the “logical” subject to occur postverbally, but now syntactically encoded as a complement.

- (81) a. **Svm wer wæs on Alexandria mægde Pafnvtivs genemned,**
 one man was in Alexandria province Pafnuntius called
 se wæs eallum mannum leof and wurd, [coeuphr:3]
 that was to-all to-men loved and valued
‘There was a certain man in the province of Alexandria named Paphnutius, who was beloved and honoured of all men.’

This results in a construction that syntactically looks suspiciously like the canonical topic-comment articulation: a pronominal subject followed by a predicate. The referential states of the individual components, however, betray that the focus domain properly only includes the syntactic complement (which holds the logical subject), since that contains the new information, and the expletive subject as well as the auxiliary *are*, in a sense, empty placeholders (which are apparently needed in order to show the core-structure). All this is not to say that expletive constructions *only* serve (or arose) to convey presentational focus; there are other functions performed by them as well, since expletive constructions can also appear with subjects that are *not* referentially new (Hartmann, 2008, Ingham, 2001). But, as far as focus strategies are concerned, the expletive construction clearly took over from the late-subject construction somewhere between the eModE and the LmodE periods, witness Figure 22. This figure confirms the hypothesis in (183), which

states that presentational focus in English remains using the area in the clause after the finite verb. This is the PostCore area for OE, but what the exact location of the area is for LmodE, when the expletive strategy is used, is not completely clear yet, since it becomes increasingly difficult to distinguish the Core area from the PostCore area with this focus articulation.

13.3.2 Constituent focus

Constituent focus is the articulation that restricts the focus domain to exactly one constituent, and this constituent can, for instance, be a noun phrase, a prepositional phrase or another kind of adverbial phrase.¹ Chapter 9 sets out by arguing that two principles of constituent focus, the “demarcation principle” in (198) and the “placement principle” in (199), remain important for the strategy used to express constituent focus in English. This strategy changes, according to the hypothesis in (197), which states that the locus of constituent focus moves from the PreCore to the PostCore area. The reason for this can be traced to the loss of the V2 system: the increasing placement of the subject before the finite verb leads to the loss of the PreCore area as an area that can be used for constituent focus.

With an open mind for alternative strategies, chapter 9 ascertained the use of various diagnostics that are *not* related to a particular position in the clause to recognize constituents that are part of this kind of focus articulation. Most effective were two diagnostics: the presence of a contrastive adverb in an NP or PP (9.2) and overt local contrast within an NP (9.5). The diagnostic of negation (9.3) would be usable too, but only if we would have a larger corpus of referentially enriched texts, since these diagnostics do need to make use of referential information. The remaining syntactic features did not appear to have a straightforward correlation with constituent focus at all: positive negation (9.4), emphatic pronouns (9.6), apposition (9.7), split constituents (9.8), contrastive left dislocation (9.9) and the different kinds of *wh* clefts (9.11). The diagnostic of constituent answers to *wh* constituent questions (9.10) proved to be unreliable for automated corpus research, since there is no way to tell whether a question is a rhetorical one, for instance, and sometimes people just do not answer a question, or if they do, they do not provide a constituent answer.

When the independent diagnostics are used to measure whether there is any preference at a particular point in time for a particular position of constituents participating in constituent focus, it becomes clear that there are two trends over the last 1000 years. Old English starts with a clear preference for constituent focus clause-initially (in the PreCore slot), but this changes into a preference for constituent focus postverbally (in the PostCore slot) by the end of the Middle English period (approximately 1500 A.D.). From early Modern English onwards, the locus of most constituent focus instances remains postverbally, but there is more room for preverbal constituent focus, as illustrated by the examples in (336).

- (336) a. **Twice only** I remember having heard it. [reade-1863:174]
 b. He made soldiers **only of the best of his men**. [long-1866:125]

The example in (336a) has a temporal NP modified by the focus adverb *only* in clause-initial position, while (336b) has a prepositional phrase modified by the same adverb in the clause-final position. It seems to be clear that throughout time, both the clause-initial position as well as the clause-final position have been able to host constituent focus, so that syntactic descriptions of English during this time period should likewise be able to facilitate this kind of focus in *both* positions (see also 13.4.2). The clause-final position has not been a problem for any syntactic description as far as I am aware of, because it is the natural host for focus: it is a natural position for prosodic marking of the focus domain to occur, it is the natural position of unestablished information to occur (satisfying the principle of natural information flow) and it is the clausal position of the object in Present-day English, which in most cases is the syntactic vehicle to contain new information anyway. The only problem with the clause-final position as in (336b) is that it may not always be clear whether we are dealing with a constituent focus articulation (in which case the remainder of the clause figures as backgrounded and established information) or with a dominant focal element within a topic-comment articulation (in which case the remainder of the VP represents unestablished information).

The most reliable diagnostics show that constituent focus can occur clause-initially and clause-finally, and that there is a rapid change in preference from the former in Old English to the latter by the end of Middle English (around 1500 A.D), after which a gradual reversal sets in. The clause-final position is still the preferred one for constituent focus by the end of late Modern English (the beginning of the 1900s), but there is still a fair amount of clause-initial constituent focus too, and this includes subjects, objects and non-argument NPs or PPs.

These changes in the position of constituent focus match up fairly nicely with the development of an increasing part of the *it*-clefts that is used to convey constituent focus, described in chapter 12. Even though the construction has been, and still is, being used for other purposes (text organization), it is fair to conclude that the *it*-clefts have taken over at least part of the constituent focus function fulfilled by the clause-initial position in Old English. The reason for this is connected with the fact that the *it*-cleft simultaneously satisfies the “demarcation principle” in (198) and the “placement principle” in (199): the former is met by the clefted constituent being in the clearly demarcated complement area of a copula clause, and the latter is met by the clefted constituent preceding the remainder of the clause.

13.4 Implications for grammar

The conclusions for changes in English focus are one thing, but the question remains what the implications of the research described in this study are for grammar in general. If we consider “grammar” to be the collection of rules and regularities that jointly determine the word order of a sentence, then there are a few issues I would like to discuss in the light of this study:

(337) *Issues for grammar*

- a. Word order is determined by a combination of syntactic and referential information
- b. Multi-phrasal prefields
- c. Syntax may depend on referentiality
- d. Mappings between syntax and focus
- e. Avoidance strategies

We will have a look of these implications one by one, although some of them are so pervasive, that they will appear in more than one part of the discussion.

13.4.1 Syntax and referential information conspire for word order

The hypothesis that I adopted in chapter 3, which is based on work from Lambrecht (1994) and Dooley & Levinsohn (2001), says that there are three different domains available for focus, which results in three different focus articulations:thetic articulation, topic-comment articulation and constituent focus articulation. These articulations combine with notions such as (a) clause-initial points of departures (Beneš, 1962, Levinsohn, 2000), (b) the Principle of Natural Information Flow (Comrie, 1989, Firbas, 1964), and (c) the presence of Dominant Focal Elements (Dooley and Levinsohn, 2001).

In the Old English narrative of Euphrosyne, we see the Principle of Natural Information Flow interacting with syntactic demands. This happens in the split constituent, and I repeat the relevant examples (retaining their original numbers) here:

- (81) a. **Svm wer** wæs on Alexandria mægðe **Pafnvtivs** **genemned**,
 one man was in Alexandria province Paphnutius called
 se wæs eallum mannum leof and wurd, [coeuφr:3]
 that was to-all to-men loved and valued
'There was a certain man in the province of Alexandria named Paphnutius, who was beloved and honoured of all men.'
- b. Ða æt nyxtan com him **anþegen** to, [coeuφr:33]
 then at last came him a noble to
se wæs weligra and wurþra þonne ealle þa oþre,
 that was wealthier and worthier than all the others
 and hire to him gyrnde.
 that her to him desired
'Then at last came to him a noble who was wealthier and worthier than all the others, and desired her for himself.'

The problem in (81a) is that a completely new and unlinked major participant is introduced in a way that is reminiscent of the topic-comment articulation (topical and linked subject, followed by new information in the predicate), but contrary to that articulation the subject is completely new. The desire to use a word order that correlates with the canonical topic-comment articulation, while there is no established topic yet, is handled in Old English by splitting the subject into two parts: the first part appears in the clause-initial (PreCore) position, where it is interpreted as the topic, and the last part of the subject appears clause-finally (in the

PostCore slot), where it satisfies the constraint to have unestablished information occur as late as possible (since we may assume that the province of Alexandria was part of the readers' world knowledge, but Paphnutius was not). The present-day English rendition of this sentence has to start with the expletive *there*, which is a placeholder for the subject; the logical subject follows the finite verb (the auxiliary *was*). This implies (and chapter 8 confirms this) that a ban on completely new subjects appearing before the finite verb has appeared in the course of history. Such a ban combines syntactic information (the fact that *a certain man* is the subject of the clause) with referential information (the fact that this man is completely unlinked to existing information in the mental model that the addressee has of the narrative's situation), in order to arrive at a particular word order.

The second example of a split constituent, the sentence in (81b), shows that compliance with the Principle of Natural Information Flow can lead to splitting a prepositional phrase into two parts: *to him* transforms into *him* + subject + *to*. Here too we see that the sentence's surface form results from combining syntactic information (the fact that *to him* is a constituent that appears after the finite verb *com* 'come', unless there are constraints overruling this) with referential information (the fact that *him* refers to an established participant, and that *an þegen* 'a noble man' is completely new).

In sum, a grammar (in the sense suggested at the preamble to 13.4) needs to be able to combine syntactic and referential information in order to arrive at the correct word orders. It needs to allow constituents to be split—even if they are as tightly knit as prepositional phrases. It not only needs to facilitate a canonical (default) word order, but also allow for deviations, based on the referential status of individual constituents.

13.4.2 Multi-phrasal prefields

There may be an issue with PreCore areas containing more than one constituent, where the first constituent has constituent focus, but no *do*-support is triggered. Syntactic descriptions of English in its different stages need to be able to host constituent focus for the first constituent (the clause-initial one; the PreCore slot), such as the *twice only* in (336a), when this first constituent is followed by a subject. A generative approach (that is: minimalism, government and binding, principles and parameters, or a derivative of any of these) to this sentence could have the focused constituent in the specifier of the CP, which is a category neutral position. A crucial complication in the example above is that the focused constituent *does* precede the pronominal subject *I*, but no *do*-support is triggered, which, in terms of generative grammar, means that a CP (or a NegP if *only* is seen as head of such a constituent) is formed with a filled specifier but an empty (or at least invisible) head, since the finite verb *remember* does not move there, nor is an auxiliary generated to occur there.

An optimality theory account in terms of Grimshaw (1997) has similar problems, since the constraint "OBHEAD" (the top most constituent must have an overt head) is

clearly violated, but it is unclear what the higher ranked constraint can be that allows for this violation.

A descriptive account similar to that used in chapter 4 for the description of the Old English and the late Modern English text could argue that *twice only* is not a focused constituent, but is a point of departure situated in the PreCore slot—one that happens to be emphatic (due to the presence of the focus particle). Such an analysis seems quite appropriate, because it leaves the topic-comment structure intact: “I” is the topic, and “remember having heard it” is the comment, the new information added to the mental model of the addressee. The analysis fits the context as well, because it correctly sets out a new (small) paragraph, as exemplified by the larger context in (338), which shows the where (336a) occurs.

- (338) a. These natives have their Naiads and Dryads; their spirits which inhabit lakes, and mountains, and forests, and high places. They have also their Typhon and their Osiris, their Evil Genius and their Good Spirit. The former Mbwiri they worship piously, being always anxious to deprecate his anger. They regard him as the Prince of this world; as a tyrant whom they hate, but before whom they must prostrate themselves.
- b. The Good Spirit, on the other hand, they do not deem it necessary to pray to in a regular way, because he will not harm them. The word by which they express this Supreme Being answers exactly to our word of God. Like the Jehovah of the Hebrews, like that word in masonry which is only known to masters, and never pronounced but in a whisper and in full lodge, this word they seldom dare to speak, and they display uneasiness if it is uttered before them.
- c. **Twice only** I remember having heard it:
- d. once, as I have related, when we were in a dangerous storm, the men threw their clenched hands upward and cried it twice;
- e. and again, when I was at Ngumbi, taking down words from an Ashira slave, I asked him what was the word for God in the language of his country. He raised his eyes, and pointing to heaven, said, in a soft voice, “Njambi.” [reade-1863:166-179]

The preverbal focused constituent of (336a) is shown here in (338c), but consider the context before this line: the paragraph in (338a) speaks about evil spirits, the paragraph in (338b) about them acknowledging one good spiritual being (notice how the PP *on the other hand* functions as an indicator of the referential point of departure), and then the section from (338c-e) focuses on the *word* used for this spiritual being. This episode divides into three smaller paragraphs, which are each signalled by a referential point of departure: “once” in (338d) and “again” in (338e).

Whatever theory of grammar is taken, it must be able to account for the kind of multi-phrasal PreCore area as illustrated by (338c): either with the first constituent understood as constituent focus (as indicated by the presence of the focus particle “only”), or as point of departure.

13.4.3 Syntax may depend on referentiality

The definition of the *it*-cleft in chapter 10 leads to a theoretical implication that I would like to draw attention to: the fact that the syntactic interpretation of a sentence can be dependent on the referential categories of its components. We have seen that when an *it*-cleft-like construction has a pronominal syntactic subject (which in present-day English usually is the pronoun *it*), then the syntax of the sentence actually works out is determined by the referential category of *it*. I repeat the relevant examples here from section 10.1.5.

- (239) a. There was someone at the door yesterday. It was **my neighbour** who had a package for me.
 b. Was that the mailman? It was **my neighbour** who had a package for me.

While the second clause in examples (239a,b) is identical, their syntax differs, depending on the referential category of the pronoun *it*: if *it* has the category “Identity”, as it does in (239a), where it links back to *someone*, then the clause is a copula construction with *my neighbour who had a package for me* as complex NP complement, but if *it* has the category “Inert”, as in (239b), then the second clause has the syntactic structure of an *it*-cleft.²

What this boils down to, then, is the fact that the referential category of one constituent (whether the subject pronoun *it* has referential category “Identity” or “Inert”) determines how the syntactic structure of a sentence will look like, irrespective of any other surface factors.

13.4.4 Mappings between syntax and focus

In this section and the following section, I would like to present more evidence for the observation that there does not need to be a one-to-one mapping between a focus articulation and the way in which it is realized (see for example Zimmermann and Onea, 2011).

I would first like to address the issue of mapping from focus to syntax. We have seen in section 13.3.1 that there are different syntactic strategies for presentational focus: a position before and a position after the finite verb. Section 13.3.2 has shown that there are different strategies for constituent focus as well: use the PreCore slot, or be part of an *it*-cleft. The implication is that grammar must be able to contain one-to-many mappings from focus to syntax.

As for the mapping from syntax to focus, the first constituent in Old English, which can be regarded as a V2 language, is, as we have seen, a good example. There is a one-to-many mapping from syntax to focus, because the first constituent may be the locus of constituent focus, it may host a point of departure in a topic-comment articulation, and it may contain a discourse link.

The *it*-cleft too is an example of a one-to-many mapping from syntax to focus. Three chapters of this dissertation (chapters 10, 11 and 12) are devoted to this construction that, at first glance, would seem to be a good diagnostic for constituent focus. The development of the *it*-cleft in English from eModE onwards does indeed indicate that the *it*-cleft primarily functions as a construction to clearly demarcate

constituent focus (either contrastive focus or emphatic prominence). A problem that has always been recognized by researchers, however, is the fact that there is no automatic guaranteed mapping from *it*-cleft to constituent focushood. In fact, the *it*-cleft appears to be capable of fulfilling several different functions, such as that of “topic-shift” in (264b), which is repeated from chapter 10, section 10.2.5.

- (264) b. (C: But really what’s happened with my sort of history is when I met uh did a little recording with Chandos Records uhm and the Ulster orchestra who was conducting there came up with enough money to do their first record and they got Chandos interested.)
 It was **then** that uh I fell in love with music like Hamilton Harty and a bit of Stanford.
 (And the Arn – the Arnold Bax Saga became something quite uh excellent.
 A: Well that’s a day we certainly want to come back to a bit later. But if we could just for a moment concentrate on the latter years of the nineteenth century.) [ICE-GB S2B-023 #61:3:A]

The clefted constituent *then* does not really appear to be set out as focused, but the clause as a whole does fulfil a clear function in the discourse: it is speaker “C”’s attempt to shift the topic of the interview to something different (the time speaker “C” fell in love with a particular kind of music). This attempt is recognized by the interviewer, speaker “A”, who explicitly indicates he wants to return to the previous topic (“the latter years of the nineteenth century”).

The non-automatic link between a construction like *it*-cleft and a function like “expressing constituent focus” becomes clear beyond doubt when we look into the Caucasian language Chechen in chapter 11. All the *it*-clefts found in this language are time-clefts, and they primarily have this discourse function, which can be either to start a story, to provide a transition between a story’s episodes, or to signal a story’s end (the “summative” function).

In sum, grammar must be able to allow for one-to-many mappings between focus articulations and syntactic constructions.

13.4.5 Grammar may have avoidance strategies

Section 10.2.3 stated that *it*-clefts can also function as an “avoidance” strategy: a strategy to arrive at a construction that may not satisfy all conditions perfectly, but avoids violating the worst constraints. This “worst case” might be the combination of focus and grammatical subject. Scandinavian languages, for instance, have a strong tendency to use clefts as a strategy to keep referentially “new” information out of the main clause’s syntactic subject position (Gundel, 2002, Hasselgård, 2004, Johansson, 2001), as illustrated with the example repeated from (260):

- (260) a. (Etter hvert som Sofie tenkte over at hun var til, kom hun også til å tenke på at hun ikke skulle være her bestandig. Jeg er i verden nå, tenkte hun. Men en dag er jeg borte vekk. Var det noe liv etter døden? Også dette spørsmålet var nok katten helt uvitende om.) (Gundel, 2002: ex. 19)
 Det var **ikke så lenge** siden Sofies farmor døde.
 that was NEG so long since Sophie’s grandmother died

(Later, when Sophie thought about her being here, she realized that she would not be here always. “I am in this world now”, she thought, “but one day I’ll be gone.” Was there life after death? This was another question the cat was probably quite unaware of.)
It wasn’t LONG ago that Sophie’s GRANDMOTHER had died.

The logical subject of the clause is *Sofies farmor* ‘Sophie’s grandmother’, which is referentially new (even though it is anchored through “Sophie”), and may therefore not occur as subject of the main clause. It gets moved into the subordinate clause by using an *it*-cleft construction.

The *it*-cleft could also be seen as an avoidance strategy in English, but then in relation with the decrease in subject-auxiliary inversion (Hasselgård, 2004). Recall the introduction, section 1.2.2, and in particular Figure 1, which show that prepositional phrases in particular are become decreasingly used as first constituents that trigger the subject and the auxiliary (in clauses that include both an auxiliary and a non-finite verb) to switch places, so that the auxiliary *precedes* rather than follows the subject. Consider for example the late Modern English subject-auxiliary inversion in (339):

- (339) a. When these prodigious Forces were throughly furnish'd, they look'd as if all the Inhabitants of the East, assembl'd together, had been going to people another Continent, rather than an Army rais'd to take one single City;
 for [pp against Athens] **was the main Quarrel**, and all these mighty Preparations chiefly design'd. [hind-1707:69-70]
 b. It was [pp against Athens] that the main Quarrel and all these mighty preparations were directed primarily.

The prepositional phrase *against Athens* in (339a) occurs clause-initially where it is accompanied by subject-auxiliary inversion. The reason for this inversion seems to be that there is constituent focus on the noun phrase *Athens*, since this provides the value of the variable that is set up by the mention of “one single city” in the preceding clause. The construction in (339a) nevertheless sounds quite archaic to modern speakers of English, and I would argue that the *it*-cleft in (339b) provides a much better alternative. One of the things the *it*-cleft in (339b) does is provide an alternative, an “avoidance” strategy, for the subject-auxiliary inversion in (339a).³

If it is true that making use of one particular construction (such as the *it*-cleft) is a strategy to avoid a more “costly” construction (such as a referentially new subject in a Swedish main clause or subject-auxiliary inversion in English), then a correct grammatical framework should be able to deal with such avoidance strategies, which may combine syntactic and referential features (such as banning “referentially new subjects after the finite verb”). One grammatical model that allows for avoidance strategies is bidirectional optimality theory (Blutner et al., 2006). Another model is the functional descriptive framework employed in the analysis of the two narrative texts in chapter 4: this too seems to be capable of dealing with more and less marked constructions.

13.5 Focus is compositional

This study has adopted a method to locate focus domains and, consequently, focused constituents (where there is constituent focus or presentational focus) that is proving to be quite successful, and that entails that focus is a *compositional* notion. Recall that chapter 2, building on the work of psycholinguistics, supports the framework where addressees build a *mental model* that includes *mental entities*. These mental entities receive properties, and sometimes link to existing entities in long term memory.

Chapter 5 took a, seemingly logical, further step in arguing that noun phrases can have a referential state, which describes their relation to the mental model, and that there is a small set of five referential state *primitives*. I have been arguing that a combination of syntactic information and information about the referential states of noun phrases is sufficient to determine the focus articulation of a particular clause. If we have, for instance, a clause with a postverbal subject that has the referential state “new”, and that does not have an anchor (see definition (193) in section 8.1), then we can safely assume a situation of presentational focus. This is one example but there are other combinations of syntactic situations and referential states of noun phrases within clauses that clearly indicate the clause has one particular focus articulation. A test case has been provided in section 5.5.3, where I undertook to derive the focus articulations of copula clauses where the subject and complement varied in terms of syntactic and referential categories.

The research done until now has not yet reached the point where I can say that we are able to automatically derive the focus articulation of *any* type of clause, but I envision future work will bring us there. If we, for a moment, assume that we reach the point where we can look at the syntactic and referential features of the elements of a clause and then determine the focus articulation of that clause based on this information, then this entails that focus has a compositional nature: it consists of the building blocks of syntax and referential states.

The grammatical “atoms” of syntax (which defines which element belongs to which constituent, the hierarchical organization of these constituents and their morphological features) and of referentiality (the referential states of constituents) can combine into all kinds of clausal “molecular” structures, but these “molecules” can only be of three basic types, which are the three different focus articulations.

I would like to take the reasoning above one step further: if we agree that focus is part of the grammar of a language, and if we agree on the conclusion I just reached that focus can be arrived at by combining syntactic and referential information, then the key elements of the rules that determine the location and size of the focus domain are syntactic and referential information.

13.6 Future work

This study is based on work in a variety of different areas, and this is also reflected in the suggestions for future work here. The narrative charting approach described in chapter 4, even though time consuming, warrants a follow-up (especially since the initial charting has been shown to lend itself for an automatic approach): texts from

three more time-periods (ME, eModE and PDE) should be scrutinized in order to arrive at an even better picture of what is going on—not only in terms of changes in focus, but also in terms of how the text organizational strategies change.

The definition of the referential state primitives in chapter 5 is quite thorough, and sections 5.4.3 and 5.4.4 indicate that the small set should be sufficient. Nevertheless, the preliminary conclusion that the generic category “Kind” (now a subset of “New” and “Inert”) does not need to be distinguished separately warrants further research. And even though the opaque contexts leading to the category of “Non-Specific” has been shown to be determinable, it would still be good to investigate whether referentially New entities created in opaque contexts lead to information structure behaviour that deviates from other referentially New entities. If research in these areas reveals that additional referential categories are needed, this has consequences for the Cesax algorithm and program. This program, described in chapter 6, is an area for further work too. The chapter itself already mentions several extensions: fine-tuning of the constraints depending on the text period, critical evaluation and possibly extension of the constraints, and fine-tuning of the suspicious situations. All of these improvements aim at increasing the percentage of correct automatically made coreferential links and the percentage of correct suggestions in the presence of suspicious situations. This last percentage can, perhaps, be increased by combining the constraint-based Cesax method with a statistical coreference resolution method.

The program CorpusStudio as described in chapter 7 proves its value in the area of information structure research described in chapters 8 and 9. There are two areas of development for CorpusStudio I would like to suggest. The first one would be to check if it is feasible to come up with a web version of CorpusStudio, making it much more platform independent. The second extension that should be made is a user-friendly interface to enter and edit queries. If such an interface would also be supplied for quick find searches in Cesax, corpus searches would get much closer to student and researcher.

The experiments in chapters 8 and 9 are valuable as they are, but the statistical significance of the results can be much improved by increasing the amount of referentially enriched texts. This is a major job, one that, in my opinion, needs doing, and I think it will return the investment in time and energy as we seek to answer more questions in the information structure research. An example of an experiment that has been put on a halt until more data is available is the use of constituent negation as a diagnostic for constituent focus, as described in section 9.3.

Future work will have to show whether the conclusions on the compositionality of focus stated in chapter 13 hold. The claim can, on the one hand, be falsified quite easily by coming up with at least one clause whose focus articulation cannot be determined on the basis of its syntax in combination with the referential states of its components. But instead of (or in addition to) a falsification attempt, it may be fruitful to see how far we can get in examining how the combination of syntactic constellations with referential categories imply particular focus domains, associating with focus articulations. This approach has already started with the examination of copula clauses, but it could continue with other clause types, such as simple

transitive or intransitive clauses. We would have to find a whole paradigm of examples with all possible combinations of syntactic and referential categories for the different components of these constructions. The next step would be to look at each of the combinations in context and determine what the focus articulation is, taking into account that there may be a point of departure, a dominant focal element or a reordering due to the principle of natural information flow. At the same time such practical approaches as sketched here are undertaken, it would be a challenge to see if we can find a *theoretical* basis for the idea that the combination of syntax and coreference information leads to particular focus articulations.

In sum: there is enough work ahead of us, and there is the tempting perspective of confirming the hypothesis that focus is compositional: that linguistics too, just like physics, has its atomic structures (such as the referential categories), which combine with other elements (the syntax) into a restricted set of meaningful molecules (the focus articulations).

¹ Constituent focus is *not* the same as the highlighting of one constituent as “dominant focal element” within the larger focus domain of a topic-comment articulation or athetic articulation clause (see 3.3.3).

² I refrain from stating exactly how the *it*-cleft structure looks like syntactically, since this is a point of much debate over the last decades, and the only thing that matters for the point I am trying to make here is the fact that the syntax of a simple equative clause with complex complement differs from that of an *it*-cleft.

³ The other thing the *it*-cleft does is provide a natural location for focus to be realized on the prepositional phrase: the predicate of an equative construction that satisfies syntax by having a subject pronoun *it*, but whose elements are otherwise referentially void.



Bibliography

- Ahlemeyer, Birgit, and Inga Kholhof. 1999. "Bridging the cleft: an analysis of the translation of English *it*-clefts into German". *Languages in Contrast* 2:1-25.
- Akmajian, Adrian. 1979. *Aspects of the grammar of focus in English*. New York: Garland publishing.
- Archibald, Elizabeth. 1991. *Apollonius of Tyre : medieval and Renaissance themes and variations : including the text of the Historia Apollonii Regis Tyri with an English translation*. Cambridge; Rochester, NY, USA: D.S. Brewer ; Boydell & Brewer.
- Ariel, Mira. 1994. "Interpreting anaphoric expressions: a cognitive versus pragmatic approach". *Journal of Linguistics* 30:3-42.
- Ariel, Mira. 1999. *Accessing noun-phrase antecedents*. London and New York: Routledge.
- Bailey, Nicholas Andrew. 2009. *Thetic constructions in Koine Greek*. Ph. D. dissertation, Vrije Universiteit
- Baker, Peter S. 2003. *The Electronic Introduction to Old English* Oxford: Blackwell, Available at: <http://www.wmich.edu/medieval/resources/IOE/index.html>.
- Ball, Catherine N. 1991. *The historical development of the it-cleft*. Ph.D., University of Pennsylvania
- Ball, Catherine N. 1994. "The origins of the informative-presupposition *it*-cleft". *Journal of Pragmatics* 22:603-628.
- Baugh, Albert C., and Thomas Cable. 2002. *A history of the English Language*, fifth. London: Routledge.
- BBC. 2009. *Eyewitness: Colombo air raid*: BBC News, Available at: http://news.bbc.co.uk/2/hi/south_asia/7902684.stm.
- BBC. 2011. *Changing the way Indians shop*: BBC News, Available at: <http://www.bbc.co.uk/news/world-asia-india-15885055>.
- Beaver, David I. 2004. "The Optimization of Discourse Anaphora". *Linguistics and Philosophy* 27:3-56. Available at: <https://webpace.utexas.edu/dib97/tooda.pdf>.
- Bech, Kristin. 1999. "Are Old English conjunct clauses really verb-final?". *Historical linguistics 1999 : selected papers from the 14th International Conference on Historical Linguistics, Vancouver, 9-13 August 1999*. ed. by Laurel J. Brinton. Amsterdam; Philadelphia: J. Benjamins.
- Beneš, Eduard. 1962. "Die Verbstellung im Deutschen, von der Mitteilungsperspektive her betrachtet". *Phonologica Pragensia* 5:6-19.
- Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Longman.
- Biberauer, Theresa, and Ans van Kemenade. 2011. "Subject Positions and Information-Structural Diversification in the History of English". *Catalan Journal of Linguistics* 10:17-69.

- Birner, Betty, and Gregory Ward. 1998. *Information Status and Canonical Word Order in English*. Amsterdam/Philadelphia: John Benjamins.
- Blutner, Reinhard, Helen de Hoop, and Petra Hendriks. 2006. *Optimal communication*. Stanford: CSLI publications.
- BNC. 2007. *The British National Corpus, version 3 (BNC XML edition)*: Oxford University Computing Services on behalf of the BNC Consortium, Available at: <http://www.natcorp.ox.ac.uk/>.
- Boag, Scott, Don Chamberlin, Mary F. Fernández, Daniela Florescu, Jonathan Robie, and Jérôme Siméon. 2010. *XQuery 1.0: An XML Query Language (Second Edition)*: W3C Recommendation, Available at: <http://www.w3.org/XML/Query/#specs>.
- Boersma, Paul, and David Weenink. 2005. "PRAAT, a system for doing phonetics by computer." *Glott International* 5:341-345.
- Boersma, Paul, and David Weenink. 2008. *Praat: doing phonetics by computer (Version 5.0.32) [Computer program]*. Retrieved December 1, 2008, from <http://www.praat.org/>
- Bolinger, Dwight. 1977. *Meaning and form*. London; New York: Longman.
- Bouma, Gerlof. 2003. "Doing Dutch pronouns automatically in Optimality Theory". Paper presented at *The EACL 2003 Workshop on The Computational Treatment of Anaphora*
- Bouma, Gosse, and Geert Kloosterman. 2007. Mining syntactically annotated corpora with XQuery. In *Proceedings of the Linguistic Annotation Workshop*. Prague, Czech Republic: Association for Computational Linguistics.
- Brants, Sabrin, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. "The Tiger treebank". *Proceedings of the Workshop on Treebanks and Linguistic Theories* Sozopol, Available at: <http://www.ims.uni-stuttgart.de/projekte/TIGER/paper/treeling2002.pdf>.
- Bresnan, Joan. 1994. "Locative inversion and the architecture of universal grammar". *Language* 70:72-131.
- Bresnan, Joan, and Jane Grimshaw. 1978. "The Syntax of Free Relatives in English". *Linguistic Inquiry* 9:331-391.
- Brewer, W., and J. C. Treyns. 1981. "Role of schemata in memory for places". *Cognitive Psychology* 13:207-230.
- Brinton, Laurel J. 1990. "The stylistic function of Middle English *gan* reconsidered". *Papers from the 5th International Conference on English Historical Linguistics*. ed. by Sylvia M. Adamson, Vivien A. Law, Nigel Vincent and Susan Wright, 31-53. Amsterdam/Philadelphia: John Benjamins Pub. Co.
- Büring, Daniel. 2005. Focus, prosody and syntax: Lessons given at École normale supérieure, Paris, Available at: <http://www.diffusion.ens.fr/index.php?res=cycles&idcycle=217>.
- Burzio, Luigi. 1986. *Italian syntax : a government-binding approach*. Dordrecht; Boston; Hingham, MA: Reidel.
- Callows, Kathleen. 1974. *Discourse considerations in translating the word of God*. Grand Rapids: Zondervan publishing house.
- Calude, Andreea S. 2008. "Clefting and extraposition in English". *International Computer Archive of Modern and Medieval English Journal* 32:7-33.

- Cann, Ronnie. 2003. "Interpreting *Be*". *Proceedings of the Conference "sub7 - Sinn und Bedeutung"*. *Arbeitspapier Nr. 114*. ed. by Matthias Weisgerber, 95-109. Germany: FB Sprachwissenschaft, Universität Konstanz, Available at: http://ling.uni-konstanz.de/pages/conferences/sub7/proceedings/download/sub7_cann.pdf.
- Chafe, Wallace L. 1976. "Givenness, contrastiveness, definiteness, subjects, topics and point of view". *Subject and topic*. ed. by Charles N. Li, 25-56. New York: Academic Press.
- Chafe, Wallace L. 1987. "Cognitive constraints on information flow". *Coherence and grounding in discourse*. ed. by Russell Tomlin, 21-52. Amsterdam: John Benjamins.
- Cheng, Lisa, and Laura J. Downing. 2009. Against FocusP: Arguments from Zulu. Ms.
- Chomsky, Noam. 1957. *Syntactic structures*. 's-Gravenhage: Mouton.
- Chomsky, Noam. 1971. "Deep structure, surface structure, and semantic interpretation". *Semantics: An interdisciplinary reader in philosophy, linguistics, and psychology*. ed. by Danny Steinberg and Leon Jakobovits, 183-216. Cambridge: Cambridge university press.
- Chomsky, Noam, and Morris Halle. 1968. *The sound pattern of English*. New York: Harper and Row.
- Clark, Stephen A. 2012. *Participant Reference in Narrative Discourse: A Comparison of Three Methodologies* Dallas, Texas: SIL International, Available at: http://www.sil.org/silepubs/Pubs/928474547863/ebook_35_Clark_Participant_Reference.pdf.
- Comrie, Bernard. 1989. *Language universals and linguistic typology*. Chicago: The university of Chicago press.
- Cornish, Francis. 1999. *Anaphora, discourse, and understanding : evidence from English and French*. Oxford; New York: Clarendon Press.
- Cowie, Jim. 2011. Personal communication. Email: 21/mar/2011.
- Crystal, David. 1980. *A first dictionary of linguistics and phonetics*. London: Deutsch.
- Davies, Mark. 2004-2012. *BYI-BNC*: (Based on the British National Corpus from Oxford University Press), Available at: <http://corpus.byu.edu/bnc/>.
- de Vries, Mark. 2007. "Dislocation and backgrounding". *Linguistics in the Netherlands 2007*. AVT publications 24, ed. by Bettelou Los and Marjo van Koppen, 235-247. Amsterdam: John Benjamins publishing company.
- Declerck, Renaat. 1983. "Predicational clefts". *Lingua* 61:9-45.
- Declerck, Renaat. 1984. "The pragmatics of it-clefts and wh-clefts". *Lingua* 64:251-289.
- Delahunty, Gerald P. 1984. "The analysis of English cleft sentences". *Linguistic Analysis* 13:63-113.
- Delin, Judy. 1990. Focus in cleft constructions. In *Internal research series "Blue Book Note" No. 5*. Edinburgh: University of Edinburgh.
- Delin, Judy. 1992. "Properties of it-cleft presupposition". *Journal of Semantics* 9:179-196.

- Dijk, T. A. Van, and W. Kintsch. 1983. *Strategies in discourse comprehension*. New York: Academic press.
- Doherty, Monika. 2006. *Structural propensities*. Amsterdam/Philadelphia: John Benjamins.
- Dooley, Robert A., and Stephen H. Levinsohn. 2001. *Analyzing discourse: basic concepts*: Summer Institute of Linguistics.
- Drach, Erich. 1937. *Grundgedanken der deutschen Satzlehre*. Frankfurt am Main: Diesterweg.
- Drubig, H. Bernhard. 2000. "Toward a typology of focus and focus constructions". *Linguistics* 41:1-50. Available at: <http://dx.doi.org/10.1515/ling.2003.003>.
- Dryer, Matthew S. 2003. "Word Order". *International encyclopedia of linguistics, 2nd edition*. ed. by William J. Frawley, 376-377. Oxford; New York: Oxford University Press.
- Dufter, Andreas. 2009. "Clefting and discourse organization: comparing Germanic and Romance". *Focus and background in Romance languages*. ed. by Andreas Dufter and Daniel Jacob, 83-122. Amsterdam: John Benjamins.
- Dumas, Alexandre. 1878. *The three musketeers*. London: George Routledge and sons.
- Eckhoff, Hanne, and Dag T. T. Haug. 2011. Personal pronouns with articles: a quantitative approach. In *Information structure and corpus annotation: theoretical and practical perspectives*. Oslo, Lysebu: University of Oslo.
- Ellegård, Alvar. 1953. *The Auxiliary Do: the Establishment and Regulation of its Use in English*. Stockholm, Almqvist, Wiksell: Göteborg.
- Enkvist, Nils Erk. 1986. "More about the textual functions of the Old English adverbial *þa*". *Linguistics across historical and geographical boundaries: in honour of Jacek Fisiak on the occasion of his fiftieth birthday*. Volume 1, ed. by I. D. Kastovsky and A. Szwedek. Berlin: Mouton de Gruyter.
- Enkvist, Nils Erk, and Brita Warvik. 1987. "Old English *tha*, temporal chains and narrative structure". *Papers from the 7th international conference on historical linguistics*. ed. by Anna Giacalone Ramat, Onofrio Carrua and Guiliano Bernini, 221-237. Amsterdam/Philadelphia: John Benjamins.
- Erteschik-Shir, Nomi. 2007. *Information Structure: the syntax-discourse interface*. Oxford: Oxford university press.
- Evenhuis, John K. 2006. *English and Spanish it-clefts in contrast and contact*. MA, California State University
- Faarlund, Jan Terje. 1990. *Syntactic change : toward a theory of historical syntax*. Berlin; New York: M. de Gruyter.
- Féry, Caroline, and Manfred Krifka. 2008. "Information structure: notional distinctions, ways of expression". *Unity and diversity of languages*. ed. by Piet van Sterkenburg, 123-135. Amsterdam/Philadelphia: John Benjamins.
- Filppula, Markku. 2009. "The rise of it-clefting in English: areal-typological and contact-linguistic considerations". *English Language and Linguistics* 13:267-293.
- Firbas, Jan. 1964. "From comparative word-order studies". *BRNO studies in English* 4:111-126.
- Fischer, Olga, Ans Van Kemenade, Willem Koopman, and Wim Van der Wurff. 2000. *The Syntax of Early English*. Cambridge: Cambridge University Press.

- Ford, Cecilia E. 1993. *Grammar in interaction : adverbial clauses in American English conversations*. Cambridge; New York: Cambridge University Press.
- Garnham, Alan. 2001. *Mental models and the interpretation of anaphora*. Hove, East Sussex: Psychology Press Ltd.
- Gegg-Harrison, Whitney, and Donna K. Byron. 2006. "PYCOT: an optimality theory-based pronoun resolution toolkit". Paper presented at *Proceedings of the Language Resources and Evaluation (LREC 2006)*.
- Givón, Talmy. 1982. "Logic versus pragmatics, with human language as the referee: toward an empirically viable epistemology". *Journal of Pragmatics* 6:81-133.
- Givón, Talmy. 1983. "Topic continuity in discourse: an introduction". *Topic continuity in discourse: a quantitative cross-language study*. ed. by Talmy Givón. Amsterdam: John Benjamins.
- Goldsmith, J.A. 1979. *Autosegmental phonology*. New York: Garland Press.
- González-Cruz, Ana Isabel. 2003. "Adverbials in focus: adverbial clauses in cleft constructions in the history of English". *Fifty years of English studies in Spain (1952-2002)*. ed. by Ignacio M. Palacios Martínez, María José López Couso, Patricia Fra López and Elena Seoane Posse, 311-319. Santiago de Compostela: Universidade de Santiago de Compostela.
- Götze, Michael, Thomas Weskott, Cornelia Endriss, Ines Fiedler, Stefan Hinterwimmer, Svetlana Petrova, Anne Schwarz, Stavros Skopeteas, and Ruben Stoel. 2007. "Information structure". *Working Papers of the SFB632, Interdisciplinary Studies on Information Structure 07*. ed. by Stefanie Dipper, Michael Götze and Stavros Skopeteas, 147-187, Available at: http://www.sfb632.uni-potsdam.de/publications/isis07_6is.pdf.
- Grimshaw, Jane. 1997. "Projection, heads, and optimality". *Linguistic Inquiry* 28:373-422.
- Grosz, Barbara J., Aravind K. Joshi, and Scott Weinstein. 1995. "Centering: a framework for modeling the local coherence of discourse". *Computational Linguistics* 21:203-226.
- Gundel, Jeanette. 1974. *The role of topic and comment in linguistic theory*. PhD dissertation, University of Texas, Austin
- Gundel, Jeanette K. 1988. "Universals of topic-comment structure". *Studies in syntactic typology*. ed. by Michael Hammond, Edith A. Moravcsik and Jessica R. Wirth, 209-239. Amsterdam/Philadelphia: John Benjamins.
- Gundel, Jeanette K. 2002. "Information structure and the use of cleft sentences in English and Norwegian". *Information structure in a cross-linguistic perspective*. 39, ed. by Hilde Hasselgård, Stig Johansson, Bergljot Behrens and Cathrine Fabricius-Hansen. Amsterdam, New York: Rodopi.
- Gundel, Jeanette K., Nancy Hedberg, and Ron Zacharski. 1993. "Cognitive status and the form of referring expressions in discourse". *Language* 69:274-307.
- Gundel, Jeanette K. 1977. "Where do cleft sentences come from?" [September 1977]. *Language* 53:543-559.
- Gussenhoven, Carlos. 1983a. "Focus, mode and the nucleus". *Journal of Linguistics* 19:377-417.
- Gussenhoven, Carlos. 1983b. "Testing the Reality of Focus Domains". *Language and speech* 26:61-80.

- Gussenhoven, Carlos. 2004. *The phonology of tone and intonation*, 2005 reprint. Cambridge: Cambridge University Press.
- Gussenhoven, Carlos. 2007. "Types of Focus in English". *Topic and Focus. Cross-Linguistic Perspectives on Meaning and Intonation* 82, ed. by Chungmin Lee, Matthew Gordon and Daniel Büring, 83-100: Springer Netherlands.
- Gussenhoven, Carlos, and Erwin R. Komen. forthcoming. "Chechen intonation".
- Haerberli, Eric. 2002. "Inflectional morphology and the loss of verb-second in English". *Syntactic effects of morphological change*. ed. by David Lightfoot, 88-106. Oxford and New York: Oxford University Press.
- Hagoort, Peter, and Jos J. A. van Berkum. 2007. "Beyond the sentence given". *Philosophical transactions of the royal society of biological sciences* 362:801-811.
- Halliday, Michael A.K. 1967. "Notes on transitivity and theme in English, part II". *Journal of Linguistics* 3:199-244.
- Hannay, Mike, and J. Lachlan Mackenzie. 2002. *Effective writing in English: a sourcebook*. Bussum: Coutinho.
- Harries-Delisle, Helga. 1978. "Contrastive emphasis and cleft sentences". *Universals of human language, Volume IV: syntax*. 4, ed. by Joseph H. Greenberg, 419-486. Stanford: Stanford University Press.
- Hartmann, Jutta M. 2008. *Expletives in existentials*. Utrecht: LOT.
- Hartmann, Katharina, and Malte Zimmermann. 2004. "Focus Strategies in Chadic: The Case of Tangale Revisited". *Interdisciplinary Studies on Information Structure*. 1, ed. by Shinichiro Ishihara, Michaela Schmitz and Anne Schwarz, 207-243. Potsdam: Universitätsverlag Potsdam.
- Hasselgård, Hilde. 2004. "Adverbials in it-cleft constructions". *Advances in corpus linguistics*. ed. by K. Aijmer and B. Altenberg, 195-211. Amsterdam & New York: Rodopi.
- Haug, Dag. 2009. Info-structural annotation in the PROIEL corpus. In *Annotating and analysing information structure in historical corpus texts*. Berlin.
- Haug, Dag T. T., Hanne M. Eckhoff, and Eirik Welo. forthcoming. "The theoretical foundations of givenness annotation". *Information structure and word order changes*. ed. by Kristin Bech and Kristine Gunn Eide. Oslo: John Benjamins.
- Haug, Dag T. T., Marius L. Jøhndal, Hanne M. Eckhoff, Eirik Welo, Mari J. B. Hertzberg, and Angelika Müth. 2009. "Computational and Linguistic Issues in Designing a Syntactically Annotated Parallel Corpus of Indo-European Languages". *TAL* 50:17-45. Available at: <http://www.atala.org/IMG/pdf/TAL-2009-50-2-01-Haug.pdf>.
- Hedberg, Nancy. 1988. The discourse function of cleft sentences in spoken discourse. In *Linguistic Society of America Meeting*. New Orleans, Louisiana.
- Hedberg, Nancy. 1990. *Discourse pragmatics and cleft sentences in English*. Ph.D. thesis, University of Minnesota
- Hedberg, Nancy. 2007. "The Information Structure of It-clefts, Wh-clefts and Reverse Wh-clefts in English". *The Grammar-Pragmatics Interface: Essays in Honor of Jeanette K. Gundel*. . ed. by Nancy Hedberg and Ron Zacharski, 49-76. Amsterdam/Philadelphia: John Benjamins, Available at: http://www.sfu.ca/~hedberg/Clefts_paper17.pdf.

- Heimerdinger, Jean-Marc. 1999. *Topic, Focus and Foreground in Ancient Hebrew Narratives*. Sheffield: Sheffield Academic Press.
- Hendriks, Petra. 2004. "Optimization in focus identification". *Optimality Theory and Pragmatics*. ed. by Reinhard Blutner and Henk Zeevat, 42-62: Palgrave Macmillan.
- Hinterhölzl, Ronald, and Ans van Kemenade. 2012. "The interaction between syntax, information structure and prosody in word order change". *The Oxford Handbook of the History of English*. ed. by Elizabeth Closs Traugott and Terttu Nevalainen, (Lead paper for the Interfaces section). New York: Oxford University Press.
- Hobbs, Jerry R. 1978. "Resolving pronoun references". *Lingua* 44:311-338.
- Hopper, Paul J., and Sandra A. Thompson. 1984. "The discourse basis for lexical categories in universal grammar". *Language* 60:703-752.
- Horn, L. 1981. Exhaustiveness and the semantics of clefts. In *Proceedings of NELS 11*, 125-142.
- Hoste, Véronique. 2005. *Optimization Issues in Machine Learning of Coreference Resolution*. Ph. D. dissertation, Universiteit Antwerpen
- Huddleston, Rodney D., and Geoffrey K. Pullum. 2002. *The Cambridge grammar of the English language*. Cambridge, UK; New York: Cambridge University Press.
- Hukari, Thomas E., and Robert D. Levine. 1995. "Adjunct extraction". *Journal of Linguistics* 31:195-226.
- Hulk, Aafke, and Ans van Kemenade. 1993. "Subjects, nominative case, agreement and functional heads". *Lingua* 89:181-215.
- ICE-GB. 2011. *The International Corpus of English, British component*: University College London, Available at: <http://www.ucl.ac.uk/english-usage/ice-gb>.
- Ingham, Richard. 2001. "The structure and function of expletive *there* in premodern English". *Reading working papers in Linguistics* 5:231-249. Available at: http://www.reading.ac.uk/acadepts/ll/app_ling/internal/workingpapers/ingham.pdf.
- Jacobs, J. 2001. "The dimensions of topic-comment". *Linguistics* 39:641-682.
- Jane, L. C., and A. M. Sellar. 2011. *Ecclesiastical History of the English People*: Wikisource, Available at: http://en.wikisource.org/wiki/Ecclesiastical_History_of_the_English_People.
- Jespersen, Otto. 1927. *A Modern English Grammar on historical principles, part III - syntax, second volume*. Heidelberg: Carl Winters Universitätsbuchhandlung.
- Jespersen, Otto. 1937. *Analytic syntax*. London: Allen and Unwin.
- Jespersen, Otto, and Niels Haislund. 1949. *A Modern English Grammar on historical principles, part VII - syntax*. Copenhagen: Ejnar Munksgaard.
- Johansson, Mats. 2001. "Clefts in contrast: a contrastive study of it clefts and wh clefts in English and Swedish texts and translations". *Linguistics* 39:547-582.
- Johansson, Mats. 2002. *Clefts in English and Swedish: a contrastive study of IT clefts and WH clefts in original texts and translations*. Ph.D. thesis, Lund University
- Johnson-Laird, P. N. 1983. *Mental models: towards a cognitive science of language, inference, and consciousness*. Cambridge, MA: Harvard University Press.

- Kaiser, Elsi, and John C. Trueswell. 2004. "The Role of Discourse Context in the Processing of a Flexible Word-Order Language". *Cognition* 94:113-147.
- Karttunen, Lauri. 1969. "Discourse referents". *COLING '69 Proceedings of the 1969 conference on Computational linguistics* 1-35. Stroudsburg, PA: Association for computational linguistics.
- Karttunen, Lauri. 2003. "Discourse referents". *Semantics: Critical Concepts in Linguistics, Vol. III*. ed. by Javeier Gutiérrez-Rexach, 20-39. Stroudsburg, PA: Routledge.
- Keenan, E. L., and Bernard Comrie. 1977. "Noun Phrase Accessibility and Universal Grammar". *Linguistic Inquiry* 8:63-99.
- Kehler, Andrew, Douglas Appelt, Lara Taylor, and Aleksandr Simma. 2004. "The (non)utility of predicate-argument frequencies for pronoun interpretation". Paper presented at *HLT/NAACL*, Boston, MA.
- Ker, N. R. editor. 1956. *The pastoral care King Alfred's translation of St. Gregory's Regula Pastoralis : Ms Hatton 20 in the Bodleian Library at Oxford : MS Cotton Tiberius B.XI in the British Museum, MS Anhang 19 in the Landesbibliothek at Kassel*. Copenhagen: Rosenkilde and Bagger.
- Kimmelman, Vadim. 2009. "On the Interpretation of èto in so-called èto-clefts". *Studies in Formal Slavic Phonology, Morphology, Syntax, Semantics and Information Structure: Proceedings of FDSL 7, Leipzig 2007*. ed. by Gerhild ybatow, Denisa Lenertová, Uwe Junghanns and Petr Biskup, 319-329. Frankfurt: Peter Lang.
- Kiss, Katalin É. 1998. "Identificational focus and information focus.". *Language*:245-273.
- Kloosterman, Geert. 2007. *An overview of the Alpino Treebank tools*: Alfa-informatica, University of Groningen. Available at: <http://www.let.rug.nl/vannoord/alp/Alpino/Treebank-Tools.html>.
- Komen, Erwin R. 2007a. Chechen stress and vowel deletion. Ms., *Rutger's Optimality Archive*, Available at: <http://roa.rutgers.edu/view.php?id=1320>.
- Komen, Erwin R. 2007b. *Focus in Chechen*. Master's Thesis, Leiden University
- Komen, Erwin R. 2008. "Branching constraints". *Optimality Theory and Minimalism: Interface Theories*. 28, ed. by Hans Broekhuis and Ralf Vogel, 157-186, Available at: <http://www.ling.uni-potsdam.de/lip/>.
- Komen, Erwin R. 2009a. CESAC: Coreference Editor for Syntactically Annotated Corpora. In *7th York-Newcastle-Holland Symposium on the History of English Syntax (SHES7)*, 8. Nijmegen, CLS/Department ENglish Language and Culture: Radboud University.
- Komen, Erwin R. 2009b. *Corpus Studio manual* Nijmegen: Radboud University Nijmegen, Available at: http://erwinkomen.ruhosting.nl/software/CorpusStudio/CorpStu_Manual.pdf.
- Komen, Erwin R. 2009c. *CorpusStudio* Nijmegen: Radboud University Nijmegen, Available at: <http://erwinkomen.ruhosting.nl/software/CorpusStudio>.
- Komen, Erwin R. 2010. "Overriding negative concord". *Linguistics in Amsterdam* 3:1-20. Available at: <http://www.linguisticsinamsterdam.nl/cgi/t/text/get-pdf?idno=m0302a02>.

- Komen, Erwin R. 2011a. *Cesax: coreference editor for syntactically annotated XML corpora* Nijmegen, Netherlands: Radboud University Nijmegen, Available at: <http://erwinkomen.ruhosting.nl/software/Cesax>.
- Komen, Erwin R. 2011b. *Cesax: coreference editor for syntactically annotated XML corpora. Reference manual*, 1.5.4 Nijmegen, Netherlands: Radboud University Nijmegen, Available at: http://erwinkomen.ruhosting.nl/software/Cesax/Cesax_Manual.pdf.
- Komen, Erwin R. 2012. "Coreferenced corpora for information structure research". *Outposts of Historical Corpus Linguistics: From the Helsinki Corpus to a Proliferation of Resources. (Studies in Variation, Contacts and Change in English 10)*. ed. by Jukka Tyrkkö, Matti Kilpiö, Terttu Nevalainen and Matti Rissanen. Helsinki, Finland: Research Unit for Variation, Contacts, and Change in English, Available at: <http://www.helsinki.fi/varieng/journal/volumes/10/index.html>.
- Krifka, Manfred. 2007. "Basic notions of information structure". *Interdisciplinary studies on information structure 06*. 06, ed. by Caroline Féry, Gisbert Fanselow and Manfred Krifka, 1-50.
- Kroch, Anthony. 1989. "Reflexes of grammar in patterns of language change". *Language variation and change* 1:199-244.
- Kroch, Anthony, Beatrice Santorini, and Ariel Diertani. 2004. *Penn-Helsinki Parsed Corpus of Early Modern English*, Available at: <http://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-2/index.html>
- Kroch, Anthony, Beatrice Santorini, and Ariel Diertani. 2010. *Penn parsed corpus of modern British English*, Available at: <http://www.ling.upenn.edu/hist-corpora/PPCMBE-RELEASE-1/index.html>.
- Kroch, Anthony, and Ann Taylor. 2000. *Penn-Helsinki Parsed Corpus of Middle English, second edition.*, Available at: <http://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-2/>
- Kruijff-Korbayová, Ivana, and Mark Steedman. 2003. "Discourse and information structure". *Language and information* 12:249-259.
- Kügler, F., and S. Skopeteas. 2006. "Interaction of lexical tone and information structure in Yucatec Maya". Paper presented at *Proceedings of the Second International Symposium on Tonal Aspects of Languages (TAL-2)*, Université de La Rochelle.
- Kuno, Susumu. 1973. *The structure of the Japanese language*. Cambridge, Massachuset: MIT Press.
- Lambrecht, Knud. 1994. *Information structure and sentence form. Topic, focus and the mental representations of discourse referents.*: Cambridge university press.
- Lambrecht, Knud. 2001. "A framework for the analysis of cleft constructions". *Linguistics* 39:463-516.
- Lambrecht, Knud. 2010. "Constraints on the subject-focus mapping in French and English". *Comparative and contrastive studies of information structure*. ed. by Carsten Breul and Edward Göbbel, 77-99. Amsterdam/Philadelphia: John Benjamins.
- Lappin, Shalom, and Herbert J. Leass. 1994. "An algorithm for pronominal anaphora resolution". *Computational Linguistics* 20:535-561.

- Leech, Geoffrey N., and Michael H. Short. 1981. *Style in Fiction: A Linguistic Introduction to English Fictional Prose*. London/New York: Longman.
- Levinsohn, Stephen. 2000. *Discourse features of New Testament Greek: A coursebook on the information structure of New Testament Greek*, second edition. Dallas: SIL-International.
- Levinsohn, Stephen. 2009. *Self-instruction materials on narrative discourse analysis*: SIL-International, Available at: www.sil.org/~levinsohns/narr.pdf.
- Levinsohn, Stephen H. 1992. "Preposed and postposed adverbials in English". *Work papers of the Summer Institute of Linguistics, University of North Dakota session*. 36, ed. by Robert A. Dooley and David F. Marshall, 19-31: University of North Dakota.
- Links, Meta M. 2010. *Exploring the transitive expletive construction in earlier English*. MA thesis, Radboud University
- Longacre, Robert, and Stephen Levinsohn. 1978. "Field analysis of discourse". *Current trends in textlinguistics*. ed. by Wolfgang U. Dressler, 103-122. Berlin, New York: Walter de Gruyter.
- Loos, Eugene E. 2003. *Glossary of linguistic terms*: SIL International, Available at: <http://www.sil.org/linguistics/GlossaryOfLinguisticTerms/>.
- Los, Bettelou. 2000. "Onginnan/beginnan with bare and to-infinitive in Ælfric". *Pathways of Change. Grammaticalization in English*. ed. by Olga Fischer, Anette Rosenbach and Dieter Stein, 251-274. Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Los, Bettelou. 2009. "The consequences of the loss of verb-second in English: information structure and syntax in interaction". *English Language and Linguistics* 13:97-125.
- Los, Bettelou. 2012. "The Loss of Verb-Second and the Switch from Bounded to Unbounded Systems". *Information structure and syntactic change in the history of English*. ed. by Anneli Meurman-Solin, María José López-Couso and Bettelou Los. Oxford: Oxford University Press.
- Los, Bettelou, and Gea Dreschler. 2012. "The loss of local anchoring: From adverbial local anchors to permissive subjects". *Rethinking Approaches to the History of English*. ed. by Terttu Nevalainen and Elizabeth Closs Traugott, 859-872. New York: Oxford University Press.
- Los, Bettelou, and Erwin R. Komen. 2012. "Clefts as resolution strategies after the loss of a multifunctional first position". *Rethinking Approaches to the History of English*. ed. by Terttu Nevalainen and Elizabeth Closs Traugott. New York: Oxford University Press.
- Lozano, Cristobal. 2006. "Focus and split-intransitivity: the acquisition of word order alternations in non-native Spanish". *Second Language Research* 22:145-187.
- Lozano, Cristóbal, and Amaya Mendikoetxea. 2010. "Interface conditions on postverbal subjects: a corpus study of L2 English". *bilingualism: language and cognition* 13:475-497.
- Marcus, Mitchell, B. Santorini, and Mary Ann Marcinkiewicz. 1993. "Building a large annotated corpus of English: the Penn treebank". *Computational Linguistics* 19.

- Martin, Andrea E., Mante S. Nieuwland, and Manuel Carreiras. 2012. "Event-related brain potentials index cue-based retrieval interference during sentence comprehension". *NeuroImage* 59:1859-1869.
- Mathesius, Vilém. 1942. "Ze srovnávacích studií slovosledných [From comparative word-order studies]". *Časopis pro moderní filologii* 28:181-190, 302-307.
- Microsoft. 2006. Microsoft Brings Programming to the Masses With Visual Studio Express: Microsoft Corporation.
- Mikkelsen, Line. 2005. *Copular clauses specification, predication and equation*. Amsterdam; Philadelphia, PA: John Benjamins.
- Miller, Jim. 2006. "Focus in the languages of Europe". *Pragmatic organization of discourse in the languages of Europe*. ed. by Giuliano Bernini and Marcia L. Schwartz, 121-214. Berlin, New York: Mouton de Gruyter.
- Mitchell, B. 1985. *Old English Syntax, Vol. 1*. Oxford: Clarendon press.
- Mitkov, Ruslan, Richard Evans, Constantin Orasan, Le An Ha, and Viktor Pekar. 2007. "Anaphora Resolution: To What Extent Does It Help NLP Applications?". Paper presented at *Anaphora: Analysis, Algorithms and Applications*, Berlin.
- Molochieva, Zarina. 2010. *Tense, aspect, and mood in Chechen*. Ph. D. dissertation, University of Leipzig
- Müller, Christoph, and Michael Strube. 2001. "Annotating Anaphoric and Bridging Relations with MMAX". *Proceedings of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*. 45-50. Seattle.
- Müller, Christoph, and Michael Strube. 2006. "Multi-Level Annotation of Linguistic Data with MMAX2". *Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods*. English Corpus Linguistics, Vol.3, ed. by Sabine Braun, Kurt Kohn and Joybrato Mukherjee, 197-214. Frankfurt: Peter Lang.
- Neeleman, Ad, E. Titov, H. van de Koot, and R. Vermeulen. 2009. "A Syntactic Typology of Topic, Focus and Contrast". *Alternatives to Cartography* ed. by J. Van Craenenbroeck, 15-52. Amsterdam: Mouton de Gruyter.
- Nevalainen, Terttu, and Elizabeth Closs Traugott. 2012. *The Oxford handbook of the history of English*. New York: Oxford University Press.
- Nichols, Johanna. 1997. "Chechen Phonology". *Phonologies of Asia and Africa*. 2, ed. by Alan S. Kaye, 941-971. Winona Lake, USA: Eisenbrauns.
- Nichols, Johanna. 2007. *An all-ASCII Latin practical orthography for Ingush*, Available at: <http://linguistics.berkeley.edu/~ingush/orthography.html#Practical>.
- Nieuwland, Mante S., and Jos J. A. van Berkum. 2006. "When Peanuts Fall in Love: N400 Evidence for the Power of Discourse". *Journal of cognitive neuroscience* 18:1098-1111. Available at: <http://www.mitpressjournals.org/doi/pdf/10.1162/jocn.2006.18.7.1098>.
- Patten, Amanda. 2010. *Cleft sentences, construction grammar and grammaticalization*. Ph.D. thesis, The University of Edinburgh
- Pérez-Guerra, Javier. 1998. "Integrating right-dislocated constituents: a study on cleaving and extraposition in the recent history of the English language". *Folia Linguistica Historica* 19:7-26.
- Pintzuk, Susan. 1996. "Old English Verb-Complement Word Order and the Change from OV to VO". *York Papers in Linguistics*. 17, ed. by J. K. Local and A. R.

- Warner, 241-264. York: York University. Department of language and linguistic science.
- Pintzuk, Susan. 2002. "Verb-object order in Old English: Variation as grammatical competition". *Syntactic Effects of Morphological Change*. ed. by David Lightfoot, 276-299. Oxford: Oxford university press.
- Pintzuk, Susan, and Ann Taylor. 2006. "The loss of OV order in the history of English". *The Blackwell Handbook of the History of English*. ed. by Ans van Kemenade and Bettelou Los. Oxford: Blackwell.
- Prince, Ellen. 1978. "A comparison of *it*-clefts and WH-clefts in discourse". *Language* 54:883-906.
- Prince, Ellen. 1981. "Toward a taxonomy of given-new information". *Radical Pragmatics*. ed. by Peter Cole, 223-255. New York: Academic Press.
- Prince, Ellen. 1992. "The ZPG letter: subjects, definiteness and information-status". *Discourse description: diverse analyses of a fund raising text*. ed. by S. Thompson and W. Mann, 295-325. Philadelphia/Amsterdam: John Benjamins B.V.
- Prince, Ellen F. 1984. "Topicalization and Left dislocation: A functional analysis". *Discourses in reading and linguistics*. 433, ed. by S. J. White and V. Teller, 213-225.
- PROIEL, the project. 2011. *Guidelines for annotation of givenness*, Available at: http://folk.uio.no/daghaug/info_guidelines.pdf.
- Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.
- Randall, Beth, Ann Taylor, and Anthony Kroch. 2005. *CorpusSearch 2*, Available at: <http://corpussearch.sourceforge.net/credits.html>.
- Recasens, Marta, Lluís Màrquez, Emili Sapena, M. Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. 2010. "SemEval-2010 Task 1: coreference resolution in multiple languages". Paper presented at *Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010*, Uppsala, Sweden.
- Reeve, Clara. 1967. *The old English baron. Edited with an introduction by James Trainer*. London: Oxford University Press.
- Reinhart, Tanya. 1981. "Pragmatics and Linguistics: An analysis of sentence topics". *Philosophica* 27:53-94.
- Riester, Arndt, David Lorenz, and Nina Seemann. 2010. A recursive annotation scheme for referential information status. In *Proceedings of the seventh international conference of language resources and evaluation (LREC)*, 717-722. Valletta, Malta.
- Roberts, Ian G. 1985. "Agreement Parameters and the Development of English Modal Auxiliaries". *Natural Language and Linguistic Theory* 3:21-58.
- Rohde, Douglas L. T. 2005. *TGrep2 user manual*, Available at: <http://tedlab.mit.edu/~dr/Tgrep2/tgrep2.pdf>.
- Roos, Liset. 2012. *The use of and differences between full inversions and existential constructions with an initial prepositional phrase*. MA thesis, Radboud University Nijmegen

- Roosevelt, President. 1936. *1936 State of the Union Address*, Available at: <http://janda.org/politxts/State%20of%20Union%20Addresses/1934-1945%20Roosevelt/FDR36.html>.
- Rooth, Mats. 1992. "A theory of focus interpretation". *Natural Language Semantics* 1:75-116.
- Ross, John Robert. 1967. *Constraints on variables in syntax*. Ph.D. thesis, MIT; Department of modern languages
- Rumelhart, D., Paul Smolensky, J. McClelland, and G. Hinton. 1986. "Schemata and sequential thought processes in PDP models". *Parallel distributed processing - vol. 2 psychological and biological models*. ed. by James L. McClelland and David E. Rumelhart. Cambridge, Mass.: MIT Press, Available at: <http://cognet.mit.edu/library/books/view?isbn=0262631105>.
- Salzmann, Martin David. 2004. *Theoretical approaches to locative inversion*. MA thesis, University of Zurich
- Sasse, Hans-Jürgen. 1987. "The thetic/categorial distinction revisited". *Linguistics* 25:511-580.
- Sasse, Hans-Jürgen. 2006. "Theticity". *Pragmatic organization of discourse in the languages of Europe*. ed. by Giuliano Bernini and Marcia L. Schwartz, 255-308. Berlin, New York: Mouton de Gruyter.
- Saxon. 2009. *The Saxon XSLT and Xquery processor*: Saxonica limited, Available at: <http://www.saxonica.com>.
- Schachter, Paul. 1973. "Focus and relativization". *Language* 49:19-46.
- Selting, M. 1994. "Emphatic speech style - with special focus on the prosodic signalling of heightened emotive involvement in conversation". *Journal of pragmatics*. 22:375.
- Skeat, Walter William, and Abbot of Eynsham Aelfric. 1835-1912. *Ælfric's Lives of Saints, volume 2*: Early English Text Society.
- Soon, Wee Meng, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. "A machine learning approach to coreference resolution of noun phrases". *Computational Linguistics* 27:521-544.
- Sornicola, Rosanna. 2006. "Interaction of syntactic and pragmatic factors on basic word order in the languages of Europe". *Pragmatic organization of discourse in the languages of Europe*. ed. by Giuliano Bernini and Marcia L. Schwartz. Berlin, New York: Mouton de Gruyter.
- Sperberg-McQueen, C.M., and Lou Burnard. 2009. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*: TEI Consortium.
- Speyer, Augustin. 2010. "German Vorfeld-filling as constraint interaction". *Constraints in discourse*. ed. by Anton Benz and Peter Kühnlein, 267-290. Berlin/New York: John Benjamins.
- Stenetorp, Pontus, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. "brat: a Web-based Tool for NLP-Assisted Text Annotation". Paper presented at *Proceedings of the Demonstrations Session at EACL 2012*.
- Stenning, K. 1977. "Articles, quantifiers, and their encoding in textual comprehension". *Discourse processes: advances in research and theory*. Vol. 1. Discourse production and comprehension, ed. by R. O. Freedle. Norwood, NJ: Ablex.

- Stenning, K. 1978. "Anaphora as an approach to pragmatics". *Linguistic theory and psychological reality*. ed. by Morris Halle, Joan Bresnan and George A. Miller. Cambridge, Mass.: MIT Press.
- Stoop, Wessel. 2011. "CLD, dat is niet contrastief". *Tabu* 39:49-61.
- Strube, Michael, and Udo Hahn. 1999. "Functional centering: grounding referential coherence in information structure". *Computational Linguistics* 25:309-344.
- Taylor, Ann, Athony Warner, Susan Pintzuk, and Frank Beths. 2003. *The York-Toronto-Helsinki Parsed Corpus of Old English Prose*, Available at: <http://www-users.york.ac.uk/~lang22/YCOE/YcoeHome.htm>
- Tesnière, Lucien. 1959. *Éléments de syntaxe structurale*. Paris: C. Klincksieck.
- Thompson, Ellen. 1987. "Subordination and narrative event structure". *Coherence and grounding in discourse*. ed. by Russell Tomlin, 435-454. Amsterdam/Philadelphia: John Benjamins.
- Thompson, Ellen. 1999. "The temporal structure of discourse: the syntax and semantics of temporal then". *Natural Language and Linguistic Theory* 17:123-160.
- Tomlin, Russell S. 1985. "Foreground-background information and the syntax of subordination". *Text - Interdisciplinary Journal for the Study of Discourse* 5:85-122.
- Truckenbrodt, Hubert. 1995. *Phonological Phrases: Their Relation to Syntax, Focus and Prominence*. Cambridge: MIT Working Papers in Linguistics.
- Vallduví, Enric. 1990. *The informational component*. Ph.D., University of Pennsylvania
- van Berkum, Jos J. A., Arnout W. Koornneef, Marte Otten, and Mante S. Nieuwland. 2007. "Establishing reference in language comprehension: an electrophysiological perspective". *Brain Research* 1146:158-171.
- van Kemenade, Ans. 1987. *Syntactic case and morphological case in the history of English*. Dordrecht: Foris publications.
- van Kemenade, Ans. 1997. "V2 and embedded topicalisation in Old and Middle English". *Parameters of Morphosyntactic Change*. ed. by Ans van Kemenade and Nigel Vincent, 326-352. Cambridge: Cambridge University Press.
- van Kemenade, Ans. 1999. "Sentential negation and word order in Old English". *Negation in the history of English*. ed. by Ingrid Tieken-Boon van Ostade, Gunnel Tottie and Wim van der Wurff 147-166. Berlin: Mouton de Gruyter.
- van Kemenade, Ans. 2000. "Jespersen's cycle revisited: formal properties of grammaticalization". *Diachronic syntax*. ed. by Susan Pintzuk, George Tsoulas and Anthony Warner, 51-74. Oxford: Oxford university press.
- van Kemenade, Ans. 2002. "Word order in Old English Prose and Poetry: the position of finite verbs and adverbs". *Studies in the History of the English Language: A Millennial Perspective*. ed. by Donka Minkova and Robert Stockwell, 355-373. Berlin: Mouton de Gruyter.
- van Kemenade, Ans. 2009. "Discourse relations and word order change". *Information structure and language change: new approaches to word order changes in Germanic*. ed. by Roland Hinterhölzl and Svetlana Petrova, 91-120. Berlin: Mouton de Gruyter.
- van Kemenade, Ans. 2011. "Secondary negation and Information Structure organization in the History of English". *The evolution of negation: beyond the*

- Jespersen cycle*. ed. by Pierre Larrivee and Richard Ingham, 77-114. Berlin: Mouton de Gruyter.
- van Kemenade, Ans. 2012. "Rethinking the loss of V2". *The Oxford Handbook of the History of English*. ed. by Elizabeth Closs Traugott and Terttu Nevalainen. New York: Oxford University Press.
- van Kemenade, Ans, and Bettelou Los. 2006a. "Discourse adverbs and clausal syntax in Old and Middle English". *The handbook of the history of English*. ed. by Ans van Kemenade and Bettelou Los, 224-248. Malden/Oxford: Blackwell publishing.
- van Kemenade, Ans, and Bettelou Los. 2006b. *The handbook of the history of English*. Malden, MA; Oxford: Blackwell.
- van Kemenade, Ans, and Tanja Milicev. 2012. "Syntax and discourse in Old English and Middle English word order". *Grammatical Change: Origins, Nature, Outcomes*. ed. by Dianne Jonas, Andrew Garrett and John Whitman, 239-254. Oxford: Oxford University Press.
- van Kemenade, Ans, and Marit Westergaard. 2012. "Syntax and Information Structure: verb second variation in Middle English". *Information Structure and Syntactic Change in the History of English*. 1, ed. by Bettelou Los, María José López-Couso and Anneli Meurman-Solin, 87-118. New York: Oxford University Press.
- van Valin, Robert D. 1999. "A Typology of the Interaction of Focus Structure and Syntax". *Typology and the Theory of Language: From Description to Explanation*. ed. by E. Raxilina and J. Testelec. Moscow.
- van Valin, Robert D. 2005. *Exploring the syntax-semantics interface*: Cambridge University Press.
- Veeninga, Maaike, Sanne Kuijper, and Petra Hendriks. 2011. "Steunpronomen die komen overal voor". *Tabu* 39:111-130.
- Versley, Y., A. Moschitti, M. Poesio, and X. Yang. 2008. "Coreference Systems based on Kernel Methods". Paper presented at *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, Manchester.
- Vieira, Renata. 1999. "Co-reference resolution of definite descriptions". Paper presented at *Proceedings of VI Simposio Internacional de comunicación Social*.
- Virtanen, Tuija. 1992. "Given and new information in adverbials: clause-initial adverbials of time and space". *Journal of Pragmatics* 17:99-115.
- Virtanen, Tuija. 2004. "Point of departure: cognitive aspects of sentence-initial adverbials". *Approaches to cognition through text and discourse*. ed. by Tuija Virtanen, 79-97. Berlin/New York: Mouton de Gruyter.
- Visser, F. Th. 1963. *An historical syntax of the English language*. Leiden: Brill.
- Vuuren, Sanne van. 2012. personal communication.
- Ward, Gregory. 1985. *The semantics and pragmatics of preposing*. Ph.D. dissertation, University of Pennsylvania
- Ward, Gregory, Betty Birner, and Rodney Huddleston. 2002. "Information packaging". *The Cambridge grammar of the English Language*. ed. by Rodney Huddleston and Geoffrey K. Pullum. Cambridge: Cambridge University Press.
- Warner, Anthony. 2007. "Parameters of variation between verb-subject and subject-verb order in late Middle English". *English Language and Linguistics* 11:81-111.

- Wedgwood, Daniel, Gergely Pethő, and Ronnie Cann. 2006. Hungarian ‘focus position’ and English it-clefts: the semantic underspecification of ‘focus’ readings. Ms., Available at: <http://www.lel.ed.ac.uk/~dan/>.
- Weil, Henri. 1844. *Question de grammair générale: De l'ordre des mots dans les langues anciennes comparées aux langues modernes*. Paris: Imprimerie de Crapelet.
- Winkler, Susanne. 2005. *Ellipsis and focus in generative grammar*. Berlin; New York: Mouton de Gruyter.
- Yule, George. 1981. “New, current and displaced entity reference”. *Lingua* 55:41-52.
- Zacharsky, Ron, and Jim Cowie. 2011. *Chechen parallel bilingual and monolingual corpus*, Available at: <http://guidetodatamining.com/appendices/corpora/>.
- Zeldes, Amir, Julia Ritz, Anke Lüdeling, and Christian Chiarcos. 2009. “ANNIS: A Search Tool for Multi-Layer Annotated Corpora”. Paper presented at *Proceedings of Corpus Linguistics 2009*, Liverpool, UK.
- Zimmerman, Odo John, and Benedict Raymund Avery. 1980. *Life and Miracles of St. Benedict: book two of the dialogues*. Westport, Conn: Greenwood press.
- Zimmermann, Malte, and Edgar Onea. 2011. “Focus marking and focus interpretation”. *Lingua Lingua* 121:1651-1670.
- Zwaan, Rolf A. 2004. “The immersed experienter: toward an embodied theory of language comprehension”. *The psychology of learning and motivation* 44:35-62.
- Zwaan, Rolf A., and Gabriel A. Radvansky. 1998. “Situation models in language comprehension and memory”. *Psychological bulletin* 123:162-185.

Samenvatting (Summary in Dutch)

Wanneer er in een boek de volgende zin staat: “Piet had in zijn leven veel voertuigen bekeken. Een oranje auto had hij nog nooit gezien”, dan weet de lezer meteen dat de nadruk op *oranje auto* ligt. Als lezer kijk je niet raar op wanneer er iets op volgt als: “Een oranje fiets wel, maar dat was in het circus”. Als taalkundigen proberen we te ontdekken welke strategieën schrijvers volgen om ervoor te zorgen dat je als lezer feilloos weet waar de nadruk ligt.

Dit boek beschrijft een zoektocht naar de manier waarop nadruk in het geschreven Engels in de loop der tijd is weergegeven. Er wordt niet alleen gekeken naar de strategieën die schrijvers in de loop der tijd hebben gebruikt om nadruk over te brengen, maar ook of en hoe die methodes beïnvloed worden door of juist invloed uitoefenen op de veranderingen in de grammatica van het Engels. Het onderzoek begint met een stuk theorie (hoofdstuk 1-3), en daarna wordt de vraag naar de ontwikkeling van nadrukstrategieën op twee manieren aangepakt:

- 1) **Met de hand:** bekijk een vroege en een late tekst zin voor zin met de hand, bepaal hoe nadruk wordt weergegeven, en vergelijk die methodes (hoofdstuk 4).
- 2) **Automatisch:** bekijk zoveel mogelijk teksten uit de hele geschiedenis met de computer, en kijk hoe de methodes om nadruk weer te geven veranderen (hoofdstuk 5-12).

De theorie

Wat is nadruk eigenlijk, en waarom willen we in onze taal dingen benadrukken? Om die vragen te beantwoorden staat hieronder een klein stukje tekst (die lijkt op de Oud-Engelse tekst uit hoofdstuk 4).

- (340) a. In de dagen van Theodosius, de zoon van Arcadius, **woonde er in Alexandrië een vrome man genaamd Paphnutius.**
b. Deze man **had één dochter, die Eufrosina heette.**
c. Eufrosina **was een mooie jongedame.**
d. Ze **hield veel van haar vader,**
e. maar er was **één ding** dat ze nog leuker vond.
f. Wat voor haar boven alles uitging was **het klooster.**
g. Ze **ging iedere week trouw naar de kerk.**

Volgens een al eerder ontwikkelde theorie is er in iedere zin een gebied dat de nadruk heeft. In de tekst boven staan die gebieden vetgedrukt. Wat steeds de nadruk krijgt is de informatie die ons brein toe moet voegen aan het model dat we maken van het verhaal dat we lezen. In de eerste zin (340a) is het gezegde en het onderwerp nieuw: het feit dat er een man met de naam Paphnutius is, en dat die in Alexandrië woonde. Dat krijgt dan ook de nadruk. Een zin als deze wordt vaak gebruikt om een nieuw persoon in een verhaal te presenteren, en daarom noemen we dit ook wel

“presentatienadruk”. Het begin van de zin (“in de dagen van Theodosius”) is op zich niet zozeer nieuwe informatie, maar helpt ons om de rest van de zin in de tijd te plaatsen; het vormt een “uitgangspunt” door gebruik te maken van een verwijzing naar “Theodosius”, waarvan we meteen aanvoelen dat de schrijver veronderstelt dat we van hem gehoord hebben.

In (340b) weten we al wie *deze man* is, en is het hele gezegde *had één dochter* nieuw. Die nieuwe informatie wordt in ons brein verbonden met de mentale voorstelling die we van “Paphnutius” hebben. We maken ook een mentale voorstelling voor de “dochter” aan, en daar komt de naam “Eufrosina” aan te hangen, evenals de informatie uit (340c,d), dat ze mooi is, jong is, en veel van haar vader hield. Zinnen als (340b,c,d) vormen het hoofdbestanddeel van verhalen; hun nadrukmethode wordt “topic-comment” genoemd, omdat er over een persoon die al genoemd is (hier Eufrosina) een opmerking wordt gemaakt.

Net als in de eerste zin wordt er in (340e) ook iets nieuws gepresenteerd: *één ding*. Van dat éne *ding* weten we eerst niet zoveel. We ruimen er in ons brein al wel een plekje voor in, en aan die (lege) voorstelling hangen we al wel een eigenschap: “Eufrosina vindt dit leuker dan haar vader”. Maar pas de volgende zin (340f) vertelt ons wat die lege voorstelling in ons brein nu precies inhoudt: *het klooster*. De zin in (340f) heeft dan ook “zinsdeelnadruk”: het nadrukgebied is beperkt tot precies één zinsdeel.

De rest van het onderzoek richt zich op de verandering in strategieën voor het weergeven van de “presentatienadruk” en de “zinsdeelnadruk”.

Methode 1: Met de hand

De eerste methode die wordt gebruikt om nadrukstrategieën te onderzoeken is die waarbij twee teksten (een vroeg Engelse en een laat Engelse) vergeleken worden. Eerst wordt daarbij gekeken wat de “standaardvolgorde” van de zinsdelen per tekst zijn. De woordvolgorde in het vroege Engels lijkt veel op die van het Nederlands, terwijl het late Engels weer heel anders werkt. Per tekst wordt zin voor zin uitgeplozen wat het domein van de nadruk is (is dat: (a) gezegde plus onderwerp, (b) alleen het gezegde, of (c) één zinsdeel?). Ook wordt gekeken of er afwijkingen van de standaardwoordvolgorde zijn, en of die dan komen door iets dat te maken heeft met (i) de grammatica, (ii) de structuur van de tekst, of (iii) met de nadruk.

De meeste zinnen in de verhalen blijken domein (b) te gebruiken (die hebben dus een “topic-comment” structuur), terwijl we juist op zoek zijn naar strategieën voor het domein (a) “presentatienadruk”, en (c) “zinsdeelnadruk”. Toch komen er wat voorzichtige resultaten voor wat betreft de verandering in nadrukmethodes naar boven. Het vroege Engels gebruikt soms “zinsdeelsplitsing” wanneer er presentatienadruk is, terwijl het latere Engels meer van “appositie” gebruik maakt.¹ In de vroege en de late Engelse tekst wordt van een gekloofde zin (in het Engels “*it-cleft*” genaamd) gebruik gemaakt, maar alleen in de late Engelse tekst wordt zo’n zin gebruikt om “zinsdeelnadruk” weer te geven. Dat fenomeen wordt uitvoerig in de hoofdstukken 10-12 besproken.

Methode 2: Automatisch

Het is erg aantrekkelijk om bij het onderzoek naar de veranderende nadrukstrategieën automatisch (met behulp van een computerprogramma) te werk te gaan, omdat er dan in principe veel meer teksten bekeken zouden kunnen worden en er duidelijkere trends tevoorschijn komen. Maar kan het ook? Het antwoord is om verschillende redenen “Ja.”

Het kan als eerste, omdat andere onderzoekers teksten uit de periode 900-1900 na Christus in de computer hebben gezet, en dat nog wel met een degelijke woorden zinsontleding. De tweede reden waarom het zou moeten kunnen, zo wordt in dit boek betoogd, is dat het in principe mogelijk moet zijn om het domein van de nadruk (en daarmee dus of een zin “presentatienadruk” of “zinsdeelnadruk” heeft) automatisch te bepalen, maar alleen als we nog iets aan de teksten in de computer toevoegen. Wanneer we bij ieder zinsdeel in de tekst aangeven of het “nieuwe” informatie bevat, dan weten we vervolgens ook welke zinsdelen in ieder geval tot het domein van de nadruk behoren.² De automatische aanpak behelst dan ook, ruwweg gesproken, de volgende stappen:

(341) *De automatische aanpak*

- Voeg aan teksten toe hoe nieuw ieder zinsdeel is.
- Bepaal per zin het bereik van nadruk.
- Selecteer alle zinnen met “zinsdeelnadruk” en alle zinnen met “presentatienadruk”
- Kijk welke strategieën gebruikt worden voor die twee nadruksoorten.

In hoofdstuk 5 wordt nader uitgewerkt hoe de “nieuwheid” van zinsdelen aangeduid kan worden. Dat blijkt namelijk af te hangen van hoe we de zinsdelen met ons brein verwerken. Er worden vijf verwijzingscategorieën afgeleid die met behulp van de tekst in (340) en de beelden in Figure 47 uitgelegd kunnen worden.

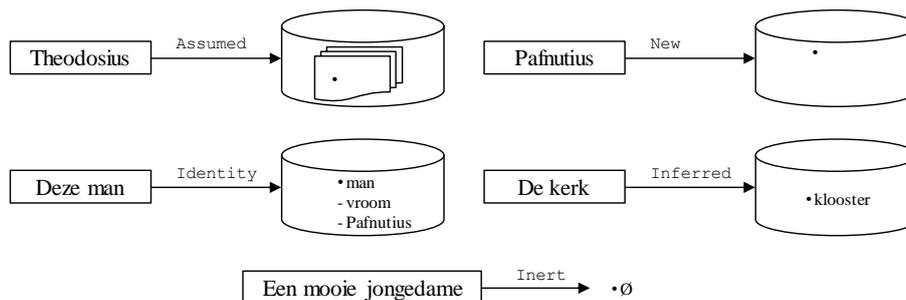


Figure 47 De vijf basismogelijkheden om te verwijzen

Het zinsdeel “Theodosius” uit (340a) heeft de verwijzingscategorie “Assumed”, omdat de schrijver er duidelijk van uit gaat dat de lezer ergens in zijn brein (in zijn lange termijn geheugen) al een plekje heeft ingeruimd voor de persoon Theodosius. De persoon “Paphnutius” is duidelijk nog niet bekend bij de lezer, en krijgt dan ook de verwijzingscategorie “New”. Wanneer (340b) verder gaat met iets te vertellen over Paphnutius wordt er met “Deze man” naar hem verwezen. Omdat Paphnutius

zelf inmiddels al een plekje heeft in onze mentale voorstelling van wat we lezen is de verwijzingscategorie van “Deze man” dan ook “Identity”. Later in het verhaal, in zin (340f), wordt “het klooster” geïntroduceerd (blijkbaar een klooster bij hen in de buurt), en “de kerk” in zin (340g) is een onderdeel van dat klooster, dus verwijst er in zekere zin naar. Maar de twee zijn niet identiek. Daarom krijgt “de kerk” de verwijzingscategorie “Inferred” (afgeleid) toegewezen: de “kerk” valt rechtstreeks uit het “klooster” af te leiden, omdat een klooster de aanwezigheid van een kerk impliceert. De laatste verwijzingscategorie is “Inert”, en die wordt gebruikt voor een zinsdeel als “een mooie jongedame” in zin (340c). Dat zinsdeel is een eigenschap van “Eufrosina”; het is geen specifiek nieuw persoon, verwijst niet terug naar iets dat al in het brein zit, en je kunt in volgende zinnen ook niet naar die “jongedame” terug verwijzen (wel naar Eufrosina, die de *eigenschap* heeft om een jongedame te zijn).

Volgens de automatische aanpak (de strategie die in 341 staat), is de eerste stap die van het toevoegen van verwijzingscategorieën aan ieder zinsdeel in de teksten die we hebben. We hebben het dan over ca. 350.000 zinnen, en hoofdstuk 6 van dit boek gaat dan ook over een programma (genaamd “Cesax”) wat probeert om de categorieën zoveel mogelijk automatisch aan te brengen. Helemaal automatisch gaat het niet (dat zou immers ook betekenen dat zelfs de referentiële categorieën af te leiden zijn van de beschikbare informatie uit de woord- en zinsontleding), maar het is wel mogelijk om onder leiding van Cesax “samen” een tekst door te wandelen en de categorieën aan te brengen. Sommige zinsdelen kan Cesax zelf verwerken, en bij andere zinsdelen doet Cesax een suggestie die je dan over kunt nemen of niet. Op deze manier zijn een paar teksten uit ieder van de vier deelperiodes van het Engels van verwijzingscategorieën voorzien. Cesax wil precies weten waar ieder zinsdeel naar terugverwijst, en slaat die informatie ook op. Daardoor ontstaan er als het ware “hyperlinkketens” in de teksten. Een voorbeeld van zo’n keten uit de tekst in (340) is die van “Eufrosina”, door middel van pijlen afgebeeld in (342).

(342) a. In de dagen van Theodosius, de zoon van Arcadius, **woonde er in Alexandrië een vrome man genaamd Paphnutius.**

b. Deze man **had één dochter, die Eufrosina heette.**

c. **Eufrosina was een mooie jongedame.**

d. **Ze** hield veel van **haar** vader,

e. maar er was **één ding** dat **ze** nog leuker vond.

f. Wat voor **haar** boven alles uitging was **het klooster.**

g. **Ze** ging iedere week trouw naar de kerk.

Het onderwerp “ze” in regel (342g) heeft als antecedent (het woord waar het naar verwijst) “haar” in (342e) en zo verder. De terugverwijzingen zijn een soort

“hyperlinks”, en de opeenvolgende antecedentenhyperlinks kunnen als een “hyperlinkketen” worden gezien.

Volgens stap (b) van de strategie in (341) zou het nadrukdomein automatisch bepaald moeten worden. Om dat mogelijk te maken is het computerprogramma “CorpusStudio” ontwikkeld. Met dat programma kun je als onderzoeker op zoek gaan naar zinnen of zinsdelen door informatie te geven over woordsoorten, zinsontleding en verwijzingscategorieën van die zinsdelen zelf en van de omliggende zinsdelen.

In hoofdstuk 8 komen we bij de stappen (c,d) van de strategie in (341), en gaan we op zoek naar zinnen met “presentatienadruk”. Het voornaamste kenmerk aan de hand waarvan dergelijke zinnen herkend kunnen worden is dat zij een onderwerp hebben met een verwijzingscategorie “New” (zie Figure 47). Er blijkt een lichte toename in het aantal zinnen met presentatienadruk te zijn waarbij het onderwerp slechts een korte hyperlinkketen genereert. Dat komt omdat het hedendaagse Engels het onderwerp veel meer dan vroeger moet gebruiken om voor de nodige samenhang in een tekst te zorgen (vroeger was dit één van de functies van het eerste zinsdeel). Wat verder nog blijkt uit het onderzoek naar presentatienadruk is de sterke afname van het aantal zinnen met een onderwerp dat zowel (a) vóór de persoonsvorm (het finiete werkwoord) staat, als (b) een referentiecategorie “New” heeft. Dat lijkt te komen door de opkomst van een alternatieve strategie, die met de expletieve *there* (bijvoorbeeld: *Once upon a time there was a young lady*). Bij deze strategie is het grammaticale onderwerp (het woordje *there*, wat overigens nergens naar terug verwijst) netjes vóór de persoonsvorm, terwijl het “logische” onderwerp met referentiecategorie “New” zich ná de persoonsvorm bevindt. Pogingen om presentatienadruk te vinden voor onderwerpen die referentieel *niet* nieuw zijn (iets wat gebeurt wanneer een persoon bijvoorbeeld op een onverwachte plek verschijnt) lopen op niets uit, omdat er gewoonweg te weinig teksten beschikbaar zijn die voorzien zijn van referentialiteitscategorieën en antecedentenhyperlinks.

Na ons op presentatienadruk te hebben gericht, gaan we in hoofdstuk 9 beginnen met het onderzoek naar “zinsdeelnadruk”. Dat hoort dus nog bij de stappen (c,d) van de strategie in (341). We hebben bij zinsdeelnadruk eigenlijk twee vragen: (1) “Zijn er één of meer plaatsen in de zin aan te wijzen waar zinsdelen met nadruk meestal geplaatst worden?” En (2) “Wat voor strategieën worden er (onafhankelijk van positie in de zin) gebruikt om zinsdeelnadruk te krijgen?” De manier om antwoord op vraag (1) te krijgen is door met vraag (2) te beginnen. Hoofdstuk 9 begint dan ook met het controleren van allerlei *kandidaten* voor zinsdeelnadruk. Een paar kandidaten blijken achteraf gezien niet (direct) met zinsdeelnadruk samen te hangen, anderen doen dat wel, maar zijn niet goed te meten met de teksten die we hebben, en dan blijft er een klein aantal kandidaten over die zowel zinsdeelnadruk impliceren als ook goed meetbaar zijn. Als eerste is daar de aanwezigheid van een bijwoord van contrast of nadruk (zoals bijvoorbeeld *only* ‘slechts’). En als tweede betrouwbare indicator is daar de aanwezigheid van contrast binnen een zinsdeel (bijvoorbeeld [*not John, but Mary*] *went to the cinema* ‘niet Jan, maar Marie is naar de film gegaan’). Door in te teksten te kijken waar constituenten met deze indicatoren in een zin voorkomen, krijgen we antwoord op vraag (1). Het onderzoek laat zien dat er

een groot verloop is van de voorkeurspositie voor zinsdeelnadruk. In het Oud Engels ligt de positie vóór het finiete werkwoord, terwijl deze na 1500 ná het finiete werkwoord ligt, en dan meest aan het einde van de zin. De verandering van deze voorkeurspositie hangt waarschijnlijk samen met de veranderingen in de Engelse grammatica: terwijl het onderwerp in het vroege Engels op twee verschillende plekken voor kon komen (als eerste zinsdeel of direct na de persoonsvorm), kan het in het late Engels, met een aantal duidelijk gedefinieerde uitzonderingen, alleen vóór de persoonsvorm voorkomen.

Er is één kandidaatsindicator voor zinsdeelnadruk die we niet in hoofdstuk 9 bekeken hebben, en dat is de gekloofde constructie (de *it*-cleft) die in hoofdstuk 4 voorbij kwam; hij kwam daar in de vroeg Engelse en de laat Engelse tekst voor, maar met een andere functie. Deze constructie is in het hedendaagse Engels niet meer weg te denken, en wordt door sommigen als dé zinsdeelnadruk strategie gezien. Daarom worden er drie hoofdstukken (10-12) van het onderzoek aan besteed. Als eerste wordt er in hoofdstuk 10 een degelijke definitie voor de constructie afgeleid, en wordt er gekeken wat andere onderzoekers gevonden hebben over het *doel* dat deze constructie dient. Daar blijken de meningen over uiteen te lopen. Duidelijk is in ieder geval wel dat de gekloofde zin niet alleen maar gebruikt wordt om zinsdeelnadruk weer te geven, maar ook om cruciale punten binnen de structuur van een tekst (het begin, het einde, een overgang) aan te duiden. Dat laatste blijkt in Scandinavische talen de boventoon te voeren. Nader onderzoek naar het Tsjetsjeens, beschreven in hoofdstuk 11, laat zien dat deze taal wel heel bijzonder is: zinsdeelnadruk wordt bereikt met woordvolgorde en niet met intonatie (er is geen aparte intonatie voor vraagzinnen of zinsdeelnadruk), en hoewel de taal een gekloofde zinsconstructie kent, wordt deze niet voor de nadruk gebruikt, maar alleen maar voor het aangeven van de structuur van de tekst. De hoofdstukken 10 en 11 vormen zo de opmaat voor 12, waarin de groei van de gekloofde zinsconstructie (die op zich al eens door eerder onderzoekers bekeken is) in een nieuw daglicht komt te vallen. In het vroege Engels werd deze constructie bijna alleen maar voor het aangeven van de tekststructuur gebruikt (de functie die het in het Tsjetsjeens nu heeft), maar met het verdwijnen van de speciale rol die het eerste zinsdeel voor zinsdeelnadruk vervult, blijkt de gekloofde constructie juist op te komen.

Wat heeft het opgeleverd?

In hoofdstuk 13 wordt teruggekeken op de resultaten die bereikt zijn in het onderzoek. Wat zijn er voor aanwijzingen gevonden over de samenhang tussen grammatica en nadruk? Als eerste is vanuit de gespleten zinsdeelconstructies in het Oud Engels gebleken dat de grammatica regels strijd voeren met het principe van de natuurlijke informatieordering (dat zegt dat relatief nieuwere informatie volgt op relatief bekendere informatie), en dit laatste principe is iets dat gemeten kan worden met de verwijzingscategorieën die zinsdelen in principe met zich meedragen. Wanneer dominante elementen die eigenlijk in de zinskern horen te blijven erbuiten geplaatst worden om ze meer nadruk te geven blijkt syntaxis overstemt te kunnen worden door nadruk. Een ander merkwaardig fenomeen is dat de grammaticale

analyse van eenzelfde zin af blijkt te kunnen hangen van de verwijzingscategorieën van de onderdelen van die zin. Het theoretische model voor een grammatica van een taal als het Engels moet dus een wisselwerking tussen syntactische regels en verwijzingscategorieën mogelijk maken. In dit laatste hoofdstuk wordt teruggeblikt op de aanname dat het wellicht mogelijk zou zijn om de nadrukdomeinen te bepalen met behulp van (a) de grammaticale analyse van een zin, (b) verwijzingscategorieën van zinsdelen, en (c) antecedenthyperlinks. Een eerste poging voor één soort zinnen in hoofdstuk 5 laat zien dat deze methode werkt, en dat leidt tot de hypothese dat het begrip “nadruk” wellicht helemaal niet zo’n atomair (of ondeelbaar) begrip is als je zou denken. Het zou heel goed mogelijk zijn dat juist de verwijzingscategorieën tot de fundamentele “deeltjes” van de taalkunde behoren. Om die hypothese te onderzoeken is echter meer fundamenteel en experimenteel werk nodig.

Hoofdstuk 13 bevat aanwijzingen voor vervolgonderzoek. Om de statistische significantie van de resultaten op het gebied van de veranderingen in de presentatie- en zinsdeelnadrukstrategieën te verhogen is het nodig om veel meer Engelse (en wellicht ook anderstalige) teksten te verrijken met verwijzingscategorieën en antecedenthyperlinks. Dit is tevens van belang om verdergaande experimentele verificatie mogelijk te maken van de mogelijkheid dat verwijzingscategorieën fundamentele taalkundige deeltjes zijn.

¹ Een voorbeeld van zinsdeelsplitsing is: “Er woonde **een man** in Alexandrië **genaamd Paphnutius**”. De vetgedrukte woorden vormen samen één zinsdeel, maar het is in twee delen gesplitst. Een voorbeeld van appositie is bijvoorbeeld: “De man had een dochter, **een vrome jonge vrouw, Eufrosina geheten**, die vaak naar het klooster ging”. De vetgedrukte zinsdelen staan in appositie tot het onderwerp “een dochter”; ze geven er in een soort opsomming een nadere omschrijving van.

² Wanneer de informatie in een zinsdeel *niet* “nieuw” is met betrekking tot het model van het verhaal dat we in ons brein opbouwen, dan weten we niet of het tot het domein van nadruk behoort of niet; beide mogelijkheden liggen nog open.

Part VI

Appendix

14 Appendix

14.1 Working with CorpusStudio

The program CorpusStudio is a stand-alone Windows program which I have written in the computer language called “Visual Basic .Net” (Microsoft, 2006).¹ The program itself and the reference manual are available on the internet (Komen, 2011a). A screenshot of a typical corpus research project in CorpusStudio is shown in Figure 48.

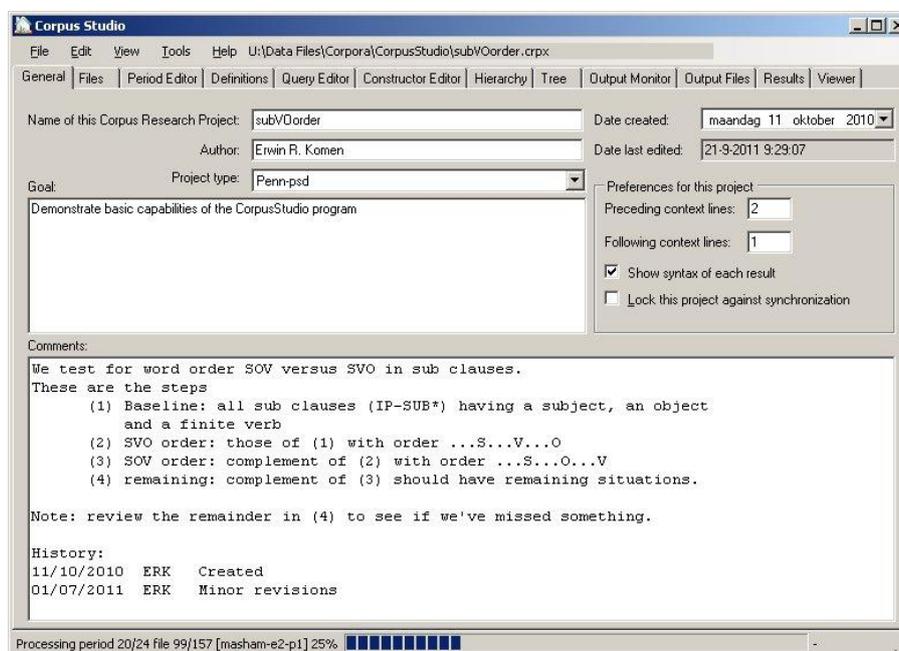


Figure 48 Main definitions of a corpus research project in CorpusStudio

The main functions of CorpusStudio are: (a) to group queries and meta information into corpus research projects (14.1.1), (b) allow a windows-interface for users to define queries and quickly locate user-defined functions (14.1.2), (c) allow users to define the order of processing queries (14.1.3), and (d) provide numerical results in table-form as well as add a user-definable context to the result lines (14.1.4).

14.1.1 Corpus research projects

A “Corpus research project file” is an *xml* file containing all information for one particular corpus research project. The “General” tab page in CorpusStudio allows defining meta information for a corpus research project, such as: its creation date,

the name of the author, the overall goal of the project, and comments. These comments could, for instance, outline the procedure followed in the project in more detail. Figure 48 shows the meta information of a project where we determine the word order (SVO versus SOV) in subclauses in all the parsed English corpora.

This particular project is of type “Penn-psd”, which means that it provides a wrapper around CorpusSearch2. The general project information says that the syntactic break-down of each line found in the output will be shown, and each line will be accompanied by two preceding and one following line. Subsequent tab pages allow for defining all necessary components of a corpus research project:

- (343) Files is used to specify the location of input and output files;
Period Editor allows dividing the input files in groups (according to time periods or to other criteria);
Definitions facilitates maintaining a common set of variables and functions;
Query Editor is the place to define all the queries needed for the project;
Constructor Editor is where the execution order of the queries is defined;
Hierarchy and Tree visualize the execution order of queries;
Output Monitor is where we can keep track of CorpusStudio processing the queries;
Results is the tab page where all the results of all the queries is shown, and where we can leaf through them.

An extensive discussion of the information that can be stored at each tab page is available in the CorpusStudio user’s manual (Komen, 2009b).

14.1.2 Defining queries

The query editor allows defining queries either for CorpusSearch2 or for Xquery. It does not only allow the researcher to define the text of the query, but also meta-information such as creation date, the query’s main goal, and any other useful comments. The queries are part of the corpus research project file they are included in, but a backup copy of each query is maintained in a user-definable directory. This copy can then be re-used in other research projects.

It is often helpful to use pre-defined shortcuts that, for instance, define which constituent labels should be regarded as those of a finite verb (finite verbs could consist of past tense verbs labelled `VBD`, present tense verbs `VBP`, past tense forms of *be*, which are labelled `BED`, present tense forms of *be* labelled `BEP` etc.) An entry in a definition file could state that the shortcut “`finiteverb`” means “`VBP|VBD|BED|BEP`”.²

Corpus research projects that work with Xquery will often make use of the same user-definable functions. These functions can be stored in the same file that contains the definitions of global variables. An example of a user-defined function is in (344).

(344) *A user-defined function that is part of a definitions file*

```

1  (: -----
2  Name : tb:SomeChildNo
3  Goal : Return the first child node of [$this]
4         having a label like $strLabel
5         and NOT having a label like $strNogo
6  History:
7  24-02-2010      ERK      Created from "SomeChild"
8  ----- :)
9  declare function tb:SomeChildNo($this as node()?, $strLabel as xs:string?,
10                                $strNogo as xs:string?) as node()?
11 { (: Get ALL the children of me :)
12   let $all := $this/child::eTree
13   (: Select those that have the indicated label :)
14   let $ok1 := $all[ru:matches(@Label, $strLabel)]
15   (: Exclude those that have the NoGo label :)
16   let $ok := $ok1[not(ru:matches(@Label, $strNogo))]
17   return
18     if (empty($ok))
19       then ()
20     else $ok[1]
21 } ;

```

Without going into a detailed description of the Xquery program language (for which see: Boag et al., 2010), there are a few things about the `tb:SomeChildNo()` function that should be mentioned here. Lines 1-8 are comment lines (everything between `(:` and `:)` is ignored by the software), and they tell us when the function was created (in “History”) and what the goal of the function is: provide the first child node of the argument `$this` that complies with two stipulations: (a) its label (which defines its syntactic category and, sometimes, its function) must match the pattern in `$strLabel`, and (b) this label may *not* match the pattern in `$strNogo`. Line 12 is the first real line. The variable `$all` gets all the child nodes of `$this`, provided these children have the `<eTree>` tag. Line 14 assigns those constituents in `$all` to the variable `$ok1`, for whom the attribute `@Label` matches the pattern in `$strLabel` (for instance that of a subject: “NP-NOM*|NP-SBJ”).³ Line 16 takes the nodes stored in `$ok1`, and takes away those that have a label matching `$strNogo` (such as non-subject NPs: “*PRD*|*LFD*|*VOC*”), storing the result in `$ok`. Lines 17-20 determine what is returned: if no nodes are left in `$ok`, then an empty node is returned, otherwise the first result in the sequence of nodes kept in `$ok` is returned. In sum, the function returns a child node that matches one pattern, while at the same time not matching another pattern.

14.1.3 Combining queries

Some corpus research projects require a number of queries to be executed in sequence. If we are looking, for instance, at the increase of the SVO (subject, finite verb, object) word order in complement clauses in English, an initial simple approach would be to have two queries: (a) one that finds all complement clauses containing a subject, a finite verb and an object in any order, and (b) one that finds complement clauses with the subject preceding the finite verb and the object in SVO order.⁴ The first query provides what I will call the “baseline”: the number of clauses that contain the basic ingredients of S, V and O. If we want to know the percentages

of clauses with a particular order of S, V and O (such as SOV order and SVO order), then we need to divide the number of clauses with this *particular* order by the baseline (the number of clauses with *all* orders). To obtain the percentage of SVO subclauses, we need to divide the numbers found in the second query by the baseline numbers in the first query.

In the example above, it is most efficient (in terms of the time needed to execute the queries) if we restrict the input of the second query (the one that detects the SVO word order) to the instances we found in the first query (the one that gives all complement clauses with S, V and O in any order). Texts will, in general, have to pass through queries in a particular order, and as stated earlier, command-line oriented corpus research programs such as Tgrep2 (Rohde, 2005) and CorpusSearch2 (Randall et al., 2005) require the linguist to use a batch file or to run the queries manually, one by one. Batch files require advanced computer skills, but the manual approach is error-prone. The program CorpusStudio offers a user-friendly alternative to the batch-file approach. The program allows defining the order in which queries are executed through the “Constructor editor”, which basically is a table where each row specifies the query to be executed as well as the input to that query. This input can be either the whole “source” (all input files selected in the “Files” tab page) or the output of a previous line in the constructor table. An example of such a constructor table is provided in Table 48.⁵

Table 48 A query execution table defined in the constructor editor

Line	Input	Query	Output	Result	Cmp	Goal
1	Source	subS+V+O	subS+V+O	subS+V+O	-	Get subclauses containing a subject, object and verb
2	1/out	subS-O-V	subS-O-V	subS-O-V	+	Get subclauses containing S, O and V in that order
3	2/cmp	subS-V-O	subS-V-O	subS-V-O	+	Get subclauses containing S, V and O in that order
4	3/cmp	subS+V+O	Remainder	Remainder	-	Get subclauses containing a subject, object and verb, but not in SVO or SOV order

The table that defines which queries should be executed starts in line 1 with the query `subS+V+O`. This line takes its input from the “source” (all the input files defined on the “Files” tab page). The second line takes the *output* of the first line (specified as `1/out`) as its input, and executes the query called `subS-O-V`, and it gets all the subclauses where the subject precedes the object NP, which, in turn, precedes the finite verb. The third line executes the query `subS-V-O` and takes its input from the *complement* of the second line. The complement contains all the `<forest>` elements that do not satisfy the conditions specified in the query, so it should contain sentences that do not have a subclause with the SOV word order.⁶ The Constructor Editor, then, allows one to define queries hierarchically.

Once the queries of a research project and the order in which queries have to be executed have been specified, query execution can take place. CorpusStudio optimizes query execution in a number of ways, and one of these ways is the *order*

in which files, periods and sentences are treated. The execution order used by CorpusStudio is shown in (345).

- (345) *Query execution order*
- a. Period (such as *Old English*, *Middle English*, or parts of these)
 - b. Text
 - c. Sentence (Xquery only)⁷
 - d. Query line (as defined in the constructor editor)

Execution starts in (345a) by taking the different periods (or genres) specified in the Period Editor into consideration: texts belonging to one period are executed one-after-another. While the CorpusStudio wrapper program loads a text into its memory, as in (345b), it walks through the text internally sentence-by-sentence (one sentence is one `<forest>` element in the *psdx* corpora), as in (345c). The query lines defined in the table at the constructor editor tab page are now, as in (345d), executed one-by-one. If the current sentence yields a match for the query defined in the first line of the constructor editor, the sentence will serve as input to the queries in those lines of the constructor editor that have “1/out” defined as their input. But if a sentence has not even passed through the query in the first line, no processing is needed for subsequent lines. This means a reduction in necessary processing: not all queries need to be fed with all the sentences in the texts. An additional advantage of the query order in (345) is that if there is an error in one of the queries (even the last one), it could surface as soon as the first sentence is processed. With complex corpus research projects sometimes taking several hours of processing, this a useful feature indeed.

14.1.4 Research project results

The results of a corpus research project are presented in several different ways. One form is that of an *html* file, which is created separately, and shown from within CorpusStudio on the Results tab page. The results presented in this way start with meta-information, such as the name of the corpus research project, the date and time of execution and the username of the researcher.

The results then come with a table offering a summary of the output of the different queries, which is broken up in the subperiods that have been specified in the “Period Editor” (see the [manual](#)), as in Table 49.⁸ Old English, for instance, is divided into sub periods O1-O4.

Table 49 Table with results provided by CorpusStudio

Description	O1	O2	O3	O4	M1	M2	M3	M4	E1	E2
subS+V+O	17	2700	3343	17	1835	524	2656	1645	3952	4467
subS-O-V	8	1519	1763	3	419	120	20	3	8	9
subS-V-O	2	691	1226	5	1255	350	2581	1617	3867	4400
Remainder	4	285	207	9	69	15	39	22	52	48
IP-MAT	83	20315	50201	365	15964	7347	26318	19839	28194	34614
IP-SUB	112	23857	37789	181	18309	5122	23678	12915	32887	35501

The numbers in each cell (such as “17” in column “O1” of row “subS+V+O”) represent the number of instances that have successfully passed the query of that particular line. Normalization of frequencies is left to the user, who can decide either to define a query that supplies a baseline, or use the numbers of main clauses (“IP-MAT”) and subclauses (“IP-SUB”) that are supplied by CorpusStudio.

The results do not only give a summary table, but they also allow “jumping” to individual examples by clicking on a cell in the table. If we were to click, for instance on the number “9” which indicates the number of sentences with SOV word order in subclauses from the period E2 (the second part of early Modern English, containing texts from 1570-1639), we end up on the part of the results page where we find the 9 instances that have been found. The individual results come with a preceding and following context (as defined for the current project), and they can be set to come with a syntactic breakdown of the relevant sentence. The example in (346), for instance, shows the sentence “but I know that God the maker hit guides” as an example of the subSOV query from the E2 time period.

(346) *Output for one hit provided by CorpusStudio*

[boethel-e2-p1] [17.132] or dost suppose that Reasons rule is in it?”
 [17.133] “I can no way think,” quoth I, “that with so rash chaunce, so certain things are moued,
[17.134] but I know that God y=e= maker hit guides,
 [17.135] nor euer shall com day that from truth of this opinion shall draw me.”

[IP-SUB [NP-SBJ God y=e= maker] [NP-OB1 hit] [VBP guides]]

The syntactic breakdown of (346) shows that it is a subordinate clause (indicated by “IP-SUB”) containing a subject (the “NP-SBJ”), a direct object (indicated by “NP-OB1”) and a finite verb in the present tense (“VBP”).

14.2 A selection of queries

This section contains the code of a number of key queries that are referred to in this book but that were not taken up with the text (some queries are provided in the text where they are discussed). The queries make use of definition files containing global variables and user-defined Xquery functions. These definition files can all be found at the author's website: <http://erwinkomen.ruhosting.nl/software/CorpusStudio>.

14.2.1 Copula clauses

The query used to get the examples for the copula clauses in section 5.5.3.1 is the following:

```

1  for $search in //eTree[ru:matches(@Label, $_matrixIP)]
2  (: The central element is a finite form of "be" :)
3  let $be := tb:SomeChild($search, $_finite_BE)
4
5  (: There has to be an "XP" of type NP, PP or ADVF preceding BE :)
6  let $first := $be/preceding-sibling::eTree[1]
7  [ru:matches(@Label, 'NP|NP-*|PP*|ADVP*|ADJ*|VAN*')]
8  let $XPSyntax := if (ru:matches($first/@Label, 'NP*')) then 'NP'
9  else if (ru:matches($first/@Label, 'PP*'))
10 then 'PP' else 'AP'
11 let $XPpenta := if ($XPSyntax = 'AP') then '-'
12 else if ($XPSyntax = 'PP') then
13 ru:feature($first/child::eTree[
14 ru:matches(@Label, 'NP*')], 'RefType')
15 else ru:feature($first, 'RefType')
16
17 (: There has to be an "YP" of type NP, PP or ADVF preceding BE :)
18 let $last := $be/following-sibling::eTree[1]
19 [ru:matches(@Label, 'NP|NP-*|PP*|ADVP*|ADJ*|VAN*')]
20 let $YPSyntax := if (ru:matches($last/@Label, 'NP*')) then 'NP'
21 else if (ru:matches($last/@Label, 'PP*')) then 'PP'
22 else 'AP'
23 let $YPpenta := if ($YPSyntax = 'AP') then '-'
24 else if ($YPSyntax = 'PP') then
25 ru:feature($last/child::eTree[
26 ru:matches(@Label, 'NP*')], 'RefType')
27 else ru:feature($last, 'RefType')
28
29 (: Determine the category for subcategorization :)
30 let $cat := concat('XP=', $XPSyntax, $XPpenta, '_YP=',
31 $YPSyntax, $YPpenta)
32
33 (: The verb, XP and YP should exist,
34 the referential category should be known :)
35 where (
36 exists($be) and exists($first) and exists($last)
37 and not($XPpenta = '')
38 and not($YPpenta = '')
39 and not(ru:matches($XPpenta, 'CrossSpeech|NewVar'))
40 and not(ru:matches($YPpenta, 'CrossSpeech|NewVar'))
41 )
42 return ru:back($search, '', $cat)

```

14.2.2 Presentational focus

The query called “any_SbjIntro”, which is from the corpus research project `SbjPosition_V2` that is used to get sentences with presentational focus, is provided here:

```

1   for $search in //eTree[ru:matches(@Label, $_anynp)]
2   (: Get my IP and get the finite verb :)
3   let $ip := $search/ancestor::eTree[ru:matches(@Label, 'IP*')][1]
4   let $vfin := $ip/child::eTree[ru:matches(@Label, $_finiteverb)]
5
6   (: See if the found NP is a subject :)
7   let $IsSbj := (ru:feature($search, 'GrRole') = 'Subject')
8
9   (: The subject node must be referentially "new" :)
10  let $IsNew := ru:isnew($search, $_newroot)
11
12  (: This next in the chain must point to me with identity :)
13  let $al := ru:chnextidt($search)
14
15  (: Get the length of the following chain :)
16  let $al_len := ru:chlen($search, 'following')
17
18  (: Determine the sentence category :)
19  let $cat := concat(tb:ChLenType($al_len), '_',
20                  tb:SbjSentType($search))
21
22  (: We only allow new subjects in finite clauses with a finite verb :)
23  where ( $IsNew
24          and $IsSbj
25          and exists($vfin)
26          and tb:IsInFinite($search)
27          )
28
29  return ru:back($ip, '', $cat)

```

The query called “matSbjIntro_expl”, which is from the corpus research project `SbjPosition_V2` that looks for presentational focus, is provided here:

```

1   for $search in //eTree[tb:IsMain(self::eTree)]
2   (: There must be a subject and an object/complement :)
3   let $sobj := tb:SomeChildNo($search, $_subject, $_nosubject)
4   let $obj := tb:SomeChild($search, $_objCompl)
5
6   (: The complement node must be referentially "new" :)
7   let $IsNew := (ru:isnew($obj, 50, $_newroot)
8                 or (ru:feature($obj, 'RefType') = 'Inert'))
9
10  (: There must be a finite verb as well as a form of "be" :)
11  let $vfin := tb:SomeChild($search, $_finiteverb)
12  let $be := tb:SomeChild($search, $_any_BE)
13
14  (: Subcategorize on subject position :)
15  let $cat := tb:SbjSentType($obj)
16
17  (: There must be an appropriate subject, complement and a finite verb :)
18  where (
19    tb:IsExpl($sobj) and exists($obj) and $IsNew
20    and not(tb:IsStarred($obj))
21    and exists($vfin)
22  )
23  return ru:back($search, '', $cat)

```

The query called “matSbjIntro_unanch”, which is from the corpus research project `SbjPosition_V2` that looks for presentational focus, is provided here:

```

1   for $search in //eTree[tb:IsMain(self::eTree)]
2   (: There must be a subject or an expletive + complement :)
3   let $sjbcand := tb:SomeChildNo($search, $_subject, $_nosubject)
4   let $sobj := if (exists($sjbcand)) then $sjbcand
5               else ( let $expl :=
6                       $search/child::eTree[tb:IsAnyExpl(self::eTree)]
7                       let $compl := tb:SomeChild($search, $_objCompl)
8                       return if (exists($expl) and exists($compl))
9                             then $compl else () )
10  let $stype := if ($sobj/@Id = $sjbcand/@Id) then '' else '_expl'
11  (: The subject node must be referentially "new" and unanchored :)
12  let $isnew := ru:isnew($sobj)
13
14  (: There must be a finite verb :)
15  let $vfin := tb:SomeChild($search, $_finiteverb)
16
17  (: Subcategorize on subject pos with respect to the finite verb :)
18  let $scat := concat( tb:SbjSentType($sobj), $stype)
19
20  (: There must be an appropriate subject and a finite verb :)
21  where (
22    exists($sobj) and $isnew and not(tb:HasAnchor($sobj))
23    and not(tb:IsStarred($sobj))
24    and not(ru:feature($sobj, 'NPtype') = 'QuantNP')
25    and exists($vfin)
26  )
27  return ru:back($search, '', $scat)
28

```

14.2.3 Focus adverb constituent position

The query called “S+V+AdvContr”, which is from the corpus research project “FocusAdvNonCesax-Xquery_V6” that looks for constituent focus provided by focus adverbs, is provided here:

```

1   for $search in //eTree[tb:HasLabel(@Label, $_matrixIP)]
2   let $sobj := tb:SomeChildNo($search, $_subject, $_nosubject)
3   let $vbj := tb:SomeChild($search, $_finiteverb)
4   let $obj := tb:AllChildren($search, 'PP*|NP*')
5   let $fp := tb:GetFP($obj, $_IsFocAdv)
6   let $fpn := tb:PPobjectOrNP($fp)
7
8   where ( exists($sobj) and
9           not(tb:IsStarred($sobj)) and
10          not(tb:Coref($sobj, 'Inert|NewVar')) and
11          exists($fpn) and
12          exists($vbj)
13        )
14  return ru:back($search, tb:NewInfo($fpn), tb:FinVerbLoc($fp))

```

The query looking for adverbs expressing emphatic prominence only differs in the variable that defines the adverb type in line #5.

The query called “finHaveS” from corpus research project “HaveOrder” looks for clauses containing “have” as main verb:

```

1   for $subj in //eTree[ru:matches(@Label, $_anynp)]
2
3   (: Check if this is a subject :)
4   let $subjOk := (   (ru:feature($subj, 'GrRole') = 'Subject') and
5                     not(ru:matches(
6                       ru:feature($subj, 'NPtype'), 'ZeroSbj|Trace'))
7                     )
8
9   (: Find clause :)
10  let $ip := $subj/parent::eTree[ru:matches(@Label, $_finiteIP)]
11  let $cpL := $ip/ancestor::eTree[ru:matches(@Label, 'CP*')][1]/@Label
12
13  (: Find out if we have a finite "have" verb :)
14  let $vb := tb:SomeChild($ip, '*HVI|*HVP|*HVD*')
15
16  (: Find out if we have a participle :)
17  let $ptc := tb:SomeChild($ip, $_nonfiniteverb)
18
19  (: Find position of subject with respect to verb :)
20  let $pos := tb:SbjSentType($subj)
21
22  (: Include non-CP sentences with a finite verb,
23     a good subject and without a participle :)
24  where (
25    exists($vb) and $subjOk and not(exists($ptc)) and
26    not(ru:matches($cpL, 'CP-QUE*|CP-ADV*|CP-REL*'))
27  )
28  return ru:back($ip, '', $pos)

```

14.2.4 Local contrast

The query called “`cfeNP_Local`”, which is from the corpus research project “`ConstFocus_xq_v1`” that looks for noun phrases with local contrast, is provided here:

```

1   for $search in //eTree[ru:matches(@Label, $_anynp)]
2   (: Get the NP's label :)
3   let $np := $search/@Label
4
5   (: Find the FIRST negator available as descendant :)
6   let $neg := $search/descendant::eTree[ru:matches(@Label, $_neg)][1]
7
8   (: Check if the NP fulfils the local-contrast condition :)
9   let $npOk := some $ch in $search/child::eTree satisfies
10      ( ru:matches($ch/@Label, 'CONJP*') and
11        (some $grch in $ch/child::eTree satisfies
12          (ru:matches($grch/@Label, 'CONJ') and
13            ru:matches($grch/child::eLeaf/@Text, $_contrast)))
14      )
15
16   (: Find the parent IP of this NP :)
17   let $ip := $search/parent::eTree[ru:matches(@Label, $_finiteIP)]
18   let $ipOk := (exists($ip) and not(ru:matches($ip/@Label, '*=[1234]'))))
19
20   (: Find verb and subject :)
21   let $vb := tb:SomeChild($ip, $_finiteverb)
22   let $sbj := tb:SomeChildNo($ip, $_subject, $_nosubject)
23   let $sbjOk := ( not(ru:matches(
24     ru:feature($sbj, 'NPtype'), 'ZeroSbj|Trace')) )
25
26   (: Define a message :)
27   let $msg := concat('NP=', tb:Labelled($search))
28
29   (: Define subcategorisation :)
30   let $cat := tb:SbjSentType($search)
31
32   (: Our examples MUST have a negator and a finite IP :)
33   where (
34     not(ru:matches($np, '*PRN*')) and
35     exists($neg) and
36     $ipOk and
37     exists($vb) and
38     $sbjOk and $npOk and $msg
39   )
40   return ru:back($ip, $msg, $cat)

```

14.2.5 Contrastive left dislocation

The query called “anyCLD_dem”, which is from the corpus research project “CLD_xq_V1” that looks for noun phrases with local contrast, is provided here:

```

1   for $search in //eTree[ru:matches(@Label, $_anyLFD)]
2   (: Find the finite clause we are part of :)
3   let $ip := $search/parent::eTree[ru:matches(@Label, $_finiteIP)]
4
5   (: Find the resumptive NP or PP :)
6   let $rsp := $ip/child::eTree[ru:matches(@Label, 'NP*RSP*|PP*RSP*')]
7
8   (: Get the NP type of this resumptive :)
9   let $npt := ru:feature($rsp, 'NPtype')
10
11  (: Get basic ingredients of the clause :)
12  let $subj := $ip/child::eTree[
13      (ru:feature(self::eTree, 'GrRole') = 'Subject')
14      and not(ru:matches
15          (ru:feature(self::eTree, 'NPtype'), 'ZeroSbj|Trace'))]
16  let $vb := tb:SomeChild($ip, $_finiteverb)
17
18  (: Subcategorize on the position of the resumptive :)
19  let $cat := concat(tb:SbjSentType($rsp), '_',
20                  ru:feature($rsp, 'GrRole'))
21
22  (: Only accept normal sentences :)
23  where (
24      exists($subj) and exists($vb) and exists($rsp)
25      and ($npt = 'Dem')
26  )
27  return ru:back($rsp, '', $cat)

```

14.2.6 Occurrence of *wh*-clefts

The query called “`matWHcleft`”, which is from the corpus research project “`WhCleft_xq_V2`” that looks for *wh*-cleft instances, is provided here:

```

1   for $search in //eTree[ru:matches(@Label, $_frlCP)]
2   (: Get the NP around the free relative :)
3   let $par := $search/parent::eTree
4   let $npWh := if ($par[tb:Like(@Label, $_anynp)]) then $par
5               else if ($par[tb:Like(@Label, 'CONJ*')])
6                   then $par/parent::eTree[tb:Like(@Label, $_anynp)]
7               else ()
8
9   (: Get the clause in which this NP is :)
10  let $ip := $npWh/parent::eTree[tb:Like(@Label, $_finiteIP)]
11  let $cp := $ip/ancestor::eTree[tb:Like(@Label, 'CP*')][1]
12  let $cpOk := not(tb:Like($cp/@Label, 'CP-QUE*|CP-ADV*|CP-REL*'))
13
14  (: Get the finite verb :)
15  let $vbFin := tb:SomeChild($ip, $_finite_BE)
16
17  (: Check for non-finite verbs :)
18  let $vbNon := tb:SomeChild($ip, $_nonfiniteverb)
19
20  (: Check for a complement NP not equal to the first NP :)
21  let $compl := $ip/child::eTree[tb:Like(@Label, 'NP*OB*|NP*PRD*')
22              and not(@Id = $npWh/@Id)]
23
24  where (
25    $cpOk
26    and ru:feature($npWh, 'GrRole') = 'Subject'
27    and not(exists($vbNon))
28    and exists($compl)
29    and exists($vbFin)
30  )
31  return ru:back($ip)

```

14.3 Statistics of tables and figures

This appendix provides the p-values of the two-tailed Fisher's exact test for all relevant tables and figures in this book. Each table here in this appendix belongs to one table or figure in the book, and each row provides the details for one period-transition: the two periods (labelled P1 and P2), the number of occurrences for each period (labelled P1# and P2#), the rest numbers (labelled P1-rest# and P2-rest#), and then the p-value.

14.3.1 The decline of subject-finite-verb inversion in main clauses

See section 1.2.2.3, Figure 1.

	P1	P2	P1#	P1-rest#	P2#	P2-rest#	p-value
matInv	OE	ME	1702	1238	2353	4288	0,0000
matInv	ME	eModE	2353	4288	1154	6592	0,0000
matInv	eModE	LmodE	1154	6592	215	3580	0,0000
matObj_Inv	OE	ME	96	89	188	141	0,2700
matObj_Inv	ME	eModE	188	141	126	205	0,0000
matObj_Inv	eModE	LmodE	126	205	29	74	0,0774
matPP_Inv	OE	ME	283	310	911	2274	0,0000
matPP_Inv	ME	eModE	911	2274	409	3619	0,0000
matPP_Inv	eModE	LmodE	409	3619	89	2342	0,0000
matAdv_Inv	OE	ME	1191	764	1028	1394	0,0000
matAdv_Inv	ME	eModE	1028	1394	484	2091	0,0000
matAdv_Inv	eModE	LmodE	484	2091	78	834	0,0000

14.3.2 Chain-starting PPs in main clauses

See section 7.4, Figure 15.

	P1	P2	P1#	P1-rest#	P2#	P2-rest#	p-value
matS+V+PPnew	OE	ME	125	110	167	96	0,0165
matS+V+PPnew	ME	eModE	167	96	239	120	0,9910
matS+V+PPnew	eModE	LmodE	239	120	427	131	0,2883

14.3.3 New and chain-starting PPs found in main clauses and subclauses

See section 7.4, Figure 16.

	P1	P2	P1#	P1-rest#	P2#	P2-rest#	p-value
strictly new PPs	OE	ME	100	312	126	277	0,0285
strictly new PPs	ME	eModE	126	277	235	516	1,0000
strictly new PPs	eModE	LmodE	235	516	188	340	0,1166
chain-starting PPs	OE	ME	181	146	222	121	0,0165
chain-starting PPs	ME	eModE	222	121	430	234	0,9910
chain-starting PPs	eModE	LmodE	430	234	292	137	0,2883

14.3.4 New subject presentational focus per chainlength category

See section 8.4.1, Figure 18.

	P1	P2	P1#	P1-rest#	P2#	P2-rest#	p-value
zero	OE	ME	161	27	419	78	0,7224
zero	ME	eModE	419	78	426	94	0,3167
zero	eModE	LmodE	426	94	781	196	0,3725
small	OE	ME	10	178	29	468	0,8558
small	ME	eModE	29	468	43	477	0,1430
small	eModE	LmodE	43	477	111	866	0,0613
medium	OE	ME	11	177	26	471	0,7091
medium	ME	eModE	26	471	36	484	0,2950
medium	eModE	LmodE	36	484	68	909	1,0000
large	OE	ME	6	182	23	474	0,5251
large	ME	eModE	23	474	15	505	0,1853
large	eModE	LmodE	15	505	17	960	0,1877

14.3.5 New subject presentational focus per clause type

See section 8.4.2, Figure 19.

	P1	P2	P1#	P1-rest#	P2#	P2-rest#	p-value
Init	OE	ME	116	72	375	122	0,0006
Init	ME	eModE	375	122	389	131	0,8280
Init	eModE	LmodE	389	131	794	183	0,0041
PreV	OE	ME	15	173	43	454	0,8782
PreV	ME	eModE	43	454	88	432	0,0001
PreV	eModE	LmodE	88	432	170	807	0,8297
VS	OE	ME	55	133	59	438	0,0001
VS	ME	eModE	59	438	35	485	0,0049
VS	eModE	LmodE	35	485	13	964	0,0001

14.3.6 New subject presentational focus for medium and large subject chains

See section 8.4.2, Figure 20.

	P1	P2	P1#	P1-rest#	P2#	P2-rest#	p-value
Init	OE	ME	6	11	28	21	0,1619
Init	ME	eModE	28	21	29	22	1,0000
Init	eModE	LmodE	29	22	61	24	0,0928
VS	OE	ME	10	7	15	34	0,0476
VS	ME	eModE	15	34	6	45	0,0272
VS	eModE	LmodE	6	45	4	81	0,1754

14.3.7 The decline of subjects occurring after the finite verb in main clauses

See section 8.4.2, Figure 21.

	P1	P2	P1#	P1-rest#	P2#	P2-rest#	p-value
PstVfLinkedSbj	OE	ME	421	260	486	158	0,0001
PstVfLinkedSbj	ME	eModE	486	158	596	53	0,0001
PstVfLinkedSbj	eModE	LmodE	596	53	1360	20	0,0001

14.3.8 Postverbal presentational focus with syntactic subjects versus expletives

See section 8.4.2, Figure 22.

	P1	P2	P1#	P1-rest#	P2#	P2-rest#	p-value
NoExpletiveSbj	OE	ME	53	207	46	118	0,0774
NoExpletiveSbj	ME	eModE	46	118	24	38	0,1469
NoExpletiveSbj	eModE	LmodE	24	38	20	45	0,3587
WithExpletiveSbj	OE	ME	0	260	6	158	0,0032
WithExpletiveSbj	ME	eModE	6	158	9	53	0,0063
WithExpletiveSbj	eModE	LmodE	9	53	33	20	0,0001

14.3.9 Main clause subjects that occur after the finite verb and that are linked

See section 8.4.4, Figure 23.

	P1	P2	P1#	P1-rest#	P2#	P2-rest#	p-value
PstVfLinkedSbj	OE	ME	213	47	113	51	0,0030
PstVfLinkedSbj	ME	eModE	113	51	29	33	0,0032
PstVfLinkedSbj	eModE	LmodE	29	33	12	41	0,0107

14.3.10 Unanchored non-quantified subjects occurring after the finite verb

See section 8.4.4, Figure 24.

	P1	P2	P1#	P1-rest#	P2#	P2-rest#	p-value
UnanchNewSbj	OE	ME	12	6	21	30	0,0988
UnanchNewSbj	ME	eModE	21	30	6	33	0,0105
UnanchNewSbj	eModE	LmodE	6	33	5	210	0,0023

14.3.11 NPs and PPs modified by a focus adverb

See section 9.2.3, Figure 25.

	P1	P2	P1#	P1-rest#	P2#	P2-rest#	p-value
Contr+Emph-PreV	O12	O34	50	5	70	51	0,0001
Contr+Emph-PreV	O34	M1	70	51	36	83	0,0001
Contr+Emph-PreV	M1	M2	36	83	6	16	0,8957
Contr+Emph-PreV	M2	M3	6	16	74	119	0,3595
Contr+Emph-PreV	M3	M4	74	119	49	120	0,0750
Contr+Emph-PreV	M4	E1	49	120	54	190	0,1326
Contr+Emph-PreV	E1	E2	54	190	70	214	0,5372
Contr+Emph-PreV	E2	E3	70	214	67	201	1,0000
Contr+Emph-PreV	E3	B1	67	201	38	93	0,3992
Contr+Emph-PreV	B1	B2	38	93	61	133	0,7127
Contr+Emph-PreV	B2	B3	61	133	69	124	0,3905

14.3.12 Postverbal subject location in main clauses with the verb *have*

See section 9.2.3, Figure 26.

	P1	P2	P1#	P1-rest#	P2#	P2-rest#	p-value
Have_Vf-Sbj	OE	ME	344	1833	233	1844	0,0001
Have_Vf-Sbj	ME	eModE	233	1844	84	2775	0,0001
Have_Vf-Sbj	eModE	LmodE	84	2775	10	1782	0,0001

14.3.13 Preverbal noun phrases with local contrast

See section 9.5.2, Figure 27.

	P1	P2	P1#	P1-rest#	P2#	P2-rest#	p-value
cfeNP_Local	OE	ME	10	7	4	7	0,4401
cfeNP_Local	ME	eModE	4	7	62	18	0,0085
cfeNP_Local	eModE	LmodE	62	18	42	5	0,1025

14.3.14 The position of CLD resumptive demonstrative pronouns

See section 9.5.2, Table 34.

	P1	P2	P1#	P1-rest#	P2#	P2-rest#	p-value
Sbj_PreVf	OE	ME	417	228	134	29	0,0000
Sbj_PreVf	ME	eModE	134	29	17	23	0,0000
Sbj_PreVf	eModE	LmodE	17	23	19	7	0,0225
Sbj_PostVf	OE	ME	48	587	1	162	0,0000
Sbj_PostVf	ME	eModE	1	162	0	40	1,0000
Sbj_PostVf	eModE	LmodE	0	40	0	26	1,0000
Obj_PreVf	OE	ME	142	503	28	135	0,1973
Obj_PreVf	ME	eModE	28	135	23	17	0,0000
Obj_PreVf	eModE	LmodE	23	17	7	19	0,0225
Obj_PostVf	OE	ME	28	617	0	163	0,0029
Obj_PostVf	ME	eModE	0	163	0	40	1,0000
Obj_PostVf	eModE	LmodE	0	40	0	26	1,0000

14.3.15 Syntactic category of the clefted constituent

See section 12.3.1, Figure 38.

	P1	P2	P1#	P1-rest#	P2#	P2-rest#	p-value
Adjunct	OE	ME	61	41	35	87	0,0000
Adjunct	ME	eModE	35	87	72	175	1,0000
Adjunct	eModE	LmodE	72	175	166	162	0,0000
NonArgNP	OE	ME	19	83	4	118	0,0002
NonArgNP	ME	eModE	4	118	11	236	0,7813
NonArgNP	eModE	LmodE	11	236	4	324	0,0309
Object	OE	ME	4	98	19	103	0,0040
Object	ME	eModE	19	103	23	224	0,0829
Object	eModE	LmodE	23	224	14	314	0,0165
PPobj	OE	ME	1	101	3	119	0,6277
PPobj	ME	eModE	3	119	5	242	0,7227
PPobj	eModE	LmodE	5	242	2	326	0,1454
Subject	OE	ME	17	85	61	61	0,0000
Subject	ME	eModE	61	61	136	111	0,3765
Subject	eModE	LmodE	136	111	142	186	0,0055

14.3.16 Clefted constituents preceding the copula

See section 12.3.1, Figure 39.

	P1	P2	P1#	P1-rest#	P2#	P2-rest#	p-value
Preceding	OE	ME	15	79	58	45	0,0000
Preceding	ME	eModE	58	45	43	137	0,0000
Preceding	eModE	LmodE	43	137	36	248	0,0023

14.3.17 Information status of the clefted constituent

See section 12.3.1, Figure 40.

	P1	P2	P1#	P1-rest#	P2#	P2-rest#	p-value
Assumed	OE	ME	0	102	3	119	0,2526
Assumed	ME	eModE	3	119	3	244	0,4015
Assumed	eModE	LmodE	3	244	9	319	0,2493
Identity	OE	ME	22	80	54	68	0,0004
Identity	ME	eModE	54	68	87	160	0,1108
Identity	eModE	LmodE	87	160	124	204	0,5417
Inferred	OE	ME	62	40	11	111	0,0000
Inferred	ME	eModE	11	111	24	223	1,0000
Inferred	eModE	LmodE	24	223	62	266	0,0021
New	OE	ME	18	84	54	68	0,0000
New	ME	eModE	54	68	128	119	0,1852
New	eModE	LmodE	128	119	133	195	0,0087

14.3.18 Information status of the cleft clause

See section 12.3.1, Figure 41.

	P1	P2	P1#	P1-rest#	P2#	P2-rest#	p-value
Assumed	OE	ME	2	100	5	117	0,2526
Assumed	ME	eModE	5	117	6	241	0,4015
Assumed	eModE	LmodE	6	241	13	315	0,2493
Inferred	OE	ME	6	96	10	112	0,2526
Inferred	ME	eModE	10	112	43	204	0,4015
Inferred	eModE	LmodE	43	204	84	244	0,2493
Known	OE	ME	11	91	33	89	0,2526
Known	ME	eModE	33	89	114	133	0,4015
Known	eModE	LmodE	114	133	123	205	0,2493
New	OE	ME	83	19	74	48	0,2526
New	ME	eModE	74	48	84	163	0,4015
New	eModE	LmodE	84	163	108	220	0,2493

14.3.19 Combined information status of clefted constituent and cleft clause

See section 12.3.1, Figure 42.

	P1	P2	P1#	P1-rest#	P2#	P2-rest#	p-value
TopCom	OE	ME	72	30	40	82	0,0000
TopCom	ME	eModE	40	82	52	195	0,0156
TopCom	eModE	LmodE	52	195	75	253	0,6135
ComTop	OE	ME	4	98	11	111	0,1802
ComTop	ME	eModE	11	111	43	204	0,0410
ComTop	eModE	LmodE	43	204	63	265	0,6641
ComCom	OE	ME	7	95	14	108	0,2599
ComCom	ME	eModE	14	108	18	229	0,2372
ComCom	eModE	LmodE	18	229	29	299	0,5414
TopTop	OE	ME	11	91	25	97	0,0669
TopTop	ME	eModE	25	97	66	181	0,2021
TopTop	eModE	LmodE	66	181	118	210	0,0191
Wh	OE	ME	61	41	32	90	0,0004
Wh	ME	eModE	32	90	68	179	0,9010
Wh	eModE	LmodE	68	179	43	285	0,0000

14.3.20 Information structure status of the cleft

See section 12.3.1, Figure 43.

	P1	P2	P1#	P1-rest#	P2#	P2-rest#	p-value
TopCom	OE	ME	61	41	18	104	0,0000
TopCom	ME	eModE	18	104	30	217	0,5123
TopCom	eModE	LmodE	30	217	38	290	0,8964
EmphAll	OE	ME	29	73	85	37	0,0000
EmphAll	ME	eModE	85	37	178	69	0,6273
EmphAll	eModE	LmodE	178	69	235	93	0,9257
Rest	OE	ME	12	90	19	103	0,4433
Emphatic Cleft Types	ME	eModE	19	103	39	208	1,0000
Emphatic Cleft Types	eModE	LmodE	39	208	55	273	0,8200

14.3.21 Emphatic cleft types

See section 12.3.1, Figure 44.

	P1	P2	P1#	P1-rest#	P2#	P2-rest#	p-value
Contrast-Context	OE	ME	10	92	40	82	0,2825
Contrast-Context	ME	eModE	40	82	72	175	0,3513
Contrast-Context	eModE	LmodE	72	175	115	213	0,0907
Contrast-Adv	OE	ME	0	102	0	122	1,0000
Contrast-Adv	ME	eModE	0	122	16	231	0,0019
Contrast-Adv	eModE	LmodE	16	231	38	290	0,0387
Contrast-Neg	OE	ME	2	100	5	117	1,0000
Contrast-Neg	ME	eModE	5	117	12	235	1,0000
Contrast-Neg	eModE	LmodE	12	235	16	312	1,0000
EmphProm	OE	ME	9	93	8	114	0,0123
EmphProm	ME	eModE	8	114	10	237	0,2981
EmphProm	eModE	LmodE	10	237	23	305	0,1440
Wh	OE	ME	8	94	32	90	0,3745
Wh	ME	eModE	32	90	68	179	1,0000
Wh	eModE	LmodE	68	179	43	285	0,0000

14.3.22 Subject-auxiliary inversion for clause-initial focused PPs

See section 12.3.1, Figure 46.

	P1	P2	P1#	P1-rest#	P2#	P2-rest#	p-value
PPfoc-S-V	OE	ME	8	3	63	29	1,0000
PPfoc-S-V	ME	eModE	63	29	39	11	0,2483
PPfoc-S-V	eModE	LmodE	39	11	43	5	0,1722

¹ The edition used for CorpusStudio is Microsoft Visual Basic 2008 Express.

² The actual definition of a finiteverb is much more complex, and can be found on the files that support working with [CorpusStudio](#).

³ The function `ru:matches()` that is used to match the label with the pattern is hard-coded in the CorpusStudio program. Its first argument is a string (a word, a label—any piece of text), and its second argument a series of patterns divided by vertical bars to which the function tries to match the string.

⁴ The approach sketched here is a simplification of reality. When looking at VO versus OV word order in subclauses there are much more factors that need to be taken into account, such as the type of verb and the heaviness of the constituents (Fischer et al., 2000, Pintzuk, 1996, Pintzuk, 2002, Pintzuk and Taylor, 2006).

⁵ The reader should realize that the corpus research project aimed at finding subclause SVO versus SOV is a simplification of reality—see footnote 13.

⁶ This example provides a good excuse for me to warn corpus researchers *against* the use of complements. If we would have a sentence that contains a number of subclauses, some of

which have the object precede the finite verb (SOV), and some follow it (SVO), then these will not be in the input to line 3. The reason for this can be illustrated by following the route of our multiple-subclause order sentence. Line 2 captures this sentence as a whole (including all its subclauses) and puts it in the *output* of line 2, and not in the *complement* of it. Line 3 takes as input the *complement* of line 2, so it does not even look at our sentence with the multiple-subclause orders. It misses out on this opportunity! In this case, it would have been better to let line 3 have the output of line 1 as input, and write a separate query for line 4.

⁷ The labelled bracketing format files used in CorpusSearch2 projects are currently not executed sentence-by-sentence, since that would cost more time (that is: to correctly split a file into sentences) than it would yield (in terms of sentences that would not need to be considered in subsequent steps, since they did not pass the first step).

⁸ The columns in Table 49 are a subset of all the time-periods defined for the parsed English corpora. See the period definition file at the CorpusStudio [homepage](#) for a full definition as to these periods.

Index

- algorithm
 - Cesax, 178, 184, 186, 187, 188, 191, 192, 194, 196, 197, 392
 - COT, 164, 178, 183, 188, 189, 191, 196, 198
- cleft
 - cleft clause, 282
 - clefted constituent, 282
 - construction, 283
- cleft feature
 - ClauseStatus, 344, 345, 348
 - CleftedCat, 344, 346, 351, 353
 - CleftedCoref, 344, 345, 347, 348
 - CleftedType, 344, 345, 347
 - CleftType, 344, 345, 346
 - FocusType, 344, 345, 348, 349, 365, 367, 374
 - information-structure status, 364, 365, 367, 442
- clefts
 - adjunct, 281, 284, 285, 286, 290, 292, 299, 305, 327, 329, 331, 342, 344, 359, 374
 - avoidance, 300, 301, 306, 389, 390
 - coindexing, 293, 294, 295, 296, 307, 325, 342
 - comment-clause, 303
 - contrastive, 299, 355
 - database, 327, 328, 329, 330, 331, 344, 346, 349, 353, 357, 362, 364, 371
 - definition, 282, 283, 286, 292, 293, 295, 323
 - diagnostics, 286, 293, 295, 325, 331, 332, 354
 - emphatic, 366, 367, 368, 369, 370, 373, 443
 - informative-presupposition, 274, 285, 299, 302, 342, 343, 344, 348, 362, 364, 373, 374, 397
 - predicational, 287, 292
 - reversed *wh*-cleft, 38, 251, 272, 273, 274, 275, 276, 277, 279, 303
 - specificational, 347
 - stressed-focus, 343, 344, 348, 364, 373
 - summative, 303, 304, 305, 334, 336
 - topic launching, 303, 304, 334
 - topic linking, 303, 304, 334, 336
 - topic-clause, 303
 - wh*-cleft, 34, 38, 239, 251, 272, 273, 274, 275, 276, 277, 278, 279, 297, 303, 309, 311, 323, 337, 399, 435
- complement clause, 83, 87, 100, 101, 130, 262, 276, 288, 295, 350, 353, 354, 425, 426
- compositionality, 14, 171, 391, 392, 393
- constraint
 - Cesax constraints, 141, 152, 153, 173, 182, 183, 189, 190, 191, 192, 193, 194, 195, 196, 429
- copula
 - copula construction, 37, 38, 39, 51, 126, 161, 164, 165, 166, 167, 168, 169, 170, 173, 187, 234, 282, 283, 286, 287, 292, 294, 295, 296, 307, 323, 324, 325, 332, 342, 346, 347, 353, 361, 384, 388, 391, 392, 429, 435
 - equative construction, 37, 38, 39, 51, 149, 150, 200, 261, 273, 274, 275, 277, 286, 287, 323, 393
- coreference, 198, 28, 137-206, 339-401
 - chain, 53, 170, 173, 190, 205, 206, 207, 211, 214, 215, 216, 217, 218, 220, 228, 232, 244, 381
 - referentiality, 267, 275, 381, 385, 388, 391
- corpulect
 - corpulect, 14
 - corpulect distribution, 17
- corpus
 - approach, 133, 171, 244
- diagnostic, 308, 235-384
- expletive, 61, 71, 119, 120, 122, 128, 131, 150, 172, 224, 234, 235, 237, 245, 246, 249, 257, 291, 353, 381, 382, 386, 403, 431, 438
 - transitive, 131, 406
- fields
 - Middle Field, 98, 114
 - PostField, 61, 69
 - PreField, 67, 68, 69, 70
- focus
 - constituent focus, 35
 - contrastive left dislocation, 251, 268, 269, 270, 278, 279, 383, 410, 434, 440
 - emphatic, 12, 13, 37, 40, 41, 100, 102, 239, 241, 251, 254, 255, 260, 264, 265, 277, 279, 334, 338, 349, 365, 366, 367, 368,

- 369, 370, 372, 373, 374, 375, 383, 387, 389, 409, 431, 442, 443
- emphatic pronouns, 40, 251, 252, 264, 265, 277, 383
- exhaustive identification, 298, 341
- local contrast, 239, 240, 251, 261, 262, 263, 278, 279, 383, 433, 434, 439
- narrow focus, 4, 13, 19, 290, 312, 321, 322, 379
- presentational focus, 44
- thetic articulation, 35, 42, 102, 106, 107, 121, 168, 199, 223, 238, 268, 385, 393
- topic-comment, 19, 35, 36, 39, 42, 45, 46, 47, 48, 49, 51, 67, 69, 83, 99, 103, 104, 105, 106, 107, 125, 128, 161, 166, 167, 199, 237, 243, 251, 265, 380, 382, 384, 385, 388, 393
- grammar**
- formal, 94
- generative, 308, 386, 412
- grammatical function, 5, 6, 48, 57, 59, 85, 379
- intonation, 312, 315, 316, 318, 322
- optimality theory, 308, 382, 386, 390, 401
- role and reference, 42, 382
- information structure**, 13, 14, 35, 36, 57, 70, 126, 134, 140, 151, 154, 158, 160, 164, 167, 171, 207, 211, 212, 217, 218, 223, 226, 253, 272, 281, 284, 296, 304, 305, 344, 357, 373, 392, 402, 403, 405, 406, 410, 449
- discourse-new, 108, 137, 143, 184, 186, 187, 198, 218, 274, 304, 305, 364
- discourse-old, 137, 305, 341
- established information, 8, 21, 35, 43, 44, 47, 49, 51, 52, 56, 65, 68, 86, 104, 106, 125, 126, 128, 136, 215, 233, 239, 241, 249, 266, 273, 274, 275, 381, 384
- hearer-new, 102, 103, 104, 120, 122, 123, 127, 128, 137
- hearer-old, 103, 137, 364
- newness, 33, 38, 42, 50, 102, 129, 200, 213, 214, 223, 224, 232, 233, 234, 236, 237, 238, 239, 240, 241, 244, 249, 250, 259, 260, 266, 267, 268, 275, 278, 362, 364, 381, 382, 390
- Principle of Natural Information Flow, 8, 43, 44, 47, 49, 52, 57, 68, 71, 101, 104, 105, 106, 108, 123, 124, 125, 127, 128, 131, 224, 249, 250, 253, 265, 266, 267, 268, 341, 380, 382, 384, 385, 386, 393
- referential point of departure, 46, 69, 73, 85, 89, 90, 93, 103, 104, 115, 116, 117, 127, 387
- unestablished information, 43, 47, 104, 125, 215, 380, 382, 384, 386
- language**
- Chechen, 4, 15, 17, 19, 20, 33, 34, 49, 190, 281, 291, 296, 297, 298, 306, 309, 311, 312, 313, 314, 316, 320, 322, 323, 324, 325, 326, 327, 328, 329, 331, 332, 333, 334, 335, 336, 337, 338, 339, 341, 373, 389, 402, 404, 407, 412, 449
- Dutch, 7, 132, 190, 269, 398, 413
- French, 6, 21, 40, 235, 258, 296, 297, 306, 307, 312, 399, 406
- German, 7, 61, 70, 164, 275, 291, 299, 300, 372, 380, 397, 409
- Italian, 398
- Norwegian, 300, 301, 306, 334, 373, 401
- Russian, 5, 20, 43, 283, 291, 295, 296, 307, 338
- Swedish, 291, 300, 301, 306, 336, 373, 390, 404
- locative**, 44, 52, 105, 122, 129, 199, 224, 236, 237, 242, 245, 328, 329, 331, 398, 409
- mapping**, 68, 121, 122, 125, 126, 141, 164, 168, 183, 235, 388, 389, 406
- mind**
- mental entity, 25, 26, 28, 29, 30, 31, 38, 52, 102, 144, 145, 146, 147, 148, 149, 150, 151, 154, 155, 156, 157, 158, 160, 168, 172, 379, 391
- mental model, 26, 27, 28, 30, 33, 36, 38, 39, 41, 42, 45, 51, 52, 86, 102, 122, 124, 133, 134, 144, 147, 148, 149, 150, 161, 223, 224, 225, 236, 242, 261, 265, 386, 387, 391
- situation model, 26, 27, 28, 29, 30, 31, 144, 145, 146, 147, 148, 150, 151, 154, 157, 158, 160, 172, 379
- morphology**, 58, 173, 307, 372, 402
- negation**, 12, 37, 39, 41, 61, 64, 65, 91, 113, 114, 130, 155, 156, 157, 158, 236, 237, 239, 242, 245, 251, 258, 259, 260, 261, 263, 264, 277, 290, 297, 298, 301, 306, 320, 327, 328, 349, 373, 383, 392, 410, 411
- positive negation, 40, 251, 260
- particle**, 13, 34, 39, 41, 43, 47, 89, 94, 130, 172, 239, 251, 253, 278, 290, 299, 300, 307, 316, 333, 339, 349, 369, 370, 371, 372, 387
- Pentaset, 143, 144, 145, 147, 148, 151, 152, 153, 154, 155, 158, 160,

- 161, 164, 165, 170, 171, 172, 177, 195, 218, 224, 226, 241
- pragmatics, 6, 8, 33, 35, 49, 51, 56, 63, 67, 68, 69, 71, 72, 75, 86, 91, 93, 94, 97, 99, 100, 105, 111, 112, 119, 120, 121, 124, 125, 127, 129, 399, 401, 402, 409, 410, 412
- principle
 - demarcation, 249, 250, 278, 383, 384
 - placement, 250, 278, 383, 384
- processing, 5, 6, 19, 25, 26, 28, 29, 30, 33, 51, 57, 124, 143, 184, 187, 201, 202, 203, 210, 213, 219, 326, 379, 409, 423, 424, 427, 449
 - load, 6, 57, 379
 - model, 25, 29, 33, 51, 143
- psycholinguistics, 13, 26, 379, 391
- reference type
 - CrossSpeech, 141, 152, 153, 192, 193, 195, 429
 - linked, 144, 237
 - NewVar, 195, 216, 429, 431
 - unanchored new, 241
 - unlinked, 144, 237
- referential state, 39, 133, 171, 172, 173
 - assumed, 144
 - Identity, 144
 - Inert, 144
 - Inferred, 144
 - New, 144
- relative clause, 37, 87, 101, 168, 173, 251, 257, 266, 267, 268, 273, 279, 282, 284, 285, 286, 287, 288, 289, 290, 291, 293, 294, 295, 296, 307, 324, 325, 327, 328, 332, 338, 350, 353, 356, 358, 373, 374
- research question, 4, 13, 14, 126, 177, 199, 217, 218, 244, 245, 249, 278, 281, 327, 379
- resumptive, 94, 208, 251, 269, 270, 271, 278, 279, 434, 440
- slots
 - Core, 59, 60, 61, 62, 64, 68, 74, 75, 84, 85, 93, 97, 98, 100, 105, 112, 113, 121, 125, 127, 131, 224, 229, 231, 232, 244, 278, 380, 383
 - CoreArgEst, 131
 - CoreArgNest, 131
 - PostCore, 59, 60, 61, 62, 68, 71, 74, 75, 84, 85, 100, 102, 105, 106, 107, 112, 113, 118, 121, 125, 127, 131, 224, 229, 231, 232, 244, 278, 380, 383, 386
 - PreCore, 59, 60, 61, 62, 64, 66, 67, 71, 74, 77, 84, 85, 93, 98, 100, 112, 113, 121, 125, 224, 229, 231, 245, 249, 250, 257, 277, 278, 281, 341, 367, 372, 380, 381, 383, 385, 386, 387, 388
 - PreSbj, 59, 60, 61, 74, 75, 77, 84, 85, 86, 90, 93, 94, 104, 108
 - Sbj, 61
 - Vb1, 59, 60, 61, 62, 64, 69, 74, 75, 77, 84, 85, 90, 94, 98, 99, 100, 105, 108, 112, 113, 117, 121, 122, 125, 126, 127, 229, 234, 244, 252, 278, 381
 - Vb2, 59, 60, 61, 62, 68, 69, 74, 75, 77, 84, 85, 94, 97, 98, 105, 106, 108, 112, 113, 121, 125, 127, 131, 229, 244, 247, 252, 278, 381
- software tools
 - Alpino, 202, 219, 404
 - CorpusSearch, 180, 181, 201, 202, 205, 218, 408, 424, 426, 444
 - CorpusStudio, 10, 12, 21, 200, 204, 205, 206, 207, 208, 209, 210, 211, 212, 213, 215, 217, 218, 219, 220, 227, 254, 326, 329, 349, 351, 381, 392, 405, 417, 423, 424, 426, 427, 428, 429, 443, 444
 - DtSearch, 202, 219
 - Saxon, 203, 409
 - Tgrep2, 202, 409, 426
 - TigerSearch, 202
- split constituent, 101, 102, 127, 128, 252, 263, 266, 267, 268, 269, 383, 385, 386
- syntax
 - English syntax, 3, 4, 8, 14, 63, 127, 223, 226, 249, 250, 277, 372, 380
 - grammatical subject, 36, 52, 107, 124, 257, 324, 325, 389
 - logical subject, 122, 128, 234, 235, 245, 297, 301, 323, 324, 382, 386, 390
 - subject-auxiliary inversion, 68, 71, 301, 372, 443
 - verb-second, 7, 8, 56, 63, 64, 65, 71, 72, 91, 94, 98, 130, 231, 250, 301, 306, 341, 345, 361, 367, 371, 373, 376, 380, 383, 388, 402, 406, 410, 430, 431, 435
 - verb-second loss, 65, 224, 234, 244, 245, 249, 250, 277, 278, 341, 357, 380, 382, 411
- text
 - charting, 55, 69, 72, 73, 75, 76, 77, 83, 85, 90, 91, 93, 100, 112, 120, 127, 129, 240, 379, 380, 391
 - syntactically parsed, 160, 171, 199, 202, 218

- text-organization, 4, 20, 51, 72, 82, 83, 85, 86, 87, 88, 90, 93, 99, 111, 112, 114, 120, 126, 127, 332, 334, 336, 337, 341, 357, 374, 384, 392
- text structuring, 129, 341
- treebank, 179, 180, 181, 189, 195, 196, 201, 202, 218, 398, 404, 407
- word order
 - AP-correlated, 83, 88, 91, 92, 95, 96, 99, 114, 117, 127
 - AP-initial, 83, 92, 93, 96, 112, 115, 117, 127
 - apposition, 40, 101, 102, 103, 104, 122, 123, 124, 128, 233, 237, 251, 252, 265, 266, 278, 383
 - clause-final, 12, 44, 55, 56, 68, 71, 114, 118, 129, 133, 229, 240, 245, 249, 256, 257, 258, 267, 271, 276, 278, 312, 384, 385
 - clause-initial, 8, 12, 13, 21, 36, 45, 46, 49, 55, 56, 58, 64, 65, 66, 68, 70, 89, 91, 93, 94, 95, 96, 97, 101, 113, 129, 230, 234, 244, 250, 252, 256, 257, 258, 270, 271, 276, 278, 281, 369, 370, 372, 373, 383, 384, 385, 386, 390, 411, 443
 - conjunct clause, 91, 97, 98, 99, 100, 101, 118, 130, 397
 - default, 5, 86, 97
 - do-support, 64, 114, 386
 - ImmPostVf, 252, 256, 263, 270, 279
 - left dislocation, 34, 76, 94, 269, 270, 271
 - postposing, 74, 76, 84
 - postverbal, 61, 128, 233, 234, 235, 236, 237, 240, 242, 244, 245, 246, 252, 255, 256, 258, 382, 383, 391, 407
 - PostVf, 228, 229, 236, 246, 252, 253, 256, 263, 279, 440
 - PostVnonF, 229, 236, 279
 - preposing, 34, 76, 107, 108, 124, 125, 210, 373, 406, 412
 - PreVf, 252, 256, 263, 270, 279, 440
 - right dislocation, 34, 76, 343
 - T-correlated, 83, 88, 91, 92, 94, 95, 96, 97, 99, 114, 117, 127, 130, 231, 241
 - then-initial, 121
 - T-initial, 65, 83, 87, 88, 90, 91, 92, 93, 94, 99, 106, 118, 121, 122, 127, 130, 232
- xml
 - Xpath, 181, 183, 219
 - Xquery, 181, 183, 200, 202, 203, 204, 205, 207, 208, 210, 211, 213, 215, 217, 218, 219, 228, 245, 254, 255, 328, 350, 351, 354, 409, 424, 425, 427, 429, 431



Curriculum Vitae

Erwin Komen was born in Utrecht, the Netherlands, on September 8, 1960. He obtained a High school diploma in 1976 and an engineer's degree (HTS) in precision mechanics in 1981. In order to pursue his interest in research, he enrolled in the Applied Physics faculty at the Delft University of Technology, where he received a master's degree for his work on robot vision in 1986. He was invited to enter a four year research project at the pattern recognition group of the Delft University of Technology, which, under the supervision of Bob Duin and Ted Young, led to his PhD thesis in 1990 on parallel computer architectures for low-level image processing, and on theoretical work aspects of recursive neighbourhood operations. He then made a career switch to linguistics, spent several years in Russia with SIL-international and then turned to Leiden University in 2006, where he investigated focus in the Chechen language and obtained a master's degree in linguistics (with distinction) under the supervision of Lisa Cheng and Anniko Liptak. His career at the Radboud University Nijmegen started in 2007, when he became PhD-candidate at the English language department. With Ans van Kemenade and Bettelou Los as supervisors, he worked in an NWO-sponsored project on the relation between information structure and syntax, as seen from the diachronic development of English. His wider interests include Chechen linguistics and corpus development in general. He is currently working part-time as PostDoc at the Radboud University Nijmegen and as linguistic consultant for SIL-international.